Ekonomia

Modele wielopoziomowe

Wykorzystanie danych regionalnych w badaniach mikroekonomicznych i socjologicznych

Wojciech Grabowski

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Modele wielopoziomowe

Wykorzystanie danych regionalnych w badaniach mikroekonomicznych i socjologicznych

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Ekonomia

Modele wielopoziomowe

Wykorzystanie danych regionalnych w badaniach mikroekonomicznych i socjologicznych

Wojciech Grabowski

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO
Łódź 2019

Wojciech Grabowski – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Katedra Modeli i Prognoz Ekonometrycznych, 90-214 Łódź, ul. Rewolucji 1905 r. 37/39

RECENZENT
Jerzy Marzec

REDAKTOR INICJUJĄCY
Monika Borowczyk

REDAKCJA
Monika Poradecka

SKŁAD I ŁAMANIE
Mateusz Poradecki

KOREKTA TECHNICZNA
Leonora Gralka

PROJEKT OKŁADKI
Katarzyna Turkowska

Zdjęcie wykorzystane na okładce: © Depositphotos.com/leungchopan

Wydrukowano z gotowych materiałów dostarczonych do Wydawnictwa UŁ

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.08951.18.0.M

Ark. druk. 16,375

ISBN 978-83-8142-418-9
e-ISBN 978-83-8142-419-6

https://doi.org/10.18778/8142-418-9

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. (42) 665 58 63

Spis treści

Wstęp **9**

Notacja wykorzystywana w monografii **13**

1. Podstawowe modele wykorzystujące dane indywidualne **33**
1.1. Wprowadzenie 33
1.2. Modele dla ciągłej zmiennej zależnej 34
1.2.1. Klasyczny model regresji liniowej 34
1.2.2. Heteroskedastyczność składnika losowego. Metody estymacji parametrów w przypadku niestałej wariancji 35
1.2.3. Metoda regresji kwantylowej 37
1.2.4. Odporna estymacja parametrów modelu regresji. Estymator *M*. Estymator *S*. Estymator *MM* 38
1.3. Model dwumianowy (dychotomiczny) 42
1.4. Model wielomianowy (polichotomiczny) kategorii uporządkowanych 51
1.5. Model wielomianowy kategorii nieuporządkowanych 52
1.6. Model regresji rankingowej 57
1.7. Problem selekcji próby w modelach ekonometrycznych. Model Heckmana 58
1.8. Model licznikowy 60
1.9. Dwurównaniowy model probitowy 65
1.10. Wielorównaniowy model probitowy 67
1.11. Endogeniczny model probitowy 69
1.12. Podsumowanie 74

2. Dane regionalne wykorzystywane w badaniach ekonomicznych **75**
2.1. Podział administracyjny, statystyczny i historyczny Polski 75
2.2. Historyczno-kulturowe zróżnicowanie terytorium obecnej Rzeczpospolitej Polskiej 78
2.3. Źródła danych, które mogą być wykorzystywane w analizach regionalnych dla Polski 84
2.3.1. Bank Danych Lokalnych 85
2.3.2. Regional Innovation Scoreboard jako źródło informacji o poziomie innowacyjności regionów 91
2.3.3. Inne źródła danych regionalnych wykorzystywane w badaniach empirycznych 95

3. Liniowe modele wielopoziomowe **99**
3.1. Wprowadzenie 99
3.2. Zmienne regionalne i sekcyjne w modelach ekonometrycznych opartych na danych indywidualnych 100
3.3. Podstawowy wielopoziomowy model regresji. Estymacja parametrów i predykcja efektów losowych 102
3.4. Efekty krzyżowe w liniowych modelach wielopoziomowych 109

3.5. Wykorzystanie liniowego modelu wielopoziomowego uwzględniającego zmienne regionalne do badania czynników wpływających na wynagrodzenia w Polsce 111
3.5.1. Przegląd literatury z zakresu czynników wpływających na wynagrodzenia 111
3.5.2. Koncepcje SBTC i RBTC i ich wykorzystanie do analizy czynników wpływających na różnice między wynagrodzeniami przedstawicieli określonych zawodów 114
3.5.3. Dane dotyczące poziomów wynagrodzeń uzyskiwanych przez pracowników w polskich przedsiębiorstwach. Podział zawodów ze względu na umiejętności posiadane przez pracowników 117
3.5.4. Specyfikacja modelu ekonometrycznego wykorzystywanego do analizy czynników wpływających na wynagrodzenia w Polsce 120

4. Uogólnione liniowe modele wielopoziomowe 141
4.1. Postać uogólnionego liniowego modelu wielopoziomowego 141
4.2. Funkcja wiarygodności w uogólnionym liniowym modelu wielopoziomowym 142
4.3. Estymacja parametrów uogólnionych liniowych modeli wielopoziomowych za pomocą metod aproksymacyjnych 149
4.4. Estymacja parametrów uogólnionych liniowych modeli wielopoziomowych za pomocą metod symulacyjnych 165
4.5. Problem selekcji próby w modelach wielopoziomowych. Estymacja parametrów wielorównaniowych modeli probitowych z efektami losowymi 170
4.6. Wykorzystanie wielopoziomowego modelu zmiennych dyskretnych do analizy zależności między wykorzystywaniem technologii informacyjnych i komunikacyjnych, innowacyjnością a produktywnością 176
4.6.1. Przegląd literatury z zakresu czynników wpływających na innowacyjność firm 176
4.6.2. Model CDM rozszerzony o wykorzystanie TIiK oraz uwzględniający czynniki regionalne 179
4.6.2.1. Dane oraz próba badawcza 179
4.6.2.2. Specyfikacja modelu ekonometrycznego 189
4.6.2.3. Wyniki estymacji i interpretacja 193

5. Wielopoziomowy polichotomiczny model logitowy. Wielopoziomowy model regresji rankingowej 215
5.1. Wprowadzenie 215
5.2. Wielopoziomowy model wielomianowy logitowy. Wielopoziomowy model regresji rankingowej 216
5.3. Wykorzystanie wielopoziomowego polichotomicznego nieuporządkowanego modelu logitowego do analizy czynników wpływających na sposób reakcji wobec zaistnienia problemu prawnego 221
5.3.1. Czynniki wpływające na sposób reakcji wobec wystąpienia problemu prawnego – przegląd literatury 221
5.3.2. Estymacja parametrów wielopoziomowego, nieuporządkowanego, polichotomicznego modelu logitowego na podstawie danych pochodzących z badania dla Polski przeprowadzonego przez Instytut Spraw Publicznych w Warszawie 223

Zakończenie 241

Bibliografia 243

Abstract **253**

Spis rysunków **257**

Spis tabel **259**

Od Redakcji **261**

Wstęp

W badaniach ekonomicznych i społecznych coraz więcej uwagi poświęca się analizie zależności występujących na poziomie indywidualnym. Powszechne są badania wykorzystujące między innymi dane dotyczące decyzji podejmowanych w przedsiębiorstwach (np. serie badań pt. *Community Innovation Survey* przeprowadzane przez urzędy statystyczne krajów Unii Europejskiej), aktywności ekonomicznej ludności (np. serie badań aktywności ekonomicznej ludności przeprowadzanych w różnych krajach Unii Europejskiej) czy też wynagrodzeń uzyskiwanych przez pracowników (badania struktury wynagrodzeń). Oprócz wymienionych wyżej oraz innych badań cyklicznych przeprowadzane są badania jednorazowe, w których jednostkami są firmy, pracownicy, gospodarstwa domowe, respondenci itp. Niektóre dane pochodzące z tych badań są publicznie dostępne, inne zaś mogą zostać zakupione lub pozyskane przez instytucje naukowo-badawcze. Zwiększa się zatem pole do zastosowań metod mikroekonometrycznych.

W badaniach mikroekonomicznych i społecznych często ignorowana jest rola kontekstu. Przyjmuje się założenia, że zależności występują między jednostkami, a lokalizacja gospodarstwa domowego czy firmy nie ma wpływu na proces podejmowania decyzji. Ewentualne różnice między zachowaniami respondentów mieszkających w innych regionach czy też różnice w procesie decyzyjnym firm znajdujących się w różnych sekcjach PKD traktuje się jako ustalone. Oznacza to zatem, że oszacowania parametrów przy odpowiednich zmiennych zero-jedynkowych odzwierciedlają te różnice. Istnieją jednak metody badania zależności między cechami przedsiębiorstw czy gospodarstw domowych umożliwiające analizę losowych różnic między jednostkami należącymi do innych sekcji czy regionów. Metody te umożliwiają również analizę różniącego się w poszczególnych grupach (którymi mogą być odpowiednie sekcje lub regiony) wpływu określonych cech firm czy gospodarstw domowych na podejmowane przez nie decyzje. Analizowane metody wykorzystuje się podczas estymacji parametrów modeli wielopoziomowych.

Niniejsza monografia poświęcona jest aspektom teoretycznym oraz badaniom empirycznym wykorzystującym modele wielopoziomowe. Prezentowane są zastosowania omawianych modeli w badaniach mikroekonomicznych (dotyczących wynagrodzeń pracowników oraz postaw innowacyjnych przedsiębiorstw wykorzystujących technologie informatyczne i komunikacyjne) oraz socjologicznych (związanych z zagadnieniami z zakresu socjologii prawa). Idea tych badań polega

na dodatkowym uwzględnieniu czynników kontekstowych (zwłaszcza związanych z przynależnością firmy, pracownika czy respondenta do regionu) w modelach wykorzystujących dane indywidualne. Jednocześnie analizowana jest rola poszczególnych czynników w wyjaśnieniu zmienności zmiennej zależnej oraz wskazywane są różnice między wynikami uzyskanymi dla modelu pełnego a tymi otrzymanymi w przypadku nieuwzględnienia zmiennych kontekstowych, czyli związanych z lokalizacją jednostki.

W badaniach empirycznych omawianych w niniejszej monografii wykorzystywane są przede wszystkim dane indywidualne uzyskane podczas realizacji grantów Narodowego Centrum Nauki (zakończonych i będących w trakcie realizacji), w których uczestniczył autor. Dane dotyczące wynagrodzeń pracowników pochodzą z baz związanych z badaniami struktury wynagrodzeń, zakupionymi podczas realizacji grantu pt. „Polaryzacja polskiego rynku pracy w kontekście zmiany technologicznej" o numerze 2016/23/B/HS4/00334. Dane związane z działalnością innowacyjną przedsiębiorstw zostały uzyskane na podstawie badania ankietowego przeprowadzonego w 2015 roku podczas realizacji grantu pt. „Wpływ technologii informacyjnych i telekomunikacyjnych na produktywność – analiza mikro- i makroekonomiczna" o numerze 2013/11/B/HS4/00661. Dane dotyczące faktu zaistnienia problemu prawnego oraz sposobu reakcji na niego pochodzą z badania przeprowadzonego w 2012 roku przez Instytut Spraw Publicznych w Warszawie pt. „Korzystający i niekorzystający z poradnictwa prawnego i obywatelskiego". Dane te zostały zakupione podczas realizacji grantu Narodowego Centrum Nauki pt. „Nieodpłatna pomoc prawna w Polsce z perspektywy ekonomicznej analizy prawa. Stan obecny i rekomendowany" o numerze 2012/07/B/HS4/02994.

W przypadku każdego z tych trzech badań oprócz informacji indywidualnych wykorzystywane są także dane kontekstowe (np. związane z lokalizacją, przynależnością firmy do sekcji PKD czy też przynależnością pracownika do grupy zawodowej). Należy jednak podkreślić, że prezentowane wyniki są komplementarne względem rezultatów uzyskanych w innych pracach autora wykorzystujących te właśnie dane (por. m.in. Arendt, Grabowski, 2017; 2018; Florczak, Grabowski, 2017; 2018a; 2018b; 2018c; Szczygielski, Grabowski, Woodward, 2017; Szczygielski, Grabowski, Pamukcu, Tandogan, 2017). Różnica polega na dodatkowym uwzględnieniu czynników kontekstowych w modelach mikroekonometrycznych.

Niniejsza monografia składa się z pięciu rozdziałów. W rozdziale pierwszym prezentowane są metody wykorzystywane do analizy zależności na poziomie indywidualnym, ale bez uwzględniania zmiennych kontekstowych. W rozdziale drugim omawiane są różnice w poziomie rozwoju ekonomiczno-społecznego między polskimi regionami historycznymi i administracyjnymi. Jednocześnie prezentowane są bazy danych regionalnych, z których część wykorzystywana jest

w badaniach empirycznych uwzględnionych w monografii. Rozdział trzeci zawiera opis metody estymacji parametrów i predykcji efektów losowych w modelach wielopoziomowych z ciągłą zmienną zależną. Oprócz tego prezentowane są rezultaty badania empirycznego poświęconego determinantom zróżnicowania wynagrodzeń oraz testowaniu hipotezy o występowaniu polaryzacji na polskim rynku pracy. W rozdziale czwartym prezentowane są uogólnione liniowe modele wielopoziomowe. Szczegółowo omawiane są metody estymacji parametrów tych modeli. Jednocześnie prezentowane są wyniki badania empirycznego mającego na celu identyfikację indywidualnych i regionalnych czynników kształtujących decyzje innowacyjne przedsiębiorstw. W rozdziale piątym rozważane są modele wielopoziomowe dla przypadku wielomianowej nieuporządkowanej oraz rankingowej zmiennej zależnej. Dodatkowo prezentowane są wyniki badania empirycznego mającego na celu identyfikację czynników wpływających na prawdopodobieństwo doświadczenia problemu prawnego oraz sposobu reakcji na niego.

Autor pragnie podziękować współautorom wcześniejszych prac, z których pochodziły inspiracje do przeprowadzenia badań empirycznych. Dzięki współpracy autor uzyskał niezbędną wiedzę, która pomogła w specyfikacji odpowiednich modeli ekonometrycznych. Autor składa podziękowania Łukaszowi Arendtowi, Karolowi Korczakowi, Krzysztofowi Szczygielskiemu, Sinanowi Tandoganowi, Teomanowi Pamukcu, Richardowi Woodwardowi. Wyniki niektórych badań były prezentowane podczas zebrań w ramach seminarium naukowego pt. „Modelowanie gospodarki narodowej". Autor pragnie podziękować uczestnikom tych zebrań, w tym przede wszystkim Aleksandrowi Welfe, Michałowi Majsterkowi, Robertowi Kelmowi, Annie Staszewskiej-Bystrovej, Piotrowi Kębłowskiemu, Piotrowi Karpowi, Emilii Gosińskiej, Katarzynie Leszkiewicz-Kędzior, Aleksandrze Majchrowskiej, Sylwii Roszkowskiej, Iwonie Świeczewskiej, Jakubowi Boratyńskiemu za wnikliwe i szczegółowe uwagi, które często przyczyniały się do poprawy jakości uzyskiwanych rezultatów. Szczególne podziękowania autor składa Ewie Stawasz-Grabowskiej oraz Michałowi Majsterkowi za cierpliwość w lekturze całości monografii. Ewentualne niedociągnięcia i błędy należy zaliczyć na konto autora.

Notacja wykorzystywana w monografii

Ze względu na dużą liczbę wzorów oraz innych symboli pojawiających się w niniejszej monografii użyteczne wydaje się przedstawienie indeksu wzorów i symboli. Czytelnik analizujący wzory i przekształcenia znajdujące się w kolejnych rozdziałach może odwoływać się do niego w celu upewnienia się, jak odczytywać określone oznaczenia.

Indeksowanie

$i = 1, \ldots, I$ – jednostki.
$j = 1, \ldots, J$ – grupy (klastry) w ogólnym modelu wielopoziomowym.
$qq = 1, \ldots, QQ$ – grupy dla efektów losowych.
$w = 1, \ldots, W$ – województwa (w modelu ogólnym).
$s = 1, \ldots, S$ – sekcje (w modelu ogólnym).
$j1 = 1, \ldots, J1$ – gminy w przykładowym modelu wielopoziomowym rozważanym w podrozdziale 3.3.
$j2 = 1, \ldots, J2$ – powiaty w przykładowym modelu wielopoziomowym rozważanym w podrozdziale 3.3.
$j3 = 1, \ldots, J3$ – województwa w przykładowym modelu wielopoziomowym rozważanym w podrozdziale 3.3.
$jj1 = 1, \ldots, JJ1$ – grupy PKD w przykładowym modelu wielopoziomowym z efektami krzyżowymi rozważanym w podrozdziale 3.4.
$jj2 = 1, \ldots, JJ2$ – działy PKD w przykładowym modelu wielopoziomowym z efektami krzyżowymi rozważanym w podrozdziale 3.4.
$jj3 = 1, \ldots, JJ3$ – sekcje PKD w przykładowym modelu wielopoziomowym z efektami krzyżowymi rozważanym w podrozdziale 3.4.
$ss = 2, \ldots, SS$ – poziomy zagnieżdżenia.
$l = 1, \ldots, L$ – wybory dokonywane przez jednostkę w uporządkowanym modelu polichotomicznym.
$r = 1, \ldots, RA$ – rankingi.
$k = 1, \ldots, K$ – zmienne egzogeniczne.
$k = 1, \ldots, K1$ – zmienne egzogeniczne wpływające na wynik jednostki na poziomie województw (podrozdział 3.3).
$k = 1, \ldots, K2$ – zmienne egzogeniczne wpływające na wynik jednostki na poziomie powiatów (podrozdział 3.3).

$k = 1, \ldots, K3$ – zmienne egzogeniczne wpływające na wynik jednostki na poziomie gmin (podrozdział 3.3).

$k = 1, \ldots, K4$ – zmienne egzogeniczne wpływające na wynik jednostki na poziomie indywidualnym (podrozdział 3.3).

$k = 1, \ldots, \widetilde{K}$ – zmienne egzogeniczne dostępne na poziomie indywidualnym, definiowane we wprowadzeniu do modelu wielopoziomowego (podrozdział 3.3).

$k = \widetilde{K} + 1, \ldots, \widetilde{K} + \widehat{K}$ – zmienne egzogeniczne obserwowalne na poziomie grupowym, definiowane we wprowadzeniu do modelu wielopoziomowego (podrozdział 3.3).

$k = 1, \ldots, \breve{K}$ – zmienne egzogeniczne, których oddziaływanie na zmienną zależną różni się między grupami.

$m = 1, \ldots, M$ – równania w standardowym i wielopoziomowym, wielorównaniowym modelu probitowym.

$n = 1, \ldots$ – iteracje w przypadkach stosowania metod symulacyjnych.

$h = 1, \ldots, H$ – replikacje w metodzie bootstrap oraz MCMC.

$p = 1, \ldots, P$ – progi w modelu polichotomicznym uporządkowanym.

$d = 1, \ldots, D$ – strukturyzacje w modelu wielopoziomowym z efektami krzyżowymi.

$g = 1, \ldots, G$ – grupy podczas omawiania testu Hosmera-Lemeshowa.

$t = 1, \ldots, T$ – indeks czasu.

$b = 1, \ldots, B$ – warianty do uszeregowania w modelu regresji rankingowej.

Notacja dla zmiennej zależnej

y – zmienna zależna.

$\boldsymbol{y} = \begin{bmatrix} y_1 & \ldots & y_I \end{bmatrix}^T$ – wektor obserwacji na zmiennej objaśnianej.

$\boldsymbol{y}_j = \begin{bmatrix} y_{j1} & \ldots & y_{jI_j} \end{bmatrix}^T$ – wektor obserwacji dla j-tej grupy (j-tego klastra).

$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1^T & \ldots & \boldsymbol{y}_J^T \end{bmatrix}^T$ – wektor wszystkich obserwacji na zmiennej objaśnianej, składający się z wektorów obserwacji dla poszczególnych klastrów.

y^* – zmienna nieobserwowalna, związana ze zmienną dwumianową lub uporządkowaną.

$y_{(m)}$ – m-ta zmienna zależna w wielorównaniowym modelu probitowym.

$\tilde{\boldsymbol{y}}$ – wektor endogenicznych regresorów w endogenicznym modelu probitowym.

Notacja dla zmiennych niezależnych

x – zmienna niezależna.

$\boldsymbol{X}$ – macierz obserwacji na wszystkich zmiennych objaśniających.

x_i – i-ty wiersz macierzy $\boldsymbol{X}$, odpowiadający wektorowi wartości na zmiennych objaśniających dla i-tej jednostki.

$\boldsymbol{x}_i^l$ – wektor obserwacji na zmiennych objaśniających w równaniu związanym z l-tym wyborem w modelu polichotomicznym nieuporządkowanym.

$\tilde{\boldsymbol{x}}_i$ – wektor wszystkich zmiennych objaśniających w endogenicznym modelu probitowym.

$\tilde{\boldsymbol{x}}_{(1)i}$ – wektor zmiennych egzogenicznych wpływających bezpośrednio na zmienną wynikową w endogenicznym modelu probitowym.

$\tilde{\boldsymbol{x}}_{(2)i}$ – wektor zmiennych instrumentalnych w endogenicznym modelu probitowym.

$\boldsymbol{x}^{(ss)}$ – wektor obserwacji na zmiennych objaśniających na *s*-tym poziomie zagnieżdżenia (w modelu wielopoziomowym).

$\boldsymbol{X}_{[1]}$ – macierz obserwacji na zmiennych objaśniających obserwowanych na poziomie indywidualnym.

$\boldsymbol{X}_{[2]}$ – macierz obserwacji na zmiennych objaśniających dostępnych na poziomie grupowym.

$\boldsymbol{X}_{[3]}$ – podmacierz macierzy $\boldsymbol{X}_{[1]}$ zawierająca zmienne obserwowalne na poziomie indywidualnym, których wpływ na regresanta losowo różni się między grupami (np. regionami).

$\boldsymbol{X}_{[3]}^{(j)}$ – podmacierz macierzy $\boldsymbol{X}_{[3]}$ zawierająca wektory zerowe dla jednostek nienależących do *j*-tej grupy oraz wektory odpowiadające wektorom macierzy $\boldsymbol{X}_{[3]}$ dla jednostek należących do *j*-tej grupy.

$\boldsymbol{X}_j$– macierz obserwacji na zmiennych objaśniających dla konkretnej *j*-tej grupy (*j*-tego klastra).

$\boldsymbol{X} = \left[\boldsymbol{X}_1^T \quad \ldots \quad \boldsymbol{X}_J^T \right]^T$ – macierz obserwacji na zmiennych objaśniających zawierająca podmacierze składające się z macierzy obserwacji dla poszczególnych grup (klastrów).

$\boldsymbol{x}_{(m)i}$ – wektor obserwacji dla *i*-tej jednostki na zmiennych występujących w *m*-tym równaniu w wielorównaniowych modelach probitowych.

$\tilde{\boldsymbol{x}}_i$ – wektor wszystkich zmiennych egzogenicznych w endogenicznym modelu probitowym.

$\tilde{\boldsymbol{x}}_{(1)i}$ – wektor zmiennych egzogenicznych niebędących instrumentami w endogenicznym modelu probitowym.

$\tilde{\boldsymbol{x}}_{(2)i}$ – wektor instrumentów w endogenicznym modelu probitowym.

$\boldsymbol{w}_i$ – wektor regresorów wpływających na selekcję w modelu Heckmana.

$\boldsymbol{zz}_i$ – wektor regresorów wpływających na to, czy zmienna licznikowa przyjmuje wartość 0 w modelu licznikowym z podwyższoną liczbą „zer".

$\tilde{\boldsymbol{x}}_i^{woj,w}$ – wektor obserwacji na zmiennych objaśniających dla *i*-tej jednostki, pod warunkiem, że należy ona do *w*-tego województwa. W przeciwnym przypadku wektor ten składa się z elementów zerowych.

$\tilde{\boldsymbol{x}}_i^{sek,s}$ – wektor obserwacji na zmiennych objaśniających dla *i*-tej jednostki, pod warunkiem, że należy ona do *s*-tej sekcji. W przeciwnym wypadku wektor ten zawiera tylko elementy zerowe.

woj – wektor zmiennych binarnych związanych z przynależnością jednostki do województwa.

sek – wektor zmiennych binarnych związanych z przynależnością jednostki do sekcji.

$\boldsymbol{xw}_i$ – wektor obserwacji na zmiennych objaśniających wyjaśniających selekcję w modelu Heckmana.

$\boldsymbol{ww}_i$ – wektor obserwacji na zmiennych różniących się ze względu na województwa.

$\boldsymbol{vv}_i$ – wektor obserwacji na zmiennych różniących się ze względu na sekcje.

***w*3** – zmienne wpływające na wartość kategorii wynikowej na poziomie województw (przykład w podrozdziale 3.3).

***w*2** – zmienne wpływające na wartość kategorii wynikowej na poziomie powiatów (przykład w podrozdziale 3.3).

***w*1** – zmienne wpływające na wartość kategorii wynikowej na poziomie gmin (przykład w podrozdziale 3.3).

***w*0** – zmienne wpływające na wartość kategorii wynikowej na poziomie indywidualnym (przykład w podrozdziale 3.3).

Notacja dla pozostałych ważnych zmiennych

UZ_i^l – użyteczność *i*-tej jednostki z wyboru *l*-tego wariantu w modelu polichotomicznym nieuporządkowanym.

VZ_i^l – część deterministyczna użyteczności *i*-tej jednostki z wyboru *l*-tego wariantu w modelu polichotomicznym nieuporządkowanym.

ID_i – zmienna binarna, która w modelu z podwyższoną liczbą „zer" informuje, czy zmienna licznikowa jest równa 0, czy też przyjmuje wartość dodatnią.

K_i – zmienna przekształcająca zmienną binarną w inną zmienną dwuwartościową, przyjmującą wartości –1 (gdy przekształcana zmienna binarna wynosi 0) oraz 1 (dla zmiennej binarnej równej 1).

p_i^l – prawdopodobieństwo wyboru *l*-tego wariantu przez *i*-tą jednostkę.

$p_{ij}(p)$ – prawdopodobieństwo, że obserwowalna zmienna zależna w uporządkowanym modelu polichotomicznym przyjmie wartość *p* dla *i*-tej jednostki z *j*-tego klastra.

d_i^l – zmienna binarna przyjmująca wartość 1, jeśli *i*-ta jednostka wybrała *l*-ty wariant i 0 w przeciwnym przypadku.

$\delta_i^{ll'}$ – zmienna binarna przyjmująca wartość 1, jeśli dla *i*-tej jednostki wariant *l*-ty jest preferowany w stosunku do wariantu *l'*.

Notacja dla elementów związanych z efektami losowymi w modelach wielopoziomowych

$\boldsymbol{Z}$ – macierz przy efektach losowych. Składa się ona głównie ze zmiennych zero-jedynkowych definiujących przynależność określonych jednostek do poszczególnych klastrów (grup), a także z tych zmiennych wchodzących w skład macierzy $\boldsymbol{X}$, których oddziaływanie na zmienną zależną różni się między grupami (klastrami).

$\boldsymbol{z}_i^{ZP}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do posiadania technologii informacyjnych i komunikacyjnych w procesach biznesowych związanych z zarządzaniem przedsiębiorstwem.

$\boldsymbol{z}_i^{ERP}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do posiadania technologii informacyjnych i komunikacyjnych w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa.

$\boldsymbol{z}_i^{CAD}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do posiadania technologii informacyjnych i komunikacyjnych w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM.

$\boldsymbol{z}_i^{SM}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do posiadania technologii informacyjnych i komunikacyjnych w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną.

$\boldsymbol{z}_i^{INW}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do inwestowania w rozwój technologii informacyjnych i telekomunikacyjnych.

$\boldsymbol{z}_i^{BR}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do posiadania własnego wydziału B+R.

$\boldsymbol{z}_i^{PROD}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do wprowadzania innowacji produktowych.

$\boldsymbol{z}_i^{PROC}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do wprowadzania innowacji procesowych lub organizacyjnych.

$\boldsymbol{z}_i^{MARK}$ – wektor przy efektach losowych w równaniu wyjaśniającym skłonność firm do wprowadzania innowacji marketingowych.

$\boldsymbol{z}_i^{PR}$ – wektor przy efektach losowych w równaniu wyjaśniającym prawdopodobieństwo zaistnienia problemu prawnego.

Notacja dla składników losowych i efektów losowych

$\boldsymbol{\varepsilon}$ – wektor składników losowych.

$\boldsymbol{\varepsilon}_{(m)}$ – wektor składników losowych dla m-tego równania w modelach zawierających więcej niż jedno równanie.

ε_i^l – składnik losowy w równaniu związanym z l-tym wyborem w modelu polichotomicznym nieuporządkowanym.

εz_{ij}^l – składnik losowy związany z i-tą jednostką należącą do j-tej grupy oraz l-tym wyborem w wielopoziomowym nieuporządkowanym modelu polichotomicznym.

$\hat{\boldsymbol{e}}$ – reszty.

$\tilde{\boldsymbol{e}}$ – składniki losowe (po ortogonalizacji) w wielorównaniowym modelu probitowym.

$\boldsymbol{u}$ – wektor wszystkich efektów losowych.

$\boldsymbol{u}_{[1]}$ – podwektor wektora efektów losowych związany z losowym wpływem poszczególnych zmiennych egzogenicznych na zmienną zależną.

$\boldsymbol{u}_{[2]}$ – podwektor wektora efektów losowych związany z losowym wyrazem wolnym.

$\boldsymbol{u}_j$ – efekty losowe dla poszczególnych klastrów.

$\boldsymbol{u}^{(ss)}$ – efekty losowe na *ss*-tym poziomie zagnieżdżenia.

$\boldsymbol{u} = \left[\boldsymbol{u}_1^T \quad \ldots \quad \boldsymbol{u}_J^T \right]^T$ – wektor efektów losowych.

$\boldsymbol{u}_{\langle h \rangle}$ – wektor efektów losowych dla *h*-tej replikacji podczas wykorzystywania metody MCMC w celu estymacji parametrów uogólnionego liniowego modelu wielopoziomowego.

$\boldsymbol{u}_{\{j\}}$ – wektor zawierający efekty losowe w przypadku *j*-tej grupy oraz elementy „zerowe" dla pozostałych grup.

$\boldsymbol{u}^{ZP}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie zarządzania produkcją.

ε^{ZP} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie zarządzania produkcją.

$\boldsymbol{u}^{ERP}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie zarządzania zasobami przedsiębiorstwa.

$\boldsymbol{\varepsilon}^{ERP}$ – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie zarządzania zasobami przedsiębiorstwa.

$\boldsymbol{u}^{CAD}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie wsparcia dla projektowania i wytwarzania CAD/CAM.

ε^{CAD} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie wsparcia dla projektowania i wytwarzania CAD/CAM.

$\boldsymbol{u}^{SM}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie sterowania maszynami lub linią produkcyjną.

ε^{SM} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wykorzystywania technologii informacyjnych i komunikacyjnych w zakresie sterowania maszynami lub linią produkcyjną.

$\boldsymbol{u}^{INW}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do inwestowania w rozwój technologii informatycznych i komunikacyjnych.

ε^{INW} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do inwestowania w rozwój technologii informatycznych i komunikacyjnych.

$\boldsymbol{u}^{BR}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do posiadania wewnętrznego wydziału B+R.

ε^{BR} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do posiadania wewnętrznego wydziału B+R.

$\boldsymbol{u}^{PROD}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji produktowych.

ε^{PROD} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji produktowych.

$\boldsymbol{u}^{PROC}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji procesowych lub organizacyjnych.

ε^{PROC} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji procesowych lub organizacyjnych.

$\boldsymbol{u}^{MARK}$ – efekty losowe w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji marketingowych.

ε^{MARK} – składnik losowy w równaniu wyjaśniającym skłonność przedsiębiorstw do wprowadzania innowacji marketingowych.

$\boldsymbol{u}^{PR}$ – wektor efektów losowych w modelu wyjaśniającym prawdopodobieństwo zaistnienia problemu prawnego.

ε^{PR} – składnik losowy w modelu wyjaśniającym prawdopodobieństwo zaistnienia problemu prawnego.

Główne parametry i estymatory

β – parametr ilustrujący wpływ zmiennej objaśniającej na zmienną zależną w większości przypadków.

β – wektor parametrów.

$\beta_{(m)}$ – wektor parametrów związany z m-tym równaniem w modelach wielorównaniowych.

$\boldsymbol{\beta}_{[1]}$ – wektor parametrów przy zmiennych obserwowalnych na poziomie jednostek.

$\boldsymbol{\beta}_{[2]}$ – wektor parametrów przy zmiennych dostępnych na poziomie grupowym.

$\boldsymbol{\beta}_{[3]}^{(j)}$ – wektor parametrów przy zmiennych wchodzących w skład macierzy $\boldsymbol{X}_{[3]}^{(j)}$.

$\boldsymbol{\beta}_{]q[}$ – wektor parametrów dla kwantyla rzędu q w metodzie regresji kwantylowej.

$\widehat{\boldsymbol{\beta}}_{OLS}$ – estymator uzyskany klasyczną metodą najmniejszych kwadratów.

$\widehat{\boldsymbol{\beta}}_{GLS}$ – estymator uzyskany uogólnioną metodą najmniejszych kwadratów.

$\widehat{\boldsymbol{\beta}}_{FGLS}$ – estymator uzyskany uogólnioną metodą najmniejszych kwadratów z estymacją.

$\widehat{\boldsymbol{\beta}}_{ML}$ – estymator uzyskany metodą największej wiarygodności.

$\widehat{\boldsymbol{\beta}}_Q$ – estymator regresji kwantylowej.

$\widehat{\boldsymbol{\beta}}_M$ – estymator *M*.

$\widehat{\boldsymbol{\beta}}_S$ – estymator *S*.

$\widehat{\boldsymbol{\beta}}_{MM}$ – estymator *MM*.

$\widehat{\boldsymbol{\beta}}_{BL}$ – estymator parametrów uogólnionego liniowego modelu wielopoziomowego, uzyskany w wyniku zastosowania korekty Breslowa i Lina.

$\widehat{\boldsymbol{\beta}}_{\{un\}}$ – estymator nieskorygowany parametrów uogólnionego liniowego modelu wielopoziomowego.

$\boldsymbol{\beta}^l$ – wektor parametrów związanych z *l*-tym wyborem w modelu polichotomicznym nieuporządkowanym.

$\widetilde{\boldsymbol{\beta}}^{woj,w}$ – wektor parametrów związanych z *w*-tym województwem

$\widetilde{\boldsymbol{\beta}}^{sek,s}$ – wektor parametrów związanych z *s*-tą sekcją.

$\boldsymbol{\beta}^{\{n\}}$ – wektor parametrów w *n*-tej iteracji.

$\boldsymbol{\beta}^{ZP}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych z zarządzaniem produkcją.

$\boldsymbol{\beta}^{ERP}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa.

$\boldsymbol{\beta}^{CAD}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM.

$\boldsymbol{\beta}^{SM}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną.

$\boldsymbol{\beta}^{INW}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do inwestowania w rozwój technologii informacyjnych i komunikacyjnych.

$\boldsymbol{\beta}^{BR}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do posiadania własnego wydziału B+R.

$\boldsymbol{\beta}^{PROD}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji produktowych.

$\boldsymbol{\beta}^{PROC}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji procesowych lub organizacyjnych.

$\boldsymbol{\beta}^{MARK}$ – wektor parametrów przy zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji marketingowych.

$\boldsymbol{\beta}^{PR}$ – wektor parametrów (stałych) w równaniu wyjaśniającym prawdopodobieństwo zaistnienia problemu prawnego.

$\breve{b}(\boldsymbol{\Theta})$ – wartość oczekiwana oszacowania uzyskanego metodą quasi-największej wiarygodności w uogólnionym liniowym modelu wielopoziomowym, gdy $\boldsymbol{\Theta}$ jest wektorem prawdziwych parametrów.

$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{W}$ – macierz kowariancji między składnikami losowymi dla różnych obserwacji.

$\widetilde{\boldsymbol{\Sigma}}$ – macierz kowariancji między składnikami losowymi z różnych równań w standardowym i wielopoziomowym, wielorównaniowym modelu probitowym.

$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}2\boldsymbol{\varepsilon}2}$ – macierz kowariancji między składnikami losowymi wchodzącymi w skład wektora $\boldsymbol{\varepsilon}_{(2)i}$ (endogeniczny model probitowy).

$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}2\varepsilon1}$ – wektor składający się z kowariancji między składnikami losowymi wchodzącymi w skład wektora $\boldsymbol{\varepsilon}_{(2)i}$ a składnikiem losowym $\boldsymbol{\varepsilon}_{(1)i}$ (endogeniczny model probitowy).

$\sigma^2_{\varepsilon 1}$ – wariancja składnika losowego $\boldsymbol{\varepsilon}_{(1)i}$ w endogenicznym modelu probitowym.

$\boldsymbol{\Omega}$ – macierz kowariancji między efektami losowymi.

$\boldsymbol{\Omega}^{(ss)}$ – macierz kowariancji między efektami losowymi na *s*-tym poziomie zagnieżdżenia.

$\tilde{\boldsymbol{Q}}^{(ss)}$ – dekompozycja Choleskiego macierzy $\boldsymbol{\Omega}^{(s)}$.

$\boldsymbol{\theta}$ – wektor zawierający unikatowe elementy macierzy $\boldsymbol{\Omega}$.

$\boldsymbol{\Theta}$ – wektor zawierający wszystkie parametry do estymacji w uogólnionym liniowym modelu wielopoziomowym.

$\boldsymbol{V}$ – macierz kowariancji między obserwacjami na zmiennej zależnej.

ρ – współczynnik korelacji między składnikami losowymi z dwóch równań w dwurównaniowym modelu probitowym.

$\rho_{mm'}$ – współczynnik korelacji w wielorównaniowym modelu probitowym między składnikami losowymi z równań *m*' oraz *m*.

τ_ρ – parametr progowy w modelu polichotomicznym uporządkowanym.

$\boldsymbol{\tau}$ – wektor składający się z parametrów progowych w modelu polichotomicznym uporządkowanym.

$\kappa\kappa$ – parametr progowy związany z funkcją straty (1) w estymacji odpornej.

cc – parametr progowy związany z funkcją straty (2) w estymacji odpornej.

ccc – punkt progowy wyznaczany podczas mierzenia jakości dopasowania w modelu dwumianowym.

$\ddot{\mathrm{M}}(s)$ – liczba efektów losowych na poziomie *s*.

μ_q – wartość kwantyla rzędu *q*.

λ – parametr intensywności w modelu Poissona.

α – parametr związany z nadwyżką wariancji ponad wartość oczekiwaną w modelu ujemnym dwumianowym.

$\boldsymbol{\varrho}$ – wektor parametrów ilustrujących wpływ przynależności do województw na wartość zmiennej wynikowej (podrozdział 3.2).

$\ddot{\boldsymbol{\varrho}}$ – wektor parametrów ilustrujących wpływ przynależności do sekcji na wartość zmiennej wynikowej (podrozdział 3.2).

$\boldsymbol{\varsigma}$ – wektor parametrów mierzących wpływ endogenicznych regresorów na wartość zmiennej wynikowej w endogenicznym modelu probitowym.

$\boldsymbol{\pi}$ – wektor parametrów odzwierciedlających wpływ zmiennych egzogenicznych na endogeniczne regresory w endogenicznym modelu probitowym.

$\boldsymbol{J}$ – macierz przekształcająca wektor obserwacji na wszystkich zmiennych egzogenicznych w wektor niezawierający instrumentów w endogenicznym modelu probitowym.

$\boldsymbol{MA}$ – macierz zawierająca „zera" i „jedynki" w modelu wielopoziomowym. Przypisanie wartości 1 lub 0 zależy od tego, czy jednostka należy do danej grupy, czy nie.

$\boldsymbol{W}_0$ – macierz wag wykorzystywana podczas aproksymacji funkcji wiarygodności za pomocą propozycji Longforda w modelu wielopoziomowym.

$\widetilde{\boldsymbol{\Theta}}$ – estymator dla wektora wszystkich parametrów uzyskany w wyniku maksymalizacji funkcji quasi-największej wiarygodności.

$\widehat{\boldsymbol{\Theta}}_{BC,KUK}^{\{n\}}$ – wektor oszacowań wszystkich parametrów modelu wielopoziomowego w n-tej iteracji, wykorzystywany do symulacji bootstrap zgodnie z propozycją Kuka.

$\widehat{\boldsymbol{\Theta}}_{BC,RM}^{\{n\}}$ – wektor oszacowań wszystkich parametrów modelu wielopoziomowego w n-tej iteracji, wykorzystywany podczas stosowania metody stochastycznej aproksymacji Robbinsa-Monro.

$\widetilde{\boldsymbol{\Theta}}_{\langle h\rangle}^{*}$ – wektor oszacowań parametrów modelu wielopoziomowego dla h-tej replikacji w modelu wielopoziomowym podczas wykorzystania metod bootstrapowych.

$\overline{\boldsymbol{\Theta}}^{*}$ – średnia z oszacowań dla wszystkich replikacji podczas wykorzystania bootstrapowych metod korekty obciążenia.

$\boldsymbol{\zeta}$ – wektor odpowiadający kryterium zbieżności.

$\ddot{\boldsymbol{\psi}}$ – wektor parametrów mierzących wpływ zmiennych wchodzących w skład wektora ww na wartość zmiennej wynikowej.

$\ddot{\boldsymbol{\omega}}$ – wektor parametrów mierzących wpływ zmiennych wchodzących w skład wektora vv na wartość zmiennej wynikowej.

$\boldsymbol{\gamma}$ – wektor parametrów przy zmiennych w równaniu selekcji w modelu Heckmana.

$\dot{\boldsymbol{\gamma}}$ – wektor parametrów przy zmiennych wpływających na fakt przyjmowania przez zmienną zależną „zerowej" wartości w modelu licznikowym z podwyższoną liczbą „zer".

$\ddot{\pi}$ – parametry mierzące wpływ zmiennych $\boldsymbol{w3}$ na wartość zmiennej wynikowej.

$\ddot{\gamma}$ – parametry mierzące wpływ zmiennych $\boldsymbol{w2}$ na wartość zmiennej wynikowej.

$\ddot{\beta}$ – parametry mierzące wpływ zmiennych $\boldsymbol{w1}$ na wartość zmiennej wynikowej.

$\ddot{\alpha}$ – parametry mierzące wpływ zmiennych $\boldsymbol{w0}$ na wartość zmiennej wynikowej.

$\overline{\hat{p}_g}$ – średnie prawdopodobieństwo, że zmienna zależna przyjmuje wartość 1 w g-tej grupie (dla testu Hosmera-Lemeshowa).

$\boldsymbol{H}$ – macierz zawierająca macierze $\boldsymbol{\pi}$ oraz $\boldsymbol{J}$ w endogenicznym modelu probitowym.

$\boldsymbol{MO} = \left[\boldsymbol{y}^T \quad \boldsymbol{u}^T \right]^T$ – macierz obserwacji na wszystkich zmiennych w modelu wielopoziomowym (obserwowalnych i nieobserwowalnych).

$\ddot{\mu}_q$ – wartość oczekiwana dla q-tego efektu losowego.

$\ddot{\tau}_q^2$ – wariancja dla q-tego efektu losowego.

uz_{ij}^l – użyteczność i-tej jednostki należącej do j-tego klastra związana z l-tym wyborem w wielopoziomowym, nieuporządkowanym modelu polichotomicznym.

fz_{ij}^l – deterministyczny składnik losowy, reprezentujący obserwowaną heterogeniczność wariantów do wyboru jednostek oraz klastrów w wielopoziomowym, nieuporządkowanym modelu polichotomicznym.

δz_{ij}^l – zmienna sztuczna, reprezentująca nieobserwowalną heterogeniczność w wielopoziomowym, nieuporządkowanym modelu polichotomicznym.

Główne funkcje

$E(\cdot|\cdot)$ – warunkowa wartość oczekiwana.

$F_{\xi}(\cdot)$ – dystrybuanta zmiennej losowej ξ.

$Q_I(\cdot)$ – funkcja celu w metodzie regresji kwantylowej.

$\rho\rho(\cdot)$ – funkcja straty w podrozdziale poświęconym odpornym metodom estymacji parametrów.

$\Phi(\cdot)$ – dystrybuanta standardowego rozkładu normalnego.

$\phi(\cdot)$ – funkcja gęstości standardowego rozkładu normalnego.

$\Phi_M(\cdot)$ – dystrybuanta M-wymiarowego rozkładu normalnego.

$\Lambda(\cdot)$ – dystrybuanta rozkładu logistycznego.

$H(\cdot)$ – dystrybuanta komplementarnego rozkładu log-log.

$I\{\cdot\}$ – funkcja wskaźnikowa przyjmująca wartość 1, gdy warunek zdefiniowany w klamrowym nawiasie jest spełniony.

$\ddot{I}_p(y_{ij})$ – funkcja przyjmująca wartość 1, jeśli zmienna uporządkowana dla i-tej obserwacji należącej do j-tego klastra przyjmuje wartość p oraz 0 w przeciwnym przypadku.

$qL(\boldsymbol{\Theta})$ – funkcja quasi-największej wiarygodności.

$qL_{S1}(\boldsymbol{\Theta})$ – funkcja quasi-największej wiarygodności wykorzystująca aproksymację Solomona-Coxa typu pierwszego.

$qL_{S2}(\boldsymbol{\Theta})$ – funkcja quasi-największej wiarygodności wykorzystująca aproksymację Solomona-Coxa typu drugiego.

$\widetilde{\ln L}_j(\boldsymbol{\Theta})_{\langle 0\rangle}^{[k]} = \frac{\partial^k \ln L_j(\boldsymbol{\Theta})}{\partial u_j^k}$ dla $u_j^k = 0$.

$\bar{\bar{r}}$ – funkcja wyrażająca ogólną postać funkcji quasi-największej wiarygodności.

$\pi_f(\cdot)$ – ogólna funkcja nieliniowego modelu wielopoziomowego.

H-L – statystyka testu Hosmera-Lemeshowa.

IMR – odwrócony iloraz Millsa.

$h(\cdot,\cdot)$ – łączna funkcja gęstości zmiennych y oraz $\tilde{\boldsymbol{y}}$ w endogenicznym modelu probitowym.

$f_{ep}(\cdot)$ – funkcja gęstości rozkładu warunkowego y względem $\tilde{\boldsymbol{y}}$ w endogenicznym modelu probitowym.

$g_{ep}(\cdot)$ – brzegowa funkcja gęstości rozkładu $\tilde{\boldsymbol{y}}$ w endogenicznym modelu probitowym.

$g_l(\cdot)$ – funkcja łączącą w uogólnionych liniowych modelach wielopoziomowych.

$g_{uu}(\cdot)$ – funkcja gęstości wektora losowego.

$h_{uu}(\cdot)$ – rozkład próbkowy efektów losowych wykorzystywany podczas stosowania algorytmu Metropolisa-Hastingsa.

$f_{wj}(\cdot|\cdot)$ – funkcja gęstości warunkowego rozkładu zmiennej wynikowej względem efektów losowych w modelach wielopoziomowych dla j-tej grupy.

$f_{\boldsymbol{w}}(\cdot|\cdot)$ – funkcja gęstości warunkowego rozkładu wektora wartości zmiennej wynikowej względem efektów losowych w modelach wielopoziomowych.

$f_{w_i}(\cdot|\cdot)$ – funkcja gęstości rozkładu warunkowego zmiennej wynikowej względem efektów losowych i parametrów dla i-tej jednostki.

$h_w(\cdot)$ – funkcja zależna od wszystkich parametrów i efektów losowych (wykorzystywana przy definiowaniu wkładu *j*-tej grupy do funkcji wiarygodności).

$\ln L$ – ogólny zapis dla logarytmu funkcji wiarygodności.

$\ln L_I^g$ – funkcja wiarygodności maksymalizowana w pierwszym kroku dla endogenicznego modelu probitowego.

$\ln L_I^f$ – funkcja wiarygodności maksymalizowana w drugim kroku dla endogenicznego modelu probitowego.

$L_j(\cdot)$ – wkład *j*-tej grupy do funkcji wiarygodności.

$L_{ij}(\cdot)$ – wkład *i*-tej jednostki należącej do *j*-tej grupy do funkcji wiarygodności.

$g_u(\cdot)$ – funkcja gęstości dla pojedynczego efektu losowego.

$\mu_i^{\boldsymbol{u}} = E(y_i \mid \boldsymbol{u})$.

$d_i\left(y_i, \mu_i^{\boldsymbol{u}}\right)$ – funkcja zależna od zmiennej wynikowej i warunkowej wartości oczekiwanej zmiennej wynikowej względem efektów losowych.

$\ddot{\boldsymbol{\kappa}} = (\boldsymbol{\beta}, \varphi)$.

$\tilde{L}_{ij} = \left\{l_{ij}^1, \ldots, l_{ij}^{L_{ij}}\right\}$ – zbiór wszystkich możliwych wyborów dla *i*-tej jednostki należącej do *j*-tego klastra w wielopoziomowym nieuporządkowanym modelu polichotomicznym.

Mierniki jakości dopasowania oraz mierniki wpływu zmiennych egzogenicznych na prawdopodobieństwo, że zmienna zależna przyjmuje określone wartości

R^2_{binary} – pseudo *R*-kwadrat.

R^2_{BL} – współczynnik determinacji Ben-Akiva i Lermana oraz Kaya i Little.

R^2_{Ef} – współczynnik determinacji R^2-Efrona.

R^2_{VZ} – współczynnik determinacji Vealla i Zimmermana.

R^2_{MZ} – współczynnik determinacji Zavoiny i McKelveya.

$R^2 - McFadden$ – współczynnik determinacji McFaddena.

$R^2 - Nagelkerke$ – współczynnik determinacji Nagelkerke'a.

CP – procent poprawnych predykcji.

SENSITIVITY – czułość.

SPECIFICITY – specyficzność.

PPV – pozytywna wartość predyktywna.

NPV – negatywna wartość predyktywna.

MPE_{ik}^{l} – wpływ marginalnej zmiany wartości *k*-tej zmiennej egzogenicznej na prawdopodobieństwo, że *i*-ta jednostka wybierze *l*-ty wariant.

$\widehat{MPE}_{k}^{l}$ – efekt krańcowy dla średniej.

$\widehat{AMPE}_{k}^{l}$ – średni efekt krańcowy.

Zmienne wykorzystywane w badaniach empirycznych
Rozdział 3

WYN_i – wynagrodzenie.
SZK_i – poziom wykształcenia mierzony liczbą lat nauki.
XZ_i – doświadczenie zawodowe mierzone liczbą przepracowanych lat.
$D1_i$ – zmienna zero-jedynkowa związana z wykształceniem podstawowym.
$D2_i$ – zmienna zero-jedynkowa związana z wykształceniem średnim.
$D3_i$ – zmienna zero-jedynkowa związana z wykształceniem wyższym.
ZK_i – zmienna kontrolna w równaniu wyjaśniającym wynagrodzenia (podrozdział 3.5).
$plac\_nom_t$ – nominalne płace przeciętne (w modelu wyjaśniającym wynagrodzenia na poziomie makro).
cen_t– indeks cen konsumpcyjnych (*CPI*) (w modelu wyjaśniającym wynagrodzenia na poziomie makro).
wyd_prac_t – wydajność pracy (w modelu wyjaśniającym wynagrodzenia na poziomie makro).
$bezr_t$ – stopa bezrobocia (w modelu wyjaśniającym wynagrodzenia na poziomie makro).
WYN_NOM_{it} – wynagrodzenie nominalne.
WYN_REL_{it} – wynagrodzenie relatywne, będące ilorazem wynagrodzenia nominalnego i mediany wynagrodzeń.
$WYZSZE_i$ – zmienna binarna przyjmująca wartość 1 dla pracownika z wyższym wykształceniem.
$SREDNIE\_ZAWODOWE_i$ – zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem średnim technicznym lub policealnym.
$ZASADNICZE\_ZAWODOWE_i$ – zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem zasadniczym zawodowym.
$PODSTAWOWE_i$ – zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem podstawowym.
$DOSW\_FIRMA_i$ – liczba przepracowanych lat przez pracownika w firmie, w której obecnie pracuje.
$DOSW\_OGOL_i$ – liczba pełnych lat przepracowanych przez pracownika.
$ROZMIAR\_10\_49_i$ – zmienna binarna przyjmująca wartość 1, jeśli firma (zatrudniająca *i*-tego pracownika) zatrudnia co najmniej 10 i mniej niż 50 osób.
$ROZMIAR\_50\_249_i$ – zmienna binarna przyjmująca wartość 1, jeśli firma (zatrudniająca *i*-tego pracownika) zatrudnia co najmniej 50 i mniej niż 250 osób.

$ROZMIAR\_250\_499_i$ – zmienna binarna przyjmująca wartość 1, jeśli firma (zatrudniająca *i*-tego pracownika) zatrudnia co najmniej 250 i mniej niż 500 osób.

$ROZMIAR\_CON500_i$ – zmienna binarna przyjmująca wartość 1, jeśli firma (zatrudniająca *i*-tego pracownika) zatrudnia co najmniej 500 osób.

$KOBIETA_i$ – zmienna binarna przyjmująca wartość 1 w przypadku kobiet oraz 0 dla mężczyzn.

$SEKTOR\_PRYWATNY_i$ – zmienna binarna przyjmująca wartość 1 w przypadku pracownika firmy z sektora prywatnego.

$NIEOKRESLONY_i$ – zmienna binarna przyjmująca wartość 1 w przypadku, gdy pracownik jest zatrudniony na czas nieokreślony.

$BEZR_{it}^{w}$ – stopa bezrobocia w okresie *t* w województwie *w*, będącym siedzibą firmy, w której pracuje *i*-ty pracownik.

$BEZR_{t}^{k}$ – stopa bezrobocia w Polsce w okresie *t*.

$\widehat{BEZR}_{it}^{w}$ – relatywna stopa bezrobocia w okresie *t* w województwie *w*, będącym siedzibą firmy, w której pracuje *i*-ty pracownik. Jest ona ilorazem stopy bezrobocia dla województwa do stopy bezrobocia dla całego kraju.

WYD_{it}^{w} – poziom wydajności pracy w okresie *t* w województwie *w*, będącym siedzibą firmy, w której pracuje *i*-ty pracownik.

$\widehat{WYD}_{it}^{w}$ – relatywny poziom wydajności pracy w okresie *t* w województwie *w*, będącym siedzibą firmy, w której pracuje *i*-ty pracownik.

Rozdział 4

ICT_KS – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w księgowości.

ICT_ZZL – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zasobami ludzkimi.

ICT_ZZ – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zaopatrzeniem.

ICT_ZP – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem produkcją.

ICT_CRM – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem sprzedażą i kontaktem z klientami.

ICT_ERP – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa.

ICT_CAD – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM.

ICT_SM – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną.

ICT_ZPAB – zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem pracami administracyjno-biurowymi.

INNOW_PROD – zmienna binarna przyjmująca wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację produktową.

INNOW_PROC – zmienna binarna przyjmująca wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację procesową lub organizacyjną.

INNOW_MARKT – zmienna binarna przyjmująca wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację marketingową.

INWESTYCJE_ICT – zmienna binarna przyjmująca wartość 1 dla firm inwestujących w technologie informatyczne i komunikacyjne.

BR – zmienna binarna przyjmująca wartość 1 dla firm posiadających własny dział B+R.

ROZMIAR – logarytm z liczby osób zatrudnionych w firmie.

WWKK – zmienna binarna przyjmująca wartość 1 w przypadku firm, w których większość kadry kierowniczej posiada wyższe wykształcenie.

WWP – zmienna binarna przyjmująca wartość 1 w przypadku firm, w których większość szeregowych pracowników posiada wyższe wykształcenie.

MSWKK – zmienna binarna przyjmująca wartość 1 w przypadku firm stosujących motywacyjny system wynagradzania kadry kierowniczej.

MSWP – zmienna binarna przyjmująca wartość 1 w przypadku firm stosujących motywacyjny system wynagradzania pracowników.

Zasieg_KZ – zmienna binarna przyjmująca wartość 1 w przypadku firm o ogólnokrajowym lub zagranicznym zasięgu oddziaływania.

Ocena_okresowa – zmienna binarna przyjmująca wartość 1 w przypadku firm prowadzących okresową ocenę kompetencji pracowników pod kątem ich przydatności do potrzeb firmy.

SZKOLENIA_TIiK – zmienna binarna przyjmująca wartość 1 w przypadku firm organizujących dodatkowe szkolenia dla pracowników w związku z wdrażaniem TIiK.

BRANZA_PRZEM – zmienna binarna przyjmująca wartość 1 w przypadku firm z branży przemysłowej.

BRANZA_BUD – zmienna binarna przyjmująca wartość 1 w przypadku firm z branży budownictwo.

BRANZA_PHU – zmienna binarna przyjmująca wartość 1 w przypadku firm z branży produkcyjno-handlowo-usługowej.

ZES_ROB – zmienna binarna przyjmująca wartość 1 dla firm, w których tworzone są zespoły robocze.

DZIEL_INF – zmienna binarna przyjmująca wartość 1 w przypadku firm, w których istnieje zwyczaj dzielenia się informacjami istotnymi dla funkcjonowania firmy z pracownikami.

NOW_UM_INF – zmienna binarna przyjmująca wartość 1 dla firm, w których wszyscy nowo przyjmowani pracownicy mają wysokie umiejętności informatyczne.

ORG – zmienna ilustrująca gotowość firmy do przeprowadzenia zmiany organizacyjnej.

WZROSTOWA – zmienna binarna przyjmująca wartość 1 dla przedsiębiorstw, które odpowiedziały, że zarówno w 2014, jak i 2013 roku przychód w firmie był większy niż w poprzednim roku.

RIS1 – udział mieszkańców (w województwie, w którym zlokalizowana jest dana firma) z wyższym wykształceniem w populacji wszystkich osób w wieku 25–64 lat.

RIS2 – wydatki na badania i rozwój w sektorze publicznym w relacji do PKB.

RIS3 – wydatki na badania i rozwój w sektorze przedsiębiorstw w relacji do PKB.

RIS4 – wydatki na innowacje firm małych i średnich (niezwiązane z wydatkami na badania i rozwój) w relacji do PKB.

RIS5 – odsetek małych i średnich przedsiębiorstw wprowadzających innowacje wewnętrzne.

RIS6 – odsetek innowacyjnych małych i średnich przedsiębiorstw współdziałających z innymi.

RIS7 – wartość zgłoszeń patentów do Europejskiego Urzędu Patentowego w relacji do PKB.

RIS8 – odsetek małych i średnich przedsiębiorstw wprowadzających innowacje produktowe lub procesowe.

RIS9 – odsetek małych i średnich przedsiębiorstw wprowadzających innowacje marketingowe.

RIS10 – odsetek zatrudnionych w przemysłach wysokiej i średniowysokiej technologii oraz usługach opartych na wiedzy.

RIS11 – relacja sprzedaży produktów stanowiących innowacje nowe dla firmy lub nowe dla rynku do całkowitych obrotów.

$\boldsymbol{x}_i^{ZP}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych z zarządzaniem produkcją.

$\boldsymbol{x}_i^{ERP}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa.

$\boldsymbol{x}_i^{CAD}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM.

$\boldsymbol{x}_i^{SM}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wykorzystywania TIiK w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną.

$\boldsymbol{x}_i^{INW}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do inwestowania w rozwój technologii informacyjnych i komunikacyjnych.

$\boldsymbol{x}_i^{BR}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do posiadania własnego wydziału B+R.

$\boldsymbol{x}_i^{PROD}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji produktowych.

$\boldsymbol{x}_i^{PROC}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji procesowych lub organizacyjnych.

$\boldsymbol{x}_i^{MARK}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym skłonność do wprowadzania innowacji marketingowych.

$\boldsymbol{x}_i^{PRODUKT}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym produktywność.

Rozdział 5

PR – zmienna dychotomiczna przyjmująca wartość 1 w przypadku osoby doświadczającej problemu prawnego oraz 0 w przeciwnym przypadku.

LBUL – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa budowlanego.

LCIV – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa cywilnego.

LROA – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa drogowego.

LPEN – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa karnego.

LCON – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa konsumenckiego.

LDAM – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy odszkodowań.

LFIN – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy problemów finansowych.

LFAM – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa rodzinnego.

LJOB – zmienna binarna przyjmująca wartość 1, jeśli problem prawny doświadczany przez respondenta dotyczy prawa pracy.

LVAL – zmienna binarna przyjmująca wartość 1, jeśli respondent uznał, że problem prawny, którego doświadczył, jest ważny.

FEM – zmienna binarna przyjmująca wartość 1 w przypadku kobiet oraz 0 dla mężczyzn.
AGE – wiek respondenta.
EDU1 – zmienna binarna przyjmująca wartość 1 w przypadku respondenta z wykształceniem ponadpodstawowym.
EDU2 – zmienna binarna przyjmująca wartość 1 w przypadku respondenta z wykształceniem ponadśrednim.
SCZO – zmienna binarna przyjmująca wartość 1, jeśli respondent jest żonatym mężczyzną lub zamężną kobietą.
SCRO – zmienna binarna przyjmująca wartość 1, jeśli respondent jest osobą rozwiedzioną.
SCWD – zmienna binarna przyjmująca wartość 1, jeśli respondent jest wdowcem lub wdową.
NFAM – liczba osób w gospodarstwie domowym.
DOCH – dochód na osobę w gospodarstwie domowym.
RESD – zmienna binarna przyjmująca wartość 1, jeśli respondent mieszka w mieście powyżej 20 000 mieszkańców.
SLAB – zmienna binarna przyjmująca wartość 1 w przypadku respondentów niezatrudnionych lub pracujących dorywczo.
PAWR – zmienna ilustrująca poziom świadomości prawnej u osoby ankietowanej.
PCPL – zmienna binarna przyjmująca wartość 1 w przypadku respondentów twierdzących, że zawsze należy przestrzegać prawa.
PUSE – zmienna związana z postawą wobec stosowania prawa.
PTRU – zmienna ilustrująca poziom zaufania do palestry.
PAVA – zmienna binarna ilustrująca subiektywną ocenę dostępności usług prawnych.
APAR – zmienna binarna przyjmująca wartość 1, jeśli respondent przynależy do organizacji społecznych.
AACT – zmienna binarna przyjmująca wartość 1 w przypadku osób, które pozytywnie odpowiedziały na pytanie dotyczące działalności społecznej.
$\boldsymbol{x}_i^{PR}$ – wektor zmiennych objaśniających w równaniu wyjaśniającym prawdopodobieństwo doświadczenia problemu prawnego przez respondenta.

Notacja dla zbieżności

$\xrightarrow{d}$ – zbieżność według rozkładu.
$\xrightarrow{p}$ – zbieżność według prawdopodobieństwa.

Inne funkcje, parametry i macierze

m(end) – liczba endogenicznych regresorów w endogenicznym modelu probitowym.

1. Podstawowe modele wykorzystujące dane indywidualne

1.1. Wprowadzenie

Zanim zaprezentowany zostanie sposób wykorzystania danych regionalnych w modelach mikroekonometrycznych krótko omówione będą główne metody służące estymacji parametrów, gdy obserwacjami są firmy, gospodarstwa domowe czy indywidualni respondenci. Dlatego też niniejszy rozdział poświęcony jest metodom wykorzystywanym podczas analizy zależności na poziomie indywidualnym. Należy podkreślić, że rozdział ten ma charakter wprowadzający do analiz problemów, w których przyjmuje się założenie, że badacz dysponuje zarówno informacjami na poziomie mikro, jak i mezo. Zagadnienia omawiane w tym rozdziale mają charakter wspomagający. Prezentowane w kolejnych podrozdziałach metody są w dużej części znane badaczom wykorzystującym statystykę i ekonometrię. Są one też dokładnie omówione w klasycznych podręcznikach i książkach z zakresu ekonometrii czy mikroekonometrii (por. Marzec, 2008; Welfe, 2009; Gruszczyński, 2012; Maddala, 2013; Wiśniewski, 2015). Dlatego też metody ekonometryczne wykorzystywane do analizy zależności na poziomie indywidualnym zostaną przedstawione skrótowo.

W podrozdziale 1.2 prezentowane są modele, w których zmienna zależna ma charakter ciągły. Dlatego też omawiane są między innymi klasyczna metoda najmniejszych kwadratów, uogólniona metoda najmniejszych kwadratów, metoda regresji kwantylowej, a także metody estymacji odpornej (np. wykorzystanie estymatora *M*, zastosowanie estymatora *MM*). Następnie w podrozdziale 1.3 omawiana jest metoda estymacji parametrów w przypadku, gdy zmienna zależna ma charakter dwumianowy (dychotomiczny). W podrozdziale 1.4 rozważany jest przypadek uporządkowanej zmiennej zależnej, natomiast w podrozdziale 1.5 omówiony jest podstawowy model dla polichotomicznej, nieuporządkowanej zmiennej zależnej.

W podrozdziale 1.6 opisany jest model regresji rankingowej. Następnie w podrozdziale 1.7 prezentowane są modele licznikowe. Dwurównaniowy oraz wielorównaniowy model probitowy prezentowane są odpowiednio w podrozdziałach 1.8 i 1.9. Przypadek estymacji parametrów modelu probitowego z endogenicznymi regresorami rozważany jest w podrozdziale 1.10.

1.2. Modele dla ciągłej zmiennej zależnej

1.2.1. Klasyczny model regresji liniowej

Ogólny model regresji z addytywnym składnikiem losowym przyjmuje postać:

$$\boldsymbol{y} = E(\boldsymbol{Y}|\boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{1.1}$$

gdzie $\boldsymbol{Y}$ oznacza zmienną, a $\boldsymbol{y}$ to obserwacje pochodzące z rozkładu zmiennej $\boldsymbol{Y}$. Wektor $E(\boldsymbol{y}|\boldsymbol{X})$ składa się z warunkowych wartości oczekiwanych zmiennej $\boldsymbol{y}$ względem wartości zmiennych wchodzących w skład macierzy $\boldsymbol{X}$. $\boldsymbol{\varepsilon}$ jest wektorem nieobserwowalnych składników losowych.

Przy założeniu, że zależność między zmienną zależną a regresorami ma charakter liniowy, model regresji przyjmuje postać:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \ldots., I, \tag{1.2}$$

gdzie $\boldsymbol{x}_i$ jest wektorem wierszowym składającym się z wartości zmiennych objaśniających dla i-tego obiektu, natomiast $\boldsymbol{\beta}$ jest kolumnowym wektorem parametrów do oszacowania. Przyjmowane są następujące założenia:

1) model jest niezmienniczy, co oznacza że zależność (1.2) jest taka sama dla wszystkich obserwacji,
2) składnik losowy $\boldsymbol{\varepsilon}_i$ ma zerową wartość oczekiwaną oraz stałą wariancję,
3) kowariancja między składnikami losowymi dla różnych obiektów jest zerowa,
4) kowariancja między każdym elementem wchodzącym w skład wektora $\boldsymbol{x}_i$ a składnikiem losowym ε_i jest równa 0,
5) wariancja składnika losowego jest taka sama dla wszystkich obiektów.

Najczęściej wykorzystywaną metodą estymacji parametrów modelu (1.2) jest klasyczna metoda najmniejszych kwadratów. Estymator uzyskany tą metodą przyjmuje postać:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.3}$$

Jeśli wymienione wyżej założenia są spełnione, wówczas estymator opisany wzorem (1.3) jest nieobciążonym i najefektywniejszym estymatorem (w klasie estymatorów liniowych i nieobciążonych) wektora $\boldsymbol{\beta}$ (por. Welfe, 2009). Przy spełnieniu powyższych założeń (będących częścią schematu Gaussa-Markowa) estymator klasycznej metody najmniejszych kwadratów także jest zgodny. Jego rozkład asymptotyczny definiowany jest następująco:

$$\sqrt{I}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{M}_{\boldsymbol{xx}}^{-1}\boldsymbol{M}_{\boldsymbol{x\Sigma x}}\boldsymbol{M}_{\boldsymbol{xx}}^{-1}), \tag{1.4}$$

gdzie:

$$\boldsymbol{M}_{\boldsymbol{xx}} = lim\frac{1}{I}\sum_{i=1}^{I} E(\boldsymbol{x}_i\boldsymbol{x}_i^T),$$

$$\boldsymbol{M}_{\boldsymbol{x\Sigma x}} = lim\frac{1}{I}\sum_{i=1}^{I} E(\varepsilon_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T),$$

natomiast $\xrightarrow{d}$ oznacza zbieżność według rozkładu.

1.2.2. Heteroskedastyczność składnika losowego. Metody estymacji parametrów w przypadku niestałej wariancji

W modelach ekonometrycznych wykorzystujących dane przekrojowe często pojawia się problem heteroskedastyczności składnika losowego. Polega on na tym, że wariancja $var(\varepsilon_i)$ różni się ze względu na obiekty. Wówczas macierz wariancji-kowariancji między składnikami losowymi przyjmuje postać:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\boldsymbol{X}) = \boldsymbol{\Sigma}, \tag{1.5}$$

gdzie $\boldsymbol{\Sigma}$ jest nieosobliwą macierzą diagonalną o zróżnicowanych elementach na diagonali. Zdefiniowany wzorem (1.3) estymator klasycznej metody najmniejszych kwadratów traci efektywność. Uzyskanie efektywnego estymatora umożliwia zastosowanie uogólnionej metody najmniejszych kwadratów. Estymator UMNK definiowany jest w następujący sposób:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \tag{1.6}$$

gdzie *GLS* oznacza *generalized least squares*. W praktyce elementy macierzy $\boldsymbol{\Sigma}$ nie są znane, więc bezpośrednie zastosowanie estymatora zdefiniowanego wzorem

(1.6) nie jest możliwe. Dlatego też przyjmując, że $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\breve{\boldsymbol{\gamma}})$, gdzie $\breve{\boldsymbol{\gamma}}$ jest wektorem parametrów o skończonym wymiarze, należy wykorzystać zgodny estymator $\widehat{\breve{\boldsymbol{\gamma}}}$ wektora parametrów $\breve{\boldsymbol{\gamma}}$, a następnie szacować parametry za pomocą uogólnionej metody najmniejszych kwadratów z estymacją. Odpowiedni estymator przyjmuje postać:

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}, \tag{1.7}$$

gdzie *FGLS* oznacza *feasible generalized least squares*.

Szczególnym przypadkiem uogólnionej metody najmniejszych kwadratów jest ważona metoda najmniejszych kwadratów. Jeśli macierz $\widehat{\boldsymbol{\Sigma}}$ jest diagonalna, mamy do czynienia z ważoną metodą najmniejszych kwadratów. Odpowiedni estymator przyjmuje postać:

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left(\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}{}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}{}^{-1}\boldsymbol{y}, \tag{1.8}$$

gdzie elementami macierzy diagonalnej $\widehat{\boldsymbol{\Sigma}}_w$ są oceny wariancji składnika losowego dla poszczególnych obserwacji.

Ponieważ heteroskedastyczność składnika losowego niesie ze sobą problem obciążoności estymatora macierzy wariancji-kowariancji estymatorów parametrów, rozważa się korektę tego obciążenia. Halbert White (1980) zaproponował wyznaczenie odpornych na heteroskedastyczność błędów standardowych w modelu regresji liniowej. W tym celu wykorzystywany jest fakt, że:

$$\boldsymbol{M}_{\boldsymbol{x\Sigma x}} = \text{plim} I^{-1}\sum_{i=1}^{I}\varepsilon_i^2\,\boldsymbol{x}_i^T\boldsymbol{x}_i. \tag{1.9}$$

Ponieważ $\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \boldsymbol{\beta}$, zachodzi zależność $y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{OLS} \xrightarrow{p} \varepsilon_i$, gdzie $\xrightarrow{p}$ oznacza zbieżność według prawdopodobieństwa. Dlatego też zgodny estymator elementu $\boldsymbol{M}_{\boldsymbol{x\Sigma x}}$ z wyrażenia (1.4) przyjmuje następującą postać:

$$\widehat{\boldsymbol{M}}_{\boldsymbol{x\Sigma x}} = I^{-1}\sum_{i=1}^{I}\hat{e}_i^2\,\boldsymbol{x}_i^T\boldsymbol{x}_i, \tag{1.10}$$

gdzie $\hat{e}_i$ oznacza resztę. Wykorzystanie metody White'a (1980) i wyznaczenie odpornych na heteroskedastyczność błędów standardowych polega na estymacji parametrów modelu regresji klasyczną metodą najmniejszych kwadratów (por. wzór 1.3), a następnie zastosowaniu wzoru (1.10) w celu wyznaczenia oszacowań elementów macierzy wariancji-kowariancji estymatorów parametrów.

1.2.3. Metoda regresji kwantylowej

Analizując kategorie ekonomiczne, badacze często są zainteresowani nie tylko wartością średnią czy jej miarą zróżnicowania, ale także rozkładem. Na przykład znajomość mediany czy kwantyli dla wynagrodzeń w określonej populacji pozwala lepiej oceniać poziom zamożności niż wiedza na temat średniej wartości cechy. Analogiczny przypadek dotyczy analizy zależności między cechami. Jeśli badacz oczekuje, że wpływ zmiennych objaśniających na zmienną objaśnianą jest różny w poszczególnych kwantylach rozkładu zmiennej zależnej, wówczas zastosowanie metody regresji kwantylowej jest uzasadnione.

Dzięki wykorzystaniu omawianej metody badacz uzyskuje pogłębiony obraz zależności między cechami. Może się bowiem okazać, że w różnych kwantylach rozkładu zmiennej zależnej inne zmienne objaśniające istotnie wpływają na regresanta. Oprócz tego estymator metody regresji kwantylowej jest bardziej odporny w przypadku pojawienia się obserwacji nietypowych w porównaniu z estymatorem MNK. Dodatkowo estymator metody regresji kwantylowej jest zgodny przy słabszych (w porównaniu z przypadkiem estymatora MNK) założeniach dotyczących struktury stochastycznej (Koenker, Bassett, 1978).

Punktem wyjścia do wyprowadzenia estymatora regresji kwantylowej jest zdefiniowanie q-tego kwantyla w populacji dla rozkładu zmiennej ciągłej y:

$$q = P\left(y \leq \mu_q\right) = F_y\left(\mu_q\right), \tag{1.11}$$

gdzie F_y oznacza dystrybuantę rozkładu zmiennej y, natomiast μ_q jest wartością kwantyla rzędu q, czyli:

$$\mu_q = F_y^{-1}\left(q\right). \tag{1.12}$$

Rozważania dotyczące wyznaczania kwantyli rozkładu można osadzić w kontekście modelu regresji. Załóżmy, że kwantyl rzędu q rozkładu zmiennej y, warunkowy ze względu na wartości przyjmowane przez zmienne z wektora (regresorów) $\boldsymbol{x}$, oznaczymy przez $\mu_q(\boldsymbol{x})$. Wówczas wartość kwantyla zależy od dystrybuanty rozkładu warunkowego y względem $\boldsymbol{x}$ $F_{y|\boldsymbol{x}}$ w następujący sposób:

$$\mu_q\left(\boldsymbol{x}\right) = F_{y|\boldsymbol{x}}^{-1}\left(q\right). \tag{1.13}$$

Zakładając, że kształtowanie się zmiennej y jako funkcji zmiennych wchodzących w skład wektora $\boldsymbol{x}$ opisuje równanie (1.2), funkcja celu (funkcja minimalizująca średnie odchylenie wartości empirycznej od teoretycznej) w metodzie regresji kwantylowej przyjmuje postać:

$$Q_I(\boldsymbol{\beta}_{]q[}) = \sum_{i:\, y_i \geq \boldsymbol{x}_i\boldsymbol{\beta}_{]q[}}^{I} q\left|y_i - \boldsymbol{x}_i\boldsymbol{\beta}_{]q[}\right| + \sum_{i:\, y_i < \boldsymbol{x}_i\boldsymbol{\beta}_{]q[}}^{I}(1-q)\left|y_i - \boldsymbol{x}_i\boldsymbol{\beta}_{]q[}\right|, \tag{1.14}$$

gdzie $|\cdot|$ oznacza wartość bezwzględną. W przypadku regresji kwantylowej wpływ zmiennych objaśniających na zmienną zależną nie jest taki sam w różnych kwantylach rozkładu zmiennej objaśnianej. Dlatego też zamiast niezmiennego wektora parametrów $\boldsymbol{\beta}$ definiowany jest zależny od kwantyla rozkładu zmiennej zależnej wektor $\boldsymbol{\beta}_{]q[}$. Estymator metody regresji kwantylowej definiowany jest następująco:

$$\widehat{\boldsymbol{\beta}}_Q = arg\, min Q_I(\boldsymbol{\beta}_{]q[}). \tag{1.15}$$

Po dokonaniu minimalizacji funkcji (1.14) dla różnych kwantyli rozkładu zmiennej zależnej należy porównać oszacowania parametrów dla poszczególnych q. Jeśli okazuje się, że te oszacowania nie różnią się od siebie znacząco, wówczas zastosowanie klasycznej metody estymacji parametrów jest bardziej uzasadnione niż wykorzystanie metody regresji kwantylowej. Wynika to z faktu, że estymator metody najmniejszych kwadratów jest najefektywniejszy. Dlatego też po oszacowaniu parametrów metodą regresji kwantylowej na ogół weryfikowana jest następująca hipoteza:

$$\begin{aligned} H_0&: \boldsymbol{\beta}_{]0,25[} = \boldsymbol{\beta}_{]0,75[}, \\ H_1&: \boldsymbol{\beta}_{]0,25[} \neq \boldsymbol{\beta}_{]0,75[}. \end{aligned} \tag{1.16}$$

Odrzucenie hipotezy zerowej oznacza, że oszacowania parametrów w pierwszym oraz trzecim kwantylu rozkładu zmiennej zależnej różnią się od siebie istotnie. Oznacza to zatem, że zastosowanie analizowanej metody estymacji parametrów jest uzasadnione. W przeciwnym przypadku lepiej zastosować klasyczne metody estymacji parametrów.

1.2.4. Odporna estymacja parametrów modelu regresji Estymator *M*. Estymator *S*. Estymator *MM*

Omówiona w punkcie 1.2.3 metoda regresji kwantylowej ma szereg zalet. Po pierwsze, dzięki jej zastosowaniu możliwa jest identyfikacja różniącego się – ze względu na kwantyle rozkładu zmiennej zależnej – wpływu regresorów na regresanta. Oprócz tego metoda regresji kwantylowej jest lepsza w przypadku pojawienia się problemu obserwacji nietypowych. Największym problemem związanym

z zastosowaniem metody omówionej w punkcie 1.2.3 jest niska efektywność estymatora. Jak pokazał Peter Huber (1964), w sytuacji normalnego rozkładu składnika losowego efektywność estymatora regresji medianowej (kwantylowej dla $q = 0{,}5$) jest równa 64% efektywności estymatora MNK.

Niedoskonałości związane z zastosowaniem kwantylowej metody estymacji parametrów przyczyniły się do powstania metod, które są odporne w sytuacji pojawienia się obserwacji nietypowych oraz prowadzą do uzyskania estymatorów efektywniejszych w porównaniu z omówionym w punkcie 1.2.3. Tymi estymatorami odpornymi są estymator *S*, estymator *M* oraz estymator *MM*. Zostaną one omówione w niniejszym punkcie.

Problem braku odporności estymatora KMNK na fakt pojawienia się obserwacji nietypowych w klasycznym modelu regresji liniowej wynika ze sposobu zdefiniowania funkcji kary. Formułę (1.3) można w alternatywny sposób zapisać następująco:

$$\widehat{\boldsymbol{\beta}}_{OLS} = argmin\left(\sum_{i=1}^{I}\left(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{OLS}\right)^2\right), \tag{1.17}$$

gdzie *OLS* oznacza *ordinary least squares*.

Ponieważ funkcja kary ma postać kwadratową, linia regresji próbuje dopasowywać się w taki sposób, aby żadna z obserwacji nie znajdowała się daleko od niej. Formułę definiującą estymator regresji medianowej, będącej szczególnym przypadkiem regresji kwantylowej, można zapisać następująco:

$$\widehat{\boldsymbol{\beta}}_{]q=0,5[} = argmin\left(\sum_{i=1}^{I}\left|y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_{]q=0,5[}\right|\right). \tag{1.18}$$

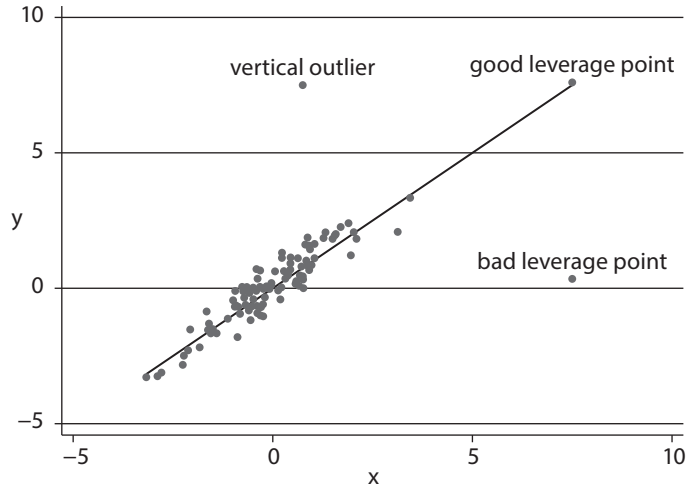
Jak widać „kara" związana z pojawieniem się wartości empirycznej „daleko" od teoretycznej jest zdecydowanie mniejsza niż w przypadku estymatora KMNK. Dlatego też estymator regresji medianowej, inaczej zwany estymatorem metody najmniejszych modułów (por. Welfe, 2009), jest bardziej odporny na problem pojawienia się obserwacji nietypowych.

Zanim zaprezentowane zostaną odporne metody estymacji parametrów należy wymienić różne typy obserwacji nietypowych. Jak wskazują Peter Rousseeuw oraz Annick Leroy (2005), w analizie regresji wyróżnia się trzy rodzaje obserwacji nietypowych, wpływających na oszacowania uzyskane klasyczną metodą najmniejszych kwadratów:

1) obserwacje nietypowe pionowe,
2) złe dźwignie,
3) dobre dźwignie.

Obserwacje nietypowe pionowe charakteryzują się tym, że wartości zmiennej objaśnianej *y* są zdecydowanie wyższe lub niższe niż wartości tej zmiennej dla innych

obserwacji, natomiast wartości dla zmiennych objaśniających nie znajdują się w ogonach ich rozkładów. Występowanie pionowych obserwacji nietypowych ma znaczny wpływ na oszacowanie wyrazu wolnego. W przypadku złych dźwigni mamy do czynienia z odbiegającymi od „standardowych" wartościami przyjmowanymi przez zmienne objaśniające oraz „typowymi" wartościami dla zmiennej zależnej. Takie obserwacje silnie wpływają na oszacowania estymatora MNK zarówno dla wyrazu wolnego, jak i pozostałych zmiennych objaśniających. Dobre dźwignie to obserwacje, które znajdują się na linii regresji, jednak wartości przyjmowane dla nich zarówno przez zmienną objaśnianą, jak i regresory są zdecydowanie wyższe lub niższe od typowych. Rysunek 1 ilustruje ideę różnych rodzajów obserwacji nietypowych.



Rysunek 1. Rodzaje obserwacji nietypowych w modelu regresji

Źródło: Verardi, Croux, 2009.

Huber (1964) zaproponował uogólnienie metody najmniejszych modułów (estymatora regresji medianowej danego wzorem 1.18), wyprowadzając estymator *M*, który definiowany jest następująco:

$$\hat{\boldsymbol{\beta}}_M = arg\ \min \sum_{i=1}^{I} \rho\rho\left(\frac{y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_M}{\sigma}\right), \tag{1.19}$$

gdzie $\rho\rho(\cdot)$ jest funkcją kary spełniającą następujące własności:

1) jest parzysta,
2) jest niemalejąca dla dodatnich argumentów,
3) rośnie wolniej w porównaniu z funkcją kwadratową.

W celu uniknięcia sytuacji wzrostu reszt wraz ze wzrostem wartości przyjmowanych przez zmienne objaśniające reszty są standaryzowane przez parametr σσ ilustrujący poziom dyspersji. Wzór definiujący postać funkcji straty jest często następujący:

$$\rho\rho(u) = \{(1 - [1 - (u/\kappa\kappa)^2])^3\} * I\{|u| \leq \kappa\kappa\} + I\{|u| > \kappa\kappa\}, \quad (1.20)$$

gdzie $\kappa\kappa = 4,685$. Estymator początkowy $\widetilde{\boldsymbol{\beta}}_M$ jest wyznaczany na podstawie następującej funkcji straty:

$$\rho\rho(u) = \{0,5(u)^2\} * I\{|u| \leq cc\} + \{cc|u| - 0,5cc^2\} * I\{|u| > cc\}, \quad (1.21)$$

gdzie *cc* = 1,345. Takie postacie funkcji straty wykorzystywane są między innymi w programie STATA podczas estymacji parametrów za pomocą estymatora *M*. Rezultaty uzyskane za pomocą estymatora *M* często nie są zadowalające. Okazuje się bowiem, że zastosowanie analizowanej metody rozwiązuje problem związany z odpornością jedynie w przypadku izolowanych obserwacji nietypowych. Jak wskazują między innymi Peter Rousseeuw oraz Bert van Zomeren (1990), metoda ta nie radzi sobie z problemem w przypadku, gdy obserwacje nietypowe występują w klastrach. Oprócz tego, jak wskazują na przykład Vincenzo Verardi i Christophe Croux (2009), w przypadku gdy obserwacje nietypowe mają charakter złych dźwigni, algorytm optymalizacyjny może być zbieżny nie do globalnego, lecz do lokalnego minimum.

Uwzględniając fakt, że estymator KMNK jest wyznaczany na podstawie minimalizacji wariancji reszt, Peter Rousseeuw oraz Victor Yohai (1987) zaproponowali estymację polegającą na minimalizacji miary zróżnicowania reszt, która jest mniej (niż wariancja) wrażliwa na wartości ekstremalne. Odporna miara zróżnicowania oznaczana jest przez $\widehat{\sigma\sigma}^S$. Spełnia ona następującą własność:

$$\frac{1}{I}\sum_{i=1}^{I}\rho\rho\left(\frac{y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_S}{\hat{\sigma}_S}\right) = b, \quad (1.22)$$

gdzie $b = E\left[\rho\rho\left(Z\right)\right]$, natomiast $Z \sim N\left(0,1\right)$. Wartość $\hat{\boldsymbol{\beta}}_S$ minimalizująca $\hat{\sigma}_S$ jest nazywana estymatorem *S*. Formalnie jest on definiowany następująco:

$$\widehat{\boldsymbol{\beta}}_S = argmin\ \hat{\sigma}^S\left(y_1 - \boldsymbol{x}_1\widehat{\boldsymbol{\beta}}_S, \ldots, y_I - \boldsymbol{x}_I\widehat{\boldsymbol{\beta}}_S\right), \quad (1.23)$$

gdzie $\hat{\sigma}^S$ jest estymatorem zdefiniowanym za pomocą formuły (1.22). Jeśli chodzi o funkcję straty $\rho\rho(\cdot)$, najczęściej przyjmowana jest zmodyfikowana postać (1.20) z $\kappa\kappa = 1,547$. Od wartości parametru $\kappa\kappa$ zależy efektywność oraz umiejętność wykrywania obserwacji nietypowych. Kosztem wysokiej efektywności wykrywania obserwacji nietypowych jest niska efektywność. Taka sytuacja ma miejsce w przypadku niskiej wartości parametru $\kappa\kappa$. Wraz ze wzrostem wartości parametru $\kappa\kappa$ następuje wzrost efektywności, natomiast obniża się wykrywalność obserwacji nietypowych.

Aby jednocześnie zapewnić wysoką wykrywalność obserwacji nietypowych oraz wysoką efektywność estymatora, Yohai (1987) zaproponowali odporny estymator *MM*. Definiuje go następująca formuła:

$$\hat{\boldsymbol{\beta}}_{MM} = arg\ \min \sum_{i=1}^{I} \rho\rho\left(\frac{y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{MM}}{\hat{\sigma}_S} \right), \tag{1.24}$$

gdzie wartość estymatora dyspersji $\hat{\sigma}_S$ jest ustalona.

1.3. Model dwumianowy (dychotomiczny)

W przypadku kiedy zmienna zależna jest zero-jedynkowa, wyjaśnianie zależności między nią a innymi kategoriami ekonomiczno-społecznymi odbywa się dzięki estymacji parametrów modelu dwumianowego. Przyjmuje on następującą postać:

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i,\ \varepsilon_i \sim F, \tag{1.25a}$$

$$y_i = I\left\{ y_i^* > 0 \right\}, \tag{1.25b}$$

gdzie *i* indeksuje obserwacje, $I\{\cdot\}$ oznacza zmienną wskaźnikową przyjmującą wartość 1, gdy warunek zdefiniowany w klamrowym nawiasie jest spełniony oraz 0 w przeciwnym przypadku, natomiast $x_{1i}, \ldots, x_{Ki}$ są zmiennymi objaśniającymi. Dla zmiennej przyjmującej dwie wartości naturalne jest zdefiniowanie prawdopodobieństw, że wynosi ona 1 lub 0. Prawdopodobieństwo, że zmienna obserwowalna y_i przyjmuje wartość 1 można zapisać następująco:

$$P(y_i = 1) = P(y_i^* > 0) = P(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i > 0) = \\ P(\varepsilon_i > -\beta_0 - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}) = 1 - F(-\beta_0 - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}), \tag{1.26}$$

gdzie F oznacza dystrybuantę rozkładu składnika losowego. Najczęściej przyjmuje się, że składnik losowy pochodzi z rozkładu normalnego lub logistycznego. Mamy wówczas do czynienia odpowiednio z modelem probitowym lub logitowym. Normalny lub logistyczny rozkład składnika losowego jest symetryczny, co powoduje że równanie (1.26) można zapisać następująco:

$$P(y_i = 1) = F(\boldsymbol{x}_i\boldsymbol{\beta}), \tag{1.27}$$

gdzie $\boldsymbol{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki}$.

Alternatywnym niesymetrycznym rozkładem składnika losowego jest komplementarny rozkład log-log. Z tym rozkładem związany jest model komplementarny log-log. Tabela 1 zawiera wzory dla dystrybuant poszczególnych rozkładów.

Tabela 1. Rozkłady składnika losowego dla różnych modeli dwumianowych

Model	Prawdopodobieństwo
Logitowy	$F(\boldsymbol{x}_i\boldsymbol{\beta}) = \dfrac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta})}$
Probitowy	$F(\boldsymbol{x}_i\boldsymbol{\beta}) = \int_{-\infty}^{\boldsymbol{x}_i\boldsymbol{\beta}} (2\pi)^{-1/2} \exp\left(-\dfrac{t^2}{2}\right) dt$
Komplementarny log-log	$F(\boldsymbol{x}_i\boldsymbol{\beta}) = 1 - \exp[-\exp(\boldsymbol{x}_i\boldsymbol{\beta})]$

Źródło: Cameron, Triviedi, 2009; Gruszczyński, 2012.

Po przyjęciu założeń dotyczących rozkładu składnika losowego szacuje się parametry metodą największej wiarygodności. W najbardziej klasycznym przypadku zakłada się niezależność rozkładów składników losowych dla kolejnych jednostek. Po skonstruowaniu funkcji wiarygodności, jej logarytmu i zróżniczkowaniu względem parametrów $\beta_0, \beta_1, \ldots, \beta_K$ uzyskuje się następujący układ równań normalnych:

$$\sum_{i=1}^{I} \frac{y_i - F(\boldsymbol{x}_i\boldsymbol{\beta})}{F(\boldsymbol{x}_i\boldsymbol{\beta})\left(1 - F(\boldsymbol{x}_i\boldsymbol{\beta})\right)} F'(\boldsymbol{x}_i\boldsymbol{\beta})\boldsymbol{x}_i^T = \boldsymbol{0}, \tag{1.28}$$

gdzie $F'(z) = \dfrac{\partial F(z)}{\partial z}$.

Nie istnieje analityczne rozwiązanie dla układu równań (1.28), więc estymator największej wiarygodności uzyskiwany jest za pomocą metod numerycznych. Zastosowanie iteracyjnego algorytmu Newtona-Raphsona zapewnia szybkie znalezienie oszacowań największej wiarygodności, ponieważ funkcja wiarygodności zarówno dla modelu logitowego, jak i probitowego jest globalnie wypukła. Estymator największej wiarygodności jest zgodny, jeśli teoretyczny rozkład składnika losowego jest zgodny z empirycznym. Oszacowanie macierzy wariancji-kowariancji estymatorów-parametrów przyjmuje postać:

$$E\left[\left(\hat{\boldsymbol{\beta}}_{ML}-\boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{ML}-\boldsymbol{\beta}\right)^{T}\right]=\left(\sum_{i=1}^{I}\frac{F'\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{ML}\right)}{F\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{ML}\right)\left(1-F\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{ML}\right)\right)}\boldsymbol{x}_i^{T}\boldsymbol{x}_i\right)^{-1}. \quad (1.29)$$

Po znalezieniu oszacowań parametrów oraz elementów macierzy wariancji-kowariancji możliwa jest weryfikacja hipotez dotyczących istotności wpływu poszczególnych zmiennych objaśniających na zmienną zależną.

W modelu dwumianowym oszacowania parametrów nie mają takiej interpretacji jak w przypadku klasycznego modelu regresji liniowej zdefiniowanego dla poziomów lub logarytmów zmiennych. Niemniej jednak po dokonaniu estymacji parametrów modelu logitowego możliwa jest interpretacja funkcji oszacowań w kontekście wpływu jednostkowej zmiany wartości zmiennej objaśniającej na iloraz szans. Rozważmy dla uproszczenia model logitowy z wyrazem wolnym oraz jedną zmienną objaśniającą. Wówczas:

$$P(y_i=1)=\frac{\exp(\beta_0+\beta_1x_{1i})}{1+\exp(\beta_0+\beta_1x_{1i})}=\Lambda(\beta_0+\beta_1x_{1i}). \quad (1.30)$$

A zatem iloraz szans definiuje się następująco:

$$\frac{P(y_i=1)}{P(y_i=0)}=\frac{\exp(\beta_0+\beta_1x_{1i})/\left(1+\exp(\beta_0+\beta_1x_{1i})\right)}{1/\left(1+\exp(\beta_0+\beta_1x_{1i})\right)}=\exp(\beta_0+\beta_1x_{1i}). \quad (1.31)$$

Rozważmy przypadek wzrostu wartości zmiennej x_{1i} o jednostkę – z poziomu c do poziomu $c + 1$. Wówczas zmianę ilorazu szans można zapisać następująco:

$$\frac{\dfrac{P\left(y_i=1\mid x_{1i}=c+1\right)}{P\left(y_i=0\mid x_{1i}=c+1\right)}}{\dfrac{P\left(y_i=1\mid x_{1i}=c\right)}{P\left(y_i=0\mid x_{1i}=c\right)}}=\frac{\exp\left(\beta_0+\beta_1 c+\beta_1\right)}{\exp\left(\beta_0+\beta_1 c\right)}=\exp\left(\beta_1\right). \tag{1.32}$$

Dlatego też jeśli w wyniku estymacji parametrów modelu logitowego uzyskane zostanie oszacowanie parametru wynoszące $\hat{\beta}_1$, należy je interpretować następująco: jednostkowy wzrost wartości zmiennej x_1 powoduje – przy innych czynnikach niezmienionych – wzrost ilorazu szans o $\left(\exp\left(\hat{\beta}_1\right)-1\right)\cdot 100\%$.

W celu oceny wpływu zmiany wartości zmiennych objaśniających na prawdopodobieństwo, że zmienna zależna przyjmuje wartość 1, wykorzystywane są tak zwane efekty krańcowe. Oblicza się je w następujący sposób:

$$\frac{\partial P\left(y_i=1\mid \mathbf{x}_i\right)}{\partial x_{ik}}=F'\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\hat{\beta}_k. \tag{1.33}$$

Tabela 2 prezentuje efekty krańcowe dla różnych rozkładów składnika losowego.

Tabela 2. Efekty krańcowe dla różnych rozkładów składnika losowego

Model	Efekt krańcowy
Logitowy	$\Lambda\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\left(1-\Lambda\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\right)\hat{\beta}_k$
Probitowy	$\phi\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\hat{\beta}_k$
Komplementarny log-log	$\exp\left(-\exp\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\right)\exp\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\hat{\beta}_k$

Źródło: Cameron, Triviedi, 2009; Gruszczyński, 2010.

Oprócz efektów krańcowych dla średnich wartości zmiennych objaśniających często oblicza się także tak zwane średnie efekty krańcowe (*average marginal effects* – AME) w następujący sposób:

$$\overline{\frac{\partial P\left(y=1\mid \overline{x}\right)}{\partial \overline{x}_k}}=\sum_{i=1}^{I}\frac{F'\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\hat{\beta}_k}{I}. \tag{1.34}$$

Współczynnik determinacji jest standardowym miernikiem jakości dopasowania w modelu regresji liniowej. W przypadku modelu dwumianowego oraz innych modeli zmiennych jakościowych często wykorzystywany jest alternatywny miernik, czyli pseudo-R^2. Zaproponowany przez Daniela McFaddena (1974) miernik przyjmuje następującą postać:

$$R^2_{binary} = 1 - \frac{\ln L\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}, \boldsymbol{y}\right)}{\ln L\left(\hat{\beta}_0 \mid \boldsymbol{X}, \boldsymbol{y}\right)} = 1 - \frac{\sum_{i=1}^{I}\left\{y_i \ln F\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right) + \left(1-y_i\right)\ln\left(1-F\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)\right)\right\}}{I\left[\bar{y}\ln\left(\bar{y}\right) + \left(1-\bar{y}\right)\ln\left(\bar{y}\right)\right]}. \quad (1.35)$$

Jak widać porównywana jest wartość funkcji wiarygodności dla modelu zawierającego wszystkie zmienne objaśniające z analogiczną wartością dla modelu zawierającego jedynie wyraz wolny.

Innym miernikiem często wykorzystywanym podczas oceniania jakości dopasowania modelu do danych empirycznych, jest następujący współczynnik determinacji zaproponowany przez Moshego Ben-Akivę i Stevena R. Lermana (1985) oraz Richarda Kaya i Sarah Little (1986):

$$R^2_{BL} = \frac{1}{I}\left(\sum_{i=1}^{I} y_i \hat{P}\left(y_i = 1 \mid \boldsymbol{x}_i\right) + \sum_{i=1}^{I}\left(1-y_i\right)\left(1-\hat{P}\left(y_i = 1 \mid \boldsymbol{x}_i\right)\right)\right). \quad (1.36)$$

Wielkość (1.36) należy interpretować jako średnie prawdopodobieństwo poprawnej predykcji. Wynika to z faktu, że dla przypadków, gdy zmienna zależna przyjmuje wartość 1, sumuje się prawdopodobieństwa przyjmowania przez nią tej wartości, natomiast dla zerowych wartości zmiennej zależnej sumuje się prawdopodobieństwa przeciwne. Istotną wadą miernika (1.36) jest to, że w przypadku prób silnie niezbilansowanych jego wartości są zawsze duże. W związku z tym mankamentem Jan Salomon Cramer (1999) zaproponował alternatywną miarę następującej postaci:

$$\lambda C = \frac{\sum_{y_i=1} P\left(y_i = 1 \mid \boldsymbol{x}_i\right)}{\#\left\{y_i = 1\right\}} + \frac{\sum_{y_i=1} P\left(y_i = 0 \mid \boldsymbol{x}_i\right)}{\#\left\{y_i = 0\right\}}, \quad (1.37)$$

gdzie na przykład $\#\{y_i = 1\}$ oznacza liczbę obserwacji, dla których $y_i = 1$. Alternatywnymi miernikami jakości, wykorzystywanymi w modelach dwumianowych, są:

1) współczynnik determinacji *R*-kwadrat Bradleya Efrona (1978):

$$R_{Ef}^2 = 1 - \frac{\sum_{i=1}^{I} \left(y_i - P\left(y_i = 1 | \boldsymbol{x}_i \right) \right)^2}{\sum_{i=1}^{I} \left(y_i - \overline{y} \right)^2}, \tag{1.38}$$

2) współczynnik determinacji *R*-kwadrat Michaela Vealla i Klausa Zimmermanna (1992):

$$R_{VZ}^2 = \frac{\delta - 1}{\delta - LRI}, \tag{1.39}$$

gdzie: $\delta = \frac{I}{2 \ln L_0}$ oraz $LRI = 1 - \frac{\ln L_U}{\ln L_0}$,

3) a także współczynnik determinacji *R*-kwadrat Richarda McKelveya i Williama Zavoiny (1975):

$$R_{MZ}^2 = \frac{\sum_{i=1}^{I} \left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}} - \frac{\sum_{i=1}^{I} \boldsymbol{x}_i \hat{\boldsymbol{\beta}}}{I} \right)^2}{I + \sum_{i=1}^{I} \left(\boldsymbol{x}_i \hat{\boldsymbol{\beta}} - \frac{\sum_{i=1}^{I} \boldsymbol{x}_i \hat{\boldsymbol{\beta}}}{I} \right)^2} \tag{1.40}$$[1]

Dla modeli dwumianowych jakość dopasowania mierzona jest za pomocą porównania teoretycznych wartości zmiennej zależnej z wartościami empirycznymi. Funkcja wykorzystywana do obliczania teoretycznych wartości zmiennej zależnej przyjmuje następującą postać:

$$\hat{y}_i = I\left\{ \hat{P}\left(y_i = 1 | \boldsymbol{x}_i \right) > ccc \right\}, \tag{1.41}$$

gdzie $\hat{P}$ oznacza empiryczne prawdopodobieństwo, że zmienna zależna przyjmuje wartość 1. Wówczas po obliczeniu wartości teoretycznych dla zmiennej zależnej definiowana jest następująca tablica trafności predykcji:

1 Formuła ta jest odpowiednia dla probitu. W przypadku logitu zamiast wartości 1 w mianowniku mamy $\pi^2/3$.

Tabela 3. Tablica trafności predykcji

	$y = 0$	$y = 1$
$\hat{y} = 0$	I_{00}	I_{01}
$\hat{y} = 1$	I_{10}	I_{11}

Źródło: opracowanie własne.

I_{00} jest liczbą obserwacji, dla których zarówno zmienna obserwowalna, jak i teoretyczna przyjmują wartość 0. I_{01} jest liczbą obserwacji, dla których wartość zmiennej obserwowalnej wynosi 1, natomiast zmienna teoretyczna przyjmuje wartość 0. I_{10} jest liczbą obserwacji, dla których wartość zmiennej obserwowalnej wynosi 0, natomiast zmienna teoretyczna przyjmuje wartość 0. W przypadku I_{11} obserwacji obie zmienne (y oraz $\hat{y}$) przyjmują wartość 1.

Mamy do czynienia z poprawną predykcją, jeśli wartość teoretyczna jest równa wartości empirycznej. Procent poprawnych predykcji obliczany jest zatem za pomocą następującej formuły:

$$CP = \frac{I_{00} + I_{11}}{I_{00} + I_{11} + I_{01} + I_{10}}. \tag{1.42}$$

Podczas generowania tablicy trafności predykcji ważnym zagadnieniem jest wybór parametru *ccc* związanego z regułą (1.41). W przypadku gdy $I\{\cdot\}$ udział „jedynek" w próbie nie jest zbyt odległy od 0,5, przyjmuje się *ccc* = 0,5. Jeśli jednak próba ma charakter niezbilansowany (udział „zer" lub „jedynek" jest bliski 100%), wówczas punkt progowy *ccc* powinien być równy udziałowi „jedynek" w próbie. Oprócz wartości zmiennej *CP* na podstawie wartości z tabeli 3 obliczane są także następujące wielkości:

1) czułość:

$$SENSITIVITY = \frac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \mathbf{x}_i\right) > ccc\right\} I\left\{y_i = 1\right\}}{\sum_{i=1}^{I} I\left\{y_i = 1\right\}}, \tag{1.43}$$

2) specyficzność:

$$SPECIFICITY = \frac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \mathbf{x}_i\right) \leq ccc\right\} I\left\{y_i = 0\right\}}{\sum_{i=1}^{I} I\left\{y_i = 0\right\}}, \tag{1.44}$$

3) pozytywna wartość predyktywna:

$$PPV = \frac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) > ccc\right\} I\left\{y_i = 1\right\}}{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) > ccc\right\}}, \tag{1.45}$$

4) negatywna wartość predyktywna:

$$NPV = \frac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) \leq ccc\right\} I\left\{y_i = 0\right\}}{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) \leq ccc\right\}}. \tag{1.46}$$

Jak widać, analizowane powyżej mierniki zależą od poziomu *ccc*. Wraz ze wzrostem parametru *ccc* wzrasta wartość miernika (1.44) i obniża się wartość miernika (1.43). Udział „jedynek" w próbie jest tak zwaną optymalną wartością graniczną Cramera (por. Gruszczyński, 2012). Możliwe jest ustalenie *ccc* na poziomie minimalizującym prawdopodobieństwo popełnienia błędu predykcji. Optymalna wartość parametru wyznaczana jest zatem na podstawie wzoru:

$$ccc_{OPT} = arg\min \left(\begin{array}{l} \dfrac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) \leq ccc\right\} I\left\{y_i = 0\right\}}{\sum_{i=1}^{I} I\left\{y_i = 0\right\}} + \\ + \dfrac{\sum_{i=1}^{I} I\left\{P\left(y_i = 1 \mid \boldsymbol{x}_i\right) > ccc\right\} I\left\{y_i = 1\right\}}{\sum_{i=1}^{I} I\left\{y_i = 1\right\}} \end{array} \right). \tag{1.47}$$

Krzywa ROC (*Receiver Operating Characteristic*) wykorzystywana jest w celu analizy zależności między czułością a specyficznością dla różnych wartości współczynnika *ccc*. Ilustruje ona wszystkie kombinacje obu rodzajów błędów dla różnych wartości progowych. Pokazuje zależność między *SENSITIVITY* oraz *1-SPECIFICITY*. Pole pod krzywą ROC wykorzystywane jest także do oceny jakości dopasowania modelu do danych. Przyjmuje ono wartości z przedziału <0,5; 1>. Jeśli powierzchnia pod krzywą ROC jest minimalna, wówczas model nie ma żadnej mocy predykcyjnej. Drugi krańcowy przypadek (pole = 1) oznacza, że model idealnie prognozuje wartości empiryczne.

Porównanie średniego prawdopodobieństwa, że zmienna zależna przyjmuje wartość 1 z wartością średnią zmiennej zależnej nie dostarcza odpowiednich informacji na temat jakości dopasowania modelu do danych. W przypadku modelu logitowego zawierającego wyraz wolny te dwie wielkości są sobie równe, co jest wymuszone przez warunki pierwszego rzędu. Zaproponowany przez Davida W. Hosmera Jr., Stanleya Lemeshowa i Susanne May (1999) test specyfikacji służy do oceny jakości dopasowania przez porównanie udziału „jedynek" w próbie z prawdopodobieństwem, że zmienna zależna przyjmuje wartość 1 w podgrupach. Sposób przeprowadzenia tego testu jest następujący:

1) po oszacowaniu parametrów modelu dwumianowego obliczane są prawdopodobieństwa, że zmienna zależna przyjmuje wartość 1 dla wszystkich obserwacji: $\hat{P}\left(y_i = 1 | \boldsymbol{x}_i\right)$;
2) następnie prawdopodobieństwa te są szeregowane od najmniejszego do największego, a zbiór obserwacji dzieli się na G grup ze względu na wartości przyjmowane przez $\hat{P}\left(y_i = 1 | \boldsymbol{x}_i\right)$;
3) dla każdej grupy $g = 1, \ldots, G$ obliczane jest średnie prawdopodobieństwo, że zmienna zależna przyjmuje wartość 1:

$$\overline{\hat{p}}_g = \frac{\sum_{i \in g} P\left(y_i = 1 | \boldsymbol{x}_i\right)}{I_g},$$

gdzie I_g oznacza liczbę obserwacji w grupie g;

4) dla każdej grupy $g = 1, \ldots, G$ obliczana jest także średnia wartość dla zmiennej zależnej, czyli $\overline{y}_g$;
5) następnie obliczana jest wartość następującej statystyki:

$$H - L = \sum_{g=1}^{G} \frac{\left(\hat{p}_g - \overline{y}_g\right)^2}{\overline{y}_g\left(1 - \overline{y}_g\right)}; \tag{1.48}$$

6) uzyskana wartość statystyki porównywana jest z wartością krytyczną dla rozkładu chi-kwadrat o $G - 2$ stopniach swobody;
7) wyższa wartość statystyki (1.48) implikuje, że należy odrzucić hipotezę zerową mówiącą o braku istotnych różnic między częstością teoretyczną a częstością empiryczną; brak podstaw do odrzucenia hipotezy zerowej wskazuje na dobre dopasowanie danych empirycznych do teoretycznych.

1.4. Model wielomianowy (polichotomiczny) kategorii uporządkowanych

W przypadku kiedy zmienna zależna przyjmuje co najmniej trzy wartości, które daje się uporządkować, mamy do czynienia z modelem wielomianowym (polichotomicznym) kategorii uporządkowanych. Wyjaśnianie zależności między tą zmienną a innymi kategoriami ekonomiczno-społecznymi odbywa się dzięki estymacji parametrów następującego modelu:

$$y_i^* = \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \,,\ \varepsilon_i \sim F \,, \tag{1.49a}$$

$$y_i = p^* I\left\{\tau_p < y_i^* \le \tau_{p+1}\right\}, \tag{1.49b}$$

dla $p = 0, \ldots, P$. Przyjmuje się, że $\tau_0 = -\infty$ oraz $\tau_{P+1} = +\infty$. Podobnie jak w przypadku modelu dwumianowego i indeksuje obserwacje, natomiast $I\{\cdot\}$ oznacza zmienną wskaźnikową przyjmującą wartość 1, gdy warunek zdefiniowany w klamrowym nawiasie jest spełniony oraz 0 w przeciwnym przypadku. Zmienne $x_{1i}, \ldots, x_{Ki}$ służą wyjaśnieniu kształtowania się zmiennej zależnej, natomiast F jest dystrybuantą rozkładu składnika losowego. Prawdopodobieństwo, że zmienna obserwowalna y_i przyjmuje wartość skrajną 0 wynosi:

$$P\left(y_i = 0\right) = P\left(y_i^* \le \tau_1\right) = P\left(\beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \le \tau_1\right) =$$

$$P\left(\varepsilon_i \le \mu_1 - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right) = F\left(\mu_1 - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right).$$

Prawdopodobieństwo, że zmienna obserwowalna przyjmuje wartość najwyższą (P) wynosi:

$$P\left(y_i = P\right) = P\left(y_i^* > \tau_P\right) = P\left(\beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i > \tau_P\right) =$$

$$1 - F\left(\tau_P - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right).$$

Prawdopodobieństwo przyjmowania przez obserwowalną zmienną zależną jednej ze środkowych wartości $p = 1, \ldots, P - 1$ wynosi:

$$\begin{aligned} P\left(y_i = p\right) &= P\left(\tau_p < y_i^* \leq \tau_{p+1}\right) = \\ &= P\left(\tau_p < \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \leq \tau_{p+1}\right) = \\ &= P\left(\tau_p - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki} < \varepsilon_i \leq \tau_{p+1} - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right) = \\ &= F\left(\tau_{p+1} - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right) - F\left(\tau_p - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki}\right). \end{aligned}$$

W przypadku modelu polichotomicznego kategorii uporządkowanych najczęściej przyjmuje się, że składnik losowy pochodzi z rozkładu normalnego lub logistycznego. Mamy wówczas do czynienia odpowiednio z uporządkowanym modelem probitowym lub uporządkowanym modelem logitowym. Parametry modelu wielomianowego kategorii uporządkowanych najczęściej szacuje się metodą największej wiarygodności.

1.5. Model wielomianowy kategorii nieuporządkowanych

Załóżmy, że jednostka (konsument, gospodarstwo domowe, respondent w badaniu ankietowym) wybiera jeden spośród L wariantów. Dodatkowo zakłada się, że warianty te wyczerpują zbiór potencjalnych możliwości wyboru oraz wykluczają się. Niech UZ_i^l będzie użytecznością, jaką posiada i-ta jednostka z wyboru l-tego wariantu. Użyteczność z l-tego wyboru jest sumą części deterministycznej oraz losowej:

$$UZ_i^l = VZ_i^l + \varepsilon_i^l, \; l = 1, \ldots, L, \quad (1.50)$$

gdzie część deterministyczna zależy od charakterystyk jednostki oraz parametrów będących przedmiotem estymacji:

$$VZ_i^l = \boldsymbol{x}_i^l \boldsymbol{\beta}^l. \quad (1.51)$$

Ponieważ dokonywany przez i-tą jednostkę (firmę, respondenta) wybór wariantu obserwowany jest przez badacza, uzyskuje się informację, która użyteczność jest najwyższa. Jeśli na przykład i-ta jednostka wybrała l-ty wariant, wówczas wiemy, że dla każdego:

$$l \neq l' \quad UZ_i^l \geq UZ_i^{l'}. \tag{1.52}$$

Dlatego też prawdopodobieństwo wyboru l-tego wariantu przez i-tą jednostkę można zapisać następująco:

$$P(y_i = l|\boldsymbol{x}_{il}) = P\left(\prod_{l \neq l'}\left(\varepsilon_i^l - \varepsilon_i^{l'}\right) \leq \left(\boldsymbol{x}_i^l\boldsymbol{\beta}^l - \boldsymbol{x}_i^{l'}\boldsymbol{\beta}^{l'}\right)\right). \tag{1.53}$$

Aby obliczyć prawdopodobieństwo (1.53), należy przyjąć założenia dotyczące rozkładu składnika losowego. Zakłada się, że dla każdej jednostki wektor losowy $\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_i^1 & \varepsilon_i^2 & \cdots & \varepsilon_i^L \end{bmatrix}$ ma rozkład o funkcji gęstości łącznego rozkładu f_ε. Jeśli łączny rozkład jest rozkładem Gumbela, wówczas prawdopodobieństwo (1.53) można zapisać następująco (por. m.in. Maddala, 1987; Gruszczyński, 2012):

$$P\left(y_i = l \mid \boldsymbol{x}_{il}\right) = p_i^l = \frac{\exp\left(\boldsymbol{x}_i^l\boldsymbol{\beta}^l\right)}{\sum_{r=1}^{L}\exp\left(\boldsymbol{x}_i^r\boldsymbol{\beta}^r\right)},\ l = 1, \ldots, L. \tag{1.54}$$

Wartości zmiennych objaśniających są jednak niezależne od kategorii wybieranej przez i-tą jednostkę, więc wzór (1.54) można inaczej zapisać:

$$P\left(y_i = l \mid \boldsymbol{x}_{il}\right) = p_i^l = \frac{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}^l\right)}{\sum_{r=1}^{L}\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}^r\right)},\ l = 1, \ldots, L. \tag{1.55}$$

Ponieważ nie wszystkie parametry są identyfikowalne, konieczna jest ich normalizacja. Najczęściej przyjmuje się założenie, że jeden z wektorów $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^L$ jest zerowy. Wówczas kategoria odpowiadająca wektorowi zerowemu nazywana jest bazową lub referencyjną. Interpretując oszacowania parametrów, kategoria bazowa wykorzystywana jest jako punkt odniesienia dla pozostałych. Jeśli dokonana zostanie następująca normalizacja $\boldsymbol{\beta}^1 = 0$, wówczas prawdopodobieństwa kolejnych wyborów przyjmują postać:

$$p_i^1 = \frac{1}{1 + \sum_{r=2}^{L} \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^r\right)}, \tag{1.56a}$$

$$p_i^l = \frac{\exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^l\right)}{1 + \sum_{r=2}^{L} \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^r\right)}, \quad l = 2, \ldots, L. \tag{1.56b}$$

Korzystając z faktu, że jednostki dokonujące swoich wyborów są niezależne, natomiast wybierane alternatywy wzajemnie się wykluczają, funkcja wiarygodności dla wielomianowego modelu logitowego przyjmuje postać:

$$L(\boldsymbol{\beta}^2, \ldots, \boldsymbol{\beta}^L | \boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{I} \prod_{l=1}^{L} \left(p_i^l\right)^{d_i^l}, \tag{1.57}$$

gdzie prawdopodobieństwa p_i^l zdefiniowane są we wzorach (1.56a)–(1.56b), natomiast d_i^l jest zmienną indykatorową przyjmującą wartość 1, gdy *i*-ta jednostka wybiera *l*-tą kategorię oraz 0 w przeciwnym przypadku. W związku z tym logarytm funkcji wiarygodności jest następujący:

$$\ln L(\boldsymbol{\beta}^2, \ldots, \boldsymbol{\beta}^L | \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{I} \sum_{l=1}^{L} d_i^l \ln\left(p_i^l\right). \tag{1.58}$$

W celu maksymalizacji funkcji (1.58) wyznaczamy i przyrównujemy do zera wektor pierwszych pochodnych funkcji względem wszystkich parametrów. Ponieważ macierz drugich pochodnych jest ujemnie określona (por. m.in. Gruszczyński, 2012), istnieje tylko jedno maksimum globalne dla funkcji (1.58). W celu wyznaczenia oszacowań największej wiarygodności wykorzystywane są klasyczne algorytmy numeryczne, na przykład Newtona-Raphsona lub BHHH (por. Greene, 2008).

Podobnie jak w przypadku modelu dwumianowego pseudo-R^2 McFaddena wykorzystywany jest do oceny dopasowania modelu do danych empirycznych. Wzór definiujący ten współczynnik determinacji jest następujący:

$$R^2 - McFadden = 1 - \frac{\ln L_{FULL}}{\ln L_0}, \tag{1.59}$$

gdzie $\ln L_{FULL}$ oraz $\ln L_0$ oznaczają wartość logarytmu funkcji wiarygodności odpowiednio dla modelu zawierającego wszystkie zmienne objaśniające oraz modelu zawierającego tylko wyraz wolny. Wartość funkcji wiarygodności dla modelu

pełnego oraz zawierającego tylko wyraz wolny jest wykorzystywana przy obliczaniu innego miernika dopasowania, zwanego R^2-Nagelkerke'a (Nagelkerke, 1991):

$$R^2 - Nagelkerke = \frac{1-\exp\left(\left(-\frac{2}{I}\right)\left(\ln L_{FULL} - \ln L_0\right)\right)}{1-\exp\left(\left(-\frac{2}{I}\right)\ln L_0\right)} \tag{1.60}$$

Podobnie jak w przypadku opisywanego w podrozdziale 1.2 modelu dwumianowego oszacowania parametrów w wielomianowym modelu logitowym nie są bezpośrednio interpretowane. Niemniej jednak w przypadku wielomianowego modelu logitowego możliwe jest obliczenie analogicznych ilorazów szans jak w modelu logitowym. Załóżmy (tak jak w podrozdziale 1.2), że pierwsza kategoria jest bazową. Wówczas dla każdej innej kategorii iloraz $\frac{p_i^l}{p_i^1}$ wyraża się wzorem:

$$\frac{p_i^l}{p_i^1} = \exp\left(\boldsymbol{x}_i\boldsymbol{\beta}^l\right), \quad l = 2, \ldots, L. \tag{1.61}$$

Rozważmy przypadek, w którym wybrany k-ty element wektora $\boldsymbol{x}_i$ przyjmuje najpierw wartość u, a następnie $u + 1$, podczas gdy pozostałe $K - 1$ elementów są takie same w obydwu przypadkach. Wówczas prawdziwa jest równość:

$$\frac{\left(\frac{p_i^l}{p_i^1} \mid x_{ik} = u+1\right)}{\left(\frac{p_i^l}{p_i^1} \mid x_{ik} = u\right)} = \exp\left(\beta^{l,k}\right), \tag{1.62}$$

Oznacza to zatem, że procentowa zmiana ilorazu szans pod wpływem jednostkowej zmiany wartości zmiennej x_{ik} wynosi:

$$\left(\frac{\left(\frac{p_i^l}{p_i^1} \mid x_{ik} = u+1\right) - \left(\frac{p_i^l}{p_i^1} \mid x_{ik} = u\right)}{\left(\frac{p_i^l}{p_i^1} \mid x_{ik} = u\right)}\right) * 100\% = \left(\exp\left(\beta^{l,k}\right) - 1\right) * 100\%. \tag{1.63}$$

Dlatego też po oszacowaniu parametrów modelu wielomianowego kategorii nieuporządkowanych przed interpretacją uzyskanych rezultatów należy wykorzystać przekształcenie (1.62).

Tak samo jak w przypadku modelu dwumianowego również dla modelu wielomianowego kategorii nieuporządkowanych obliczane są efekty krańcowe. Rozważmy sytuację, w której następuje zmiana wartości *k*-tej zmiennej o x_{ik}, a pozostałe zmienne nie ulegają zmianie, czyli: $\left[0 \ldots 0\, x_{ik}\, 0 \ldots 0\right]$. Wówczas zmiana prawdopodobieństwa dokonania *l*-tego wyboru przez *i*-tą jednostkę wynosi:

$$\begin{aligned} \Delta p_i^l &= P\left(y_i = l \mid \boldsymbol{x}_i + \Delta \boldsymbol{x}_i\right) - P\left(y_i = l \mid \boldsymbol{x}_i\right) = \\ &= \frac{\exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^l + \Delta x_{ik} \beta^{l,k}\right)}{1 + \sum_{r=2}^{L} \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^r + \Delta x_{ik} \beta^{l,k}\right)} - \frac{\exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^l\right)}{1 + \sum_{r=2}^{L} \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}^r\right)}. \end{aligned} \tag{1.64}$$

Wielkość (1.64) należy interpretować jako wpływ zmiany wartości *k*-tej zmiennej o x_{ik} na prawdopodobieństwo, że *i*-ta jednostka wybiera *l*-tą kategorię. Jeśli *k*-ta zmienna ma charakter dyskretny lub jest nawet zmienną zero-jedynkową, interpretacja wielkości (1.64) jest uzasadniona. W przypadku ciągłej zmiennej objaśniającej uzasadnione jest obliczanie wpływu marginalnej zmiany wartości zmiennej objaśniającej na prawdopodobieństwo, że *i*-ta jednostka wybiera *j*-tą kategorię:

$$MPE_{ik}^l = \frac{\partial p_{il}}{\partial x_{ik}} = p_{il}\left(\beta^{l,k} - \sum_{r=1}^{L} p_{ir} \beta^{r,k}\right). \tag{1.65}$$

Oszacowania efektów krańcowych dla średnich wartości zmiennych objaśniających przyjmują następującą postać:

$$\hat{MPE}_k^l = \tilde{p}^l\left(\hat{\beta}^{l,k} - \sum_{r=1}^{L} \tilde{p}^r \hat{\beta}^{r,k}\right), \tag{1.66}$$

gdzie:

$$\tilde{p}_l = \frac{\exp\left(\overline{\boldsymbol{x}} \hat{\boldsymbol{\beta}}^l\right)}{1 + \sum_{r=2}^{L} \exp\left(\overline{\boldsymbol{x}} \hat{\boldsymbol{\beta}}^r\right)}.$$

Oprócz tego możliwe jest obliczenie oczekiwanych przeciętnych efektów krańcowych:

$$\hat{AMPE}_k^l = \frac{1}{I} \sum_{i=1}^{I} MPE_{ik}^l,\ l = 1, \ldots, L. \tag{1.67}$$

1.6. Model regresji rankingowej

Czasami w badaniach społecznych mamy do czynienia z rankingami. Respondent stojący przed wyborem jednego z kilku wariantów ma za zadanie nie tylko wskazanie najlepszego, lecz także uszeregowanie ich od najbardziej do najmniej pożądanego. Model ekonometryczny próbujący wyjaśnić, w jaki sposób na przykład cechy socjoekonomiczne wpływają na uszeregowanie, nazywany jest modelem regresji rankingowej. Załóżmy, że respondent ma za zadanie uszeregowanie R elementów. Niech b_i^r oznacza wariant, któremu został przyporządkowany r-ty ranking przez i-tego respondenta. Ranking dla i-tego respondenta definiowany jest $\boldsymbol{RA}_i \equiv \left(b_i^1, b_i^2, \ldots, b_i^R\right)$. Model mający na celu identyfikację czynników wpływających na sposób uszeregowania wariantów nazywany jest modelem regresji rankingowej lub modelem logistycznym dla rankingów (por. Plackett, 1975; Luce, 2005). Jeśli $\tilde{R}$ jest zbiorem możliwych wariantów, wówczas prawdopodobieństwo określonego rankingu dla i-tego obiektu (np. respondenta) dane jest wzorem:

$$P\left(\boldsymbol{RA}_i \mid \tilde{R}\right) = \prod_{r=1}^{R-1} \frac{\exp\left(v_i^{b_i^r}\right)}{\sum_{s=r}^{R} \exp\left(v_i^{b_i^s}\right)}, \tag{1.68}$$

gdzie $v_i^{b_i^r}$ jest funkcją liniową zmiennych objaśniających wpływających na sposób uszeregowania wariantów. Na przykład $v_i^{b_i^r} = \boldsymbol{x}_i^r \boldsymbol{\beta}^r$.

Analizowany model często nazywany jest eksplodującym modelem logitowym (por. np. Chapman, Staelin, 1982). Wynika to z faktu, że prawdopodobieństwo określonego rankingu jest iloczynem prawdopodobieństw wyboru jednego z sukcesywnie zmniejszającej się liczby wariantów.

Funkcja wiarygodności dla modelu regresji rankingowej przyjmuje postać (por. m.in. Allison, Christakis, 1994):

$$\ln L = \sum_{i=1}^{I}\sum_{r=1}^{R} \boldsymbol{x}_i^r \boldsymbol{\beta}^r - \sum_{i=1}^{I}\sum_{r=1}^{L} \ln\left[\sum_{r'=1}^{L} \delta_i^{rr'} \exp\left(\boldsymbol{x}_i^r \boldsymbol{\beta}^r\right)\right], \tag{1.69}$$

gdzie $\delta_i^{rr'}$ jest zmienną wskaźnikową przyjmującą wartość 1, jeśli r'-ty wariant znajduje się wyżej w rankingu stworzonym przez i-tego respondenta w porównaniu z r-tym. Funkcja wiarygodności (1.69) może być maksymalizowana za pomocą numerycznych algorytmów optymalizacyjnych (por. m.in. Maddala, 2013). Jest ona globalnie wklęsła, co oznacza, że uzyskiwane maksimum jest globalne (por. Beggs, Cardell, Hausman, 1981). Jak pokazują Robert Keener i Donald Waldman (1985), estymator uzyskany metodą największej wiarygodności jest zgodny i ma rozkład asymptotycznie normalny.

1.7. Problem selekcji próby w modelach ekonometrycznych. Model Heckmana

W badaniach mikroekonomicznych i socjologicznych często pojawia się problem selekcji próby. Załóżmy, że dysponujemy danymi pochodzącymi z badania aktywności ekonomicznej ludności. Respondenci na początku są pytani o aktywność zawodową. Dopiero osoby aktywne zawodowo udzielają odpowiedzi na pytanie dotyczące wysokości wynagrodzenia. Spłacalność kredytu obserwowana jest tylko w grupie osób, którym bank go udzielił. Decyzja o wyborze partii politycznej podejmowana jest tylko przez osoby biorące udział w głosowaniu.

Nieuwzględnienie problemu selekcji próby ma poważne konsekwencje dla estymacji parametrów. Dobry przykład dotyczący negatywnych konsekwencji nieuwzględnienia faktu selekcji próby można znaleźć między innymi w pracy Marcina Owczarczuka (2010). Przypuśćmy, że w populacji wszystkich osób ubiegających się o kredyt reprezentujący grupę zawodową X są złymi kredytobiorcami, co oznacza, że często nie spłacają kredytów. Bank nie stosuje żadnych ograniczeń i każdy ubiegający się o kredyt otrzymuje go. Wówczas estymacja parametrów modelu scoringowego bazującego na pełnej próbie dostarczy zapewne następującego wniosku dla pracowników banku: „udzielenie kredytu osobie należącej do grupy zawodowej X wiąże się z niższym prawdopodobieństwem spłacenia w porównaniu z osobami z innej grupy zawodowej". Po skorzystaniu z takich rekomendacji w dalszej kolejności bank decyduje się selekcjonować osoby, którym udzielany jest kredyt. Dlatego też niewielki odsetek osób, którym przyznany został kredyt, należy do grupy zawodowej X. Załóżmy, że następnie analizowana jest spłacalność kredytu w grupie osób, którym został on przyznany. Osoby, które należą do grupy zawodowej X i jednocześnie otrzymały kredyt, muszą charakteryzować się odpowiednimi wartościami innych zmiennych. Wynika to z faktu, że skoro przynależność do grupy zawodowej negatywnie wpływa na fakt przyznania im kredytu, to wartości innych zmiennych musiały oddziaływać w odwrotną stronę, skoro osoby te znalazły się w grupie kredytobiorców. Odpowiednie wartości innych zmiennych (np. dochód w gospodarstwie domowym, liczba osób na utrzymaniu) sprawiają, że osoby te są solidnymi kredytobiorcami. Gdyby parametry modelu scoringowego były estymowane tylko na grupie tych, którym kredytu udzielono, doszlibyśmy do błędnych wniosków, że przynależność do grupy zawodowej X ma pozytywny wpływ na prawdopodobieństwo spłacenia kredytu. Nieuwzględnienie problemu selekcji próby prowadzi do uzyskania nieprawidłowych wyników.

W celu analitycznego wyprowadzenia wielkości obciążenia w modelu selekcji próby należy zapisać odpowiednie równania:

$$y^{*}_{(2)i} = \boldsymbol{w}_i\boldsymbol{\gamma} + \varepsilon_{(2)i}, \tag{1.70a}$$

$$y_{(2)i} = I\left\{y^{*}_{(2)i} > 0\right\}, \tag{1.70b}$$

$$y_{(1)i} = \boldsymbol{x}_i\boldsymbol{\beta} + \tilde{\sigma}\varepsilon_{(1)i}, \text{ dla } y_{(2)i} = 1, \tag{1.70c}$$

$$\begin{bmatrix}\varepsilon_{(1)i}\\ \varepsilon_{(2)i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right). \tag{1.70d}$$

W modelu (1.70a)–(1.70d) pierwsze dwa równania dotyczą selekcji próby. $\boldsymbol{w}_i$ jest wektorem obserwacji na zmiennych wpływających na fakt selekcji, natomiast $\boldsymbol{\gamma}$ zawiera parametry przy tych zmiennych. Równanie (1.70c) jest wynikowe. Zmienna $y_{(1)i}$, której wartości obserwowane są tylko wówczas, gdy $y^{*}_{(2)i} > 0$, może mieć charakter zarówno ciągły, jak i dyskretny. Równanie (1.70d) składa się z założeń dotyczących łącznego rozkładu składników losowych $\varepsilon_{(1)i}$ oraz $\varepsilon_{(2)i}$. Zgodnie ze wzorem (1.70d) warunkowa wartość oczekiwana $E(\varepsilon_{(1)i}|\varepsilon_{(2)i})$ wynosi:

$$E\left(\varepsilon_{(1)i} \mid \varepsilon_{(2)i}\right) = \rho\varepsilon_{(2)i}. \tag{1.71}$$

Dlatego też warunkową wartość oczekiwaną zmiennej losowej $\varepsilon_{(1)i}$ w rozkładzie uciętym można zapisać następująco:

$$E\left(\varepsilon_{(1)i} \mid \varepsilon_{(2)i} > C\right) = E\left(\rho\varepsilon_{(2)i} \mid \varepsilon_{(2)i} > C\right) = \rho E\left(\varepsilon_{(2)i} \mid \varepsilon_{(2)i} > C\right) = \rho\frac{\phi(C)}{1-\Phi(C)}. \tag{1.72}$$

W modelu regresji mamy zatem:

$$E\left(y_{(1)i} \mid y^{*}_{(2)i} > 0\right) = \boldsymbol{x}_i\boldsymbol{\beta} + E\left[\varepsilon_{(1)i} \mid \varepsilon_{(2)i} > -\boldsymbol{w}_i\boldsymbol{\gamma}\right] = \boldsymbol{x}_i\boldsymbol{\beta} + \rho\frac{\phi\left(\boldsymbol{w}_i\boldsymbol{\gamma}\right)}{1-\Phi\left(\boldsymbol{w}_i\boldsymbol{\gamma}\right)}. \tag{1.73}$$

W związku z tym podczas estymacji parametrów modelu regresji danego równaniem (1.70c), bez uwzględnienia faktu selekcji próby, nie uwzględnia się ważnego predyktora $\rho\frac{\phi\left(\boldsymbol{w}_i\boldsymbol{\gamma}\right)}{1-\Phi\left(\boldsymbol{w}_i\boldsymbol{\gamma}\right)}$. Wraz ze wzrostem wartości współczynnika korelacji między składnikami losowymi ρ następuje wzrost wpływu selekcji. Z równania (1.73)

wynika, że w przypadku braku korelacji między składnikami losowymi z dwóch równań, selekcja nie ma wpływu na wartość oczekiwaną zmiennej wynikowej.

Jedną z metod estymacji parametrów modelu selekcji próby Heckmana jest zastosowanie zgodnego estymatora dwustopniowego. W pierwszym kroku szacowane są parametry modelu probitowego, danego równaniami (1.70a)–(1.70b). Następnie obliczany jest odwrócony iloraz Millsa $\frac{\phi\left(\boldsymbol{w}_i\hat{\boldsymbol{\gamma}}\right)}{1-\Phi\left(\boldsymbol{w}_i\hat{\boldsymbol{\gamma}}\right)}$. Jest on dalej wykorzystywany jako zmienna objaśniająca w równaniu (1.73). Parametry $\boldsymbol{\beta}$ oraz ρ w równaniu (1.73) szacowane są w drugim kroku.

1.8. Model licznikowy

W niektórych przypadkach zmienna, której kształtowanie się chcemy wyjaśnić, przyjmuje tylko nieujemne wartości całkowite. Mamy wówczas do czynienia ze zmienną licznikową, a model regresji wyjaśniający jej kształtowanie się nazywamy licznikowym. Zastosowanie klasycznych metod estymacji w przypadku licznikowej zmiennej zależnej nie jest właściwym rozwiązaniem. Wynika to z faktu, że rozkład zmiennej zależnej nie jest ciągły. Przyjmuje ona wartości dyskretne z określonymi prawdopodobieństwami.

Najpopularniejszym modelem licznikowym jest regresja Poissona. Punktem wyjścia do wyprowadzenia tego modelu jest założenie, że zmienna zależna przyjmuje określone wartości zgodnie z rozkładem dyskretnym Poissona:

$$P\left(y_i = a\right) = \frac{\exp\left(-\lambda\right)\lambda^a}{a!},\ a = 0, 1, 2, \ldots \tag{1.74}$$

λ jest parametrem intensywności, natomiast wartość oczekiwana i wariancja dla rozkładu Poissona wynoszą odpowiednio:

$$E\left(y_i\right) = Var\left(y_i\right) = \lambda\,. \tag{1.75}$$

Cechą charakterystyczną rozkładu Poissona jest równość wartości oczekiwanej i wariancji. W modelu regresji Poissona przyjmuje się założenie, że wartość parametru intensywności jest różna dla poszczególnych jednostek i wynosi λ_i. Przyjmuje się również, że zmienna zależna y_i, której kształtowanie się zależy od wektora obserwacji na zmiennych objaśniających $\boldsymbol{x}_i$, ma warunkowy rozkład Poissona:

$$E\left(y_i \mid \boldsymbol{x}_i\right) = Var\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i . \tag{1.76}$$

Wartość parametru intensywności zależy od zmiennych objaśniających w następujący sposób:

$$\lambda_i = \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}\right). \tag{1.77}$$

Dlatego też podstawą tak zwanej regresji poissonowskiej jest formuła:

$$\ln E\left(y_i \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i \boldsymbol{\beta} . \tag{1.78}$$

Parametry modelu Poissona szacuje się metodą największej wiarygodności. Jak wskazuje między innymi Marek Gruszczyński (2012), przybliżanie regresji Poissona standardową regresją liniową jest uzasadnione, gdy wartości zmiennej licznikowej przekraczają 15–20. Wówczas rozkład Poissona zbiega do normalnego.

W modelu Poissona krańcowy efekt zmiennej x_k na warunkową wartość oczekiwaną zmiennej y wynosi:

$$\frac{\partial E\left(y_i \mid \boldsymbol{x}_i\right)}{\partial x_{ik}} = \exp\left(\boldsymbol{x}_i \boldsymbol{\beta}\right)\beta_k . \tag{1.79}$$

Oznacza to zatem, że – podobnie jak w przypadku modelu dwumianowego – efekt krańcowy zależy od wartości przyjmowanych przez zmienne objaśniające. Dlatego też na ogół oblicza się średni efekt krańcowy lub efekt krańcowy dla pewnych ustalonych wartości zmiennych objaśniających. Relatywna zmiana warunkowej wartości oczekiwanej zmiennej y_i, związana z bardzo małą zmianą jednej ze zmiennych objaśniających, wynosi:

$$\frac{\dfrac{\partial E\left(y_i \mid \boldsymbol{x}_i\right)}{\partial x_{ik}}}{E\left(y_i \mid \boldsymbol{x}_i\right)} = \frac{\exp\left(\boldsymbol{x}_i \boldsymbol{\beta}\right)\beta_k}{\exp\left(\boldsymbol{x}_i \boldsymbol{\beta}\right)} = \beta_k . \tag{1.80}$$

Dlatego też, jeśli na przykład oszacowanie parametru β_k wyniesie 0,1, należy interpretować je następująco: wzrost wartości zmiennej x_k o jednostkę powoduje, przy innych czynnikach niezmienionych, wzrost wartości oczekiwanej zmiennej y o 10%. Jeśli w modelu Poissona zmienna objaśniająca x_k jest zlogarytmowana, wówczas oszacowanie parametru β_k interpretowane jest następująco: wzrost wartości

zmiennej x_k o 1% powoduje, przy innych czynnikach niezmienionych, wzrost wartości oczekiwanej zmiennej zależnej o około β_k%. W przypadku gdy zmienna x_k jest dyskretna, wówczas wzrost wartości tej zmiennej o jednostkę powoduje relatywny wzrost wartości oczekiwanej dla zmiennej zależnej o:

$$\frac{E\left(y_i \mid x_{ik}+1\right)-E\left(y_i \mid x_{ik}\right)}{E\left(y_i \mid x_{ik}\right)}=\frac{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}+\beta_k\right)-\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}=\exp\left(\beta_k\right)-1. \quad (1.81)$$

Jeśli zmienna x_k jest binarna, wówczas interpretacja parametru β_k jest następująca: wartość oczekiwana zmiennej zależnej jest o $100 \cdot (\exp(\beta_k) - 1)$ procent wyższa w przypadku, gdy x_k wynosi 1, w porównaniu z sytuacją, w której przyjmuje wartość 0.

Budzącym wątpliwości założeniem modelu regresji Poissona jest równość warunkowej wartości oczekiwanej oraz warunkowej wariancji. W przypadku gdy wariancja zmiennej licznikowej przekracza wartość oczekiwaną, mamy do czynienia ze zjawiskiem „nadmiernego rozproszenia" w modelu licznikowym (por. Wu, 1999). Wówczas zamiast szacowania parametrów modelu Poissona można wykorzystać model regresji ujemnej dwumianowej. W modelu tym (jak sama nazwa wskazuje) zakłada się, że rozkład prawdopodobieństwa zmiennej zależnej jest ujemny dwumianowy:

$$P\left(y_i\right)=\frac{\tilde{\Gamma}\left(y_i+\alpha^{-1}\right)}{\tilde{\Gamma}\left(y_i+1\right)\tilde{\Gamma}\left(\alpha^{-1}\right)}\left(\frac{\lambda_i}{\lambda_i+\alpha^{-1}}\right)^{y_i}\left(\frac{\alpha^{-1}}{\lambda_i+\alpha^{-1}}\right)^{\alpha^{-1}}. \quad (1.82)$$

Wartość oczekiwana i wariancja zmiennej zależnej wynoszą odpowiednio:

$$E\left(y_i\right)=\lambda_i \quad (1.83a)$$

oraz

$$Var\left(y_i\right)=\lambda_i+\alpha\lambda_i^2. \quad (1.83b)$$

Parametr α jest dodatni. Im niższe są jego wartości, tym mniejsze są różnice między prawdopodobieństwem dla rozkładu Poissona a liczonym na podstawie rozkładu ujemnego dwumianowego. Model regresji ujemnej dwumianowej pierwszego typu zakłada:

$$\lambda_i=\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right), \quad (1.84a)$$

$$\alpha_i = \frac{\sigma^2}{\lambda_i}, \tag{1.84b}$$

natomiast analogiczny model drugiego typu różni się założeniem dotyczącym parametru α (zakładana jest jego stałość). W obu modelach mamy:

$$E\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i . \tag{1.85}$$

Wzór na wariancję w modelu ujemnym dwumianowym pierwszego typu jest następujący:

$$Var\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i + \sigma^2 \lambda_i^2, \tag{1.86a}$$

podczas gdy w przypadku alternatywnego wariantu przyjmuje on postać:

$$Var\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i \left(1 + \sigma^2\right). \tag{1.86b}$$

Interpretacja parametrów modelu regresji ujemnej dwumianowej jest podobna do tej z modelu Poissona.

Wybór między modelem licznikowym Poissona a modelem regresji ujemnej dwumianowej wymaga określenia, czy wariancja zmiennej licznikowej znacząco przekracza jej wartość oczekiwaną. Jeśli taka sytuacja ma miejsce, zastosowanie modelu regresji ujemnej dwumianowej jest bardziej uzasadnione. W celu sprawdzenia, który model jest bardziej zasadny, wykorzystywany jest standardowy test ilorazu wiarygodności (por. Bazyl, 2010). Odpowiednia statystyka testowa przyjmuje postać:

$$G^2 = 2\left(\ln L_{MRUD} - \ln L_{MP}\right), \tag{1.87}$$

gdzie $\ln L_{MRUD}$ oraz $\ln L_{MP}$ oznaczają wartość logarytmu funkcji wiarygodności odpowiednio dla modelu ujemnego dwumianowego oraz modelu Poissona. Przy prawdziwości hipotezy zerowej statystyka (1.87) ma rozkład chi-kwadrat o jednym stopniu swobody.

W przypadku niektórych zmiennych licznikowych może mieć miejsce taka sytuacja, że często przyjmują one wartość 0. Wówczas zastosowanie zarówno modelu regresji Poissona, jak i modelu regresji ujemnej dwumianowej prowadzi do niedoszacowania prawdopodobieństwa uzyskania przez zmienną zależną wartości 0. Wówczas w celu analizy wpływu zmiennych objaśniających na wartość zmiennej zależnej wykorzystywany jest model uwzględniający podwyższoną liczbę zer

zmiennej objaśnianej (Khoshgoftaar, Gao, Szabo, 2005). Model ten przyjmuje postać:

$$y_i = \begin{cases} 0 & \text{gdy} \quad ID_i = 1, \\ y_i^* & \text{gdy} \quad ID_i = 0, \end{cases} \tag{1.88}$$

gdzie y_i^* jest zmienną o rozkładzie Poissona, natomiast prawdopodobieństwo, że zmienna ID_i przyjmuje wartość 1 opisane jest wzorem:

$$P\left(ID_i = 1 \mid \boldsymbol{zz}_i\right) = \frac{\exp\left(\boldsymbol{zz}_i\boldsymbol{\gamma}\right)}{1+\exp\left(\boldsymbol{zz}_i\boldsymbol{\gamma}\right)}, \tag{1.89}$$

gdzie $\boldsymbol{zz}_i$ jest wektorem zawierającym kategorie wpływające na to, czy zmienna ID_i przyjmuje wartość 1, czy 0.

Warunkowy rozkład prawdopodobieństwa zmiennej zależnej przy założeniu, że $ID_i = 0$, dany jest następującym równaniem:

$$P\left(Y = y_i \mid \boldsymbol{x}_i, ID_i = 0\right) = \frac{\exp\left(-\lambda_i\right)\lambda_i^{y_i}}{y_i!}. \tag{1.90}$$

W związku z tym funkcja prawdopodobieństwa obserwowanej zmiennej licznikowej przyjmuje postać:

$$P\left(Y = y_i \mid \boldsymbol{x}_i\right) = P\left(ID_i = 1\right) * 0 + P\left(ID_i = 0\right) * \frac{\exp\left(-\lambda_i\right)\lambda_i^{y_i}}{y_i!}. \tag{1.91}$$

Wartość oczekiwana zmiennej licznikowej wynosi zatem:

$$E\left(y_i \mid \boldsymbol{x}_i, \boldsymbol{zz}_i\right) = \frac{\exp\left(\boldsymbol{x}_i\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{zz}_i\dot{\boldsymbol{\gamma}}\right)}. \tag{1.92}$$

Test Vuonga (por. Long, Freese, 2014) wykorzystywany jest w celu porównania modelu Poissona z podwyższoną liczbą zer z jego standardową postacią. Polega on na porównaniu prawdopodobieństw prognozowanych przez oba modele i wyliczeniu następującego logarytmu z ilorazu:

$$m_i = \ln\left[\frac{\hat{P}_1\left(y_i \mid \boldsymbol{x}_i\right)}{\hat{P}_2\left(y_i \mid \boldsymbol{x}_i\right)}\right], \tag{1.93}$$

gdzie $\hat{P}_1\left(y_i \mid \boldsymbol{x}_i\right)$ jest prognozowanym przez model (z podwyższoną liczbą zer) prawdopodobieństwem, że $Y = y_i$, natomiast $\hat{P}_2\left(y_i \mid \boldsymbol{x}_i\right)$ jest prognozowanym przez standardowy model Poissona prawdopodobieństwem, że $Y = y_i$. W dalszej kolejności obliczana jest wartość następującej statystyki:

$$V = \frac{\sqrt{I}}{s_m}\overline{m}, \tag{1.94}$$

gdzie I jest liczbą obserwacji, natomiast $\overline{m}$ oraz s_m oznaczają oceny próbkowe estymatorów dla odpowiednich parametrów. Hipoteza zerowa oraz alternatywna dla testu przedstawiają się następująco:

$$\begin{aligned} H_0 &: E\left(m_i\right) = 0, \\ H_1 &: E\left(m_i\right) \neq 0. \end{aligned} \tag{1.95}$$

Statystyka (1.94) ma rozkład asymptotycznie normalny (por. Bazyl, 2010). Brak podstaw do odrzucenia hipotezy zerowej implikuje, że obydwa modele są jednakowo dobre. Jeśli odrzucona zostanie hipoteza zerowa, należy uznać, że model z podwyższoną liczbą zer lub model Poissona jest lepszy. Ujemna wartość statystyki (1.94) przemawia za modelem Poissona, natomiast wartość dodatnia informuje o wyższości modelu z podwyższoną liczbą zer.

1.9. Dwurównaniowy model probitowy

W badaniach mikroekonomicznych i socjologicznych możemy mieć do czynienia z sytuacją, w której jednostka (np. gospodarstwo domowe, przedsiębiorstwo, osoba) dokonuje jednocześnie dwóch wyborów. Na przykład firma może wprowadzić innowację tylko procesową, tylko produktową, oba rodzaje lub żadnej z nich. W takich przypadkach mamy dwuwymiarową zależną zmienną dwumianową. Dwurównaniowy model probitowy nadaje się do wyjaśnienia wpływu zmiennych objaśniających na obie zmienne zależne. Przyjmuje on następującą postać:

$$y^{*}_{(1)i} = \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)} + \varepsilon_{(1)i}\,, \tag{1.96}$$

$$y^{*}_{(2)i} = \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} + \varepsilon_{(2)i}\,,$$

$$y_{(a)i} = I\left\{y^{*}_{(a)i} > 0\right\},\ a = 1, 2\,,$$

$$\begin{bmatrix}\varepsilon_{(1)i}\\ \varepsilon_{(2)i}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right).$$

Parametry modelu (1.96) najczęściej szacowane są metodą największej wiarygodności. Wykorzystanie tego modelu jest uzasadnione, gdy korelacja ρ między składnikami losowymi $\varepsilon_{(1)i}$ oraz $\varepsilon_{(2)i}$ jest niezerowa. Dlatego też po estymacji parametrów modelu (1.96) testowana jest hipoteza $\rho = 0$. Jeśli nie ma podstaw do jej odrzucenia, zamiast estymacji parametrów dwurównaniowego modelu probitowego należy osobno szacować parametry obu równań. W przypadku dwurównaniowego modelu probitowego prawdopodobieństwa, że dwie zmienne przyjmują wartości 1 i 0, są następujące:

$$\begin{aligned}
P\left(y_{(1)i} = 1, y_{(2)i} = 1\right) &= \Phi_2\left(\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}, \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \rho\right),\\
P\left(y_{(1)i} = 0, y_{(2)i} = 0\right) &= \Phi_2\left(-\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}, -\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \rho\right),\\
P\left(y_{(1)i} = 1, y_{(2)i} = 0\right) &= \Phi_2\left(\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}, -\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \rho\right),\\
P\left(y_{(1)i} = 0, y_{(2)i} = 1\right) &= \Phi_2\left(-\boldsymbol{x}_{(1)i}\beta_{(1)}, \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \rho\right),
\end{aligned} \tag{1.97}$$

gdzie $\Phi_2(a, b, c)$ oznacza dystrybuantę dwuwymiarowego rozkładu normalnego o zerowych wartościach oczekiwanych oraz jednostkowych wariancjach w punktach a oraz b.

1.10. Wielorównaniowy model probitowy

Badacze spotykają się również z problemami więcej niż dwóch zależnych zmiennych dwumianowych. Mamy wówczas do czynienia z wielorównaniowym modelem probitowym. Przyjmuje on następującą postać:

$$\begin{gathered} y^{*}_{(m)i} = \boldsymbol{x}_{(m)i}\boldsymbol{\beta}_{(m)} + \varepsilon_{(m)i},\ m = 1, 2, \ldots, M, \\ y_{(m)i} = I\left\{y^{*}_{(m)i} > 0\right\}, \\ \left[\varepsilon_{(1)i} \quad \cdots \quad \varepsilon_{(M)i}\right]^{T} \sim N\left(0, \tilde{\boldsymbol{\Sigma}}\right), \end{gathered} \tag{1.98}$$

gdzie $\tilde{\boldsymbol{\Sigma}}$ jest dodatnio określoną macierzą wariancji-kowariancji między składnikami losowymi. Przyjmuje się założenie, że diagonalne elementy tej macierzy wynoszą 1, a pozadiagonalne interpretowane są jako korelacje między składnikami losowymi z różnych równań. Podobnie jak w przypadku dwurównaniowego modelu probitowego różność od zera elementów macierzy wariancji-kowariancji implikuje zasadność łącznej estymacji. Na przykład jeśli mamy do czynienia z trójrównaniowym modelem probitowym, weryfikowana jest następująca hipoteza:

$$\begin{gathered} H_0 : \rho_{12} = \rho_{13} = \rho_{23} = 0, \\ H_1 : \rho_{12} \neq 0 \vee \rho_{13} \neq 0 \vee \rho_{23} \neq 0. \end{gathered} \tag{1.99}$$

W celu weryfikacji hipotezy (1.99) wykorzystywany jest test ilorazu wiarygodności. Porównywane są funkcje wiarygodności dla modelu bez restrykcji oraz modelu, w którym przyjmuje się założenie, że składniki losowe pochodzące z różnych równań są od siebie niezależne.

Parametry wielorównaniowego modelu probitowego najczęściej szacuje się za pomocą symulacyjnej metody największej wiarygodności. W celu zilustrowania tej metody rozważmy przypadek trzech równań. Przy założeniu niezależności poszczególnych obserwacji funkcja wiarygodności przyjmuje postać:

$$\ln L = \sum_{i=1}^{I} \ln\left(\Phi_3\left(\vartheta_i; \tilde{\boldsymbol{\Sigma}}\right)\right), \tag{1.100}$$

gdzie $\Phi_3\left(\vartheta_i; \tilde{\boldsymbol{\Sigma}}\right)$ jest dystrybuantą trójwymiarowego rozkładu normalnego o zerowych wartościach oczekiwanych macierzy wariancji-kowariancji $\tilde{\boldsymbol{\Sigma}}$ w punktach zdefiniowanych przez wektor ϑ_i, gdzie:

$$\vartheta_i = \left[\mathrm{K}_{(1)i} \boldsymbol{x}_{(1)i} \boldsymbol{\beta}_{(1)} \quad \mathrm{K}_{(2)i} \boldsymbol{x}_{(2)i} \boldsymbol{\beta}_{(2)} \quad \mathrm{K}_{(3)i} \boldsymbol{x}_{(3)i} \boldsymbol{\beta}_{(3)} \right] \text{ oraz } \mathrm{K}_{(m)i} = 2y_{(m)i} - 1 \text{ dla } m = 1, 2, 3.$$

Najpopularniejszą metodą symulacyjną wykorzystywaną do szacowania wartości dystrybuanty funkcji wiarygodności jest algorytm Geweke-Hajivassiliou-Keane (GHK), który został opisany między innymi w pracach Axela Börscha-Supana i Vassilisa Hajivassiliou (1993), Michaela Keane'a (1994) oraz Vassilisa Hajivassiliou i Paula Ruuda (1994). Wykorzystywany jest fakt, że dystrybuanta rozkładu *M*-wymiarowego może zostać alternatywnie zapisana jako iloczyn *M* dystrybuant warunkowych z jednowymiarowych rozkładów normalnych. Dla przykładu rozważmy przypadek, w którym wszystkie trzy zmienne dwumianowe przyjmują wartość 1. Odpowiednie prawdopodobieństwo można zdefiniować następująco:

$$\begin{aligned} P\left(y_{(1)i}=1, y_{(2)i}=1, y_{(3)i}=1\right) &= P\left(\varepsilon_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}, \varepsilon_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \varepsilon_{(3)i} \le \boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)}\right) = \\ & P\left(\varepsilon_{(3)i} \le \boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)} \mid \varepsilon_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, \varepsilon_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right) \times \\ & \times P\left(\varepsilon_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} \mid \varepsilon_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{1}\right) \times P\left(\varepsilon_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right). \end{aligned} \tag{1.101}$$

Podobne dekompozycje prawdopodobieństwa łącznego można zastosować dla każdego z pozostałych siedmiu wariantów.

W celu aproksymacji odpowiednich prawdopodobieństw wchodzących w skład wyrażenia (1.101) rozważana jest dekompozycja Choleskiego macierzy wariancji-kowariancji między składnikami losowymi:

$$E\left(\begin{bmatrix} \varepsilon_{(1)i} \\ \varepsilon_{(2)i} \\ \varepsilon_{(3)i} \end{bmatrix} \begin{bmatrix} \varepsilon_{(1)i} & \varepsilon_{(2)i} & \varepsilon_{(3)i} \end{bmatrix}\right) = \tilde{\boldsymbol{\Sigma}} = \boldsymbol{C}\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^{T}\boldsymbol{C}, \tag{1.102}$$

gdzie $\boldsymbol{C}$ jest trójkątną dolną macierzą Choleskiego odpowiadającą macierzy $\tilde{\boldsymbol{\Sigma}}$ oraz $\tilde{\boldsymbol{e}} \sim \boldsymbol{\Phi}_3\left(0, \boldsymbol{I}_3\right)$, gdzie $\boldsymbol{I}_3$ oznacza macierz jednostkową o wymiarach 3 × 3. W związku z dekompozycją (1.102) prawdziwe są następujące zależności:

$$\varepsilon_{(1)i} = C_{11}\tilde{e}_{1i},$$

$$\varepsilon_{(2)i} = C_{21}\tilde{e}_{1i} + C_{22}\tilde{e}_{2i},$$

$$\varepsilon_{(3)i} = C_{31}\tilde{e}_{1i} + C_{32}\tilde{e}_{2i} + C_{33}\tilde{e}_{3i},$$

gdzie C_{ij} jest elementem z i-tego wiersza oraz j-tej kolumny macierzy $\boldsymbol{C}$. Dlatego też wzór definiujący prawdopodobieństwo (1.101) można inaczej zapisać następująco:

$$P\left(y_{(1)i}=1, y_{(2)i}=1, y_{(3)i}=1\right)=P\left(\varepsilon_{(3)i} \le \frac{\left(\boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)}-C_{32}\tilde{e}_2^*-C_{31}\tilde{e}_1^*\right)}{C_{33}}\right)\times$$
$$\times P\left(\varepsilon_{(2)i} \le \frac{\left(\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}-C_{21}\tilde{e}_1^*\right)}{C_{22}}\right)\times P\left(\varepsilon_{(1)i} \le \frac{\left(\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right)}{C_{11}}\right), \tag{1.103}$$

gdzie $\tilde{e}_1^*$ jest zmienną o standardowym rozkładzie normalnym, uciętą w punkcie $\boldsymbol{x}_{(1)i}\beta_{(1)}$, natomiast $\tilde{e}_2^*$ jest zmienną o standardowym rozkładzie normalnym, uciętą w punkcie $\left(\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}-C_{21}\tilde{e}_1^*\right)/C_{22}$. Wartość trzeciego czynnika wyrażenia (1.103) oblicza się natychmiastowo. Po wylosowaniu wartości zmiennych losowych $\tilde{e}_1^*$ oraz $\tilde{e}_2^*$ obliczany jest pierwszy i drugi czynnik wyrażenia (1.103) i tym samym znajdowane jest prawdopodobieństwo, że zmienne dyskretne przyjmują określone wartości.

Zaproponowany w pracach Borscha-Supana i Hajivassiliou (1993), Keane'a (1994) oraz Hajivassiliou i Ruuda (1994) symulator GHK polega na wylosowaniu wartości zmiennych $\tilde{e}_1^*$ oraz $\tilde{e}_2^*$ z odpowiednich rozkładów uciętych, a następnie rekursywnym obliczeniu prawdopodobieństwa danego wzorem (1.103). Po wykonaniu odpowiedniej liczby replikacji obliczana jest wartość symulowanego prawdopodobieństwa. Następnie w danym kroku algorytmu iteracyjnego obliczona wartość prawdopodobieństwa jest uwzględniana w funkcji wiarygodności zdefiniowanej wzorem (1.100). Jak pokazali Borsch-Supan i Hajivassiliou (1993), symulator GHK posiada pożądane własności. Symulowane prawdopodobieństwa są nieobciążone, ograniczone do przedziału (0; 1), natomiast symulowana funkcja największej wiarygodności jest ciągła i nieobciążona względem parametrów.

1.11. Endogeniczny model probitowy

W modelach probitowych może pojawić się problem endogeniczności regresorów. Dlatego też zaproponowane zostały metody estymacji parametrów uwzględniające brak egzogeniczności zmiennych objaśniających. Endogeniczny model probitowy w swojej klasycznej postaci przyjmuje postać (por. Rivers, Vuong, 1988):

$$\begin{aligned} y_i^* &= \tilde{\boldsymbol{y}}_i^T \boldsymbol{\varsigma} + \breve{\boldsymbol{x}}_{(1)i}\, \boldsymbol{\beta} + \varepsilon_{(1)i}, \\ \tilde{\boldsymbol{y}}_i &= \breve{\boldsymbol{x}}_i\, \boldsymbol{\pi} + \boldsymbol{\varepsilon}_{(2)i}, \\ y_i &= I\left\{ y_i^* > 0 \right\}, \end{aligned} \tag{1.104}$$

gdzie $\breve{\boldsymbol{x}}_i = \left[\breve{\boldsymbol{x}}_{(1)i} \quad \breve{\boldsymbol{x}}_{(2)i} \right]$. Elementy wektora $\breve{\boldsymbol{x}}_{(2)i}$ są instrumentami i ich liczba nie może być niższa od liczby endogenicznych zmiennych objaśniających wchodzących w skład wektora $\tilde{\boldsymbol{y}}_i$. W odniesieniu do macierzy wariancji-kowariancji między składnikami losowymi przyjmowane jest założenie, że:

$$E\left(\begin{bmatrix} \varepsilon_{(1)i} \\ \boldsymbol{\varepsilon}_{(2)i} \end{bmatrix} \begin{bmatrix} \varepsilon_{(1)i} & \boldsymbol{\varepsilon}_{(2)i}^T \end{bmatrix} \right) = \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \sigma_{\varepsilon 1}^2 & \boldsymbol{\Sigma}_{\varepsilon 1 \varepsilon 2} \\ \boldsymbol{\Sigma}_{\varepsilon 1 \varepsilon 2} & \boldsymbol{\Sigma}_{\varepsilon 2} \end{bmatrix}. \tag{1.105}$$

W celu oszacowania parametrów modelu (1.104) rozważane były cztery główne metody. Jednym z podejść, zaproponowanym przez Lesliego Godfreya i Michaela Wickensa (1982), jest zastosowanie metody największej wiarygodności z ograniczoną informacją. Okazało się, że metoda ta napotyka na szereg problemów numerycznych, zwłaszcza w przypadku dużych modeli. Alternatywnym podejściem do estymacji parametrów endogenicznych modeli probitowych jest zastosowanie zaproponowanej przez Lung-Fei Lee (1981) probitowej metody zmiennych instrumentalnych. Punktem wyjścia do zastosowania tej metody jest postać zredukowana modelu (1.104):

$$y_i^* = \left(\breve{\boldsymbol{x}}_i\, \boldsymbol{\pi} \right) \boldsymbol{\varsigma} + \breve{\boldsymbol{x}}_{(1)i}\, \boldsymbol{\beta} + \varepsilon_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}^T \boldsymbol{\varsigma}. \tag{1.106}$$

Dlatego też brzegowa funkcja wiarygodności dla zmiennej *y*, pod warunkiem, że dane są elementy macierzy $\boldsymbol{X}$, dana jest wzorem:

$$\begin{aligned} \ln L_I^*\left(\frac{\boldsymbol{\varsigma}}{\omega}, \frac{\boldsymbol{\beta}}{\omega}, \boldsymbol{\pi} \right) = \sum_{i=1}^{I} y_i \ln \Phi\left(\frac{1}{\omega}\left(\breve{\boldsymbol{x}}_i\, \boldsymbol{\pi} \right) \boldsymbol{\varsigma} + \frac{1}{\omega} \breve{\boldsymbol{x}}_{(1)i}\, \boldsymbol{\beta} \right) + \\ \left(1 - y_i\right) \ln\left(1 - \Phi\left(\frac{1}{\omega}\left(\breve{\boldsymbol{x}}_i\, \boldsymbol{\pi} \right) \boldsymbol{\varsigma} + \frac{1}{\omega} \breve{\boldsymbol{x}}_{(1)i}\, \boldsymbol{\beta} \right) \right), \end{aligned} \tag{1.107}$$

gdzie:

$$\omega^2 = 1 + \left(\varsigma + \boldsymbol{\lambda\lambda}\right)^T \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \left(\varsigma + \boldsymbol{\lambda\lambda}\right),$$
$$\boldsymbol{\lambda\lambda} = \left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}\right)^{-1} \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 1}.$$

Ideą probitowej metody zmiennych instrumentalnych jest oszacowanie parametrów równań dla wszystkich zmiennych tworzących wektor $\tilde{\boldsymbol{y}}_i$, a następnie wykorzystanie zgodnego estymatora $\hat{\boldsymbol{\pi}}$ w celu maksymalizacji wyrażenia $\ln L_I^* \left(\frac{\varsigma}{\omega}, \frac{\boldsymbol{\beta}}{\omega}, \hat{\boldsymbol{\pi}} \right)$ ze względu na wektory $\frac{\varsigma}{\omega}$ oraz $\frac{\boldsymbol{\beta}}{\omega}$.

Alternatywną metodą estymacji parametrów endogenicznego modelu probitowego jest zaproponowana przez Takeshiego Amemiyę (1978) uogólniona dwustopniowa metoda probitowa. Polega ona na maksymalizacji formy zredukowanej (1.107) bez nakładania restrykcji:

$$\ln L_I^* \left(\boldsymbol{\tau}_*\right) = \sum_{i=1}^{I} \left\{ y_i \ln \Phi \left(\breve{\boldsymbol{x}}_i \boldsymbol{\tau}_* \right) + \left(1 - y_i\right) \ln \left(1 - \Phi \left(\breve{\boldsymbol{x}}_i \boldsymbol{\tau}_* \right) \right) \right\}. \quad (1.108)$$

Wyrażenie (1.108) maksymalizowane jest ze względu na $\boldsymbol{\tau}_*$, gdzie:

$$\boldsymbol{\tau}_* = \boldsymbol{\pi} \left(\frac{\varsigma}{\omega} \right) + \boldsymbol{J} \left(\frac{\boldsymbol{\beta}}{\omega} \right), \quad (1.109)$$

natomiast $\boldsymbol{J}$ jest macierzą selekcji przekształcającą wektor $\breve{\boldsymbol{x}}_{\boldsymbol{i}}$ w wektor $\breve{\boldsymbol{x}}_{(1)\boldsymbol{i}}$. Zastąpienie $\boldsymbol{\tau}_*$ oraz π oszacowaniami z próby w równaniu (1.109) implikuje następującą zależność:

$$\hat{\boldsymbol{\tau}}_* = \left[\hat{\boldsymbol{\pi}} \quad \boldsymbol{J}\right] \begin{bmatrix} \frac{\boldsymbol{\gamma}}{\omega} \\ \frac{\boldsymbol{\beta}}{\omega} \end{bmatrix} + \left(\hat{\boldsymbol{\tau}}_* - \boldsymbol{\tau}_*\right) - \left(\hat{\boldsymbol{\pi}}_* - \boldsymbol{\pi}\right) \left(\frac{\varsigma}{\omega} \right) = \hat{\boldsymbol{H}} \begin{bmatrix} \frac{\varsigma}{\omega} \\ \frac{\boldsymbol{\beta}}{\omega} \end{bmatrix} + \boldsymbol{e}, \quad (1.110)$$

gdzie:

$$\hat{\boldsymbol{H}} = \left[\hat{\boldsymbol{\pi}} \quad \boldsymbol{J}\right],$$
$$\boldsymbol{e} = \left(\hat{\boldsymbol{\tau}}_* - \boldsymbol{\tau}_*\right) - \left(\hat{\boldsymbol{\pi}}_* - \boldsymbol{\pi}\right)\left(\frac{\boldsymbol{\varsigma}}{\omega}\right).$$

Jak widać zatem, problem estymacyjny sprowadza się do oszacowania parametrów modelu regresji liniowej. Zastosowanie MNK-estymatora dla równania (1.110) prowadzi do uzyskania zgodnych oszacowań parametrów $\frac{\boldsymbol{\varsigma}}{\omega}$ oraz $\frac{\boldsymbol{\beta}}{\omega}$. Uzyskanie estymatora o wyższej efektywności wymaga zastosowania uogólnionej metody najmniejszych kwadratów. Jeśli $\hat{\boldsymbol{V}}$ jest zgodnym estymatorem asymptotycznej macierzy kowariancji składników losowych $\boldsymbol{e}$, wówczas wyprowadzony przez Amemiyę (1978) estymator uogólnionej dwustopniowej metody probitowej przyjmuje postać:

$$\begin{bmatrix} \frac{\hat{\boldsymbol{\varsigma}}}{\omega} \\ \frac{\hat{\boldsymbol{\beta}}}{\omega} \end{bmatrix} = \left(\hat{\boldsymbol{H}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{H}}\right)^{-1} \hat{\boldsymbol{H}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{\tau}}_*. \tag{1.111}$$

Douglas Rivers oraz Quang H. Vuong (1988) zaproponowali estymację parametrów dynamicznego modelu probitowego za pomocą dwustopniowej warunkowej metody największej wiarygodności. Punktem wyjścia jest łączna funkcja gęstości zmiennych $\boldsymbol{y}$ oraz $\tilde{\boldsymbol{y}}$, której postać jest następująca:

$$\begin{aligned} & h\left(y_i, \tilde{\boldsymbol{y}}_i \mid \breve{\boldsymbol{x}}_i; \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\lambda\lambda}, \boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}\right) = \\ & = \left(2\tilde{\pi}\right)^{-\frac{(m(end)+1)}{2}} \left(det\left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}\right)\right)^{-1} \times \\ & \times \left[\int_{c_1}^{\infty} exp\left\{-\frac{1}{2}\left[u^2 - 2\boldsymbol{\lambda\lambda}^T \boldsymbol{\varepsilon}_{(2)i} u + \boldsymbol{\varepsilon}_{(2)i}^T \left(\left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}\right)^{-1} + \boldsymbol{\lambda\lambda}^T\right) \boldsymbol{\varepsilon}_{(2)i}\right]\right\} du\right]^{y_i} \\ & \times \left[\int_{-\infty}^{c_1} exp\left\{-\frac{1}{2}\left[u^2 - 2\boldsymbol{\lambda\lambda}^T \boldsymbol{\varepsilon}_{(2)i} u + \boldsymbol{\varepsilon}_{(2)i}^T \left(\left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}\right)^{-1} + \boldsymbol{\lambda\lambda}^T\right) \boldsymbol{\varepsilon}_{(2)i}\right]\right\} du\right]^{1-y_i}, \end{aligned} \tag{1.112}$$

gdzie:

$$c_i = -\left(\tilde{\boldsymbol{y}}_i^T \boldsymbol{\varsigma} - \breve{\boldsymbol{x}}_{(1)i} \boldsymbol{\beta} \right).$$

Funkcja gęstości (1.112) może zostać zapisana jako iloczyn funkcji gęstvości dla modelu probitowego oraz funkcji gęstości dla standardowego rozkładu normalnego:

$$\begin{aligned} h\left(y_i, \tilde{\boldsymbol{y}}_i \mid \boldsymbol{x}_i; \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\lambda\lambda}, \boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right) = \\ = f_{ep}\left(y_i \mid \tilde{\boldsymbol{y}}_i, \breve{\boldsymbol{x}}_i; \boldsymbol{\varsigma}, \boldsymbol{\beta}, \boldsymbol{\lambda\lambda}, \boldsymbol{\pi} \right) g_{ep}\left(\tilde{\boldsymbol{y}}_i \mid \breve{\boldsymbol{x}}_i; \boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right), \end{aligned} \tag{1.113}$$

gdzie omawiane funkcje gęstości przyjmują następujące postacie:

$$\begin{aligned} f_{ep}\left(y_i \mid \tilde{\boldsymbol{y}}_i, \breve{\boldsymbol{x}}_i; \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\pi} \right) = \\ = \Phi\left(\tilde{\boldsymbol{y}}_i^T \boldsymbol{\varsigma} + \breve{\boldsymbol{x}}_{(1)i} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{2i} \boldsymbol{\lambda\lambda} \right)^{y_i} \left[1 - \Phi\left(\tilde{\boldsymbol{y}}_i^T \boldsymbol{\varsigma} + \breve{\boldsymbol{x}}_{(1)i} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{(2)i} \boldsymbol{\lambda\lambda} \right) \right]^{1-y_i} \end{aligned} \tag{1.114}$$

$$\begin{aligned} g_{ep}\left(\tilde{\boldsymbol{y}}_i \mid \breve{\boldsymbol{x}}_i; \boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right) = \\ = \left(2\tilde{\pi} \right)^{-\frac{m(end)}{2}} \left(det\left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right) \right)^{-1/2} \\ exp\left\{ -\frac{1}{2} \left(\left(\tilde{\boldsymbol{y}}_i - \boldsymbol{\pi}^T \breve{\boldsymbol{x}}_i^T \right)^T \left(\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right)^{-1} \left(\tilde{\boldsymbol{y}}_i - \boldsymbol{\pi}^T \breve{\boldsymbol{x}}_i^T \right) \right) \right\} \end{aligned} \tag{1.115}$$

Estymator dwustopniowej warunkowej metody największej wiarygodności uzyskuje się w dwóch krokach. W pierwszym kroku maksymalizowana jest funkcja wiarygodności:

$$\ln L_I^g \left(\boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right) = \sum_{i=1}^{I} \ln \left\{ g\left(\tilde{\boldsymbol{y}}_i \mid \breve{\boldsymbol{x}}_i; \boldsymbol{\pi}, \boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2} \right) \right\} \tag{1.116}$$

względem π oraz $\boldsymbol{\Sigma}_{\varepsilon 2 \varepsilon 2}$. Następnie maksymalizuje się funkcję wiarygodności:

$$\ln L_I^f\left(\varsigma, \boldsymbol{\beta}, \boldsymbol{\lambda}, \hat{\boldsymbol{\pi}}\right) = \sum_{i=1}^{I} \ln\left\{ f\left(y_i \mid \tilde{\boldsymbol{y}}_i, \breve{\boldsymbol{x}}_i; \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\lambda}\boldsymbol{\lambda}, \hat{\boldsymbol{\pi}} \right) \right\} \tag{1.117}$$

względem pozostałych parametrów.

Rivers i Vuong (1988), wyznaczając analitycznie macierze wariancji-kowariancji estymatorów parametrów oraz przeprowadzając symulacje Monte Carlo, pokazali, że zarówno w dużych, jak i małych próbach estymacja dwustopniową warunkową metodą największej wiarygodności prowadzi do uzyskania bliższych prawdziwym wartościom (w porównaniu z pozostałymi estymatorami) oszacowań parametrów.

1.12. Podsumowanie

W kolejnych rozdziałach niniejszej monografii rozważane są różne warianty modeli wielopoziomowych (liniowe modele wielopoziomowe, uogólnione liniowe modele wielopoziomowe, modele wielopoziomowe dla zmiennych wielomianowych nieuporządkowanych, wielopoziomowe wielorównaniowe modele probitowe), prezentowane jest również ich wykorzystanie w kontekście analizy danych indywidualnych. Ich idea polega na rozszerzeniu modeli tradycyjnych (bez aspektów wielopoziomowych) i uwzględnieniu zmiennych kontekstowych.

W związku z tym wydaje się, że prezentacja podstawowych metod analizy danych indywidualnych jest uzasadniona. Właściwe zrozumienie metod estymacji parametrów modeli wielopoziomowych wymaga odpowiedniej wiedzy z zakresu estymacji parametrów modeli wykorzystywanych podczas analizy danych indywidualnych. A zatem zaprezentowane zostały wszystkie metody analizy danych indywidualnych, które w dalszej części książki pojawiają się w kontekście modeli wielopoziomowych. Przedstawione zostały też inne metody, które nie są rozszerzane przez uwzględnienie zmiennych kontekstowych. Prezentacja tych metod jest jednak niezbędna ze względu na powiązania występujące między różnymi modelami wykorzystywanymi w celu analizy danych indywidualnych.

2. Dane regionalne wykorzystywane w badaniach ekonomicznych

2.1. Podział administracyjny, statystyczny i historyczny Polski

Z administracyjnego punktu widzenia Polska dzieli się na szesnaście województw. W każdym z nich wyróżnia się powiaty, w obrębie których znajdują się gminy. Obecna trójstopniowa struktura podziału terytorialnego obowiązuje od 1 stycznia 1999 roku i jest wynikiem reformy administracyjnej. Województwa mają charakter rządowo-samorządowy, natomiast powiaty i gminy są samorządowe.

Wyróżnia się rządowe oraz samorządowe stolice województw. W stolicy rządowej urzęduje wojewoda, natomiast w stolicy samorządowej znajduje się Urząd Marszałkowski, w którym władzę sprawuje marszałek województwa. W przypadku czternastu województw to samo miasto pełni funkcję rządowej i samorządowej stolicy. W województwie kujawsko-pomorskim oraz lubuskim funkcje te są podzielone. Tabela 4 prezentuje nazwy województw oraz miasta siedziby urzędów wojewódzkich i marszałkowskich.

W ramach poszczególnych województw wyróżnia się powiaty. Wraz z wprowadzeniem reformy administracyjnej utworzono 373 powiaty (308 ziemskich, 65 grodzkich). W przypadku powiatów grodzkich mamy do czynienia z miastami na prawach powiatu. Wszystkie miasta wojewódzkie wymienione w tabeli 4 mają status miast na prawach powiatu. Oprócz tego zdecydowana większość miast, które w latach 1975–1998 pełniły funkcję stolic województw oraz utraciły ją w wyniku wprowadzenia reformy administracyjnej, zostało miastami na prawach powiatu po 1999 roku. Wyjątek stanowią Piła, Ciechanów oraz Sieradz. Po reformie administracyjnej z 1999 roku miastami na prawach powiatu stały się także inne miasta, liczące wówczas powyżej 100 000 mieszkańców (Sosnowiec, Grudziądz, Gdynia, Gliwice, Zabrze, Bytom, Tychy, Dąbrowa Górnicza, Ruda Śląska, Rybnik, Chorzów), a także niektóre miasta w dużych aglomeracjach lub konurbacjach

(Jastrzębie-Zdrój, Jaworzno, Mysłowice, Piekary Śląskie, Świętochłowice, Siemianowice Śląskie, Żory, Sopot, Świnoujście). 1 stycznia 2003 roku Wałbrzych przestał być miastem na prawach powiatu i został włączony do powiatu wałbrzyskiego. Od 1 stycznia 2013 roku znowu istnieje powiat Miasto Wałbrzych.

Tabela 4. Województwa w Polsce i ich stolice

Województwo	Siedziba urzędu wojewódzkiego	Siedziba urzędu marszałkowskiego
Dolnośląskie	Wrocław	Wrocław
Kujawsko-pomorskie	Bydgoszcz	Toruń
Lubelskie	Lublin	Lublin
Lubuskie	Gorzów Wielkopolski	Zielona Góra
Łódzkie	Łódź	Łódź
Małopolskie	Kraków	Kraków
Mazowieckie	Warszawa	Warszawa
Opolskie	Opole	Opole
Podkarpackie	Rzeszów	Rzeszów
Podlaskie	Białystok	Białystok
Pomorskie	Gdańsk	Gdańsk
Śląskie	Katowice	Katowice
Świętokrzyskie	Kielce	Kielce
Warmińsko-mazurskie	Olsztyn	Olsztyn
Wielkopolskie	Poznań	Poznań
Zachodniopomorskie	Szczecin	Szczecin

Źródło: opracowanie własne.

1 stycznia 2002 roku wprowadzono zmiany w podziale administracyjnym państwa. Utworzony w 1999 roku powiat olecko-gołdapski został podzielony na powiat olecki oraz gołdapski. Zmieniła się nazwa oraz siedziba powiatu tyskiego. Zamiast Tychów stolicami powiatu bieruńsko-lędzińskiego zostały Bieruń oraz Lędziny. Z części innych powiatów utworzono następujące powiaty: brzeziński, leski, łobeski, wschowski.

Ze względu na przystąpienie Polski do Unii Europejskiej konieczne było dostosowanie systemu statystycznych jednostek terytorialnych w Polsce do analogicznego systemu obowiązującego w UE. Dlatego też 26 listopada 2005 roku, czyli półtora roku po akcesji, zostały formalnie w Polsce wprowadzone jednostki NUTS. Odpowiadały one przyjętej w 2000 roku w Polsce nomenklaturze jednostek terytorialnych do celów statystycznych. Wówczas wyróżnionych zostało 6 jednostek na poziomie NUTS-1 (regiony), 16 jednostek na poziomie NUTS-2 (województwa) oraz 45 jednostek na poziomie NUTS-3 (podregiony). Na początku podział na jednostki NUTS-1 oraz NUTS-2 nie ulegał zmianie. Zmieniał się natomiast podział

na jednostki NUTS-3. 1 stycznia 2008 roku zwiększono liczbę jednostek NUTS-3 do 66. 1 stycznia 2015 roku liczba ta wzrosła do 72. Natomiast 1 stycznia 2018 roku podział statystyczny Polski został zmieniony na wszystkich poziomach NUTS. Jednostki poziomu NUTS-1 nazywają się obecnie makroregionami i jest ich siedem. Makroregion południowy składa się z województwa małopolskiego oraz śląskiego. W makroregionie północno-zachodnim znajdują się województwa wielkopolskie, zachodniopomorskie i lubuskie. Województwo dolnośląskie oraz opolskie znajdują się w makroregionie południowo-zachodnim. Województwa kujawsko-pomorskie, warmińsko-mazurskie oraz pomorskie tworzą makroregion północny. Do makroregionu centralnego należą obecnie województwa łódzkie i świętokrzyskie. Makroregion wschodni składa się z województw lubelskiego, podkarpackiego i podlaskiego. Ostatnim makroregionem, który powstał dopiero w 2018 roku, jest warszawski. Jednostki poziomu NUTS-2 to regiony, które są województwami. Wyjątek stanowią regiony wchodzące w skład makroregionu województwa mazowieckiego. Województwo mazowieckie, które przed 2018 rokiem stanowiło jedną jednostkę statystyczną poziomu NUTS-2, zostało podzielone na dwie jednostki statystyczne: region warszawski stołeczny (podregiony: Miasto Warszawa, warszawski wschodni, warszawski zachodni) oraz region mazowiecki regionalny (podregiony: radomski, ciechanowski, płocki, ostrołęcki, siedlecki, żyrardowski). Od 2018 roku wyróżnia się w Polsce 73 jednostki poziomu NUTS-3, czyli podregiony. Jednostkami poziomu NUTS-4 są powiaty. Gminy wchodzące w skład powiatów są jednostkami poziomu NUTS-5.

Jak widać zatem, zgodność między jednostkami statystycznymi a administracyjnymi występuje na poziomie NUTS-4 oraz NUTS-5. Jednostki poziomu NUTS-1 oraz NUTS-3 mają charakter jednostek statystycznych i nie mają swoich odpowiedników administracyjnych. Z wyjątkiem województwa mazowieckiego istnieje zgodność między jednostkami statystycznymi a administracyjnymi na poziomie NUTS-2.

Oprócz przedstawienia podziału administracyjnego i statystycznego terytorium Rzeczypospolitej Polskiej, warto także wspomnieć o podziale Polski na cztery części w związku z jego XIX- i XX-wieczną historią. W okresie między Kongresem Wiedeńskim a momentem rozpoczęcia pierwszej wojny światowej (lata 1815–1914) polskie terytorium było podzielone na trzy strefy ze stabilnymi granicami. Stanowiły je zabór austriacki, zabór pruski oraz zabór rosyjski. Czwarta część to Ziemie Zachodnie i Północne – byłe tereny niemieckie, gdzie polska ludność osiedliła się głównie po drugiej wojnie światowej. Wyniki badań ekonomicznych (por. np. Tokarski, 2013) oraz socjologicznych (por. m.in. Jałowiecki, 1996; Krzemiński, 2009; Grabowski, 2018) wskazują, że na terenie czterech analizowanych obszarów uformowane zostały odmienne wzorce kulturowe oraz orientacje światopoglądowe,

natomiast kapitał społeczny i aktywność obywatelska odmiennie kształtowały się w różnych częściach kraju. Położenie danej miejscowości na terenie jednego z czterech obszarów sprawiało, że jej mieszkańcy musieli dostosować się do wzorców kulturowych tam obowiązujących. Doprowadziło to do wyodrębnienia się różnych norm społecznych oraz poglądów na kwestię polityki i państwowości (por. Krzemiński, 2009). Dodatkowo polityka zaborców wobec ludności polskiej miała istotny wpływ na proces formowania się tych wzorców (por. Raciborski, 1997). Oddziaływanie okresu zaborów na trwałość zróżnicowania przestrzennego wzorców kulturowych dominujących w poszczególnych częściach Polski wynika z faktu, że kiedy wraz z rewolucją przemysłową następowały istotne przemiany społeczne, polskie tereny włączone były do trzech odrębnych mocarstw. Ludność, która osiedliła się na byłych terenach niemieckich po drugiej wojnie światowej, również różniła się od ludności zamieszkującej „stare" tereny Polski. Na terenie Ziem Zachodnich i Północnych uformowała się tak zwana zbiorowość postmigracyjna (por. Bartkowski, 2003). Konieczność migracji po zakończeniu drugiej wojny światowej oraz fakt uwolnienia się od kontroli grupowej spowodowały, że w analizowanej zbiorowości uformowały się wzorce kulturowe odmienne od tych, które obserwuje się w innych regionach Polski (Jasiewicz, 1977).

Ukształtowane w historii wzorce kulturowe (np. postawy względem instytucji, państwa i prawa) przetrwały do obecnych czasów. Przekazywane z pokolenia na pokolenie postawy światopoglądowe sprawiły, że wewnątrzregionalne zróżnicowanie wzorców zachowań jest zdecydowanie niższe niż zróżnicowanie międzyregionalne. Dlatego też przynależność jednostki administracyjnej do określonego regionu historycznego jest ważnym czynnikiem wpływającym na poziom rozwoju ekonomicznego oraz postawy względem instytucji, państwa i prawa.

2.2. Historyczno-kulturowe zróżnicowanie terytorium obecnej Rzeczpospolitej Polskiej

W badaniach empirycznych wykorzystujących dane regionalne, które omówione są w dalszych częściach niniejszej monografii, często występują zmienne zero-jedynkowe związane z przynależnością miejscowości do jednej z czterech części Polski. Krótkie nawiązanie do XIX- i XX-wiecznej historii Polski i zrozumienie jej wpływu na różnice w kulturze obserwowanej w różnych częściach kraju wydaje się więc bardzo ważne. Oczywiście niniejsza monografia nie ma charakteru książki historycznej. Nie jest w stanie zastąpić tak znanej pozycji dotyczącej historii gospodarczej Polski jak na przykład książka Andrzeja Jezierskiego i Cecylii Leszczyńskiej (2011). Krótkie omówienie historii Polski z okresu zaborów oraz

analiza różnic między zbiorowością postmigracyjną terenów Polski zachodniej i północnej a zamieszkującymi inne części naszego kraju wydaje się niezbędne.

Teren byłego zaboru austriackiego (dawna Galicja) określa się jako obszar charakteryzujący się najwyższym poziomem świadomości narodowej mieszkańców oraz największą dojrzałością demokratyczną (por. Bartkowski, 2003; Zarycki, 2015). Wynika to z faktu, że mieszkańcy tego zaboru mieli zdecydowanie więcej swobód obywatelskich i politycznych w porównaniu z osobami mieszkającymi w XIX i na początku XX wieku na terenie zaboru pruskiego i rosyjskiego. Po 1861 roku na terenie analizowanego zaboru nastąpiła demokratyzacja praw wyborczych i rozszerzenie wolności politycznych. W rezultacie rozprzestrzeniła się polska kultura, powstały krajowe organizacje polityczne, wzmocniona została świadomość narodowa mieszkańców. Dowodem na wysoki poziom swobód obywatelskich mieszkańców zaboru austriackiego był fakt, że język polski został dopuszczony na wszystkich poziomach edukacji (od szkół podstawowych po uniwersytety), natomiast osoby polskiego pochodzenia dominowały w galicyjskim samorządzie i sejmie (por. Bartkowski, 2003).

Ważnym dziedzictwem zaboru austriackiego obserwowanym w południowej części drugiej Rzeczpospolitej w latach 1918–1939 oraz w Polsce południowo-wschodniej po II wojnie światowej była bardzo silna pozycja kościoła katolickiego. Wynikała ona z faktu, że zabór austriacki był jedyną częścią XIX-wiecznego terytorium ziem polskich, w której kościół katolicki nie był prześladowany. Ówczesne Austro-Węgry były monarchią katolicką.

Analizując obszar obecnej Galicji (tereny byłego zaboru austriackiego), należy zwrócić uwagę na obserwowany tam typ osadnictwa. Gminy leżące w województwie podkarpackim oraz dużej części województwa małopolskiego na ogół składają się z małej liczby dużych wsi. Liczebność wsi jest wystarczająca do podejmowania kolektywnych działań, ale nie jest wystarczająco duża, aby umożliwić mieszkańcom bycie anonimowym. Należy zwrócić uwagę także na fakt, że odsetek rdzennych mieszkańców na terenie byłego zaboru austriackiego jest zdecydowanie wyższy niż w innych częściach Polski. Czynniki te sprawiają, że jednostkom tworzącym zbiorowości w wielu miejscowościach trudniej jest odrzucić wzorce kulturowe obowiązujące na danym terenie.

Analizując historyczne dziedzictwo zaboru austriackiego, niektórzy badacze (np. Gorzelak, Jałowiecki, 1996) wskazywali na takie cechy regionu i jego mieszkańców, jak zaściankowość, kolektywizm, inklinacje autorytarne oraz klerykalizm, argumentując, że integracja tego obszaru z zachodnią Europą jest utrudniona. Jednocześnie badacze ci wskazywali, że zła sytuacja ekonomiczna województwa podkarpackiego (którego gminy leżą na terenie byłego zaboru austriackiego) wynika

z faktu, że jego mieszkańcy mają skłonność do myślenia w kategoriach moralnych i ideologicznych, a nie pragmatycznych.

Inni badacze wskazują natomiast na pozytywne aspekty dziedzictwa okresu rozbiorów obserwowane na terytorium obecnego województwa małopolskiego i podkarpackiego. Wysoki poziom frekwencji wyborczej odnotowywany w gminach położonych w Polsce południowo-wschodniej wynika z dłuższych tradycji demokratycznych w porównaniu z innymi częściami kraju. Ponieważ poziom zasiedziałości (udział rdzennych mieszkańców w populacji) jest zdecydowanie wyższy niż w innych częściach Polski, a wiele rodzin z tego terenu dziedziczy nieruchomości od wielu pokoleń, poszanowanie dla własności wśród mieszkańców byłej Galicji jest bardzo silne (Bartkowski, 2003). Dodatkowo, jak wskazuje Janusz Majcherek (1995), wysoki wskaźnik religijności mieszkańców prowadzi do wzrostu moralności, szacunku wobec obowiązującego prawa oraz niskiej przestępczości. Brak anonimowości mieszkańców i silniejsza kontrola grupowa wynikające z faktu, że typowe wsie postgalicyjskie charakteryzują się zdecydowanie większą liczbą mieszkańców w porównaniu z ich odpowiednikami położonymi na terenach byłego zaboru pruskiego czy rosyjskiego sprawiają, że warunki do rozwoju kapitału społecznego są lepsze (Hann, Magocsi, 2014).

Wysoki poziom rozwoju ekonomicznego obecnych ziem zachodniej i centralnej Wielkopolski, zachodniej części województwa kujawsko-pomorskiego, a także Trójmiasta i okolic może być dziedzictwem okresu rozbiorów. W tamtym czasie analizowany tereny należały do zaboru pruskiego. Jak wskazują między innymi Ryszard Żukowski (2004) oraz Tomasz Zarycki (2015), rolnictwo na terenie zaboru pruskiego było w XIX wieku zdecydowanie bardziej zmodernizowane niż w dwóch pozostałych zaborach. Jego funkcjonowanie opierało się na lokalnym kredycie, rozwoju spółdzielczości oraz działalności stowarzyszeń samopomocowych. Mieszkający na terenie zaboru pruskiego polscy przedsiębiorcy musieli konkurować z zamieszkującymi ten obszar właścicielami firm pochodzącymi z Prus. Zmusiło to polskie przedsiębiorstwa do zwiększenia efektywności prowadzonej działalności.

Analizując wpływ dziedzictwa zaboru pruskiego, należy zwrócić uwagę na rolę, jaką odegrała pruska reforma rolnictwa z 1823 roku. Po jej przeprowadzeniu nastąpiła zmiana struktury gospodarstw ze względu na rozmiar. Dominowały gospodarstwa średnie i duże. Doprowadziło to do pojawienia się nadwyżki ludności mieszkającej na wsi w stosunku do poziomu zatrudnienia w rolnictwie. W związku z tym nastąpiła absorpcja tej nadwyżki przez sektor przemysłu i usług, który rozwinął się w różnych częściach Wielkopolski, Kujaw i Pomorza Gdańskiego. Porównując reformę rolnictwa przeprowadzoną na terenie byłego zaboru pruskiego z analogiczną na terenie zaboru rosyjskiego czy austriackiego, należy zwrócić

uwagę na to, że tylko w pierwszym przypadku jej celem była racjonalizacja wielkości gospodarstw rolnych. Dlatego też omawiana reforma sprawiła, że na analizowanym obszarze nie występowały konflikty między szlachtą a chłopami. Na terenie zaboru pruskiego trudno było znaleźć niewielkie gospodarstwa rolne. Obserwowana obecnie struktura gospodarstw rolnych w różnych częściach Polski jest dziedzictwem okresu rozbiorów. Na terenie byłego zaboru pruskiego dominują gospodarstwa duże, podczas gdy wielkość typowego gospodarstwa rolnego znajdującego się w Polsce centralnej, wschodniej czy południowej jest zdecydowanie mniejsza (Wajda, 1990). Brak konfliktów między szlachtą a chłopami sprawił, że w analizowanym regionie nie zyskiwały na popularności partie chłopskie. Obecnie poziom poparcia dla tych partii obserwowany w zachodniej Wielkopolsce, na Pomorzu Gdańskim czy w okolicach Bydgoszczy i Torunia jest znacznie niższy niż w innych częściach Polski.

Należy także zauważyć, że intensywność konfliktów o podłożu klasowym na terenie byłego zaboru pruskiego była w okresie drugiej Rzeczpospolitej oraz po II wojnie światowej zdecydowanie niższa niż w innych częściach Polski. Wyjaśnienia tej prawidłowości również można poszukiwać w przeszłości historycznej związanej z okresem zaborów. Konflikty, które pojawiały się na terenie zaboru pruskiego w okresie rozbiorów, miały zabarwienie narodowe, a nie klasowe. Były one postrzegane jako konflikty między pruskimi właścicielami a polskimi pracownikami. W okresie zaborów obserwowany był spadek zróżnicowania międzyklasowego w społeczeństwie zamieszkującym tereny byłego zaboru pruskiego. Zdolne jednostki pochodzące z ubogich rodzin mogły liczyć na wsparcie dzięki systemowi stypendiów dla utalentowanej młodzieży. Mieszkańcy analizowanego regionu mieli świadomość, że dzięki intensywnej pracy są w stanie awansować i poprawić jakość swojego życia. Dlatego też na terenie Wielkopolski powstała klasa średnia w skali niespotykanej w innych regionach. Duży udział klasy średniej w społeczeństwie sprawił, że po 1989 roku na relatywnie wysokie poparcie mogły liczyć partie deklarujące prowadzenie polityki liberalnej gospodarczo, wspierające rozwój przedsiębiorczości i położenie właścicieli firm (Matykowski, 2007; Matykowski, Kulczyńska, 2016).

Kulturowy wymiar dziedzictwa zaboru pruskiego nie jest łatwy do zidentyfikowania. Państwo pruskie było bardzo represyjne wobec ludności polskiej. Wszelkie próby manifestacji tożsamości narodowej były zwalczane. Wynika to z faktu, że na terenie Prus wykształciły się ideologie wspierające homogeniczne kulturowo społeczeństwo. Dlatego też powstawały organizacje (np. Hakata) zajmujące się walką z polską kulturą, językiem itp. W związku z tym na terenie zaboru pruskiego nie istniało szkolnictwo wyższe, a nauka na poziomie średnim i podstawowym odbywała się w języku niemieckim. Z jednej

strony brakowało osób z wyższym wykształceniem, a z drugiej mieszkańcy tych ziem mieli dobre możliwości nauki na poziomie podstawowym.

Żukowski (2004) argumentuje, że wysoki poziom etyki pracy obserwowany na terenie Prus w XIX wieku został „przekazany" polskim mieszkańcom ówczesnego zaboru pruskiego. Należy także podkreślić, że rozwój ekonomiczny oraz poprawa sytuacji materialnej mieszkańców tego zaboru były sposobami dbania o „polskość" (Bartkowski, 2003). Powstałe w okresie zaborów wzorce kulturowe dotyczące pracowitości i dbania o własną sytuację materialną zostały następnie przekazane młodszym mieszkańcom Wielkopolski i miały wpływ na dominujący światopogląd wśród przedstawicieli następnych pokoleń (Podemski, Ziółkowski, 2007).

Ważną kwestią podczas analizy wyników głosowania na terenie byłego zaboru pruskiego jest obserwowany w okresie powojennym wyższy poziom frekwencji wyborczej w porównaniu z resztą Polski. Najczęściej spotyka się dwa uzasadnienia tego zjawiska (por. Zarycki, 2015). Po pierwsze, likwidacja analfabetyzmu na terenie zaboru pruskiego w XIX wieku spowodowała, że po II wojnie światowej najstarsi mieszkańcy tego regionu potrafili czytać i pisać, mieli więc lepsze kompetencje do oddawania głosów w wyborach. Po drugie, wskazuje się, że legendarna „pruska dyscyplina", z którą mieszkańcy Wielkopolski spotkali się w okresie zaborów, zakorzeniona została w ich świadomości. Zgodnie z tą interpretacją uczestnictwo w głosowaniu było traktowane jako obywatelski obowiązek (Żukowski, 2004).

Zabór rosyjski był najdalej wysuniętą na zachód częścią Imperium Rosyjskiego. Położenie tych terenów przyczyniło się do lokalizacji wielu istotnych centrów przemysłowych. Inwestorzy, pochodzący przede wszystkim z Prus, chętnie lokowali swoje zakłady przemysłowe w Łodzi, Dąbrowie Górniczej i innych miastach zaboru rosyjskiego położonych blisko granicy prusko-rosyjskiej. Budowali oni fabryki niedaleko swojego macierzystego kraju, korzystali z tańszej siły roboczej dostępnej na terenach rosyjskich i dodatkowo mieli dostęp do bardzo dużego rynku rosyjskiego. Oznacza to zatem, że fakt podziału terytorium polskiego na zabory oraz sprzyjająca rozwojowi przemysłu polityka taryfowa Imperium Rosyjskiego przyczyniły się do powstania w zachodniej części ówczesnego zaboru rosyjskiego kilku ważnych okręgów przemysłowych (np. łódzkiego, częstochowskiego, Zagłębia Dąbrowskiego). Rozwinęły się wielkie aglomeracje (np. łódzka), które różniły się zdecydowanie od otaczających je wsi i małych miasteczek. Dlatego też ważnym dziedzictwem zaboru rosyjskiego są silne kontrasty między poziomem rozwoju dużych miast a sytuacją ekonomiczną gmin wiejskich położonych relatywnie blisko wielkich aglomeracji.

Zabór rosyjski pozostawił spuściznę w postaci skorumpowanej administracji, autorytaryzmu oraz zacofania cywilizacyjnego (Kochanowicz, 2018). Aspekty

te w drugiej Rzeczpospolitej oraz w okresie powojennym wymieniane były jako główne czynniki hamujące rozwój tych terenów (Zarycki, 2015). Tradycje demokratyczne nie zostały zakorzenione w kulturze terenów zaboru rosyjskiego. Oprócz tego na analizowanym obszarze nie rozwinęła się dobrze infrastruktura.

Duża liczba prężnych uniwersytetów działających na terenie zaboru rosyjskiego w drugiej połowie XIX wieku przyczyniła się do powstania na analizowanym obszarze lewicowo-liberalnej wykształconej inteligencji. Omawiana warstwa społeczna miała ambicje do zachowania przywilejów posiadanych przez szlachtę w okresie przedrozbiorowym. To na terenie tego zaboru w XIX wieku mieszkało wielu sławnych polskich pisarzy i poetów (Władysław Reymont, Henryk Sienkiewicz, Adam Mickiewicz, Maria Konopnicka). Z drugiej strony udział analfabetów był zdecydowanie wyższy niż w innych częściach Polski. W rezultacie wariancja poziomu wykształcenia wśród mieszkańców zaboru rosyjskiego na początku XX wieku była bardzo wysoka.

Analizując wpływ historii na poziom rozwoju społeczno-ekonomicznego, światopogląd, postawy i preferencje polityczne mieszkańców różnych części Polski, nie można zapominać o tak zwanych Ziemiach Zachodnich i Północnych. Po II wojnie światowej tereny te zostały zasiedlone przez ludność napływową. Pochodziła ona głównie z terenów przez Polskę utraconych oraz z Wielkopolski.

Ludność, która zamieszkała północne i zachodnie tereny trzeciej Rzeczpospolitej, okazała się bardziej innowacyjna i mobilna. Dlatego też na przykład wyniki dotyczące regionalnego zróżnicowania poziomu innowacyjności przedsiębiorstw wskazują, że firmy zlokalizowane na terenach obecnego województwa dolnośląskiego czy zachodniopomorskiego są – przy innych czynnikach niezmienionych – bardziej innowacyjne niż przedsiębiorstwa ulokowane w innych częściach Polski (Arendt, Grabowski, 2017). Oprócz tego badania dotyczące szybkości pojawiania się nowinek technologicznych w gospodarstwach domowych w latach siedemdziesiątych czy osiemdziesiątych wskazywały, że na czele rankingu województw (przy podziale na 49 jednostek administracyjnych) były takie województwa jak zielonogórskie, gorzowskie, leszczyńskie, poznańskie czy szczecińskie (por. Bartkowski, 2003).

Analizy dotyczące aspektów światopoglądowych, przeprowadzone jeszcze w latach siedemdziesiątych XX wieku, wskazały, że mieszkańcy tak zwanych Ziem Odzyskanych byli mniej konserwatywni i bardziej tolerancyjni względem konkubinatów, rozwodów, faktu posiadania nieślubnych dzieci w porównaniu z osobami zamieszkującymi inne części Polski (Jasiewicz, 1977). Omawiany niski poziom konserwatyzmu i wyższy poziom innowacyjności mogą wynikać z osłabienia roli tradycyjnych wzorców oraz słabszej kontroli grupowej.

Negatywnymi aspektami związanymi z faktem, że analizowany obszar zamieszkuje społeczność postmigracyjna, są wyższe wskaźniki przestępczości

obserwowane na Ziemiach Zachodnich i Północnych w porównaniu z innymi częściami Polski. Tych tendencji nie da się wyjaśnić tylko i wyłącznie czynnikami strukturalnymi powiązanymi z wysokimi poziomami bezrobocia w niektórych powiatach województwa zachodniopomorskiego i lubuskiego ani też przestępczością związaną z przekraczaniem granicy państw. Jak argumentuje Jerzy Bartkowski (2003), ten sam czynnik może prowadzić zarówno do wyższej otwartości i innowacyjności, jak i wyższej skłonności do łamania prawa. Jest nim brak kontroli grupowej. Migracja mieszkańców różnych części Polski na tereny Ziem Zachodnich i Północnych po II wojnie światowej sprawiła, że nastąpiło osłabienie roli wzorców obowiązujących w miejscu pochodzenia.

Ponieważ jest to relatywnie „nowa" część Polski, słabszy jest poziom identyfikacji narodowej. Na Ziemiach Zachodnich i Północnych mniejszy jest zdecydowanie wpływ kościoła katolickiego w porównaniu z innymi częściami Polski. Omówione czynniki sprawiają, że poziom poparcia dla partii liberalnych oraz lewicowych jest zdecydowanie wyższy niż na przykład w centralnej, południowej i wschodniej części kraju.

Zróżnicowana historia omawianych czterech regionów miała również wpływ na skłonność do formalizowania działań, postawę wobec stosowania i przestrzegania prawa oraz sposób reakcji wobec zaistnienia problemu prawnego (por. Bartkowski, 2003). Ze względu na fakt, że na terenie byłego zaboru pruskiego zlikwidowany został problem analfabetyzmu, mieszkańcom dużej części obecnego województwa wielkopolskiego, kujawsko-pomorskiego czy pomorskiego łatwiej było załatwiać sprawy w urzędach jeszcze w dwudziestoleciu międzywojennym. Jednocześnie rozwój instytucji na terenie analizowanego obszaru był zdecydowanie silniejszy niż w zaborze rosyjskim czy austriackim. Doprowadziło to do wyższej skłonności do formalizowania działań wśród mieszkańców Poznania, Bydgoszczy, Torunia, Gdańska i okolic niż w innych częściach Polski. Stopień formalizacji działań na tak zwanych Ziemiach Zachodnich i Północnych po II wojnie światowej był zdecydowanie niższy niż w innych częściach kraju. Wynika to z faktu, że formalne instytucje zostały stworzone na analizowanych terenach dopiero po zakończeniu wojny, więc mieszkańcy obecnego województwa zachodniopomorskiego, lubuskiego czy zachodniopomorskiego mieli słabsze możliwości formalizacji działań.

2.3. Źródła danych, które mogą być wykorzystywane w analizach regionalnych dla Polski

Badania empiryczne, w których wykorzystywane są dane regionalne, opierają się na informacjach pochodzących z różnych baz. Informacje te mogą być wykorzystywane zarówno wtedy, gdy wszystkie kategorie dostępne są na poziomie mezo,

jak i wówczas gdy wiele zmiennych obserwowalnych jest na poziomie indywidualnym. W modelach tych przyjmuje się, że informacje obserwowalne na poziomie mezo determinują decyzje firm czy gospodarstw domowych. W wielu omawianych w niniejszej monografii badaniach empirycznych wykorzystywane są dane obserwowalne na poziomie mezo. Dlatego też niniejszy podrozdział służy prezentacji baz danych zawierających informacje o poszczególnych jednostkach administracyjnych. Omawiane są przede wszystkim te bazy, z których pochodzą dane wykorzystywane w przedstawionych w kolejnych rozdziałach badaniach empirycznych. Wymienione są także inne powszechnie znane źródła danych o sytuacji w regionach.

2.3.1. Bank Danych Lokalnych

Podstawowym źródłem danych, które mogą być wykorzystywane w analizach regionalnych dla Polski, jest Bank Danych Lokalnych (BDL) Głównego Urzędu Statystycznego. Jest on największą w Polsce bazą danych o gospodarce, społeczeństwie i środowisku. Oferuje ponad 40 tysięcy cech statystycznych pogrupowanych tematycznie. Pierwsze dane dostępne w Banku Danych Lokalnych pochodzą z 1995 roku. Aby móc skorzystać z danych BDL-u, należy odwiedzić jego stronę internetową: https://bdl.stat.gov.pl/BDL/start. Podstawowym podziałem danych regionalnych dostępnych na stronie Głównego Urzędu Statystycznego jest podział na dane według dziedzin oraz według jednostek administracyjnych.

W Banku Danych Lokalnych wyróżnia się następujące kategorie:

1. Ceny.
2. Finanse przedsiębiorstw.
3. Finanse publiczne.
4. Fundusze unijne.
5. Gospodarka mieszkaniowa i komunalna.
6. Handel i gastronomia.
7. Inwestycje i środki trwałe.
8. Kultura fizyczna, sport i rekreacja.
9. Kultura i sztuka.
10. Ludność.
11. Narodowe spisy powszechne.
12. Nauka i technika.
13. Ochrona zdrowia i opieka społeczna.
14. Organizacja państwa i wymiar sprawiedliwości.
15. Podmioty gospodarcze i przekształcenia własnościowe i strukturalne.
16. Podział terytorialny.

17. Powszechne spisy rolne.
18. Przemysł i budownictwo.
19. Rachunki regionalne.
20. Rolnictwo, leśnictwo i łowiectwo.
21. Rynek materiałowy i paliwowo-energetyczny.
22. Rynek pracy.
23. Samorząd terytorialny.
24. Sektor non-profit.
25. Stan i ochrona środowiska.
26. Szkolnictwo.
27. Szkolnictwo wyższe.
28. Transport i łączność.
29. Turystyka.
30. Wychowanie przedszkolne.
31. Wynagrodzenia i świadczenia społeczne.

Każda z wymienionych wyżej kategorii składa się z grup, które następnie dzielą się na podgrupy. Wewnątrz podgrupy rozróżnia się poszczególne zmienne. W niniejszym rozdziale omówione są kategorie i grupy, ponieważ liczba podgrup i zmiennych jest tak duża, że wymienienie ich wszystkich zdecydowanie zwiększyłoby objętość monografii. Tabele 5–8 zawierają informacje dotyczące kategorii oraz grup zmiennych dostępnych w Banku Danych Lokalnych.

Tabela 5. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część pierwsza

Kategorie	Ceny	Finanse przedsiębiorstw	Finanse publiczne	Fundusze unijne	Gospodarka mieszkaniowa i komunalna	Handel i gastronomia	Inwestycje i środki trwałe	Kultura fizyczna, sport i rekreacja
Grupy	1. Ceny w rolnictwie 2. Przeciętne ceny detaliczne towarów i usług konsumpcyjnych 3. Przeciętne ceny detaliczne towarów niekonsumpcyjnych 4. Przeciętne ceny producentów niektórych wyrobów spożywczych na rynku krajowym 5. Wskaźniki cen	1. Wyniki finansowe przedsiębiorstw 2. Dane archiwalne	1. Dochody budżetów gmin i miast na prawach powiatu 2. Dochody budżetów powiatów 3. Dochody budżetów województw 4. Dochody i wydatki budżetów jst – wskaźniki 5. Wydatki budżetów gmin i miast na prawach powiatu 6. Wydatki budżetów powiatów 7. Wydatki budżetów województw 8. Dane archiwalne	1. Fundusze unijne – prow. 2007–2013 2. Fundusze unijne – prow. 2014–2020 3. Fundusze unijne – systemy wsparcia bezpośredniego 4. NSRO podpisane umowy o dofinansowanie – 2007–2013 5. NSRO wartość projektów – 2007–2013 6. NSRO wnioski o dofinansowanie – 2007–2013 7. Umowy partnerstwa – umowy/decyzje o dofinansowanie – 2014–2020 8. Umowy partnerstwa – wartość projektów zakończonych (wydatki kwalifikowalne) – 2014–2020 9. Umowy partnerstwa – wartość umów/decyzji o dofinansowanie – 2014–2020 10. Umowy partnerstwa – wnioski o dofinansowanie – 2014–2020 11. Umowy w ramach programu „Po ryby” 2007–2013 12. Umowy w ramach programu „Po ryby” 2014–2020	1. Ciepłownictwo 2. Dodatki mieszkaniowe 3. Remonty komunalnych zasobów mieszkaniowych 4. Remonty zasobów mieszkaniowych według form własności 5. Ubytki w zasobach mieszkaniowych 6. Urządzenia sieciowe 7. Zasoby mieszkaniowe 8. Dane archiwalne	1. Placówki gastronomiczne 2. Sklepy i stacje paliw 3. Sprzedaż detaliczna 4. Sprzedaż hurtowa 5. Targowiska 6. Dane archiwalne	1. Nakłady inwestycyjne 2. Nakłady inwestycyjne i środki trwałe w przedsiębiorstwach wg PKD 2007 3. Rzeczowy majątek trwały 4. Źródła finansowania nakładów inwestycyjnych w przedsiębiorstwach 5. Dane archiwalne	1. Sport

Źródło: opracowanie własne na podstawie Banku Danych Lokalnych GUS.

Tabela 6. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część druga

Kategorie	Kultura i sztuka	Ludność	Narodowe spisy powszechne	Nauka i technika	Ochrona zdrowia i opieka społeczna	Organizacja państwa i wymiar sprawiedliwości	Podmioty gospodarcze i przekształcenia własnościowe i strukturalne	Podział terytorialny
Grupy	1. Biblioteki 2. Działalność centrów, domów, ośrodków kultury, klubów i świetlic 3. Działalność sceniczna i wystawiennicza 4. Kina 5. Muzea 6. Organizacja imprez masowych	1. Gospodarstwa domowe 2. Małżeństwa rozwody i separacje 3. Migracje wewnętrzne i zagraniczne 4. Prognozy 5. Stan ludności 6. Urodzenia i zgony 7. Dane archiwalne	1. NSP 1998 – Gospodarstwa domowe 2. NSP 1998 – Ludność 3. NSP 1998 – Mieszkania 4. NSP 2002 – Aktywność ekonomiczna ludności 5. NSP 2002 – Budynki 6. NSP 2002 – Gospodarstwa domowe i rodziny 7. NSP 2002 – Ludność 8. NSP 2002 – Mieszkania niezamieszkane 9. NSP 2002 – Mieszkania zamieszkane 10. NSP 2011 – Aktywność ekonomiczna ludności 11. NSP 2011 – Budynki 12. NSP 2011 – Dojazdy do pracy 13. NSP 2011 – Gospodarstwa domowe i rodziny 14. NSP 2011 – Ludność 15. NSP 2011 – Mieszkania zamieszkane	1. Biotechnologia 2. Działalność badawcza i rozwojowa 3. Działalność innowacyjna 4. Ochrona własności przemysłowej w Polsce 5. Społeczeństwo informacyjne 6. Dane archiwalne	1. Ambulatoryjna opieka zdrowotna 2. Apteki i punkty apteczne 3. Kadra medyczna 4. Lecznictwo uzdrowiskowe, stacjonarne zakłady rehabilitacji leczniczej 5. Opieka nad dziećmi i młodzieżą 6. Placówki stacjonarnej pomocy społecznej 7. Pomoc doraźna i ratownictwo medyczne 8. Stacjonarne zakłady opieki długoterminowej i paliatywno-hospicyjnej 9. Stan zdrowia ludności 10. Szpitale 11. Środowiskowa pomoc społeczna 12. Świadczenia rodzinne 13. Świadczenia z pomocy społecznej 14. Żłobki 15. Dane archiwalne	1. Organy przedstawicielskie 2. Przestępstwa stwierdzone przez policję w zakończonych postępowaniach przygotowawczych 3. Dane archiwalne	1. Nowo zarejestrowane w rejestrze regon podmioty gospodarki narodowej 2. Podmioty gospodarki narodowej – wskaźniki 3. Podmioty gospodarki narodowej wg rejestru REGON (dane kwartalne) 4. Podmioty gospodarki narodowej wpisane do rejestru REGON 5. Podmioty niefinansowe prowadzące działalność gospodarczą (pkd 2007) 6. Podmioty z udziałem kapitału zagranicznego 7. Wyrejestrowane z rejestru REGON podmioty gospodarki narodowej 8. Dane archiwalne	1. Podział administracyjny, sieć osadnicza 2. Powierzchnia geodezyjna kraju (dane GUGIK)

Źródło: opracowanie własne na podstawie Banku Danych Lokalnych GUS.

Tabela 7. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część trzecia

Kategorie	Powszechne spisy rolne	Przemysł i budownictwo	Rachunki regionalne	Rolnictwo, leśnictwo i łowiectwo	Rynek materiałowy i paliwowo-energetyczny	Rynek pracy	Samorząd terytorialny	Sektor non-profit
Grupy	1. PSR 1996 – aktywność ekonomiczna i źródła utrzymania w gospodarstwach 2. PSR 1996 – indywidualne gospodarstwa rolne 3. PSR 1996 – infrastruktura, wyposażenie techniczne i środki produkcji 4. PSR 1996 – powierzchnia, użytkowanie gruntów i zwierzęta gospodarskie 5. PSR 2002 – PSR wg siedziby gospodarstwa 6. PSR 2002 – PSR wg siedziby użytkownika	1. Budownictwo mieszkaniowe 2. Budynki 3. Pozwolenia wydane na budowę i zgłoszenia z projektem budowlanym (dane kwartalne) 4. Pozwolenia wydane na budowę i zgłoszenia z projektem budowlanym budowy obiektów budowlanych 5. Produkcja budowlano-montażowa 6. Produkcja sprzedana przemysłu i budownictwa 7. Dane archiwalne	1. Koszty związane z zatrudnieniem (ceny bieżące) – PKD 2007 – ESA 2010 2. Nadwyżka operacyjna brutto (ceny bieżące) – PKD 2007 – ESA 2010 3. Nominalne dochody w sektorze gospodarstw domowych – PKD 2007 – ESA 2010 4. Produkcja globalna (ceny bieżące) – PKD 2007 – ESA 2010 5. Produkt krajowy brutto (ceny bieżące) – PKD 2007 – ESA 2010 6. Produkt krajowy brutto (ceny bieżące) – PKD 2007 – ESA 2010 – szacunki wstępne 7. Produkt krajowy brutto (ceny stałe) – PKD 2007 – ESA 2010 8. Realne dochody w sektorze gospodarstw domowych – PKD 2007 – ESA 2010 9. Wartość dodana brutto (ceny bieżące) – PKD 2007 – ESA 2010 10. Wartość dodana brutto (ceny bieżące) – PKD 2007 – ESA 2010 – szacunki wstępne 11. Dane archiwalne	1. Gospodarcze wykorzystanie lasu 2. Gospodarstwa rolne 3. Lasy prywatne i gminne 4. Leśnictwo wszystkich form własności 5. Pogłowie zwierząt 6. Produkcja rolnicza 7. Produkcja zwierzęca 8. Skup produktów rolnych 9. Uprawy rolnicze 10. Użytkowanie gruntów 11. Zadrzewienia 12. Zagrożenie i ochrona środowiska leśnego 13. Zwierzęta łowne 14. Dane archiwalne	1. Rynek materiałowy 2. Dane archiwalne	1. Aktywność ekonomiczna ludności (dane kwartalne) 2. Aktywność ekonomiczna ludności (dane średnioroczne) 3. Bezrobocie rejestrowane 4. Miejsca pracy 5. Pracujący i zatrudnieni w przedsiębiorstwach o liczbie pracujących do 49 osób 6. Pracujący według innego podziału niż PKD 7. Warunki pracy 8. Dane archiwalne	1. Grunty komunalne 2. Organy dzielnic m. st. Warszawy 3. Organy gminy 4. Organy miast na prawach powiatu 5. Organy powiatu 6. Organy województwa 7. Planowanie przestrzenne 8. Samorząd terytorialny – wskaźnik 9. Dane archiwalne	1. Aktywne organizacje i stowarzyszenia

Źródło: opracowanie własne na podstawie Banku Danych Lokalnych GUS.

Tabela 8. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część czwarta

Kategorie	Stan i ochrona środowiska	Szkolnictwo	Szkolnictwo wyższe	Transport i łączność	Turystyka	Wychowanie przedszkolne	Wynagr. i świadczenia społeczne
Grupy	1. Efekty rzeczowe inwestycji ochrony środowiska i gospodarki wodnej oddane w roku sprawozdawczym 2. Ekonomiczne aspekty ochrony środowiska 3. Emisja zanieczyszczeń powietrza z zakładów szczególnie uciążliwych 4. Gospodarka wodno-ściekowa w przemyśle 5. Nakłady na środki trwałe służące ochronie środowiska i gospodarce wodnej wg kierunków inwestowania 6. Nakłady na środki trwałe służące ochronie środowiska i gospodarce wodnej wg źródeł finansowania 7. Nieczystości ciekłe 8. Ochrona powierzchni ziemi i gleby 9. Ochrona przyrody i różnorodności biologicznej 10. Oczyszczanie ścieków komunalnych 11. Odpady komunalne 12. Odpady wytworzone i dotychczas składowane (nagromadzone z wyłączeniem odpadów komunalnych) 13. Tereny zieleni 14. Zasoby eksploatacyjne wód podziemnych 15. Zużycie wody i oczyszczalnie ścieków	1. Edukacja dzieci i młodzieży ze specjalnymi potrzebami edukacyjnymi 2. Nauczanie języków obcych w szkołach dla dzieci i młodzieży 3. Ośrodki dla dzieci i młodzieży 4. Skolaryzacja 5. Szkolnictwo branżowe I stopnia 6. Szkolnictwo gimnazjalne 7. Szkolnictwo ogólnokształcące 8. Szkolnictwo podstawowe 9. Szkolnictwo policealne 10. Szkolnictwo ponadgimnazjalne zawodowe i artystyczne 11. Szkolnictwo zasadnicze zawodowe 12. Szkoły ponadgimnazjalne i policealne 13. Zdawalność egzaminów 14. Dane archiwalne	1. Nauczyciele akademiccy 2. Studenci i absolwenci 3. Studia podyplomowe i doktoranckie 4. Szkoły wyższe 5. Wskaźniki	1. Drogi publiczne 2. Drogi publiczne gminne 3. Drogi publiczne powiatowe 4. Działalność transportowa 5. Komunikacja miejska 6. Linie regularnej komunikacji autobusowej 7. Łączność 8. Pojazdy 9. Ścieżki rowerowe 10. Transport kolejowy 11. Transport lotniczy 12. Transport morski 13. Transport przybrzeżny 14. Wypadki drogowe 15. Dane archiwalne	1. Placówki gastronomiczne w turystycznych obiektach noclegowych 2. Stopień wykorzystania turystycznych obiektów noclegowych 3. Turystyczne obiekty noclegowe (dane miesięczne) 4. Turystyczne obiekty noclegowe (stan w dniu 31 lipca) 5. Turystyczne obiekty noclegowe i ich wykorzystanie 6. Wyposażenie turystycznych obiektów noclegowych	1. Nauczyciele 2. Przedszkola 3. Punkty przedszkolne 4. Zespoły wychowania przedszkolnego	1. Świadczenia społeczne 2. Wynagrodzenia 3. Dane archiwalne

Źródło: opracowanie własne na podstawie Banku Danych Lokalnych GUS.

2.3.2. Regional Innovation Scoreboard jako źródło informacji o poziomie innowacyjności regionów

European Innovation Scoreboard (EIS) wykorzystywany jest w celu oceny jakości narodowych systemów innowacji w krajach Unii Europejskiej. Innowacyjność poszczególnych krajów jest mierzona za pomocą wskaźnika kompozytowego, który obliczany jest na podstawie 27 współczynników. Współczynniki te należą do jednej z czterech grup (warunki ramowe, inwestycje, aktywności innowacyjne, oddziaływanie).

Regional Innovation Scoreboard (RIS) powstał na bazie European Innovation Scoreboard (EIS). Jego celem jest ocena regionalnych systemów innowacji w regionach Unii Europejskiej. Niektóre wskaźniki pochodzące z EIS nie są dostępne na poziomie regionów. Dlatego też RIS obliczany jest na podstawie nie 27, lecz 18 wskaźników wykorzystywanych przy konstrukcji EIS. W przypadku niektórych zmiennych zastosowane zostały nieco inne definicje. Tabela 9 prezentuje grupy, mierniki znajdujące się wewnątrz konkretnych grup, a także informacje o różnicach między zmiennymi wykorzystywanymi do mierzenia innowacyjności na poziomie krajowym oraz na poziomie regionalnym. Informacje dotyczą edycji z 2017 roku.

Tabela 9. Zmienne wykorzystywane do mierzenia innowacyjności na poziomie regionalnym dostępne w Regional Innovation Scoreboard 2017

Warunki ramowe		
Temat	Zmienna EIS 2017	Zmienna RIS 2017
Kapitał ludzki	Absolwenci studiów doktoranckich w wieku 25–34 lata na 1000 mieszkańców	Brak danych regionalnych
	Udział osób w wieku 25–34 lata posiadających wyższe wykształcenie	Udział osób w wieku 30–34 lata posiadających wyższe wykształcenie
	Udział osób w wieku 25–64 lata, uczących się lub przechodzących szkolenia mające na celu poprawę ich wiedzy, umiejętności i kompetencji	Tak samo jak w EIS 2017
Atrakcyjność systemu badań	Relacja liczby artykułów pisanych we współpracy międzynarodowej do liczby ludności	Tak samo jak w EIS 2017
	Relacja liczby artykułów opublikowanych w czasopismach należących do 10% najlepiej cytowanych periodyków do liczby wszystkich opublikowanych artykułów	Tak samo jak w EIS 2017
	Procentowy udział zagranicznych doktorantów we wszystkich doktorantach	Brak danych regionalnych

Tabela 9 (cd.)

Przyjazność otoczenia dla innowacji	Udział przedsiębiorstw mających dostęp do szybkiego internetu (szybkość przesyłu wynosi 100 Mb/s lub jest wyższa)	Brak danych regionalnych
	Udział osób w wieku 18–64 lata, którzy są w momencie zakładania firmy albo rozpoczęli działalność w ciągu ostatnich 42 miesięcy	Brak danych regionalnych
Inwestycje		
Temat	Zmienna EIS 2017	Zmienna RIS 2017
Finansowanie i wsparcie	Relacja wydatków na badania i rozwój w sektorze publicznym do PKB	Tak samo jak w EIS 2017
	Relacja wydatków na kapitał podwyższonego ryzyka do PKB	Brak danych regionalnych
Inwestycje firm	Relacja wydatków na badania i rozwój w sektorze prywatnym do PKB	Tak samo jak w EIS 2017
	Relacja wydatków innych niż wydatki na badania i rozwój do całkowitych obrotów	Dostępne tylko dla małych i średnich przedsiębiorstw
	Liczba przedsiębiorstw organizujących szkolenia mające na celu poprawienie umiejętności informatycznych wśród personelu	Brak danych regionalnych
Aktywności innowacyjne		
Temat	Zmienna EIS 2017	Zmienna RIS 2017
Wprowadzający innowacje	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje produktowe lub procesowe	Tak samo jak w EIS 2017
	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje marketingowe lub organizacyjne	Tak samo jak w EIS 2017
	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje wewnątrz swojej firmy	Tak samo jak w EIS 2017
Współpraca w zakresie innowacji	Odsetek innowacyjnych małych i średnich przedsiębiorstw współpracujących z otoczeniem	Tak samo jak w EIS 2017
	Relacja publikacji powstałych we współpracy nauki z biznesem do populacji	Tak samo jak w EIS 2017
	Udział współfinansowania publicznych wydatków na badania i rozwój ze strony sektora prywatnego	Brak danych regionalnych
Aktywa intelektualne	Wartość Patent Cooperation Treaty aplikacji patentowych w relacji do PKB	Projekty patentowe EPO
	Wartość zgłoszeń o znaki towarowe w relacji do PKB	Zgłoszenia europejskich znaków towarowych
	Wartość zgłoszeń indywidualnych projektów w relacji do PKB	Zgłoszenia projektów

Oddziaływanie		
Temat	Zmienna EIS 2017	Zmienna RIS 2017
Wpływ na zatrudnienie	Odsetek osób zatrudnionych w przemysłach i usługach intensywnie wykorzystujących wiedzę	Zatrudnienie w przemysłach wysokiej i średniowysokiej technologii oraz usługach intensywnie wykorzystujących wiedzę
	Zatrudnienie w szybko rosnących firmach innowacyjnych sekcji	Brak danych regionalnych
Wpływ na sprzedaż	Relacja eksportu produktów wysokich i średniowysokich technologii do całkowitego eksportu	Eksport produktów wytworzonych przez przemysły wysokiej i średniowysokiej technologii
	Udział eksportu usług intensywnie wykorzystujących wiedzę w całkowitym eksporcie usług	Brak danych regionalnych
	Relacja sprzedaży innowacji nowych dla firmy albo nowych dla rynku w całkowitym obrocie	Dostępne tylko dla małych i średnich przedsiębiorstw

Źródło: opracowanie własne na podstawie RIS, 2017.

Ostatnia (rok 2018) edycja Regional Innovation Scoreboard pochodzi z 2017 roku i jest ósmą z kolei. Wcześniejsze edycje RIS dotyczą lat 2009, 2010, 2011, 2012, 2013, 2014, 2016. Regional Innovation Scoreboard obejmuje swoim zasięgiem 220 regionów z 22 krajów Unii Europejskiej, a także Norwegię, Serbię i Szwajcarię na różnych poziomach klasyfikacji regionalnej NUTS. Dostępność danych umożliwia uwzględnienie w ramach Regional Innovation Scoreboard 28 regionów na poziomie NUTS1 oraz 192 regionów na poziomie NUTS2. W przypadku Wielkiej Brytanii, Francji, Austrii, Bułgarii oraz Belgii miernik innowacyjności regionalnej obliczany jest na poziomie NUTS1. W przypadku Czech, Danii, Niemiec, Irlandii, Grecji, Hiszpanii, Chorwacji, Włoch, Holandii, Polski, Portugalii, Rumunii, Słowenii, Słowacji, Finlandii, Szwecji, Wielkiej Brytanii, Norwegii, Szwajcarii i Serbii innowacyjność mierzona jest na poziomie NUTS2.

W przypadku ośmiu wskaźników, takich jak:

1) udział ludności w wieku 30–34 lata posiadających wyższe wykształcenie,
2) umiejętność długotrwałej nauki,
3) wydatki na badania i rozwój w sektorze publicznym,
4) wydatki na badania i rozwój w sektorze przedsiębiorstw,
5) wnioski patentowe EPO,
6) wnioski o znaki towarowe,

7) zgłoszone projekty,
8) zatrudnienie w przemyśle wysokich technologii lub usługach opartych na wiedzy,

RIS wykorzystuje dane z Eurostatu. Jeśli chodzi o eksport produktów przemysłów średniowysokiej i wysokiej technologii, jego wartości dla poszczególnych regionów zostały oszacowane na podstawie opracowania pt. *Identifying revealed comparative advantages in an EU regional context*, przygotowanego przez Instytut Badań Ekonomicznych Dolnej Saksonii, Wiedeński Instytut Międzynarodowych Studiów Ekonomicznych oraz Centrum Europejskich Badań Ekonomicznych (ZEW). W przypadku sześciu wskaźników, związanych z danymi pochodzącymi z badania Community Innovation Survey, dane regionalne nie pochodzą z bazy Eurostatu. Autorzy RIS 2017 uzyskali odpowiednie zmienne, wykorzystując bazy danych krajowych urzędów statystycznych. Jeśli chodzi o mierniki związane z danymi bibliometrycznymi, autorzy raportu uzyskali je dzięki współpracy z Leiden University.

Dane dotyczące innowacyjności i pochodzące z badania Community Innovation Survey (CIS) dostępne są dla 22 krajów (Austria, Belgia, Bułgaria, Chorwacja, Czechy, Dania, Finlandia, Francja, Niemcy, Grecja, Węgry, Włochy, Norwegia, Polska, Portugalia, Rumunia, Serbia, Słowacja, Słowenia, Hiszpania, Szwajcaria, Wielka Brytania). W przypadku krajów, dla których NUTS1 i NUTS2 pokrywają się z całkowitym terytorium albo też krajów, dla których próba badania CIS jest za mała, aby móc analizować dane regionalne, nie obliczono wartości odpowiednich zmiennych. Należy zwrócić uwagę na niedoskonałość danych pochodzących z badania CIS do mierzenia innowacyjności regionalnej. Wynika to z faktu, że w przypadku innowacyjnych przedsiębiorstw, mających swoje siedziby w różnych regionach, aktywność innowacyjna przyporządkowywana jest do regionu, w którym znajduje się główna siedziba firmy. Dlatego też pojawia się ryzyko, że w regionach obejmujących stolice krajów poziom innowacyjności jest przeszacowany ze względu na fakt, że firmy wprowadzające działania innowacyjne w różnych regionach są klasyfikowane jako reprezentanci regionów posiadających swoje centrale. Dlatego też w Regional Innovation Scoreboard obliczana jest innowacyjność w grupie firm małych i średnich.

W przypadku tylko jednego wskaźnika pochodzącego z RIS2017 (relacja liczby artykułów pisanych we współpracy międzynarodowej do liczby ludności) dane dotyczą 2016 roku. W przypadku pięciu mierników (udział osób w wieku 30–34 lata posiadających wyższe wykształcenie; udział osób w wieku 25–64 lata, uczących się lub przechodzących szkolenia mające na celu poprawę ich wiedzy, umiejętności i kompetencji; relacja wydatków na badania i rozwój w sektorze publicznym do PKB; relacja publikacji powstałych we współpracy nauki z biznesem do populacji; poziom zatrudnienia w przemysłach średniowysokiej i wysokiej technologii

lub usługach opartych na wiedzy) dane pochodzą z 2015 roku, a w przypadku kolejnych dziewięciu mierników (relacja liczby artykułów opublikowanych w czasopismach należących do 10% najlepiej cytowanych periodyków do liczby wszystkich opublikowanych artykułów; relacja wydatków innych niż wydatki na badania i rozwój do całkowitych obrotów; odsetek małych i średnich przedsiębiorstw wprowadzających innowacje produktowe lub procesowe; odsetek małych i średnich przedsiębiorstw wprowadzających innowacje wewnątrz swojej firmy; odsetek małych i średnich przedsiębiorstw wprowadzających innowacje marketingowe lub organizacyjne; odsetek innowacyjnych małych i średnich przedsiębiorstw współpracujących z innymi; wartość zgłoszeń znaków towarowych w relacji do PKB; wartość zgłoszeń indywidualnych projektów w relacji do PKB, a także relacja sprzedaży innowacji nowych dla firmy albo nowych dla rynku w całkowitym obrocie) dane pochodzą z 2014 roku. Dane z 2011 roku wykorzystywane są w przypadku dwóch mierników (zgłoszenia patentowe EPO oraz eksport produktów wytworzonych przez przemysły wysokiej i średniowysokiej technologii).

Jedno z badań empirycznych rozważanych w niniejszej monografii dotyczy innowacyjności przedsiębiorstw. Jego celem jest sprawdzenie roli czynników regionalnych w wyjaśnieniu decyzji dotyczących wprowadzania innowacyjnych rozwiązań. Ponieważ dane dotyczące kategorii omówionych w tabeli 9 są dostępne dla polskich województw, możliwe jest obliczenie indeksów innowacyjności dla tych jednostek administracyjnych. Poziomy innowacyjności regionów można uzyskać, korzystając z opracowań RIS 2017, RIS 2015 itd. Należy jednak pamiętać, że na przykład wartości, które można znaleźć w opracowaniu RIS 2017, informują o poziomie innowacyjności obserwowanym w latach poprzedzających 2017 rok. Analogicznie należy interpretować wartości dostępne w poprzednich edycjach Regional Innovation Scoreboard. Jak widać zatem, Regional Innovation Scoreboard jest cennym źródłem informacji o innowacyjności polskich regionów. Dlatego też wydaje się, że uzasadnione jest wykorzystanie tej bazy w celu skonstruowania zmiennej wyjaśniającej skłonność do wprowadzania innowacji w polskich firmach. Wynika to z faktu, że otoczenie zewnętrzne firmy może odgrywać istotną rolę przy podejmowaniu decyzji dotyczących wprowadzania innowacji (por. m.in. Sternberg, Arndt, 2001; Kosała, Wach, 2011; Golejewska, 2018).

2.3.3. Inne źródła danych regionalnych wykorzystywane w badaniach empirycznych

Chociaż Bank Danych Lokalnych Głównego Urzędu Statystycznego oraz Regional Innovation Scoreboard są ważnymi źródłami informacji na temat sytuacji ekonomiczno-społeczno-demograficznej i środowiskowej regionów oraz jakości

regionalnych systemów innowacji, istnieją bazy zawierające dane regionalne, które są niedostępne w wyżej wymienionych. W związku z tym warto krótko omówić alternatywne źródła informacji o sytuacji w poszczególnych regionach.

Ważnym źródłem informacji dotyczących przestrzennego zróżnicowania wyników głosowania w wyborach samorządowych, parlamentarnych, prezydenckich czy do Parlamentu Europejskiego jest strona internetowa Państwowej Komisji Wyborczej. Zawarte są na niej informacje dotyczące frekwencji wyborczej, liczby głosów oddanych na określonych kandydatów, jak również poziomu poparcia dla poszczególnych komitetów wyborczych. Z punktu widzenia przestrzennego zróżnicowania wyników głosowania dostępność danych umożliwia analizę różnic nie tylko ze względu na województwa, powiaty czy gminy. Dane są osiągalne na poziomie obwodowych komisji wyborczych.

Ważnym źródłem informacji dotyczącej zdolności poszczególnych jednostek administracyjnych do pozyskiwania środków z Unii Europejskiej jest strona internetowa www.mapadotacji.gov.pl. Dzięki wykorzystaniu danych pochodzących z niej analizowane jest przestrzenne zróżnicowanie wysokości dotacji. Możliwa jest także analiza wysokości dofinansowania w określonych tematach (nauka i edukacja, energetyka, transport, turystyka, bezpieczeństwo, badania, rozwój, innowacje, kultura i sztuka, ochrona zdrowia, rozwój firm, telekomunikacja i e-usługi, praca i integracja społeczna, współpraca międzynarodowa, administracja, rewitalizacja lub ochrona środowiska). Projekty wyszukuje się także po nazwie beneficjenta, tytule projektu albo latach, na które został on przyznany (np. 2004–2006, 2007–2013 lub 2014–2020).

Kolejnym źródłem informacji, które warto wymienić, jest strona internetowa www.polskawliczbach.pl. Dzięki skorzystaniu z omawianego źródła możliwe jest uzyskanie informacji na temat przestrzennego zróżnicowania sytuacji demograficznej, tendencji zachodzących na rynku pracy, poziomu przestępczości, sytuacji w edukacji, szkolnictwie i kulturze oraz stanu i zmian w nieruchomościach. Możliwa jest analiza rankingów ze względu na określone kategorie ekonomiczne, jak również zasięgnięcie informacji dotyczącej wartości określonych zmiennych dla poszczególnych województw, powiatów, gmin czy miast.

Instytut Statystyki Kościoła Katolickiego jest pierwszym w Polsce ośrodkiem badań nad religijnością. Uzyskanie danych zebranych przez ten instytut wymaga wejścia na stronę internetową iskk.pl. Informacje na temat określonych zmiennych można otrzymać jednak na poziomie diecezji. W Polsce oprócz ordynariatu polowego wyróżnia się 41 diecezji. Podział na diecezje nie pokrywa się jednak z podziałem administracyjnym kraju na województwa, powiaty i gminy. Dlatego też nie jest możliwe poszukiwanie związków korelacyjnych między zmiennymi obserwowanymi na poziomie diecezji a innymi kategoriami dostępnymi na poziomie

jednostek administracyjnych. Niemniej jednak właściwe rozumienie prawidłowości zachodzących w określonych regionach Polski wymaga wiedzy na temat postaw religijnych mieszkańców poszczególnych obszarów. A zatem korzystanie z danych pochodzących ze strony internetowej Instytutu Statystyki Kościoła Katolickiego powinno pomóc socjologom w zrozumieniu zróżnicowania przestrzennego określonych zachowań i postaw społecznych.

Podstawowe informacje, jakie można uzyskać na omawianej strony internetowej, dotyczą praktyk niedzielnych. Wskaźniki „dominicantes" oraz „comunicantes" są dostępne na poziomie diecezji dla okresu od 1980 do 2016 roku. Informują one odpowiednio o udziale katolików uczęszczających co tydzień na mszę świętą oraz odsetku chrześcijan przyjmujących komunię w każdą niedzielę w całej Polsce oraz w poszczególnych diecezjach. Dzięki ISKK dostępne są także informacje dotyczące liczby sakramentów świętych w relacji do liczby mieszkańców.

3. Liniowe modele wielopoziomowe

3.1. Wprowadzenie

W rozdziale pierwszym omówione zostały podstawowe modele zmiennych jakościowych, w których dane obserwowane są na poziomie jednostek (np. firm, pracowników, gospodarstw domowych). Często jednak oprócz informacji dotyczących cech indywidualnych poszczególnych jednostek badania dostępne są również dane na temat lokalizacji firmy czy gospodarstwa domowego, a także numeru PKD związanego z działalnością firmy. Okazuje się, że wartości kategorii ekonomicznych dostępnych na poziomie mezo czy na poziomie sekcji mogą być istotnymi determinantami dla decyzji podejmowanych przez poszczególne jednostki. Na przykład decyzja dotycząca uczestnictwa w podaży pracy nierejestrowanej może być warunkowana sytuacją na rynku pracy w miejscowości zamieszkiwanej przez respondenta. Uwzględnienie odpowiedniej zmiennej w modelu dwumianowym sprawia, że oprócz danych, które są unikatowe dla wszystkich jednostek, mamy zmienne przyjmujące tę samą wartość dla osób zamieszkujących ten sam obszar.

Wprowadzenie zmiennych obserwowalnych na poziomie mezo w modelu ekonometrycznym opartym na danych jednostkowych nie wymaga jednak zastosowania niestandardowych metod estymacji. Dopiero gdy uwzględnione zostaną efekty losowe związane z poszczególnymi grupami (np. gminami, powiatami, sekcjami), konieczne jest zastosowanie nieklasycznych metod estymacji i niestandardowych modeli. Modelowanie wielopoziomowe polega między innymi na uwzględnianiu efektów losowych związanych z poszczególnymi grupami w modelu opartym na danych indywidualnych.

W naukach społecznych często mamy do czynienia z hierarchicznie ustrukturyzowanymi danymi. Różne czynniki oddziałują na wartości tak zwanej zmiennej wynikowej. Celem badacza jest określenie wpływu poszczególnych czynników na wartości wynikowe. Ideą modelowania wielopoziomowego jest jednoczesna analiza powiązań międzygrupowych i wewnątrzgrupowych.

W badaniach ekonomicznych często analizowane są powiązania na podstawie danych zagregowanych (np. dla powiatów, województw, państw). Jeśli jednak w badaniu ankietowym dysponujemy danymi indywidualnymi dla firm, gospodarstw domowych, respondentów, możliwa jest analiza powiązań na podstawie danych indywidualnych (zdezagregowanych). Modelowanie wielopoziomowe jest rozwiązaniem umożliwiającym połączenie wnioskowania na poziomie mikro i makro.

Niniejszy rozdział poświęcony jest modelom wielopoziomowym dla ciągłych zmiennych zależnych. W podrozdziale 3.2 omawiane są sposoby uwzględniania zmiennych regionalnych oraz związanych z przynależnością do sekcji w klasycznych modelach ekonometrycznych. W podrozdziale 3.3 opisany jest stan wiedzy z zakresu podstawowego wielopoziomowego modelu regresji, a także prezentowane są metody estymacji parametrów oraz oceny jakości dopasowania. Modele wielopoziomowe z krzyżowymi efektami losowymi omawiane są w podrozdziale 3.4. Zastosowanie modelu wielopoziomowego do testowania hipotez dotyczących zmian zachodzących na polskim rynku pracy prezentowane jest w podrozdziale 3.5. W tym celu wykorzystywane są dane pochodzące z badania struktury wynagrodzeń.

3.2. Zmienne regionalne i sekcyjne w modelach ekonometrycznych opartych na danych indywidualnych

W badaniach opartych na danych indywidualnych często pojawiają się informacje na temat lokalizacji firmy czy gospodarstwa domowego. Jeśli dane te dotyczą firm lub pracowników zatrudnionych w określonych przedsiębiorstwach, dostępne są informacje na temat przynależności tego podmiotu gospodarczego do grupy czy sekcji PKD.

Rozważmy przypadek, w którym przedmiotem zainteresowania badacza jest kształtowanie się ciągłej zmiennej y_i, zależnej od regresorów tworzących wektor $\boldsymbol{x}_i$. Wartości zmiennych (y_i oraz $\boldsymbol{x}_i$) obserwowane są na poziomie firm. Załóżmy, że badacz dysponuje dodatkową informacją dotyczącą lokalizacji (ze względu na województwa) firm oraz ich przynależności do sekcji PKD. Załóżmy także, że istnieje przypuszczenie, iż lokalizacja oraz przynależność do określonej sekcji PKD mają wpływ na zmienną zależną. Wówczas odpowiedni model regresji, uwzględniający zarówno wpływ lokalizacji, jak i przynależności do sekcji, przyjmuje następującą postać:

$$yi = \boldsymbol{xi\beta} + \boldsymbol{woji\rho} + \boldsymbol{seki}\ddot{\boldsymbol{\rho}} + \varepsilon i, \tag{3.1}$$

gdzie $\boldsymbol{woj}_i$ składa się z piętnastu zmiennych binarnych[1] przyjmujących wartość 1 w zależności od przynależności firmy do określonego województwa, natomiast $\boldsymbol{sek}_i$ zawiera zmienne binarne określające przynależność przedsiębiorstwa do sekcji PKD.

Oszacowania parametrów $\boldsymbol{\rho}$ oraz $\ddot{\boldsymbol{\rho}}$ należy interpretować jako różnice w wartościach zmiennej y_i wynikające z przynależności firmy odpowiednio do określonego województwa czy sekcji, przy innych czynnikach niezmienionych. Należy jednak pamiętać, że oszacowane różnice mają charakter ustalony. Przyjmowane jest zatem założenie, że różnice między wartościami zmiennej zależnej dla firm położonych w tym samym województwie i należących do tych samych sekcji PKD wynikają tylko z wartości zmiennych objaśniających. Niewyjaśnione różnice mają charakter losowy.

W modelu (3.1) przyjmowane jest założenie, że wpływ poszczególnych zmiennych wchodzących w skład wektora $\boldsymbol{x}_i$ na zmienną zależną jest taki sam we wszystkich województwach oraz wszystkich sekcjach. Możliwa jest jednak taka parametryzacja, która umożliwia oszacowanie parametrów ilustrujących zależność między zmienną y_i a zmiennymi $\boldsymbol{x}_i$ oddzielnie dla każdego województwa i sekcji. Odpowiedni model przyjmuje postać:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \sum_w woj_{iw}\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{woj,w}\tilde{\boldsymbol{\beta}}^{woj,w} + \sum_s sek_{is}\tilde{\boldsymbol{x}}_{\boldsymbol{i}}^{sek,s}\tilde{\boldsymbol{\beta}}^{sek,s} + \boldsymbol{woj}_i\varrho + \boldsymbol{sek}_i\boldsymbol{\theta} + \varepsilon_i \,. \quad (3.2)$$

Wektor $\boldsymbol{\beta}$ ilustruje wpływ regresorów na regresanta w referencyjnym województwie i sekcji. Zależność między zmienną zależną a zmiennymi objaśniającymi dla firm z sekcji *s*-tej zlokalizowanych w *w*-tym województwie ilustruje wektor $\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}^{woj,w} + \tilde{\boldsymbol{\beta}}^{sek,s}$.

Do modelu opartego na danych indywidualnych możliwe jest włączenie innych kategorii różniących się ze względu na województwa lub sektory. Wówczas model regresji przyjmuje postać:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{ww}_{iw}\,\ddot{\boldsymbol{\psi}} + \boldsymbol{vv}_{is}\,\ddot{\boldsymbol{\omega}} + \varepsilon_i \,. \quad (3.3)$$

$\boldsymbol{ww}_{iw}$ jest wektorem obserwacji na zmiennych różniących się w województwach. *w* indeksuje województwa, natomiast przedsiębiorstwa zlokalizowane w tej samej jednostce administracyjnej charakteryzują się tymi samymi wartościami poszczególnych kategorii. Analogiczna sytuacja dotyczy zmiennych wchodzących w skład wektora $\boldsymbol{vv}_{is}$, przyjmujących te same wartości dla wszystkich firm reprezentujących daną sekcję.

1 W celu uniknięcia współliniowości pomija się zmienną binarną związaną z referencyjnym województwem czy sekcją.

3.3. Podstawowy wielopoziomowy model regresji. Estymacja parametrów i predykcja efektów losowych

Załóżmy, że $i = 1, 2, \ldots, I$ indeksuje poszczególne jednostki. Przyjmijmy, że są to firmy. Niech $j1 = 1, 2, \ldots, J1$, $j2 = 1, \ldots, J2$ oraz $j3 = 1, \ldots, J3$ indeksują odpowiednio gminy, powiaty oraz województwa, w których przedsiębiorstwa te są zlokalizowane. Jeśli $y_{i,j1,j2,j3}$ oznacza wynik i-tej firmy zlokalizowanej w $j1$-tej gminie, $j2$-tym powiecie i $j3$-cim województwie, wówczas wynik ten może zostać zdekomponowany następująco (por. Goldstein, 1986):

$$y_{i,j1,j2,j3} = \ddot{\alpha}^{*}_{i,j1,j2,j3} + \ddot{\beta}^{*}_{j1,j2,j3} + \ddot{\gamma}^{*}_{j2,j3} + \ddot{\pi}^{*}_{j3}. \tag{3.4}$$

Na każdym poziomie hierarchii formułowany jest model uzależniający elementy równania (3.4) od zmiennych objaśniających. Wówczas na poziomie województw mamy:

$$\pi^{*}_{j3} = \sum_{k=0}^{K1} \pi_k \, w3_{k,j3} + u3_{j3}, \tag{3.5}$$

gdzie $u3_{j3}$ jest zmienną losową, dla której $E\left(u3_{j3}\right) = 0$, $var\left(u3_{j3}\right) = \sigma^2_{u3}$, natomiast π^{*}_{k} jest współczynnikiem na poziomie województwa przy k-tej zmiennej objaśniającej $w3_{k,j3}$ dla $j3$-ego województwa. Przyjmując, że $w_{0,j3} = 1$, $K1$ jest liczbą zmiennych wpływających na wynik obserwowany na poziomie województw. Na poziomie powiatów, mamy:

$$\ddot{\gamma}^{*}_{j2,j3} = \sum_{k=0}^{K2} \ddot{\gamma}^{*}_{k,j3} w2_{k,j2,j3} + u2_{j2,j3}, \tag{3.6}$$

gdzie $u2_{j2,j3}$ jest zmienną losową, dla której $E\left(u2_{j2,j3}\right) = 0$ oraz $var\left(u2_{j2,j3}\right) = \sigma^2_{u2}\left(j3\right)$, natomiast $\ddot{\gamma}^{*}_{k,j3}$ jest współczynnikiem na poziomie powiatu przy k-tej zmiennej objaśniającej $w2_{k,j2,j3}$ dla $j2$-ego powiatu zlokalizowanego w $j3$-cim województwie. Przyjmując, że $w2_{0,j2,j3} = 1$, $K2$ jest liczbą zmiennych wpływających na wynik obserwowany na poziomie powiatów. Następnie na poziomie gmin mamy:

$$\ddot{\beta}^{*}_{j1,j2,j3} = \sum_{k=0}^{K3} \ddot{\beta}^{*}_{k,j2,j3} w1_{k,j1,j2,j3} + u1_{j1,j2,j3}, \tag{3.7}$$

gdzie $u1_{j1,j2,j3}$ jest zmienną losową, dla której $E\left(u1_{j1,j2,j3}\right)=0$ oraz $var\left(u1_{j1,j2,j3}\right)=\sigma_{u1}^2\left(j2,j3\right)$, natomiast $\beta_{k,j2,j3}$ jest współczynnikiem przy *k*-tej zmiennej objaśniającej $w1_{k,j1,j2,j3}$ dla *j*1-ej gminy zlokalizowanej w *j*2-im powiecie i *j*3-cim województwie. Przyjmując, że $w1_{0,j1,j2,j3}=1$, $K3$ jest liczbą zmiennych wpływających na wynik obserwowany na poziomie gmin. Ostatecznie na poziomie firm (jednostek) mamy:

$$\ddot{\alpha}_{i,j1,j2,j3}^{*}=\sum_{k=0}^{K4}\ddot{\alpha}_{k,j1,j2,j3}^{*}w0_{k,i,j1,j2,j3}+u0_{i,j1,j2,j3}, \tag{3.8}$$

gdzie w odniesieniu do zmiennej losowej $u0_{i,j1,j2,j3}$ zakłada się, że ma ona zerową wartość oczekiwaną oraz wariancję równą $var\left(u0_{i,j1,j2,j3}\right)=\sigma_{u0}^2\left(j1,j2,j3\right)$. $\ddot{\alpha}_{k,j1,j2,j3}^{*}$ jest współczynnikiem przy *k*-tej zmiennej objaśniającej $w0_{k,i,j1,j2,j3}$ dla *i*-tej firmy zlokalizowanej w *j*1-ej gminie, *j*2-im powiecie oraz *j*3-cim województwie. Przyjmuje się, że $w0_{0,i,j1,j2,j3}=1$, $K4$ jest liczbą zmiennych wpływających na wynik obserwowany na poziomie indywidualnym.

Zakładając, że parametry $\ddot{\alpha},\ddot{\beta},\ddot{\gamma},\ddot{\pi}$ są niezależne od firm i jednostek administracyjnych, a następnie zestawiając ze sobą równania (3.4)–(3.8), uzyskujemy:

$$\begin{aligned}y_{i,j1,j2,j3}=\ddot{\alpha}_0+\ddot{\beta}_0+\ddot{\gamma}_0+\ddot{\pi}_0+\sum_{k=1}^{K4}\ddot{\alpha}_{k,j1,j2,j3}w0_{k,i,j1,j2,j3}+\\+\sum_{k=1}^{K3}\ddot{\beta}_{k,j2,j3}w1_{k,j1,j2,j3}+\sum_{k=1}^{K2}\ddot{\gamma}_{q,m}w2_{k,j2,j3}+\\+\sum_{k=1}^{K1}\ddot{\pi}_k w3_{k,j3}+\left(u3_{j3}+u2_{j2,j3}+u1_{j1,j2,j3}+u0_{i,j1,j2,j3}\right)\end{aligned} \tag{3.9}$$

Przy założeniu, że składniki losowe $u3_{j3}$, $u2_{j2,j3}$, $u1_{j1,j2,j3}$ oraz $u0_{i,j1,j2,j3}$ są ze sobą parami nieskorelowane, wariancja zmiennej wynikowej przedstawia się następująco:

$$var\left(y_{i,j1,j2,j3}\right)=\sigma_{u3}^2+\sigma_{u2}^2\left(j3\right)+\sigma_{u1}^2\left(j2,j3\right)+\sigma_{u0}^2\left(j1,j2,j3\right). \tag{3.10}$$

Jak widać zatem na podstawie wzoru (3.10), całkowita wariancja zmiennej wynikowej *y* może zostać zdekomponowana na:

1) zróżnicowanie między firmami,
2) zróżnicowanie między gminami,
3) zróżnicowanie między powiatami,
4) zróżnicowanie między województwami.

Jak pokazują między innymi Harvey Goldstein (1986) oraz Edyta Łaszkiewicz (2016), wykorzystując notację ogólną, model (3.9) może zostać alternatywnie zapisany jako następujący liniowy model mieszany:

$$\boldsymbol{y} = \boldsymbol{X}_{[1]}\boldsymbol{\beta}_{[1]} + \boldsymbol{X}_{[2]}\boldsymbol{\beta}_{[2]} + \left[\boldsymbol{X}_{[3]}^{(1)} \middle| \boldsymbol{X}_{[3]}^{(2)} \middle| \ldots \middle| \boldsymbol{X}_{[3]}^{(J)} \right] \boldsymbol{u}_1 + \mathbf{M}\mathbf{A}\boldsymbol{u}_2 + \boldsymbol{\varepsilon}\,, \tag{3.11}$$

gdzie $\boldsymbol{y}$ jest $I\times1$-wymiarowym wektorem obserwacji na zmiennej zależnej, $\boldsymbol{X}_{[1]}$ jest pełną macierzą obserwacji na zmiennych objaśniających różniących się ze względu na jednostki o wymiarze $I \times \tilde{K}$, natomiast $\beta_{[1]}$ jest wektorem parametrów mierzących wpływ poszczególnych zmiennych dostępnych na poziomie jednostkowym o wymiarze $\tilde{K} \times 1$. Macierz $\boldsymbol{X}_{[2]}$ ma wymiary $I \times \hat{K}$ i składa się ze zmiennych kontekstowych przyjmujących te same wartości dla wszystkich firm zlokalizowanych w tej samej jednostce administracyjnej. Poszczególne kolumny tej macierzy mogą dotyczyć zmiennych kontekstowych obserwowalnych na poziomie województw, powiatów oraz gmin. Macierz $\boldsymbol{\beta}_{[2]}$ o wymiarze $\hat{K} \times 1$ składa się z parametrów ustalonych przy zmiennych kontekstowych. $\boldsymbol{X}_{[3]}$ jest podmacierzą macierzy $\boldsymbol{X}_{[1]}$ o wymiarze $I \times \tilde{K}$. Zawiera ona te zmienne, których wpływ na regresanta losowo różni się między jednostkami administracyjnymi. *i*-ty wiersz macierzy $\boldsymbol{X}_{[3]}^{(j)}$ odpowiada *i*-temu wierszowi macierzy $\boldsymbol{X}_{[3]}$, jeśli *i*-ta firma zlokalizowana jest w *j*-tej jednostce administracyjnej oraz wektorowi zerowemu w przeciwnym przypadku. Macierz $\boldsymbol{u}_{[1]}$ o wymiarze $\tilde{K}^* J \times 1$ składa się z efektów losowych związanych z różnicami w oddziaływaniu regresorów na regresanta między regionami. Macierz $\boldsymbol{MA}$ o wymiarach $N \times J$ składa się z zer i jedynek. Element (i, j) wynosi 1, jeśli *i*-ta firma zlokalizowana jest w *j*-tej jednostce administracyjnej oraz jest równy 0 w przeciwnym przypadku. Wektor $\boldsymbol{u}_{[2]}$ o wymiarach $J \times 1$ składa się z efektów losowych wskazujących na różnice w wielkościach zmiennej wynikowej między jednostkami administracyjnymi. $I\times1$-wymiarowy wektor ε zawiera składniki losowe. Model (3.11) w najbardziej ogólny sposób można zapisać następująco:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}\,, \tag{3.12}$$

gdzie $\boldsymbol{X} = \left[\boldsymbol{X}_{[1]} \quad \boldsymbol{X}_{[2]} \right]$, $\boldsymbol{\beta} = \left[\boldsymbol{\beta}_{[1]}^T \quad \boldsymbol{\beta}_{[2]}^T \right]^T$, $\boldsymbol{Z} = \left[\boldsymbol{X}_{[3]}^{(1)} \middle| \boldsymbol{X}_{[3]}^{(2)} \middle| \ldots \middle| \boldsymbol{X}_{[3]}^{(J)} \middle| \boldsymbol{MA} \right]$, $\boldsymbol{u} = \left[\boldsymbol{u}_{[1]}^T \quad \boldsymbol{u}_{[2]}^T \right]^T$. W odniesieniu do macierzy wariancji-kowariancji dla składników i efektów losowych przyjmuje się założenie:

$$E\left(\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}\begin{bmatrix}\boldsymbol{u}^T & \boldsymbol{\varepsilon}^T\end{bmatrix}\right)=\begin{bmatrix}\boldsymbol{\Omega} & 0\\ 0 & \sigma_\varepsilon^2\boldsymbol{I}\end{bmatrix}. \tag{3.13}$$

Ponieważ $E(\boldsymbol{u}) = 0$ oraz $E(\varepsilon) = 0$, zachodzi $E(\boldsymbol{y}) = \boldsymbol{X\beta}$. Ze wzoru (3.13) wynika, że macierz wariancji-kowariancji dla zmiennej wynikowej przyjmuje postać:

$$E\left(\left(\boldsymbol{y}-1\bar{y}\right)\left(\boldsymbol{y}-1\bar{y}\right)^T\right)=\boldsymbol{V}=\boldsymbol{Z\Omega Z}^T+\sigma_\varepsilon^2\boldsymbol{I}\,, \tag{3.14}$$

gdzie 1 jest wektorem „jedynek" o wymiarach $I \times 1$.

Jedną z metod estymacji parametrów modelu wielopoziomowego jest metoda największej wiarygodności. Estymacja parametrów modelu (3.12) metodą największej wiarygodności rozważana była między innymi przez Davida Harville'a (1977) oraz Shayle'a Searle'a, Georga Casellę i Charlesa McCullocha (1992). Przyjmując, że θ jest wektorem składającym się unikatowych elementów macierzy Ω, funkcję wiarygodności można zapisać następująco:

$$\ln L\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma_\varepsilon^2\right)=-\frac{I}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(det\left(\boldsymbol{V}\right)\right)-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^T\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right). \tag{3.15}$$

Jak wskazuje między innymi Łaszkiewicz (2016), maksymalizacja funkcji (3.15) jest równoważna minimalizacji wyrażenia:

$$ld\left(\boldsymbol{\beta},\boldsymbol{\theta},\sigma_\varepsilon^2\right)=I\ln\left(2\pi\right)+\ln\left(det\left(\boldsymbol{V}\right)\right)+\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^T\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right). \tag{3.16}$$

Na początku dla ustalonych elementów wektora $\boldsymbol{\theta}$ wyznaczany jest estymator $\boldsymbol{\beta}$ zgodnie ze wzorem:

$$\hat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}^{-1}\boldsymbol{y}\,. \tag{3.17}$$

Po wyznaczeniu oszacowań parametrów efektów stałych można dokonać profilowania funkcji (3.16), to znaczy wyznaczyć funkcję zależną od parametrów $\boldsymbol{\theta}$ oraz σ_ε^2 (por. Greene, 2008):

$$ld\left(\boldsymbol{\theta},\sigma_\varepsilon^2\right)=I\ln\left(2\pi\right)+I\ln\left(\sigma_\varepsilon^2\right)+\ln\left(det\left(\boldsymbol{V}\right)\right)+\frac{\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T\boldsymbol{V}^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)}{\sigma_\varepsilon^2}+I\,. \tag{3.18}$$

Przez analogię, wykorzystując ustalone elementy wektora $\boldsymbol{\theta}$, dokonuje się profilowania funkcji (3.18) w taki sposób, że zależy ona tylko od parametru σ_ε^2. Maksymalizując tę funkcję względem σ_ε^2, uzyskuje się następujące oszacowanie:

$$\hat{\sigma}_\varepsilon^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{V}^*\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) / I\,, \tag{3.19}$$

gdzie $\boldsymbol{V}^* = \boldsymbol{I} + \boldsymbol{Z}\hat{\boldsymbol{\Omega}}\boldsymbol{Z}^T$. Podstawiając wyrażenie (3.19) do funkcji (3.18), uzyskuje się następującą funkcję wektora $\boldsymbol{\theta}$:

$$\begin{aligned} ld\left(\boldsymbol{\theta}\right) &= I\ln\left(2\pi / I\right) + I\ln\left(\sigma_\varepsilon^2\right) + \ln\left(det\left(\boldsymbol{V}\right)\right) + I \\ &+ I\ln\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) / \sigma_\varepsilon^2\,. \end{aligned} \tag{3.20}$$

Iteracyjna metoda Newtona-Raphsona jest w dalszej kolejności wykorzystywana w celu minimalizacji funkcji (3.20) ze względu na wektor parametrów $\boldsymbol{\theta}$.

Kolejnym ważnym algorytmem wykorzystywanym podczas estymacji parametrów liniowych modeli wielopoziomowych jest algorytm EM (*Expectation Maximization*). Aby wykorzystać analizowaną metodę, przyjmuje się założenie, że dane wchodzące w skład wektora $\boldsymbol{u}$ mają charakter brakujących. Logarytm funkcji wiarygodności dla pełnych danych ($\boldsymbol{y}$, $\boldsymbol{u}$) przyjmuje postać:

$$\ln L_F\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2\right) = \sum_{j=1}^{J}\left\{\ln\left\{f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}, \boldsymbol{\beta}, \sigma_\varepsilon^2\right)\right\} + \ln\left\{g_{uu}\left(\boldsymbol{u} \mid \boldsymbol{\Omega}\right)\right\}\right\}, \tag{3.21}$$

gdzie $f_{wj}\left(\cdot\right)$ jest funkcją gęstości zmiennej losowej pochodzącej z wielowymiarowego rozkładu normalnego o wartości oczekiwanej $\boldsymbol{X}_j\boldsymbol{\beta} + \boldsymbol{Z}_j\boldsymbol{u}$ i wariancji $\sigma_\varepsilon^2\boldsymbol{I}_{I_j}$, natomiast $g_{uu}\left(\cdot\right)$ jest funkcją gęstości zmiennej losowej pochodzącej z wielowymiarowego rozkładu normalnego o zerowej wartości oczekiwanej i macierzy wariancji-kowariancji $\boldsymbol{\Omega}$. Kolejne kroki algorytmu EM są następujące:

Krok 1. Dla bieżącej wartości $\boldsymbol{\Omega}$ w iteracji numer $n(\boldsymbol{\Omega}^{\{n\}})$ wyznaczane są $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}\left(\boldsymbol{\Omega}^{\{n\}}\right)$ oraz $\hat{\sigma}_\varepsilon^2 = \hat{\sigma}_\varepsilon^2\left(\boldsymbol{\Omega}^{\{n\}}\right)$ zgodnie z wzorami (3.17) oraz (3.19).

Krok 2. Wyznaczana jest wartość oczekiwana funkcji wiarygodności, zależna od elementów macierzy $\boldsymbol{\Omega}$:

$$\begin{aligned} D\left(\boldsymbol{\Omega}\right) &\equiv E\left\{\ln L_F\left(\hat{\boldsymbol{\beta}}, \boldsymbol{\Omega}, \hat{\sigma}_\varepsilon^2\right) \mid \boldsymbol{y}\right\} = \\ &= C - \frac{QQ}{2}\ln\left(det\left(\boldsymbol{\Omega}\right)\right) - \frac{1}{2}\sum_{j=1}^{J} E\left(\boldsymbol{u}_{\{j\}}^T \boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}} \mid \boldsymbol{y}\right), \end{aligned} \tag{3.22}$$

gdzie C jest stałą niezależną od $\boldsymbol{\Omega}$, natomiast wartość oczekiwana formy kwadratowej $\boldsymbol{u}_{\{j\}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{u}_{\{j\}}$ jest obliczana ze względu na warunkową funkcję gęstości $g_{uu}\left(\boldsymbol{u}_{\{j\}} \mid \boldsymbol{y}, \hat{\boldsymbol{\beta}}, \boldsymbol{\Omega}^{\{n\}}, \hat{\sigma}_\varepsilon^2\right)$.

Krok 3. Dokonuje się maksymalizacji funkcji (3.22) ze względu na elementy macierzy $\boldsymbol{\Omega}$ i wyznaczane są oszacowania dla $n + 1$-ego kroku algorytmu iteracyjnego, czyli $\boldsymbol{\Omega}^{\{n+1\}}$.

Najpopularniejszą metodą estymacji parametrów liniowych modeli wielopoziomowych jest zaproponowana przez Goldsteina (1986) iteracyjna uogólniona metoda najmniejszych kwadratów. Idea tej metody polega na tym, że model (3.12) można traktować jako standardowy model liniowy z macierzą kowariancji między składnikami losowymi, która nie spełnia klasycznych założeń schematu Gaussa-Markowa (por. Welfe, 2009). Jeśli macierz $\boldsymbol{V}$ ze wzoru (3.12) jest znana, wówczas estymator uogólnionej metody najmniejszym kwadratów przyjmuje postać:

$$\hat{\boldsymbol{\beta}}_{GLS} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}. \tag{3.23}$$

Estymator macierzy wariancji-kowariancji estymatorów parametrów przyjmuje zaś postać:

$$E\left(\left(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}\right)^T\right) = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1}. \tag{3.24}$$

Jeśli macierz $\boldsymbol{V}$ nie jest znana, w celu znalezienia oszacowań nieznanych parametrów modelu (3.8) wykorzystywana jest iteracyjna uogólniona metoda najmniejszych kwadratów. Polega ona na tym, że na początku przyjmuje się oszacowanie dla macierzy $\boldsymbol{V}$ wynoszące $\hat{\boldsymbol{V}}^{\{0\}} = \sigma^2 \boldsymbol{I}_{IXI}$. Wówczas zgodny estymator wektora $\boldsymbol{\beta}$ uzyskany w pierwszym kroku algorytmu iteracyjnego przyjmuje postać:

$$\hat{\boldsymbol{\beta}}^{\{1\}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{3.25}$$

Po wyznaczeniu tego estymatora wyznacza się oszacowanie macierzy $\boldsymbol{V}$ w pierwszym kroku w następujący sposób:

$$\hat{\boldsymbol{V}}^{\{1\}} = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\{1\}}\right)\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\{1\}}\right)^T + \boldsymbol{X}\left(\boldsymbol{X}^T \left(\hat{\boldsymbol{V}}^{\{0\}}\right)^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T. \tag{3.26}$$

Następnie w drugim kroku algorytmu iteracyjnego uzyskuje się estymator $\boldsymbol{\beta}$ w następujący sposób:

$$\hat{\boldsymbol{\beta}}^{\{2\}} = \left(\boldsymbol{X}^T \left(\hat{\boldsymbol{V}}^{\{1\}} \right)^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \left(\hat{\boldsymbol{V}}^{\{1\}} \right)^{-1} \boldsymbol{y} . \tag{3.27}$$

Procedura jest kontynuowana aż do osiągnięcia zbieżności.

Jak pokazał Goldstein (1986; 1989), zastosowanie iteracyjnej uogólnionej metody najmniejszych kwadratów jest asymptotycznie numerycznie równoważne z wykorzystaniem estymatora największej wiarygodności dla modelu wielopoziomowego, jak zostało zaproponowane w pracy Harville'a (1977).

Zgodnie z powyższymi rozważaniami, nie szacuje się efektów losowych. Estymacji podlega wariancja tych efektów. Zależą one jednak od predykcji. Predykcja efektów losowych (elementów wektora $\boldsymbol{u}$) polega na wyznaczeniu najlepszego, liniowego i nieobciążonego predyktora (*Best Linear Unibiased Predictor*). Wykorzystywany jest fakt, że łączny rozkład dla wektorów $\boldsymbol{y}$ oraz $\boldsymbol{u}$ przyjmuje następującą postać:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{X\beta} \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z\Omega Z}^T + \sigma_\varepsilon^2 \boldsymbol{I} & \boldsymbol{Z\Omega} \\ \boldsymbol{\Omega Z}^T & \boldsymbol{\Omega} \end{bmatrix} \right). \tag{3.28}$$

Z zależności (3.28) wynika, że wartość oczekiwana efektów losowych, warunkowa ze względu na wartości zmiennej wynikowej, wynosi:

$$E\left(\boldsymbol{u} \mid \boldsymbol{y}\right) = \boldsymbol{\Omega Z}^T \left(\boldsymbol{Z\Omega Z}^T + \sigma_\varepsilon^2 \boldsymbol{I} \right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X\beta} \right). \tag{3.29}$$

Ponieważ wariancja σ_ε^2 oraz elementy macierzy $\boldsymbol{\Omega}$ i $\boldsymbol{V}$ nie są znane, do predykcji efektów losowych wykorzystywana jest następująca formuła:

$$\hat{\mathbf{u}} = \hat{\boldsymbol{\Omega}} \boldsymbol{Z}^T \left(\boldsymbol{Z} \hat{\boldsymbol{\Omega}} \boldsymbol{Z}^T + \hat{\sigma}_\varepsilon^2 \boldsymbol{I} \right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}} \right). \tag{3.30}$$

Predykcje efektów losowych należy interpretować jako oszacowania różnic w poziomie zmiennej wynikowej między jednostkami należącymi do różnych grup. Należy jednak zaznaczyć, że wzory (3.29) oraz (3.30) są prawdziwe przy założeniu rozkładu normalnego zmiennej wynikowej. Ze wzoru (3.30) wynika, że po oszacowaniu parametrów $\boldsymbol{\beta}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{\Omega}}$ oraz po dokonaniu predykcji efektów losowych wektor wartości teoretycznych dla zmiennej wynikowej wyznaczany jest na podstawie wzoru:

$$\boldsymbol{y} = \boldsymbol{X} \hat{\boldsymbol{\beta}} + \boldsymbol{Z} \hat{\boldsymbol{\Omega}} \boldsymbol{Z}^T \left(\boldsymbol{Z} \hat{\boldsymbol{\Omega}} \boldsymbol{Z}^T + \hat{\sigma}_\varepsilon^2 \boldsymbol{I} \right)^{-1} \left(\boldsymbol{y} - \boldsymbol{X} \hat{\boldsymbol{\beta}} \right). \tag{3.31}$$

W modelach wielopoziomowych ważnym zagadnieniem jest testowanie obecności efektów losowych. W celu omówienia tego zagadnienia rozważmy model (3.12), w którym efekty losowe związane są z jednostkami administracyjnymi. Załóżmy, że oprócz efektów losowych występują efekty stałe. Zmienne zero-jedynkowe związane z efektami stałymi wchodzą w skład macierzy $\boldsymbol{X}$. Niech $\overline{\overline{\boldsymbol{\beta}}}$ będzie podwektorem wektora $\boldsymbol{\beta}$ przy zmiennych binarnych dla jednostek administracyjnych. Wówczas mamy do czynienia z następującymi parami modeli zagnieżdżonych:

1) model bez efektów vs. model z efektami stałymi,
2) model bez efektów vs. model z efektami losowymi,
3) model z efektami stałymi vs. model z efektami mieszanymi (stałymi i losowymi).

Aby dokonać wyboru jednego modelu z pierwszej pary, należy zweryfikować prawdziwość hipotezy $\overline{\overline{\boldsymbol{\beta}}} = 0$. W tym celu można skorzystać na przykład ze standardowej statystyki F lub ze standardowych testów wykorzystujących wartości funkcji wiarygodności (np. testu ilorazu wiarygodności) (por. Welfe, 2009). Test ilorazu wiarygodności może zostać również wykorzystany, gdy sprawdzamy, który model jest lepszy dla pary drugiej i trzeciej. Z nieco większym problemem mamy do czynienia w przypadku, gdy model z efektami losowymi jest alternatywą dla zawierającego efekty stałe. Wówczas zasadne jest wykorzystanie testu dla hipotez niezagnieżdżonych Vuonga (1989), porównanie wartości informacyjnych obu modeli lub porównanie granicznych poziomów istotności dla weryfikacji hipotez $\overline{\overline{\boldsymbol{\beta}}} = 0$ oraz $\boldsymbol{u} = 0$.

3.4. Efekty krzyżowe w liniowych modelach wielopoziomowych

Hierarchiczna strukturyzacja danych nie jest ograniczona do wykorzystania informacji o lokalizacji obiektów wewnątrz poszczególnych jednostek administracyjnych. Innym przykładem hierarchicznej strukturyzacji jest wykorzystanie informacji dotyczącej rodzaju działalności prowadzonej przez firmę. Największa grupa obejmuje sekcje, które następnie dzielą się na działy. Wewnątrz poszczególnych działów wyróżnia się grupy, które następnie dzielą się na klasy. Załóżmy, że dysponujemy informacjami na temat wyniku firmy zlokalizowanej w określonym województwie, powiecie czy gminie. Mamy jednak także informacje dotyczące grupy, działu oraz sekcji dla działalności firmy. Niech $y_{i,j1,j2,j3}^{a1,a2,a3}$ oznacza wartość zmiennej wynikowej dla *i*-tej firmy zlokalizowanej w *j*1-ej gminie, *j*2-im powiecie, *j*3-cim województwie, której kod działalności należy do *a*1-ej grupy, *a*2-ego działu oraz *a*3-ciej sekcji. Wówczas równanie analogiczne do równania (3.1), wykorzystujące dekompozycję tej zmiennej, przyjmuje postać:

$$y_{i,j1,j2,j3}^{a1,a2,a3} = \ddot{\alpha}_{i,j1,j2,j3}^{a1,a2,a3} + \ddot{\beta}_{j1,j2,j3}^{*} + \ddot{\gamma}_{j2,j3}^{*} + \ddot{\pi}_{j3}^{*} + \ddot{\beta}_{a1,a2,a3}^{**} + \ddot{\gamma}_{a2,a3}^{**} + \ddot{\pi}_{a3}^{**}. \quad (3.32)$$

Model analogiczny do (3.12), uwzględniający podwójną strukturyzację, przyjmuje postać:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{\{1\}}\boldsymbol{u}_{\{1\}} + \boldsymbol{Z}_{\{2\}}\boldsymbol{u}_{\{2\}} + \boldsymbol{\varepsilon}\,, \tag{3.33}$$

gdzie $\boldsymbol{Z}_{\{1\}}$ jest macierzą zmiennych zero-jedynkowych związanych z przyporządkowaniem firm do odpowiednich jednostek administracyjnych, macierz $\boldsymbol{Z}_{\{2\}}$ zawiera zmienne binarne związane z przyporządkowaniem firm do odpowiednich grup, działów oraz sekcji. Wektory $\boldsymbol{u}_{\{1\}}$ oraz $\boldsymbol{u}_{\{2\}}$ zawierają efekty losowe związane odpowiednio z jednostkami administracyjnymi oraz klasyfikacją działalności.

Model (3.33) można uogólnić na przypadek *D*-krotnej strukturyzacji. Przyjmuje on wówczas następującą postać:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum_{\boldsymbol{d}=1}^{\boldsymbol{D}} \boldsymbol{Z}_{\{\boldsymbol{d}\}}\boldsymbol{u}_{\{\boldsymbol{d}\}} + \boldsymbol{\varepsilon}\,, \tag{3.34}$$

gdzie $\boldsymbol{Z}_{\{d\}}$ jest macierzą obserwacji na zmiennych zero-jedynkowych odpowiadających *d*-tej strukturyzacji, natomiast $\boldsymbol{Z}_{\{d\}}$ jest odpowiadającym tej macierzy wektorem efektów losowych. Założenia dotyczące elementów losowych z równania (3.34) są następujące:

$$E\left(\boldsymbol{u}_{\{d\}}\boldsymbol{u}_{\{d\}}{}^{T}\right) = \boldsymbol{\Omega}_{\{d\}}\,,\ d = 1, 2, \ldots, D, \tag{3.35a}$$

$$E\left(\boldsymbol{u}_{\{d\}}\right) = 0\,,\ d = 1, 2, \ldots, D, \tag{3.35b}$$

$$E\left(\boldsymbol{\varepsilon\varepsilon}^{\boldsymbol{T}}\right) = \sigma^2\boldsymbol{W}\,, \tag{3.35c}$$

$$E\,(\varepsilon) = 0, \tag{3.35d}$$

$$E\left(u_{\{d\}p}u_{\{d'\}r}\right) = 0\,,\ d, d' = 1, 2, \ldots, D \text{ oraz } d \neq d'\,, \tag{3.35e}$$

$$E\left(u_{\{d\}p} \mid u_{\{d'\}r}\right) = 0\,,\ d, d' = 1, 2, \ldots, D \text{ oraz } d \neq d'\,, \tag{3.35f}$$

$$E\left(u_{\{d\}pp}\varepsilon_i\right) = 0\,, \text{ dla każdego } d, pp, i. \tag{3.35g}$$

Jeśli założenia (3.35a)–(3.35f) są spełnione, wówczas prawdziwe są zależności:

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}, \tag{3.36a}$$

$$E\left(\left(\boldsymbol{y} - 1\bar{y}\right)\left(\boldsymbol{y} - 1\bar{y}\right)^{T}\right) = \tilde{\boldsymbol{V}} = \sigma^2 \boldsymbol{W} + \sum_{\boldsymbol{d}=1}^{\boldsymbol{D}} \boldsymbol{Z}_{\{d\}} \boldsymbol{\Omega}_{\{d\}} \boldsymbol{Z}_{\{\boldsymbol{d}\}}{}^{T}. \tag{3.36b}$$

W celu estymacji parametrów modelu (3.34) wykorzystywane są te same techniki jak w przypadku pojedynczej strukturyzacji.

3.5. Wykorzystanie liniowego modelu wielopoziomowego uwzględniającego zmienne regionalne do badania czynników wpływających na wynagrodzenia w Polsce

3.5.1. Przegląd literatury z zakresu czynników wpływających na wynagrodzenia

Prace poświęcone czynnikom wpływającym na wynagrodzenia należy podzielić na dwie zasadnicze grupy. Pierwsza z nich obejmuje modele mikroekonomiczne, w których przedmiotem zainteresowania są wynagrodzenia poszczególnych pracowników, natomiast zmienne objaśniające są związane z ich charakterystykami (na przykład: wiek, płeć, wykształcenie, stan cywilny, wielkość zamieszkiwanej miejscowości). Druga grupa badań dotyczy czynników makroekonomicznych oraz obserwowanych na poziomie mezo jako zmiennych wpływających na średnie wynagrodzenia obserwowane w danym kraju lub jednostce administracyjnej.

Aby zrozumieć modele ekonometryczne wykorzystywane w celu identyfikacji czynników wpływających na regionalne zróżnicowanie między średnimi płacami, należy zapisać model równania płac wchodzącego w skład dwurównaniowego modelu sprzężenia inflacyjnego (por. m.in. Welfe, Majsterek, Florczak, 1994; Welfe, Welfe, 2004; Majsterek, 2008):

$$plac\_nom_t = \delta_0 + \delta_1 cen_t + \delta_2 wyd\_prac_t + \delta_3 bezr_t + \varepsilon_t, \tag{3.37}$$

gdzie:
$plac\_nom_t$ – nominalne płace przeciętne,
cen_t – indeks cen konsumpcyjnych (*CPI*),
wyd_prac_t – wydajność pracy,
$bezr_t$ – stopa bezrobocia.

Model opisany równaniem (3.37) jest na ogół wykorzystywany w celu wyjaśnienia kształtowania się płac w gospodarce narodowej (por. np. Welfe, Karp, 2017). Stanowi też punkt wyjścia w modelach służących identyfikacji determinant międzyregionalnego zróżnicowania płac (por. np. Adamczyk, Tokarski, Włodarczyk, 2009; Cieślik, Rokicki, 2016). Poziom wynagrodzeń w danym powiecie powinien zależeć od sytuacji przetargowej pracowników. Jest ona zależna nie tylko od ich cech indywidualnych, ale również od sytuacji obserwowanej na lokalnym rynku pracy. Spośród dwóch osób charakteryzujących się tym samym poziomem wykształcenia wyższe wymagania płacowe powinna mieć osoba mająca świadomość, że niski poziom bezrobocia przyczynia się do tego, że pracodawca ma trudności ze znalezieniem odpowiednich pracowników. W przypadku wysokiego poziomu bezrobocia w danym regionie nawet osoba bardzo dobrze wykształcona nie może liczyć na wysokie wynagrodzenie. Dlatego też uwzględnienie stopy bezrobocia w modelu wyjaśniającym zróżnicowanie przestrzenne płac jest jak najbardziej uzasadnione. Wzrost wydajności pracy może prowadzić zarówno do wzrostu rentowności firm, jak i zwiększenia wynagrodzeń. Wynika to z faktu, że przedsiębiorstwa są w stanie przeznaczać wypracowaną nadwyżkę na wynagrodzenia dla pracowników. W związku z tym wydajność pracy, którą najczęściej mierzy się za pomocą PKB na zatrudnionego, jest często stosowana w modelach wyjaśniających przestrzenne zróżnicowanie płac. Próbując zidentyfikować czynniki wpływające na różnice w przeciętnych wynagrodzeniach między powiatami lub podregionami, konieczne jest stosowanie alternatywnych zmiennych skorelowanych z PKB, gdyż analizowana kategoria dostępna jest tylko na poziomie województw.

Ważnym wkładem do wiedzy na temat mikroekonomicznych determinant wynagrodzeń jest równanie płac, które zostało opisane w monografii pt. *Schooling, Experience and Earnings* napisanej przez Jacoba Mincera w 1974 roku[2]. Publikacja ta wywarła istotny wpływ w obszarze empirycznych badań nad mikroekonomicznymi czynnikami oddziałującymi na wysokość indywidualnych dochodów. W klasycznym równaniu Mincera zmienną objaśnianą jest logarytm dochodów lub płac, natomiast regresorami są zmienne związane z poziomami wykształcenia oraz poziomem doświadczenia (por. Mincer, Polachek, 1974; Mincer, 1993). Przyjmowane jest *implicite* założenie, że różnice w wynagrodzeniach pracowników o różnym poziomie wykształcenia odzwierciedlają różnice w uzyskiwanych przez nich krańcowych produktywnościach pracy. Dodatkowo przyjmuje się, że zależność między poziomem doświadczenia a krańcową produktywnością pracy najsilniejsza jest w grupie osób o najmniejszym doświadczeniu. Różnica między produktywnością pracy osoby pracującej dwa lata i osoby z rocznym doświadczeniem jest

2 Pierwsze wydanie wspomnianej monografii pojawiło się w 1974 roku. Odwołania w dalszej części dotyczą wznowienia z 1993 roku.

zdecydowanie wyższa niż analogiczna różnica dla osób z doświadczeniem dziesięcio- i jedenastoletnim. Dlatego też przyjmowany jest kwadratowy charakter zależności między płacami a doświadczeniem mierzonym w latach. Postać klasycznego równania płac Mincera jest następująca:

$$\ln\left(WYN_i\right) = \beta_0 + \beta_1 SZK_i + \beta_2 XZ_i + \beta_3 XZ_i^2 + \varepsilon_i\,, \tag{3.38}$$

gdzie:
WYN_i – wynagrodzenie *i*-tego pracownika,
SZK_i – poziom wykształcenia *i*-tego pracownika mierzony liczbą lat nauki,
XZ_i – doświadczenie zawodowe *i*-tego pracownika mierzone liczbą przepracowanych przez niego lat,
ε_i – składnik losowy spełniający klasyczne założenia schematu Gaussa-Markowa.

Niektóre założenia klasycznego równania płac Mincera wydają się mało realistyczne. Najsilniejsza krytyka dotyczy założenia stałości stopy zwrotu względem wszystkich szczebli wykształcenia. W odpowiedzi na tę krytykę Geogre Psacharopoulos i Ying Chu Ng (1994) zaproponowali szacowanie parametrów zmodyfikowanego równania płac Mincera, w którym wśród regresorów wyróżnia się zmienne zero-jedynkowe związane z poziomem wykształcenia uzyskanym przez pracownika. Równanie płac zaproponowane przez omawianych autorów przyjmuje postać:

$$\ln\left(WYN_i\right) = \beta_0 + \beta_1 D1_i + \beta_2 D2_i + \beta_3 D3_i + \beta_4 XZ_i + \beta_5 XZ_i^2 + \varepsilon_i\,, \tag{3.39}$$

gdzie $D1_i$, $D2_i$ oraz $D3_i$ są zmiennymi zero-jedynkowymi mierzącymi poziom wykształcenia uzyskany przez *i*-tego pracownika (odpowiednio: podstawowy[3], średni, wyższy). Bardziej zaawansowane – w stosunku do równania zaproponowanego przez Psacharopoulosa i Nga (1994) – podejście do mierzenia wpływu jakości wykształcenia na poziom wynagrodzeń zaproponowali Charlotte Lauer i Viktor Steiner (2000). Wyróżnili oni uniwersytety, wyższe szkoły zawodowe, szkoły średnie ogólne oraz zawodowe szkoły średnie. Na inną wadę klasycznego równania płac Mincera wskazują między innymi Flavio Cunha i James Heckman (2007), a także Jacek Liwiński i Emilia Bedyk (2016). Autorzy ci argumentują, że w omawianym równaniu nie uwzględnia się wrodzonych zdolności pracowników, inwestycji rodziców w kształcenie dzieci, a także środowiska, w którym wychowują się osoby decydujące o dalszym kształceniu się.

3 Przyjmuje się, że brak wykształcenia jest kategorią referencyjną.

Opisane wzorem (3.39) równanie płac Mincera służy jako podstawa analityczna badań nad wpływem czynników innych niż wykształcenie i doświadczenie zawodowe na wynagrodzenia. Poszerzone równanie płac Mincera przyjmuje postać:

$$\ln\left(WYN_i\right) = \alpha_0 + \alpha_1 D1_i + \alpha_2 D2_i + \alpha_3 D3_i + \alpha_4 X_i + \alpha_5 X_i^2 + \sum_{k=1}^{K} \beta_k ZK_{ki} + \varepsilon_i \tag{3.40}$$

gdzie ZK_{ki} oznacza wartość k-tej zmiennej kontrolnej[4] dla i-tego pracownika (k = 1, 2, …, K), natomiast β_k (k = 1, 2, …, K) są parametrami strukturalnymi przy odpowiednich zmiennych kontrolnych. Zbiór zmiennych kontrolnych w równaniu (3.40) zależy od celu badania empirycznego wykorzystującego poszerzone równanie Mincera. Spośród ważnych zastosowań należy wymienić prace służące identyfikacji zróżnicowania dochodów ze względu na płeć (por. Machin, Puhani, 2002; Roszkowska, Majchrowska, 2014; Majchrowska, Strawiński, 2018), miejsce zamieszkania (np. Clemens, Montenegro, Prichett, 2009), branżę (King, 1978), sektor (Ehrenberg, Schwarz, 1987), religię (por. np. Kortt, Dollery, 2012; Sinnewe, Kortt, Steen, 2016), stan cywilny (Nakosteen, Zimmer, 1987) czy też fakt doświadczenia problemu prawnego oraz sposób reakcji na niego (Florczak, Grabowski, 2018b; Grabowski, 2019).

3.5.2. Koncepcje SBTC i RBTC i ich wykorzystanie do analizy czynników wpływających na różnice między wynagrodzeniami przedstawicieli określonych zawodów

W ostatnich latach na rynkach pracy krajów rozwiniętych zaobserwowano tendencje polegające na zmianie struktury zawodów i rozkładu płac. Nastąpił gwałtowny wzrost różnic między wynagrodzeniami uzyskiwanymi przez pracowników wykwalifikowanych a otrzymywanymi przez osoby niewykwalifikowane. Model koncepcyjny, który próbuje wyjaśnić te zmiany, wykorzystuje zaproponowane przez Richarda Freemana i Lawrenca'a Katza (1994) oraz Lawrence'a Katza i Davida Autora (1999) podejście podażowo-popytowo-instytucjonalne. Czynnikami podażowymi są: zmiany poziomu wykształcenia pracowników, programy szkoleniowe, spadek poziomu kształcenia w określonych profesjach oraz migracje. Do czynników instytucjonalnych zalicza się między innymi uzwiązkowienie,

4 Zmienne kontrolne oznaczają w tym przypadku inne zmienne objaśniające.

zmiany wysokości płac minimalnych oraz elastyczność rynku pracy. Jeśli zaś chodzi o czynniki popytowe, są one związane ze zmianami technologicznymi i handlem międzynarodowym. O ile czynniki podażowe i instytucjonalne mają wpływ na sytuację na rynku pracy w krótkim okresie, o tyle postęp technologiczny kształtuje popyt na określone zawody i wynagrodzenia w długim okresie (por. Arendt, 2018).

Różne koncepcje odnoszące się do obserwowanych i prognozowanych zmian popytu na pracę i zarobków nie prowadzą do tej samej konkluzji. Niemniej jednak dominują dwie ważne hipotezy, których prawdziwość zostanie zweryfikowana w kolejnym podrozdziale. Jedną z nich jest hipoteza zmiany technologicznej faworyzującej wysokie kwalifikacje (*Skill-Biased Technical Change*; dalej hipoteza SBTC). Zgodnie z hipotezą SBTC w wyniku postępu technologicznego pracownicy wysoko wykwalifikowani zastępują cechujących się niższymi kwalifikacjami. Dlatego też długotrwały wzrost popytu na pracowników wysoko wykwalifikowanych prowadzi do wzrostu relacji ich płac do płac uzyskiwanych przez osoby cechujące się niższymi kwalifikacjami. Hipoteza SBTC wykorzystana została przez Darona Acemoglu (2002) w celu wyjaśnienia przeszłych zmian na rynku pracy, wynikających z postępu technologicznego. Sformułował on hipotezę endogenicznego SBTC, zgodnie z którą rewolucja przemysłowa z XIX wieku faworyzowała pracowników o niższych kwalifikacjach. Niemniej jednak, zdaniem Maartena Goosa i Alana Manninga (2007), hipoteza SBTC była w stanie wyjaśnić tylko i wyłącznie zmiany w zarobkach pracowników najlepiej sytuowanych, natomiast nie potrafiła wytłumaczyć tendencji obserwowanych w zmianach wynagrodzeń osób z mediany oraz „ogonów" rozkładów wynagrodzeń.

Jak argumentuje James Wickham (2011), nawet w XXI wieku w krajach rozwiniętych odnotowuje się wzrost produkcji, która nie wymaga wykwalifikowanej siły roboczej. Wynika to w szczególności z faktu, że wybór technologii produkcji dokonywany przez firmy nie zależy tylko i wyłącznie od jej dostępności, ale również od podaży siły roboczej. Ze względu na obserwowany w ostatnich latach wynikający z globalizacji i liberalizacji przepływu towarów i siły roboczej wzrost natężenia procesów migracyjnych firmy często preferują produkcję mniej zaawansowaną technologicznie. Imigranci wybierają na ogół zawody charakteryzujące się brakiem konieczności posiadania wysokich kwalifikacji i niższym wynagrodzeniem (Oesch, Rodriguez Menes, 2010). Dlatego też, jak argumentują Goos i Maning (2007), następuje polaryzacja rynku pracy polegająca na wzroście zapotrzebowania na osoby charakteryzujące się bardzo wysokimi i bardzo niskimi kwalifikacjami, przy jednoczesnym spadku popytu na pracowników o średnich kwalifikacjach. Hipoteza mówiąca o tym, że w wyniku postępu technologicznego na rynku pracy

faworyzowane są osoby charakteryzujące się bardzo wysokimi i bardzo niskimi kwalifikacjami, jest hipotezą postępu technicznego ukierunkowanego na rutynizację (*Routinisation-Biased Technical Change*; dalej hipoteza RBTC). Pierwszą pracą, w której została ona sformułowana, jest artykuł Davida Autora, Franka Levy'ego i Richarda Murnane'a (2003). Hipoteza ALM (od pierwszych liter nazwisk autorów) wskazuje, że kapitał inwestowany w rozwój technologii informatycznych i komunikacyjnych jest substytucyjny względem czynności rutynowych i komplementarny wobec czynności nierutynowych. Dlatego też w przypadku hipotezy RBTC główna uwaga skupia się nie na kwalifikacjach, lecz na wykonywanych czynnościach. Ponieważ czynności rutynowe wykonywane są na ogół przez osoby charakteryzujące się średnim poziomem kwalifikacji, następuje wzrost popytu na pracowników o najwyższych i najniższych kwalifikacjach. David Autor i David Dorn (2013) oraz Antonio Accetturo, Alberto. Dalmazzo i Guido Blasio (2013) argumentują, że polaryzacja na rynku pracy może wynikać ze zmiany preferencji konsumenckich wśród najlepiej wykształconych mieszkańców wielkich aglomeracji. Menedżerowie oraz profesjonaliści, przez prowadzenie odpowiedniego stylu życia, generują popyt na usługi wykonywane przez osoby posiadające niskie kwalifikacje. Prowadzi to do polaryzacji rynku pracy.

Obydwie hipotezy prowadzą do różnych konkluzji dotyczących zmian rozkładu płac w czasie. Obie są zgodne co do tego, że wynagrodzenia osób charakteryzujących się najwyższym poziomem kwalifikacji powinny rosnąć najszybciej. Niemniej jednak brakuje konsensusu w odniesieniu do wpływu postępu technologicznego na relację zarobków pracowników o średnich[5] kwalifikacjach do wynagrodzeń osób charakteryzujących się najniższym poziomem kwalifikacji. Zgodnie z hipotezą SBTC wraz z postępem technologicznym dynamika płac powinna być liniowo skorelowana z poziomem kwalifikacji. Dlatego też płace osób charakteryzujących się średnimi kwalifikacjami powinny rosnąć szybciej niż płace pracowników najniżej wykwalifikowanych. Zgodnie z hipotezą RBTC wzrost popytu na osoby wykonujące fizyczne prace rutynowe (osoby najniżej wykwalifikowane), przy jednoczesnym spadku popytu na pracowników umysłowych wykonujących prace rutynowe (osoby o średnim poziomie kwalifikacji), prowadzi do relatywnego wzrostu wynagrodzeń tych pierwszych w relacji do płac uzyskiwanych przez tych drugich.

5 Osobami o średnich kwalifikacjach można nazwać pracowników będących absolwentami liceów ogólnokształcących, techników oraz szkół pomaturalnych.

3.5.3. Dane dotyczące poziomów wynagrodzeń uzyskiwanych przez pracowników w polskich przedsiębiorstwach. Podział zawodów ze względu na umiejętności posiadane przez pracowników

Istnieją dwa ważne źródła informacji statystycznej na temat wysokości wynagrodzeń. Jednym z nich są dane dotyczące aktywności zawodowej i wynagrodzeń pochodzące z badań aktywności ekonomicznej (BAEL). Badania BAEL są przeprowadzane przez Główny Urząd Statystyczny od 1992 roku z kwartalną częstotliwością. Są one realizowane metodą reprezentacyjną. Dlatego też informacje uzyskane od gospodarstw w nich uczestniczących uogólniane są następnie dla ludności całego kraju. Pytania dotyczące aktywności zawodowej zadawane są osobom, które ukończyły 15 lat. Celem tego badania jest uzyskanie informacji o wielkości oraz strukturze zasobów pracy. W efekcie przeprowadzenia tego badania ustalona zostaje zarówno liczba osób biernych, jak i aktywnych zawodowo. Warto zaznaczyć, że wykonane przez GUS badanie BAEL jest częścią przeprowadzanego na szczeblu Unii Europejskiej badania LFS (European Union Labour Force Survey). Omawiane badanie jest przeprowadzane w 28 krajach członkowskich strefy euro, dwóch krajach kandydujących oraz trzech krajach EFTA (Szwajcaria, Norwegia, Islandia).

Oprócz pytań dotyczących aktywności ekonomicznej w kwestionariuszu BAEL zawarte są pytania związane z wynagrodzeniami i dochodami osób ankietowanych. Problem polega jednak na tym, że respondenci często odmawiają odpowiedzi na pytanie dotyczące wysokości uzyskiwanych wynagrodzeń. Dlatego też korzystając z danych pochodzących z badania aktywności ekonomicznej ludności, należy uwzględnić problem występowania braków w danych.

Bardziej dokładną analizę wynagrodzeń umożliwia reprezentacyjne badanie struktury wynagrodzeń według zawodów, realizowane na formularzu Z-12. Jest ono odpowiednikiem badania Stucture of Earnings Survey prowadzonego w Unii Europejskiej co cztery lata. W Polsce badanie wynagrodzeń przeprowadzane jest z częstotliwością dwuletnią. Jego celem jest określenie struktury i poziomu miesięcznych i godzinowych wynagrodzeń brutto według cech osób (płeć, wiek, poziom wykształcenia, staż pracy, zawód, rodzaj działalności, sektor własności). Zakresem podmiotowym badania są wylosowane podmioty gospodarcze (zaliczane do sekcji PKD2007 od A do S) o liczbie pracujących przekraczającej dziewięć osób.

Zaletą badania struktury wynagrodzeń jest to, że informacje dotyczące wysokości wynagrodzeń przekazywane są przez pracodawców. Dlatego też nie występuje problem związany z unikaniem przekazywania informacji. Oprócz danych dotyczących wynagrodzeń badacz posiada informacje związane z płcią, poziomem wykształcenia, doświadczeniem zawodowym, sekcją itp. Dlatego też możliwa jest identyfikacja wpływu poszczególnych cech pracowników na wysokość uzyskiwanych wynagrodzeń.

Informacje dotyczące kodów zawodów pochodzące z Klasyfikacji Zawodów i Specjalności umożliwiają potraktowanie struktury danych jako hierarchicznej. Rozważmy osobę wykonującą zawód fizyka medycznego (sześciocyfrowy kod zawodów: 211104). Osoba ta należy do grupy fizyków i astronomów (czterocyfrowy kod zawodów: 2111), która zawiera się w grupie fizyków, chemików i specjalistów nauk o ziemi (trzycyfrowy kod zawodów: 211), należącej do grupy specjalistów nauk fizycznych, matematycznych, technicznych i informatycznych (dwucyfrowy kod zawodów: 21). Tabela 10 prezentuje główne grupy zawodów dla kodów jednocyfrowych.

Tabela 10. Kody zawodów jednocyfrowych

Numer grupy	Nazwa grupy	Poziom umiejętności
1	Przedstawiciele władz publicznych, wyżsi urzędnicy i kierownicy	III[a)], IV
2	Specjaliści	IV
3	Technicy i inny średni personel	III
4	Pracownicy biurowi	II
5	Pracownicy usług i sprzedawcy	II
6	Rolnicy, ogrodnicy, leśnicy i rybacy	II
7	Robotnicy przemysłowi i rzemieślnicy	II
8	Operatorzy i monterzy maszyn i urządzeń	II
9	Pracownicy wykonujący prace proste	I

[a)] Pracownicy wykonujący zawody należące do grupy dwucyfrowej 14 są zaliczani do osób o wysokich, ale nie najwyższych kwalifikacjach. Przedstawiciele pozostałych zawodów należących do grupy jednocyfrowej numer 1 są zaliczani do osób o najwyższych kwalifikacjach.

Źródło: opracowanie własne.

Szczegóły dotyczące podziałów zawodów na grupy i kodów zawodów co najmniej dwucyfrowych można znaleźć na stronie internetowej http://www.klasyfikacje.gofin.pl/kzis/6,0.html.

Aby móc weryfikować prawdziwość hipotez dotyczących tendencji zachodzących na rynku pracy w Polsce, konieczne jest powiązanie zawodów z umiejętnościami wymaganymi do ich wykonywania. Jednym ze sposobów jest wykorzystanie metody zaproponowanej przez Międzynarodową Organizację Pracy (ILO, 2012). Zgodnie z tą propozycją (por. tabela 10) w zależności od wykonywanego zawodu definiowane są cztery poziomy umiejętności. Pierwszy, najniższy poziom umiejętności wiąże się z zawodami zaczynającymi się cyfrą „9", czyli odnoszącymi się do prac prostych. Drugi poziom umiejętności jest wymagany do wykonywania prac związanych z zawodami z grup jednocyfrowych 4, 5, 6, 7 oraz 8. Trzeci poziom umiejętności posiadają przedstawiciele zawodów z grupy jednocyfrowej numer 3

oraz grupy dwucyfrowej numer 14 (kierownicy w branży hotelarskiej, handlu i innych branżach usługowych), podczas gdy do zawodów należących do najwyższej grupy (z punktu widzenia wymaganych umiejętności) zalicza się przedstawicieli drugiej grupy jednocyfrowej oraz pierwszej grupy jednocyfrowej z pominięciem tych z czternastej grupy dwucyfrowej.

Tabela 11. Przyporządkowanie trzycyfrowych grup zawodów do grup zadaniowych

Grupa zadaniowa	Trzycyfrowe grupy zawodowe należące do danej grupy zadaniowej
NRUI	111, 112, 121, 122, 131, 132, 133, 134, 141, 142, 143, 222, 223, 224, 225, 229, 232, 233, 234, 235
NRUA	211, 212, 213, 214, 215, 216, 227, 228, 241, 242, 251, 252, 261, 262, 263, 322
NRUIA	221, 226, 231, 244, 264, 265, 335
NRF	311, 312, 313, 314, 315, 321, 323, 324, 325
RU	331, 341, 411, 412, 413, 421, 422, 431, 432, 441, 531
RF	511, 512, 513, 514, 515, 516, 521, 522, 523, 524, 532, 541, 611, 612, 613, 621, 622, 631, 632, 633, 634, 711, 712, 713, 721, 722, 723, 731, 732, 741, 742, 751, 752, 753, 754, 811, 812, 813, 814, 815, 816, 817, 818, 821, 831, 832, 833, 834, 835, 911, 912, 921, 931, 932, 933, 941, 951, 952, 961, 962

Źródło: opracowanie własne.

Alternatywny podział odwołujący się do pracy Acemoglu i Autora (2011) i wykorzystujący dane pochodzące z badania aktywności ekonomicznej ludności i bazy O-NET został zaproponowany w pracy Wojciecha Hardego, Romy Keister i Piotra Lewandowskiego (2018). Jest on oparty na analizie ewolucji „zawartości" zawodów wykonywanych przez pracowników w krajach Europy Środkowo-Wschodniej. Punktem wyjścia takiego podziału jest przyjęcie założenia, że charakter wykonywanych przez pracowników zadań ma większy wpływ na ich płace niż posiadane przez nich kwalifikacje. Zgodnie z koncepcją zaproponowaną przez Autora, Levy'ego i Murane'a (2003) wykonywane zadania dzielą się przede wszystkim na rutynowe i nierutynowe. Ponieważ wykonywanie zadań nierutynowych i powtarzalnych nie wymaga posiadania dodatkowych umiejętności, a postęp technologiczny w długim okresie przyczynia się do spadku popytu na pracowników wykonujących takie prace, nie mogą oni liczyć na wysokie płace. Arendt i Grabowski (2018), rozszerzając koncepcję wprowadzoną przez Hardego, Keister i Lewandowskiego (2018), zaproponowali następujący podział zawodów ze względu na charakter wykonywanych zadań:

1) zawody nierutynowe wymagające pracy umysłowej oraz umiejętności interpersonalnych (NRUI),
2) zawody nierutynowe wymagające pracy umysłowej oraz umiejętności analitycznych (NRUA),

3) zawody nierutynowe wymagające pracy umysłowej oraz zarówno umiejętności interpersonalnych, jak i analitycznych (NRUIA),
4) zawody nierutynowe wymagające pracy fizycznej (NRF),
5) zawody rutynowe wymagające pracy umysłowej (RU),
6) zawody rutynowe wymagające pracy fizycznej (RF).

Tabela 11 prezentuje przyporządkowanie trzycyfrowych grup zawodów do sześciu wymienionych grup zadaniowych.

3.5.4. Specyfikacja modelu ekonometrycznego wykorzystywanego do analizy czynników wpływających na wynagrodzenia w Polsce

Badanie empiryczne omawiane w niniejszym podrozdziale oparte jest na danych pochodzących z badania struktury wynagrodzeń. Na początku należy zdefiniować wykorzystywaną w nich zmienną zależną. Ponieważ na skutek wzrostu cen, wydajności pracy oraz zmieniającej się stopy bezrobocia zmieniają się płace w gospodarce realnej, definiowana jest zmienna odporna na te zmiany. Proponuje się normalizację nominalnego wynagrodzenia przez podzielenie go przez medianę ze wszystkich wynagrodzeń raportowanych w danej edycji badania aktywności ekonomicznej ludności lub badania struktury wynagrodzeń. Oczywiście w celu porównania dwóch wynagrodzeń należy mieć pewność, że są one uzyskiwane przez dwóch pracowników pracujących na taką samą część etatu. Dlatego też rozważane są pełnoetatowe wynagrodzenia pracowników. Jeśli na przykład dany pracownik pracuje na pół etatu, wówczas w celu uzyskania wynagrodzenia pełnoetatowego rzeczywiste wynagrodzenie należy przemnożyć przez 2. Wynagrodzenie relatywne uzyskiwane przez *i*-tego pracownika w okresie *t* definiuje zatem następująca formuła:

$$WYN\_REL_{it} = \frac{WYN\_NOM_{it}}{med\left(WYN\_NOM_{it}\right)}. \tag{3.41}$$

W analizie wykorzystującej dane pochodzące ze badania struktury wynagrodzeń rozważane są zmienne objaśniające, w stosunku do których przyjmuje się założenie o stałości parametrów z nimi związanych (tabela 12).

Oprócz czynników związanych z wykształceniem, doświadczeniem zawodowym czy płcią należy uwzględnić zmienne związane z rodzajem działalności wykonywanej przez firmę. Wydaje się, że uzasadnione jest zastosowanie efektów losowych związanych z przynależnością firmy, w której pracuje dany pracownik, do sekcji PKD. Wynika to z faktu, że osoby pracujące w firmie wykonującej określoną działalność porównują swoje wynagrodzenia z oferowanymi przez inne firmy wykonujące

tę samą działalność. Dlatego też należy oczekiwać, że zróżnicowanie wynagrodzeń wśród pracowników firm należących do tej samej sekcji PKD powinno być niższe niż zróżnicowanie między sekcjami. W przypadku tej kategorii mamy do czynienia ze strukturą hierarchiczną opisaną na rysunku 2.

Tabela 12. Zmienne objaśniające wpływające na wysokość wynagrodzeń pracowników

Nazwa zmiennej	Opis
Zmienne związane z poziomem wykształcenia danego pracownika	
WYZSZE[a)]	Zmienna binarna przyjmująca wartość 1 dla pracownika z wyższym wykształceniem
SREDNIE_ZAWODOWE	Zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem średnim technicznym lub policealnym
ZASADNICZE_ZAWODOWE	Zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem zasadniczym zawodowym
PODSTAWOWE	Zmienna binarna przyjmująca wartość 1 dla pracownika z wykształceniem podstawowym
Zmienne związane z doświadczeniem pracownika	
DOSW_FIRMA	Liczba pełnych przepracowanych lat przez pracownika w przedsiębiorstwie, w którym obecnie pracuje
DOSW_OGOL	Liczba pełnych przepracowanych lat przez pracownika
Zmienne związane z płcią pracownika, sektorem własności firmy oraz jej rozmiarem	
ROZMIAR_FIRMY	Liczba pracowników pracujących w danej firmie
KOBIETA	Zmienna binarna przyjmująca wartość 1, jeśli badany pracownik jest kobietą
SEKTOR_PRYWATNY	Zmienna binarna przyjmująca wartość 1 w przypadku pracownika firmy z sektora prywatnego
NIEOKRESLONY	Zmienna binarna przyjmująca wartość 1 w przypadku, gdy pracownik jest zatrudniony na czas nieokreślony

[a)] W badaniach struktury wynagrodzeń przeprowadzanych od 2010 roku wyróżnia się dodatkową kategorię „Doktorat". Ponieważ w badaniu uwzględniane są dane obejmujące okres od 2004 do 2016 roku, pracownicy posiadający stopień naukowy co najmniej doktora klasyfikowani są jako osoby z wyższym wykształceniem.

Źródło: opracowanie własne.

Rysunek 2. Struktura hierarchiczna związana z podziałem działalności wykonywanych przez firmy na sekcje

Źródło: opracowanie własne.

Dlatego też w modelu ekonometrycznym należy uwzględnić efekty losowe związane z przynależnością firmy, w której pracuje ankietowany, do odpowiedniej sekcji/grupy. Ponieważ losowe różnice między sekcjami/grupami mogą być różne w kolejnych latach, w modelu ekonometrycznym wprowadzone zostaną efekty losowe zarówno dla sekcji, jak i lat.

W badaniu struktury wynagrodzeń możliwe jest uzyskanie ograniczonej informacji dotyczącej lokalizacji firmy, w której pracuje dany pracownik. Jedyną informacją dostępną w tym badaniu jest województwo, w którym zlokalizowana jest firma zatrudniająca pracownika. Na podstawie zaprezentowanego w podrozdziale 3.5.1 przeglądu literatury z zakresu makroekonomicznych determinant wynagrodzeń wydaje się, że stopa bezrobocia w danym województwie oraz wydajność pracy w regionie determinują poziomy wynagrodzeń uzyskiwanych przez pracowników. Dlatego też konieczne jest uwzględnienie odpowiednich zmiennych różniących się ze względu na województwa oraz lata. Ponieważ zmienna zależna ma charakter zmiennej relatywnej, a normalizacja odbywa się przez podzielenie wartości nominalnej przez wartość dla odpowiedniego roku, podobnej normalizacji należy dokonać w przypadku stopy bezrobocia. Dlatego też w badaniu empirycznym wykorzystywana jest następująca zmienna znormalizowana:

$$\widetilde{BEZR}_{it}^{w} = \frac{BEZR_{it}^{w}}{BEZR_{it}^{k}},\ w = 1, 2, \ldots, 16, \tag{3.42}$$

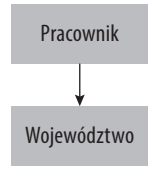
gdzie $w = 1, 2, \ldots, 16$ numeruje poszczególne województwa, $BEZR_{it}^{k}$ jest stopą bezrobocia w całym kraju, natomiast $BEZR_{it}^{w}$ oznacza stopę bezrobocia w w-tym województwie w okresie t. W podobny sposób wprowadzana jest zmienna związana z relatywnym poziomem wydajności pracy w w-tym województwie. Dla każdego województwa definiowana jest następująca kategoria:

$$\widetilde{WYD}_{it}^{j} = \frac{WYD_{it}^{w}}{WYD_{it}^{k}},\ w = 1, 2, \ldots, 16, \tag{3.43}$$

gdzie WYD_{it}^{k} jest poziomem wydajności pracy w całym kraju (mierzonym jako iloraz Produktu Krajowego Brutto do liczby zatrudnionych), natomiast WYD_{it}^{w} oznacza wydajność pracy odnotowaną w w-tym województwie w okresie t.

Oprócz zmieniających się w czasie zmiennych relatywnych związanych ze stopą bezrobocia oraz poziomem wydajności pracy można uwzględnić dodatkowo losowe zróżnicowanie wynagrodzeń między województwami.

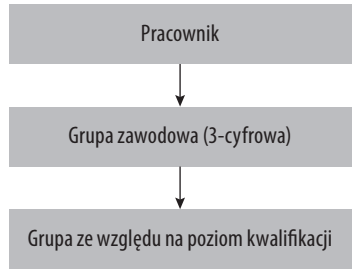
Wydaje się, że gdyby istniały duże różnice między wynagrodzeniami osób wykonujących ten sam zawód, charakteryzującymi się tym samym poziomem wykształcenia i doświadczenia zawodowego oraz pracujących w tych samych województwach, pracownicy z firmy oferującej niższe zarobki przechodziliby do firmy umożliwiającej uzyskanie wyższego dochodu. Wysokie różnice są jednak możliwe, gdy dwie firmy zlokalizowane są daleko od siebie (np. w innych województwach). Tak więc oczekuje się wysokiej korelacji między wynagrodzeniami osób z tych samych województw oraz zróżnicowania płac między województwami. Dlatego też w modelu ekonometrycznym wprowadzane są efekty losowe związane z województwami, które różnią się w poszczególnych latach. Hierarchiczna struktura danych związanych z lokalizacją opisana jest na rysunku 3.



Rysunek 3. Struktura hierarchiczna związana z województwami, w których zlokalizowane są firmy

Źródło: opracowanie własne.

Oprócz dotychczas wymienionych czynników na wysokość wynagrodzeń również powinna mieć wpływ przynależność do określonej grupy zawodowej. Wynika to z faktu, że osoby wykonujące określony zawód porównują swoje wynagrodzenia z innymi osobami wykonującymi podobny zawód. Gdyby zarobki pracowników wykonujących podobne zadania różniły się od siebie istotnie, następowałby przepływ zatrudnionych z firm płacących mniej do przedsiębiorstw oferujących wyższe wynagrodzenia. Z drugiej strony wykonywanie różnych zawodów wiąże się z posiadaniem innych umiejętności i kwalifikacji. Dlatego też oczekuje się wysokiego współczynnika korelacji między zarobkami osób należących do tej samej grupy zawodowej oraz dużego zróżnicowania między wynagrodzeniami osób należących do różnych grup zawodowych. W tej sytuacji zasadne wydaje się być zastosowanie modelu wielopoziomowego w badaniach nad wynagrodzeniami pracowników i predykcje efektów losowych dla określonych grup zawodowych. Trzeba jednak pamiętać, że poszczególne grupy zawodowe należą do szerszych grup. W podrozdziale 3.5.3 omówiony został podział zawodów ze względu na kwalifikacje pracowników oraz charakter wykonywanych zadań. W obu przypadkach mamy do czynienia z hierarchiczną strukturą danych, którą opisują rysunki 4 i 5.



Rysunek 4. Hierarchiczna struktura danych ze względu na przynależność pracowników do grup zawodowych oraz ze względu na przynależność tych grup do grup ze względu na poziom kwalifikacji

Źródło: opracowanie własne.

Rysunek 5. Hierarchiczna struktura danych ze względu na przynależność pracowników do grup zawodowych oraz grup ze względu na przynależność tych grup do grup zadaniowych

Źródło: opracowanie własne.

Na podstawie dotychczasowych rozważań proponowane są dwie specyfikacje modelu ekonometrycznego wykorzystywanego do identyfikacji czynników wpływających na wynagrodzenia uzyskiwane przez polskich pracowników w latach 2004–2016.

$$WYN\_REL_{it} = \sum_{k=1}^{K} \beta_k x_{kit} + \gamma_1 \widetilde{WYD}_{it}^{w} + \gamma_2 \widetilde{BEZR}_{it}^{w} + \\ + \sum_{t'=2004}^{2016} \sum_{j=1}^{16} z1_{it}^{jt'} u1_{t'}^{j} + \\ + \sum_{t'=2004}^{2016} \sum_{s} z2_{it}^{st'} u2_{t'}^{s} + \sum_{t'=2004}^{2016} \sum_{w} z3_{it}^{wt'} u3_{t'}^{w} + \sum_{t'=2008}^{2016} \sum_{p} z4_{it}^{pt'} u4_{t'}^{p} + \varepsilon_{it}, \tag{3.44}$$

gdzie $x_{1it}, x_{2it}, \ldots, x_{Kit}$ są zmiennymi objaśniającymi (zdefiniowanymi w tabeli 12), przy których parametry są nielosowe i nie różnią się od siebie w poszczególnych

latach, zmienne regionalne $\widetilde{WYD}_{it}^{w}$ oraz $\widetilde{BEZR}_{it}^{j}$ są zdefiniowane odpowiednio wzorami (3.43) i (3.42), $u1_t^w$ oznacza efekty losowe związane z poszczególnymi województwami w kolejnych latach. $u2_t^s$ oznaczają efekty losowe związane z poszczególnymi sekcjami w kolejnych latach. Efekty losowe $u3_t^{\breve{w}}$, $u4_t^p$ są związane odpowiednio z trzycyfrowymi grupami oraz grupami ze względu na poziom kwalifikacji. $K1^w$ jest zbiorem pracowników reprezentujących firmy zlokalizowane w *w*-tym województwie, $K2^s$ jest zbiorem osób pracujących w firmach należących do *s*-tej sekcji, $K3^{\breve{w}}$ jest zbiorem pracowników reprezentujących $\breve{w}$-tą trzycyfrową grupę zawodów, natomiast $K4^p$ jest zbiorem pracowników reprezentujących *p*-tą grupę pracowników ze względu na posiadane kwalifikacje. Zmienna zero-jedynkowa $z1_{it}^{wj'}$ przyjmuje wartość 1, jeśli $t = t'$ oraz $i \in K1^w$. Zmienna $z2_{it}^{st'}$ przyjmuje wartość 1, jeśli $t = t'$ oraz $i \in K2^s$. Przez analogię $z3_{it}^{\breve{w}t'} = 1$, gdy $t = t'$ oraz $i \in K3^{\breve{w}}$, natomiast $z4_{it}^{pt'} = 1$, gdy $t = t'$ oraz $i \in K4^p$.

W drugiej proponowanej specyfikacji efekty losowe związane z grupami ze względu na poziom kwalifikacji zastępowane są grupami zadaniowymi. Dlatego też wielopoziomowy model wyjaśniający kształtowanie się relatywnych wynagrodzeń przyjmuje postać:

$$WYN\_REL_{it} = \sum_{k=1}^{K}\beta_k x_{kit} + \gamma_1 \widetilde{WYD}_{it}^{j} + \gamma_2 \widetilde{BEZR}_{it}^{w} + \sum_{t'=2004}^{2016}\sum_{w=1}^{16} z1_{it}^{wt'} u1_{t'}^{w} + \\ + \sum_{t'=2004}^{2016}\sum_{s} z2_{it}^{st'} u2_{t'}^{s} + \sum_{t'=2004}^{2016}\sum_{\breve{w}} z3_{it}^{\breve{w}t'} u3_{t'}^{\breve{w}} + \sum_{t'=2008}^{2016}\sum_{p} z5_{it}^{bt'} u5_{t'}^{b} + \varepsilon_{it}, \tag{3.45}$$

gdzie efekty losowe $u5_t^b$ są związane z grupami zadaniowymi, $z5_{it}^{bt'} = 1$, gdy $t = t'$ oraz $i \in K5^b$, natomiast $K5^b$ jest zbiorem pracowników należących do *b*-tej grupy zadaniowej. Definicje zmiennych $z1_{it}^{wt'}$, $z2_{it}^{st'}$ oraz $z3_{it}^{wt'}$ są takie same jak w przypadku modelu (3.44).

Tabela 13 prezentuje oszacowania parametrów modelu wielopoziomowego (specyfikacja 3.44) przy zmiennych indywidualnych oraz regionalnych, uzyskane z wykorzystaniem procedur w programie STATA15.

Oszacowania parametrów przy zmiennych związanych z poziomem wykształcenia są zgodne z oczekiwaniami. Pracownicy posiadający wyższe wykształcenie zarabiają więcej niż osoby o wykształceniu średnim ogólnokształcącym, przy innych czynnikach niezmienionych. Ta sama sytuacja dotyczy osób z wykształceniem średnim zawodowym. Osoby z wykształceniem podstawowym oraz zasadniczym zawodowym zarabiają, przy innych czynnikach niezmienionych, mniej w porównaniu z osobami posiadającymi wykształcenie średnie ogólnokształcące.

Tabela 13. Oszacowania parametrów modelu wielopoziomowego wyjaśniającego kształtowanie się wynagrodzeń w polskiej gospodarce

Zmienna	Oszacowanie	Błąd standardowy	Graniczny poziom istotności
WYZSZE	0,6586	0,0014	0,000
SREDNIE_ZAWODOWE	0,0596	0,0009	0,000
ZASADNICZE_ZAWODOWE	–0,1804	0,0015	0,000
PODSTAWOWE	–0,1992	0,0020	0,000
DOSW_FIRMA	0,0051	0,0001	0,000
DOSW_OGOL	0,0081	0,0001	0,000
ROZMIAR_10_49	0,1179	0,0032	0,000
ROZMIAR_50–249	0,1892	0,0032	0,000
ROZMIAR_250_499	0,2170	0,0033	0,000
ROZMIAR_CON500	0,2772	0,0032	0,000
KOBIETA	–0,2279	0,0008	0,000
SEKTOR_PRYWATNY	0,0666	0,0008	0,000
NIEOKRESLONY	0,2040	0,0009	0,000
$\widetilde{WYD}_{it}^{j}$	0,3289	0,0020	0,000
$\widetilde{BEZR}_{it}^{j}$	–0,0759	0,0063	0,000
Liczba obserwacji	4 288 026		

Źródło: opracowanie własne.

Wyniki zawarte w tabeli 13 umożliwiają jedynie statystyczną weryfikację różnic między wynagrodzeniami osób o odpowiednim poziomie wykształcenia a wynagrodzeniami uzyskiwanymi przez osoby z wykształceniem średnim ogólnokształcącym. Możliwe jest jednak zweryfikowanie hipotez dotyczących równości odpowiednich par parametrów. Dlatego też tabela 14 zawiera wyniki w zakresie testowania, czy wynagrodzenia osób z dwóch grup istotnie różnią się od siebie.

Tabela 14. Weryfikacja hipotez o równości „stóp zwrotu" dla osób o różnych poziomach wykształcenia

	WYZSZE	*SREDNIE_ ZAWODOWE*	*SRED- NIE_OGOL- NOKSZTAL- CACE*	*ZASADNI- CZE_ZAWO- DOWE*	*PODSTA- WOWE*
WYZSZE	X	+	+	+	+
SREDNIE_ZAWODOWE	–	X	+	+	+
SREDNIE_ OGOLNOKSZTALCACE	–	–	X	+	+
ZASADNICZE_ZAWODOWE	–	–	–	X	0
PODSTAWOWE	–	–	–	0	X

Źródło: opracowanie własne.

Na przykład „+" w komórce dla pary *WYZSZE-SREDNIE_OGOLNOKSZTALCACE* oznacza, że przyjmując poziom istotności 0,01, należy odrzucić hipotezę o równości wynagrodzeń osób z wyższym wykształceniem z wynagrodzeniami osób o wykształceniu średnim ogólnokształcącym. Z drugiej strony „–" w komórce *SREDNIE_OGOLNOKSZTALCACE-WYZSZE* oznacza, że pracownicy z wykształceniem średnim ogólnokształcącym zarabiają znacznie (przyjmując poziom istotności 0,01) gorzej w porównaniu z pracownikami posiadającymi wyższe wykształcenie. „0" w komórce dla pary *ZASADNICZE_ZAWODOWE-PODSTAWOWE* oznacza, że nie istnieją istotne (przyjmując poziom istotności 0,01) różnice w wynagrodzeniach osób z wykształceniem zasadniczym zawodowym i podstawowym.

Uzyskane wyniki wskazują, że wraz ze wzrostem poziomu wykształcenia obserwowany jest wzrost wynagrodzeń, co jest zgodne z teorią zwrotu z edukacji opracowaną przez Mincera (1993). Wynik ten potwierdza także rezultaty innych badań mikroekonomicznych, opartych na danych dla Polski, wskazujących na rolę poziomu wykształcenia w determinowaniu wynagrodzeń (por. m.in. Florczak, Grabowski, 2018b).

Istotnie ujemne oszacowania parametru przy zmiennej *Kobieta* są zgodne z rezultatami innych badań poświęconych między innymi problemowi dyskryminacji płacowej kobiet (por. np. Majchrowska, Strawiński, 2018). Osoby zatrudnione na czas nieokreślony zarabiały, przy innych czynnikach niezmienionych, w całym analizowanym okresie więcej w porównaniu z zatrudnionymi na czas określony. Dodatkowo w całym analizowanym okresie sektor prywatny oferował pracownikom lepsze warunki płacowe w porównaniu z sektorem państwowym. Wraz ze wzrostem doświadczenia zawodowego oraz stażu pracy pracownika w badanym przedsiębiorstwie następował wzrost wynagrodzenia. Rezultaty te nie budzą wątpliwości i są zgodne z wynikami innych badań poświęconych czynnikom determinującym wynagrodzenia pojedynczych pracowników w polskiej gospodarce (por. np. Majchrowska, Strawiński, 2016; 2018; Strawiński, Majchrowska, Broniatowska, 2016; Domański, 2018; Florczak, Grabowski, 2018b; Grabowski, 2019).

W tabeli 15 prezentowane są zmieniające się w czasie efekty losowe dla województw.

Predykcje efektów losowych dla poszczególnych województw charakteryzuje niska zmienność w czasie. W przypadku niektórych województw – na przykład wielkopolskiego i śląskiego – mamy do czynienia z istotnie dodatnimi oszacowaniami efektów losowych w całym okresie 2004–2016. Najwyższe dodatnie predykcje efektów losowych dotyczą jednak województwa mazowieckiego. W całym analizowanym okresie kształtują się one między 0,19 a 0,32. W przypadku niektórych regionów mamy do czynienia z negatywnymi oszacowaniami efektów losowych w całym analizowanym okresie. Dotyczy to takich województw jak: kujawsko-pomorskie, lubelskie, łódzkie, podkarpackie, podlaskie, świętokrzyskie czy warmińsko-mazurskie. Negatywne efekty losowe dla

świętokrzyskiego, podkarpackiego czy podlaskiego są szczególnie wysokie co do modułu. Oznacza to, że mieszkańcy tych regionów mogą liczyć na zdecydowanie niższe zarobki, niż wynikałoby to z ich cech społeczno-demograficznych, charakteru umowy o pracę, faktu wykonywania określonego zawodu czy przynależności przedsiębiorstwa do określonej sekcji czy sektora. Uzyskane wyniki są zgodne z podejściem Nowej Geografii Ekonomicznej, wskazującym na dodatnią korelację między płacami a ekonomicznym potencjałem regionu (Cieślik, Rokicki, 2016). Regiony charakteryzujące się dużym poziomem koncentracji mieszkańców w aglomeracjach (np. mazowieckie, pomorskie, śląskie, dolnośląskie) oraz pozytywnymi tendencjami migracyjnymi mogą oferować swoim pracownikom wyższe płace w porównaniu z województwami mniej zurbanizowanymi i charakteryzującymi się ujemnym saldem migracji (podkarpackie, podlaskie, świętokrzyskie). Analizując kształtowanie się efektów losowych w województwie dolnośląskim oraz lubuskim w latach 2008–2016, a także w województwie zachodniopomorskim między 2010 a 2016 rokiem, należy zwrócić uwagę na tendencję rosnącą. Można zatem mówić o poprawiających się warunkach płacowych pracowników z firm zlokalizowanych w województwach graniczących z Niemcami. Może to wynikać z faktu, że firmy zlokalizowane w województwach przygranicznych korzystają ze swojego położenia, zwiększając potencjał rozwojowy, co następnie przekłada się na istotny wzrost relatywnych wynagrodzeń mieszkańców tych regionów.

Tabela 15. Efekty losowe dla województw w poszczególnych latach
Odchylenia standardowe zawarte są w nawiasach

Województwo	2004	2006	2008	2010	2012	2014	2016
dolnośląskie	0,011 (0,004)	0,053 (0,004)	0,015 (0,003)	0,041 (0,003)	0,061 (0,003)	0,063 (0,003)	0,059 (0,003)
kujawsko- -pomorskie	–0,026 (0,005)	–0,004 (0,005)	–0,020 (0,004)	–0,033 (0,004)	–0,025 (0,004)	–0,038 (0,004)	–0,048 (0,004)
lubelskie	–0,082 (0,006)	–0,093 (0,005)	–0,074 (0,004)	–0,064 (0,004)	–0,086 (0,004)	–0,075 (0,004)	–0,077 (0,004)
lubuskie	–0,010 (0,008)	–0,005 (0,007)	–0,017 (0,005)	–0,013 (0,005)	–0,005 (0,005)	0,004 (0,006)	0,015 (0,005)
łódzkie	–0,054 (0,005)	–0,069 (0,004)	–0,025 (0,003)	–0,011 (0,003)	–0,012 (0,003)	–0,007 (0,003)	–0,020 (0,003)
małopolskie	0,024 (0,005)	–0,006 (0,004)	0,008 (0,003)	–0,004 (0,003)	0,002 (0,003)	–0,000 (0,003)	0,020 (0,003)
mazowieckie	0,319 (0,003)	0,255 (0,003)	0,204 (0,002)	0,213 (0,002)	0,197 (0,002)	0,191 (0,002)	0,209 (0,002)
opolskie	–0,015 (0,008)	–0,013 (0,008)	0,006 (0,006)	–0,017 (0,006)	–0,031 (0,006)	–0,014 (0,005)	–0,021 (0,005)
podkarpackie	–0,080 (0,005)	–0,105 (0,005)	–0,086 (0,004)	–0,077 (0,004)	–0,073 (0,004)	–0,090 (0,004)	–0,077 (0,003)

Województwo	2004	2006	2008	2010	2012	2014	2016
podlaskie	–0,128 (0,008)	–0,138 (0,007)	–0,088 (0,006)	–0,031 (0,005)	–0,084 (0,005)	–0,056 (0,005)	–0,067 (0,005)
pomorskie	0,064 (0,005)	0,078 (0,005)	0,059 (0,004)	0,043 (0,004)	0,066 (0,003)	0,052 (0,004)	0,050 (0,003)
śląskie	0,023 (0,003)	0,097 (0,003)	0,083 (0,002)	0,049 (0,002)	0,072 (0,002)	0,051 (0,002)	0,026 (0,002)
świętokrzyskie	–0,059 (0,007)	–0,059 (0,008)	–0,055 (0,005)	–0,053 (0,005)	–0,076 (0,005)	–0,068 (0,005)	–0,086 (0,005)
warmińsko-mazurskie	–0,039 (0,007)	–0,040 (0,006)	–0,038 (0,005)	–0,033 (0,005)	–0,012 (0,005)	–0,040 (0,005)	–0,016 (0,005)
wielkopolskie	0,026 (0,004)	0,029 (0,004)	0,011 (0,003)	0,002 (0,003)	0,009 (0,003)	0,026 (0,002)	0,018 (0,002)
zachodniopomorskie	0,025 (0,006)	0,021 (0,006)	0,016 (0,005)	–0,012 (0,004)	–0,002 (0,004)	0,002 (0,004)	0,015 (0,004)

Źródło: opracowanie własne.

W tabeli 16 prezentowane są efekty losowe dla poszczególnych sekcji PKD w latach 2004–2016.

Tabela 16. Efekty losowe dla sekcji PKD w poszczególnych latach

	2004	2006	2008	2010	2012	2014	2016
Sekcja A Rolnictwo, leśnictwo, łowiectwo i rybactwo	–0,031 (0,008)	–0,035 (0,008)	–0,089 (0,009)	–0,057 (0,008)	–0,033 (0,009)	–0,015 (0,008)	0,014 (0,009)
Sekcja B Górnictwo i wydobywanie	0,476 (0,006)	0,485 (0,005)	0,455 (0,005)	0,336 (0,005)	0,353 (0,005)	0,288 (0,005)	0,157 (0,006)
Sekcja C Przetwórstwo przemysłowe	–0,042 (0,003)	–0,041 (0,002)	–0,050 (0,002)	–0,047 (0,002)	–0,032 (0,002)	–0,028 (0,002)	–0,035 (0,002)
Sekcja D Wytwarzanie i zaopatrywanie w energię elektryczną, gaz, parę wodną, gorącą wodę i powietrze do układów klimatyzacyjnych	0,053 (0,006)	0,042 (0,006)	0,078 (0,006)	0,116 (0,005)	0,134 (0,006)	0,162 (0,006)	0,171 (0,006)
Sekcja E Dostawa wody: gospodarowanie ściekami i odpadami oraz działalność związana z rekultywacją	–0,031 (0,007)	–0,021 (0,006)	–0,086 (0,007)	–0,046 (0,006)	–0,067 (0,006)	–0,073 (0,006)	–0,080 (0,006)
Sekcja F Budownictwo	–0,025 (0,004)	–0,021 (0,004)	0,042 (0,004)	–0,042 (0,004)	–0,036 (0,004)	–0,040 (0,004)	–0,042 (0,004)
Sekcja G Handel hurtowy i detaliczny; Naprawa pojazdów samochodowych, włączając motocykle	–0,041 (0,002)	–0,046 (0,003)	–0,019 (0,003)	–0,024 (0,002)	–0,049 (0,002)	0,002 (0,002)	–0,008 (0,002)

Tabela 16 (cd.)

	2004	2006	2008	2010	2012	2014	2016
Sekcja H Transport i gospodarka magazynowa	–0,055 (0,003)	–0,060 (0,003)	–0,042 (0,004)	–0,036 (0,003)	–0,032 (0,003)	–0,040 (0,003)	–0,051 (0,003)
Sekcja I Działalność związana z zakwaterowaniem i usługami gastronomicznymi	–0,091 (0,008)	–0,103 (0,008)	–0,088 (0,008)	–0,076 (0,008)	–0,085 (0,007)	–0,076 (0,007)	–0,054 (0,007)
Sekcja J Informacja i komunikacja	0,378 (0,005)	0,397 (0,006)	0,253 (0,006)	0,289 (0,006)	0,248 (0,005)	0,252 (0,005)	0,290 (0,005)
Sekcja K Działalność finansowa i ubezpieczeniowa	0,256 (0,004)	0,275 (0,005)	0,285 (0,005)	0,268 (0,004)	0,238 (0,004)	0,208 (0,004)	0,319 (0,004)
Sekcja L Działalność związana z obsługą rynku nieruchomości	–0,094 (0,007)	–0,105 (0,007)	–0,080 (0,007)	–0,083 (0,007)	–0,111 (0,007)	–0,052 (0,007)	–0,071 (0,007)
Sekcja M Działalność profesjonalna, naukowa i techniczna	0,104 (0,005)	0,102 (0,005)	0,088 (0,005)	0,087 (0,005)	0,119 (0,005)	0,118 (0,005)	0,125 (0,005)
Sekcja N Działalność w zakresie usług administrowania i działalność wspierająca	–0,195 (0,005)	–0,231 (0,004)	–0,250 (0,005)	–0,214 (0,004)	–0,147 (0,005)	–0,130 (0,005)	–0,140 (0,004)
Sekcja O Administracja publiczna i obrona narodowa; Obowiązkowe zabezpieczenia społeczne	–0,115 (0,003)	–0,105 (0,003)	–0,096 (0,003)	–0,173 (0,003)	–0,212 (0,003)	–0,225 (0,003)	–0,231 (0,003)
Sekcja P Edukacja	–0,021 (0,002)	–0,023 (0,002)	–0,028 (0,002)	0,094 (0,002)	0,081 (0,002)	0,080 (0,002)	0,023 (0,002)
Sekcja Q Opieka zdrowotna i pomoc społeczna	–0,093 (0,003)	–0,115 (0,003)	–0,112 (0,003)	–0,089 (0,003)	–0,136 (0,002)	–0,181 (0,003)	–0,172 (0,003)
Sekcja R Działalność związana z kulturą, rozrywką i rekreacją	–0,205 (0,007)	–0,197 (0,007)	–0,170 (0,007)	–0,196 (0,007)	–0,239 (0,007)	–0,237 (0,007)	–0,205 (0,006)
Sekcja S Pozostała działalność usługowa	–0,087 (0,015)	–0,095 (0,016)	–0,093 (0,018)	–0,104 (0,018)	0,005 (0,013)	–0,015 (0,014)	–0,011 (0,012)

Źródło: opracowanie własne.

Analizując wyniki zawarte w tabeli 16, należy zwrócić uwagę na wysokie i dodatnie predykcje efektów losowych dla sekcji J (Informacja i komunikacja) oraz K (Działalność finansowa i ubezpieczeniowa). Wynik ten nie budzi wątpliwości, ponieważ przedsiębiorstwa zajmujące się działalnością z zakresu informacji i komunikacji na ogół zatrudniają wysoko wykwalifikowanych pracowników, osiągają

wysokie przychody, mogą więc oferować wysokie zarobki. Wzrastający popyt na specjalistów w branżach informatycznych i finansowych prowadzi do wzrostu wynagrodzeń osób świadczących taką działalność. Relatywnie wysokie wynagrodzenia w polskim górnictwie i wydobywaniu są historycznie uwarunkowane ze względu na silną pozycję związków zawodowych. Była ona znaczna jeszcze w okresie poprzedzającym transformację systemową. Okazuje się, że poziom uzwiązkowienia w polskim górnictwie jest zdecydowanie wyższy niż w innych branżach. Strategia negocjacyjna związków zawodowych w Polsce ma na celu ochronę miejsc pracy i zapewnienie wysokich wynagrodzeń pracowników. Sektorowa analiza ilorazu wynagrodzeń do produktywności pracy wskazuje, że jest ona w górnictwie zdecydowanie wyższa niż w innych branżach (Jonek-Kowalska, 2014). Analizując kształtowanie się efektów losowych w czasie dla sekcji B, J, K, należy zauważyć różnice. Predykcje efektów losowych dla lat 2004–2008 dla sekcji B były zdecydowanie wyższe niż dla sekcji J oraz K. Potem nastąpiło wyrównanie efektów losowych, natomiast w 2016 roku wyniki dla górnictwa i wydobywania były zdecydowanie gorsze niż w przypadku informacji i komunikacji, a także działalności finansowej i ubezpieczeniowej. Wynika to z faktu, że w ostatnich latach prowadzona była polityka wygaszania nierentownych kopalń. Odprawy oraz emerytury pomostowe oferowane były górnikom z dużym stażem (a więc zarabiającym więcej), co przyczyniło się do spadku średniego wynagrodzenia w analizowanej sekcji. Oprócz trzech wyżej wymienionych sekcji PKD należy jeszcze zwrócić uwagę na sekcje D oraz M, w których również odnotowano dodatnie efekty losowe. Ujemne predykcje efektów losowych dotyczą większości spośród pozostałych sekcji PKD. Szczególną uwagę należy zwrócić jednak na następujące sekcje: N (Działalność w zakresie usług administrowania i działalność wspierająca), O (Administracja publiczna i obrona narodowa; Obowiązkowe zabezpieczenia społeczne), Q (Opieka zdrowotna i pomoc społeczna), R (Działalność związana z kulturą, rozrywką i rekreacją). Efekty losowe dla tych sekcji są istotnie ujemne i wysokie co do modułu. Oznacza to zatem, że – przy innych czynnikach niezmienionych – pracownicy przedsiębiorstw należących do sekcji N, O, Q, R mogą liczyć na zdecydowanie niższe wynagrodzenia w porównaniu z osobami pracującymi w innych branżach. Wynik ten również nie budzi wątpliwości. Podmioty zajmujące się działalnością związaną z kulturą, rozrywką i rekreacją, opieką zdrowotną i pomocą społeczną, administracją publiczną czy też działalnością z zakresu obowiązkowego zabezpieczenia społecznego na ogół finansowane są przez budżet państwa i ich pracownicy nie mogą liczyć na wysokie wynagrodzenia.

W dalszej kolejności prezentowane są efekty losowe dla grup ze względu na poziom kwalifikacji (tabela 17).

Tabela 17. Średnie efekty losowe dla grup ze względu na poziom kwalifikacji w poszczególnych latach (w nawiasach podane są odchylenia standardowe)

Poziom kwalifikacji	Rok						
	2004	2006	2008	2010	2012	2014	2016
I	–0,371 (0,142)	–0,243 (0,125)	–0,179 (0,079)	–0,188 (0,087)	–0,184 (0,090)	–0,173 (0,091)	–0,161 (0,087)
II	–0,305 (0,057)	–0,212 (0,052)	–0,153 (0,034)	–0,159 (0,040)	–0,160 (0,042)	–0,161 (0,042)	–0,148 (0,040)
III	0,123 (0,106)	0,079 (0,096)	0,045 (0,063)	0,049 (0,060)	0,035 (0,062)	0,028 (0,062)	0,018 (0,059)
IV	0,553 (0,084)	0,376 (0,075)	0,287 (0,050)	0,298 (0,047)	0,309 (0,049)	0,306 (0,049)	0,290 (0,046)

Źródło: opracowanie własne.

Wyniki zawarte w tabeli 17 wskazują, że pracownicy wykonujący zawody wymagające niskich kwalifikacji mogą liczyć na niższe wynagrodzenia niż reprezentanci zawodów wymagających wysokich umiejętności. Różnice należy interpretować przy innych czynnikach niezmienionych. Tak więc osoby wykonywujące prace proste na ogół zarabiają mniej – zarówno ze względu na niski poziom wykształcenia, jak i fakt przynależności do grupy zawodowej skupiającej profesje wymagające niskiego poziomu kwalifikacji. Natomiast specjaliści, którzy na ogół mają wyższe wykształcenie i wykonują zawody wymagające najwyższego poziomu kwalifikacji, zarabiają zdecydowanie więcej. Należy jednak zauważyć, że zmieniające się w czasie predykcje efektów losowych nie wskazują na poprawność hipotezy SBTC. Gdyby tak było, obserwowany byłby wzrost wartości bezwzględnych dla oszacowań efektów losowych – obserwowany jest jednak ich ciągły spadek. Wynik ten można uzasadnić przez zdecydowany spadek podaży pracy w zawodach charakteryzujących się niskim poziomem wymaganych umiejętności. W ostatnich dwudziestu latach nastąpił zdecydowany spadek zainteresowania szkołami zawodowymi wśród polskich uczniów. Procesowi temu towarzyszył wzrost zainteresowania szkolnictwem wyższym. Ze względu na niskie koszty pracy oraz fakt obecności specjalnych stref ekonomicznych na terenie Polski wiele firm międzynarodowych przenosiło tu swoją produkcję. W rezultacie popyt na nisko wykwalifikowaną siłę roboczą był stabilny, a podaż zmniejszała się. Musiało to doprowadzić do relatywnego wzrostu płac w grupie osób o niskim poziomie kwalifikacji. Uzyskany wynik jest zgodny z rezultatami innych badań, prowadzonych między innymi przez Pawła Strawińskiego, Paulinę Broniatowską i Aleksandrę Majchrowską (2016; 2018). Analiza wpływu poziomu wykształcenia na wysokość wynagrodzeń wskazuje, że chociaż płace osób z wykształceniem zawodowym są niższe w porównaniu z zarobkami osób lepiej wykształconych, różnice te zmniejszają się. Z drugiej strony spadek oszacowanych efektów losowych dla pracowników z grupy III i IV może wynikać z faktu, że za wzrostem podaży osób

charakteryzujących się wysokim poziomem wykształcenia nie nastąpił wzrost popytu na osoby wykonujące zawody wymagające najwyższych kwalifikacji. Dlatego też „premia" za wykonywanie zawodu wymagającego najwyższych kwalifikacji była niższa w 2016 roku w porównaniu z 2004 rokiem.

Innym uzasadnieniem dla wzrostu prognozowanych efektów losowych w 2016 roku w przypadku pracowników wykonujących prace o najniższym poziomie kwalifikacji może być wprowadzenie programu „Rodzina 500+" przez nowy polski rząd wybrany jesienią 2015 roku. Jego rezultatem był między innymi spadek podaży pracy w Polsce. Był on szczególnie widoczny w zawodach wymagających pracy fizycznej i niskich kwalifikacji od pracowników. Spadek podaży pracy w grupie osób o najniższym poziomie kwalifikacji mógł, przy innych czynnikach niezmienionych, bezpośrednio prowadzić do wzrostu prognozowanych efektów losowych dla pracowników z grupy I.

Tabela 18 prezentuje średnie efekty losowe dla trzycyfrowych grup zawodowych w okresie 2004–2016. Należy zwrócić przede wszystkim uwagę na te trzycyfrowe grupy zawodowe, dla których relatywne wynagrodzenia były, przy innych czynnikach niezmienionych, zdecydowanie wyższe lub zdecydowanie niższe

W przypadku grup zawodowych 111 (Przedstawiciele władz publicznych i wyżsi urzędnicy) oraz 121 (Kierownicy do spraw obsługi biznesu i zarządzania) odnotowywane są bardzo wysokie efekty losowe na początku analizowanego okresu. W latach 2010–2016 efekty te są nadal dodatnie, jednak zdecydowanie niższe w porównaniu z okresem 2004–2008. Szczególnie wysokie predykcje efektów losowych w latach 2010–2016 odnotowuje się w przypadku grupy zawodowej 112 (Dyrektorzy generalni i zarządzający). Na nieco niższą, ale również wysoką premię płacową mogą liczyć reprezentanci grupy zawodowej 133 (Kierownicy do spraw technologii informatycznych i telekomunikacyjnych). W latach 2004–2008 na bardzo wysoką premię płacową mogli liczyć pracownicy należący do grupy zawodowej 314 (Technicy nauk biologicznych, rolniczych i technologii żywności), natomiast w okresie 2010–2016 szczególnie wysokimi zarobkami charakteryzowali się przedstawiciele grupy zawodowej 315 (Pracownicy transportu morskiego, żeglugi śródlądowej i lotnictwa (z wyłączeniem sił zbrojnych)). W zdecydowanej większości przypadków oszacowane efekty losowe były ujemne. Należy jednak zwrócić uwagę na kilka grup zawodowych, dla których oszacowania te były wysokie co do modułu. Reprezentanci trzycyfrowych grup zawodowych 262 (Bibliotekoznawcy, archiwiści i muzealnicy), 264 (Literaci, dziennikarze i filolodzy) oraz 265 (Twórcy i artyści) mogli liczyć na zdecydowanie niższe wynagrodzenia niż wynikałoby to z ich cech, charakterystyk podmiotów ich zatrudniających oraz przynależności do grupy ze względu na poziom posiadanych kwalifikacji. Wynik ten nie budzi jednak wątpliwości. Osoby wykonujące te zawody są na ogół zatrudniane przez jednostki budżetowe. Dlatego też wynagrodzenia oferowane im nie są wysokie.

Tabela 18. Średnie efekty losowe dla trzycyfrowych grup zawodowych

Grupa	2004	2006	2008	2010	2012	2014	2016	Grupa	2004	2006	2008	2010	2012	2014	2016
111	1,610	1,304	0,704	0,521	0,590	0,575	0,525	411	0,251	0,135	0,072	0,030	0,015	0,002	0,011
112	–0,159	0,094	–0,067	1,528	1,605	1,609	1,604	412	0,201	0,124	0,096	0,060	0,050	0,056	0,043
121	1,808	1,764	1,422	0,358	0,417	0,454	0,471	413	0,004	–0,042	–0,020	–0,108	–0,090	–0,116	–0,090
122	0,145	0,155	0,249	0,434	0,467	0,537	0,564	414	–0,095	–0,180	–0,029	–	–	–	–
131	–0,119	0,319	0,333	0,257	0,232	0,378	0,209	419	0,183	0,066	0,044	–	–	–	–
132	–	–	–	0,018	0,058	0,047	0,087	421	0,112	0,048	0,094	0,115	0,053	0,018	0,066
133	–	–	–	0,784	0,881	0,842	1,065	422	–0,037	–0,022	–0,012	–0,032	0,000	–0,026	–0,000
134	–	–	–	0,181	0,175	0,114	0,159	431	–	–	–	0,127	0,070	0,073	0,079
141	–	–	–	–0,014	0,049	0,009	0,040	432	–	–	–	–0,040	–0,058	–0,069	–0,065
142	–	–	–	0,009	–0,058	–0,067	–0,037	441	–	–	–	0,005	–0,015	–0,021	–0,024
143	–	–	–	0,236	0,296	0,370	0,290	511	0,250	0,080	0,148	0,146	0,257	0,235	0,178
211	–0,500	–0,527	–0,358	–0,291	–0,243	–0,199	–0,255	512	0,015	–0,012	–0,015	–0,014	–0,027	–0,018	0,002
212	–0,781	–0,551	–0,351	–0,460	–0,373	–0,305	–0,331	513	–0,098	–0,067	–0,045	0,004	–0,021	–0,033	–0,004
213	0,042	0,112	0,080	–0,353	–0,383	–0,306	–0,298	514	0,055	–0,057	0,008	0,003	0,071	0,035	0,020
214	–0,257	–0,246	–0,120	–0,212	–0,182	–0,192	–0,187	515	–0,323	–0,348	–0,281	–0,088	–0,099	–0,068	–0,068
215	–	–	–	–0,102	–0,056	–0,018	0,037	516	–	–	–	0,072	–0,077	0,061	–0,005
216	–	–	–	–0,285	–0,409	–0,333	–0,319	521	0,042	0,148	–0,065	–	–	–	–
221	–0,604	–0,480	–0,399	0,240	0,212	0,127	0,124	522	–0,076	–0,052	–0,039	–0,043	–0,067	–0,045	–0,039
222	–0,507	–0,442	–0,348	–0,304	–0,365	–0,414	–0,295	523	–	–	–	–0,076	–0,096	–0,084	–0,057
223	–0,512	–0,264	0,048	–0,314	–0,398	–0,433	–0,294	524	–	–	–	0,025	0,016	0,018	0,067
224	–0,829	–0,479	–0,304	–0,381	–0,478	–0,392	–0,401	531	–	–	–	0,036	0,017	0,026	0,021
225	–	–	–	–0,133	–0,157	–0,266	–0,234	532	–	–	–	–0,068	–0,121	–0,126	–0,087
226	–	–	–	–0,096	0,047	–0,086	–0,336	541	–	–	–	–0,258	–0,204	–0,205	–0,215
227	–	–	–	–0,340	–0,361	–0,381	–0,313	611	–0,041	–0,081	–0,075	–0,112	–0,062	–0,041	–0,055
228	–	–	–	–0,326	–0,326	–0,339	–0,127	612	–0,007	–0,097	–0,026	–0,023	0,102	–0,035	0,033
229	–	–	–	–	–	–	–0,396	613	–0,166	–0,114	0,025	–0,075	–0,085	–0,038	0,192

231	0,154	–0,130	–0,271	–0,368	–0,390	–0,262	–0,286	621	–0,077	–0,032	–0,068	–0,050	–0,049	–0,044	–0,065
232	0,632	0,551	0,257	0,283	0,380	0,313	0,213	622	–	–	–	–0,135	–0,156	–0,122	–0,115
233	0,606	0,505	0,256	0,485	0,568	0,496	0,420	631	–0,070	0,012	–0,102	–	–	–	–
234	0,956	0,864	0,443	0,454	0,470	0,377	0,294	632	–0,078	–0,091	–0,108	–	–	–	–
235	0,157	0,087	–0,002	0,260	0,253	0,188	0,124	633	–0,313	–0,236	–0,295	–	–	–	–
241	–0,107	–0,055	–0,055	–0,096	–0,124	–0,144	–0,044	711	0,462	0,627	0,630	–0,053	–0,047	–0,041	–0,043
242	0,790	0,554	0,347	–0,218	–0,200	–0,223	–0,195	712	–0,106	–0,054	–0,004	–0,007	–0,012	–0,006	0,015
243	–0,688	–0,599	–0,528	–0,035	–0,077	–0,035	–0,048	713	–0,024	0,016	0,004	–0,017	–0,024	0,004	0,036
244	–0,287	–0,242	–0,183	–0,079	–0,115	–0,090	–0,065	714	–0,058	0,027	0,017	–	–	–	–
245	0,206	–0,522	–0,317	–	–	–	–	721	0,008	0,065	0,118	0,043	0,054	0,063	0,033
246	–1,093	–1,093	–0,571	–	–	–	–	722	–0,056	–0,044	0,006	–0,030	–0,017	–0,019	–0,029
247	–0,314	–0,317	–0,143	–	–	–	–	723	0,012	0,093	0,112	0,076	0,106	0,092	0,038
251	–	–	–	0,148	0,116	0,157	0,295	724	0,034	0,078	0,103	–	–	–	–
252	–	–	–	–0,066	–0,062	–0,019	0,041	725	0,040	–0,007	–0,028	–	–	–	–
261	–	–	–	0,438	0,429	0,456	0,533	731	0,007	–0,002	0,106	–0,076	–0,078	–0,052	–0,056
262	–	–	–	–0,446	–0,479	–0,334	–0,431	732	–0,055	–0,062	–0,098	0,039	0,035	0,039	0,034
263	–	–	–	–0,158	–0,252	–0,245	–0,251	741	–0,147	–0,110	–0,064	0,079	0,097	0,087	0,053
264	–	–	–	–0,380	–0,422	–0,501	–0,586	742	–0,187	–0,159	–0,094	–0,061	–0,046	0,001	0,005
265	–	–	–	–0,307	–0,362	–0,453	–0,391	751	–	–	–	–0,057	–0,063	–0,068	–0,071
311	–0,138	–0,066	–0,023	–0,089	–0,072	–0,084	–0,064	752	–	–	–	–0,121	–0,113	–0,110	–0,125
312	–0,186	–0,192	–0,012	–0,000	0,045	0,013	0,030	753	–	–	–	–0,074	–0,062	–0,062	–0,084
313	0,410	–0,365	–0,124	–0,058	–0,025	–0,047	–0,048	754	–	–	–	0,178	0,092	0,058	0,007
314	2,300	2,105	1,089	0,034	0,043	0,110	0,089	811	0,214	0,336	0,309	0,324	0,367	0,312	0,245
315	–0,346	–0,303	–0,199	1,860	1,755	1,825	1,474	812	0,010	0,097	0,053	0,024	0,006	–0,027	–0,031
321	–0,227	–0,022	–0,060	–0,012	–0,050	–0,155	–0,119	813	–0,014	–0,081	–0,089	0,076	0,114	0,062	0,041
322	–0,323	–0,145	–0,055	–0,254	–0,250	–0,279	–0,229	814	–0,024	–0,100	–0,062	–0,017	–0,026	–0,020	–0,013
323	–0,372	–	–	–0,203	–0,146	–0,150	–0,267	815	0,150	0,122	0,120	–0,105	–0,069	–0,086	–0,067
324	–	–	–	–0,175	–0,205	–0,197	–0,200	816	0,099	0,109	0,074	0,013	–0,013	–0,011	–0,046

Tabela 18 (cd.)

Grupa	2004	2006	2008	2010	2012	2014	2016	Grupa	2004	2006	2008	2010	2012	2014	2016
325	–	–	–	–0,177	–0,176	–0,203	–0,183	817	0,061	–0,054	–0,121	–0,083	–0,078	–0,091	–0,092
331	0,131	0,091	–0,041	–0,046	–0,055	–0,052	0,005	818	–	–	–	–0,028	–0,029	–0,053	–0,045
332	–	–	–	–0,005	–0,014	0,049	0,060	821	–0,039	–0,024	–0,018	–0,055	–0,049	–0,062	–0,075
333	–	–	–	–0,086	–0,086	–0,147	–0,091	831	–0,012	–0,088	0,053	0,031	0,105	0,125	0,134
334	–	–	–	–0,088	–0,125	–0,104	–0,068	832	–0,039	–0,025	–0,024	–0,037	–0,070	–0,062	–0,075
335	–	–	–	–0,154	–0,167	–0,187	–0,153	833	–0,040	–0,001	0,018	–0,018	–0,016	–0,031	–0,023
341	–0,000	0,084	0,056	–0,156	–0,177	–0,190	–0,145	834	0,220	0,134	–0,012	–0,002	–0,027	–0,031	–0,043
342	–0,092	–0,129	–0,058	–0,160	–0,139	–0,101	–0,121	835	–	–	–	0,120	0,087	0,186	0,107
343	–0,080	–0,079	–0,041	–0,188	–0,212	–0,191	–0,169	911	–	–	–0,058	–0,038	–0,059	–0,047	–0,047
344	–0,117	–0,110	–0,058	–	–	–	–	912	–0,143	–	0,016	–0,027	–0,028	–0,015	–0,063
345	–	–	0,115	–	–	–	–	921	–0,035	–0,065	–0,017	0,004	–0,016	–0,044	–0,049
346	–0,238	–0,251	–0,154	–	–	–	–	931	0,001	0,071	0,128	0,061	0,048	0,008	0,005
347	–0,293	–0,168	–0,132	–	–	–	–	932	–0,046	–0,057	–0,015	–0,038	–0,041	–0,070	–0,077
348	–0,266	–0,274	–0,223	–	–	–	–	933	–0,043	–0,065	–	–0,059	–0,059	–0,081	–0,069
351	–	–	–	–0,035	–0,082	–0,028	0,071	941	–	–	–	0,007	–0,015	–0,003	0,005
352	–	–	–	–0,133	–0,072	–0,129	–0,121	951	–	–	–	–0,196	–0,139	–0,047	–
								961	–	–	–	–0,051	–0,025	–0,024	–0,030
								962	–	–	–	–0,065	–0,075	–0,073	–0,052

Źródło: opracowanie własne.

Następnie oszacowane zostały parametry modelu (3.45), w którym grupy ze względu na poziom kwalifikacji zastąpione są przez grupy zadaniowe. Oszacowania parametrów związanych z efektami stałymi są podobne do oszacowań uzyskanych dla modelu (3.44). Nie są one zatem prezentowane w niniejszej monografii. Podobna sytuacja dotyczy oszacowań efektów losowych związanych z trzycyfrowymi grupami zawodowymi. Prezentowane i omawiane są natomiast predykcje efektów losowych dla grup zadaniowych (por. tabela 19).

Tabela 19. Predykcje efektów losowych dla grup zadaniowych w poszczególnych latach (w nawiasach podane są odchylenia standardowe)

Grupa zadaniowa	Rok						
	2004	2006	2008	2010	2012	2014	2016
NRUI	0,44 (0,09)	0,45 (0,08)	0,31 (0,07)	0,35 (0,06)	0,35 (0,07)	0,39 (0,06)	0,39 (0,05)
NRUA	0,03 (0,14)	0,05 (0,12)	–0,04 (0,07)	–0,00 (0,07)	0,03 (0,06)	0,07 (0,06)	0,11 (0,05)
NRUIA	0,07 (0,18)	0,09 (0,11)	–0,06 (0,10)	0,02 (0,09)	0,01 (0,08)	–0,01 (0,08)	–0,02 (0,09)
NRF	0,03 (0,11)	0,04 (0,10)	0,02 (0,09)	0,02 (0,07)	0,01 (0,06)	0,01 (0,07)	0,05 (0,07)
RU	–0,20 (0,12)	–0,22 (0,10)	–0,21 (0,09)	–0,17 (0,08)	–0,23 (0,08)	–0,22 (0,07)	–0,26 (0,08)
RF	–0,37 (0,05)	–0,33 (0,06)	–0,23 (0,05)	–0,22 (0,04)	–0,19 (0,04)	–0,18 (0,04)	–0,18 (0,04)

Źródło: opracowanie własne.

Pozytywne wartości dla prognozowanych efektów losowych w całym okresie odnotowuje się w przypadku pracowników wykonujących prace nierutynowe, umysłowe, interpersonalne oraz przedstawicieli zawodów nierutynowych i fizycznych. Wysokie i istotne oszacowania w przypadku przedstawicieli grupy zadaniowej NRUI wskazują, że umiejętności miękkie (interpersonalne) są szczególnie ważne na polskim rynku pracy. Współczesne teorie dotyczące rynków pracy dowodzą, że rośnie znaczenie tego typu umiejętności (por. Pellegrino, Hilton, 2012).

Negatywna „premia płacowa" odnotowywana jest w przypadku pracowników wykonujących rutynowe prace umysłowe. Wynik ten jest zgodny między innymi z wynikami badań uzyskanymi przez Aleksandrę Partekę (2018) i Łukasza Arendta (2018). Uzasadnieniem tego zjawiska może być postępująca informatyzacja w finansach i bankowości. Popyt na pracowników wykonujących proste prace w bankach, firmach ubezpieczeniowych czy innych instytucjach finansowych zmniejszył się w ostatnich latach. Osoby te charakteryzują się na ogół średnim poziomem kwalifikacji i przeważnie nie uzyskują wysokich wynagrodzeń. Kolejnym

z wyjaśnień tego zjawiska może być fakt, że w ostatnich latach w Polsce obserwowany był gwałtowny wzrost liczby firm offshoringowych. Umieszczenie tego typu przedsiębiorstw na terytorium Polski miało na celu redukcję kosztów przez korporacje transnarodowe. Prace w tych firmach na ogół miały charakter rutynowy, wykonywane były jednak przez osoby charakteryzujące się wyższym wykształceniem. Wynagrodzenia pracowników w firmach offshoringowych były jednak niskie.

W przypadku pracowników wykonujących zadania rutynowe i manualne obserwowana jest negatywna „premia płacowa" w całym analizowanym okresie. Należy zwrócić jednak uwagę, że z okresu na okres staje się ona coraz wyższa (niższa co do modułu). Obniżenie „negatywnej" premii płacowej może być związane z gwałtownym pogorszeniem się systemu kształcenia zawodowego, które nastąpiło wraz z reformą edukacji narodowej z 1998 roku. Obniżyła się podaż pracowników wykonujących rutynowe manualne czynności, natomiast popyt na ich usługi pozostał stabilny, co mogło wynikać z faktu, iż międzynarodowe korporacje lokowały zakłady produkcyjne w Polsce w celu redukcji kosztów (por. Parteka, 2018). Ponieważ czynności rutynowe i manualne są bardzo często wykonywane przez osoby z wykształceniem zawodowym, analiza kształtowania się ich płac w dużej części wyjaśnia obserwowane tendencje. Jak wskazują między innymi Strawiński, Broniatowska i Majchrowska (2016; 2018), w ostatnich latach nastąpiło zmniejszenie różnicy między wynagrodzeniami osób z wykształceniem zawodowym a otrzymywanymi przez pracowników z wykształceniem średnim ogólnokształcącym. Polepszenie się sytuacji osób wykonujących prace fizyczne rutynowe względem pracowników w zawodach z grupy RU może także wynikać z faktu wzrostu aglomeracji i migracji do największych polskich miast. Zgodnie z koncepcją autorstwa Acceturo, Dalmazzo i de Blasio (2014) obserwowane w ostatnich latach tendencje wzrostu liczby profesjonalistów pracujących w przedsiębiorstwach zlokalizowanych w największych polskich aglomeracji (warszawska, krakowska, wrocławska, poznańska, trójmiejska) sprawiają, że wzrasta popyt na usługi proste wykonywane przez pracowników o najniższych kwalifikacjach.

Wyniki przedstawione w tabeli 19 są częściowo zgodne z hipotezą polaryzacji. Niemniej jednak, jak wskazują Arendt i Grabowski (2018), zjawisko polaryzacji na polskim rynku pracy nie może być w pełni wyjaśnione postępem technologicznym, tak jak w przypadku Kanady (Green, Sand, 2015). Oprócz tego uzyskane wyniki nie wskazują na trwały wzrost nierówności płacowych. Wynik ten jest zgodny z rezultatem otrzymanym przez Philippa Hühnego i Dierka Herzera (2017). Jak wskazują wspomniani autorzy, odpowiednia polityka edukacyjna, mająca na celu wyrównanie szans społecznych, jak również złagodzenie wymogów egzaminacyjnych podczas rekrutacji na wyższe uczelnie, może prowadzić do redukcji różnicy między wynagrodzeniami osób wykwalifikowanych i niewykwalifikowanych.

W celu sprawdzenia, czy uwzględnienie zmiennych kontekstowych (związanych z regionami) poprawiło jakość uzyskanych wyników, porównywane były błędy standardowe reszt dla następujących czterech modeli:

1) modelu pełnego,
2) modelu nieuwzględniającego efektów losowych związanych z przynależnością firm zatrudniających pracowników do województw oraz zmiennych regionalnych,
3) modelu uwzględniającego efekty losowe związane z przynależnością firm zatrudniających pracowników do województw, lecz nieuwzględniającego zmiennych regionalnych,
4) modelu nieuwzględniającego efektów losowych związanych z przynależnością firm zatrudniających pracowników do województw, lecz uwzględniającego zmienne regionalne.

Tabela 20 zawiera średnie błędy szacunku dla wszystkich czterech modeli.

Tabela 20. Średnie błędy szacunku dla czterech porównywanych modeli

Model	Średni błąd szacunku
Pełny	0,0035
Nieuwzględniający efektów losowych związanych z przynależnością firm zatrudniających pracowników do województw oraz zmiennych regionalnych	0,0057
Uwzględniający efekty losowe związane z przynależnością firm zatrudniających pracowników do województw, lecz nieuwzględniający zmiennych regionalnych	0,0046
Nieuwzględniający efektów losowych związanych z przynależnością firm zatrudniających pracowników do województw, lecz uwzględniający zmienne regionalne	0,0043

Źródło: opracowanie własne.

Uzyskane rezultaty wskazują, że nieuwzględnienie efektów losowych lub zmiennych regionalnych prowadzi do znaczącego pogorszenia jakości dopasowania modelu do danych empirycznych. Okazuje się jednak, że pogorszenie jakości dopasowania jest silniejsze, jeśli w modelu nieuwzględniane są zmienne regionalne.

4. Uogólnione liniowe modele wielopoziomowe

4.1. Postać uogólnionego liniowego modelu wielopoziomowego

W niniejszym podrozdziale rozważane są uogólnione modele liniowe rozszerzone o obecność efektów losowych. Formuła określająca wielopoziomowy uogólniony model liniowy przyjmuje następującą postać:

$$g_l\left\{E\left(\boldsymbol{y}\mid\boldsymbol{u}\right)\right\}=\boldsymbol{X\beta}+\boldsymbol{Zu}\,,\ \ \boldsymbol{y}\sim F\,, \tag{4.1}$$

gdzie $\boldsymbol{y}$ jest I×1-wymiarowym wektorem obserwacji pochodzących z rozkładu o dystrybuancie F, $\boldsymbol{X}$ jest macierzą obserwacji na zmiennych objaśniających o wymiarach I×K, β jest K×1-wymiarowym wektorem parametrów, $\boldsymbol{Z}$ jest I×QQ-wymiarową macierzą składającą się ze zmiennych binarnych oraz ewentualnie kategorii, których wpływ na zmienną wynikową losowo różni się między klastrami. Wektor $\boldsymbol{u}$ ma wymiar QQ × 1 i zawiera efekty losowe, $g_l\left(\cdot\right)$ jest funkcją łączącą. Posiada ona własność odwracalności i spełniona jest następująca równość:

$$E\left(\boldsymbol{y}\mid\boldsymbol{u}\right)=g_l^{-1}\left(\boldsymbol{X\beta}+\boldsymbol{Zu}\right). \tag{4.2}$$

Przyjmując różne założenia dotyczące dystrybuanty F oraz funkcji łączącej $g_l\left(\cdot\right)$, uzyskuje się poszczególne warianty wielopoziomowych uogólnionych modeli liniowych. Są one prezentowane w tabeli 21.

Tabela 21. Rodzaje wielopoziomowych uogólnionych modeli liniowych

Funkcja łącząca gl(*)	Dystrybuanta F	Model
Funkcja logistyczna $g_l\left(x\right)=\frac{\exp\left(x\right)}{1+\exp\left(x\right)}$	Rozkład zero-jedynkowy	Wielopoziomowy model logitowy

Tabela 21 (cd.)

Funkcja łącząca gl(*)	Dystrybuanta F	Model
Dystrybuanta rozkładu normalnego $g_l(x) = \Phi(x)$	Rozkład zero-jedynkowy	Wielopoziomowy model probitowy
Dystrybuanta komplementarnego rozkładu log-log $g_l(x) = 1 - \exp(-\exp(-x))$	Rozkład zero-jedynkowy	Wielopoziomowy komplementarny model log-log
Funkcja logistyczna $g_l(x) = \frac{\exp(x)}{1+\exp(x)}$	Rozkład dyskretny (skokowy) dla zmiennej losowej mierzonej na skali porządkowej	Wielopoziomowy model logitowy kategorii uporządkowanych
Dystrybuanta rozkładu normalnego $g_l(x) = \Phi(x)$	Rozkład dyskretny (skokowy) dla zmiennej losowej mierzonej na skali porządkowej	Wielopoziomowy model probitowy kategorii uporządkowanych
$g_l(x) = \ln(x)$	Rozkład Poissona	Wielopoziomowy model regresji Poissona
$g_l(x) = \ln(x)$	Rozkład ujemny dwumianowy	Wielopoziomowy model ujemny dwumianowy

Źródło: opracowanie własne.

W kolejnych podrozdziałach omawiane są metody estymacji parametrów uogólnionych liniowych modeli wielopoziomowych.

4.2. Funkcja wiarygodności w uogólnionym liniowym modelu wielopoziomowym

Analizując dotychczasową wiedzę na temat estymacji parametrów uogólnionych liniowych modeli wielopoziomowych, należy wyodrębnić dwie główne grupy metod. Pierwsza obejmuje metody aproksymacyjne. W drugiej maksymalizacja funkcji wiarygodności uogólnionych liniowych modeli wielopoziomowych odbywa się za pomocą metod symulacyjnych. Analizowane grupy metod zostaną szczegółowo omówione w podrozdziałach 4.3 oraz 4.4.

W niniejszym podrozdziale prezentowane są funkcje wiarygodności w wielopoziomowym modelu logitowym, probitowym, komplementarnym log-log, uporządkowanym probitowym, uporządkowanym logitowym, Poissona i ujemnym dwumianowym. W tym celu wyprowadzane są funkcje gęstości warunkowej rozkładu $\boldsymbol{y}$ względem $\boldsymbol{u}$. Przyjmuje się założenie, że $j = 1, ..., J$ indeksuje klastry, a $i = 1, ..., I_j$ indeksuje kolejne obiekty w j-tym klastrze.

Na początku rozważmy dwupoziomowy model probitowy. W takim przypadku rozkład warunkowy wektora obserwacji na zmiennej zależnej ze względu na efekty losowe (dla *j*-tego klastra) wynosi:

$$f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right) = \prod_{i=1}^{I_j}\left[\left\{\Phi\left(\eta_{ij}\right)\right\}^{y_{ij}}\left\{1-\Phi\left(\eta_{ij}\right)\right\}^{1-y_{ij}}\right] = \\ = \exp\left(\sum_{i=1}^{I_j}\left[y_{ij}\ln\left\{\Phi\left(\eta_{ij}\right)\right\} - \left(1-y_{ij}\right)\ln\left\{\Phi\left(-\eta_{ij}\right)\right\}\right]\right), \tag{4.3}$$

gdzie $\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}$. Wyrażenie (4.3) można alternatywnie zapisać w następującej formie macierzowej:

$$f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right) = \exp\left[\boldsymbol{y}_j^T \ln\left\{\Phi\left(\boldsymbol{\eta}_j\right)\right\} - \left(1_{I_j\times 1} - \boldsymbol{y}_j\right)^T \ln\left\{\Phi\left(-\boldsymbol{\eta}_j\right)\right\}\right], \tag{4.4}$$[1]

gdzie wektor $\boldsymbol{\eta}_j$ zawiera elementy η_{ij}.

Przyjmując założenie, że rozkład wektora $\boldsymbol{u}$ jest wielowymiarowym rozkładem normalnym o wartości oczekiwanej 0 i macierzy wariancji-kowariancji $\boldsymbol{\Omega}$, wkład jednostek należących do *j*-tego klastra do funkcji wiarygodności wynosi:

$$L_j\left(\boldsymbol{\beta},\boldsymbol{\Omega}\right) = \left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int f\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right)\exp\left(\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2\right)d\boldsymbol{u} = \\ = \left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int\exp\left(h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right)\right)d\boldsymbol{u}, \tag{4.5}$$

gdzie:

$$h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right) = \boldsymbol{y}_j^T\ln\left\{\Phi\left(\boldsymbol{\eta}_j\right)\right\} - \left(1_{I_j\times 1} - \boldsymbol{y}_j\right)^T\ln\left\{\Phi\left(-\boldsymbol{\eta}_j\right)\right\} - \boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2,$$

natomiast $\boldsymbol{u}_{\{j\}}$ jest wektorem zawierającym efekty losowe dla *j*-tego klastra oraz wartości 0 na miejscu efektów losowych dla pozostałych klastrów.

---

1 W analizowanym podrozdziale przyjmowane jest założenie, że funkcja od wektora jest kolumnowym wektorem składającym się z pojedynczych funkcji. Tak więc na przykład w przypadku analizowanego wzoru *i*-ty element wektora y_j jest mnożony przez $\Phi(\eta_{ij})$.

W przypadku dwupoziomowego modelu logitowego warunkowa funkcja gęstości dla *j*-tego klastra $f_{wj}(\boldsymbol{y}_j|\boldsymbol{u}_j)$ przyjmuje postać:

$$\begin{aligned} f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right) &= \prod_{i=1}^{I_j}\left[\left\{\Lambda\left(\eta_{ij}\right)\right\}^{y_{ij}}\left\{1-\Lambda\left(\eta_{ij}\right)\right\}^{1-y_{ij}}\right] = \\ &= \exp\left(\sum_{i=1}^{I_j}\left[y_{ij}\ln\left\{\Lambda\left(\eta_{ij}\right)\right\}+\left(1-y_{ij}\right)\ln\left\{1-\Lambda\left(\eta_{ij}\right)\right\}\right]\right), \end{aligned} \tag{4.6}$$

gdzie $\boldsymbol{\eta}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}$. Wzór (4.6) w notacji macierzowej przyjmuje postać:

$$f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right) = \exp\left[\boldsymbol{y}_j^T \ln\left\{\Lambda\left(\boldsymbol{\eta}_j\right)\right\} + \left(1_{I_j\times 1} - \boldsymbol{y}_j\right)^T \ln\left\{1-\Lambda\left(\boldsymbol{\eta}_j\right)\right\}\right], \tag{4.7}$$

Wkład jednostek należących do *j*-tego klastra do funkcji wiarygodności wynosi:

$$L_j\left(\boldsymbol{\beta},\boldsymbol{\Omega}\right) = \left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int \exp\left(h\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right)\right)d\boldsymbol{u}_j, \tag{4.8}$$

gdzie:

$$h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right) = \boldsymbol{y}_j^T \ln\left\{\Lambda\left(\boldsymbol{\eta}_j\right)\right\} - \left(1_{I_j\times 1} - \boldsymbol{y}_j\right)^T \ln\left\{1-\Lambda\left(\boldsymbol{\eta}_j\right)\right\} - \boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}} / 2 .$$

W przypadku dwupoziomowego komplementarnego modelu log-log postać funkcji $f_{wj}(\boldsymbol{y}_j|\boldsymbol{u}_j)$ jest następująca:

$$\begin{aligned} f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right) &= \prod_{i=1}^{I_j}\left[\left\{\mathrm{H}\left(\boldsymbol{\eta}_{ij}\right)\right\}^{y_{ij}}\left\{1-\mathrm{H}\left(\boldsymbol{\eta}_{ij}\right)\right\}^{r_{ij}-y_{ij}}\right] = \\ &= \exp\left(\sum_{i=1}^{I_j}\left[y_{ij}\ln\left\{\mathrm{H}\left(\boldsymbol{\eta}_{ij}\right)\right\}+\left(1-y_{ij}\right)\left\{1-\ln\left\{\mathrm{H}\left(\boldsymbol{\eta}_{ij}\right)\right\}\right\}\right]\right), \end{aligned} \tag{4.9}$$

gdzie $\boldsymbol{\eta}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}$ oraz $H(v) = 1 - \exp\{-\exp(v)\}$. Wzór (4.9) można zapisać w notacji macierzowej w następujący sposób:

$$f_{wj}(\boldsymbol{y}_j \mid \boldsymbol{u}_j) = \exp\left[\boldsymbol{y}_j^T \ln\{\mathrm{H}(\boldsymbol{\eta}_j)\} - \left(1_{I_j \times 1} - \boldsymbol{y}_j\right)^T \ln\{1 - \mathrm{H}(\boldsymbol{\eta}_j)\}\right], \quad (4.10)$$

Wkład jednostek należących do *j*-tego klastra do funkcji wiarygodności wynosi:

$$L_j(\boldsymbol{\beta}, \boldsymbol{\Omega}) = (2\pi)^{-\frac{QQ}{2}} \left(det(\boldsymbol{\Omega})\right)^{-\frac{1}{2}} \int \exp\left(h_{wj}\left(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{u}_{\{j\}}\right)\right) d\boldsymbol{u}_j, \quad (4.11)$$

gdzie:

$$h_{wj}\left(\boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{u}_{\{j\}}\right) = \boldsymbol{y}_j^T \ln\{\mathrm{H}(\boldsymbol{\eta}_j)\} - \left(1_{I_j \times 1} - \boldsymbol{y}_j\right)^T \ln\{1 - \mathrm{H}(\boldsymbol{\eta}_j)\} - \boldsymbol{u}_{\{j\}}^T \boldsymbol{\Omega}^{-1} \boldsymbol{u}_{\{j\}} / 2.$$

Kolejną klasą modeli wielopoziomowych zmiennych jakościowych są uporządkowane modele z efektami losowymi. Dodatkowy wektor parametrów w analizowanych modelach składa się z progów. Oznaczmy go $\boldsymbol{\tau}$. Przy założeniu logistycznego rozkładu składnika losowego ε_{ij} rozkład warunkowy wektora obserwacji na zmiennej zależnej ze względu na efekty losowe (dla *j*-tego klastra) przedstawia się następująco:

$$f_{wj}(\boldsymbol{y}_j \mid \boldsymbol{\kappa}, \boldsymbol{u}_j) = \prod_{i=1}^{I} p_{ij}(p)^{\ddot{I}_p(y_{ij})} = \exp \sum_{i=1}^{I_j} \left\{ \ddot{I}_p(y_{ij}) \ln\left(p_{ij}(p)\right)\right\}, \quad (4.12)$$

gdzie:

$$\ddot{I}_p(y_{ij}) = \begin{cases} 1 & \text{gdy } y_{ij} = p, \\ 0 & \text{w przec. przypadku,} \end{cases}$$

natomiast $p_{ij}(p)$ oznacza prawdopodobieństwo, że zmienna uporządkowana y dla i-tej jednostki z j-tego klastra przyjmie wartość p. Wkład jednostek należących do j-tego klastra do funkcji wiarygodności wynosi:

$$L_j\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{\kappa}\right)=$$

$$=\left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int f_{wj}\left(\boldsymbol{y}_j\mid\boldsymbol{u}_j,\boldsymbol{\kappa}\right)\exp\left(\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2\right)d\boldsymbol{u}_j= \tag{4.13}$$

$$=\left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int\exp\left(h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}},\boldsymbol{\kappa}\right)\right)d\boldsymbol{u}_j,$$

gdzie:

$$h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}},\boldsymbol{\kappa}\right)=\sum_{i=1}^{I_j}\left\{\ddot{I}_p\left(y_{ij}\right)\ln\left(p_{ij}\left(p\right)\right)\right\}-\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2.$$

Prawdopodobieństwo $p_{ij}(p)$, w zależności od tego, czy rozkład składnika losowego jest normalny, czy logistyczny, obliczane jest zgodnie ze wzorem:

$$p_{ij}\left(p\right)=\frac{1}{1+\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}-\tau_p\right)}-\frac{1}{1+\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_j-\tau_{p-1}\right)}, \tag{4.14}$$

lub

$$p_{ij}\left(p\right)=\Phi\left(-\boldsymbol{x}_{ij}\boldsymbol{\beta}-\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}+\tau_p\right)-\Phi\left(-\boldsymbol{x}_{ij}\boldsymbol{\beta}-\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}+\tau_{p-1}\right), \tag{4.15}$$

gdzie $\tau_0=-\infty$, $\tau_P=+\infty$, a P jest maksymalną wartością przyjmowaną przez zmienną wielomianową kategorii uporządkowanych.

Dla dwupoziomowego modelu Poissona rozkład warunkowy wektora obserwacji na zmiennej zależnej ze względu na efekty losowe (dla j-tego klastra) przedstawia się następująco:

$$f_{wj}\left(\boldsymbol{y}_j\mid\boldsymbol{u}_j\right)=$$

$$=\prod_{i=1}^{I_j}\left[\left\{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}\right)\right\}^{y_{ij}}\exp\left\{-\left\{\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}\right)\right\}\right\}/\,y_{ij}!\right]= \tag{4.16}$$

$$=\exp\left[\sum_{i=1}^{I_j}\left\{y_{ij}\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}\right)-\exp\left(\boldsymbol{x}_{ij}\boldsymbol{\beta}+\boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}\right)-\ln\left(y_{ij}!\right)\right\}\right].$$

Oznaczając $c\left(\boldsymbol{y}_j\right)=\sum_{i=1}^{I_j}\ln\left(y_{ij}!\right)$ jako funkcję niezależną od parametrów modelu, wyrażenie (4.16) może zostać inaczej zapisane z wykorzystaniem notacji macierzowej:

$$\begin{aligned} &f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right)= \\ &=\exp\left\{\boldsymbol{y}_j^T\left(\boldsymbol{X}_j\boldsymbol{\beta}+\boldsymbol{Z}_j\boldsymbol{u}_{\{j\}}\right)-1^T\exp\left(\boldsymbol{X}_j\boldsymbol{\beta}+\boldsymbol{Z}_j\boldsymbol{u}_{\{j\}}\right)-\ c\left(\boldsymbol{y}_j\right)\right\}, \end{aligned} \tag{4.17}$$

gdzie macierze $\boldsymbol{X}_j$ oraz $\boldsymbol{Z}_j$ powstają w wyniku ustawienia wektorów $\boldsymbol{x}_{ij}$ oraz $\boldsymbol{z}_{ij}$ w stos.

Wkład jednostek należących do *j*-tego klastra do funkcji wiarygodności wynosi:

$$\begin{aligned} &L_j\left(\boldsymbol{\beta},\boldsymbol{\Omega}\right)= \\ &=\left(2\pi\right)^{-\frac{QQ}{2}}det\left(\boldsymbol{\Omega}\right)^{-\frac{1}{2}}\int f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j\right)\exp\left(\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2\right)d\boldsymbol{u}_j= \\ &=\exp\left\{-\ c\left(\boldsymbol{y}_j\right)\right\}\left(2\pi\right)^{-\frac{IQ}{2}}det\left(\boldsymbol{\Omega}\right)^{-\frac{1}{2}}\int\exp\left(h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right)\right)d\boldsymbol{u}_j, \end{aligned} \tag{4.18}$$

gdzie:

$$h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}}\right)=\boldsymbol{y}_j^T\left(\boldsymbol{X}_j\boldsymbol{\beta}+\boldsymbol{Z}_j\boldsymbol{u}_{\{j\}}\right)-1^T\exp\left(\boldsymbol{X}_j\boldsymbol{\beta}+\boldsymbol{Z}_j\boldsymbol{u}_{\{j\}}\right)-\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2\,.$$

W przypadku dwupoziomowego modelu ujemnego dwumianowego z warunkowym parametrem nadmiernej dyspersji α rozkład warunkowy wektora obserwacji na zmiennej zależnej ze względu na efekty losowe (dla *j*-tego klastra) przedstawia się następująco:

$$\begin{aligned} &f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j,\alpha\right)=\prod_{i=1}^{I_j}\left\{\frac{\Gamma\left(y_{ij}+r\right)}{\Gamma\left(y_{ij}+r\right)\Gamma\left(r\right)}\,p_{ij}^r\left(1-p_{ij}\right)^{y_{ij}}\right\}= \\ &=\exp\left[\sum_{i=1}^{I_j}\left\{\ln\Gamma\left(y_{ij}+r\right)-\ln\Gamma\left(y_{ij}+1\right)-\ln\Gamma\left(r\right)+c\left(y_{ij},\alpha\right)\right\}\right], \end{aligned} \tag{4.19}$$

gdzie:

$$c\left(y_{ij},\alpha\right)=-\frac{1}{\alpha}\ln\left\{1+\exp\left(\eta_{ij}+\ln\left(\alpha\right)\right)\right\}-y_{ij}\ln\left\{1+\exp\left(-\eta_{ij}-\ln\left(\alpha\right)\right)\right\},$$

a także $r = \frac{1}{\alpha}$, $p_{ij} = \frac{1}{\left(1+\alpha\mu_{ij}\right)}$ oraz $\eta_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{u}_{\{j\}}$.

Dla wariantu dwupoziomowego modelu ujemnego dwumianowego ze stałym parametrem nadmiernej dyspersji δ rozkład warunkowy $\boldsymbol{y}_j$ dany jest wzorem:

$$\begin{aligned} f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j, \delta\right) &= \prod_{i=1}^{I_j}\left\{\frac{\Gamma\left(y_{ij}+r\right)}{\Gamma\left(y_{ij}+1\right)\Gamma\left(r_{ij}\right)} p^{r_{ij}}\left(1-p\right)^{y_{ij}}\right\} = \\ &= \exp\left[\sum_{i=1}^{I_j}\left\{\ln\Gamma\left(y_{ij}+r_{ij}\right) - \ln\Gamma\left(y_{ij}+1\right) - \ln\Gamma\left(r_{ij}\right) + c\left(y_{ij},\delta\right)\right\}\right], \end{aligned} \tag{4.20}$$

gdzie:

$$c\left(y_{ij},\delta\right) = -\left(\frac{\mu_{ij}}{\delta} + y_{ij}\right)\ln\left(1+\delta\right),$$

a także $r_{ij} = \frac{\mu_{ij}}{\delta}$ oraz $p = \frac{1}{1+\delta}$.

Ponieważ $\boldsymbol{u}_j$ pochodzi z wielowymiarowego rozkładu normalnego o wartości oczekiwanej 0 i macierzy wariancji-kowariancji $\boldsymbol{\Omega}$, wkład jednostek należących do *j*-tego klastra do funkcji wiarygodności wynosi:

$$\begin{aligned} L_j\left(\boldsymbol{\beta},\boldsymbol{\Omega},\gamma\right) &= \\ = \left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int f_{wj}\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j,\gamma\right)\exp\left(\boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2\right)d\boldsymbol{u}_j &= \\ = \left(2\pi\right)^{-\frac{QQ}{2}}\left(det\left(\boldsymbol{\Omega}\right)\right)^{-\frac{1}{2}}\int\exp\left(h_{wj}\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_j,\gamma\right)\right)d\boldsymbol{u}_j, \end{aligned} \tag{4.21}$$

gdzie:

$$h\left(\boldsymbol{\beta},\boldsymbol{\Omega},\boldsymbol{u}_{\{j\}},\gamma\right) = f\left(\boldsymbol{y}_j \mid \boldsymbol{u}_j,\gamma\right) - \boldsymbol{u}_{\{j\}}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}_{\{j\}}/2.$$

Jak pokazują między innymi Searle, Casella i McCulloch (1992), całkowita wartości logarytmu funkcji wiarygodności jest sumą logarytmów funkcji wiarygodności obliczonych dla poszczególnych klastrów. Dlatego też w przypadku każdego wielopoziomowego modelu zmiennych jakościowych prawdziwa jest równość:

$$\ln L\left(\boldsymbol{\Theta}\right) = \sum_{j=1}^{J}\ln L_j\left(\boldsymbol{\Theta}\right), \tag{4.22}$$

gdzie $\boldsymbol{\Theta}$ jest wektorem wszystkich parametrów podlegających estymacji. Na przykład w przypadku wielopoziomowego modelu logitowego wektor ten obejmuje parametry wchodzące w skład wektora $\boldsymbol{\beta}$ oraz macierzy $\boldsymbol{\Omega}$. Dla wielopoziomowego modelu ujemnego dwumianowego analizowany wektor zawiera dodatkowo parametr γ.

4.3. Estymacja parametrów uogólnionych liniowych modeli wielopoziomowych za pomocą metod aproksymacyjnych

Poszczególne komponenty funkcji wiarygodności związane z odpowiednimi klastrami zawierają całki wielokrotne. Istnieją dwie główne grupy metod obliczania tych całek. Po pierwsze, są to metody aproksymacyjne. Oprócz nich należy zwrócić uwagę na wykorzystanie metod symulacyjnych. W niniejszym podrozdziale prezentowane są te metody maksymalizacji funkcji wiarygodności, które należą do pierwszej z grup. Ponieważ celem monografii nie jest omawianie numerycznych metod obliczania całek wielokrotnych, niezbędne metody aproksymacji są jedynie zapowiadane. Szerzej piszą o nich na przykład Zenon Fortuna, Bohdan Macukow i Janusz Wąsowski (2017) oraz Jerzy Klamka i Zbigniew Ogonowski (2015).

Metoda maksymalizacji funkcji wiarygodności wykorzystująca standardową oraz adaptacyjną kwadraturę Gaussa-Hermite'a została dokładnie opisana między innymi w pracy Sophii Rabe-Hesketh, Andersa Skrondala i Andrew Picklesa (2005). Jej zastosowanie prezentowane jest na początku dla modelu wielopoziomowego zawierającego efekty losowe związane z wyrazem wolnym. Oznacza to zatem, że dla każdego klastra mamy pojedynczy efekt losowy u_j. Na podstawie rozważań z podrozdziału 4.3 należy wnioskować, że wkład j-tego klastra do funkcji wiarygodności w modelu z losowym wyrazem wolnym może zostać ogólnie opisany wzorem:

$$L_j\left(\boldsymbol{\Theta}\right) = \int g\left(u_j; 0; \sigma^2\right) \prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid u_j\right) du_j, \tag{4.23}$$

gdzie $\boldsymbol{\Theta}$ jest wektorem wszystkich parametrów podlegających estymacji, $g(u_j;0;\sigma^2)$ oznacza funkcję gęstości rozkładu normalnego o zerowej wartości oczekiwanej i wariancji σ^2, natomiast $L_{ij}(\boldsymbol{\Theta}|u_j)$ oznacza wkład do funkcji wiarygodności dla i-tej jednostki z j-tego klastra, warunkowy ze względu na efekt losowy u_j. Na przykład w przypadku wielopoziomowego modelu probitowego:

$$L_{ij}\left(\boldsymbol{\Theta} \mid u_j\right) = y_{ij}\Phi\left(\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_j\right) + \left(1 - y_{ij}\right)\Phi\left(-\boldsymbol{x}_{ij}\boldsymbol{\beta} - u_j\right). \tag{4.24}$$

Zastosowanie kwadratury Gaussa-Hermite'a w celu wyznaczenia całki wielokrotnej w wyrażeniu (4.23) proponowali J.S. Butler oraz Robert Moffitt (1982). Zamiast całkowania względem u_j proponuje się wykonanie całkowania względem $v_j = \frac{u_j}{\sigma}$, wykorzystując funkcję gęstości standardowego wielowymiarowego rozkładu normalnego $\phi(\cdot)$. Wówczas równanie (4.23) można alternatywnie zapisać następująco:

$$L_j(\boldsymbol{\Theta}) = \int \phi(v_j) \prod_{i=1}^{I} L_{ij}(\boldsymbol{\Theta} \mid v_j) dv_j . \tag{4.25}$$

Zaproponowana w pracy Butlera oraz Moffitta (1982) standardowa aproksymacja Gaussa-Hermite'a przyjmuje postać:

$$L_j(\boldsymbol{\Theta}) = \int \phi(v_j) \prod_{i=1}^{I_j} L_{ij}(\boldsymbol{\Theta} \mid v_j) dv_j \approx \sum_{rr=1}^{RR} pp_{rr} \prod_{i=1}^{I_j} L_{ij}(\boldsymbol{\Theta} \mid aa_{rr}), \tag{4.26}$$

gdzie pp_{rr} oraz aa_{rr} są odpowiednio wagami oraz węzłami RR-wymiarowej kwadratury Gaussa dla całek postaci $\int exp(-x) f(x) dx$.

Omówiona metoda wykorzystująca standardową kwadraturę Gaussa-Hermite'a była wielokrotnie stosowana w badaniach empirycznych opartych na modelach wielopoziomowych (por. np. Blundell, Windmeijer, 1997; Rice, Jones, 1997; Cardoso, 2000; Carey, 2000; Nelder, Pawitan, Lee, 2006). Okazało się, że zastosowanie kwadratury Gaussa-Hermite'a prowadzi do uzyskania właściwych oszacowań w przypadku niewielkich liczebności w poszczególnych klastrach. Jednak w przypadku gdy wykorzystywane są bazy danych o dużej liczbie obserwacji (np. dane pochodzące z badania aktywności ekonomicznej ludności lub badania struktury wynagrodzeń), a liczba klastrów nie jest zbyt duża, pojawia się problem obciążoności estymatora. Badania symulacyjne wskazujące na analizowany problem zostały przeprowadzone między innymi przez Geogre'a Borjasa i Glenna Sueyoshiego (1994), Lee (2000), Paula Alberta i Deana Follmanna (2000), a także Emmanuela Lesaffre'a i Barta Spiessensa (2001).

Rozwiązaniem problemu związanego z obciążonością estymatora MNW w przypadku stosowania aproksymacji za pomocą kwadratury Gaussa-Hermite'a jest zastosowanie jej adaptacyjnej wersji. Podejście to zostało zaproponowane między innymi w pracach José Pinheiro i Douglasa Batesa (1995), a także Rabe-Hesketh, Skrondala i Picklesa (2005). Podejście to oparte jest na propozycji Johna Naylora i Adriana Smitha (1982), która była pierwotnie wykorzystywana w statystyce

bayesowskiej podczas obliczania gęstości *a posteriori* za pomocą całkowania numerycznego. Polega ono na skalowaniu i zamienianiu węzłów, tak aby znalazły się pod maksimum funkcji podcałkowej.

Punktem wyjścia do zastosowania zmodyfikowanej metody przybliżania całki danej wzorem (4.25) jest obserwacja, że funkcję podcałkową można interpretować jako iloczyn gęstości *a priori* zmiennej losowej v_j oraz łącznego prawdopodobieństwa realizacji warunkowego ze względu na v_j. Dlatego też po normalizacji względem v_j uzyskuje się gęstość *a posteriori* dla efektu losowego v_j warunkową ze względu na wartości zmiennych obserwowalnych. Jeśli klastry składają się z dużej liczby obserwacji, rozkłady brzegowe są zbieżne do gęstości rozkładu normalnego, co wynika z bayesowskiej wersji centralnego twierdzenia granicznego (por. Carlin, Louis, 2000). Zakładając, że $\hat{\mu}_j$ oraz $\hat{\tau}_j^2$ są odpowiednio wartością oczekiwaną i wariancją dla gęstości *a posteriori*, wyrażenie (4.25) można inaczej zapisać:

$$L_j\left(\boldsymbol{\Theta}\right) = \int g\left(v_j; \hat{\mu}_j, \hat{\tau}_j^2\right)\left(\frac{\phi\left(v_j\right)\prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid \boldsymbol{v}_j\right)}{g\left(v_j; \hat{\mu}_j, \hat{\tau}_j^2\right)}\right) dv_j \,. \tag{4.27}$$

Po zamianie całkowanej zmiennej $z_j = \frac{\left(v_j - \hat{\mu}_j\right)}{\hat{\tau}_j}$, a następnie zastosowaniu standardowej kwadratury, otrzymujemy:

$$L_j\left(\boldsymbol{\Theta}\right) \approx \sum_{rr=1}^{RR} \pi_{jrr} \prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid \ \alpha\alpha_{jrr}\right), \tag{4.28}$$

gdzie:

$$\alpha\alpha_{jrr} = \hat{\mu}_j + \hat{\tau}_j aa_{rr} \,,$$

$$\pi\pi_{jrr} = \sqrt{2\pi}\hat{\tau}_j exp\left(\frac{aa_{rr}^2}{2}\right)\phi\left(\hat{\mu}_j + \hat{\tau}_j a_{rr}\right) pp_{rr} \,.$$

$\alpha\alpha_{rr}$ oraz $\pi\pi_{jrr}$ są odpowiednio przekształconymi węzłami oraz przekształconymi wagami kwadratury.

Wykorzystując wartości startowe dla średniej i odchylenia standardowego *a posteriori* $\hat{\mu}_j^{\{0\}}$ oraz $\hat{\tau}_j^{\{0\}}$, definiowane są wartości startowe dla parametrów

wykorzystywanych w formule (4.28), czyli $\alpha\alpha_{jrr}^{\{0\}}$ oraz $\pi\pi_{jrr}^{\{0\}}$. Wkład *j*-tego klastra do funkcji wiarygodności w pierwszej iteracji obliczany jest zgodnie ze wzorem:

$$L_j\left(\boldsymbol{\Theta}\right)^{\{1\}} = \sum_{rr=1}^{RR} \pi\pi_{jrr}^{\{0\}} \prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid \ \alpha\alpha_{jrr}^{\{0\}}\right). \tag{4.29}$$

Następnie średnia i odchylenie standardowe *a posteriori* aktualizowane są na podstawie formuł:

$$\hat{\mu}_j^{\{1\}} = \frac{\sum_{rr=1}^{RR}\left(\alpha\alpha_{jrr}^{\{0\}}\right)\pi\pi_{jrr}^{0}\prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid \ \alpha\alpha_{jrr}^{\{0\}}\right)}{L_j^1\left(\boldsymbol{\theta}\right)}, \tag{4.30a}$$

$$\hat{\tau}_j^{\{1\}} = \sqrt{-\left(\hat{\mu}_j^{\{1\}}\right)^2 + \frac{\sum_{rr=1}^{RR}\left(\alpha\alpha_{jrr}^{\{0\}}\right)^2 \pi\pi_{jrr}^{\{0\}}\prod_{i=1}^{I_j} L_{ij}\left(\boldsymbol{\Theta} \mid \ \alpha\alpha_{jrr}^{\{0\}}\right)}{L_j\left(\boldsymbol{\Theta}\right)^{\{1\}}}}. \tag{4.30b}$$

W dalszej kolejności $\pi\pi_{jrr}^{\{1\}}$ oraz $\alpha\alpha_{jrr}^{\{1\}}$ są wykorzystywane w celu wyznaczenia $L_j(\Theta)^{\{1\}}$. Iteracyjna procedura jest kontynuowana aż do osiągnięcia zbieżności, czyli do momentu, kiedy różnice między oszacowaniami uzyskanymi w dwóch kolejnych krokach są bardzo niskie. Zbieżność analizowanego algorytmu została pokazana między innymi w pracy Rabe-Hesketh, Skrondala i Picklesa (2005).

Omówiona dotychczas metoda maksymalizacji funkcji wiarygodności, wykorzystująca standardową i adaptacyjną kwadraturę Gaussa-Hermite'a, dotyczy przypadku modelu dwupoziomowego, w którym występuje pojedynczy efekt losowy związany z wyrazem wolnym. W takiej wersji analizowana metoda została opisana w pracy Pinheiro i Batesa (1995). Ogólny model wielopoziomowy zakładający występowanie *S* poziomów zagnieżdżenia oraz uwzględniający obecność efektów losowych związanych ze zmiennymi objaśniającymi został przedstawiony w pracy Rabe-Hesketh, Skrondala i Picklesa (2005). W ogólnym modelu wielopoziomowym wyrażenie $\boldsymbol{x}_{ij}\boldsymbol{\beta} + u_j$ wchodzące w skład wzoru (4.24) jest zastępowane następująco:

$$\boldsymbol{\eta} = \boldsymbol{x}\boldsymbol{\beta} + \sum_{ss=2}^{SS} \boldsymbol{x}^{(ss)}\boldsymbol{u}^{(ss)}. \tag{4.31}$$

Wzór definiujący funkcję wiarygodności (4.22) można zapisać następująco:

$$\ln L(\boldsymbol{\Theta}) = \sum_{i} \ln L_i^{(SS)}(\boldsymbol{\Theta}), \tag{4.32}$$

gdzie $\ln L_i^{(SS)}(\boldsymbol{\Theta})$ należy interpretować jako wkład jednostki na najwyższym poziomie zagnieżdżenia *SS* do funkcji wiarygodności. Definiując $\boldsymbol{U}^{(ss)} = \left[\left(\boldsymbol{u}^{(ss)}\right)^T \quad \cdots \quad \left(\boldsymbol{u}^{(SS)}\right)^T \right]^T$ dla $ss \leq SS$ oraz przyjmując założenie o niezależności między jednostkami na poziomie *ss* – 1 przy danych efektach losowych $\boldsymbol{U}^{(ss)}$ na poziomach *ss* i wyższych, wkład (niezlogarytmowany) jednostki na poziomie *ss* do funkcji wiarygodności uzyskuje się rekursywnie w następujący sposób:

$$\begin{aligned} L_i^{(ss)}\left(\boldsymbol{\Theta} \mid \boldsymbol{U}^{(ss+1)}\right) = \\ = \int g\left(\boldsymbol{u}^{(ss)}; 0, \boldsymbol{\Omega}^{(ss)}\right) \prod L_i^{(ss-1)}\left(\boldsymbol{\Theta} \mid \boldsymbol{U}^{(ss+1)}\right) d\boldsymbol{u}^{(ss)}, \; ss = \\ = 2, \ldots, SS-1, \end{aligned} \tag{4.33a}$$

$$L_i^{(SS)}(\boldsymbol{\Theta}) = \int g\left(\boldsymbol{u}^{(SS)}; 0, \boldsymbol{\Omega}^{(SS)}\right) \prod L_i^{(SS-1)}\left(\boldsymbol{\Theta} \mid \boldsymbol{u}^{(SS)}\right) d\boldsymbol{u}^{(SS)}, \tag{4.33b}$$

gdzie $L_i^{(1)}\left(\boldsymbol{\Theta} \mid \boldsymbol{U}^{(2)}\right)$ jest wkładem do funkcji wiarygodności na poziomie indywidualnym, $g_{uu}\left(\boldsymbol{u}^{(ss)}; 0, \boldsymbol{\Omega}^{(ss)}\right)$ jest funkcją gęstości dla wektora losowego $\boldsymbol{u}^{(ss)}$ pochodzącego z wielowymiarowego rozkładu normalnego o zerowych wartościach oczekiwanych i macierzy kowariancji $\boldsymbol{\Omega}^{(ss)}$. Mnożenie we wzorach (4.33a)–(4.33b) odbywa się po wszystkich jednostkach na poziomie *ss* – 1 znajdujących się „wewnątrz" jednostki na poziomie *s*. Całkowanie odbywa się ze względu na niezależne zmienne losowe o standardowym rozkładzie normalnym $\boldsymbol{v}^{(ss)}$, które z efektami losowymi związane są w następujący sposób:

$$\boldsymbol{u}^{(ss)} = \tilde{\boldsymbol{Q}}^{(ss)} \boldsymbol{v}^{(ss)}, \tag{4.34}$$

gdzie $\tilde{\boldsymbol{Q}}^{(ss)}$ jest dekompozycją Choleskiego macierzy $\Omega^{(ss)}$. Iloczynowa kwadratura Gaussa-Hermite'a przyjmuje postać:

$$L^{(ss)}\left(\boldsymbol{\Theta}\,|\,\boldsymbol{V}^{(ss+1)}\right)=$$

$$=\int\phi\left(v_{M(ss)}\right)\ldots\int\phi\left(v_1\right)\prod L^{(ss-1)}\left(\boldsymbol{\Theta}\,|\,v_1,\ldots,v_{\ddot{M}(ss)},\boldsymbol{V}^{(ss+1)}\right)dv_1\ldots dv_{\ddot{M}(ss)}\approx$$

$$\approx\sum_{rr_{\ddot{M}}}pp_{rr_{\ddot{M}}}\ldots\sum_{r_{r1}}pp_{rr_1}\prod L^{(ss-1)}\left(\boldsymbol{\Theta}\,|\,aa_{rr1},\ldots,aa_{rr\ddot{M}},\boldsymbol{V}^{(ss+1)}\right),\tag{4.35}$$

gdzie $\boldsymbol{V}^{(ss)}=\left[\left(\boldsymbol{v}^{(ss)}\right)^T\quad\cdots\quad\left(\boldsymbol{v}^{(SS)}\right)^T\right]^T$, a $\ddot{\mathrm{M}}(ss)$ oznacza liczbę efektów losowych na poziomie *ss*.

W celu zastosowania adaptacyjnej kwadratury Gaussa-Hermite'a efekty losowe $v_1,\ldots,v_{\ddot{\mathrm{M}}(ss)}$ transformowane są w nowe zmienne losowe charakteryzujące się zerową korelacją *a posteriori*, zgodnie z propozycją Naylora i Smitha (1988):

$$w_1=v_1,\tag{4.36a}$$

$$w_{ss}=v_{ss}+\sum_{o=1}^{ss-1}\left(-\frac{cov\left(v_{ss},w_o\right)}{var\left(w_o\right)}\right)w_t,\ ss=2,\ldots,SS,\tag{4.36b}$$

gdzie $SS=\sum_{ss}\ddot{\mathrm{M}}(ss)$.

Punktem wyjścia w procedurze iteracyjnej są efekty losowe z_s, charakteryzujące się zerową wartością oczekiwaną, jednostkową wariancją i zerową kowariancją w rozkładzie *a posteriori*. Są one szacowane w węzłach kwadratury Gaussa-Hermite'a aa_{rr}, $rr = 1, \ldots, RR$. Następnie dokonuje się transformacji efektów losowych zgodnie ze wzorem:

$$w_{ss}=\hat{\mu}_{ss}+\hat{\tau}_{ss}z_{ss},\tag{4.37}$$

oraz przekształcenia ich w efekty losowe v_{ss} zgodnie z wzorami (4.36a)–(4.36b), a także oblicza się $\alpha\alpha_{ssrr}$ zgodnie z następującą formułą:

$$\alpha\alpha_{ssrr}=\hat{\mu}_{ss}+\hat{\tau}_{ss}aa_{rr}.\tag{4.38}$$

Węzły oraz wagi dla adaptacyjnej kwadratury oblicza się odpowiednio na podstawie wzorów:

$$A_{ssrr} = \alpha\alpha_{ssrr} - \sum_{o=1}^{ss-1}\left(-\frac{cov\left(v_{ss}, w_o\right)}{var\left(w_o\right)}\right)\alpha\alpha_{orr} \tag{4.39}$$

oraz

$$PP_{ssrr} = \sqrt{2\pi}\hat{\tau}_{ss}\exp\left(\frac{\left(aa_{rr}\right)^2}{2}\right)\phi\left(A_{ssrr}\right)pp_{rr}\,. \tag{4.40}$$

Wagi należy uszeregować od tych na najwyższym poziomie do tych na najniższym. Sposób uszeregowania wag związanych z efektami losowymi z tego samego poziomu nie ma znaczenia. Elementy $\frac{cov\left(v_{ss}, w_o\right)}{var\left(w_o\right)}$, które są niezbędne zarówno przy wykonywaniu transformacji (4.36a)–(4.36b), jak również do obliczania momentów *a posteriori* dla efektów losowych *w*, są wyznaczane na podstawie średnich, wariancji i kowariancji *a posteriori* dla efektów losowych *v*. Wariancje i kowariancje *a posteriori* dla efektów losowych *v* na różnych poziomach zagnieżdżenia wyznaczane są na podstawie wzoru:

$$cov\left(v_j^{(ss)}v_{j'}^{(ss')}\right) = E\left[v_j^{(ss)}v_{j'}^{(ss')}\right] - E\left[v_j^{(ss)}\right]E\left[v_{j'}^{(ss')}\right]. \tag{4.41}$$

Wkład *ss*-tego poziomu do funkcji wiarygodności obliczany jest za pomocą formuły:

$$L^{(ss)}\left(\boldsymbol{\Theta}\,|\,\boldsymbol{V}^{(ss+1)}\right) =$$

$$= \sum_{rr_{\ddot{M}^{(ss)}}} PP_{\ddot{M}^{(ss)}rr_{\ddot{M}^{(ss)}}} \ldots \sum_{rr_1} PP_{1_{rr_1}} \prod L^{(ss-1)}\left(\boldsymbol{\Theta}\,|\,A_{1rr1},\ldots,A_{\ddot{M}^{(ss)}rr\ddot{M}^{(ss)}},\boldsymbol{V}^{(ss+1)}\right). \tag{4.42}$$

Zaproponowana przez Roberta Wedderburna (1974) metoda quasi-największej wiarygodności polega na maksymalizacji funkcji o tych samych własnościach co standardowa funkcja wiarygodności. Nazwa tej metody wynika z faktu, że funkcja quasi-największej wiarygodności nie odpowiada rozważanemu rozkładowi prawdopodobieństwa. Inne aproksymacyjne metody estymacji parametrów uogólnionych liniowych modeli wielopoziomowych są oparte właśnie na funkcji quasi-największej wiarygodności. Dlatego też w dalszej części podrozdziału definiowane

będą postacie tej funkcji. W związku z tym definiowana będzie funkcja quasi-największej wiarygodności przez $q\ln L(\Theta)$. Norman Breslow oraz David Clayton (1993) zaproponowali szacowanie parametrów hierarchicznego uogólnionego modelu liniowego na podstawie całkowanej funkcji quasi-największej wiarygodności, zdefiniowanej w następujący sposób:

$$\exp\left\{q\ln\left(L\left(\boldsymbol{\Theta}\right)\right)\right\} \propto det\left(\boldsymbol{\Omega}\right)^{-1/2} \int \exp\left[-\frac{1}{2\varphi}\sum_{i=1}^{I} d_i\left(y_i, \mu_i^{\boldsymbol{u}}\right) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{\Omega}^{-1}\boldsymbol{u}\right] d\boldsymbol{u}, \quad (4.43)$$

gdzie:

$$d_i\left(y_i, \mu_i^{\boldsymbol{u}}\right) = -2\int_{y_i}^{\mu_i^{\boldsymbol{u}}} \frac{y_i - s}{a_i \nu\nu(s)} ds,$$

$$\mu_i^{\boldsymbol{u}} = E\left(y_i \mid \boldsymbol{u}\right),$$

natomiast parametry φ, a_i oraz funkcja $\nu\nu(s)$ pochodzą z równania wariancji następującej postaci:

$$var\left(y_i \mid \boldsymbol{u}\right) = \varphi a_i \nu\nu\left(\mu_i^{\boldsymbol{u}}\right).$$

Ponieważ równanie (4.43) ma następującą formę:

$$\hat{r} = c det\left(\boldsymbol{\Omega}\right)^{-1/2} \int \exp\left(-\pi\left(\boldsymbol{u}\right)\right) d\boldsymbol{u}, \quad (4.44)$$

proponuje się wykorzystanie metody Laplace'a w celu aproksymacji wartości całki. Aproksymowana wartość logarytmu funkcji wiarygodności danej wzorem (4.43) wynosi:

$$q\ln\left(L\left(\boldsymbol{\Theta}\right)\right) \approx -\frac{1}{2}\ln\left(det\left(\boldsymbol{\Omega}\right)\right) - \frac{1}{2}\ln\left(det\left(\kappa\kappa^{''}\left(\tilde{\boldsymbol{u}}\right)\right)\right) - \kappa\kappa\left(\tilde{\boldsymbol{u}}\right), \quad (4.45)$$

gdzie $\tilde{\boldsymbol{u}} = \tilde{\boldsymbol{u}}(\theta)$ minimalizuje $\kappa\kappa\left(\boldsymbol{u}\right)$, czyli jest rozwiązaniem następującego równania:

$$\kappa\kappa^{'(\boldsymbol{u})} = -\sum_{i=1}^{I} \frac{\left(y_i - \mu_i^{\boldsymbol{u}}\right)\mathbf{z}_i}{\varphi a_i v v\left(\mu_i^{\boldsymbol{u}}\right) g_l^{'}\left(\mu_i^{\boldsymbol{u}}\right)} + \boldsymbol{\Omega}^{-1}\boldsymbol{u} = 0\,, \tag{4.46}$$

gdzie $g_l(\cdot)$ jest funkcją łączącą uogólnionego liniowego modelu hierarchicznego. W zależności od wyboru rodzaju uogólnionego liniowego modelu wielopoziomowego (np. hierarchiczny model logitowy, uporządkowany hierarchiczny model probitowy, hierarchiczny model licznikowy) mamy różne postacie równania (4.46). Macierz drugich pochodnych $\kappa\kappa^{''}(\boldsymbol{u})$ obecna we wzorze (4.45) przyjmuje postać:

$$\kappa\kappa^{''}(\boldsymbol{u}) = \sum_{i=1}^{I} \frac{\mathbf{z}_i\,\mathbf{z}_i^T}{\varphi a_i\, v v\left(\mu_i^{\boldsymbol{u}}\right)\left[g_l^{'}\left(\mu_i^{\boldsymbol{u}}\right)\right]^2} + \boldsymbol{\Omega}^{-1} + \boldsymbol{R} \approx \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{\Omega}^{-1}\,, \tag{4.47}$$

gdzie macierz $\boldsymbol{W}$ jest macierzą diagonalną, której i-ty diagonalny element wynosi:

$$w_i = \left\{\varphi a_i v\left(\mu_i^{\boldsymbol{u}}\right)\left[g_l^{'}\left(\mu_i^{\boldsymbol{u}}\right)\right]^2\right\}^{-1}.$$

Macierz $\boldsymbol{R}$, która przyjmuje następującą postać:

$$\boldsymbol{R} = -\sum_{\boldsymbol{i}=1}^{\boldsymbol{I}}\left(y_{\boldsymbol{i}} - \mu_{\boldsymbol{i}}^{\boldsymbol{u}}\right)\mathbf{z}_{\boldsymbol{i}}\frac{\partial}{\partial \boldsymbol{u}}\left[\frac{1}{\varphi a_i \boldsymbol{v}\left(\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{u}}\right) g_l^{'}\left(\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{u}}\right)}\right], \tag{4.48}$$

charakteryzuje się tym, że jej wszystkie elementy mają zerową wartość oczekiwaną oraz niższy rząd (*order of magnitude*) niż pozostałe składniki równania (4.47). Dlatego też łącząc ze sobą równania (4.45) oraz (4.47), a także ignorując macierz $\boldsymbol{R}$, uzyskuje się następujące przybliżenie dla logarytmu funkcji quasi-największej wiarygodności:

$$\mathrm{qlnL}(\boldsymbol{\Theta}) \approx -\frac{1}{2}\ln\left(det\left(\mathbf{I} + \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Omega}\right)\right) - \frac{1}{2\varphi}\sum_{i=1}^{I} d_i\left(y_i, \mu_i^{\tilde{\boldsymbol{u}}}\right) - \frac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{\Omega}^{-1}\tilde{\boldsymbol{u}}\,, \tag{4.49}$$

gdzie $\tilde{\boldsymbol{u}}$ maksymalizuje sumę dwóch ostatnich składników wyrażenia (4.49). Ponieważ iteracyjne wagi uogólnionego modelu liniowego zmieniają się powoli (por.

Breslow, Clayton, 1993), ignoruje się pierwszy składnik wyrażenia (4.53) i $\boldsymbol{\beta}$ jest wybierane tak, aby jego pozostała część była jak najwyższa. Dlatego też maksymalizowana jest funkcja pseudo-quasi-największej wiarygodności, której postać jest następująca (por. Green, 1987):

$$\mathrm{qlnL}(\boldsymbol{\Theta}) = -\frac{1}{2\varphi}\sum_{i=1}^{I} d_i\left(y_i, \mu_i^{\tilde{\boldsymbol{u}}}\right) - \frac{1}{2}\tilde{\boldsymbol{u}}^T\boldsymbol{\Omega}^{-1}\tilde{\boldsymbol{u}}\,. \tag{4.50}$$

Różniczkując wyrażenie (4.20) ze względu na parametry $\boldsymbol{\beta}$ oraz $\boldsymbol{u}$ i przyrównując pierwsze pochodne do zera, uzyskujemy:

$$\frac{\partial q\mathrm{ln}L}{\partial \hat{\boldsymbol{a}}} = \sum_{i=1}^{I} \frac{\left(y_i - \mu_i^{\tilde{\boldsymbol{u}}}\right)\boldsymbol{x}_i^T}{\varphi a_i v\left(\mu_i^{\tilde{\boldsymbol{u}}}\right) g_l^{'}\left(\mu_i^{\tilde{\boldsymbol{u}}}\right)} = 0 \tag{4.51a}$$

oraz

$$\frac{\partial q\mathrm{ln}L}{\partial \boldsymbol{u}} = \sum_{i=1}^{I} \frac{\left(y_i - \mu_i^{\tilde{\boldsymbol{u}}}\right)\boldsymbol{z}_i^T}{\varphi a_i v\left(\mu_i^{\tilde{\boldsymbol{u}}}\right) g_l^{'}\left(\mu_i^{\tilde{\boldsymbol{u}}}\right)} = \boldsymbol{\Omega}^{-1}\boldsymbol{u}\,. \tag{4.51b}$$

Jedną z metod rozwiązania układu równań (4.51a)–(4.51b) jest zastosowanie algorytmu scoringowego Fishera. Podejście to zostało zaproponowane w pracy Petera Greena (1987). Polega ono na wykorzystaniu iteracyjnej ważonej metody najmniejszych kwadratów, gdzie zmienna zależna oraz macierz wag zmieniają się w każdym kroku algorytmu. Ponieważ komponenty wektora $\boldsymbol{y}$ można zdefiniować następująco:

$$y_i = g_l^{'}\left(\mu_i^{\boldsymbol{u}}\right) + \left(y_i - \mu_i^{\boldsymbol{u}}\right) g_l^{'}\left(\mu_i^{\boldsymbol{u}}\right), \tag{4.52}$$

rozwiązanie układu równań (4.55a)–(4.55b) za pomocą algorytmu scoringowego Fishera jest następujące:

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Omega} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{I} + \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\boldsymbol{\Omega} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{y} \end{bmatrix}, \tag{4.53}$$

gdzie zależność między zdefiniowanym powyżej wektorem efektów losowych $\boldsymbol{u}$ a wektorem $\boldsymbol{v}$ ze wzoru (4.53) wygląda następująco: $\boldsymbol{u} = \boldsymbol{\Omega v}$. Dlatego też estymator wektora parametrów $\boldsymbol{\beta}$ wyznacza się na podstawie równania:

$$\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}\,, \tag{4.54}$$

gdzie: $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{Z\Omega Z}^T$. Predykcji efektów losowych dokonuje się na podstawie formuły:

$$\hat{\boldsymbol{u}} = \hat{\boldsymbol{\Omega}}\boldsymbol{Z}^T\hat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right). \tag{4.55}$$

W celu wnioskowania o istotności poszczególnych zmiennych wchodzących w skład macierzy $\boldsymbol{X}$ wykorzystuje się aproksymowany estymator macierzy wariancji-kowariancji następującej postaci:

$$E\left(\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T\right) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}. \tag{4.56}$$

Alternatywne aproksymacje logarytmu funkcji wiarygodności w uogólnionym liniowym modelu wielopoziomowym zostały zaproponowane między innymi w pracy Patricii Solomon i Davida Coxa (1992). Zaprezentowali oni postacie funkcji wiarygodności wykorzystujące rozwinięcie w szereg Maclaurina. Rozważany był uogólniony liniowy model wielopoziomowy z pojedynczym efektem losowym związanym z wyrazem wolnym. Aproksymowana funkcja wiarygodności przyjmuje postać:

$$q\ln L_{S1}(\boldsymbol{\Theta}) = \sum_{j=1}^{J}\left\{\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^0 - \frac{1}{2}\ln\left(1 - det(\boldsymbol{\Omega})\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^2\right) + \right.$$
$$\left. + \frac{det(\boldsymbol{\Omega})\left(\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^1\right)^2}{2\left(1 - det(\boldsymbol{\Omega})\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^2\right)}\right\}, \tag{4.57}$$

gdzie $\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^k$ definiowane jest następująco $\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^k = \dfrac{\partial^k \ln L_j(\boldsymbol{\Theta})}{\partial u_j^k}$ i obliczane dla $u_j = 0$. Druga aproksymacja zaproponowana w pracy Solomona i Coxa (1992) wykorzystuje trzecią i czwartą potęgę u_j. Aproksymowana funkcja wiarygodności przyjmuje postać:

$$q\ln L_{S2}(\boldsymbol{\theta}) = q\ln L_{S1}(\boldsymbol{\theta}) + \frac{\left(det(\boldsymbol{\Omega})\right)^2}{2}\sum_{j=1}^{IQ}\left(\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^1\,\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^3 + \frac{1}{4}\widetilde{\ln L}_j(\boldsymbol{\Theta})_0^4\right). \tag{4.58}$$

Zaproponowane przez Solomona i Coxa (1992) estymatory maksymalizują wartości funkcji quasi-największej wiarygodności dane wzorami (4.57) oraz (4.58).

Jedna z metod aproksymacji funkcji wiarygodności wykorzystywana w celu estymacji parametrów logitowego modelu wielopoziomowego zaproponowana została przez Nicholasa Longforda (1987; 1994). Wartość logarytmu funkcji wiarygodności aproksymowana jest przy wykorzystaniu rozwinięcia w szereg Taylora drugiego rzędu wokół wektora $\boldsymbol{u} = 0$:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\Omega}) \approx \ln L(\boldsymbol{\beta} \mid \boldsymbol{u}) + (\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{\mu}_0)^T \boldsymbol{Z}\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T\left(\boldsymbol{Z}^T\boldsymbol{W}_0\boldsymbol{Z}\right)\boldsymbol{u}, \tag{4.59}$$

gdzie i-ty element wektora kolumnowego $\boldsymbol{\mu}\boldsymbol{\mu}_0$ wyznacza się w następujący sposób $\mu\mu_{0,i} = \dfrac{\exp(\boldsymbol{x}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i\boldsymbol{\beta})}$, natomiast $\boldsymbol{W}_0$ jest macierzą diagonalną, której i-ty diagonalny element zdefiniowany jest następująco: $\mu\mu_{0,i}(1 - \mu\mu_{0,i})$. Po dokonaniu pewnych przekształceń uzyskuje się następującą aproksymację logarytmicznej funkcji wiarygodności:

$$\ln L(\boldsymbol{\beta}, \boldsymbol{\Omega}) \approx \ln L(\boldsymbol{\beta} \mid 0) - \frac{1}{2}\ln\left|\hat{\boldsymbol{G}}\right| + \frac{1}{2}\breve{\mathbf{e}}^T\left(\boldsymbol{W}_0 - \boldsymbol{V}_0^{-1}\right)\breve{\mathbf{e}}, \tag{4.60}$$

gdzie:

$$\hat{\boldsymbol{G}} = \boldsymbol{I} + \left(\boldsymbol{Z}^T\boldsymbol{W}_0\right)\left(\boldsymbol{Z}\boldsymbol{\Omega}\right), \tag{4.61}$$

$$\breve{\mathbf{e}} = \boldsymbol{W}_0^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{\mu}_0\right), \tag{4.62}$$

$$\boldsymbol{V}_0 = \boldsymbol{Z}\boldsymbol{\Omega}\boldsymbol{Z}^T + \left(\boldsymbol{W}_0\right)^{-1}. \tag{4.63}$$

Wektor pierwszych pochodnych i macierz drugich pochodnych dla aproksymowanej funkcji wiarygodności (4.60) oraz przy zignorowaniu faktu, że elementy macierzy $\boldsymbol{W}_0$ zależą od parametrów $\boldsymbol{\beta}$, przyjmują postać:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} \approx \boldsymbol{X}^T \left(\boldsymbol{V}_0\right)^{-1} \breve{\mathbf{e}}, \tag{4.64}$$

$$\frac{\partial^2 \ln L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \approx -\boldsymbol{X}^T \left(\boldsymbol{V}_0\right)^{-1} \boldsymbol{X} \ . \tag{4.65}$$

Pierwsza i druga pochodna aproksymowanej funkcji wiarygodności (4.60) względem parametru ϖ z macierzy $\boldsymbol{\Omega}$ są następujące:

$$\frac{\partial \ln L}{\partial \varpi} = \frac{1}{2}\left\{\breve{\mathbf{e}}^T \boldsymbol{V}_0^{-1} \frac{\partial \boldsymbol{V}_0}{\partial \varpi} \boldsymbol{V}_0^{-1} \breve{\mathbf{e}} - tr\left(\boldsymbol{V}_0^{-1} \frac{\partial \boldsymbol{V}_0}{\partial \varpi}\right)\right\} \tag{4.66}$$

oraz

$$-\boldsymbol{E}\left(\frac{\partial^2 \ln L}{\partial \varpi_i \partial \varpi_j}\right) \approx \frac{1}{2} tr\left(\boldsymbol{V}_0^{-1} \frac{\partial \boldsymbol{V}_0}{\partial \varpi_i} \boldsymbol{V}_0^{-1} \frac{\partial \boldsymbol{V}_0}{\partial \varpi_i}\right). \tag{4.67}$$

Formuły (4.64)–(4.67) mogą zostać wykorzystane w celu zastosowania procedury scoringowej Fishera. Jak zauważyli German Rodriguez oraz Noreen Goldman (1995), zastosowanie algorytmu scoringowego Fishera jest równoważne estymacji ważoną metodą najmniejszych kwadratów parametrów następującego modelu:

$$\boldsymbol{y}_0 = \boldsymbol{X}\boldsymbol{\beta} \ + \boldsymbol{W}_0^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\boldsymbol{\mu}_0\right), \tag{4.68}$$

gdzie wagami są odwrotności $\boldsymbol{V}_0$, które definiuje równanie (4.63).

Alternatywne podejście do estymacji parametrów nieliniowych modeli wielopoziomowych zaproponowane zostało przez Goldsteina (1991). Nieliniowy model wielopoziomowy w sposób ogólny można zapisać następująco:

$$\boldsymbol{y} = \pi\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}\right) + \boldsymbol{\varepsilon}, \tag{4.69}$$

gdzie ε jest składnikiem losowym o zerowej wartości oczekiwanej. Goldstein (1991) zaproponował wykorzystanie rozwinięcia w szereg Taylora pierwszego rzędu wokół

$\boldsymbol{\beta} = \boldsymbol{\beta}^{\{0\}}$ oraz $\boldsymbol{u} = 0$, gdzie $\boldsymbol{\beta}^{\{0\}}$ jest pewną wartością startową. Po rozwinięciu wyrażenia (4.69) w szereg Taylora mamy:

$$\boldsymbol{y} \approx \pi\left(\boldsymbol{X}\boldsymbol{\beta}^{\{0\}}\right) + \frac{\partial \pi}{\partial\left(\boldsymbol{X}\boldsymbol{\beta}_0\right)}\boldsymbol{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{\{0\}}\right) + \frac{\partial \pi}{\partial\left(\boldsymbol{X}\boldsymbol{\beta}^{\{0\}}\right)}\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{4.70}$$

gdzie $\dfrac{\partial \pi}{\partial\left(\boldsymbol{X}\boldsymbol{\beta}^{\{0\}}\right)}$ jest macierzą diagonalną pierwszych pochodnych. W przypadku gdy mamy do czynienia z wielopoziomowym modelem logitowym, wykorzystuje się fakt, że dla funkcji logitowej pierwsza pochodna wynosi $\boldsymbol{W}_0$ i po wykonaniu drobnych przekształceń uzyskujemy:

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{Z}^*\boldsymbol{u} + \boldsymbol{\varepsilon}, \tag{4.71}$$

gdzie:

$$\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{\mu}_0 + \boldsymbol{X}^*\boldsymbol{\beta}^{\{0\}},$$

$$\boldsymbol{X}^* = \boldsymbol{W}_0\boldsymbol{X},$$

$$\boldsymbol{Z}^* = \boldsymbol{W}_0\boldsymbol{Z}.$$

Równanie (4.71) ma postać standardowego liniowego modelu wielopoziomowego. Zgodnie z tym równaniem prawdziwe są następujące zależności:

$$E\left(\boldsymbol{y}^*\right) = \boldsymbol{X}^*\boldsymbol{\beta}, \tag{4.72a}$$

$$E\left(\boldsymbol{y}^*\left(\boldsymbol{y}^*\right)^T\right) \approx \boldsymbol{V}^* = \boldsymbol{Z}^*\boldsymbol{\Omega}\left(\boldsymbol{Z}^*\right)^T + \boldsymbol{W}_0. \tag{4.72b}$$

Goldstein (1991) proponował szacowanie parametrów liniowego modelu wielopoziomowego (4.71) z macierzą kowariancji daną wzorem (4.72b) za pomocą uogólnionej metody najmniejszych kwadratów. Po wykonaniu pojedynczej iteracji należy jeszcze raz wyznaczyć macierz wag $\boldsymbol{W}_0$, a następnie zmodyfikowane macierze dla zmiennych objaśniających $\boldsymbol{X}^* = \boldsymbol{W}_0\boldsymbol{X}$ oraz $\boldsymbol{Z}^* = \boldsymbol{W}_0\boldsymbol{Z}$. Po zastąpieniu

odpowiednich elementów równania (4.71) kolejny raz szacowane są jego parametry. Omawiana procedura iteracyjna nosi nazwę iteracyjnie przeważanej metody najmniejszych kwadratów (*iteratively reweighted least squares*). Zbieżność proponowanego algorytmu pokazana została w pracy Petera McCullagha i Johna Neldera (2000). Jak pokazali Rodriguez oraz Goldman (1995), potrzeba modyfikacji zmiennych w każdej iteracji może zostać zlikwidowana. Jeśli równanie (4.71) zostanie przemnożone przez $(\boldsymbol{W}_0)^{-1}$, wówczas uzyskujemy:

$$\boldsymbol{y}_0 = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}^*, \tag{4.73}$$

gdzie zmodyfikowana zmienna zależna $\boldsymbol{y}_0$ przyjmuje następującą postać:

$$\boldsymbol{y}_0 = \boldsymbol{X\beta}_0 + \boldsymbol{W}_0^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu\mu}_0\right), \tag{4.74}$$

natomiast jej wartość oczekiwana i wariancja wynoszą odpowiednio:

$$E\left(\boldsymbol{y}_0\right) = \boldsymbol{X\beta} \tag{4.75}$$

oraz

$$var\left(\boldsymbol{y}_0\right) = \boldsymbol{Z\Omega Z}^{\boldsymbol{T}} + \boldsymbol{W}_0^{-1}. \tag{4.76}$$

Symulacje Monte Carlo przeprowadzone przez Rodrigueza i Goldman (1995) wskazują, że zarówno estymacja parametrów wielopoziomowego modelu logitowego za pomocą aproksymacji Longforda (1987; 1994), jak i wykorzystanie metody Goldsteina (1991) prowadzą do uzyskania obciążonych estymatorów parametrów. Dlatego też zaproponowane zostały metody korekty obciążenia.

W celu przezwyciężenia problemu obciążoności estymatorów w niektórych badaniach empirycznych wykorzystywano aproksymację Laplace'a wyższego rzędu. Podejście tego typu można znaleźć między innymi w pracach Stephena Raudenbusha, Meng-Li Yang i Matheosa Yosefa (2000), Petera Congdona (2005), Evangelosa Evangelou, Zhengyuan Zhu i Richarda Smitha (2011), Evangelosa Evangelou i Jo Eidsvika (2017), Rezy Hosseiniego Shojaei, Yadolloha Waghei i Mohsena Mohammadzadeha (2018). Najważniejszymi metodami korekty obciążenia dla estymatorów parametrów uogólnionych liniowych modeli wielopoziomowych są propozycja Anthony'ego Y.C. Kuka (1995), a także wykorzystanie metody stochastycznej aproksymacji Robbinsa-Monro (RM) (por. Wetherill, Glazerbrook, 1986). Metoda Kuka (1995) polega na bootstrapowej korekcie obciążenia. Punktem wyjścia jest wektor $\boldsymbol{\Theta}$ zawierający zarówno parametry ustalone, jak i losowe.

Estymator dla tego wektora parametrów, uzyskany w wyniku maksymalizacji funkcji quasi-największej wiarygodności, oznaczamy przez $\tilde{\boldsymbol{\Theta}}$. Na początku algorytmu iteracyjnego przyjmuje się, że estymator skorygowany jest równy temu bez zastosowania procedury korekty obciążenia. Prawdziwa jest zatem równość:

$$\hat{\boldsymbol{\Theta}}_{BC,KUK}^{\{0\}} = \tilde{\boldsymbol{\Theta}}. \tag{4.77}$$

W *i*-tym kroku algorytmu iteracyjnego wykonuje się następujące elementy:

1) wykorzystuje się parametryczny bootstrap w celu symulacji *H* zbiorów obserwacji z modelu z parametrami $\hat{\boldsymbol{\Theta}}_{BC,KUK}^{\{n-1\}}$;
2) szacuje się parametry dla każdego z *H* zbiorów obserwacji w celu uzyskania *H* wektorów oszacowań $\tilde{\boldsymbol{\Theta}}_{[h]}^{*}$, $h = 1, \ldots, H$;
3) oblicza się średnią ze wszystkich wektorów oszacowań: $\bar{\boldsymbol{\Theta}}^{*} = \sum_{h=1}^{H} \frac{\tilde{\boldsymbol{\Theta}}_{[h]}^{*}}{H}$;
4) oblicza się oszacowania z korektą obciążenia na podstawie następującej formuły: $\hat{\boldsymbol{\Theta}}_{BC,KUK}^{\{n\}} = \hat{\boldsymbol{\Theta}}_{BC,KUK}^{\{n-1\}} + \left(\tilde{\boldsymbol{\Theta}} - \bar{\boldsymbol{\Theta}}^{*}\right)$;
5) kroki 1–4 powtarza się do momentu, kiedy $\left|\hat{\boldsymbol{\theta}}_{BC,KUK}^{\{n\}} - \hat{\boldsymbol{\theta}}_{BC,KUK}^{\{n-1\}}\right| < \boldsymbol{\zeta}$, gdzie $\boldsymbol{\zeta}$ oznacza bardzo niską wartość.

Punktem wyjścia do zastosowania metody stochastycznej aproksymacji RM jest przyjęcie założenia, że początkowy estymator skorygowany jest równy temu bez zastosowania procedury korekty obciążenia. Oznacza to zatem, że:

$$\hat{\boldsymbol{\Theta}}_{BC,RM}^{\{0\}} = \tilde{\boldsymbol{\Theta}}. \tag{4.78}$$

W *n*-tym kroku algorytmu iteracyjnego wykonuje się następujące elementy:

1) symuluje się pojedynczy zbiór danych dla wektora oszacowań $\hat{\boldsymbol{\theta}}_{BC,RM}^{\{n-1\}}$, a następnie znajduje się oszacowanie $\tilde{\boldsymbol{\theta}}^{*}$;
2) oblicza się $\hat{\boldsymbol{\theta}}_{BC,RM}^{\{n\}} = \hat{\boldsymbol{\theta}}_{BC,RM}^{\{n-1\}} + \ddot{a}_n\left(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^{*}\right)$, gdzie $\ddot{a}_i = \frac{\ddot{c}}{i}$,, natomiast stałą $\ddot{c}$ na ogół wybiera się tak, aby zachodziła równość $\ddot{c} = \left\{\nabla\breve{\mathrm{k}}\left(\boldsymbol{\theta}\right)\right\}^{-1}$, gdzie $\breve{\mathrm{k}}\left(\boldsymbol{\theta}\right) = \tilde{\boldsymbol{\theta}} - \breve{b}\left(\boldsymbol{\theta}\right)$, natomiast $\breve{b}\left(\boldsymbol{\theta}\right)$ jest wartością oczekiwaną oszacowania uzyskanego metodą quasi-największej wiarygodności, jeśli θ jest prawdziwym wektorem parametrów;
3) kroki 1–2 powtarzane są aż do osiągnięcia zbieżności.

Największą wadą obydwu analizowanych metod korekty obciążenia jest brak możliwości obliczenia odchyleń standardowych dla skorygowanych estymatorów.

Wzór na asymptotyczne obciążenie w uogólnionym liniowym modelu wielopoziomowym wyprowadzili Norman Breslow i Xihong Lin (1995). Zaproponowali oni zastosowanie analitycznej metody korekty obciążenia. Wzór definiujący skorygowany estymator parametrów strukturalnych przyjmuje postać:

$$\hat{\boldsymbol{\beta}}_{BL} = \hat{\boldsymbol{\beta}}_{\{un\}} - \frac{1}{2} det\left(\boldsymbol{\Omega}\right)\left(\boldsymbol{X}^T \boldsymbol{W}_0 \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{u}\,, \tag{4.79}$$

gdzie $\hat{\boldsymbol{\beta}}_{\{un\}}$ jest estymatorem nieskorygowanym, uzyskanym za pomocą aproksymacji funkcji wiarygodności zaproponowanej przez Solomona i Coxa (1992) lub Breslowa i Claytona (1993).

4.4. Estymacja parametrów uogólnionych liniowych modeli wielopoziomowych za pomocą metod symulacyjnych

Najpopularniejsza metoda estymacji parametrów hierarchicznych modeli logitowych jest oparta na propozycji McCullocha (1997). W analizowanej pracy zaproponowane zostało wykorzystanie trzech algorytmów symulacyjnych w celu estymacji parametrów uogólnionych liniowych modeli mieszanych. Punktem wyjścia do estymacji parametrów za pomocą metod symulacyjnych jest założenie, że elementy wektora $\boldsymbol{y}$ są niezależne i pochodzą z rozkładu należącego do rodziny wykładniczej:

$$f_{w_i}\left(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\beta}, \varphi\right) = \exp\left\{\frac{\left(y\eta_i - c\left(\eta_i\right)\right)}{a\left(\varphi\right)} + d\left(y, \varphi\right)\right\}, \tag{4.80a}$$

$$\boldsymbol{u} \sim g_{uu}\left(\boldsymbol{u} \mid \boldsymbol{\Omega}\right), \tag{4.80b}$$

gdzie:

$$\eta_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{u}\,.$$

Funkcja wiarygodności dana wzorem (4.22) dla modelu (4.80a)–(4.80b) przyjmuje postać:

$$L(\boldsymbol{\Theta}\,|\,\boldsymbol{y}) = \int \prod_{i=1}^{I} f_{w_i}\left(y_i\,|\,\boldsymbol{u},\boldsymbol{\beta},\varphi\right) g_{uu}\left(\boldsymbol{u}\,|\,\boldsymbol{\Omega}\right) d\boldsymbol{u}\,, \tag{4.81}$$

gdzie wymiar całki zależy od liczby poziomów dla efektów losowych.

Jedną z metod estymacji parametrów modelu (4.80a)–(4.80b) jest zastosowanie algorytmu MCEM (Monte Carlo Expectation Maximization). Był on wcześniej wykorzystywany w pracy K.S. Chana i Johannesa Ledoltera (1994), jednak jego zastosowanie do szacowania parametrów wielopoziomowych modeli logitowych zasugerowane zostało przez McCullocha (1997). Aby zastosować algorytm EM do estymacji parametrów modeli wielopoziomowych, należy przyjąć założenie, że efekty losowe $\boldsymbol{u}$ są danymi brakującymi. Wówczas jeśli macierz obserwacji na wszystkich zmiennych (obserwowalnych i brakujących) zdefiniuje się jako $\boldsymbol{MO} = \left[\boldsymbol{y}^T \quad \boldsymbol{u}^T\right]^T$, logarytm funkcji wiarygodności wynosi:

$$\ln L_{\boldsymbol{MO}} = \sum_i \ln\left\{ f_{w_i}\left(y_i\,|\,\boldsymbol{u},\boldsymbol{\beta},\varphi\right)\right\} + \ln\left\{ g_{uu}\left(\boldsymbol{u}\,|\,\boldsymbol{\Omega}\right)\right\}. \tag{4.82}$$

Na początku wybierane są wartości startowe $\boldsymbol{\beta}^{\{0\}}$, $\boldsymbol{\varphi}^{\{0\}}$ oraz $\boldsymbol{\Omega}^{\{0\}}$. Kolejne kroki algorytmu EM są następujące:

Krok 1. Obliczane są $\boldsymbol{\beta}^{\{n+1\}}$, $\boldsymbol{\varphi}^{\{n+1\}}$ jako oszacowania maksymalizujące $E\left(\ln\left(f_{w_i}\left(y_i\,|\,\boldsymbol{u},\boldsymbol{\beta}^{\{n\}},\varphi^{\{n\}}\right)\right)\right)$.

Krok 2. Szacowane są parametry rozkładu efektów losowych w kolejnej iteracji $\boldsymbol{\Omega}^{\{n+1\}}$. W celu znalezienia tych oszacowań maksymalizowana jest następująca wartość oczekiwana $E\left(\ln\left(g_{uu}\left(\boldsymbol{u}\,|\,\boldsymbol{\Omega}\right)\right)\right)$.

Krok 3. Ustala się $n = n + 1$ i następuje powrót do pierwszego kroku.

Osiągnięcie zbieżności implikuje, że uzyskane oszacowania maksymalizują wartość funkcji wiarygodności.

Analityczne obliczenie wartości oczekiwanych $E\left(\ln\left(f_{w_i}\left(y_i\,|\,\boldsymbol{u},\boldsymbol{\beta},\varphi\right)\right)\right)$ oraz $E\left(\ln\left(g_{\boldsymbol{uu}}\left(\boldsymbol{u}\,|\,\boldsymbol{\Omega}\right)\right)\right)$ nie jest możliwe, ponieważ gęstość warunkowa $\boldsymbol{u}|y$ obejmuje funkcję f_y. Niemniej jednak możliwe jest zastosowanie algorytmu Metropolis (por. Tanner, 1993) w celu wylosowania z rozkładu warunkowego $\boldsymbol{u}|y$. Takie rozwiązanie nie wymaga specyfikacji f_y. Aby zastosować ten algorytm, należy wybrać rozkład próbkowy $h_{uu}(\boldsymbol{u})$, z którego następnie losowane są nowe wartości oraz obliczana jest wartość funkcji akceptacji (prawdopodobieństwa akceptacji nowej wartości). Niech $\boldsymbol{u}$ oznacza wektor wartości aktualnie wylosowanych z rozkładu warunkowego $u|y$, natomiast losowany jest qq-ty element tego wektora u^*_{qq} w taki sposób, że wektor

$\boldsymbol{u}^*$ definiowany jest następująco: $\boldsymbol{u}^* = \left(u_1, u_2, \ldots, u_{qq-1}, u^*_{qq}, u_{qq+1}, \ldots, u_{QQ}\right)$. Wówczas $\boldsymbol{u}^*$ akceptowany jest jako nowy właściwy wektor z prawdopodobieństwem:

$$\ddot{A}_j\left(\boldsymbol{u}, \boldsymbol{u}^*\right) = \min\left\{1, \frac{g_{\boldsymbol{uu}}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}\right)}{g_{uu}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}^*\right)}\right\}. \tag{4.83}$$

Jeśli iloraz $\frac{g_{\boldsymbol{uu}}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}\right)}{g_{uu}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}^*\right)}$ jest niższy niż 1, wówczas losowana jest wartość z rozkładu $U(0, 1)$. Jeśli wartość wylosowana przekracza $\frac{g_{uu}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}\right)}{g_{uu}\left(\boldsymbol{u}^* \mid \boldsymbol{y}, \boldsymbol{\beta}, \varphi, \boldsymbol{\Omega}\right) h_{uu}\left(\boldsymbol{u}^*\right)}$, wówczas należy zachować wektor $\boldsymbol{u}$.

Łącząc ze sobą algorytmy Metropolis oraz EM, uzyskuje się algorytm iteracyjny MCEM. Po wyborze wartości startowych $\boldsymbol{\beta}^{\{0\}}$, $\varphi^{\{0\}}$ oraz $\boldsymbol{\Omega}^{\{0\}}$ kontynuowana jest następująca procedura:

Krok 1. Generowanych jest H wektorów wartości $\boldsymbol{u}_{[1]}$, $\boldsymbol{u}_{[2]}$, …, $\boldsymbol{u}_{[H]}$ z rozkładu $g_{uu}(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}^{\{n\}}, \boldsymbol{\varphi}^{\{n\}}, \boldsymbol{\Omega}^{\{n\}})$ przy użyciu opisanego wzorem (4.83) algorytmu Metropolis.

Krok 2. $\boldsymbol{\beta}^{\{n+1\}}$ oraz $\boldsymbol{\varphi}^{\{n+1\}}$, wybierane są w taki sposób, aby zmaksymalizować wartość wyrażenia:

$$\frac{1}{H}\sum_{h=1}^{H} \ln\left(f_w\left(\boldsymbol{y} \mid \boldsymbol{u}_{[h]}, \boldsymbol{\beta}^{\{n+1\}}, \varphi^{\{n+1\}}\right)\right). \tag{4.84}$$

Krok 3. $\boldsymbol{\Omega}^{\{n+1\}}$ wybierane jest w taki sposób, aby zmaksymalizować wartość wyrażenia:

$$\frac{1}{H}\sum_{h=1}^{H} \ln\left(g_{uu}\left(\boldsymbol{u}_{[h]} \mid \boldsymbol{\Omega}^{\{n+1\}}\right)\right). \tag{4.85}$$

Krok 4. Należy sprawdzić, czy zbieżność została osiągnięta. Jeśli tak, wówczas wyznaczone w $n + 1$ kroku oszacowania są ocenami największej wiarygodności. W przeciwnym przypadku należy wrócić do kroku 1 i procedurę iteracyjną kontynuować tak długo, aż zbieżność zostanie osiągnięta.

Kolejna zaproponowana przez McCullocha (1997) symulacyjna metoda estymacji parametrów uogólnionych liniowych modeli mieszanych nazywana jest metodą

Monte Carlo Newtona-Raphsona. Punktem wyjścia do jej zastosowania jest obserwacja, że gęstość brzegowa $\boldsymbol{y}$ jest iloczynem funkcji gęstości f_w oraz g_{uu} zależnych od różnych parametrów. Wówczas wartości oczekiwane z wektorów pierwszych pochodnych względem $\ddot{\boldsymbol{\kappa}} = (\boldsymbol{\beta}, \varphi)$ oraz $\boldsymbol{\Omega}$ wynoszą odpowiednio:

$$E\left[\frac{\partial \ln\left\{f_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \ddot{\boldsymbol{\kappa}}\right)\right\}}{\partial \ddot{\boldsymbol{\kappa}}} \mid \boldsymbol{y}\right] = 0 \tag{4.86a}$$

oraz

$$E\left[\frac{\partial \ln\left\{g_{uu}\left(\boldsymbol{u} \mid \boldsymbol{\Omega}\right)\right\}}{\partial \boldsymbol{\Omega}} \mid \boldsymbol{y}\right] = 0\,. \tag{4.86b}$$

Ponieważ równanie (4.86b) obejmuje jedynie rozkład $\boldsymbol{u}$, jest ono często łatwe do rozwiązania. W celu rozwiązania równania (4.86a) McCulloch (1997) zaproponował rozwinięcie wyrażenia $\frac{\partial \ln f_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \ddot{\boldsymbol{\kappa}}\right)}{\partial \ddot{\boldsymbol{\kappa}}}$ jako funkcję $\boldsymbol{\beta}$ dookoła wartości $\boldsymbol{\beta}^{\{0\}}$:

$$\begin{aligned}\frac{\partial \ln\left\{f_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \ddot{\boldsymbol{\kappa}}\right)\right\}}{\partial \boldsymbol{\beta}} \cong \frac{\partial \ln\left\{f_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{\kappa}\right)\right\}}{\partial \boldsymbol{\beta}}\Big|_{\ddot{\boldsymbol{\kappa}}=\ddot{\boldsymbol{\kappa}}^{\{0\}}} + \\ + \frac{\partial^2\left\{\ln \boldsymbol{f}_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \ddot{\boldsymbol{\kappa}}\right)\right\}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\Big|_{\ddot{\boldsymbol{\kappa}}=\ddot{\boldsymbol{\kappa}}^{\{0\}}} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{\{0\}}\right).\end{aligned} \tag{4.87}$$

W związku z tym formuła algorytmu scoringowego dla rozkładu należącego do rodziny wykładniczej (4.80a) przyjmuje postać:

$$\begin{aligned}\frac{\partial \ln f_w\left(\boldsymbol{y} \mid \boldsymbol{u}, \ddot{\boldsymbol{\kappa}}\right)}{\partial \boldsymbol{\beta}} \cong \boldsymbol{X}^T \ddot{\boldsymbol{W}}\left(\ddot{\boldsymbol{\kappa}}^{\{0\}}, \boldsymbol{u}\right) \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\Big|_{\ddot{\boldsymbol{\kappa}}=\ddot{\boldsymbol{\kappa}}^{\{0\}}} \left(\frac{\left(\boldsymbol{y} - \boldsymbol{\mu}\left(\ddot{\boldsymbol{\kappa}}^{\{0\}}, \boldsymbol{u}\right)\right)}{a(\varphi)}\right) - \\ - \boldsymbol{X}^T \boldsymbol{W}\left(\ddot{\boldsymbol{\kappa}}^{\{0\}}, \boldsymbol{u}\right) \boldsymbol{X} \frac{\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{\{0\}}\right)}{a(\varphi)},\end{aligned} \tag{4.88}$$

gdzie:

$$\mu_i\left(\boldsymbol{\Theta},\boldsymbol{u}\right)=E\left[Y_i\mid\boldsymbol{u}\right],\ \boldsymbol{W}\left(\ddot{\boldsymbol{\kappa}},\boldsymbol{u}\right)^{-1}=$$

$$=diag\left\{\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2 var\left(Y_i\mid\boldsymbol{u}\right)\right\},\ \frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}=diag\left(\frac{\partial\eta_i}{\partial\mu_i}\right).$$

Wykorzystując aproksymację (4.88) oraz równanie (4.86a), uzyskujemy następujące równanie oszacowania uzyskiwanego w $n+1$ kroku algorytmu iteracyjnego:

$$\boldsymbol{\beta}^{\{n+1\}}=\boldsymbol{\beta}^{\{n\}}+E\left[\boldsymbol{X}^T\boldsymbol{W}\left(\ddot{\boldsymbol{\kappa}}^{\{n\}},\boldsymbol{u}\right)\boldsymbol{X}\mid\boldsymbol{y}\right]^{-1}\boldsymbol{X}^T$$

$$\left(E\left[\boldsymbol{W}\left(\ddot{\boldsymbol{\kappa}}^{\{n\}},\boldsymbol{u}\right)\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}|_{\ddot{\boldsymbol{\kappa}}=\ddot{\boldsymbol{\kappa}}^{\{n\}}}\left(\boldsymbol{y}-\mu\left(\boldsymbol{\beta}^{\{n\}},\boldsymbol{u}\right)\right)\mid\boldsymbol{y}\right]\right). \tag{4.89}$$

Wartość oczekiwana z równania (4.89) obliczana jest za pomocą metody Monte Carlo. Ostatecznie algorytm poszukiwania oszacowań dla wielopoziomowego modelu logitowego zgodnie z metodą MCNR (Monte Carlo Newton-Raphson) składa się z następujących kroków:

Krok 1. Wybierane są wartości startowe $\boldsymbol{\beta}^{\{0\}}$, $\boldsymbol{\varphi}^{\{0\}}$ oraz $\boldsymbol{\Omega}^{\{0\}}$.

Krok 2. Generowanych jest H wartości $\boldsymbol{u}_{[1]}$, $\boldsymbol{u}_{[2]}$, …, $\boldsymbol{u}_{[H]}$ z rozkładu warunkowego $g_{uu}(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\beta}^{\{n\}}, \boldsymbol{\varphi}^{\{n\}}, \boldsymbol{\Omega}^{\{n\}})$ przy użyciu algorytmu Metropolis opisanego powyżej wzorem (4.80).

Krok 3. Obliczane jest $\boldsymbol{\beta}^{\{1\}}$ na podstawie wzoru (4.79) po podstawieniu $n=0$.

Krok 4. Dla danego $\boldsymbol{\beta}^{\{1\}}$ obliczane jest $\boldsymbol{\varphi}^{\{1\}}$ jako rozwiązanie równania $E\left[\frac{\partial\ln f_w\left(\boldsymbol{y}\mid\boldsymbol{u},\ddot{\boldsymbol{\kappa}}\right)}{\partial\varphi}\mid\boldsymbol{y}\right]=0$.

Krok 5. Gdy dane są $\beta^{\{1\}}$ oraz $\varphi^{\{1\}}$, obliczane są elementy $\boldsymbol{\Omega}^{\{1\}}$ w taki sposób, aby wartość wyrażenia $\frac{\sum_{h=1}^{H}\ln\left\{g_{\boldsymbol{uu}}\left(\boldsymbol{u}_{[h]}\mid\boldsymbol{\Omega}\right)\right\}}{H}$ była najwyższa.

Krok 6. Następuje powrót do kroku trzeciego i obliczanie $\boldsymbol{\beta}^{\{2\}}$ – tym razem dla $n=1$.

Procedura kontynuowana jest tak długo, aż zbieżność zostanie osiągnięta.

Symulacyjna metoda maksymalizacji funkcji wiarygodności dla wielopoziomowego modelu logitowego została zaproponowana także w pracach Charlesa

J. Geyera i Elizabeth A. Thompson (1992) oraz Alana Gelfanda i Bradleya Carlina (1993). Startując od funkcji wiarygodności (4.51), wspomniani autorzy zaproponowali wykorzystanie rozkładu ważności próbkowania $h_{uu}(\boldsymbol{u})$ i zdefiniowanie funkcji wiarygodności w następujący sposób:

$$\begin{aligned} L\left(\boldsymbol{\beta},\varphi,\boldsymbol{\Omega}\mid\boldsymbol{y}\right) &= \int f_{w_i}\left(y_i\mid\boldsymbol{u},\boldsymbol{\beta},\varphi\right) g_{uu}\left(\boldsymbol{u}\mid\boldsymbol{\Omega}\right) d\boldsymbol{u} = \\ &= \int \frac{f_{w_i}\left(y_i\mid\boldsymbol{u},\boldsymbol{\beta},\varphi\right) g_{uu}\left(\boldsymbol{u}\mid\boldsymbol{\Omega}\right) du}{h_{uu}\left(\boldsymbol{u}\right)} h_{uu}\left(\boldsymbol{u}\right) d\boldsymbol{u} \cong \\ &\cong \frac{1}{H}\sum_{h=1}^{H} \frac{f_{\boldsymbol{w}}\left(\boldsymbol{y}\mid\boldsymbol{u}_{[h]},\boldsymbol{\beta},\varphi\right) g_{\boldsymbol{uu}}\left(\boldsymbol{u}_{[h]}\mid\boldsymbol{\Omega}\right)}{h_{\boldsymbol{u}}\left(\boldsymbol{u}_{[h]}\right)}, \end{aligned} \tag{4.90}$$

gdzie wektor $\boldsymbol{u}$ jest losowany z rozkładu ważności próbkowania, natomiast H jest liczbą wysymulowanych wartości. W rezultacie uzyskuje się nieobciążone oszacowanie wartości funkcji wiarygodności, niezależnie od wyboru funkcji $h_{uu}(\boldsymbol{u})$. Wartość symulowanej funkcji wiarygodności jest następnie maksymalizowana numerycznie, wykorzystując pojedynczą symulację lub wiele symulacji w kolejnych iteracjach.

W celu porównania właściwości estymatorów uzyskanych metodami symulacyjnymi z innymi rozważanymi na początku niniejszego podrozdziału McCulloch (1997) wykonał symulację Monte Carlo. Wyniki badania symulacyjnego wskazują, że zastosowanie metod MCEM oraz MCNR prowadzi do uzyskania oszacowań zdecydowanie bliższych prawdziwym wartościom parametrów w porównaniu z alternatywnymi procedurami rozważanymi w niniejszym podrozdziale. Dotyczy to zarówno oszacowania dla wektora parametrów $\boldsymbol{\beta}$, jak i elementów macierzy $\boldsymbol{\Omega}$.

4.5. Problem selekcji próby w modelach wielopoziomowych Estymacja parametrów wielorównaniowych modeli probitowych z efektami losowymi

Rozważmy standardowy model probitowy:

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \\ y_i &= I\left\{y_i^* > 0\right\}, \end{aligned} \tag{4.91}$$

gdzie – jak wiadomo – składnik losowy pochodzi ze standardowego rozkładu normalnego. Wartości oczekiwane zmiennej nieobserwowalnej w rozkładzie uciętym wynoszą:

$$E\left(y_i^* \mid y_i^* > 0\right) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \frac{\phi\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)}{\boldsymbol{\Phi}\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)} \tag{4.92a}$$

oraz

$$E\left(y_i^* \mid y_i^* < 0\right) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} - \frac{\phi\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)}{1-\Phi\left(\boldsymbol{x}_i\hat{\boldsymbol{\beta}}\right)}. \tag{4.92b}$$

W przypadku gdy do modelu (4.91) dołączane są efekty losowe związane z poszczególnymi klastrami, zmienia się zarówno wartość oczekiwana zmiennej wynikowej, jak i wariancja części losowej. Załóżmy, że po estymacji parametrów wielopoziomowego modelu probitowego znane są oszacowania $\hat{\boldsymbol{\beta}}$, oszacowanie wariancji składnika losowego ε_i wynoszące $\hat{\sigma}_\varepsilon^2$, predykcje efektów losowych $\hat{\boldsymbol{u}} = \left[\hat{u}_1 \quad \cdots \quad \hat{u}_{QQ}\right]^T$, a także ich odchylenia standardowe $\left[s\left(\hat{u}_1\right) \quad \cdots \quad s\left(\hat{u}_{QQ}\right)\right]^T$. Wówczas estymator wartości oczekiwanej i wariancji warunkowych ze względu na efekty losowe wynoszą odpowiednio:

$$E\left(y_i^*\right) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}, \tag{4.93a}$$

$$Var\left(y_i^*\right) = \hat{\sigma}_\varepsilon^2 + \sum_{qq=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_{qq}\right). \tag{4.93b}$$

W związku z tym warunkowa wartość oczekiwana zmiennej nieobserwowalnej w zależności od wartości przyjmowanych przez zmienną obserwowalną wynosi:

$$E\left(y_i^* \mid y_i^* > 0\right) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}} + \frac{\phi\left(\dfrac{\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{q=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_q\right)}}\right)}{\boldsymbol{\Phi}\left(\dfrac{\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{j=1}^{IQ} z_{ij}^2 s^2\left(\hat{u}_j\right)}}\right)} \tag{4.94a}$$

oraz

$$E\left(y_i^* \mid y_i^* < 0\right) = \boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}} - \frac{\phi\left(\dfrac{\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{qq=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_{qq}\right)}}\right)}{1 - \boldsymbol{\Phi}\left(\dfrac{\boldsymbol{x}_i\hat{\boldsymbol{\beta}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{qq=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_{qq}\right)}}\right)}. \tag{4.94b}$$

Efekty losowe związane z przynależnością jednostek lub firm do określonych regionów czy sekcji mogą występować także w przypadkach związanych z selekcją próby. W przypadku selekcji próby w równaniu wyjaśniającym kształtowanie się zmiennej wynikowej należy uwzględnić odwrócony iloraz Millsa. Dlatego też w modelach wielopoziomowych uwzględniających selekcję próby również należy wyznaczyć wartości odwróconego ilorazu Millsa, aby dołączyć tę zmienną do równania wyjaśniającego kształtowanie się zmiennej wynikowej. W celu wyznaczenia odwróconego ilorazu Millsa zapiszmy równanie selekcji próby, w którym efekty losowe związane z przynależnością do określonych grup mają wpływ na prawdopodobieństwo selekcji:

$$\begin{aligned} y_{(2)i}^{*} &= \boldsymbol{w}_i\boldsymbol{\gamma} + \boldsymbol{z}_i\boldsymbol{u} + \varepsilon_{(2)i}, \\ y_{(2)i} &= I\left\{y_{(2)i}^{*} > 0\right\}. \end{aligned} \tag{4.95}$$

Po oszacowaniu parametrów równania selekcji (4.95) oraz predykcji efektów losowych i ich błędów standardowych wyznaczany jest odwrócony iloraz Millsa w następujący sposób:

$$IMR_i = \frac{\phi\left(\dfrac{\boldsymbol{w}_i\hat{\boldsymbol{\gamma}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{qq=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_{qq}\right)}}\right)}{1 - \boldsymbol{\Phi}\left(\dfrac{\boldsymbol{w}_i\hat{\boldsymbol{\gamma}} + \boldsymbol{z}_i\hat{\boldsymbol{u}}}{\sqrt{\hat{\sigma}_\varepsilon^2 + \sum_{qq=1}^{QQ} z_{iqq}^2 s^2\left(\hat{u}_{qq}\right)}}\right)}. \tag{4.96}$$

Wyrażenie (4.96) jest następnie włączane do równania wyjaśniającego kształtowanie się zmiennej wynikowej i uwzględniającego problem selekcji próby.

Przynależność firmy do określonej sekcji czy regionu może mieć również losowy wpływ na proces decyzyjny podejmowany równocześnie na kilku szczeblach. Na przykład przedsiębiorstwo może rozważać wprowadzenie innowacji produktowej, procesowej, organizacyjnej lub marketingowej. Jednocześnie firmy prowadzące ten sam rodzaj działalności lub zlokalizowane niedaleko siebie mogą porozumiewać się w zakresie wprowadzania innowacji bądź zaniechania odpowiedniego działania. W takich przypadkach estymacja parametrów wielopoziomowego, wielorównaniowego modelu probitowego jest uzasadniona.

Podobnie jak w podrozdziale 1.10 również w niniejszym podrozdziale rozważany będzie trójrównaniowy probitowy model wielopoziomowy. Przejście na przypadek ogólny nie jest skomplikowane. Zdefiniujmy trójrównaniowy, wielopoziomowy model probitowy:

$$y^{*}_{(1)i} = \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)} + \boldsymbol{z}_{(1)i}\boldsymbol{u}_{(1)} + \varepsilon_{(1)i}, \tag{4.97a}$$

$$y^{*}_{(2)i} = \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} + \boldsymbol{z}_{(2)i}\boldsymbol{u}_{(2)} + \varepsilon_{(2)i}, \tag{4.97b}$$

$$y^{*}_{(3)i} = \boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)} + \boldsymbol{z}_{(3)i}\boldsymbol{u}_{(3)} + \varepsilon_{(3)i}, \tag{4.97c}$$

$$\begin{bmatrix} \varepsilon_{(1)i} & \varepsilon_{(2)i} & \varepsilon_{(3)i} \end{bmatrix}^{T} \sim \boldsymbol{N}\left(0, \tilde{\boldsymbol{\Sigma}}\right), \tag{4.97d}$$

gdzie $\tilde{\boldsymbol{\Sigma}}$ jest dodatnio określoną macierzą wariancji-kowariancji między składnikami losowymi. Zdefiniujmy następujące składniki losowe:

$$v_{(1)i} = \boldsymbol{z}_{(1)i}\boldsymbol{u}_{(1)} + \varepsilon_{(1)i}, \tag{4.98a}$$

$$\begin{aligned} v_{(2)i} &= \boldsymbol{z}_{(2)i}\boldsymbol{u}_{(2)} + \varepsilon_{(2)i}, \\ v_{(3)i} &= \boldsymbol{z}_{(3)i}\boldsymbol{u}_{(3)} + \varepsilon_{(3)i}. \end{aligned} \tag{4.98b}$$

Po dokonaniu predykcji efektów losowych składniki losowe o zerowej wartości oczekiwanej dane są wzorem:

$$s_{(1)i} = v_{(1)i} - \boldsymbol{z}_{(1)i}\hat{\boldsymbol{u}}_{(1)}, \tag{4.99a}$$

$$s_{(2)i} = v_{(2)i} - \boldsymbol{z}_{(2)i}\hat{\boldsymbol{u}}_{(2)}, \tag{4.99b}$$

$$s_{(3)i} = v_{(3)i} - \boldsymbol{z}_{(3)i}\hat{\boldsymbol{u}}_{(3)}. \tag{4.99c}$$

Aby opisać procedurę estymacji parametrów wielopoziomowego, wielorównaniowego modelu probitowego, rozważmy prawdopodobieństwo, że wszystkie trzy zmienne obserwowalne przyjmują wartość 1:

$$\begin{aligned} &P\left(y_{(1)i}=1, y_{(2)i}=1, y_{(3)i}=1\right) = \\ &= P\left(v_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}, v_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_2, v_{(3)i} \le \boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)}\right), \end{aligned} \tag{4.100}$$

Zgodnie ze wzorem (2.73) prawdopodobieństwo łączne można zapisać jako iloczyn prawdopodobieństw warunkowych:

$$\begin{aligned} &P\left(y_{(1)i}=1, y_{(2)i}=1, y_{(3)i}=1\right) = \\ &= P\left(v_{(3)i} \le \boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)} \mid v_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)}, v_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right) * \\ &* P\left(v_{(2)i} \le \boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} \mid v_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right) * P\left(v_{(1)i} \le \boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}\right). \end{aligned} \tag{4.101}$$

Podobne dekompozycje prawdopodobieństwa łącznego można zastosować dla pozostałych siedmiu wariantów.

W celu aproksymacji odpowiednich prawdopodobieństw wchodzących w skład wyrażenia (4.101) dokonywana jest dekompozycja Choleskiego macierzy wariancji-kowariancji między składnikami losowymi $s_{(1)i}$, $s_{(2)i}$ oraz $s_{(3)i}$:

$$E\left(\begin{bmatrix} s_{(1)i} \\ s_{(2)i} \\ s_{(3)i} \end{bmatrix}\begin{bmatrix} s_{(1)i} & s_{(2)i} & s_{(3)i} \end{bmatrix}\right) = \tilde{\boldsymbol{\Sigma}} = \boldsymbol{C}\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^{T}\boldsymbol{C}, \tag{4.102}$$

gdzie $\boldsymbol{C}$ jest trójkątną dolną macierzą Choleskiego, odpowiadającą macierzy $\tilde{\boldsymbol{\Sigma}}$ oraz $\tilde{\boldsymbol{e}} \sim \Phi_3\left(0, \boldsymbol{I}_3\right)$. Dekompozycja (4.102) implikuje prawdziwość następujących zależności:

$$s_{(1)i} = C_{11}e_{(1)i},$$
$$s_{(2)i} = C_{21}e_{(1)i} + C_{22}e_{(2)i},$$
$$s_{(3)i} = C_{31}e_{1i} + C_{32}e_{2i} + C_{33}e_{3i},$$

gdzie C_{ij} jest elementem z i-tego wiersza oraz j-tej kolumny macierzy $\mathbf{C}$. W związku z tym wzór definiujący prawdopodobieństwo (4.100) można inaczej zapisać następująco:

$$P\left(y_{(1)i}=1, y_{(2)i}=1, y_{3i}=1\right) =$$
$$= P\left(s_{(1)i} \le \frac{\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)}}{C_{11}}\right) * P\left(s_{(2)i} \le \frac{\left(\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} - C_{21}e^{*}_{(1)}\right)}{C_{22}}\right) * \tag{4.103}$$
$$* P\left(s_{(3)i} \le \frac{\boldsymbol{x}_{(3)i}\boldsymbol{\beta}_{(3)} - C_{32}e^{*}_{(2)} - C_{21}e^{*}_{(1)}}{C_{33}}\right),$$

gdzie $e^{*}_{(1)}$ jest zmienną losową o standardowym rozkładzie normalnym, uciętą w punkcie $\boldsymbol{x}_{(1)i}\boldsymbol{\beta}_{(1)} - \boldsymbol{z}_{(1)i}\hat{\boldsymbol{u}}_{(1)}$, natomiast $e^{*}_{(2)}$ jest zmienną losową o standardowym rozkładzie normalnym, uciętą w punkcie $\frac{\left(\boldsymbol{x}_{(2)i}\boldsymbol{\beta}_{(2)} - \boldsymbol{z}_{(2)i}\hat{\boldsymbol{u}}_{(2)} - C_{21}e^{*}_{(1)}\right)}{C_{22}}$. Należy jednak zauważyć, że zmienne losowe $s_{(1)i}$, $s_{(2)i}$ oraz $s_{(3)i}$ zależą od predykcji efektów losowych. Dlatego też proponowana jest procedura polegająca na iteracyjnej predykcji efektów losowych i szacowaniu parametrów. Jej kroki są następujące:

Krok 1. Przy danych predykcjach efektów losowych szacowane są elementy macierzy $\boldsymbol{\Omega}$. Krok 2. Parametry wektorów $\boldsymbol{\beta}_{(1)}$, $\boldsymbol{\beta}_{(2)}$ oraz $\boldsymbol{\beta}_{(3)}$ szacowane są z wykorzystaniem symulatora GHK opisanego w podrozdziale 1.10.

Krok 3. Opisany w podrozdziale 4.4 algorytm Metropolis wykorzystywany jest do predykcji efektów losowych. Polega ona na wylosowaniu próby pseudolosowej z pewnego rozkładu (warunkowego).

Procedura kontynuowana jest tak długo, aż zbieżność zostanie osiągnięta.

4.6. Wykorzystanie wielopoziomowego modelu zmiennych dyskretnych do analizy zależności między wykorzystywaniem technologii informacyjnych i komunikacyjnych, innowacyjnością a produktywnością

4.6.1. Przegląd literatury z zakresu czynników wpływających na innowacyjność firm

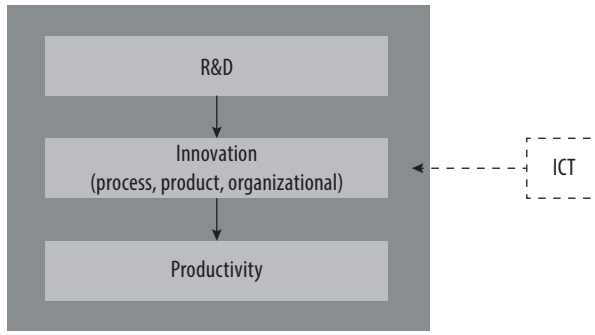
Istotne zmiany w przedsiębiorstwach obserwowane w ostatnich latach (m.in. wzrost udziału osób zatrudnionych posiadających wysokie kwalifikacje, zmiana organizacji zatrudnienia) przyczyniły się do powstania nowego paradygmatu w zakresie funkcjonowania firmy (por. Hollenstein, Stucki, 2012). Wskazuje się, że rozwój i dyfuzja technologii informatycznych i telekomunikacyjnych to główny motor zmian obserwowanych na poziomie mikro. Oczekuje się, że wraz ze wzrostem intensywności wykorzystania technologii informatycznych i telekomunikacyjnych powinien nastąpić wzrost produktywności przyczyniający się następnie do wzrostu gospodarczego.

Innym ważnym czynnikiem wzrostu, na który wskazuje się w pracach poświęconych tej tematyce, są wydatki na badania i rozwój. Przyczyniają się one do powstania innowacji, których wprowadzenie powinno wpływać na wzrost produktywności. Model wskazujący na zależność między cechami firm, wydatkami ponoszonymi na B+R, innowacyjnością a produktywnością został zaproponowany przez Bruno Crepona, Emmanuela Dugueta i Jacques'a Mairesse'a (1998), a jego nazwa CDM pochodzi o nazwisk autorów. Model CDM jest powszechnie wykorzystywany w pracach poświęconych decyzjom innowacyjnym i ich wpływowi na produktywność w przedsiębiorstwach (por. Szczygielski, Grabowski, 2014). Jego ideą jest wykorzystanie trzystopniowej procedury estymacyjnej. W pierwszym kroku szacowane są parametry modelu wyjaśniającego wydatki na badania i rozwój w zależności od cech firm. Następnie skłonność do wprowadzania innowacji jest wyjaśniana za pomocą intensywności wydatków na badania i rozwój. W trzecim kroku szacowane są parametry równania wyjaśniającego produktywność w zależności od poziomu innowacyjności firm.

Innowacje oraz wykorzystanie technologii informatycznych i telekomunikacyjnych powinny być traktowane jako czynniki komplementarne, których współpraca jest konieczna w celu uzyskania wyższej produktywności. Jak wskazują między innymi Paul Gretton, Jyothi Gali i Dean Parham (2004), Philipp Koellinger (2005), a także Vincenzo Spiezia (2011), można wyróżnić różne kanały wpływu technologii informacyjnych i komunikacyjnych (dalej TIiK) na innowacyjność:

1) dzięki wykorzystaniu TIiK następuje redukcja kosztów transakcyjnych, poprawa procesów biznesowych i koordynacji działania, a także wzrost dywersyfikacji produkcji,
2) celem technologii jest dostarczenie platformy do wprowadzania innowacji produktowych i procesowych oraz umożliwienie ich transferu między firmami,
3) technologie informatyczne i telekomunikacyjne ułatwiają kontakt między producentami, dostawcami i konkurentami.

W klasycznym modelu CDM nie uwzględnia się roli czynników związanych z wykorzystaniem technologii informatycznych i telekomunikacyjnych w podnoszeniu poziomu innowacyjności i produktywności przedsiębiorstw. Dlatego też w badaniach empirycznych poświęconych decyzjom innowacyjnym firm często wykorzystuje się rozszerzony model CDM zamiast klasycznego. Analizowane rozszerzenie polega na uwzględnieniu zmiennej wskazującej na intensywność stosowania technologii informatycznych i telekomunikacyjnych w każdym z trzech równań (por. np. Hall, Lotti, Mairesse, 2013). Przyjmuje się, że wykorzystanie technologii informatycznych i telekomunikacyjnych wpływa na intensywność wydatków na badania i rozwój, zachowania innowacyjne firm oraz produktywność. Rysunek 6 przedstawia rozszerzony model CDM.



Rysunek 6. Rozszerzony (o wykorzystanie TIiK) model CDM

Źródło: Hall, Lotti, Mairesse, 2013, zob. także Arendt, Grabowski, 2017.

Oprócz badań prowadzonych na poziomie firm, poświęconych czynnikom wpływającym na innowacyjność przedsiębiorstw, warto zwrócić uwagę na prace, w których analizowany jest wpływ przynależności do określonych regionów na postawy innowacyjne. Jak wskazują Małgorzata Kosała i Krzysztof Wach (2011), wśród determinant zdolności innowacyjnej przedsiębiorstw wyróżnia się czynniki wewnętrzne i zewnętrzne. Wśród czynników wewnętrznych wyróżnia się kategorie związane z poziomem zasobów ludzkich (wiedza i doświadczenie kadry

kierowniczej oraz pracowników), a także zmienne związane z historyczną oraz bieżącą działalnością firmy (szkolenia, inwestycje, nakłady na badania i rozwój, eksperymenty, wprowadzanie zmian, usuwanie błędów). Do czynników zewnętrznych wpływających na zdolność innowacyjną przedsiębiorstw zalicza się między innymi politykę terytorialną, politykę rządową, stowarzyszenie w organizacjach branżowych oraz klastrach, a także kontakty. Chodzi tu na przykład o kontakty z klientami, dostawcami, konkurentami, fundacjami, agencjami państwowymi, stowarzyszeniami branżowymi, firmami szkoleniowymi, instytucjami finansowymi czy agencjami konsultingowymi. Jak widać, czynniki zewnętrzne wynikają ze specyfiki oraz charakteru otoczenia danej organizacji. Określają one, czy dane środowisko jest przyjazne dla innowacyjności. Wśród badaczy wskazujących na ważną rolę aspektów regionalnych w zarządzaniu przedsiębiorstwem należy wskazać Michaela Portera (1998). Według tego autora różne regiony konkurują ze sobą, starając się zaoferować najbardziej korzystne otoczenie biznesu. Zgodnie z koncepcją zaproponowaną przez Portera i Scota Sterna (2001), wyróżnia się cztery główne obszary otoczenia regionalnego: uwarunkowania zasobowe, uwarunkowania popytowe, kontekst działania firmy, a także branże pokrewne i wspierające. Zdaniem autora elementy te tworzą system, a zatem muszą się wzajemnie wzmacniać i uzupełniać. Ważny wkład do teorii z zakresu wpływu przynależności regionalnej na innowacyjność firm miały prace Olivera Pfirrmanna (1994), Arnouda Langendijka (2001), oraz Christine Oughton, Mikela Landabaso i Kevina Morgana (2002). Wyniki badań przeprowadzonych przez tych autorów wskazują, że czynniki regionalne w największym stopniu determinują innowacyjne działania przedsiębiorstw oraz ich rozwój. Ostatni wymienieni autorzy definiują nawet zjawisko zwane regionalnym paradoksem innowacyjnym. Analizując wpływ zmiennych związanych z otoczeniem firmy na jej zachowania innowacyjne, Rolf Sternberg oraz Olaf Arndt (2001) rozróżniają czynniki regionalne oraz ponadregionalne. Wśród tych pierwszych autorzy wymieniają wykwalifikowaną kadrę pracowniczą w regionie, transfer technologii i zaplecze badawcze, wsparcie instytucjonalne, regionalną strukturę gospodarczą. Do drugich zaliczają politykę innowacyjną (inicjatywy badawcze oraz wsparcie badań, wsparcie kooperacji oraz działań w nowych obszarach badawczych), a także czynniki makrootoczenia (stan i rozwój branży, wielkość rynku, popyt, konkurencja, globalizacja i regionalizacja, postęp techniczny). Heiko Bergmann, Andrea *Japsen* i Christine *Tamasy* (2002) w badaniach poświęconych roli otoczenia regionalnego w kształtowaniu innowacyjności przedsiębiorstw wyróżnili następujące czynniki warunkujące:

1) transfer technologii i wiedzy do małych i średnich przedsiębiorstw,
2) jakość wykształcenia oraz sposób kształcenia uwzględniający specyficzne potrzeby potencjalnych przedsiębiorców,

3) dostępność doradztwa biznesowego dla małych i średnich przedsiębiorstw,
4) dostępność wysoko wykwalifikowanej kadry pracowniczej na lokalnym rynku pracy.

Powstało wiele prac poświęconych mierzeniu potencjału innowacyjnego polskich regionów i jego wpływu na innowacyjność firm zlokalizowanych na określonych terenach. Wyniki analiz prowadzonych dla województw (Kosała, Wach, 2011) oraz dla podregionów (por. Golejewska, 2018) wskazują na znaczące różnice między obserwowanym poziomem innowacyjności w różnych częściach Polski. Wartości wskaźników syntetycznych innowacyjności najwyższe są w przypadku takich województw jak mazowieckie, dolnośląskie, małopolskie, śląskie czy pomorskie. Porównanie poziomu innowacyjności między mniejszymi jednostkami administracyjnymi wskazuje na zdecydowaną przewagę aglomeracji (np. Wrocław, Warszawa, Kraków, Trójmiasto) nad mniej zurbanizowanymi podregionami.

Dysponując danymi indywidualnymi dotyczącymi aktywności innowacyjnej przedsiębiorstw oraz informacjami na temat ich lokalizacji, możliwa jest estymacja parametrów modelu uwzględniającego zarówno wewnętrzne, jak i zewnętrzne determinanty. Jednoczesne uwzględnienie zmiennych indywidualnych i regionalnych, identyfikacja podobieństw w postawach innowacyjnych przedsiębiorstw zlokalizowanych w tej samej jednostce administracyjnej, a także oszacowanie zróżnicowania między decyzjami firm mających swoje siedziby w innych województwach jest możliwe dzięki zastosowaniu modelu wielopoziomowego.

4.6.2. Model CDM rozszerzony o wykorzystanie TIiK oraz uwzględniający czynniki regionalne

W niniejszym punkcie prezentowany jest model CDM rozszerzony o wykorzystanie TIiK, uwzględniający jednocześnie wpływ czynników związanych z otoczeniem firmy.

4.6.2.1. Dane oraz próba badawcza

Dane wykorzystywane w badaniu empirycznym zostały zebrane w ramach projektu pt. „Wpływ technologii informatycznych i telekomunikacyjnych na produktywność – analiza mikro- i makroekonomiczna", finansowanego przez Narodowe Centrum Nauki. Tysiąc pięćdziesiąt firm zostało wylosowanych tak, aby próba była reprezentatywna ze względu na rozmiar firmy (wyróżniono firmy małe, średnie i duże, zgodnie z definicjami przyjętymi w Unii Europejskiej), sektor (przemysł, budownictwo, usługi), a także region (województwa). Wywiady bezpośrednie zostały przeprowadzone w 2015 roku z personelem zarządzającym firmami. W kwestionariuszu zawarte były pytania dotyczące wykorzystania technologii informatycznych

i telekomunikacyjnych przez firmy, innowacyjności, zmiany organizacyjnej oraz kapitału ludzkiego. W badaniu empirycznym uwzględnione zostały tylko przedsiębiorstwa wykorzystujące technologie informatyczne i telekomunikacyjne. Warunkiem koniecznym udziału w badaniu było posiadanie komputera oraz wykorzystywanie TIiK w co najmniej dwóch spośród następujących procesów biznesowych:

1) księgowość (fakturowanie, finanse i podatki),
2) zarządzanie zasobami ludzkimi w firmie,
3) zarządzanie zaopatrzeniem (magazyn, inwentaryzacja, dostawy),
4) zarządzanie produkcją,
5) zarządzanie sprzedażą i kontaktem z klientami (CRM),
6) zarządzanie zasobami przedsiębiorstwa (ERP),
7) wsparcie dla projektowania i wytwarzania CAD/CAM,
8) sterowanie maszynami lub linią produkcyjną,
9) zarządzanie pracami administracyjno-biurowymi.

Tabela 22. Definicje zmiennych binarnych związanych z wykorzystanie technologii informatycznych i telekomunikacyjnych

Nazwa zmiennej	Definicja
ICT_KS	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w księgowości
ICT_ZZL	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zasobami ludzkimi
ICT_ZZ	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zaopatrzeniem
ICT_ZP	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem produkcją
ICT_CRM	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem sprzedażą i kontaktem z klientami
ICT_ERP	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa
ICT_CAD	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM
ICT_SM	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną
ICT_ZPAB	Zmienna binarna przyjmująca wartość 1 dla firm wykorzystujących TIiK w procesach biznesowych związanych z zarządzaniem pracami administracyjno-biurowymi

Źródło: opracowanie własne.

W badaniu nie zostały uwzględnione firmy niewykorzystujące technologii informatycznych i telekomunikacyjnych oraz wykorzystujące te technologie w niewielkiej skali. Szczegóły dotyczące ankiety wykorzystywanej do przeprowadzenia badania empirycznego, cech firm wykorzystujących technologie informatyczne i komunikacyjne oraz analiz niektórych (przeprowadzonych na podstawie innych metod) zależności między cechami firm, wykorzystaniem TIiK, innowacyjnością a produktywnością można znaleźć między innymi w pracach Łukasza Arendta, Elżbiety Kryńskiej (2015) oraz Arendta, Grabowskiego (2017; 2018).

Na podstawie odpowiedzi firm dotyczących wykorzystania technologii informatycznych i komunikacyjnych zdefiniowane zostały następujące zmienne zero-jedynkowe (tabela 22).

Tabela 23 zawiera rozkład empiryczny dla zmiennych zero-jedynkowych zdefiniowanych w tabeli 22.

Tabela 23. Empiryczny rozkład dla zmiennych binarnych zdefiniowanych w tabeli 22

	ICT_KS	*ICT_ZZL*	*ICT_ZZ*	*ICT_ZP*	*ICT_CRM*	*ICT_ERP*	*ICT_CAD*	*ICT_SM*	*ICT_ZPAB*
0	0,08	0,45	0,30	0,70	0,37	0,72	0,76	0,77	0,34
1	0,92	0,55	0,70	0,30	0,63	0,28	0,24	0,23	0,66

Źródło: opracowanie własne.

Wyniki zawarte w tabeli 23 wskazują, że technologie informatyczne i telekomunikacyjne z zakresu księgowości, zarządzania zasobami ludzkimi, zarządzania zaopatrzeniem, zarządzania sprzedażą i kontaktem z klientami oraz zarządzania pracami administracyjno-biurowymi są powszechnie wykorzystywane w polskich przedsiębiorstwach. Większość spośród badanych przedsiębiorstw wykorzystuje TIiK w wymienionych procesach biznesowych. Nieliczne przedsiębiorstwa wykorzystują technologie informatyczne i komunikacyjne w procesach związanych z zarządzaniem produkcją, zarządzaniem zasobami przedsiębiorstwa, wsparciem dla projektowania i wytwarzania CAD/CAM oraz sterowaniem maszynami lub linią produkcyjną. Dlatego też wydaje się, że osiągnięcie przewagi konkurencyjnej, umożliwiającej wprowadzanie innowacji i wzrost produktywności, pozwala na wykorzystanie bardziej zaawansowanych i rzadziej stosowanych TIiK. Wykorzystując technologie informatyczne i komunikacyjne w procesach biznesowych związanych z księgowością przedsiębiorstwo nie jest w stanie uzyskać przewagi nad konkurencją, gdyż potencjalni konkurenci wykorzystują TIiK w analizowanych procesach. Jeśli jednak komputery i oprogramowanie używane są na przykład w celu sterowania maszynami i linią produkcyjną, wówczas przedsiębiorstwo może osiągnąć przewagę nad firmami, które wykonują te procesy bez ich użycia. Dlatego też w dalszej części badania empirycznego podjęta zostanie próba wyjaśnienia czynników

wpływających na użycie zaawansowanych TIiK, czyli dotyczących takich procesów biznesowych, jak zarządzanie produkcją, zarządzanie zasobami przedsiębiorstwa, wsparcie dla projektowania i wytwarzania CAD/CAM, sterowanie maszynami lub linią produkcyjną.

Firmy objęte badaniem sprawdzane były także pod kątem inwestycji w technologie informatyczne i komunikacyjne. Każdy z respondentów miał za zadanie udzielenie odpowiedzi na następujące pytanie: „Czy w ciągu minionych 24 miesięcy Państwa firma poczyniła jakieś inwestycje w rozwój TIiK?". Respondenci mogli na nie odpowiedzieć tylko „Tak" lub „Nie". W związku z tym na podstawie odpowiedzi na powyższe pytanie skonstruowana została zmienna binarna *INWETYCJE_ICT*, przyjmująca wartość 1 w przypadku firm, które odpowiedziały pozytywnie. Tabela 24 prezentuje rozkład empiryczny zmiennej *INWETYCJE_ICT*.

Tabela 24. Empiryczny rozkład zmiennej ilustrującej fakt dokonania inwestycji w rozwój TIiK przez firmy

Wartość	Udział
0	0,50
1	0,50

Źródło: opracowanie własne.

Jak widać, około połowa badanych firm poczyniła inwestycje w rozwój technologii informatycznych i komunikacyjnych w okresie co najwyżej 24 miesięcy przed momentem przeprowadzenia badania. W dużej części były to inwestycje związane z zakupem komputerów i odpowiedniego oprogramowania, rozbudową strony WWW oraz wewnętrznej sieci internetowej, a także zakupem maszyn sterowanych numerycznie. Zmienna *INWETYCJE_ICT* oraz kategorie związane z wykorzystaniem technologii informatycznych i komunikacyjnych w zdefiniowanych powyżej czterech procesach biznesowych (zarządzanie produkcją, zarządzanie zasobami przedsiębiorstwa, wsparcie dla projektowania i wytwarzania CAD/CAM, sterowanie maszynami lub linią produkcyjną) ilustrują intensywność nakładów na TIiK w omawianym modelu koncepcyjnym.

Zgodnie z modelem CDM (por. Crepon, Duguet, Mairessec, 1998) poziom innowacyjności firm powinien zależeć od nakładów na badania i rozwój. W kwestionariuszu nie znalazło się pytanie dotyczące intensywności wydatków na B+R. Niemniej jednak osoby ankietowane udzielały odpowiedzi na następujące pytanie: „Czy Państwa firma posiada wyodrębniony wydział B+R – badawczo-rozwojowy, zajmujący się tworzeniem i testowaniem nowych rozwiązań?". Respondenci mogli na nie odpowiedzieć tylko „Tak" lub „Nie". Wydaje się, że firmy

posiadające własny dział B+R powinny charakteryzować się wyższym poziomem innowacyjności w porównaniu z firmami, które takiego działu nie posiadają. Dlatego też intensywność nakładów na B+R ilustrowana jest za pomocą analizowanej zmiennej, której empiryczny rozkład prezentowany jest w tabeli 25.

Tabela 25. Empiryczny rozkład zmiennej wskazującej, czy firma posiada wyodrębniony wydział B+R

Wartość	Udział
0	0,77
1	0,23

Źródło: opracowanie własne.

Jak widać, nieco mniej niż jedna czwarta badanych firm posiadała wyodrębniony wydział B+R zajmujący się tworzeniem i testowaniem nowych rozwiązań.

Kolejna grupa pytań, które należy omówić, dotyczy wprowadzania innowacji produktowych, procesowych, organizacyjnych oraz marketingowych. Pytanie dotyczące innowacyjności firm miało następującą postać: „Czy w ciągu ostatnich 24 miesięcy w Państwa firmie wprowadzono rozwiązania, które można uznać za innowacyjne?". Tutaj również firmy odpowiadały „Tak" lub „Nie". Jeśli udzielona została odpowiedź „Tak", wówczas respondenci odpowiadali na kolejne pytanie: „Czy te rozwiązania dotyczyły: produktów, procesów zachodzących w firmie, organizacji firmy, marketingu?".

Firmy, które na pytanie dotyczące wprowadzenia innowacji odpowiadały pozytywnie, zaznaczały co najmniej jedną z czterech odpowiedzi dotyczących rodzaju innowacji. Respondenci mogli zaznaczyć wiele rodzajów innowacji, co wynika z faktu, że firmy mogły wprowadzać zarówno innowacje produktowe, jak i procesowe czy marketingowe. Na podstawie odpowiedzi udzielonych przez respondentów na zdefiniowane powyżej pytania skonstruowane zostały następujące zmienne zero-jedynkowe:

1) *INNOW_PROD* – przyjmuje wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację produktową,
2) *INNOW_PROC* – przyjmuje wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację procesową lub organizacyjną,
3) *INNOW_MARKT* – przyjmuje wartość 1 dla firm, które w ciągu ostatnich 24 miesięcy wprowadziły innowację marketingową.

Zamiast tworzenia dwóch odrębnych zmiennych (osobno związanych z wprowadzaniem innowacji procesowej i organizacyjnej) stworzona została pojedyncza zmienna, która informuje o wprowadzeniu co najmniej jednej z tych dwóch innowacji. Wynika to z faktu, że innowacje procesowe mają podobny charakter

do innowacji organizacyjnych. Tabela 26 prezentuje rozkłady empiryczne dla analizowanych zmiennych zero-jedynkowych.

Tabela 26. Empiryczny rozkład zmiennych binarnych związanych z wprowadzeniem określonych typów innowacji

Wartość	*INNOW_PROD*	*INNOW_PROC*	*INNOW_MARKT*
0	0,72	0,75	0,84
1	0,28	0,25	0,16

Źródło: opracowanie własne.

Jakąś innowację produktową w okresie co najwyżej 24 miesięcy przed badaniem wprowadziło 28% badanych firm. Nieco mniej, bo 25% badanych firm, wprowadziło innowację procesowo-organizacyjną, natomiast najniższy odsetek firm – około 16% – zdecydowało się na wprowadzenie innowacji z zakresu marketingu.

W dalszej kolejności należy omówić pytanie wykorzystywane do mierzenia produktywności firm. Według definicji ekonomicznej produktywność jest ilorazem wytworzonej oraz sprzedanej produkcji do ilości wytworzonych (zużytych) zasobów. Tymi zasobami mogą być na przykład pracownicy. Dlatego też często wykorzystywanym miernikiem produktywności jest wartość sprzedanych towarów na jednego zatrudnionego pracownika. Ponieważ informacje dotyczące wartości sprzedanych towarów mogą być traktowane jako dane poufne, w kwestionariuszu nie zostały ujęte pytania dotyczące tej wielkości. Dlatego jako aproksymanta produktywności przyjęte zostało średnie wynagrodzenie w firmie. Takie podejście zostało zastosowane między innymi w badaniu Thierry'ego *Lallemanda*, Roberta *Plasmana* i Francois *Rycxa* (2009), Giulii Faggio, Kjella Salvanesa i Johna Van Reenena (2010), Benoît Mahy'ego, Francois Rycxa i Mélanie Volral (2011) czy też w pracy Arendta i Grabowskiego (2017). W związku z tym pytanie wykorzystywane do mierzenia produktywności firm brzmiało: „Jakie jest średnie wynagrodzenie brutto pracowników pełnoetatowych w Państwa firmie?".

Tabela 27 prezentuje informacje dotyczącej średniej wartości, odchylenia standardowego oraz kwartyli dla analizowanej zmiennej.

Należy zwrócić uwagę, że średnie wynagrodzenie w próbie, wynoszące 2856 PLN, było zdecydowanie niższe od średniego wynagrodzenia w gospodarce krajowej w Polsce w 2015 roku. Nie świadczy to jednak absolutnie o braku reprezentatywności badania. Wynika to z co najmniej trzech powodów. Po pierwsze, średnie wynagrodzenie podawane przez Główny Urząd Statystyczny jest liczone bez uwzględnienia mikroprzedsiębiorstw. Wynagrodzenia

w mikroprzedsiębiorstwach są z reguły niższe. W analizowanym badaniu poświęconym innowacyjności firm i ich skłonności do korzystania z technologii informatycznych i komunikacyjnych przedsiębiorstwa zatrudniające mniej niż 10 osób także zostały uwzględnione. Po drugie, średnia z tabeli 27 ma charakter nieważony. Oznacza to zatem, że przedsiębiorstwa małe i duże uwzględnione są z tą samą wagą. Ponieważ korelacja między rozmiarem firmy a wynagrodzeniami pracowników jest dodatnia, uwzględnienie wag związanych z rozmiarem firm doprowadziłoby do wzrostu średniego wynagrodzenia. Oprócz tego należy pamiętać, że średnie wynagrodzenia w Polsce są często zawyżane przez uwzględnienie wynagrodzeń uzyskiwanych w takich dużych firmach państwowych jak na przykład KGHM Polska Miedź SA, PKN Orlen SA, Grupa Lotos SA. Największe firmy państwowe nie zostały uwzględnione w badaniu empirycznym.

Tabela 27. Podstawowe statystyki opisowe dla zmiennej *PRODUKTYWNOSC*

Miernik	Średnia	Odchylenie standardowe	Pierwszy kwartyl	Mediana	Trzeci kwartyl
Wartość	2856	1179	2000	2520	3352

Źródło: opracowanie własne.

Oprócz zmiennych endogenicznych związanych z decyzjami przedsiębiorstw dotyczącymi wykorzystania technologii informatycznych i komunikacyjnych, dokonania inwestycji w rozwój tych technologii, ich innowacyjności i produktywności, należy omówić zmienne, które traktowane są jako egzogeniczne w modelu ekonometrycznym. Wybór zmiennych jest rezultatem studiów literaturowych poświęconych czynnikom wpływającym na wykorzystanie TIiK, innowacyjność i produktywność (por. m.in. Ark, Piątkowski, 2004; Loof, Heshmati, 2006; Aghion, Howitt, 2008; Hall, Mairesse, Mohnen, 2010; Szczygielski, Grabowski, 2014; Bronzini, Piselli, 2016; Coad, Segarra, Teruel, 2016; Arendt, Grabowski, 2017). Jest on także uwarunkowany dostępnością danych, które mogą być uzyskane na podstawie odpowiedniego kwestionariusza. Tabela 28 prezentuje opis zmiennych rozważanych w badaniu empirycznym. Zawarte są w niej informacje dotyczące nazw zmiennych, sposobu mierzenia ich wartości oraz uzasadnienie (wraz z odwołaniem do literatury) dotyczące wykorzystania tych zmiennych w modelu wyjaśniającym skłonność do inwestowania w TIiK i korzystania z nich, innowacyjność i produktywność.

Tabela 28. Zmienne egzogeniczne rozważane w badaniu empirycznym

Nazwa zmiennej	Definicja	Oczekiwany wpływ zmiennej na wykorzystywanie TIiK, innowacyjność i produktywność
ROZMIAR	Logarytm z liczby osób zatrudnionych w firmie.	Oczekuje się dodatniego oszacowania parametru w każdym z równań. Wyniki innych badań wskazują na dodatnią zależność między rozmiarem a skłonnością do wprowadzania innowacji (Coad, Segarra, Teruel, 2016).
WWKK	Zmienna binarna przyjmująca wartość 1 w przypadku firm, w których większość kadry kierowniczej posiada wyższe wykształcenie.	Wysoki poziom wykształcenia kadry kierowniczej oraz pracowników szeregowych powinien prowadzić do wzrostu skłonności do wprowadzania zaawansowanych rozwiązań z zakresu TIiK, a zarazem do wzrostu innowacyjności i produktywności w przedsiębiorstwach.
WWP	Zmienna binarna przyjmująca wartość 1 w przypadku firm, w których większość szeregowych pracowników posiada wyższe wykształcenie.	
MSWKK	Zmienna binarna przyjmująca wartość 1 w przypadku firm stosujących motywacyjny system wynagradzania kadry kierowniczej.	Jeśli w przedsiębiorstwie funkcjonuje motywacyjny system wynagradzania, zarówno pracownicy, jak i kadra kierownicza mają większą skłonność do wprowadzania innowacyjnych rozwiązań (por. Grabowski, Skorupińska, 2015).
MSWP	Zmienna binarna przyjmująca wartość 1 w przypadku firm stosujących motywacyjny system wynagradzania szeregowych pracowników.	
Zasieg_KZ	Zmienna binarna przyjmująca wartość 1 w przypadku firm o ogólnokrajowym lub zagranicznym zasięgu oddziaływania.	Firmy, które są zmuszone konkurować z innymi na rynku ogólnokrajowym lub międzynarodowym, powinny inwestować w badania i rozwój, wprowadzać innowacje czy zaawansowane rozwiązania z zakresu TIiK (por. Arendt, Grabowski, 2017).
Ocena_ okresowa	Zmienna binarna przyjmująca wartość 1 w przypadku firm prowadzących okresową ocenę kompetencji pracowników pod kątem ich przydatności do potrzeb firmy.	Oddziaływanie faktu prowadzenia okresowej oceny kompetencji pracowników powinno być takie samo jak oddziaływanie motywacyjnego systemu wynagradzania.
SZKOLENIA_ TIiK	Zmienna binarna przyjmująca wartość 1 w przypadku firm organizujących dodatkowe szkolenia dla pracowników w związku z wdrażaniem TIiK.	Jeżeli pracownicy są lepiej przeszkoleni, inwestowanie w TIiK oraz wprowadzanie zaawansowanych narzędzi w tym zakresie jest uzasadnione (Ark, Piątkowski, 2004).

Nazwa zmiennej	Definicja	Oczekiwany wpływ zmiennej na wykorzystywanie TIiK, innowacyjność i produktywność
PRYWATNA	Zmienna binarna przyjmująca wartość 1 w przypadku firm prywatnych.	Oczekuje się dodatniego oszacowania parametru przy tej zmiennej we wszystkich równaniach.
BRANZA_ PRZEM	Zmienna binarna przyjmująca wartość 1 w przypadku firm z branży przemysłowej.	Trudno przewidzieć, czy przynależność do określonej branży ma wpływ zarówno na wykorzystanie zaawansowanych TIiK, jak i na wprowadzanie innowacji.
BRANZA_BUD	Zmienna binarna przyjmująca wartość 1 w przypadku firm z branży budowniczej.	
BRANZA_PHU	Zmienna binarna przyjmująca wartość 1 w przypadku firm z branży produkcyjno-handlowo-usługowej.	
ZES_ROB	Zmienna binarna przyjmująca wartość 1 dla firm, w których tworzone są zespoły robocze.	Oczekuje się dodatniego oszacowania parametru przy tej zmiennej w każdym z równań.
DZIEL_INF	Zmienna binarna przyjmująca wartość 1 w przypadku firm, w których istnieje zwyczaj dzielenia się z pracownikami informacjami istotnymi dla funkcjonowania firmy.	Dzielenie się informacjami z pracownikami powinno pozytywnie stymulować skłonność do inwestowania w technologie informatyczne i telekomunikacyjne oraz wprowadzanie innowacyjnych rozwiązań.
NOW_UM_INF	Zmienna binarna przyjmująca wartość 1 dla firm, w których wszyscy nowo przyjmowani pracownicy mają wysokie umiejętności informatyczne.	Wysoki poziom umiejętności informatycznych nowo przyjmowanych pracowników sprawia, że jest im łatwiej wykorzystywać nową aparaturę z zakresu TIiK. Dlatego też oczekuje się dodatniego oszacowania parametru przy omawianej zmiennej.
ORG	Zmienna ilustrująca gotowość firmy do przeprowadzenia zmiany organizacyjnej.	Jak wskazują m.in. Arendt i Grabowski (2017), gotowość do przeprowadzenia zmiany organizacyjnej jest ważnym czynnikiem komplementarnym wobec wykorzystywania TIiK.
WZROSTOWA	Zmienna przyjmująca wartość 1 dla przedsiębiorstw, które odpowiedziały, że zarówno w 2014, jak i 2013 roku przychód w firmie był większy niż w poprzednim roku.	Oczekuje się, że w firmach „rosnących" wyższa jest skłonność zarówno do inwestowania w technologie informatyczne i komunikacyjne oraz badania i rozwój, jak i wprowadzania innowacji (por. Szczygielski, Grabowski, Woodward, 2017).

Źródło: opracowanie własne.

Zgodnie z zawartym w punkcie 4.6.1 przeglądem literatury z zakresu czynników wpływających na innowacyjność firm wydaje się, że nie tylko cechy wewnętrzne determinują decyzje podejmowane przez zarządzających. Wyniki wielu badań pokazują, że nie można pomijać roli czynników zewnętrznych w kształtowaniu postaw innowacyjnych przedsiębiorstw. Dlatego też oprócz wyszczególnionych w tabeli 28 zmiennych obserwowalnych na poziomie firm warto uwzględnić kategorie dostępne na poziomie województw, informujące o tym, czy otoczenie zewnętrzne przedsiębiorstw sprzyja ich innowacyjności. Najbardziej naturalnym zbiorem informacji na temat innowacyjności otoczenia wydają się być dane pochodzące z Regional Innovation Scorebaord. Ponieważ badanie ankietowe wykorzystywane w niniejszej pracy zostało przeprowadzone w 2015 roku, dane dotyczące innowacyjności otoczenia pochodzą z okresu wcześniejszego. Dlatego też wykorzystywane są dane dotyczące innowacyjności regionalnej z 2014 roku. Tabela 29 prezentuje nazwy i opis zmiennych rozważanych w badaniu empirycznym. W przypadku każdej z nich oczekuje się dodatniej zależności między jej wartościami a skłonnością do wprowadzania innowacji i inwestowania w technologie informatyczne i telekomunikacyjne.

Tabela 29. Zmienne ilustrujące innowacyjność otoczenia rozważane w badaniu empirycznym

Zmienna	Definicja
RIS1	Udział mieszkańców z wyższym wykształceniem w populacji wszystkich osób w wieku 25–64 lat
RIS2	Wydatki na badania i rozwój w sektorze publicznym w relacji do PKB
RIS3	Wydatki na badania i rozwój w sektorze przedsiębiorstw w relacji do PKB
RIS4	Wydatki na innowacje firm małych i średnich (niezwiązane z wydatkami na badania i rozwój) w relacji do PKB
RIS5	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje wewnętrzne
RIS6	Odsetek innowacyjnych małych i średnich przedsiębiorstw współpracujących z innymi
RIS7	Aplikacje patentowe EPO w relacji do PKB
RIS8	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje produktowe lub procesowe
RIS9	Odsetek małych i średnich przedsiębiorstw wprowadzających innowacje marketingowe
RIS10	Odsetek zatrudnionych w przemysłach wysokiej i średniowysokiej technologii oraz usługach opartych na wiedzy
RIS11	Relacja sprzedaży produktów stanowiących innowacje nowe dla firmy lub nowe dla rynku do całkowitych obrotów

Źródło: opracowanie własne.

4.6.2.2. Specyfikacja modelu ekonometrycznego

Omówione powyżej zależności między cechami firm, poziomem innowacyjności regionu, w którym dana firma funkcjonuje, nakładami na badania i rozwój, nakładami na rozwój technologii informatycznych i komunikacyjnych, innowacjami w zakresie TIiK, innowacjami produktowymi, procesowo-organizacyjnymi i marketingowymi a produktywnością stanowią punkt wyjścia do badania.

Na początku szacowane są parametry równania wyjaśniającego skłonność do inwestowania w badania i rozwój oraz technologie informatyczne i telekomunikacyjne, a także wykorzystywania zaawansowanych narzędzi TIiK w zależności od cech firm. W celu uwzględnienia roli zdolności innowacyjnych regionów oraz losowych różnic między skłonnością do inwestowania w badania i rozwój oraz technologie informatyczne i komunikacyjne szacowane są parametry następującego wielopoziomowego wielorównaniowego modelu probitowego:

$$ICT\_ZP_i^* = \boldsymbol{x}_i^{ZP}\boldsymbol{\beta}^{ZP} + \boldsymbol{z}_i^{ZP}\boldsymbol{u}^{ZP} + \varepsilon_i^{ZP}, \tag{4.104a}$$

$$ICT\_ERP_i^* = \boldsymbol{x}_i^{ERP}\boldsymbol{\beta}^{ERP} + \boldsymbol{z}_i^{ERP}\boldsymbol{u}^{ERP} + \varepsilon_i^{ERP}, \tag{4.104b}$$

$$ICT\_CAD_i^* = \boldsymbol{x}_i^{CAD}\boldsymbol{\beta}^{CAD} + \boldsymbol{z}_i^{CAD}\boldsymbol{u}^{CAD} + \varepsilon_i^{CAD}, \tag{4.104c}$$

$$ICT\_SM_i^* = \boldsymbol{x}_i^{SM}\boldsymbol{\beta}^{SM} + \boldsymbol{z}_i^{SM}\boldsymbol{u}^{SM} + \varepsilon_i^{SM}, \tag{4.104d}$$

$$INWESTYCJE\_ICT_i^* = \boldsymbol{x}_i^{INW}\boldsymbol{\beta}^{INW} + \boldsymbol{z}_i^{INW}\boldsymbol{u}^{INW} + \varepsilon_i^{INW}, \tag{4.104e}$$

$$BR_i^* = \boldsymbol{x}_i^{BR}\boldsymbol{\beta}^{BR} + \boldsymbol{z}_i^{BR}\boldsymbol{u}^{BR} + \varepsilon_i^{BR}, \tag{4.104f}$$

$$ZM_i = I\left\{ZM_i^* > 0\right\}, \tag{4.104g}$$

dla *ZM* = *ICT_ZP*, *ICT_ERP*, *ICT_CAD*, *ICT_SM*, *INWESTYCJE_ICT*, *BR*.

Dla składników losowych przyjmowane jest założenie, że:

$$\left[\varepsilon_i^{ZP}\ \varepsilon_i^{ERP}\ \varepsilon_i^{CAD}\ \varepsilon_i^{SM}\quad \varepsilon_i^{INW}\ \varepsilon_i^{BR}\right]^T \sim N\left(0, \boldsymbol{\Sigma}\right), \tag{4.104h}$$

natomiast dla efektów losowych zakłada się, że:

$$\boldsymbol{u}^{ZM} \sim N\left(0, \boldsymbol{\Omega}^{ZM}\right) \tag{4.104i}$$

również dla *ZM* = *ICT_ZP*, *ICT_ERP*, *ICT_CAD*, *ICT_SM*, *INWESTYCJE_ICT*, *BR*.

Po oszacowaniu parametrów oraz efektów losowych dla wielopoziomowego wielorównaniowego modelu probitowego uwzględniane są one w celu obliczenia wartości teoretycznych dla zmiennych nieobserwowalnych. Wykorzystuje się w tym celu estymator wartości oczekiwanej zmiennej ukrytej, warunkowej względem efektów losowych:

$$\begin{aligned} \widehat{ICT\_ZP}_i^{*} &= E\left(ICT\_ZP_i^{*} \mid \hat{\boldsymbol{u}}^{ZP}\right) = \\ &= \boldsymbol{x}_i^{ZP}\hat{\boldsymbol{\beta}}^{ZP} + ICT\_ZP_i * a^1\left(ICT\_ZP_i\right) + \\ &\quad -\left(1-ICT\_ZP_i\right) * a^2\left(ICT\_ZP_i\right) \end{aligned} \tag{4.105a}$$

gdzie:

$$a^1\left(k_i\right) = \frac{\phi\left(\dfrac{\boldsymbol{x}_i^k\hat{\boldsymbol{\beta}}^k + \boldsymbol{z}_i^k\hat{\boldsymbol{u}}^k}{\sqrt{var\left(\boldsymbol{z}_i^k\boldsymbol{u}^k + \varepsilon_i^k\right)}}\right)}{\Phi\left(\dfrac{\boldsymbol{x}_i^k\hat{\boldsymbol{\beta}}^k + \boldsymbol{z}_i^k\hat{\boldsymbol{u}}^k}{\sqrt{var\left(\boldsymbol{z}_i^k\boldsymbol{u}^k + \varepsilon_i^k\right)}}\right)},$$

$$a^2\left(k_i\right) = \frac{\phi\left(\dfrac{\boldsymbol{x}_i^k\hat{\boldsymbol{\beta}}^k + \boldsymbol{z}_i^k\hat{\boldsymbol{u}}^k}{\sqrt{var\left(\boldsymbol{z}_i^k\boldsymbol{u}^k + \varepsilon_i^k\right)}}\right)}{1-\Phi\left(\dfrac{\boldsymbol{x}_i^k\hat{\boldsymbol{\beta}}^k + \boldsymbol{z}_i^k\hat{\boldsymbol{u}}^k}{\sqrt{var\left(\boldsymbol{z}_i^k\boldsymbol{u}^k + \varepsilon_i^k\right)}}\right)}.$$

Przez analogię definiowane są estymatory wartości oczekiwanych zmiennych ukrytych dla pozostałych zmiennych:

$$\begin{aligned} \widehat{ICT\_ERP}_i^{*} &= E\left(ICT\_ERP_i^{*} \mid \hat{\boldsymbol{u}}^{ERP}\right) = \\ &= \boldsymbol{x}_i^{ERP}\hat{\boldsymbol{\beta}}^{ERP} + ICT\_ERP_i * a^1\left(ERP_i\right) + \\ &\quad -\left(1-ICT_{ERPi}\right) * a^2\left(ERP_i\right) \end{aligned} \tag{4.105b}$$

$$\begin{aligned} ICT\_\hat{C}AD_i^* &= E\left(ICT\_CAD_i^* \mid \hat{\boldsymbol{u}}^{CAD}\right) = \\ &= \boldsymbol{x}_i^{CAD}\hat{\boldsymbol{\beta}}^{CAD} + ICT\_CAD_i * a^1\left(CAD_i\right) + \\ &\quad -\left(1-ICT\_CAD_i\right) * a^2\left(CAD_i\right) \end{aligned} \tag{4.105c}$$

$$\begin{aligned} ICT\_\hat{S}M_i^* &= E\left(ICT\_SM_i^* \mid \hat{\boldsymbol{u}}^{SM}\right) = \\ &= \boldsymbol{x}_i^{SM}\hat{\boldsymbol{\beta}}^{SM} + ICT\_SM_i * a^1\left(SM_i\right) + \\ &\quad -\left(1-ICT\_SM_i\right) * a^2\left(SM_i\right) \end{aligned} \tag{4.105d}$$

$$\begin{aligned} &INWESTYCJE\_\hat{I}CT_i^* = \\ &= E\left(INWESTYCJE\_ICT_i^* \mid \hat{\boldsymbol{u}}^{INW}\right) = \boldsymbol{x}_i^{INW}\hat{\boldsymbol{\beta}}^{INW} + \\ &\quad + INWESTYCJE\_ICT_i * a^1\left(INW_i\right) + \\ &\quad + \left(1-INWESTYCJE\_ICT_i\right) * a^2\left(INW_i\right) \end{aligned} \tag{4.105e}$$

$$B\hat{R}_i^* = E\left(BR_I^* \mid \hat{\boldsymbol{u}}^{CAD}\right) = \boldsymbol{x}_i^{BR}\hat{\boldsymbol{\beta}}^{BR} + BR_i * a^1\left(BR_i\right) + \left(1-BR_i\right) * a^2\left(BR_i\right).$$

Zgodnie z koncepcją modelu CDM, a także z ideą modelu opisanego na rysunku 6, nakłady na badania i rozwój oraz zaawansowane technologie informatyczne i telekomunikacyjne przyczyniają się do wzrostu innowacyjności przedsiębiorstw. Dlatego też wartości teoretyczne (4.105a)–(4.105f) wykorzystywane są jako zmienne objaśniające w modelu wyjaśniającym innowacyjność. Ponieważ wyróżnia się innowacje produktowe, procesowo-organizacyjne oraz marketingowe, w kolejnym kroku szacowane są parametry trójrównaniowego modelu probitowego:

$$\begin{aligned} INNOW\_PROD_i^* &= \boldsymbol{x}_i^{PROD}\boldsymbol{\beta}^{PROD} + \\ &+ \boldsymbol{ICT\_\hat{T}EOR}_i\boldsymbol{\lambda}_1 + \boldsymbol{z}_i^{PROD}\boldsymbol{u}^{PROD} + \varepsilon_i^{PROD}, \end{aligned} \tag{4.106a}$$

$$\begin{aligned} INNOW\_PROC_i^* &= \boldsymbol{x}_i^{PROC}\boldsymbol{\beta}^{PROC} + \boldsymbol{ICT\_\hat{T}EOR}_i\boldsymbol{\lambda}_2 + \\ &+ \boldsymbol{z}_i^{PROC}\boldsymbol{u}^{PROC} + \varepsilon_i^{PROC}, \end{aligned} \tag{4.106b}$$

$$INNOW\_MARK_i^* = \boldsymbol{x}_i^{MARK}\boldsymbol{\beta}^{MARK} + \\ + \boldsymbol{ICT\_\hat{T}EOR}_i\lambda_3 + \boldsymbol{z}_i^{MARK}\boldsymbol{u}^{MARK} + \varepsilon_i^{MARK}, \tag{4.106c}$$

$$ZM2_i = I\left\{ZM2_i^* > 0\right\}, \tag{4.106d}$$

dla *ZM2=INNOW_PROD, INNOW_PROCORG, INNOW_MARKT*, gdzie:

$$\boldsymbol{ICT\_\hat{T}EOR}_i = \\ \left[ICT\_\hat{Z}P_i^*\ ICT\_\hat{E}RP_i^*\ ICT\_\hat{C}AD_i^*\ ICT\_\hat{S}M_i^*\ INWESTYCJ\hat{E}\_ICT_i^*\ B\hat{R}_i^*\right].$$

Po oszacowaniu parametrów oraz efektów losowych dla wielopoziomowego, wielorównaniowego modelu probitowego wartości teoretyczne dla zmiennych nieobserwowalnych związanych z innowacyjnością oblicza się następująco:

$$INNOW\_\hat{P}ROD_i^* = \\ E\left(INNOW\_PROD_i^* \mid \hat{\boldsymbol{u}}^{PROD}\right) = \boldsymbol{x}_i^{PROD}\hat{\boldsymbol{\beta}}^{PROD} + \\ + INNOW\_PROD_i * a^1\left(PROD_i\right) - \\ - \left(1 - INNOW\_PROD_i\right) * a^2\left(PROD_i\right) \tag{4.107a}$$

$$INNOW\_\hat{P}ROC_i^* = \\ = E\left(INNOW\_PROC_i^* \mid \hat{\boldsymbol{u}}^{PROC}\right) = \boldsymbol{x}_i^{PROC}\hat{\boldsymbol{\beta}}^{PROC} + \\ + INNOW\_PROC_i * a^1\left(PROC_i\right) - \\ - \left(1 - INNOW\_PROC_i\right) * a^2\left(PROC_i\right), \tag{4.107b}$$

$$INNOW\_\hat{M}ARK_i^* = E\left(INNOW\_MARK_i^* \mid \hat{\boldsymbol{u}}^{MARK}\right) = \\ = \boldsymbol{x}_i^{MARK}\hat{\boldsymbol{\beta}}^{MARK} + INNOW\_MARK_i * a^1\left(MARK_i\right) + \\ - \left(1 - INNOW\_MARK_i\right) * a^2\left(MARK_i\right) \tag{4.107c}$$

Wartości teoretyczne (4.107a)–(4.107c) wykorzystywane są jako zmienne objaśniające w modelu wyjaśniającym produktywność. Rozważana jest estymacja parametrów następującego modelu:

$$PRODUKTYWNOSC_i = \boldsymbol{x}_i^{PRODUKT}\boldsymbol{\beta}^{PRODUKT} + \\ + \boldsymbol{INNOW\_\hat{TEOR}}_i\boldsymbol{\theta} + \varepsilon_i^{PRODUKT}, \tag{4.108}$$

gdzie:

$$\boldsymbol{INNOW\_\hat{TEOR}}_i = \\ \left[INNOW\_\hat{PROD}_i^* \quad INNOW\_\hat{PROC}_i^* \quad INNOW\_\hat{MARK}_i^* \right].$$

4.6.2.3. Wyniki estymacji i interpretacja

Tabela 30 prezentuje wyniki estymacji parametrów wielopoziomowego, wielorównaniowego modelu probitowego (4.104a)–(4.104i).

Tabela 30. Wyniki estymacji parametrów wielopoziomowego, wielorównaniowego modelu probitowego (w nawiasach podano średnie błędy szacunku)

Zmienne objaśniające	Równanie $ICT\_ZP_i^*$	Równanie $ICT\_ERP_i^*$	Równanie $ICT\_CAD_i^*$	Równanie $ICT\_SM_i^*$	Równanie $INWEST\_TIK_i^*$	Równanie BR_i^*
ROZMIAR	0,223*** (0,034)	0,287*** (0,032)	0,088** (0,036)	0,217*** (0,038)	0,144*** (0,030)	0,231*** (0,044)
Zasieg_KZ	0,303** (0,123)	–	0,417*** (0,129)	0,273** (0,138)	–	0,569*** (0,162)
ZES_ROB	0,240** (0,111)	0,185* (0,113)	0,289** (0,119)	0,221* (0,121)	–	0,587*** (0,207)
MSWKK	–	–	–	–	0,368** (0,155)	–
MSWP	–	0,163# (0,112)	–	–	–	–
WWKK	0,212* (0,130)	–	0,568*** (0,144)	–	0,394*** (0,111)	0,376** (0,182)
NOW_UM_INF	–	–	0,232** (0,117)	–	–	–
BRANZA_ PRZEM	1,232*** (0,148)	–	0,892*** (0,153)	1,283*** (0,156)	–	0,383*** (0,149)
BRANZA_BUD	0,459*** (0,161)	–	1,059*** (0,158)	0,380** (0,181)	–	–

Tabela 30 (cd.)

Zmienne objaśniające	Równanie $ICT\_ZP_i^*$	Równanie $ICT\_ERP_i^*$	Równanie $ICT\_CAD_i^*$	Równanie $ICT\_SM_i^*$	Równanie $INWEST\_TIK_i^*$	Równanie BR_i^*
BRANZA_PHU	1,197*** (0,123)	–	0,561*** (0,133)	1,093*** (0,136)	–	–
Ocena_ okresowa	–	0,200* (0,112)	–	–	–	0,422*** (0,136)
SZKOLENIA_ TIiK	–	0,167* (0,101)	0,180* (0,103)	0,163# (0,119)	0,857*** (0,103)	0,407*** (0,127)
DZIEL_INF	–	–	–	–	0,400*** (0,122)	–
WZROSTOWA	–	0,140# (0,098)	–	–	–	0,287** (0,119)
RIS4	–	–	–	–	–	2,781** (1,247)
RIS5	–	10,728*** (3,613)	–	–	–	–
RIS7	4,329** (2,020)	–	–	–	–	–
RIS9	–	–	–	–	3,832* (2,360)	–
RIS10	–	–	–	1,646** (0,684)	–	–
RIS11	–	–	2,184# (1,522)	–	–	–
Testowanie obecności efektów losowych (wnioskowanie na poziomie istotności 0,05)	Obecność efektów losowych dla wyrazu wolnego	Obecność efektów losowych dla wyrazu wolnego	Obecność efektów losowych dla wyrazu wolnego	Obecność efektów losowych dla wyrazu wolnego	Obecność efektów losowych dla wyrazu wolnego oraz parametru przy zmiennej *MSWKK*	Obecność efektów losowych dla wyrazu wolnego oraz parametru przy zmiennej *ZES_ROB*

***, **, *, # oznaczają odpowiednio istotność na poziomie 0,01, 0,05, 0,1 oraz 0,2.

Wyniki testowania wskazują, że w każdym z równań uwzględnienie części związanej z efektami losowymi jest uzasadnione. W przypadku skłonności do wprowadzania technologii informatycznych i komunikacyjnych w zakresie zarządzania produkcją, zarządzania zasobami przedsiębiorstwa, wsparcia dla projektowania i wytwarzania CAD/CAM, sterowania maszynami lub linią produkcyjną model z losowym wyrazem wolnym okazał się najlepszy. Oznacza to zatem, że między polskimi regionami występują losowe różnice w skłonności firm do wykorzystywania zaawansowanych narzędzi informatycznych. Wpływ poszczególnych zmiennych na skłonność do wykorzystywania

określonych narzędzi nie różni się istotnie między województwami. W przypadku równań wyjaśniających skłonność do inwestowania w technologie informatyczne i telekomunikacyjne oraz posiadania wyodrębnionego wydziału B+R okazało się, że model uwzględniający losowe zróżnicowanie parametrów strukturalnych jest lepszy.

Analizując wpływ poszczególnych kategorii na prawdopodobieństwa, że zmienne zależne przyjmują wartość 1, należy zwrócić uwagę na zmienną *ROZMIAR*. Była ona istotna w każdym z równań. Okazuje się, że przedsiębiorstwa są w stanie wprowadzać zaawansowane narzędzia informatyczne i telekomunikacyjne, inwestować w ich rozwój oraz posiadać własny dział B+R dopiero po osiągnięciu odpowiedniego rozmiaru (por. Coad, Segarra, Teruel, 2016). Wynik ten niesie ze sobą ważne implikacje dla polityki promowania rozwoju przedsiębiorczości. Powinna ona uwzględniać fakt, że mikroprzedsiębiorstwa i małe firmy często mają problem z pozyskaniem niezbędnych środków w celu zakupienia odpowiednich narzędzi. Dlatego też polityka wspierania małych firm powinna być nastawiona na udzielenie pomocy w celu pozyskania zaawansowanych narzędzi informatycznych i telekomunikacyjnych, tak aby możliwy był dalszy wzrost niewielkich podmiotów gospodarczych.

Fakt, że firma jest aktywna na rynku ogólnokrajowym oraz na rynkach zagranicznych, ma pozytywny wpływ na prawdopodobieństwo posiadania własnego wydziału B+R oraz niektórych zaawansowanych narzędzi z zakresu TIiK (w procesach biznesowych związanych z zarządzaniem produkcją, wsparciem dla projektowania i wytwarzania CAD/CAM oraz ze sterowaniem maszynami lub linią produkcyjną). Wynik ten nie budzi wątpliwości, ponieważ aktywność na rynkach innych niż lokalny zmusza przedsiębiorstwa do bycia konkurencyjnymi (por. Grabowski, Skorupińska, 2015). Konkurencyjność wymusza inwestowanie w badania i rozwój oraz zaawansowane narzędzia informatyczne i telekomunikacyjne.

Zmienna *ZES_ROB* okazała się być istotna w pięciu z sześciu równań. Uzyskane wyniki estymacji wskazują, że firmy, w których tworzone są zespoły robocze, mają większą skłonność do inwestowania w badania i rozwój oraz wprowadzania zaawansowanych narzędzi informatycznych i telekomunikacyjnych. Rezultat ten stanowi wskazówkę dotyczącą skuteczności metod zarządzania w przedsiębiorstwach. Okazuje się, że organizowanie zespołów roboczych jest czynnikiem wspomagającym dla zastosowania zaawansowanych narzędzi informatycznych i telekomunikacyjnych. Pracownicy tworzący te zespoły mogą dzielić się własną wiedzą z innymi. Współpraca w ramach tych zespołów może prowadzić do pojawienia się pomysłów wprowadzenia zaawansowanych narzędzi informatycznych i telekomunikacyjnych (por. Arendt, Grabowski, 2017).

Fakt wprowadzenia w przedsiębiorstwie motywacyjnego systemu wynagradzania kadry kierowniczej przyczynia się do wzrostu skłonności do inwestowania w rozwój technologii informatycznych i telekomunikacyjnych. Wynik ten stanowi wskazówkę dla właścicieli firm, którzy powinni mieć świadomość, że wprowadzenie

motywacyjnego systemu wynagradzania zarządzających może pozytywnie wpłynąć na ich decyzje w zakresie inwestowania w technologie przynoszące dalszy rozwój. Właściciele powinni dbać nie tylko o poziom motywacji wśród kadry kierowniczej, lecz także powinni zwrócić uwagę na poziom wykształcenia zarządzających. Okazuje się, że w firmach, w których większość kadry zarządzającej posiada wyższe wykształcenie, większa jest skłonność do inwestowania zarówno w badania i rozwój, jak i w narzędzia informatyczne i telekomunikacyjne. Wyższy poziom wykształcenia u większości osób z kadry zarządzającej sprzyja także wprowadzaniu narzędzi informatycznych i telekomunikacyjnych w zakresie zarządzania produkcją oraz wsparcia dla projektowania i wytwarzania CAD/CAM. Wynik ten nie budzi wątpliwości, ponieważ osoby lepiej wykształcone mają świadomość, że inwestowanie w technologie informatyczne i telekomunikacyjne oraz wprowadzanie zaawansowanych rozwiązań w tym zakresie może doprowadzić do poprawy wyników firmy. Fakt posiadania motywacyjnego systemu wynagradzania pracowników przyczynia się jedynie do wzrostu skłonności do wprowadzania technologii informatycznych i telekomunikacyjnych w procesach biznesowych związanych z zarządzaniem zasobami przedsiębiorstwa.

Zastanawiające jest to, że fakt, iż przedsiębiorstwo wymaga od nowo przyjmowanych pracowników wysokich umiejętności informatycznych, przyczynia się tylko i wyłącznie do wzrostu skłonności wprowadzania technologii w procesach biznesowych związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM. Skłonność do wprowadzania innych technologii informatycznych i telekomunikacyjnych, a także inwestowania w technologie nie jest istotnie wyższa w grupie firm zatrudniających tylko osoby o wysokich umiejętnościach informatycznych. Wynik ten nie jest zgodny z rezultatami uzyskanymi między innymi przez Marcina Piątkowskiego (2004).

Fakt, że przedsiębiorstwo zatrudnia tylko i wyłącznie pracowników na pełen etat, ma również wpływ na wykorzystanie zaawansowanych technologii informatycznych i telekomunikacyjnych w firmie. Dotyczy to przede wszystkim technologii wykorzystywanych w procesie zarządzania produkcją oraz zarządzania zasobami przedsiębiorstwa.

Przynależność do branży również determinuje skalę wykorzystania zaawansowanych technologii informatycznych i telekomunikacyjnych w przedsiębiorstwach. Firmy z branży przemysłowej oraz PHU najchętniej wprowadzają technologie z zakresu zarządzania produkcją. Nieco rzadziej technologie te wykorzystywane są w branży budowalnej. Najniższa skłonność do korzystania z analizowanych technologii obserwowana jest w firmach zajmujących się handlem oraz pozostałymi usługami. Podobny związek między przynależnością do branży a wykorzystywaniem technologii informatycznych i komunikacyjnych dotyczy procesów biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną. Firmy z branży budowalnej dominują w wykorzystaniu technologii informatycznych i komunikacyjnych w procesach biznesowych

związanych ze wsparciem dla projektowania i wytwarzania CAD/CAM. W przypadku tych technologii najniższa skłonność do ich wykorzystywania dotyczy przedsiębiorstw z branży handlowej oraz usług pozostałych. Przedsiębiorstwa przemysłowe istotnie częściej mają własne wydziały B+R w porównaniu z firmami z innych branż.

Kolejną kategorią związaną ze sposobem zarządzania firmą, mającą wpływ na wykorzystanie zaawansowanych technologii informatycznych i telekomunikacyjnych, jest prowadzenie okresowej oceny pracowników. Przedsiębiorstwa, które stosują taką formę ewaluacji, istotnie częściej posiadają własny dział B+R oraz wykorzystują technologie informatyczne i komunikacyjne w procesach związanych z zarządzaniem zasobami przedsiębiorstwa. Wynik ten nie powinien budzić wątpliwości, ponieważ pracownicy podlegający ocenie okresowej mają świadomość, że proponując innowacyjne rozwiązania, zyskują aprobatę ze strony zarządzających (Hall, Lotti, Mairesse, 2013).

Zmienna *SZKOLENIA_TIiK* okazała się istotna w większości rozważanych równań. Fakt prowadzenia szkoleń związanych z wykorzystywaniem technologii informatycznych i komunikacyjnych, przy innych czynnikach niezmienionych, przyczynia się do wzrostu skłonności do wykorzystywania tych technologii w takich procesach biznesowych, jak zarządzanie zasobami przedsiębiorstwa, wsparcie dla projektowania i wytwarzania CAD/CAM, sterowanie maszynami lub linią produkcyjną. Firmy organizujące szkolenia istotnie częściej inwestują w technologie oraz posiadają własny wydział B+R. Uzyskany rezultat nie budzi wątpliwości, jednak kierunek zależności może być przeciwny. Firmy inwestujące w zaawansowane TIiK są zmuszane przeprowadzać szkolenia, aby móc te technologie wykorzystywać (por. Ark, Piątkowski, 2004).

Strategia organizacyjna polegająca na dzieleniu się informacjami istotnymi dla funkcjonowania firmy z pracownikami ma wpływ na skłonność do inwestowania w technologie informatyczne i telekomunikacyjne. Dobrze poinformowani pracownicy mogą przekazać informację dotyczącą zasadności poczynienia konkretnych inwestycji na rzecz lepszego funkcjonowania firmy w przyszłości. Po raz kolejny widać, że nie tylko cechy, których nie da się łatwo zmienić (takie jak rozmiar czy branża), mają wpływ na skłonność do wprowadzania zaawansowanych rozwiązań z zakresu technologii informatycznych i komunikacyjnych. Sposób zarządzania wewnątrz przedsiębiorstwa również ma duże znaczenie. Wskazuje na to istotność zmiennej *DZIEL_INF* oraz innych kategorii związanych ze strategią organizacyjną.

Firmy „rosnące" istotnie częściej wykorzystują technologie z zakresu zarządzania zasobami przedsiębiorstwa oraz posiadają własny wydział B+R. Wynik ten nie budzi wątpliwości, zwłaszcza jeśli uwzględniony zostanie fakt, że rozmiar firmy również istotnie wpływa na decyzję przedsiębiorstwa o posiadaniu wydziału B+R. Istnieje zatem bariera dla firm małych oraz wolno rosnących. Dopiero przedsiębiorstwa duże lub zwiększające swoje przychody w sposób znaczący są w stanie zainwestować w badania i rozwój.

Spośród zmiennych związanych z poziomem innowacyjności regionów należy zwrócić uwagę na udział zatrudnienia w działalnościach wymagających wiedzy. Wraz ze wzrostem tego udziału obserwowany jest wzrost skłonności do wykorzystywania technologii informatycznych i komunikacyjnych w procesach biznesowych związanych ze sterowaniem maszynami lub linią produkcyjną. Jeśli w otoczeniu przedsiębiorstwa jest więcej firm sprzedających innowacje produktowe nowe dla firmy lub nowe dla rynku, wzrasta skłonność do wykorzystywania technologii w procesach zarządzania zasobami przedsiębiorstwa oraz wsparcia dla projektowania i wytwarzania CAD/CAM.

W tabeli 31 prezentowane są średnie oraz odchylenia standardowe efektów losowych dla poszczególnych województw w równaniach związanych z korzystaniem z zaawansowanych narzędzi informatycznych i telekomunikacyjnych. Są one wyznaczane na podstawie propozycji McCullocha (1997) omówionej w podrozdziale 4.4. Tabela 32 zawiera średnie oraz odchylenia standardowe efektów losowych dla poszczególnych województw, związane ze zmiennymi wyrazami wolnymi oraz zmiennymi parametrami przy określonych zmiennych w równaniach $INWEST\_TIK_i^*$ oraz BR_i^*.

Tabela 31. Średnie efekty losowe dla poszczególnych województw

Województwo	Równanie $ICT\_ZP_i^*$	Równanie $ICT\_ERP_i^*$	Równanie $ICT\_CAD_i^*$	Równanie $ICT\_SM_i^*$
dolnośląskie	0,141	0,191	0,247	0,130
kujawsko-pomorskie	0,171	0,421	0,977	0,253
lubelskie	0,074	0,009	–0,115	0,043
lubuskie	–0,040	0,142	0,250	0,007
łódzkie	0,140	–0,267	–0,313	0,150
małopolskie	–0,059	–0,017	0,174	–0,182
mazowieckie	–0,013	0,152	–0,205	–0,067
opolskie	0,079	0,255	0,057	–0,046
podkarpackie	0,049	0,562	–0,100	0,029
podlaskie	–0,071	–0,116	0,264	0,254
pomorskie	0,113	–0,399	–0,666	0,373
śląskie	–0,133	0,079	0,013	–0,246
świętokrzyskie	–0,158	–0,131	–0,116	–0,050
warmińsko-mazurskie	–0,030	0,012	0,273	–0,123
wielkopolskie	–0,081	–0,203	–0,147	–0,170
zachodniopomorskie	–0,183	–0,690	–0,593	–0,355

Źródło: opracowanie własne.

Tabela 32. Wartości oczekiwane oraz odchylenia standardowe efektów losowych dla poszczególnych województw

Województwo	Równanie $INWEST\_TIK_i^*$ – wyraz wolny	Równanie $INWEST\_TIK_i^*$ – parametr przy zmiennej MSWKK	Równanie BR_i^* – wyraz wolny	Równanie BR_i^* – parametr przy zmiennej ZES_ROB
dolnośląskie	–0,193	–0,693	–0,185	–0,478
kujawsko-pomorskie	–0,108	–0,253	–0,062	–0,030
lubelskie	–0,067	–0,134	0,140	–0,521
lubuskie	–0,018	–0,405	0,025	–0,104
łódzkie	0,193	0,517	–0,050	0,242
małopolskie	–0,121	–0,151	0,143	–0,370
mazowieckie	0,277	0,427	–0,191	0,593
opolskie	0,078	–0,033	–0,224	0,321
podkarpackie	0,041	0,165	–0,219	–0,573
podlaskie	–0,231	0,185	–0,026	0,188
pomorskie	0,077	0,530	0,124	0,830
śląskie	0,042	–0,143	0,036	–0,462
świętokrzyskie	–0,020	–0,228	–0,058	0,406
warmińsko-mazurskie	0,118	–0,672	0,364	–0,837
wielkopolskie	–0,131	0,608	–0,234	–0,061
zachodniopomorskie	0,063	0,278	0,417	0,857

Źródło: opracowanie własne.

Analizując wyniki zawarte w tabeli 31, należy zwrócić uwagę przede wszystkim na te województwa, w których wartości średnie efektów losowych są tylko dodatnie lub tylko ujemne. Przedsiębiorstwa zlokalizowane w województwie dolnośląskim oraz kujawsko-pomorskim, przy innych czynnikach niezmienionych, istotnie częściej wykorzystują zaawansowane technologie informatyczne i telekomunikacyjne w procesach biznesowych. Szczególnie dotyczy to firm z województwa mającego dwie stolice, gdyż predykcje efektów losowych są dla tego regionu naprawdę duże. Na drugim biegunie znajdują się firmy mające swoje siedziby w województwie świętokrzyskim, wielkopolskim oraz zachodniopomorskim. Przedsiębiorstwa zlokalizowane w tych regionach istotnie rzadziej korzystają z zaawansowanych technologii informatycznych i telekomunikacyjnych w porównaniu z firmami z innych części Polski. Istotność i wysokie co do modułu wartości efektów losowych świadczą o tym, że firmy zlokalizowane w tych samych regionach mają skłonność do przekazywania sobie informacji dotyczących zastosowania niektórych zaawansowanych rozwiązań z zakresu TIiK. Dzięki temu skłonność do korzystania z zaawansowanych technologii informatycznych i komunikacyjnych zależy nie tylko od cech firm i sposobu zarządzania, ale także wynika z jakości otoczenia przedsiębiorstwa.

Predykcje efektów losowych związanych z parametrami równania $INWEST\_TIK_i^*$ wskazują, że istnieją istotne międzyregionalne różnice w skłonnościach do inwestowania w technologie informatyczne i komunikacyjne. Skłonność ta, przy innych czynnikach niezmienionych, okazuje się być najwyższa w województwie łódzkim i mazowieckim. Z drugiej strony skłonność do inwestowania w analizowane technologie jest zdecydowanie niższa wśród firm z województwa dolnośląskiego i podlaskiego. Rezultat ten jest zgodny z wynikami innych badań dla polskich podregionów (por. m.in. Golejewska, 2018). Dodatkowo predykcje efektów losowych przy parametrze ilustrującym wpływ faktu posiadania motywacyjnego systemu wynagradzania kadry kierowniczej na skłonność do inwestowania w TIiK różni się między województwami. O ile w przypadku niektórych województw wpływ ten wydaje się być znikomy lub niezgodny z intuicją (dla województwa dolnośląskiego, lubuskiego i warmińsko-mazurskiego suma oszacowania parametru przy zmiennej *MSWKK* oraz oszacowania efektu losowego jest ujemna), o tyle w przypadku przedsiębiorstw zlokalizowanych w województwach wielkopolskim, pomorskim, łódzkim i mazowieckim fakt motywacyjnego wynagradzania kadry kierowniczej ma bardzo wyraźny wpływ na decyzje związane z inwestowaniem w technologie informatyczne i telekomunikacyjne.

Okazuje się, że najbliższe otoczenie firmy ma istotny wpływ na fakt posiadania wewnętrznego wydziału B+R. Jeśli weźmiemy dwie firmy o takich samych cechach, ale zlokalizowane w różnych województwach, to zauważymy, że skłonności do posiadania analizowanego wydziału różnią się od siebie w tych przedsiębiorstwach. Posiadanie własnego działu zajmującego się badaniami i rozwojem jest modne w przypadku przedsiębiorstw z województwa zachodniopomorskiego oraz warmińsko-mazurskiego. Z drugiej strony zdecydowanie niższa skłonność do posiadania analizowanego wydziału obserwowana jest w grupie przedsiębiorstw z województw: podkarpackiego, opolskiego i wielkopolskiego. Zależność między faktem tworzenia zespołów roboczych wewnątrz firmy a decyzją o posiadaniu wewnętrznego wydziału B+R również różni się między regionami. O ile w przypadku niektórych województw suma oszacowania parametru przy zmiennej *ZES_ROB* oraz oszacowania efektu losowego związanego z parametrem przy tej zmiennej jest tylko nieznaczenie wyższa od 0 lub nawet ujemna (dotyczy to takich województw jak: dolnośląskie, lubelskie, podkarpackie, świętokrzyskie, warmińsko-mazurskie), o tyle fakt tworzenia zespołów roboczych bardzo silnie wpływa na prawdopodobieństwo posiadania własnego wydziału B+R w przedsiębiorstwach z województwa mazowieckiego, pomorskiego i zachodniopomorskiego.

W kolejnym kroku szacowane są parametry modelu wyjaśniającego skłonność firm do wprowadzania różnych rodzajów innowacji. Wyniki estymacji parametrów wielopoziomowego, wielorównaniowego modelu probitowego prezentowane są w tabeli 33.

Tabela 33. Oszacowania parametrów wielopoziomowego, wielorównaniowego modelu probitowego wyjaśniającego skłonność do wprowadzania innowacji

Zmienne objaśniające	Równanie $INNOW\_PROD_i^*$	Równanie $INNOW\_PROC_i^*$	Równanie $INNOW\_MARK_i^*$
$ICT\_\hat{ZP}_i^*$	0,043** (0,021)	–	–
$ICT\_\hat{CAD}_i^*$	0,048# (0,037)	–	–
$ICT\_\hat{SM}_i^*$	–	0,066** (0,029)	–
$INWEST\_\hat{TIK}_i^*$	0,064** (0,027)	0,064** (0,027)	–
$\hat{BR}_i^*$	0,181*** (0,037)	0,166*** (0,035)	0,122*** (0,035)
ROZMIAR	0,055# (0,041)	0,224*** (0,039)	0,145*** (0,037)
WWKK	0,456*** (0,155)	–	–
MSWKK	0,202* (0,129)	–	–
BRANZA_PRZEM	0,512*** (0,172)	–	–
BRANZA_BUD	0,326* (0,185)	–	–
BRANZA_PHU	0,585*** (0,147)	–	–
UPWZ	0,394** (0,193)	0,599*** (0,195)	–
Zasieg_KZ	0,259* (0,139)	–	–
ZES_ROB	0,482*** (0,127)	0,480*** (0,127)	0,429*** (0,132)
Ocena_okresowa	–	0,324*** (0,121)	0,285** (0,126)
DZIEL_INF	–	0,451*** (0,168)	0,265# (0,173)
WZROSTOWA	–	0,223** (0,114)	0,248** (0,111)
RIS3	–	2,267# (1,562)	2,165* (1,309)
RIS4	3,058** (1,479)	–	–

Tabela 33 (cd.)

Zmienne objaśniające	Równanie $INNOW\_PROD_i^*$	Równanie $INNOW\_PROC_i^*$	Równanie $INNOW\_MARK_i^*$
Testowanie obecności efektów losowych (wnioskowanie na poziomie istotności 0,05)	Model z losowym wyrazem wolnym	Model z losowym wyrazem wolnym	Model z losowym wyrazem wolnym

Źródło: opracowanie własne.

Zawarte w tabeli 33 wyniki estymacji parametrów wskazują, że wprowadzenie zaawansowanych technologii informatycznych i telekomunikacyjnych pozytywnie wpływa na innowacyjność przedsiębiorstw. Najważniejszą determinantą okazał się jednak fakt posiadania wewnętrznego wydziału B+R. Firmy posiadające taki dział zdecydowanie częściej wprowadzają innowacje produktowe, procesowo-organizacyjne oraz marketingowe w porównaniu z pozostałymi przedsiębiorstwami. Wynik ten nie budzi wątpliwości i jest zgodny z rezultatami uzyskanymi na podstawie tych samych danych, zawartymi między innymi w pracach Arendta i Grabowskiego (2017; 2018). Celem wydziałów zajmujących się badaniem i rozwojem jest przedstawianie innowacyjnych rozwiązań zarządzającym przedsiębiorstwem. Dlatego też posiadanie takich wydziałów pozytywnie stymuluje innowacyjność.

Inwestowanie w technologie informatyczne i komunikacyjne jest czynnikiem sprzyjającym dla wprowadzania innowacji produktowych, procesowych oraz organizacyjnych. Dzięki posiadaniu nowoczesnych technologii zdecydowanie łatwiej jest wprowadzać na rynek zmodyfikowane produkty czy nowe procesy (Hall, Lotti, Mairesse, 2013). Zmienna związana z faktem inwestowania w technologie informatyczne i komunikacyjne okazała się jednak nieistotna w równaniu wyjaśniającym skłonność do wprowadzania innowacji marketingowych.

Posiadanie zaawansowanych narzędzi informatycznych i telekomunikacyjnych również nie ma wpływu na skłonność do wprowadzania innowacji marketingowych. Jest jednak czynnikiem sprzyjającym wprowadzaniu innowacji produktowych oraz procesowo-organizacyjnych. Przedsiębiorstwa wykorzystujące oprogramowanie do zarządzania produkcją oraz wspierające projektowanie i wytwarzanie CAD/CAM znacznie częściej wprowadzają innowacje produktowe w porównaniu z firmami niekorzystającymi z tych programów. Wynik ten nie budzi wątpliwości, ponieważ analizowane oprogramowanie ułatwia zarządzanie produkcją oraz pomaga wyspecyfikować zapotrzebowanie rynkowe. Sprzyja to powstawaniu nowych produktów. Posiadanie oprogramowania do sterowania maszynami lub linią produkcyjną pozytywnie wpływa na prawdopodobieństwo wprowadzenia innowacji procesowo-organizacyjnej (Wielicki, Arendt, 2010).

Wynik ten również nie powinien dziwić, ponieważ analizowane oprogramowanie umożliwia poprawę procesów produkcyjnych i wprowadzanie zmian w zakresie organizacji procesu produkcji.

Okazuje się, że rozmiar firmy ma istotny wpływ na skłonność do wprowadzania innowacji. Dotyczy to przede wszystkim innowacji procesowo-organizacyjnych oraz marketingowych. Okazuje się, że firma powinna być nie tylko duża, ale i „wzrostowa". W przedsiębiorstwach zatrudniających więcej pracowników, charakteryzujących się rosnącymi obrotami w całym okresie 2012–2014, prawdopodobieństwo wprowadzania innowacji marketingowych i procesowo-organizacyjnych jest zdecydowanie wyższe niż w grupie małych, „niewzrostowych" przedsiębiorstw. Firmy duże mogą więcej i to właśnie bardzo często młody wiek i zbyt mały rozmiar są barierami dla innowacyjności polskich przedsiębiorstw (por. Grabowski, Stawasz, 2017). Wynik ten jest ważną wskazówką dla polityki wspierania małych firm i mikroprzedsiębiorstw. Dotychczasowe (prowadzone przed 2015 rokiem) działania z zakresu polityki wsparcia okazały się niewystarczające do poprawienia innowacyjności małych firm.

Jakość kadry kierowniczej i polityka prowadzona wobec niej okazały się istotnymi determinantami w równaniu wyjaśniającym prawdopodobieństwo wprowadzenia innowacji produktowej. Prawdopodobieństwo wprowadzenia innowacji produktowej jest zdecydowanie wyższe w grupie firm, w których większość kadry kierowniczej ma wyższe wykształcenie oraz istnieje motywacyjny system wynagradzania zarządzających.

Przynależność do branży nie ma większego wpływu na decyzję o wprowadzeniu innowacji procesowo-organizacyjnej oraz marketingowej. W istotny sposób jednak determinuje prawdopodobieństwo wprowadzenia innowacji produktowej. Okazuje się, że przedsiębiorstwa handlowo-usługowe najczęściej decydują się na wprowadzenie tego rodzaju innowacji. Wynik ten nie jest zaskakujący, ponieważ głównym celem działalności tego typu firm jest wytwarzanie i sprzedaż produktów. Chcąc być konkurencyjnymi, firmy te zmuszone są wprowadzać nowinki do swojego asortymentu. Wprowadzanie nowych produktów na rynek jest też głównym celem większości firm przemysłowych. Dlatego też przedsiębiorstwa należące do tej branży relatywnie często wprowadzają innowacje produktowe. Nieco rzadziej innowacje produktowe wprowadzane są przez firmy z branży budowlanej. Najniższa skłonność do wprowadzania innowacji produktowych obserwowana jest w grupie firm zajmujących się handlem i pozostałymi usługami. Biorąc pod uwagę specyfikę działalności tego typu przedsiębiorstw, uzyskany rezultat nie budzi wątpliwości i jest zgodny z rezultatami innych badań dla Polski (por. m.in. Szczygielski, Grabowski, 2014; Lewandowska, Kowalski, 2015; Lewandowska, 2016; Świadek, Szajt, 2018).

Kolejna zmienna, która ma wpływ na prawdopodobieństwo wprowadzenia tylko innowacji produktowej, jest związana z zasięgiem działalności firmy. Przedsiębiorstwa

aktywne nie tylko na rynku lokalnym mają większą skłonność do wprowadzania innowacji produktowych. Wynika to zapewne z faktu, że konkurowanie na rynku krajowym i międzynarodowym wymaga ulepszania swoich produktów. Dlatego też przedsiębiorstwa aktywne na większej liczbie rynków znacznie częściej wprowadzają innowacje produktowe. Potrzeba wprowadzania innowacji marketingowych lub mających na celu poprawę procesu produkcji czy też nowych metod zarządzania ma niewiele wspólnego z zasięgiem aktywności firmy (por. Arendt, Grabowski, 2018). Dlatego też zmienna *Zasieg_KZ* okazała się nieistotna w równaniu wyjaśniającym prawdopodobieństwo wprowadzenia innowacji procesowo-organizacyjnej oraz marketingowej.

Warto zwrócić uwagę na ważną rolę, jaką tworzenie zespołów roboczych odgrywa w stymulowaniu innowacyjnych zachowań w grupie polskich przedsiębiorstw. Zmienna *ZES_ROB* okazała się być istotna w każdym z trzech równań. Jeśli wewnątrz przedsiębiorstwa organizowane są zespoły robocze, ich pracownicy skutecznie współpracują, wpadają na nowe pomysły, co następnie sprzyja wprowadzaniu innowacyjnych rozwiązań. Uzyskany rezultat jest ważną wskazówką dla tych przedsiębiorstw, w których nie są tworzone zespoły robocze. Firmy te powinny zastanowić się nad powołaniem takich zespołów, co w dalszej kolejności zwiększy wykorzystanie technologii informatycznych i telekomunikacyjnych oraz wprowadzanie innowacyjnych rozwiązań.

Kategoriami z zakresu metod zarządzania zasobami ludzkimi, które okazały się mieć istotny wpływ na prawdopodobieństwo wprowadzenia innowacji procesowo-organizacyjnej lub marketingowej, okazały się: fakt prowadzenia oceny okresowej oraz sytuacja, w której zarząd dzieli się informacjami z pracownikami. Ocena okresowa zmusza pracowników do aktywnego działania i poszukiwania nowych rozwiązań podnoszących wydajność firmy (por. Sidor-Rządkowska, 2015). Dlatego też oceniani pracownicy częściej wpadają na ciekawe pomysły, co prowadzi do wzrostu poziomu innowacyjności firmy. Jeśli zarząd przekazuje informacje dotyczące funkcjonowania firmy pracownikom szeregowym, to – przy innych czynnikach niezmienionych – następuje wzrost poziomu innowacyjności przedsiębiorstwa. Może to wynikać z dwóch przyczyn. Po pierwsze, pracownicy, którym przekazywana jest informacja na temat funkcjonowania firmy, mają większą świadomość, że wprowadzenie określonych rozwiązań może poprawić funkcjonowanie przedsiębiorstwa. Dlatego też osoby te częściej zgłaszają swoim pracodawcom pomysły, które mogą przyczynić się do wprowadzenia nowych rozwiązań, również tych innowacyjnych. Po drugie, wydaje się, że pracownicy lepiej poinformowani czują większą więź ze swoimi pracodawcami.

Oprócz wpływu zmiennych indywidualnych obserwowalnych na poziomie przedsiębiorstwa należy zwrócić uwagę na ważną rolę kategorii obserwowalnych na poziomie regionów. Relacja wydatków na badania i rozwój w sektorze

przedsiębiorstw do PKB oraz iloraz wydatków (niezwiązanych ze środkami przeznaczonymi na B+R) na innowacje firm małych i średnich okazały się istotnymi determinantami w wielorównaniowym modelu probitowym. Istotność zmiennych regionalnych wskazuje, że rola otoczenia w stymulowaniu innowacyjności firm jest bardzo ważna. Zachowania innowacyjne istotnie częściej obserwowane są w przedsiębiorstwach zlokalizowanych w województwach charakteryzujących się wyższym poziomem innowacyjności.

O znaczeniu otoczenia dla innowacyjności firm świadczy także istotność efektów losowych. Wyniki testu ilorazu wiarygodności wskazują, że wielopoziomowy model wielorównaniowy probitowy jest lepszy od modelu nieuwzględniającego obecności efektów losowych. Oznacza to zatem, że nie tylko omówione czynniki indywidualne oraz zmienne regionalne wpływają na skłonność firm do wprowadzania innowacji. Skłonności te różnią się między województwami w sposób losowy. Kooperacja i konkurencja między firmami zlokalizowanymi w tej samej jednostce administracyjnej sprawiają, że przedsiębiorstwa analizują zachowania swoich konkurentów w momencie podejmowania decyzji dotyczących wprowadzenia lub zaniechania innowacji. W tabeli 34 prezentowane są predykcje efektów losowych dla trzech równań oraz wszystkich województw.

Tabela 34. Średnie efekty losowe dla poszczególnych województw

Województwo	Równanie $INNOW\_PROD_i^*$	Równanie $INNOW\_PROC_i^*$	Równanie $INNOW\_MARK_i^*$
dolnośląskie	0,657	0,496	0,342
kujawsko-pomorskie	–0,215	0,364	0,068
lubelskie	–0,266	0,287	0,221
lubuskie	–0,223	–0,220	–0,122
łódzkie	–0,038	0,010	–0,153
małopolskie	0,231	–0,019	0,050
mazowieckie	0,276	–0,178	–0,094
opolskie	0,119	–0,256	–0,260
podkarpackie	–0,488	–0,042	0,003
podlaskie	0,202	0,311	0,302
pomorskie	0,064	0,233	0,082
śląskie	–0,477	–0,481	–0,293
świętokrzyskie	0,025	–0,357	–0,276
warmińsko-mazurskie	–0,071	0,004	0,002
wielkopolskie	–0,249	–0,032	–0,096
zachodniopomorskie	0,453	–0,118	0,224

Źródło: opracowanie własne.

Predykcje efektów losowych zawarte w tabeli 34 wskazują, że – przy innych czynnikach niezmienionych – najwyższa skłonność do wprowadzania innowacji produktowych obserwowana jest w województwie dolnośląskim, zachodniopomorskim i mazowieckim. Wyniki analiz polityki innowacyjnej przedsiębiorstw pochodzące z Community Innovation Survey rzeczywiście wskazują, że firmy zlokalizowane w województwie mazowieckim oraz dolnośląskim najczęściej deklarują wprowadzanie innowacji produktowych. Wysokie oszacowanie efektu losowego dla województwa zachodniopomorskiego wynika z faktu, że firmy z tego regionu również często wprowadzają analizowany rodzaj innowacji, a wartości przyjmowane przez zmienne objaśniające nie należą w tym przypadku do najwyższych. Dotyczy to zwłaszcza inwestowania w zaawansowane oprogramowanie z zakresu TIiK. Firmy z analizowanego regionu relatywnie rzadko wykorzystują oprogramowanie do zarządzania produkcją oraz wspierające projektowanie i wytwarzanie CAD/CAM, a mimo tego często wprowadzają innowacje produktowe. Najniższa skłonność do wprowadzania innowacji produktowych dotyczy takich województw, jak śląskie, podkarpackie czy lubelskie. Uzasadnienia dla takiego wyniku różnią się między regionami. W przypadku województwa śląskiego należy zauważyć, że działalności dominujące w tym regionie nie znajdują się w fazie wzrostowej. Dlatego też przedsiębiorstwa wykonujące te działalności nie wprowadzają innowacji. Oprócz tego aktywność firm zlokalizowanych w analizowanym województwie charakteryzuje się niską dynamiką. Na ogół mają one od lat ustalonych kooperantów. Stopa „urodzeń" nowych firm jest niska, co powoduje, że istniejące przedsiębiorstwa mają niższą motywację do konkurowania przez wprowadzanie innowacji produktowych (por. Arendt, Grabowski, 2017). Niższa skłonność do wprowadzania innowacji produktowych w przypadku firm z województwa lubelskiego może wynikać z faktu, że natrafiają one w swoim rozwoju na barierę finansową. Przedsiębiorstwa zlokalizowane w tym regionie charakteryzują się gorszą sytuacją finansową w porównaniu z firmami w innych częściach Polski. Ujemne saldo migracji nie zachęca do konkurowania o lokalnych odbiorców przez wprowadzanie innowacji produktowych. Podobna tendencja dotyczy przedsiębiorstw zlokalizowanych w województwie podkarpackim. Przedsiębiorstwa z województwa dolnośląskiego cechuje nie tylko wysoka skłonność do wprowadzania innowacji produktowych, ale także większe zaangażowanie w udoskonalanie procesów produkcji czy też metod zarządzania. Pewną rolę może tu odgrywać przygraniczne położenie analizowanego regionu. Województwo dolnośląskie graniczy zarówno z charakteryzującymi się bardzo wysokim poziomem innowacyjności Niemcami, jak i bardziej innowacyjnymi od Polski Czechami[2].

2 Wyższy poziom innowacyjności Czech (w porównaniu z Polską) potwierdzają wyniki badań przeprowadzonych w ramach European Innovation Scoreboard.

Współpraca i rywalizacja o rynek zbytu z firmami zlokalizowanymi pod drugiej stronie granicy wymusza wzrost konkurencyjności. Dlatego też przedsiębiorstwa z analizowanego regionu udoskonalają metody produkcji oraz chętniej (w porównaniu z firmami z innych części Polski) udoskonalają metody zarządzania (por. Grabowski, Stawasz, 2017). Oprócz tego wysoką skłonnością do wprowadzania innowacji procesowych charakteryzują się przedsiębiorstwa z województwa kujawsko-pomorskiego i podlaskiego. Niski poziom innowacyjności dotyczy natomiast firm znajdujących się w województwie śląskim i świętokrzyskim. Działalności dominujące w tych regionach (związane z górnictwem, hutnictwem czy przetwarzaniem) na ogół znajdują się w fazie schyłkowej. Dlatego też funkcjonujące tam przedsiębiorstwa relatywnie rzadko unowocześniają procesy produkcji. Jak się okazuje, województwo dolnośląskie również dominuje, jeśli chodzi o skłonności firm do wprowadzania innowacji marketingowych. Ten rodzaj innowacji jest też często wprowadzany przez przedsiębiorstwa z województwa podlaskiego. Prawdopodobieństwo wprowadzenia innowacji marketingowej jest, przy innych czynnikach niezmienionych, najniższe w grupie firm z województw: opolskiego, śląskiego oraz świętokrzyskiego.

Tabela 35. Oszacowania parametrów modelu wyjaśniającego produktywność w polskich przedsiębiorstwach wykorzystujących technologie informatyczne i telekomunikacyjne[a)]

Zmienna	Oszacowanie	Błąd standardowy	Graniczny poziom istotności
$INNOW\_\hat{P}ROD_i^*$	0,013	0,008	0,097
$INNOW\_\hat{M}ARK_i^*$	0,013	0,006	0,036
$ICT\_\hat{E}RP_i^*$	0,026	0,007	0,000
WWP	0,127	0,041	0,002
MSWP	0,099	0,021	0,000
ECP	0,145	0,045	0,002
SZKOLENIA_TIiK	0,069	0,023	0,003
BRANZA_BUD	0,091	0,032	0,004

[a)] Zmienną zależną jest logarytm ze średniego poziomu wynagrodzenia w firmie.

Źródło: opracowanie własne.

Po obliczeniu teoretycznych wartości dla zmiennych nieobserwowalnych związanych z wprowadzaniem innowacji wykonywany jest trzeci krok analizowanej procedury. Szacowane są parametry liniowego modelu wielopoziomowego wyjaśniającego poziom produktywności w badanych przedsiębiorstwach. Oszacowania parametrów prezentowane są w tabeli 35, natomiast średnie wartości i odchylenia standardowe dla efektów losowych zawiera tabela 36.

Tabela 36. Predykcje efektów losowych dla liniowego modelu wielopoziomowego

Województwo	Wartość oczekiwana efektu losowego dla wyrazu wolnego	Odchylenie standardowe efektu losowego dla wyrazu wolnego	Wartość oczekiwana efektu losowego dla parametru przy zmiennej WWP	Wartość oczekiwana efektu losowego dla parametru przy zmiennej WWP
dolnośląskie	0,131	0,063	–0,060	0,037
kujawsko-pomorskie	0,004	0,102	–0,142	0,045
lubelskie	–0,119	0,081	–0,090	0,047
lubuskie	0,041	0,102	–0,148	0,044
łódzkie	–0,051	0,079	0,135	0,061
małopolskie	0,098	0,065	0,065	0,041
mazowieckie	0,181	0,059	0,064	0,030
opolskie	–0,052	0,083	0,053	0,047
podkarpackie	0,087	0,109	–0,219	0,047
podlaskie	0,003	0,080	0,180	0,044
pomorskie	–0,020	0,065	0,038	0,042
śląskie	0,102	0,065	–0,203	0,030
świętokrzyskie	0,011	0,071	–0,004	0,045
warmińsko-mazurskie	–0,064	0,079	–0,041	0,046
wielkopolskie	–0,119	0,071	0,073	0,036
zachodniopomorskie	–0,234	0,075	0,298	0,059

Źródło: opracowanie własne.

Wyniki estymacji parametrów liniowego modelu wielopoziomowego wskazują, że wraz ze wzrostem innowacyjności przedsiębiorstw obserwowany był wzrost produktywności. Dotyczyło to innowacji produktowych oraz marketingowych. Przeciętne wynagrodzenia w przedsiębiorstwach wprowadzających analizowane rodzaje innowacji okazały się, przy innych czynnikach niezmienionych, istotnie wyższe. Wynik ten nie budzi wątpliwości, ponieważ efektem innowacji produktowych powinien być wzrost sprzedaży, co następnie powinno prowadzić do wzrostu zysków firmy i podwyższenia wynagrodzeń. Innowacje marketingowe, polegające na przykład na opracowaniu nowej metody dostawy produktów, powinny prowadzić do wzrostu wydajności, co następnie może mieć pozytywny wpływ na wynagrodzenia. Brak istotności zmiennej związanej ze skłonnością do wprowadzania innowacji procesowych w równaniu wyjaśniającym produktywność jest zgodny z wynikami badań uzyskanymi przez Arendta i Grabowskiego (2018). Może to wynikać z faktu, że unowocześnienie procesu produkcji często wiąże się z zastąpieniem siły roboczej przez maszyny – wówczas popyt na pracę jest w takich przedsiębiorstwach niższy. Pracownicy tych firm mogą liczyć na gorsze warunki płacowe.

Zmienna $ICT\_\hat{ERP}_i^*$ okazała się mieć pozytywny wpływ na poziom produktywności w polskich przedsiębiorstwach. Wykorzystywanie oprogramowania do zarządzania zasobami przedsiębiorstwa (ERP) przyczynia się, przy innych czynnikach niezmienionych, do wzrostu produktywności. Wynik ten jest zgodny z rezultatem uzyskanym między innymi w pracy Arendta i Grabowskiego (2018), którzy pokazali, że posiadanie tego typu narzędzia z zakresu TIiK w największym stopniu determinuje wynagrodzenia. Oznacza to zatem, że firmy, które nie zakupiły oprogramowania do zarządzania zasobami przedsiębiorstwa, powinny rozważyć jego nabycie. Taka strategia pozwoli w przyszłości efektywniej zarządzać zasobami przedsiębiorstwa, zredukować koszty działalności i doprowadzić do wyższej produktywności. Zmienne związane z wykorzystaniem innych technologii informatycznych i komunikacyjnych okazały się być nieistotne w równaniu produktywności. Należy jednak zauważyć, że korzystanie z innego rodzaju oprogramowania w sposób pośredni wpływa na produktywność firmy. Firmy wykorzystujące oprogramowanie do zarządzania produkcją oraz programy wspierające projektowanie i wytwarzanie CAD/CAM istotnie częściej decydują się na wprowadzenie innowacji produktowej, co następnie prowadzi do wzrostu ich produktywności.

Poziom wykształcenia pracowników szeregowych oraz system ich wynagradzania również mają istotny wpływ na poziom produktywności w firmie. Pozytywna zależność między średnimi wynagrodzeniami a udziałem pracowników szeregowych posiadających wyższe wykształcenie jest zgodna z Mincerowską (Mincer, Polachek, 1974) koncepcją dodatniej stopy zwrotu z wykształcenia oraz potwierdza wyniki badania empirycznego analizowanego w rozdziale trzecim. Wyższe wynagrodzenia obserwowane są również – *ceteris paribus* – w firmach stosujących motywacyjny system wynagradzania pracowników. Rezultat ten można interpretować na dwa sposoby. Po pierwsze, jeśli w danej firmie oprócz zasadniczego wynagrodzenia pracownicy uzyskują premie i nagrody, ich średnie pensje są wyższe. Po drugie, system premii i zachęt powinien motywować pracowników do bardziej efektywnej pracy, co następnie może doprowadzić do wzrostu produktywności (por. Pocztowski, Pauli, 2013).

Należy również zwrócić uwagę na rolę wprowadzania elastycznego czasu pracy jako czynnika podnoszącego poziom produktywności. Wynik ten również można interpretować na dwa sposoby. Po pierwsze, pracownicy, którym łatwiej jest połączyć wykonywanie zawodu z obowiązkami domowymi, są bardziej wydajni. Prowadzi to do wzrostu produktywności firmy, a zarząd jest w stanie zapewnić swoim pracownikom wyższe wynagrodzenia. Po drugie, przywilej elastycznego czasu pracy często przysługuje specjalistom wykonującym swoją pracę w sposób zdalny. Osoby te mogą obiektywnie liczyć na wyższe wynagrodzenia.

Fakt prowadzenia regularnych szkoleń z zakresu wykorzystania technologii informatycznych i telekomunikacyjnych również jest ważną stymulantą dla poziomu

produktywności. Wynika to z faktu, że dobrze przeszkoleni pracownicy lepiej wykonują swoje zadania i są bardziej efektywni (Wielicki, Arendt, 2010). Wzrost zysków sprawia, że przedsiębiorstwo jest w stanie zaoferować wyższe wynagrodzenia swoim pracownikom.

Chociaż podział na branże uzyskany podczas prowadzenia tego badania jest zdecydowanie mniej szczegółowy niż podział na sekcje PKD wykorzystywany w badaniu prezentowanym w rozdziale trzecim, należy zwrócić uwagę na zróżnicowanie średnich wynagrodzeń w poszczególnych branżach. Istotnie wyższe średnie wynagrodzenia odnotowano, przy innych czynnikach niezmienionych, w branży budowlanej. Między pozostałymi analizowanymi branżami nie zaobserwowano istotnych różnic w poziomie produktywności.

Predykcje efektów losowych związanych z wyrazem wolnym dla poszczególnych województw wskazują, że poziom produktywności zależy nie tylko od ustalonych cech firm, ale także w sposób losowy jest zdeterminowany przynależnością przedsiębiorstwa do określonego regionu. Firmy zlokalizowane blisko siebie nie mogą oferować zdecydowanie różniących się wynagrodzeń za tę samą pracę. Gdyby tak było, pracownicy opuszczaliby jedno przedsiębiorstwo i zatrudnialiby się w innym. Z drugiej strony, z punktu widzenia przedsiębiorstwa, oferowanie zbyt wysokich wynagrodzeń prowadzi do spadku ich konkurencyjności kosztowej. Dlatego też występowanie efektów losowych wskazujących na międzyregionalne zróżnicowanie średnich wynagrodzeń jest zgodne z oczekiwaniami oraz rezultatami badania empirycznego przeprowadzonego w rozdziale trzecim. Najwyższe predykcje efektów losowych odnoszą się do firm z województw mazowieckiego, dolnośląskiego i śląskiego. Wynik ten jest zgodny z intuicją, ponieważ w analizowanych regionach wynagrodzenia są rzeczywiście najwyższe (por. Tokarski, 2013; Arendt, Grabowski, 2018). Z drugiej strony najniższe efekty losowe odnotowywane są dla województw lubelskiego, wielkopolskiego i zachodniopomorskiego. O ile w przypadku pierwszego z nich uzyskany rezultat pokrywa się z analizą międzyregionalnego zróżnicowania wynagrodzeń, o tyle w przypadku województw ze stolicami w Poznaniu i Szczecinie średnie wynagrodzenia oferowane w przedsiębiorstwach tam zlokalizowanych nie należą do najniższych w kraju. Niemniej jednak analizowane regiony charakteryzują się wysokim poziomem innowacyjności, której nie towarzyszą najwyższe wynagrodzenia. Dlatego też efekty losowe związane z wyrazem wolnym są w przypadku analizowanych województw istotnie ujemne.

Wartości oczekiwane efektów losowych dla parametru przy zmiennej *WWP* wskazują, że zależność między poziomem wykształcenia kadry niekierowniczej a produktywnością różni się między województwami. Najsłabszą zależność odnotowuje się w województwach podkarpackim, śląskim, kujawsko-pomorskim i lubuskim. Charakteryzują się one relatywnie niskim udziałem pracowników z wyższym

wykształceniem. Oprócz tego w analizowanych regionach w strukturze zatrudnienia dominują sekcje charakteryzujące się niskim zapotrzebowaniem na wykwalifikowaną kadrę. Dlatego też we wspomnianych województwach najczęściej obserwowana jest taka sytuacja, że osoby posiadające wyższe wykształcenie wykonują prace proste. Tak więc poziom wykształcenia pracowników szeregowych nie ma większego wpływu na wynagrodzenia. Z drugiej strony województwa łódzkie, zachodniopomorskie oraz podlaskie charakteryzują się bardzo silną zależnością między poziomem wykształcenia pracowników szeregowych a produktywnością. Analizując te regiony z punktu widzenia struktury zatrudnienia ze względu na sekcje PKD, widoczne jest wyraźne zróżnicowanie. Dużo osób pracuje w sekcjach charakteryzujących niską wydajnością pracy i zatrudniających pracowników o niskim poziomie kwalifikacji. Z drugiej strony zatrudnienie w sekcjach wymagających wysokich kwalifikacji od pracowników również nie jest małe. Dlatego też obserwowane są duże różnice między wynagrodzeniami pracowników usług opartych na wiedzy i przemysłów średniowysokiej i wysokiej technologii a zarobkami osób pracujących w firmach wymagających od zatrudnionych niższego poziomu kwalifikacji.

Oszacowania parametrów wielopoziomowego, wielorównaniowego, rekurencyjnego modelu probitowego nie mają dokładnej interpretacji. Po znakach oszacowań możliwe jest stwierdzenie, że określone cechy firm wpływają pozytywnie lub negatywnie na skłonność do posiadania zaawansowanych technologii informacyjnych i komunikacyjnych, inwestowania w ich rozwój czy też wprowadzania innowacji. Oprócz tego niektóre zmienne występują w różnych równaniach. Dlatego też, aby zaprezentować efekt netto zmiany zmiennej objaśniającej na zmienną zależną, dla wszystkich binarnych zmiennych objaśniających szacowane są następujące wielkości:

$$ef\_net_f = \frac{\sum_{i=1}^{I}\left(P\left(y_i = 1 \mid x_f = 1, \boldsymbol{x}_{i\{\setminus f\},} z_i, \hat{\boldsymbol{u}}\right) - P\left(y_i = 1 \mid x_f = 0, \boldsymbol{x}_{i\{\setminus f\},} z_i, \hat{\boldsymbol{u}}\right)\right)}{I}, \tag{4.109}$$

gdzie $\boldsymbol{x}_{i\{/f\}}$ jest wektorem obserwacji dla wszystkich zmiennych z pominięciem zmiennej *f*-tej. Należy dodać, że wzór (4.109) przedstawia średnią wartości efektu krańcowego względem *f*-tej zmiennej. Dzięki zastosowaniu wzoru (4.109) można wskazać zmienne binarne o najsilniejszym i najsłabszym wpływie na zmianę prawdopodobieństwa sukcesu dla zmiennej endogenicznej. Średnie efekty krańcowe netto można policzyć także względem ciągłych zmiennych objaśniających. Wówczas należy skorzystać z formuły różniczkowej.

Tabela 37. Oszacowania efektów netto wskazujących na wpływ egzogenicznych zmiennych binarnych na prawdopodobieństwo, iż dana zmienna zależna przyjmuje wartość 1

	ICT_ZP	*ICT_ERP*	*ICT_CAD*	*ICT_SM*	*INVEST_TIiK*	*BR*	*INNOW_ PROD*	*INNOW_ PROC*	*INNOW_ MARK*
Zasieg_KZ	0,29	–	0,22	0,24	–	0,27	0,29	0,16	0,09
ZES_ROB	0,29	0,24	0,19	0,25	–	0,39	0,37	0,33	0,23
MSWKK	–	–	–	–	0,32	–	0,23	–	–
MSWP	–	0,18	–	–	–	–	–	–	–
WWKK	0,22	–	0,19	–	0,33	0,23	0,25	0,15	0,12
NOW_UM_INF	–	–	0,07	–	–	–	0,09	–	–
BRANZA_ PRZEM	0,46	–	0,31	0,44	–	0,28	0,27	0,20	0,08
BRANZA_BUD	–0,04	–	0,17	–0,06	–	–	–0,02	0,07	–
BRANZA_PHU	0,30	–	0,07	0,21	–	–	0,24	0,02	–
Ocena_ okresowa	–	0,20	–	–	–	0,30	–	0,26	0,16
SZKOLENIA_ TIiK	–	0,19	0,14	0,15	0,43	–	0,27	–	–
DZIEL_INF	–	–	–	–	0,25	–	0,23	0,17	0,11
WZROSTOWA	–	0,11	–	–	–	0,16	–	0,14	0,12
UPWZ	–	–	–	–	–	–	0,18	0,09	–

Źródło: opracowanie własne.

Uzyskane oszacowania parametrów wskazują na średnie różnice w prawdopodobieństwach przyjęcia przez zmienną zależną wartości 1 w zależności od wartości przyjmowanych przez zmienne egzogeniczne. Na przykład wartość 0,29 należy interpretować jako średnią różnicę między prawdopodobieństwem wykorzystywania TIiK w procesach związanych z zarządzaniem produkcją między firmami o zasięgu co najmniej krajowym w porównaniu z przedsiębiorstwami o zasięgu lokalnym. Znaki dla tych różnic są zgodne ze znakami oszacowań w wielopoziomowym, wielorównaniowym modelu rekurencyjnym. Jednak w przeciwieństwie do oszacowań prezentowanych w tabelach 30–36 wielkości z tabeli 37 mają interpretację. Należy zauważyć, że prezentowana strategia (estymacja parametrów modelu rekurencyjnego i szacowanie średnich efektów krańcowych netto) jest uzasadniona, jeśli badacz chciałby przeanalizować zależności między kategoriami endogenicznymi i jednocześnie ocenić wpływ kategorii egzogenicznych na każdą zmienną wynikową. Wynika to z faktu, że uwzględnienie wartości teoretycznych zależnych od zmiennych egzogenicznych sprawia, iż wpływ poszczególnych regresorów uwzględniany jest więcej niż jednokrotnie. Dlatego też na podstawie pojedynczych oszacowań trudno wnioskować na temat wpływu poszczególnych regresorów na zmienną zależną. W związku z tym należy w takich sytuacjach skorzystać z formuły (4.109) dla przypadku dyskretnych regresorów lub analogicznej dla regresorów ciągłych. Takie techniki są powszechnie stosowane, gdy zadaniem badacza jest analiza złożonego schematu zależności, a w systemie dodatkowo występują zmienne egzogeniczne (por. np. Grabowski, Stawasz, 2017).

Tabela 38. Spadek wartości predykcyjnej modelu po usunięciu określonych grup zmiennych

	Model bez efektów losowych dla województw oraz zmiennych regionalnych pochodzących z RIS (w %)	Model bez efektów losowych dla województw, ale ze zmiennymi regionalnymi pochodzącymi z RIS (w %)	Model z efektami losowymi dla województw, ale bez zmiennych regionalnych pochodzących z RIS (w %)
ICT_ZP	–14,7	–12,3	–2,8
ICT_ERP	–20,6	–14,1	–7,1
ICT_CAD	–16,0	–13,5	–3,1
ICT_SM	–18,5	–16,2	–2,7
INWEST_TIK	–18,1	–15,2	–3,2
BR	–16,2	–13,5	–3,0
INNOW_PROD	–23,7	–19,2	–5,1
INNOW_PROC	–24,5	–20,1	–5,3
INNOW_MARK	–25,1	–20,6	–5,4

Źródło: opracowanie własne.

W celu sprawdzenia, czy uwzględnienie efektów losowych dla województw oraz zmiennych ilustrujących jakość regionalnych systemów innowacji przyczynia się do poprawy jakości dopasowania modelu do danych empirycznych, udział poprawnych predykcji został obliczony dla czterech wariantów modeli:

1) modelu oryginalnego (pełnego),
2) modelu nieuwzględniającego efektów losowych dla województw oraz zmiennych regionalnych pochodzących z RIS,
3) modelu nieuwzględniającego efektów losowych dla województw, ale uwzględniającego zmienne regionalne pochodzące z RIS,
4) modelu nieuwzględniającego efektów losowych dla województw, ale uwzględniającego zmienne regionalne pochodzące z RIS.

Tabela 38 prezentuje różnice między udziałem poprawnych predykcji dla modelu oryginalnego oraz modelu, w którym usunięto określone grupy zmiennych.

Wyniki omawianego badania wskazują, że zarówno usunięcie efektów regionalnych, jak i nieuwzględnienie efektów losowych prowadzi do spadku dopasowania modelu do danych empirycznych. Rezultat ten wskazuje na zasadność wykorzystania modelu wielopoziomowego oraz uwzględniania efektów losowych i zmiennych wskazujących na jakość regionalnych systemów innowacji w modelach wyjaśniających zachowania innowacyjne przedsiębiorstw. Okazuje się jednak, że jeśli zignorowany zostanie fakt występowania losowych różnic między regionalnymi skłonnościami do wykorzystania zaawansowanych technologii informatycznych i komunikacyjnych oraz inwestowania w nie, spadek jakości dopasowania jest silniejszy. Podobna sytuacja dotyczy równań innowacyjności. Spadek jakości dopasowania w równaniach wyjaśniających skłonność do wprowadzania innowacji okazał się wyższy niż w pozostałych sześciu równaniach. Wynik ten nie budzi wątpliwości, ponieważ parametry równań innowacyjności szacowane są po estymacji parametrów wcześniejszych sześciu równań. Dlatego też spadek dopasowania w równaniach decyzji inwestycyjnych jest związany z nieuwzględnieniem efektów losowych/zmiennych regionalnych w dwóch krokach.

5. Wielopoziomowy polichotomiczny model logitowy Wielopoziomowy model regresji rankingowej

5.1. Wprowadzenie

Prezentowane w rozdziale czwartym metody estymacji parametrów uogólnionych liniowych modeli wielopoziomowych można wykorzystać do analizy hierarchicznie uporządkowanych danych, gdzie regresant ma charakter zmiennej dwumianowej, wielomianowej kategorii uporządkowanych czy licznikowej. W badaniach mikroekonomicznych lub socjologicznych często pojawia się jednak problem, w którym zmienna zależna przyjmuje kilka wartości niedających się uporządkować. Jeśli na przykład próbuje się znaleźć związek między cechami społeczno-demograficznymi respondentów a ich sympatiami politycznymi, zmienna zależna na ogół przyjmuje więcej niż dwie wartości.

Na poglądy polityczne mogą mieć wpływ zarówno cechy indywidualne (np. wiek, płeć, poziom wykształcenia), jak i czynniki historyczne i kulturowe. Jak pokazują wyniki niektórych badań, fakt przynależności miejsca zamieszkania respondenta do określonej jednostki administracyjnej czy regionu historycznego może silniej wpływać na decyzje podejmowane przy urnach niż cechy indywidualne (Grabowski, 2018). Rozróżnienie między czynnikami kontekstowymi a niekontekstowymi jest ważnym zadaniem badacza z zakresu socjologii polityki. Zastosowanie wielopoziomowych modeli wielomianowych logitowych jest użytecznym sposobem rozwiązania analizowanego problemu. Po oszacowaniu parametrów oraz efektów losowych dla takiego modelu możliwa jest ewaluacja znaczenia czynników kontekstowych i niekontekstowych.

W niniejszym rozdziale omawiane są metody estymacji parametrów wielopoziomowych modeli wielomianowych logitowych, a także wielopoziomowych modeli regresji rankingowej. Metody te są następnie wykorzystywane do analizy czynników wpływających na sposób reakcji wobec wystąpienia problemu prawnego.

5.2. Wielopoziomowy model wielomianowy logitowy Wielopoziomowy model regresji rankingowej

Rozważmy dwupoziomowy wielomianowy model logitowy. Uogólnienie na przypadek trzech i więcej poziomów jest natychmiastowe. Niech $j = 1, \ldots, J$ numeruje klastry na drugim poziomie. Na pierwszym poziomie mamy I_j obiektów w ramach każdego klastra indeksowanych $i = 1, \ldots, I_j$. Przez $\tilde{L}_{ij} = \left\{ l_{ij}^1, \ldots, l_{ij}^{L_{ij}} \right\}$ oznaczamy zbiór L_{ij} możliwych wyborów dla i-tej jednostki należącej do j-tego klastra.

Parametryczny model nieobserwowanych losowych użyteczności przyjmuje następującą postać (por. Skrondal, Rabe-Hesketh, 2003):

$$uz_{ij}^l = fz_{ij}^l + \delta z_{ij}^l + \varepsilon z_{ij}^l , \tag{5.1}$$

gdzie fz_{ij}^l jest deterministycznym składnikiem losowym reprezentującym obserwowaną heterogeniczność wariantów do wyboru, jednostek oraz klastrów, a δz_{ij}^l jest zmienną sztuczną reprezentującą nieobserwowalną heterogeniczność i zależną od wariantów, dekomponowaną w następujący sposób:

$$\delta z_{ij}^l = \delta z_{ij}^{l(1)} + \delta z_{ij}^{l(2)} , \tag{5.2}$$

gdzie $\delta z_{ij}^{l(1)}$ oraz $\delta z_{ij}^{l(2)}$ są zmiennymi sztucznymi związanymi z heterogenicznością odpowiednio na pierwszym i drugim poziomie. εz_{ij}^l jest specyficznym dla l-tego wariantu składnikiem losowym reprezentującym nieobserwowalną heterogeniczność, niezależną ze względu na wybory, jednostki, klastry.

Część ustalona ze wzoru (5.1) dekomponowana jest następująco:

$$fz_{ij}^l = \breve{m}^l + \boldsymbol{x}_{ij}^l \breve{\mathbf{b}} + \boldsymbol{x}_{ij} \breve{\mathbf{g}}^a , \tag{5.3}$$

gdzie $\breve{m}^l$ jest stałą, $\boldsymbol{x}_{ij}^l$ jest wektorem zmiennych objaśniających, różniących się ze względu na wybory i mogących się różnić ze względu na jednostki oraz klastry. Zmienne objaśniające wchodzące w skład wektora x_{ij} różnią się ze względu na jednostki lub klastry, jednak nie różnią się ze względu na wybory. $\breve{\mathbf{b}}$ oraz $\breve{\mathbf{g}}^a$ są odpowiednimi wektorami parametrów.

Komponent $\delta z_{ij}^{l(1)}$, związany z zależnościami między użytecznościami wśród jednostek, dekomponowany jest następująco:

$$\delta z_{ij}^{l(1)} = \breve{\mathbf{z}}_{ij}^{l(1)} \breve{\boldsymbol{\beta}}_{ij}^{(1)} + \breve{\boldsymbol{\lambda}}^{l(1)} \breve{\boldsymbol{\eta}}_{ij}^{(1)}, \tag{5.4}$$

gdzie $\breve{\boldsymbol{\beta}}_{ij}^{(1)}$ jest wektorem parametrów związanych z efektami losowymi przy zmiennych wchodzących w skład wektora $\breve{\mathbf{z}}_{ij}^{l(1)}$, natomiast $\breve{\boldsymbol{\eta}}_{ij}^{(1)}$ zawiera czynniki nieobserwowalne na poziomie jednostek. Wpływ tych czynników na użyteczność mierzy parametr $\breve{\boldsymbol{\lambda}}^{l(1)}$.

Wartości oczekiwane efektów losowych definiowane są następująco:

$$\begin{aligned} E\left[\breve{\boldsymbol{\beta}}_{ij}^{(1)}\right] &= \boldsymbol{x}_{ij} \breve{\mathbf{G}}_{\beta}^{(1)}, \\ E\left[\breve{\boldsymbol{\eta}}_{ij}^{(1)} \mid \boldsymbol{x}_{ij}\right] &= \boldsymbol{x}_{ij} \breve{\mathbf{G}}_{\eta}^{(1)}, \end{aligned} \tag{5.5}$$

gdzie $\breve{\mathbf{G}}_{\beta}^{(1)}$ oraz $\breve{\mathbf{G}}_{\eta}^{(1)}$ są macierzami parametrów przy tych zmiennych.

Komponent $\delta z_{ij}^{l(2)}$ związany z zależnościami między użytecznościami wewnątrz klastrów dekomponowany jest następująco:

$$\delta z_{ij}^{l(2)} = \breve{\mathbf{z}}_{ij}^{l(2)} \breve{\boldsymbol{\beta}}_{j}^{(2)} + \breve{\boldsymbol{\lambda}}^{l(2)} \breve{\boldsymbol{\eta}}_{j}^{(2)} + \breve{\mathbf{z}}_{ij}^{(2)} \breve{\boldsymbol{\gamma}}_{j}^{l(2)}, \tag{5.6}$$

gdzie wektor parametrów $\breve{\boldsymbol{\beta}}_{j}^{(2)}$ ilustruje różniący się w poszczególnych klastrach wpływ zmiennych objaśniających związanych z klastrami na użyteczność, natomiast $\breve{\boldsymbol{\gamma}}_{j}^{l(2)}$ reprezentuje zróżnicowanie wpływu regresorów różniących się ze względu na jednostki $\breve{\mathbf{z}}_{ij}^{(2)}$ na użyteczność. $\breve{\boldsymbol{\eta}}_{j}^{(2)}$ są efektami losowymi na poziomie klastrów, a ich wpływ na użyteczność mierzą parametry wchodzące w skład wektora $\breve{\boldsymbol{\lambda}}^{l(2)}$.

Wartości oczekiwane odpowiednich efektów losowych zapisuje się następująco:

$$\begin{aligned} E\left[\breve{\boldsymbol{\beta}}_{j}^{(2)}\right] &= \boldsymbol{x}_{j} \breve{\mathbf{G}}_{\beta}^{(2)}, \\ E\left[\breve{\boldsymbol{\eta}}_{j}^{(2)}\right] &= \boldsymbol{x}_{j} \breve{\mathbf{G}}_{\eta}^{(2)}, \\ E\left[\breve{\boldsymbol{\gamma}}_{j}^{l(2)}\right] &= \boldsymbol{x}_{j} \breve{\mathbf{G}}_{\gamma}^{l(2)}, \end{aligned} \tag{5.7}$$

gdzie $\boldsymbol{x}_j$ są zmiennymi objaśniającymi różniącymi się na poziomie klastrów. Macierze $\boldsymbol{G}_{\beta}^{(2)}$, $\boldsymbol{G}_{\eta}^{(2)}$ oraz $\boldsymbol{G}_{\gamma}^{l(2)}$ składają się z parametrów przy tych zmiennych.

Model (5.1) można w macierzowy sposób zapisać następująco:

$$\boldsymbol{uz}_{ij} = \boldsymbol{fz}_{ij} + \boldsymbol{\delta z}_{ij}^{(1)} + \boldsymbol{\delta z}_{ij}^{(2)} + \boldsymbol{\varepsilon z}_{ij}, \tag{5.8}$$

gdzie $\boldsymbol{uz}_{ij}$ jest wektorem kolumnowym o wymiarach $L_{ij} \times 1$, składającym się z użyteczności i-tej jednostki z j-tego klastra, związanych ze wszystkimi wyborami. Wektor efektów stałych $\boldsymbol{fz}_{ij}$ dekomponowany jest następująco:

$$\boldsymbol{f}_{ij} = \breve{\mathbf{m}} + \boldsymbol{X}_{ij}^{(1)}\breve{\mathbf{b}} + \left(\boldsymbol{I}_{L_{ij}} \otimes \boldsymbol{x}_{ij}\right)\breve{\mathbf{g}}, \tag{5.9}$$

gdzie wektor kolumnowy $\breve{\mathbf{m}}$ składa się ze stałych m^l dla poszczególnych wyborów, macierz $\boldsymbol{X}_{ij}^{(1)}$ składa się z wierszy odpowiadających $\boldsymbol{x}_{ij}^{l}$, natomiast wektor kolumnowy $\breve{\mathbf{g}}$ składa się z wektorów kolumnowych $\breve{\mathbf{g}}^a$ dla poszczególnych wyborów.

Wektor komponentów związany z zależnościami między użytecznościami wśród jednostek $\boldsymbol{\delta z}_{ij}^{(1)}$ dekomponuje się następująco:

$$\boldsymbol{\delta z}_{ij}^{(1)} = \breve{\mathbf{Z}}_{ij}^{(1)}\breve{\boldsymbol{\beta}}_{ij}^{(1)} + \breve{\boldsymbol{\Lambda}}_{[ij]}\breve{\boldsymbol{\eta}}_{ij}^{(1)}, \tag{5.10}$$

gdzie wierszami macierzy $\breve{\mathbf{Z}}_{ij}^{(1)}$ są wektory $\breve{\mathbf{z}}_{ij}^{l(1)}$ ze wzoru (5.4) związane z poszczególnymi wyborami, natomiast macierz $\breve{\boldsymbol{\Lambda}}_{[ij]}$ składa się z wierszy $\breve{\boldsymbol{\lambda}}^{l(1)}$.

Wektor komponentów $\boldsymbol{\delta z}_{ij}^{(2)}$ związany z zależnościami między użytecznościami wewnątrz klastrów dekomponowany jest następująco:

$$\boldsymbol{\delta z}_{ij}^{(2)} = \breve{\mathbf{Z}}_{ij}^{(2)}\breve{\boldsymbol{\beta}}_{ij}^{(2)} + \boldsymbol{\Lambda}_{[ij]}^{(2)}\boldsymbol{\eta}_{j}^{(2)} + \left(\boldsymbol{I}_{A_{ij}} \otimes \boldsymbol{z}_{ij}^{(2)}\right)\boldsymbol{\gamma}_{j}^{(2)}, \tag{5.11}$$

gdzie macierz $\breve{\mathbf{Z}}_{ij}^{(2)}$ składa się z wektorów $\breve{\mathbf{z}}_{ij}^{l(2)}$ ze wzoru (5.6), macierz $\boldsymbol{\Lambda}_{[ij]}^{(2)}$ zawiera wektory $\breve{\boldsymbol{\lambda}}^{l(2)}$, natomiast wektor kolumnowy $\breve{\boldsymbol{\gamma}}_{j}^{(2)}$ obejmuje wektory $\breve{\boldsymbol{\gamma}}_{j}^{l(2)}$. Jeśli chodzi o wartość oczekiwaną wektora $\breve{\boldsymbol{\gamma}}_{j}^{(2)}$, zakłada się, że:

$$E\left[\breve{\boldsymbol{\gamma}}_{j}^{(2)} \mid \boldsymbol{x}_{j}\right] = \boldsymbol{x}_{j}\breve{\mathbf{G}}_{\gamma}^{(2)}, \tag{5.12}$$

gdzie macierz $\breve{\mathbf{G}}_{\gamma}^{(2)}$ powstaje przez ułożenie pod sobą macierzy $\breve{\mathbf{G}}_{\gamma}^{l(2)}$ dla różnych wariantów.

Dla składników losowych εz_{ij}^{l} przyjmuje się założenie, że pochodzą one z rozkładu Gumbela, tak jak to jest w przypadku modelu omówionego w podrozdziale 1.5. Wówczas rozkład efektów losowych na poziomie jednostek jest następujący:

$$\boldsymbol{EL}_{ij}^{(1)} = \begin{bmatrix} \breve{\boldsymbol{\beta}}_{ij}^{(1)} \\ \breve{\boldsymbol{\eta}}_{ij}^{(1)} \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \boldsymbol{x}_{ij}\breve{\boldsymbol{G}}_{\beta}^{(1)} \\ \boldsymbol{x}_{ij}\breve{\boldsymbol{G}}_{\eta}^{(1)} \end{bmatrix}, \begin{bmatrix} \breve{\boldsymbol{\Psi}}_{\beta}^{(1)} & 0 \\ 0 & \breve{\boldsymbol{\Psi}}_{\eta}^{(1)} \end{bmatrix} \right). \tag{5.13}$$

Rozkład efektów losowych na poziomie klastrów jest następujący:

$$\boldsymbol{EL}_{j}^{(2)} = \begin{bmatrix} \breve{\boldsymbol{\beta}}_{j}^{(2)} \\ \breve{\boldsymbol{\eta}}_{j}^{(2)} \\ \breve{\boldsymbol{\gamma}}_{j}^{(2)} \end{bmatrix} \sim \boldsymbol{N}\left(\begin{bmatrix} \boldsymbol{x}_{j}\breve{\boldsymbol{G}}_{\beta}^{(2)} \\ \boldsymbol{x}_{j}\breve{\boldsymbol{G}}_{\eta}^{(2)} \\ \boldsymbol{x}_{j}\breve{\boldsymbol{G}}_{\gamma}^{(2)} \end{bmatrix}, \begin{bmatrix} \breve{\boldsymbol{\Psi}}_{\beta}^{(2)} & 0 & 0 \\ 0 & \breve{\boldsymbol{\Psi}}_{\eta}^{(2)} & 0 \\ 0 & 0 & \breve{\boldsymbol{\Psi}}_{\gamma}^{(2)} \end{bmatrix} \right). \tag{5.14}$$

W analizowanym modelu część losowa zawiera elementy rozkładu normalnego oraz rozkładu Gumbela. Dlatego też prezentowany model może być nazywany Gumbel-normalnym.

W celu wyprowadzenia postaci funkcji wiarygodności dla każdej jednostki *ij* definiowana jest macierz porównań $\boldsymbol{H}_{ij}^{F}$ o wymiarach $\left(L_{ij}-1\right)\times L_{ij}$ zawierająca elementy –1, 0 oraz 1 w taki sposób, aby każdy wybór był wyrażony przez dodatnią różnicę użyteczności. Załóżmy, że *i*-ta jednostka należąca do *j*-tego klastra dokonała wyboru pierwszego wariantu spośród trzech. Wówczas prawdziwe są następujące nierówności:

$$\begin{aligned} uz_{ij}^{1} > uz_{ij}^{2}, \\ uz_{ij}^{1} > uz_{ij}^{3}, \end{aligned} \tag{5.15}$$

natomiast postać macierzy porównań jest następująca:

$$\boldsymbol{H}_{ij}^{F} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}, \tag{5.16}$$

gdyż w przypadku takiej macierzy spełniona jest nierówność $\boldsymbol{H}_{ij}^{F}\boldsymbol{uz}_{ij} > 0$. Prawdopodobieństwo wyboru danego wariantu przez *i*-tą jednostkę z *j*-tego klastra można zatem wyrazić w następujący sposób:

$$P\left(l_{ij}^{l} \mid \tilde{L}_{ij}\right) = P\left(\boldsymbol{H}_{ij}^{F}\boldsymbol{uz}_{ij} > 0\right). \tag{5.17}$$

Na podstawie równania (5.17) widzimy, że w celu wyznaczenia prawdopodobieństwa, że *i*-ta jednostka należąca do *j*-tego klastra dokona *l*-tego wyboru, należy obliczyć wartość A_{ij} – 1-wymiarowej całki dla wielowymiarowej zmiennej losowej o rozkładzie logistyczno-normalnym. Skrondal oraz Rabe-Hesketh (2003) proponują zastosowanie alternatywnego podejścia, polegającego na wykorzystaniu faktu, że Gumbela-normalny model użyteczności redukuje się do niezależnego modelu użyteczności Gumbela, przy danych wartościach zmiennych nieobserwowalnych. Wówczas prawdopodobieństwo warunkowe, że *i*-ta jednostka należąca do *j*-tego klastra dokona *l*-tego wyboru, można zapisać następująco:

$$P\left(a_{ij}^{l} \mid \tilde{L}_{ij}, \boldsymbol{EL}_{ij}, \boldsymbol{EG}_{ij}\right) = \frac{\exp\left(fz_{ij}^{l} + \delta z_{ij}^{l}\right)}{\sum_{s=1}^{L_{ij}} \exp\left(fz_{ij}^{s} + \delta z_{ij}^{s}\right)}, \tag{5.18}$$

gdzie $\boldsymbol{EL}_{ij}$ oraz $\boldsymbol{EG}_{ij}$ oznaczają odpowiednio zbiór zmiennych nieobserwowalnych oraz egzogenicznych. W celu uzyskania wkładu do funkcji wiarygodności związanego z *j*-tym klastrem należy dokonać całkowania niezależnych wielowymiarowych zmiennych nieobserwowalnych $\boldsymbol{EL}_{ij}^{(1)}$ oraz $\boldsymbol{EL}_{j}^{(2)}$. Wówczas odpowiednie prawdopodobieństwo wynosi:

$$\begin{aligned} &P\left(l_{1j}^{l} \cap \ldots \cap l_{I_j j}^{l} \mid \boldsymbol{EG}_{ij}\right) = \\ &= \int \prod_{i=1}^{I_j} \left\{ \int P\left(l_{ij}^{l} \mid \tilde{L}_{ij}, \boldsymbol{EL}_{ij}, \boldsymbol{EG}_{ij}\right) \varphi\left(\boldsymbol{EL}_{ij}^{(1)}\right) d\boldsymbol{EL}_{ij}^{(1)} \right\} \varphi\left(\boldsymbol{EL}_{j}^{(2)}\right) d\boldsymbol{EL}_{j}^{(2)} \end{aligned} \tag{5.19}$$

Funkcja wiarygodności dla wielopoziomowego, wielomianowego modelu logitowego przyjmuje następującą postać:

$$\ln L = \sum_{l=1}^{L} \sum_{j=1}^{J} \ln\left(P\left(l_{1j}^{l} \cap \ldots \cap l_{I_j j}^{l} \mid \boldsymbol{EG}_{ij}\right)\right). \tag{5.20}$$

W przypadku wielopoziomowego modelu regresji rangowanej wybór *l*-tego wariantu przez *i*-ty obiekt należący do *j*-tego klastra zastępowany jest przez uszeregowanie wariantów według *r*-tego rankingu $R_{ij}^{(r)}$. Zamiast prawdopodobieństwa (5.19) rozważane jest prawdopodobieństwo rankingu dane wzorem:

$$P\left(R_{1j}^{(r)} \cap \ldots \cap R_{N_j j}^{(r)} \mid \boldsymbol{E}_{ij}\right) = \tag{5.21}$$
$$= \int \prod_{i=1}^{I_j} \left\{ \int P\left(R_{1j}^{(r)} \mid \tilde{A}_{ij}, \boldsymbol{EL}_{ij}, \boldsymbol{EG}_{ij}\right) \varphi\left(\boldsymbol{L}_{ij}^{(1)}\right) d\boldsymbol{L}_{ij}^{(1)} \right\} \varphi\left(\boldsymbol{L}_{ij}^{(1)}\right) d\boldsymbol{L}_{j}^{(2)}$$

Podobnie jak w przypadku wielopoziomowego, wielomianowego modelu logitowego funkcja wiarygodności składa się z sum funkcji wiarygodności dla poszczególnych klastrów oraz rankingów:

$$\ln L = \sum_{r=1}^{R} \sum_{j=1}^{J} \ln\left(P\left(R_{1j}^{(r)} \cap \ldots \cap R_{N_j j}^{(r)} \mid \boldsymbol{EG}_{ij}\right)\right). \tag{5.22}$$

5.3. Wykorzystanie wielopoziomowego polichotomicznego nieuporządkowanego modelu logitowego do analizy czynników wpływających na sposób reakcji wobec zaistnienia problemu prawnego

5.3.1. Czynniki wpływające na sposób reakcji wobec wystąpienia problemu prawnego – przegląd literatury

Osoby doświadczające problemu prawnego mogą zareagować na niego na trzy sposoby (por. Kritzer, 2008; Florczak, 2016; Florczak, Grabowski, 2018a). Po pierwsze, mogą zaniechać wszelkich działań lub próbować rozwiązać problem w drodze nieformalnych starań własnych czy też zasięgać informacji/porady w sposób niezinstytucjonalizowany. Po drugie, mogą udać się do odpowiedniej instytucji, gdzie taką usługę uzyskuje się nieodpłatnie. Po trzecie, mogą starać się o uzyskanie komercyjnej porady prawnej w kancelarii radcowskiej lub adwokackiej.

Każda z analizowanych reakcji wymaga innego wysiłku ze strony osoby doświadczającej problemu. W przypadku braku reakcji mamy do czynienia z „zerowym" wysiłkiem. Udanie się do instytucji, od której można uzyskać poradę nieodpłatnie, wiąże się z poniesieniem kosztu alternatywnego (np. kosztu przemieszczania, utraconych dochodów ze względu na poświęcenie czasu na poszukiwanie stosownego rozwiązania). W przypadku udania się do kancelarii adwokackiej lub komorniczej mamy do czynienia z koniecznością uiszczenia opłaty za poradę komercyjną. W praktyce obserwowane są wszystkie wymienione postawy. Z jednej strony świadczy to o zróżnicowanej percepcji korzyści wynikających z uzyskania pomocy (lub zaniechania rozwiązania problemu), z drugiej zaś

o zróżnicowaniu owych korzyści. Gdyby faktyczne zróżnicowanie korzyści nie miało miejsca, wówczas nie obserwowano by również zróżnicowania ludzkich reakcji na wystąpienie problemu prawnego (Florczak, Grabowski, 2018b).

Jeśli jednak postępowanie osób dotkniętych problemem prawnym jest racjonalne, pojawia się pytanie o czynniki determinujące ich zachowania. W niektórych przypadkach koszt zaniechania może być „wyższy" niż koszt związany z udaniem się do kancelarii adwokackiej lub notarialnej. W innych przypadkach ewentualne straty wynikające z braku reakcji są niższe od kosztów uzyskania porady ze strony adwokata czy notariusza. Zadaniem socjologii prawa jest identyfikacja czynników wpływających na zróżnicowanie reakcji wobec wystąpienia problemu prawnego. Przegląd literatury teoretycznej i empirycznej pozwala wyróżnić następujące grupy czynników determinujących wybór metody postępowania w obliczu wystąpienia problemu prawnego (por. m.in. Murayama, 2007; Kritzer, 2008; Winczorek, 2015):

1) przedmiot problemu prawnego i jego waga,
2) cechy społeczno-ekonomiczno-demograficzne osób dotkniętych problemem,
3) indywidualne postawy wobec przestrzegania i stosowania prawa,
4) aktywność społeczna respondentów,
5) środowisko społeczno-ekonomiczne respondenta i bariery dostępu do poradnictwa prawnego.

Dotychczas przeprowadzone badania empiryczne dotyczące omawianego zagadnienia wykorzystywały w większości nieskomplikowane metody statystyki opisowej (np. Murayama, 2007; Preisert, Schimanek, Waszak, Winiarska, 2013) lub też proste narzędzia statystyki matematycznej, ograniczone do analiz dla przestrzeni dwuwymiarowych (Kritzer, 2008). Taka metodyka nie jest odpowiednia, co wynika z faktu, że sposób reakcji wobec wystąpienia problemu prawnego zależy od szeregu czynników. Nieuwzględnienie ich w modelu wyjaśniającym znaczenie tylko jednego czynnika może prowadzić do uzyskania błędnych wniosków. Pojawia się ryzyko powstania obciążenia związanego z pominięciem ważnych zmiennych. Dlatego też estymacja parametrów modelu ekonometrycznego uwzględniającego jednocześnie wpływ wielu czynników jest właściwym rozwiązaniem.

W pracy Waldemara Florczaka i Wojciecha Grabowskiego (2018a) podjęto – zgodnie ze stanem wiedzy autorów – pierwszą próbę zastosowania wielomianowego modelu logitowego do kwantyfikacji wpływu wszystkich czynników wskazanych w literaturze przedmiotu na wybór postępowania wobec doświadczanego problemu prawnego. Analizę empiryczną przeprowadzono przy użyciu reprezentatywnych danych mikroekonomicznych uzyskanych w 2012 roku przez Instytut Spraw Publicznych w Warszawie na podstawie ogólnopolskiego badania ankietowego usługodawców i beneficjentów poradnictwa prawnego. Wyniki estymacji parametrów wielomianowego modelu logitowego wskazały, że najważniejszymi

determinantami decyzji dotyczącej sposobu rozwiązania problemu były: jego rodzaj, ranga, postawa wobec prawa oraz niektóre cechy socjoekonomiczne. Wyniki estymacji pokazują, że najważniejszym ograniczeniem popytu na usługi poradnictwa prawnego nie jest zbyt niski dochód, lecz ich słaba przestrzenna dostępność.

Mimo dostępności informacji o powiecie zamieszkiwanym przez respondenta w badaniu empirycznym nie zostały wykorzystane żadne dane regionalne. Wydaje się jednak, że sposób postępowania osoby dotkniętej problemem prawnym może zależeć od podaży usług prawnych w miejscowości, w której zamieszkuje. Oprócz tego sposoby reakcji wobec wystąpienia problemu prawnego mogą wynikać z utartych przekonań czy też czynników kulturowych. Na przykład w jednej społeczności mogą występować przekazywane z pokolenia na pokolenie wzorce braku zaufania do prawników, w innej zaś poziom zaufania może być zdecydowanie większy. Dlatego też decyzja jednostki może w dużym stopniu zależeć od kontekstu kulturowego. To powoduje, że model wyjaśniający sposób reakcji wobec wystąpienia problemu prawnego powinien uwzględniać znaczenie zmiennych związanych z lokalizacją i podażą usług prawnych w miejscu zamieszkania osoby ankietowanej. Takie badanie przeprowadzone jest w niniejszym rozdziale.

5.3.2. Estymacja parametrów wielopoziomowego, nieuporządkowanego, polichotomicznego modelu logitowego na podstawie danych pochodzących z badania dla Polski przeprowadzonego przez Instytut Spraw Publicznych w Warszawie

Punktem wyjścia do przeprowadzenia weryfikacji empirycznej są dane mikroekonomiczne uzyskane na podstawie ogólnopolskiego badania ankietowego beneficjentów i usługodawców poradnictwa prawnego przeprowadzonego w 2012 roku przez Instytut Spraw Publicznych w Warszawie. Jeśli chodzi o badanie beneficjentów, zrealizowanych zostało 1050 wywiadów bezpośrednich (PAPI) na ogólnopolskiej próbie. W badaniu zastosowana została próba losowo-warstwowa, co pozwoliło uzyskać jej zgodność ze strukturą populacji ze względu na wiek, płeć, wielkość miejscowości zamieszkania oraz województwo. Dobór respondentów na terenie był realizowany zgodnie z formułą *random route*. Szczegóły dotyczące analizowanego badania można znaleźć między innymi w pracy Stanisława Burdzieja i Marka Dudkiewicza (2013). Analizowane dane mikroekonomiczne zostały wykorzystane podczas realizacji grantu Narodowego Centrum Nauki pt. „Nieodpłatna pomoc prawna w Polsce z punktu widzenia ekonomicznej analizy prawa. Stan obecny i rekomendowany" o numerze 2012/07/B/HS4/02994. Wyniki analiz można znaleźć na przykład w pracach Florczaka i Grabowskiego (2017; 2018a; 2018b; 2018c).

W tabeli 39 prezentowane jest zestawienie czynników potencjalnie wpływających na wybór postępowania w obliczu zaistnienia problem prawnego. Uwzględniane są te kategorie, których uzyskanie jest możliwe dzięki danym z badania przeprowadzonego przez Instytut Spraw Publicznych.

Tabela 39. Potencjalne determinanty sposobu reakcji wobec wystąpienia problemu prawnego

Symbol i nazwa zmiennej	Oczekiwany znak oszacowania parametru przy zmiennej Uzasadnienie teoretyczne Odwołanie do literatury
Przedmiot prawa i waga problemu prawnego	
LBUL – prawo budowalne *LCIV* – prawo cywilne *LROA* – prawo drogowe *LPEN* – prawo karne *LCON* – prawo konsumenckie *LDAM* – odszkodowania *LFIN* – problemy finansowe *LFAM* – prawo rodzinne *LJOB* – prawo pracy/rozwiązanie umowy o pracę *LLEG* – prawo spadkowe *LPRO* – prawo rzeczowe	Jak pokazują badania odnoszące się do innych niż Polska krajów, przynależność problemu do określonego przedmiotu prawa jest jedną z kluczowych determinant sposobu reakcji (por. Murayama, 2007; Kritzer, 2008; Pleasance, Balmer, Reimers, 2011; Winczorek, 2015). Na gruncie rozważań teoretycznych trudno jest przesądzić, które problemy są mniej, a które bardziej ważne dla osób ich doświadczających.
LVAL – zmienna binarna przyjmująca wartość 1 w przypadku, gdy respondent ocenił problem prawny jako ważny	Oczekuje się, że im większe jest znaczenie problemu prawnego dla osoby go doświadczającej, tym wyższa jest skłonność do poszukiwania profesjonalnej pomocy prawnej (Kritzer, 2008; Pleasance, Balmer, Reimers, 2011; Winczorek, 2015).
Cechy społeczno-demograficzne	
FEM – zmienna binarna przyjmująca wartość 1 dla kobiet oraz 0 dla mężczyzn	Jak wskazują wyniki badań przeprowadzonych m.in. przez Masayukiego Murayamę (2007), Rosemarty Hunter i Tracey de Simeone (2009) oraz Arkadiusza Preiserta i in. (2013), kobiety przyjmują na ogół bardziej aktywną postawę wobec zaistnienia problemów prawnych.
AGE – wiek w latach	Postawa wobec przestrzegania i stosowania prawa zmienia się wraz z wiekiem, za sprawą większego doświadczenia życiowego (por. Legal Service Corporation, 1994; Murayama, 2007; Preisert i in., 2013; Winczorek, 2015). Dlatego też starsze osoby powinny częściej przybierać aktywną postawę wobec zaistnienia problemu prawnego. Oczekiwany znak oszacowania jest zatem dodatni.

Symbol i nazwa zmiennej	Oczekiwany znak oszacowania parametru przy zmiennej Uzasadnienie teoretyczne Odwołanie do literatury
EDU1 – zmienna binarna, przyjmująca wartość 1 w przypadku respondenta z wykształceniem ponadpodstawowym *EDU2* – zmienna binarna przyjmująca wartość 1 dla respondenta z wykształceniem ponadśrednim	*Ceteris paribus* lepsze wykształcenie powinno skutkować obszerniejszą wiedzą ogólną i tym samym bardziej aktywną postawą wobec problemu prawnego (por. Legal Service Corporation, 1994; Murayama, 2007; Preisert i in., 2013).
SCZO – zmienna binarna, przyjmująca wartość 1, jeśli stanem cywilnym respondenta jest żonaty/mężatka *SCRO* – zmienna binarna przyjmująca wartość 1, jeśli stanem cywilnym respondenta jest rozwodnik/rozwódka *SCWD* – zmienna binarna przyjmująca wartość 1, jeśli stanem cywilnym respondenta jest wdowiec/wdowa	Stan cywilny respondenta ma duży wpływ na inne cechy osobnicze, mogące oddziaływać na percepcję rzeczywistości i warunkujące przyjmowanie określonych postaw w momencie wystąpienia problemu prawnego (por. Legal Service Corporation, 1994; Murayama, 2007; Preisert i in., 2013). Trudno stwierdzić, w jaki sposób fakt pozostawania w określonym stanie cywilnym wpływa na wybór sposobu reakcji wobec wystąpienia problemu prawnego.
NFAM – liczba osób w gospodarstwie domowym	Wraz ze wzrostem liczby osób w gospodarstwie domowym można oczekiwać, przy innych czynnikach niezmienionych, zmian postaw i zachowań wobec pojawiających się problemów (Preisert i in., 2013)
Cechy ekonomiczno-społeczne	
DOCH – dochód na osobę w gospodarstwie domowym	Wraz ze wzrostem dochodu znikają bariery dostępu do usług prawnych, zwłaszcza komercyjnych (por. Legal Service Corporation, 1994; Murayama, 2007; Preisert i in., 2013).
RESD – zmienna binarna przyjmująca wartość 0, jeśli respondent zamieszkuje wieś i miasto poniżej 20 tys. mieszkańców oraz 1 dla mieszkańców większych miast	Mieszkańcy miast mają łatwiejszy dostęp zarówno do niekomercyjnych, jak i komercyjnych usług prawnych. Oznacza to, że mniejsze bariery w dostępie do porad prawnych dotyczą osób zamieszkujących większe miasta. Dlatego też należy oczekiwać dodatniego oszacowania parametru dla równań związanych z przyjęciem aktywnej postawy wobec zaistnienia problemu prawnego.
SLAB – zmienna binarna związana z sytuacją zawodową. Przyjmuje ona wartość 0, jeśli respondent ma stałą pracę oraz 1 dla osób niezatrudnionych lub pracujących dorywczo	Brak pracy może prowadzić do negatywnych konsekwencji społecznych, psychologicznych oraz ekonomicznych, a to z kolei do zmniejszenia podmiotowości osób nieaktywnych zawodowo (Murayama, 2007; Preisert i in., 2013). Dlatego też oczekuje się, że prawdopodobieństwo przyjęcia aktywnej postawy wobec zaistnienia problemu prawnego jest niższe dla tych osób, dla których analizowana zmienna przyjmuje wartość 1.

Tabela 34 (cd.)

Symbol i nazwa zmiennej	Oczekiwany znak oszacowania parametru przy zmiennej Uzasadnienie teoretyczne Odwołanie do literatury
Świadomość prawna i postawa wobec przestrzegania i stosowania prawa	
PAWR – zmienna ilustrująca poziom świadomości prawnej u osoby ankietowanej. Przyjmuje ona wartości całkowite z przedziału 0–15	Wraz ze wzrostem świadomości prawnej należy oczekiwać adekwatnej oceny zaistniałej sytuacji i pełniejszego uświadomienia niekorzystnych następstw pozostawiania problemu nierozwiązanym (Pleasance i in., 2003; Pleasance, Balmer, Reimers, 2011; Winczorek, 2015).
PCPL – zmienna związana z postawą wobec przestrzegania prawa. Przyjmuje wartość 1 w przypadku respondentów twierdzących, że zawsze należy przestrzegać prawa	Postawa wobec przestrzegania prawa powinna mieć wpływ na sposób rozwiązywania problemów prawnych (Kurczewski, Fuszara, 2004). Oczekuje się, że osoby twierdzące, że zawsze należy przestrzegać prawa, chętniej poszukują niekomercyjnego lub komercyjnego rozwiązania problemu prawnego.
PUSE – zmienna związana z postawą wobec stosowania prawa. Zmienna przyjmuje wartości z przedziału 0–3. 0 oznacza postawę ugodową, natomiast 3 postawę skargliwą	Jak wskazują wyniki badań empirycznych przeprowadzonych m.in. przez Jacka Kurczewskiego i Małgorzatę Fuszarę (2004), postawa wobec stosowania prawa jest ważką i trwałą cechą osobniczą wpływającą na sposób rozwiązywania problemów prawnych.
PTRU – zmienna ilustrująca poziom zaufania do palestry. Przyjmuje wartości z przedziału <0; 2> – najniższa wartość oznacza brak zaufania, a najwyższa pełne zaufanie	Uprzedzenia czy stereotypy wobec palestry mogą powodować spadek gotowości do korzystania z komercyjnej pomocy prawnej (Pleasance i in., 2003; Kritzer, 2008).
PAVA – zmienna binarna ilustrująca subiektywną ocenę dostępności usług prawnych. Przyjmuje ona wartość 0, jeśli respondent uważa, że są one trudno dostępne oraz 1, gdy jest innego zdania	Subiektywne postrzeganie dostępności usług prawnych (bez względu na ich obiektywną podaż) wpływa na decyzję dotyczącą sposobu reakcji na zaistnienie problemu prawnego. Osoby uważające, że usługi prawne są łatwo dostępne, powinny mieć wyższą skłonność do przyjmowania aktywnych postaw.
Aktywność społeczna	
APAR – zmienna binarna przyjmująca wartość 1, jeśli respondent przynależy do organizacji społecznych	Aktywność społeczna świadczy o inicjatywności, operatywności, upodmiotowieniu i uspołecznieniu. Dlatego też osoby aktywne społecznie powinny częściej aktywnie reagować w obliczu wystąpienia problemu prawnego w porównaniu z osobami mniej społecznie aktywnymi.
AACT – zmienna binarna przyjmująca wartość 1 w przypadku osób, które pozytywnie odpowiedziały na pytanie dotyczące działalności społecznej	

Źródło: opracowanie własne.

Badając wpływ określonych czynników na sposób reakcji wobec wystąpienia problemu prawnego, należy uwzględnić fakt, że wielu ankietowanych nie doświadczyło problemu prawnego. Dlatego też zmienna związana ze sposobem reakcji wobec wystąpienia problemu prawnego w przypadku zdecydowanej większości badanych nie była obserwowalna. W ankiecie przeprowadzonej przez Instytut Spraw Publicznych na początku respondentom było zadawane następujące pytanie: „Czy doświadczył/a Pan/Pani problemu prawnego w okresie ostatnich 5 lat?". Jeśli ankietowany odpowiedział twierdząco, zadawane było pytanie dotyczące sposobu reakcji. Florczak i Grabowski (2018a), szacując parametry wielomianowego modelu logitowego, ograniczyli się do próby osób doświadczających problemu prawnego. Stanowiła ona nieco ponad 20% próby pełnej. Takie postępowanie nie jest właściwe, ponieważ osoby doświadczające problemów prawnych charakteryzują się pewnymi cechami, które również mają wpływ na sposób reakcji wobec zaistnienia zdarzenia. Dlatego też, podobnie jak w klasycznym modelu Heckmana (1979), należy zastosować odpowiednią korektę podczas estymacji parametrów.

Wyodrębnienia czynników wpływających na fakt wystąpienia problemu prawnego można dokonać na podstawie analiz wyników estymacji parametrów modeli w badaniach eksploracyjnych przeprowadzonych między innymi przez Florczaka i Grabowskiego (2017, 2018c). W pierwszym kroku rozważana jest estymacja parametrów następującego dychotomicznego modelu wielopoziomowego:

$$PR_i^* = \boldsymbol{x}_i^{PR}\boldsymbol{\beta}^{PR} + \boldsymbol{z}_i^{PR}\boldsymbol{u}^{PR} + \varepsilon_i^{PR}, \tag{5.20a}$$

$$PR_i = I\left\{PR_i^* > 0\right\}, \tag{5.20b}$$

gdzie PR_i jest zmienną dychotomiczną, przyjmującą wartość 1 w przypadku respondentów doświadczających problemu prawnego oraz 0 dla pozostałych osób, $\boldsymbol{x}_i^{PR}$ jest wektorem zawierającym czynniki wpływające na fakt wystąpienia odpowiedniego zdarzenia. $\boldsymbol{z}_i^{PR}$ składa się ze zmiennych binarnych dla poszczególnych jednostek administracyjnych oraz ewentualnie zmiennych, których wpływ na zmienną zależną losowo różni się między województwami/powiatami, a $\boldsymbol{u}^{PR}$ jest wektorem efektów losowych ilustrujących różną skłonność mieszkańców poszczególnych jednostek administracyjnych do doświadczenia problemu prawnego. Po dokonaniu estymacji parametrów modelu (5.20a)–(5.20b) obliczana jest wartość odwróconego ilorazu Millsa w następujący sposób:

$$IMR_i = a^2\left(PR_i\right). \tag{5.21}$$

Następnie wielkość (5.21) wykorzystywana jest jako zmienna objaśniająca w wielopoziomowym, wielomianowym modelu logitowym wyjaśniającym sposób reakcji wobec zaistnienia problemu prawnego.

W celu estymacji parametrów równania (5.20a)–(5.20b) wykorzystywane są wyniki badania eksploracyjnego przeprowadzonego przez Florczaka i Grabowskiego (2018a), mającego na celu identyfikację czynników wpływających na prawdopodobieństwo wystąpienia problemu prawnego. W analizowanym badaniu wykorzystano klasyczny model logitowy oraz model logitowy z warunkowym uśrednianiem. Nie uwzględniano wówczas zmiennych związanych z lokalizacją respondenta. Wydaje się jednak, że mogą występować losowe różnice między prawdopodobieństwami doświadczenia problemu prawnego wśród osób charakteryzujących się tymi samymi cechami indywidualnymi a mieszkającymi w różnych powiatach/województwach. Wyższa/niższa skłonność do doświadczenia problemu prawnego w określonym powiecie czy województwie może wynikać z czynników kulturowych. Zakorzenione nawyki oraz wartości grupowe mogą wpływać na wybór sposobu rozwiązywania problemów. Niektóre metody rozwiązywania problemów prowadzą do spadku napięcia, inne zaś mogą przyczyniać się do powstania konfliktów, których rezultatem jest pojawienie się problemów prawnych. Dlatego też wydaje się, że rozszerzenie modelu zaproponowanego przez Florczaka i Grabowskiego (2018a) o efekty losowe związane z lokalizacją respondenta jest dobrym rozwiązaniem. Wyniki estymacji parametrów wielopoziomowego modelu logitowego wyjaśniającego skłonność do wystąpienia problemu prawnego prezentowane są w tabeli 40.

Tabela 40. Wyniki estymacji parametrów wielopoziomowego modelu logitowego wyjaśniającego skłonność do doświadczania problemu prawnego

Zmienna	Oszacowanie	Błąd standardowy	Graniczny poziom istotności
EDU1[a)] (wykształcenie ponadpodstawowe)	0,712	0,391	0,068
EDU2 (wykształcenie ponadśrednie)	1,140	0,440	0,009
SCRO[b)] (osoba rozwiedziona)	1,743	0,357	0,000
SCWD (wdowiec, wdowa)	1,682	0,390	0,001
PCPL (postawa wobec przestrzegania prawa)	0,630	0,334	0,060

Zmienna	Oszacowanie	Błąd standardowy	Graniczny poziom istotności
PUSE (postawa wobec stosowania prawa)	0,412	0,097	0,000
AACT (działalność społeczna)	0,501	0,252	0,047
AGE (wiek)	0,126	0,036	0,000
AGE2	–0,001	0,0004	0,000
Stała	–3,296	0,373	0,000
Wariancja efektów losowych dla województw	0,673	0,431	–
Wariancja efektów losowych dla powiatów	1,342	0,405	–
Testowanie obecności efektów losowych	Wartość statystyki LR = 104,32 Graniczny poziom istotności = 0,000		

a) Kategorią referencyjną są osoby z wykształceniem niepełnym podstawowym, podstawowym i gimnazjalnym; b) Kategorią referencyjną są respondenci stanu wolnego oraz żonaci mężczyźni i zamężne kobiety.

Źródło: obliczenia własne.

Uzyskane wyniki estymacji są w dużej części zgodne z rezultatami otrzymanymi przez Florczaka i Grabowskiego (2018a). Taka sytuacja nie powinna budzić wątpliwości, ponieważ w obu przypadkach wykorzystane zostały dane pochodzące z tego samego badania Instytutu Spraw Publicznych. Należy jednak zauważyć, że niektóre zmienne (np. związane z dochodem), które okazały się istotne w pracy Florczaka i Grabowskiego (2018a), były nieistotne w omawianym badaniu. Może to wynikać z faktu, że podobne czynniki kontekstowe wpływały na wysokość dochodów uzyskiwanych przez respondentów oraz ich skłonność do doświadczania problemu prawnego. Jak wiadomo, wysokość dochodów na osobę w gospodarstwach domowych jest silnie zróżnicowana regionalnie. W tej sytuacji uwzględnienie efektów losowych związanych z lokalizacją przyczynia się do redukcji roli dochodów.

Uzyskane oszacowania parametrów wskazują, że wraz ze wzrostem poziomu wykształcenia obserwowany jest wzrost prawdopodobieństwa doświadczenia problemu prawnego. Rezultat ten jest zgodny z oczekiwaniami, ponieważ lepiej wykształcone osoby charakteryzują się wyższą aktywnością w życiu gospodarczym. Dobrze wykształcone osoby są na ogół aktywne zawodowo. Częściej mogą doświadczać problemów związanych z brakiem wynagrodzenia za wykonaną pracę czy wadliwym wykonaniem usługi na rzecz ich gospodarstwa domowego. Dlatego

też osoby te mogą częściej poszukiwać porady u adwokata czy radcy prawnego (por. Winczorek, 2015). Analizując zależność między stanem cywilnym a prawdopodobieństwem doświadczenia problemu prawnego, należy zwrócić uwagę, że analizowana skłonność najczęściej dotyczy osób rozwiedzionych. Uzyskane oszacowanie jest zgodne z oczekiwaniami, ponieważ osoby rozwiedzione często doświadczają problemów prawnych związanych z prawem do opieki nad dziećmi, alimentami, podziałem majątku (por. Murayama, 2007). Z drugiej strony wysoka skłonność do doświadczania problemów prawnych dotyczy osób owdowiałych. Mogą one doświadczać problemów związanych z otrzymaniem spadku po zmarłej osobie lub koniecznością spłaty zadłużenia powstałego w okresie życia współmałżonka. Relatywnie niska skłonność do doświadczania problemów prawnych dotyczy żonatych mężczyzn, zamężnych kobiet, kawalerów, panien oraz osób żyjących w konkubinacie.

Postawa wobec przestrzegania prawa ma statystycznie istotny wpływ na prawdopodobieństwo doświadczenia problemu prawnego. Respondenci, którzy twierdzą, że prawa należy zawsze przestrzegać, istotnie rzadziej doświadczają problemów prawnych w porównaniu z osobami, które mają zdecydowanie mniej restrykcyjny stosunek do przestrzegania prawa. Wynik ten również nie budzi zastrzeżeń, gdyż wydaje się, że osoby charakteryzujące się bardzo restrykcyjnym podejściem do przestrzegania prawa wcielają w życie swój światopogląd. Dlatego też nie doświadczają one zazwyczaj problemów prawnych z własnej winy (por. Kritzer, 2008). Skargliwa postawa wobec stosowania prawa ma również wpływ na prawdopodobieństwo doświadczenia problemu prawnego. Osoby, które uważają, że zawsze należy stosować prawo, częściej domagają się rekompensaty ze strony kooperantów lub zwracają uwagę na nieprawidłowo wykonaną usługę. Skargliwi respondenci zdecydowanie częściej domagają się swoich praw i nie zawsze są w stanie je wyegzekwować. Dlatego też niższa skłonność do doświadczania problemu prawnego w grupie osób bardziej ugodowych może wynikać z faktu, że są one w stanie częściej pójść na kompromis i zaakceptować wynik transakcji, chociaż nie zawsze może on być po ich myśli (por. Hunter, de Simeone, 2009).

Fakt bycia aktywnym społecznie okazał się istotnym czynnikiem wpływającym na skłonność do doświadczenia problemu prawnego. Osoby aktywne społecznie działają w różnych organizacjach. Starają się reagować na nieprawidłowości obserwowane w środowisku lokalnym. Działalność organizacji społecznych często wymaga podejmowania działań mających na celu dochodzenie swoich praw. Realizacja określonych celów powoduje nierzadko wejście w konflikt z lokalnymi władzami czy określonymi grupami interesu. Dlatego też aktywność społeczna prowadzi *ceteris paribus* do wzrostu prawdopodobieństwa doświadczenia problemu prawnego.

Ostatni czynnik wpływający na prawdopodobieństwo doświadczenia problemu prawnego związany jest z wielkością zamieszkiwanej miejscowości. Osoby mieszkające w miastach liczących 50–100 tysięcy mieszkańców istotnie częściej doświadczają problemów prawnych w porównaniu z respondentami zamieszkującymi miejscowości z innych klas wielkości. Rezultat ten może wynikać z faktu, że miejscowości liczące 50–100 tysięcy mieszkańców są najczęściej tak zwanymi miastami archipelagu (Springer, 2016), które utraciły status stolic województw po 1998 roku. Jednocześnie wzrost udziału osób z wyższym wykształceniem oraz wzrost liczby studentów w populacji osób w wieku 19–25 lat sprawiły, że takie miasta jak Konin, Jelenia Góra czy Łomża straciły znaczenie odpowiednio względem Poznania, Wrocławia czy Białegostoku, a liczba mieszkańców w tych miejscowościach zaczęła drastycznie maleć. Niepewna sytuacja przedsiębiorstw zlokalizowanych w miastach wyludniających się oraz niekorzystna sytuacja na lokalnych rynkach pracy mogły prowadzić do powstawania zatorów płatniczych czy też konfliktów na linii pracodawca–pracownik. Zdarzenia te mogły spowodować konflikty prawne i problemy prawne w tego typu miastach zdecydowanie częściej niż na wsiach czy w innych miejscowościach. Jednocześnie patologie społeczne wynikające z relatywnie wysokiej stopy bezrobocia w miastach należących do analizowanej grupy sprawiły, że mogły (częściej niż w miejscowościach innego typu) pojawiać się problemy prawne związane z rozwodami czy prawem karnym.

Tabela 41. Oszacowania efektów losowych dla poszczególnych województw

Województwo	Średni efekt losowy	Województwo	Średni efekt losowy
dolnośląskie	–1,200	podkarpackie	0,177
kujawsko-pomorskie	–1,303	podlaskie	0,228
lubelskie	0,803	pomorskie	0,840
lubuskie	0,417	śląskie	0,629
łódzkie	0,043	świętokrzyskie	–0,383
małopolskie	–0,089	warmińsko-mazurskie	–0,149
mazowieckie	0,597	wielkopolskie	1,148
opolskie	–0,619	zachodniopomorskie	–1,137

Źródło: obliczenia własne.

Wyniki testu ilorazu wiarygodności wskazują, że uwzględnienie efektów losowych związanych z przynależnością gospodarstwa domowego zamieszkiwanego przez respondenta do określonego województwa czy powiatu istotnie poprawia model. Oznacza to zatem, że czynniki kulturowe, takie jak zakorzenione nawyki czy wartości grupowe, mają istotny wpływ na skłonność do zaistnienia problemu prawnego. Sposoby reakcji na określone sytuacje mogą różnić się w zależności od zamieszkiwanej miejscowości. Niektóre metody postępowania mogą częściej

prowadzić do pojawienia się konfliktu prawnego. Uzyskany wynik stanowi ważny wkład do rozważań dotyczących teorii determinizmu społecznego oraz dyskusji na temat roli przeszłości w postawach obserwowanych w polskim społeczeństwie. Okazuje się, że wzorce zachowań obserwowane w poszczególnych społecznościach lokalnych mają ważny wpływ na wybór sposobu postępowania w określonych sprawach. Mniej ugodowe wzorce zachowań mogą prowadzić do konfliktów prawnych, które często przyczyniają się do powstania problemów prawnych. Tabela 41 prezentuje oszacowania dla efektów losowych w poszczególnych województwach, natomiast tabela 42 zawiera analogiczne oszacowania dla poszczególnych powiatów.

Tabela 42. Oszacowania efektów losowych dla poszczególnych powiatów

Powiat	Efekt	Powiat	Efekt	Powiat	Efekt
Dzierżoniowski	–0,617	m. Tarnów	0,439	m. Częstochowa	1,041
Strzeliński	–0,454	Ciechanowski	0,029	m. Dąbrowa Górnicza	0,019
Wałbrzyski	–0,522	Makowski	0,095	m. Gliwice	0,325
m. Jelenia Góra	1,379	Płoński	–0,855	m. Jaworzno	–1,025
m. Legnica	–0,364	Radomski	–0,011	m. Katowice	–0,554
m. Wrocław	–0,619	Sokołowski	0,209	m. Mysłowice	–0,967
Średzki (woj. dolnośląskie)	–0,527	m. Płock	–0,678	m. Rybnik	–0,016
Inowrocławski	–0,292	m. Radom	0,075	m. Sosnowiec	0,277
Lipnowski	–0,418	m. Siedlce	–0,638	m. Tychy	–1,097
Toruński	–0,509	m. Warszawa	0,648	m. Żory	–1,004
m. Bydgoszcz	–0,619	Żyrardowski	1,985	Buski	–0,773
m. Toruń	0,145	Brzeski	–0,113	Opatowski	0,093
m. Włocławek	–0,180	Krapkowicki	–0,016	Ostrowiecki	–0,945
Hrubieszowski	0,594	Namysłowski	–0,964	Włoszczowski	0,805
Kraśnicki	–0,205	m. Opole	0,204	m. Kielce	0,269
Opolski (woj. lubelskie)	–0,630	Dębicki	–0,587	Działdowski	–0,938
m. Lublin	0,448	Leżajski	0,289	Iławski	0,474
m. Zamość	0,675	Rzeszowski	0,916	Olsztyński	–0,078
Łukowski	0,273	Stalowowolski	–0,719	Szczytnieński	1,289
Krośnieński (woj. lubuskie)	0,367	m. Rzeszów	0,355	m. Elbląg	–0,276
Nowosolski	–0,233	Moniecki	–0,119	m. Olsztyn	–0,685
m. Gorzów Wielkopolski	–0,802	Sokólski	–0,352	Gostyński	–1,449
m. Zielona Góra	0,535	m. Białystok	–0,324	Koniński	–0,110
Żarski	0,733	m. Łomża	1,122	Kolski	1,483
Bełchatowski	0,421	Chojnicki	0,109	Krotoszyński	0,497

Powiat	Efekt	Powiat	Efekt	Powiat	Efekt
Pabianicki	–0,418	Kościerzyński	1,011	Rawicki	0,174
Poddębicki	0,478	Malborski	0,453	m. Kalisz	–0,024
Sieradzki	–0,374	Tczewski	0,324	m. Konin	0,506
m. Skierniewice	0,358	m. Gdańsk	–0,690	m. Leszno	0,027
m. Łódź	0,970	m. Gdynia	–0,148	m. Poznań	–0,453
Łaski	–1,371	m. Słupsk	0,150	Białogardzki	–0,369
Krakowski	–0,469	Częstochowski	1,619	Stargardzki	0,963
Miechowski	–0,971	Gliwicki	2,640	Sławieński	–0,719
Oświęcimski	–0,533	Mikołowski	–0,807	m. Koszalin	–0,474
Tarnowski	0,447	Pszczyński	–0,117	m. Szczecin	–0,617
m. Kraków	0,960	m. Bielsko Biała	0,570	Świdwiński	–0,417

Źródło: opracowanie własne.

Wyniki zawarte w tabeli 41 wskazują, że prawdopodobieństwo doświadczenia problemu prawnego jest zdecydowanie niższe w grupie respondentów zamieszkujących województwo dolnośląskie, kujawsko-pomorskie oraz zachodniopomorskie. Oszacowania efektów losowych są w przypadku analizowanych regionów niższe niż –1. Najwyższe oszacowanie efektu losowego obserwowane jest dla województwa wielkopolskiego. Oznacza to zatem, że respondenci z tego regionu (*ceteris paribus*) najczęściej doświadczali problemów prawnych. Wynik ten może być wyjaśniony przez przeszłość historyczną analizowanego regionu. Duża część obecnego województwa wielkopolskiego należała w przeszłości do zaboru pruskiego. W kulturze pruskiej w XIX wieku wykształcone zostały wzorce formalnego rozwiązywania problemów, podpisywania umów przed wykonaniem usługi i częstego wchodzenia na drogę prawną w sytuacji niezadowolenia ze sposobu realizacji zadań. Dlatego też wyższy poziom skargliwości Wielkopolan może być uwarunkowany kulturowo (por. Podemski, Ziółkowski, 2007). Należy zauważyć, że wyższa skłonność do doświadczania problemów prawnych dotyczy również mieszkańców województwa pomorskiego. Obecne terytorium tego regionu również należało w dużej części do zaboru pruskiego w XIX wieku. Wynik dla województwa kujawsko-pomorskiego nie jest jednak zgodny z dotychczasowymi rozważaniami. Część tego regionu również należała do zaboru pruskiego w XIX wieku, a oszacowanie efektu losowego okazało się być ujemne i wysokie co do modułu.

Oszacowania efektów losowych dla poszczególnych powiatów wskazują na silne zróżnicowanie skłonności do doświadczania problemów prawnych między mniejszymi jednostkami administracyjnymi zlokalizowanymi w tym samym województwie. Należy zwrócić uwagę na sumę efektów losowych dla największych miast na prawach powiatu oraz województw, których są one stolicami. Po dodaniu odpowiednich oszacowań mamy: Warszawa: +1,245; Kraków: +0,871; Łódź: +1,013; Poznań: +0,695. Oznacza to zatem, że mieszkańcy czterech z pięciu największych

polskich miast charakteryzują się zdecydowanie większą skłonnością do doświadczania problemów prawnych w porównaniu z respondentami z mniejszych miejscowości. Intensywność zawierania transakcji i podpisywania umów w największych miastach, a także wyższa skłonność do formalizowania działań może przyczynić się do powstania konfliktów prawnych, co następnie prowadzi do pojawienia się problemów prawnych. Bardzo wysokie oszacowania efektów losowych obserwowane są w przypadku powiatu żyrardowskiego, ale także takich miast jak Konin, Częstochowa, Łódź, Zamość. Analizowane powiaty charakteryzują się także niekorzystnymi tendencjami migracyjnymi oraz wysokim odsetkiem rozwodów. Problemy mieszkańców tych powiatów mogą wynikać w dużym stopniu z faktu upadku wielu zlokalizowanych tam zakładów przemysłowych w okresie transformacji systemowej (powiat żyrardowski, Konin, Częstochowa, Łódź) lub też z degradacji miast w wyniku reformy administracyjnej (Konin, Częstochowa, Zamość). Wzorce kulturowe, które wykształciły się w okresie rozbiorów, a dotyczą powiatów położonych na terenie byłego zaboru rosyjskiego, również mogą mieć wpływ na wyższą skłonność do doświadczania problemów prawnych przez mieszkańców tamtych terenów. Bardzo wysokie średnie efekty losowe obserwowane są w przypadku większości powiatów z województwa lubelskiego (np. powiat hrubieszowski, miasto Lublin, miasto Zamość). Jak wskazuje między innymi Bartkowski (2003), konflikty między szlachtą a chłopami w XIX wieku obserwowano przede wszystkim na terenie folwarków, których było bardzo dużo na terenie obecnej Lubelszczyzny. Chociaż czasy zmieniły się zdecydowanie, skłonność do rozwiązywania konfliktów na drodze prawnej może mieć charakter zwyczaju przekazywanego z pokolenia na pokolenie. Bardzo wysoka skłonność do doświadczania problemów prawnych dotyczy również mieszkańców Łomży. Wyjaśnienia dla takiego stanu można poszukiwać zarówno we współczesności, jak i historii. Po pierwsze, Łomża utraciła status miasta wojewódzkiego w wyniku reformy administracyjnej z 1998 roku. Liczba mieszkańców w analizowanej miejscowości zmniejszała się relatywnie szybko, a stopa bezrobocia zdecydowanie przekraczała wartość ogólnopolską oraz dla województwa podlaskiego. Wiązało się to z wysokim poziomem przestępczości, który od wielu lat obserwowany jest w stolicy byłego województwa łomżyńskiego. Jednak oprócz współczesnych źródeł problemów z zakresu prawa rodzinnego czy karnego dotykających mieszkańców Łomży należy zwrócić uwagę na czynniki historyczno-kulturowe. Wydaje się, że poprzednie pokolenia obecnych mieszkańców Łomży w dużej części pochodzą z Grajewa, Wysokiego Mazowieckiego, Zambrowa czy Kolna. Są to tereny, na których drobna szlachta stanowiła dużą część ludności w XIX wieku. Analizując efekty losowe dla poszczególnych miast na prawach powiatu, należy zwrócić uwagę na ujemne i wysokie co do modułu średnie efektów krańcowych dla średnich i dużych miast położonych na obszarze byłych terenów niemieckich. Obserwowana

tendencja dotyczy takich miast jak Szczecin, Koszalin, Gorzów Wielkopolski, Legnica, Wrocław. Wynik ten również może znajdować uzasadnienie w powojennej historii tamtych terenów. Formalne instytucje służące rozwiązywaniu problemów prawnych były tworzone na analizowanym obszarze dopiero po II wojnie światowej. Dlatego też pierwsze pokolenia polskich mieszkańców Szczecina, Wrocławia i innych podobnych miast nie korzystały z wypracowanych wzorców formalizacji działań. W rezultacie poziom zaufania między ludźmi był wyższy i nie było konieczności podpisywania formalnych umów przed realizacją usługi. Odpowiednie wzorce kulturowe zostały przekazane kolejnym pokoleniom. W rezultacie respondenci pochodzący z byłych terenów niemieckich charakteryzują się niższą skłonnością do doświadczania problemów prawnych.

Tabela 43. Oszacowania parametrów dla efektów stałych wielopoziomowego, wielomianowego modelu logitowego

Zmienna	Rozwiązanie niekomercyjne		Rozwiązanie komercyjne	
	Oszacowanie	GPI	Oszacowanie	GPI
FEM (kobieta)	8,877	0,049	0,322	0,568
LVAL (problem ważny)	3,347	0,000	4,202	0,76
LLEG (problem z zakresu prawa spadkowego)	–0,427	0,507	2,265	0,003
LPRO (problem z zakresu prawa rzeczowego)	0,721	0,407	2,537	0,011
LFAM (problem z zakresu prawa rodzinnego)	0,150	0,830	1,964	0,025
LPEN (problem z zakresu prawa karnego)	0,247	0,852	2,824	0,055
PUSE (postawa wobec stosowania prawa)	0,269	0,199	0,584	0,030
PAWR (świadomość prawna)	0,167	0,083	0,268	0,050
Testowanie istotności efektów losowych GPI = 0,000		R^2-Mc Faddena = 0,52		R^2-Nagelkerke'a = 0,73

Źródło: obliczenia własne.

Po oszacowaniu odpowiednich parametrów obliczony został odwrócony iloraz Millsa, zgodnie ze wzorem (5.20). Wielkość ta jest następnie wykorzystywana jako

zmienna objaśniająca w równaniu wyjaśniającym sposób reakcji wobec zaistnienia problemu prawnego. Tabela 43 prezentuje wyniki estymacji parametrów wielopoziomowego modelu wielomianowego logitowego. Prezentowane są oszacowania parametrów przy zmiennych istotnych na poziomie istotności 0,1 dla co najmniej jednego z równań oraz graniczne poziomy istotności. Oprócz tego prezentowane są wartości funkcji eksponencjalnych dla oszacowań w celu dokonania interpretacji wpływu określonych cech społeczno-ekonomicznych respondentów na prawdopodobieństwo wyboru odpowiedniego sposobu reakcji wobec zaistnienia problemu prawnego.

Oszacowania parametrów wielopoziomowego, wielomianowego modelu logitowego są zgodne z oczekiwaniami. Są one także częściowo zgodne z rezultatami otrzymanymi przez Florczaka i Grabowskiego (2018b). Należy jednak pamiętać, że w cytowanym artykule pominięty został problem selekcji próby. Nie uwzględniono także losowego zróżnicowania sposobu reakcji wobec zaistnienia problemu prawnego między regionami. W tabeli 43 brakuje oszacowania przy zmiennej *IMR*, co wynika z faktu, że okazała się ona nieistotna zarówno w równaniu związanym z wyborem niekomercyjnego sposobu rozwiązania problemu, jak i w równaniu wyjaśniającym iloraz szans dla wyboru komercyjnego rozwiązania względem braku reakcji. Wyniki testowania efektów losowych wskazują, że model zakładający brak losowego zróżnicowania sposobu reakcji jest istotnie gorszy. Dlatego też zastosowanie modelu wielopoziomowego jest lepszym sposobem identyfikacji determinant sposobu reakcji wobec wystąpienia problemu prawnego.

Zgodnie z oczekiwaniami oraz wynikami badań dla innych krajów (por. Kritzer, 2008; Pleasance, Balmer, Reimers, 2011) waga problemu odgrywa bardzo dużą rolę w wyborze sposobu reakcji. Jeśli respondent uznał problem prawny (którego doświadczył w ciągu ostatnich pięciu lat) za ważny, wówczas relacja prawdopodobieństwa przyjęcia aktywnej postawy względem postawy pasywnej była zdecydowanie wyższa w porównaniu z sytuacją, gdy waga problemu była niższa. Dotyczy to przede wszystkim wyboru komercyjnej metody rozwiązania problemu. Skłonność do szukania porady prawnej u adwokata czy radcy prawnego była zdecydowanie wyższa w grupie osób, które doświadczyły istotnego problemu. Rodzaj problemu prawnego również miał znaczny wpływ na poszukiwanie sposobu jego rozwiązania. Skłonność do ponoszenia kosztów materialnych w związku z rozwiązaniem problemu prawnego okazała się zdecydowanie wyższa w grupie osób, które doświadczyły problemów związanych z prawem spadkowym, rzeczowym, rodzinnym czy karnym. Ewentualne korzyści (również w postaci uniknięcia strat) z rozwiązania problemów z zakresu prawa spadkowego, rzeczowego czy rodzinnego często okazują się zdecydowanie wyższe niż ponoszone koszty. Dlatego też osoby doświadczające problemów z analizowanej grupy chętnie korzystają z pomocy adwokata czy radcy prawnego.

Zdają sobie oni sprawę z tego, że zaniechanie rozwiązywania problemu może wiązać się z poniesieniem znacznych kosztów. Dobra reprezentacja strony sprawy przed sądem może pomóc w uzyskaniu korzyści zdecydowanie wyższych od poniesionych kosztów (por. Winczorek, 2015). Jeśli zaś chodzi o prawo karne, to brak rozwiązania problemów z jego zakresu może wiązać się z koniecznością zapłaty wysokiej grzywny, ograniczeniem lub nawet utratą wolności. Dlatego też respondenci, którzy doświadczyli problemu z zakresu prawa karnego, zdecydowanie częściej prosili o pomoc adwokata niż osoby doświadczające innych problemów.

Analizując wpływ pozostałych kategorii, należy zwrócić uwagę na istotność zmiennej *FEM* w równaniu związanym z komercyjnym sposobem rozwiązania problemu. Oznacza to zatem, że kobiety częściej niż mężczyźni przyjmują aktywną postawę. Wynik ten jest zgodny z rezultatami innych badań przeprowadzonych między innymi przez Murayamę (2007), Hunter i de Simeone (2009), a także Preiserta i in. (2013). W tamtych pracach również wskazano na istotną statystycznie rolę płci w determinowaniu sposobu reakcji na wystąpienie problemu. Dwie zmienne – związane ze świadomością prawną, a także postawą wobec przestrzegania i stosowania prawa – również okazały się istotne statystycznie. Wraz ze wzrostem świadomości prawnej następuje, przy innych czynnikach niezmienionych, wzrost prawdopodobieństwa wyboru komercyjnego lub niekomercyjnego rozwiązania. Osoba charakteryzująca się wyższym poziomem świadomości prawnej zdecydowanie lepiej zna swoje prawa oraz ewentualne konsekwencje związane z brakiem reakcji w sytuacji wystąpienia problemu prawnego. Wydaje się także, że osoby o wyższym poziomie wiedzy z zakresu prawa lepiej potrafią wskazywać instytucje zajmujące się rozwiązywaniem konkretnych problemów. Dlatego też charakteryzujący się wysoką świadomością prawną obywatele próbują znaleźć rozwiązanie powstałych problemów prawnych u adwokata, radcy prawnego lub w innej kompetentnej instytucji. Podobne rezultaty, odnoszące się do zależności między poziomem świadomości prawnej a wyborem sposobu reakcji względem problemu prawnego, uzyskali w swoich pracach między innymi Pascoe Pleasance i in. (2003) oraz Pascoe Pleasance, Nigel Balmer i Stian Reimers (2011). Zmienna wskazująca na poziom skargliwości również okazała się mieć istotny wpływ na sposób reakcji wobec wystąpienia problemu prawnego. Okazuje się, że osoby bardziej skargliwe względem prawa chętniej decydują się na wybór komercyjnego sposobu rozwiązania problemu w porównaniu z obywatelami bardziej ugodowymi. Wynika to zapewne z faktu, że wraz ze wzrostem skargliwości następuje wzrost skłonności do dochodzenia swoich praw. Dlatego też bierna postawa wobec wystąpienia problemu prawnego jest rzadziej przyjmowana.

Po przeprowadzeniu testu ilorazu wiarygodności okazało się, że istnieją istotne losowe różnice w sposobie reakcji wobec wystąpienia problemu prawnego między mieszkańcami różnych makroregionów. Dlatego też rozważany był

wielopoziomowy, wielomianowy model logitowy. Chociaż podejmowano próby estymacji parametrów modelu uwzględniającego efekty losowe dla makroregionów i województw, wyniki nie były zadowalające. Może to wynikać z faktu, że liczba osób doświadczających problemu prawnego jest stosunkowo niewielka (286 obserwacji) i silnie zróżnicowana między województwami. Zdarzają się województwa, w których liczba osób podejmujących konkretny rodzaj reakcji nie przekracza trzech. Uzyskane wyniki estymacji parametrów takiego modelu byłyby mało precyzyjne, a co za tym idzie – mało wiarygodne. Dlatego też oszacowano parametry modelu z efektami losowymi dla regionów historycznych omówionych w podrozdziale 2.2. W tabeli 44 prezentowane są predykcje efektów losowych dla poszczególnych makroregionów.

Tabela 44. Oszacowania efektów losowych dla poszczególnych regionów historycznych

Region historyczny	Oszacowania efektów losowych w równaniu wyboru rozwiązania niekomercyjnego	Oszacowania efektów losowych w równaniu wyboru rozwiązania komercyjnego
Zabór pruski	1,021	1,510
Zabór rosyjski	–0,130	–0,235
Zabór austriacki	–0,515	–0,561
Ziemie Zachodnie i Północne	–0,376	–0,714

Źródło: obliczenia własne.

Oszacowania dla efektów losowych wskazują, że lokalizacja osoby napotykającej na problem prawny ma ważny wpływ na wybór sposobu reakcji. Osoby zamieszkujące tereny byłego zaboru pruskiego charakteryzują się zdecydowanie większą skłonnością do poszukiwania aktywnego sposobu rozwiązania zaistniałego problemu prawnego. Dotyczy to zarówno rozwiązań komercyjnych, jak i niekomercyjnych. Uzyskany rezultat może być wyjaśniony XIX-wieczną historią obecnych terenów Polski. W kulturze pruskiej w XIX wieku wykształcone zostały wzorce formalnego rozwiązywania problemów, podpisywania umów przed wykonaniem usługi i częstego wchodzenia na drogę prawną w sytuacji niezadowolenia ze sposobu realizacji zadań. Sądownictwo i formalne instytucje mające na celu rozwiązywanie problemów prawnych wykształciły się w tamtym okresie na skalę nieodnotowywaną w innych częściach Polski (por. Zarycki, 2015). Poszukiwanie aktywnego sposobu rozwiązania problemu prawnego wpisane zostało w kulturę osób zamieszkujących tereny zaboru pruskiego. Te normy kulturowe były następnie przekazywane z pokolenia na pokolenie. Dlatego też (przy innych czynnikach niezmienionych) mieszkańcy Poznania, Torunia, Bydgoszczy, Trójmiasta i okolic istotnie częściej wybierają aktywną metodę rozwiązania problemu prawnego. Osoby zamieszkujące teren byłego zaboru rosyjskiego, austriackiego czy też

Ziemie Zachodnie i Północne częściej przyjmują bierną postawę w sytuacji zaistnienia problemu prawnego. Dotyczy to zwłaszcza mieszkańców takich województw jak dolnośląskie, lubuskie czy zachodniopomorskie. Dopiero po II wojnie światowej analizowane tereny zostały zamieszkane przez ludność polską. Formalne i nieformalne instytucje mające na celu rozwiązywanie konfliktów prawnych zaczęły być tam tworzone dopiero po 1950 roku. Dlatego też mniejsza skłonność do wybierania formalnego sposobu rozwiązania problemu prawnego jest niższa w analizowanym regionie w porównaniu z pozostałymi częściami Polski.

W celu sprawdzenia wpływu nieuwzględnienia efektów losowych na jakość dopasowania modelu do danych współczynnik R^2-Nagelkerke został policzony zarówno dla modelu uwzględniającego efekty losowe w obu krokach, jak i dla modelu bez efektów losowych. W przypadku zignorowania efektów losowych w obu krokach wartość współczynnika R^2-Nagelkerke wyniosła 0,49, a więc okazała się o około 33% niższa w porównaniu z modelem uwzględniającym efekty losowe. Wynik ten, jak również rezultat testowania obecności efektów losowych, wskazują, że w modelu wyjaśniającym prawdopodobieństwo zaistnienia problemu prawnego oraz wybór sposobu reakcji na niego uwzględnienie czynników kontekstowych jest ważne. Oznacza to zatem, że wartości analizowanych kategorii z zakresu socjologii prawa zależą nie tylko od cech indywidualnych mieszkańców, ale także od czynników kulturowych.

Zakończenie

W niniejszej monografii rozważano modele wielopoziomowe dla różnych wariantów z punktu widzenia kształtowania się zmiennej zależnej. Analizowano przypadki ciągłej, binarnej, polichotomicznej uporządkowanej, polichotomicznej nieuporządkowanej, licznikowej oraz rankingowej zmiennej zależnej. Metody estymacji parametrów tych modeli zostały szczegółowo omówione. Jednocześnie zaprezentowane zostały wyniki badań empirycznych, w których wykorzystane były metody estymacji parametrów modeli wielopoziomowych.

Rezultaty badań empirycznych oraz porównanie jakości dopasowania między modelami zawierającymi zmienne kontekstowe, a ignorującymi rolę kontekstu wskazują, że brak uwzględnienia zmiennej związanej z przynależnością do regionu powoduje pogorszenie jakości modelu. Uzyskane wyniki pokazują, że uwzględnianie losowych różnic w wartościach zmiennej wynikowej między grupami oraz rozważanie losowych różnic we wpływie określonych kategorii mikroekonomicznych na zmienną zależną mogą przyczynić się do poprawy jakości wnioskowania i sprawić, że prognozowane wartości są bliższe rzeczywistym.

Oszacowania parametrów modelu wielopoziomowego wyjaśniającego wysokość wynagrodzeń w polskich przedsiębiorstwach wskazują, że modele nieuwzględniające czynników kontekstowych nie są w stanie dokładnie zilustrować zależności między charakterystykami pracowników a wysokością wynagrodzeń. Rezultaty wskazujące na losowe różnice między wynagrodzeniami osób o tych samych charakterystykach indywidualnych i pracującymi w dwóch różnych regionach nie są zaskoczeniem. Międzyregionalne zróżnicowanie poziomu rozwoju ekonomiczno-społecznego wydaje się być stabilne. Uzyskane wyniki estymacji wskazują jednak, że w grupie regionów bogatych w latach 2004–2016 nastąpiły zmiany w klasyfikacji, jeśli chodzi o sytuację na rynku pracy. Predykcje efektów losowych dla województwa dolnośląskiego wyprzedziły odpowiednie wartości dla regionów ze stolicami w Gdańsku, Poznaniu i Katowicach. Jednocześnie obserwowana konwergencja do regionów bogatych jest silniejsza w przypadku województwa lubuskiego czy zachodniopomorskiego niż w regionach ze stolicami w Białymstoku czy Kielcach. Uzyskane rezultaty wskazują także na istotne, ale zmieniające się w czasie efekty losowe dla sekcji PKD. Wykorzystanie modelu wielopoziomowego i klasyfikacja trzycyfrowych grup zawodowych ze względu na poziom kwalifikacji pracowniczych oraz charakter wykonywanych zadań umożliwiła weryfikację

hipotezy dotyczącej polaryzacji na polskim rynku pracy. Uzyskane rezultaty wskazują na występowanie analizowanego zjawiska. Między 2004 a 2016 rokiem nastąpił wzrost różnicy w wynagrodzeniach osób o najwyższych i najniższych kwalifikacjach. Okazało się jednak, że – w związku z tendencjami zachodzącymi na rynku pracy w ostatnich latach – najbardziej poszkodowane są osoby o średnich kwalifikacjach.

Wyniki estymacji parametrów modelu wyjaśniającego skłonność do inwestowania w technologie informatyczne i komunikacyjne, innowacyjność przedsiębiorstw oraz ich produktywność wskazują na ważną rolę czynników komplementarnych. Chociaż czynniki indywidualne okazały się być ważniejszymi determinantami postaw innowacyjnych polskich firm, należy podkreślić, że istnieją istotne losowe różnice między przedsiębiorstwami zlokalizowanymi w różnych regionach. Różnice te dotyczą również wpływu poszczególnych cech firm na ich zachowania innowacyjne. Jednocześnie jakość regionalnych systemów innowacji okazała się istotną determinantą w modelu wyjaśniającym prawdopodobieństwo inwestowania w technologie informatyczne i komunikacyjne, wprowadzania różnych typów innowacji oraz produktywności.

Wyniki estymacji parametrów modelu wyjaśniającego skłonność do doświadczania problemu prawnego oraz sposób reakcji nań również wskazują na ważną rolę kontekstu. Okazuje się, że aktywną postawę wobec zaistnienia problemu prawnego częściej przyjmowały osoby mieszkające na terenie byłego zaboru pruskiego, co mogło wynikać z faktu, że na analizowanym obszarze system prawny rozwinięty był już w XIX wieku. Mogło to spowodować, że mieszkańcy Wielkopolski, Kujaw i Pomorza jeszcze w dwudziestoleciu międzywojennym chętniej odwoływali się do instytucji w celu rozwiązania problemu prawnego. Taka postawa mogła być przekazywana następnie z pokolenia na pokolenie. W pozostałych częściach Polski, gdzie instytucje zajmujące się rozwiązywaniem konfliktów prawnych powstały w dwudziestoleciu międzywojennym (tereny byłego zaboru rosyjskiego oraz austriackiego) lub po II wojnie światowej (byłe tereny niemieckie), skłonność do przyjmowania aktywnej postawy wobec zaistnienia problemu prawnego jest zdecydowanie niższa.

Bibliografia

Accetturo A., Dalmazzo A., de Blasio G. (2013), *Skill polarization in local labor markets under share-altering technical change*, „Journal of Regional Science", vol. 54(2), s. 249–272.

Acemoglu D. (2002), *Technical Change, Inequality, and the Labor Market*, „Journal of Economic Literature", vol. 40(1), s. 7–72.

Acemoglu D., Autor D. (2011), *Skills, tasks and technologies: Implications for employment and earnings*, [w:] O. Ashenfelter, D. Card (red.), *Handbook of Labor Economics*, Elsevier, Amsterdam, s. 1043–1166.

Adamczyk A., Tokarski T., Włodarczyk R. (2009), *Przestrzenne zróżnicowanie płac w Polsce*, „Gospodarka Narodowa", nr 1(1), s. 1–20.

Aghion P., Howitt P. (2008), *The Economics of Growth*, MIT Press, Cambridge.

Albert P., Follmann D. (2000), *Modeling Repeated Count Data Subject to Informative Dropout*, „Biometrics", vol. 56(3), s. 667–677.

Allison P., Christakis N. (1994), *Logit Models for Sets of Ranked Items*, „Sociological Methodology", no. 24, s. 199–228.

Amemiya T. (1978), *The Estimation of a Simultaneous Equation Generalized Probit Model*, „Econometrica", vol. 46(5), s. 1193–1205.

Arendt Ł. (2018), *Is the Polish Labour Market heading towards Polarisation?*, „Olsztyn Economic Journal", vol. 13(3), s. 309–322.

Arendt Ł., Grabowski W. (2017), *Innovations, ICT and ICT-driven labour productivity in Poland*, „Economics of Transition", vol. 25(4), s. 723–758.

Arendt Ł., Grabowski W. (2018), *Impact of ICT Utilization on Innovations and on Labor Productivity: Micro-level Analysis for Poland*, [w:] A. Dias, B. Salmelin, D. Pereira, J. Dias (red.), *Modeling Innovation Sustainability and Technologies*, 1st ed., Springer International Publishing, New York, s. 225–247.

Arendt Ł., Kryńska E. (2015), *Technologie informacyjne i komunikacyjne a produktywność w Polsce i krajach Europy Środkowo-Wschodniej*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Ark B., Piątkowski M. (2004), *Productivity Innovation and TIK in Old and New Europe*, „International Economics and Economic Policy", vol. 1(2–3), s. 215–246.

Autor D., Dorn D. (2013), *The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market*, „American Economic Review", vol. 103(5), s. 1553–1597.

Autor D., Levy F., Murnane R. (2003), *The Skill Content of Recent Technological Change: An Empirical Exploration*, „The Quarterly Journal of Economics", vol. 118(4), s. 1279–1333.

Bartkowski J. (2003), *Tradycja i polityka*, Wydawnictwo Akademickie Żak, Warszawa.

Bazyl M. (2010), *Modele zmiennych licznikowych*, [w:] M. Gruszczyński (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Oficyna a Wolters Kluwer business Warszawa, s. 217–230.

Beggs S., Cardell S., Hausman J. (1981), *Assessing the potential demand for electric cars*, „Journal of Econometrics", vol. 17(1), s. 1–19.

Ben-Akiva M., Lerman S.R. (1985), *Discrete choice analysis. Theory and application to travel demand*, MIT Press, Cambridge.

Bergmann H., Japsen A., Tamasy C. (2002), *Regionaler Entrepreneurship Monitor (REM). Grundungsaktivitaten and Rahmenbedingungen in zehn deutschen Regionen*, Wirtschafts- and Sozialgeographisches Institut, Köln.

Blundell R., Windmeijer F. (1997), *Cluster effects and simultaneity in multilevel models*, „Health Economics", vol. 6(4), s. 439–443.

Borjas G., Sueyoshi G. (1994), *A two-stage estimator for probit models with structural group effects*, „Journal of Econometrics", vol. 64(1–2), s. 165–182.

Börsch-Supan A., Hajivassiliou V. (1993), *Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models*, „Journal of Econometrics", vol. 58(3), s. 347–368.

Breslow N., Clayton D. (1993), *Approximate Inference in Generalized Linear Mixed Models*, „Journal of the American Statistical Association", vol. 88(421), s. 9–25.

Breslow N., Lin X. (1995), *Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion*, „Biometrika", vol. 82(1), s. 81–91.

Bronzini R., Piselli P. (2016), *The impact of R&D subsidies on firm innovation*, „Research Policy", vol. 45(2), s. 442–457.

Burdziej S., Dudkiewicz M. (2013), *Korzystający i niekorzystający z poradnictwa prawnego i obywatelskiego*, Instytut Spraw Publicznych, Warszawa.

Butler J., Moffitt R. (1982), *A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model*, „Econometrica", vol. 50(3), s. 761–764.

Cameron A., Trivedi P. (2009), *Microeconometrics*, 8th ed., Cambridge University Press, Cambridge.

Cardoso A. (2000), *Wage differentials across firms: an application of multilevel modelling*, „Journal of Applied Econometrics", vol. 15(4), s. 343–354.

Carey K. (2000), *A multilevel modelling approach to analysis of patient costs under managed care*, „Health Economics", vol. 9(5), s. 435–446.

Carlin B.P., Louis T.A. (2000), *Bayes and empirical bayes methods for data analysis*, 2nd ed., Chapman and Hall/CRC, New York.

Chan K., Ledolter J. (1995), *Monte Carlo EM Estimation for Time Series Models Involving Counts*, „Journal of the American Statistical Association", vol. 90(429), s. 242–252.

Chapman R., Staelin R. (1982), *Exploiting Rank Ordered Choice Set Data within the Stochastic Utility Model*, „Journal of Marketing Research", vol. 19(3), s. 288–301.

Cieślik A., Rokicki B. (2016), *Individual wages and regional market potential*, „Economics of Transition", vol. 24(4), s. 661–682.

Clemens M., Montenegro C., Prichett L. (2009), *The Place Premium: Wage Differences for Identical Workers Across the US Border*, Harvard Library Research Working Paper, 09–004, s. 1–68.

Coad A., Segarra A., Teruel M. (2016), *Innovation and firm growth: Does firm age play a role?*, „Research Policy", vol. 45(2), s. 387–400.

Congdon P. (2005), *Bayesian Models for Categorical Data*, John Wiley & Sons, London.

Cramer J. (1999), *Predictive Performance of the Binary Logit Model in Unbalanced Samples*, „Journal of the Royal Statistical Society: Series D (The Statistician)", vol. 48(1), s. 85–94.

Crepon B., Duguet E., Mairessec J. (1998), *Research, Innovation and Productivity: An Econometric Analysis at the Firm Level*, „Economics of Innovation And New Technology", vol. 7(2), s. 115–158.

Cunha F., Heckman J. (2007), *The Technology of Skill Formation*, „American Economic Review", vol. 97(2), s. 31–47.

Domański H. (2018), *Wpływ wykształcenia na rozkład zarobków w Polsce w latach 1988–2013*, „Ekonomista", nr 1, s. 7–24.

Efron B. (1978), *Regression and ANOVA with Zero-One Data: Measures of Residual Variation*, „Journal of the American Statistical Association", vol. 73(361), s. 113–121.

Ehrenberg R., Schwarz J. (1987), *Public Sector and Labor Markets*, Cornell University ILR School Working Paper.

Evangelou E., Eidsvik J. (2017), *The value of information for correlated GLMs*, „Journal of Statistical Planning And Inference", no. 180, s. 30–48.

Evangelou E., Zhu Z., Smith R. (2011), *Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation*, „Journal of Statistical Planning And Inference", vol. 141(11), s. 3564–3577.

Faggio G., Salvanes K., Van Reenen J. (2010), *The evolution of inequality in productivity and wages: panel data evidence*, „Industrial And Corporate Change", vol. 19(6), s. 1919–1951.

Florczak W. (2016), *Modelling Effective Legal Aid System*, „Ekonomia i Prawo", nr 15(3), s. 317–334.

Florczak W., Grabowski W. (2017), W*ystąpienie problemu prawnego jako funkcja czynników indywidualnych i kontekstowych. Analiza ekonometryczna z wykorzystaniem hierarchicznego modelu logitowego*, „Studia Prawno-Ekonomiczne", nr 105, s. 193–213.

Florczak W., Grabowski W. (2018a), *Analiza czynników determinujących reakcję na zaistnienie problemu prawnego przy użyciu wielomianowego modelu logitowego*, „Wiadomości Statystyczne", nr 63(1), s. 57–76.

Florczak W., Grabowski W. (2018b), *Czy warto korzystać z porad prawnych? Szacunki mikroekonomicznych efektów poradnictwa prawno-obywatelskiego*, „Ekonomista", nr 2, s. 185–208.

Florczak W., Grabowski W. (2018c), *Co wpływa na wielkość popytu na porady prawne? Analiza logitowa z wykorzystaniem metody klasycznego uśredniania międzymodelowego*, „Przegląd Statystyczny", nr 65, s. 53–80.

Fortuna Z., Macukow B., Wąsowski J. (2017), *Metody numeryczne*, Wydawnictwa Naukowo-Techniczne, Warszawa.

Freeman R.B., Katz L.F. (1994), *Rising wage inequality: the United States vs. other advanced countries*, [w:] R.B. Freeman (red.), *Working under different rules*, Russell Sage Foundation, New York, s. 29–62.

Gelfand A., Carlin B. (1993), *Maximum-likelihood estimation for constrained- or missing-data models*, „Canadian Journal of Statistics", vol. 21(3), s. 303–311.

Geyer C.J., Thompson E.A. (1992), *Constrained Monte Carlo Maximum Likelihood for Dependent Data*, „Journal of the Royal Statistical Society B", vol. 54(3), s. 657–699.

Godfrey L., Wickens M. (1982), *A simple derivation of the limited information maximum likelihood estimator*, „Economics Letters", vol. 10(3–4), s. 277–283.

Goldstein H. (1986), *Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares*, „Biometrika", vol. 73(1), s. 43–56.

Goldstein H. (1989), *Models for multilevel response variables with an application to growth curves*, [w:] R.D. Bock (red.), *Multilevel analysis of educational data*, Academic Press, San Diego, s. 107–125.

Goldstein H. (1991), *Nonlinear Multilevel Models, with an Application to Discrete Response Data*, „Biometrika", vol. 78(1), s. 45–51.

Golejewska A. (2018), *Innovativeness of Enterprises in Poland in the Regional Context*, „Journal of Entrepreneurship, Management and Innovation", vol. 14(1), s. 29–44.

Goos M., Manning A. (2007), *Lousy and Lovely Jobs: The Rising Polarization of Work in Britain*, „Review of Economics and Statistics", vol. 89(1), s. 118–133.

Gorzelak G., Jałowiecki B. (1996), *Koniunktura gospodarcza i mobilizacja społeczna w gminach '95*, Uniwersytet Warszawski – Europejski Instytut Rozwoju Regionalnego i Lokalnego, Warszawa.

Grabowski W. (2018), *Determinanty przestrzennego zróżnicowania wyników głosowania w wyborach parlamentarnych z 2015 roku*, „Studia Socjologiczne", vol. 228(1), s. 35–64.

Grabowski W. (2019), *Does the use of professional legal assistance bring measurable benefits?*, „Applied Economics Letters", w druku.

Grabowski W., Skorupińska A. (2015), *TIK i czynniki komplementarne – ujęcie modelowe*, [w:] Ł. Arendt, E. Kryńska (red.), *Technologie informacyjne i komunikacyjne a produktywność w Polsce i krajach Europy Środkowo-Wschodniej*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, s. 138–146.

Grabowski W., Stawasz E. (2017), *The Role of Business Consulting in Creating Knowledge and Formulating a Strategy of Development in Polish Micro-Enterprises*, „Journal of East European Management Studies", vol. 22(3), s. 374–396.

Green D., Sand B. (2015), *Has the Canadian labour market polarized?*, „Canadian Journal of Economics", vol. 48(2), s. 612–646.

Green P. (1987), *Penalized Likelihood for General Semi-Parametric Regression Models*, „International Statistical Review/Revue Internationale De Statistique", vol. 55(3), s. 245–259.

Greene W. (2008), *Econometric analysis*, Pearson, Upper Saddle River.

Gretton P., Gali J., Parham D. (2004), *The effects of ICTs and complementary innovations on Australian productivity growth*, [w:] *The Economic Impact of ICT: Measurement, Evidence and Implications*, OECD Publishing, Paris, s. 105–130.

Gruszczyński M. (2012), *Mikroekonometria*, Wolters Kluwer Polska, Warszawa.

Hajivassiliou V., Ruud P. (1994), *Classical estimation methods for LDV models using simulation*, [w:] R. Engle, D. McFadden (red.), *Handbook of Econometrics*, Elsevier Science, Amsterdam, s. 2384–2438.

Hall B., Lotti F., Mairesse J. (2013), *Evidence on the impact of R&D and ICT investments on innovation and productivity in Italian firms*, „Economics of Innovation And New Technology", vol. 22(3), s. 300–328.

Hall B., Mairesse J., Mohnen P. (2010), *Measuring the Returns to R&D*, [w:] B. Hall, N. Rosenberg (red.), *Handbook of the Economics of Innovation*, Elsevier Science, Amsterdam, s. 1033–1082.

Hann C., Magocsi P. (2014), *Galicia*, University of Toronto Press, Toronto.

Hardy W., Keister R., Lewandowski P. (2018), *Educational upgrading, structural change and the task composition of jobs in Europe*, „Economics of Transition", vol. 26(2), s. 201–231.

Harville D. (1977), *Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems*, „Journal of the American Statistical Association", vol. 72(358), s. 320–338.

Heckman J. (1979), *Sample Selection Bias as a Specification Error*, „Econometrica", vol. 47(1), s. 153–161.

Hollenstein H., Stucki T. (2012), *The 'new firm paradigm' and the provision of training: The impact of ICT, workplace organization and human capital*, „Swiss Journal of Economics and Statistics", vol. 148(4), s. 557–595.

Hosmer D., Lemeshow S., May S. (1999), *Applied Survival Analysis: Regression Modelling of Time-to-Event Data*, Wiley & Sons, London.

Hosseini Shojaei R., Waghei Y., Mohammadzadeh M. (2018), *Parameter Estimation in Spatial Generalized Linear Mixed Models with Skew Gaussian Random Effects using Laplace Approximation*, „Journal of Statistical Research of Iran", vol. 14(2), s. 157–169.

Huber P. (1964), *Robust Estimation of a Location Parameter*, „The Annals of Mathematical Statistics", vol. 35(1), s. 73–101.

Hunter R., Simone T. de (2009), *Women, Legal Aid and Social Inclusion*, „Australian Journal of Social Issues", vol. 44(4), s. 379–398.

Hühne P., Herzer D. (2017), *Is inequality an inevitable by-product of skill-biased technical change?*, „Applied Economics Letters", vol. 24(18), s. 1346–1350.

Jałowiecki B. (1996), *Przestrzeń historyczna, regionalizm, regionalizacja*, [w:] tenże (red.), *Oblicza polskich regionów*, Warszawa, s. 19–88.

Jasiewicz Z. (1977), *Rodzina wiejska na Ziemi Lubuskiej*, Państwowe Wydawnictwo Naukowe, Warszawa.

Jezierski A., Leszczyńska C. (2011), *Historia gospodarcza Polski*, Wydawnictwo Key Text, Warszawa.
Jonek-Kowalska I. (2014), *Employment and Renumeration Trends in Polish Hard Coal Mines in the Context of the Relations Between Boards and Trade Unions*, „International Journal of Synergy and Research", vol. 3, s. 27–43.
Katz L., Autor D. (1999), *Changes in the wage structure and earnings inequality*, [w:] O. Ashenfelter, D. Card (red.), *Handbook of Labor Economics*, Elsevier, Amsterdam, s. 1463–1555.
Kay R., Little S. (1986), *Assessing the Fit of the Logistic Model: A Case Study of Children with the Haemolytic Uraemic Syndrome*, „Applied Statistics", vol. 35(1), s. 16–30.
Keane M. (1994), *A Computationally Practical Simulation Estimator for Panel Data*, „Econometrica", vol. 62(1), s. 95–116.
Keener R.W., Waldman D.M. (1985), *Maximum Likelihood Regression of Rank-Censored Data*, „Journal of the American Statistical Association", vol. 80(390), s. 385–392.
Khoshgoftaar T., Gao K., Szabo R. (2005), *Comparing software fault predictions of pure and zero-inflated Poisson regression models*, „International Journal of Systems Science", vol. 36(11), s. 705–715.
King A. (1978), *Industrial Structure, the Flexibility of Working Hours, and Women's Labor Force Participation*, „The Review of Economics and Statistics", vol. 60(3), s. 399–407.
Klamka J., Ogonowski Z. (2015), *Metody numeryczne*, Wydawnictwo Politechniki Śląskiej, Gliwice.
Kochanowicz J. (2018), *The Polish Kingdom: A periphery as a leader*, XIV International Economic History Congress in Helsinki, Helsinki.
Koellinger P. (2005), *Why IT matters? An empirical study of e-business usage, innovation and firm performance*, German Institute for Economic Research Discussion Paper, no. 495.
Koenker R., Bassett G. (1978), *Regression Quantiles*, „Econometrica", vol. 46(1), s. 33–50.
Kortt M., Dollery B. (2012), *Religion and the rate of return to human capital: evidence from Australia*, „Applied Economics Letters", vol. 19(10), s. 943–946.
Kosała M., Wach K. (2011), *Regionalne determinanty rozwoju innowacyjności przedsiębiorstw*, „Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie", z. 866, s. 19–35.
Kritzer H. (2008), *To Lawyer, or Not to Lawyer, is That the Question?*, „Journal of Empirical Legal Studies", no. 5, s. 875–906.
Krzemiński P. (2009), *Zachowania wyborcze w wyborach parlamentarnych i prezydenckich w Polsce w latach 2005–2007 – wzory przestrzennych zróżnicowań*, „Przegląd Geograficzny", vol. 81(2), s. 259–281.
Kuk A.Y.C. (1995), *Asymptotically unbiased estimation in generalized linear models with random effects*, „Journal of the Royal Statistical Society B", no. 57, s. 395–407.
Kurczewski J., Fuszara M. (2004), *Polskie spory i sądy*, Ośrodek Badań Społecznych Instytut Stosowanych Nauk Społecznych Uniwersytetu Warszawskiego, Warszawa.
Lallemand T., Plasman R., Rycx F. (2009), *Wage structure and firm productivity in Belgium*, [w:] E.P. Lazear, K.L. Shaw (red.), *The Structure of Wages: An International Comparison*, Chicago University Press, Chicago, s. 179–215.
Langendijk A. (2001), *Regional Learning between Variation and Convergence: The Concept of 'Mixed-land-use' in Regional Spatial Planning in the Netherland*s, „Canadian Journal of Regional Science", vol. 24(1), s. 135–154.
Lauer C., Steiner V. (2000), *Returns to Education in West Germany – An Empirical Assessment*, ZEW Discussion Paper, 00–04.
Lee L. (1981), *Simultaneous equations models with discrete and censored dependent variables*, [w:] C. Manski, D. McFadden (red.), *Structural analysis of discrete data with economic applications*, MIT Press, Cambridge.
Lee L. (2000), *A numerically stable quadrature procedure for the one-factor random-component discrete choice model*, „Journal of Econometrics", vol. 95(1), s. 117–129.

Legal Service Corporation (1994), *Legal Needs and Civil Justice. A survey of Americans. Major Findings from comprehensive Legal Needs Study*, American Bar Association Working Paper, no. 1.

Lesaffre E., Spiessens B. (2001), *On the effect of the number of quadrature points in a logistic random effects model: an example*, „Journal of the Royal Statistical Society: Series C (Applied Statistics)", vol. 50(3), s. 325–335.

Lewandowska M.S. (2016), *Do Government Policies Foster Environmental Performance of Enterprises from CEE Region?*, „Comparative Economic Research", vol. 19(3), s. 45–67.

Lewandowska M.S., Kowalski A.M. (2015), *Współpraca polskich przedsiębiorstw w sferze innowacji a wsparcie z funduszy unijnych*, „Gospodarka Narodowa", nr 4(278), s. 69–89.

Liwiński J., Bedyk E. (2016), *Does it pay to invest in the education of children?*, „Ekonomia. Rynek, Gospodarka, Społeczeństwo", nr 47, s. 53–77.

Long J., Freese J. (2014), *Regression models for categorical dependent variables using Stata*, Stata Press Publ., College Station.

Longford N. (1987), *A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects*, „Biometrika", vol. 74(4), s. 817–827.

Longford N. (1994), *Logistic regression with random coefficients*, „Computational Statistics & Data Analysis", vol. 17(1), s. 1–15.

Loof H., Heshmati A. (2006), *On the Relationship between Innovation and Performance: A Sensitivity Analysis*, „Economics of Innovation and New Technology", vol. 13, s. 317–344.

Luce R. (2005), *Individual choice behavior*, Dover Publications, Mineola.

Łaszkiewicz E. (2016), *Ekonometria przestrzenna III. Modele wielopoziomowe – teoria i zastosowania*, Wydawnictwo C.H. Beck, Warszawa.

Machin S., Puhani P. (2002), *Subject of Degree and the Gender Wage Differential. Evidence from the UK and Germany*, IZA Discussion Paper, no. 553.

Maddala G. (1987), *Limited-dependent and qualitative variables in econometrics*, Cambridge University Press, Cambridge.

Maddala G. (2013), *Ekonometria*, Polskie Wydawnictwo Naukowe, Warszawa.

Mahy B., Rycx F., Volral M. (2011), *Wage Dispersion and Firm Productivity in Different Working Environments*, „British Journal of Industrial Relations", vol. 49(3), s. 460–485.

Majcherek J. (1995), *Fenomen Małopolski*, „Rzeczpospolita", 1 grudnia 1995.

Majchrowska A., Strawiński P. (2016), *Regional Differences in Gender Wage Gaps in Poland: New Estimates Based on Harmonized Data for Wages*, „Central European Journal of Economic Modelling and Econometrics", vol. 8(2), s. 115–141.

Majchrowska A., Strawiński P. (2018), *Impact of minimum wage increase on gender wage gap: Case of Poland*, „Economic Modelling", no. 70, s. 174–185.

Majsterek M. (2008), *Wielowymiarowa analiza kointegracyjna w ekonomii*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Marzec J. (2008), *Bayesowskie modele zmiennych jakościowych i ograniczonych w badaniach niespłacalności kredytów*, Wydawnictwo Uniwersytetu Ekonomicznego, Kraków.

Matykowski R. (2007), *Zachowania wyborcze Wielkopolan: Czy występują odmienności w przestrzeni geograficzno-historycznej?*, [w:] J. Schmidt (red.), *Granica*, Wydawnictwo AWEI, Poznań, s. 75–92.

Matykowski R., Kulczyńska K. (2016), *Wybory do Sejmu w 2015 roku w województwie wielkopolskim: odmienności przestrzenne w kontekście subregionalnym i lokalnym*, „Rozwój Regionalny i Polityka Regionalna", nr 36, s. 163–178.

McCullagh P., Nelder J. (2000), *Generalized linear models*, Champman and Hall/CRC, London.

McCulloch C. (1997), *Maximum Likelihood Algorithms for Generalized Linear Mixed Models*, „Journal of the American Statistical Association", vol. 92(437), s. 162–170.

McFadden D. (1974), *Conditional Logit Analysis of Qualitative Choice Behaviour*, [w:] P. Zarembka (red.), *Frontiers in Econometrics*, Academic Press, New York, s. 105–142.
McKelvey R., Zavoina W. (1975), *A statistical model for the analysis of ordinal level dependent variables*, „The Journal of Mathematical Sociology", vol. 4(1), s. 103–120.
Mincer J. (1993), *Schooling, experience, and earnings*, Gregg Revivals, Aldershot.
Mincer J., Polachek S. (1974), *Family Investments in Human Capital: Earnings of Women*, „Journal of Political Economy", vol. 82(2, part 2), s. S76–S108.
Murayama M. (2007), *Experiences of Problems and Disputing Behavior in Japan*, „Meji Law Journal", no. 14, s. 1–59.
Nagelkerke N. (1991), *A Note on a General Definition of the Coefficient of Determination*, „Biometrika", vol. 78(3), s. 691–697.
Nakosteen R., Zimmer M. (1987), *Marital Status and Earnings of Young Men: A Model with Endogenous Selection*, „The Journal of Human Resources", vol. 22(2), s. 248–257.
Naylor J., Smith A. (1982), *Applications of a Method for the Efficient Computation of Posterior Distributions*, „Applied Statistics", vol. 31(3), s. 214–225.
Naylor J., Smith A. (1988), *Econometric illustrations of novel numerical integration strategies for Bayesian inference*, „Journal of Econometrics", vol. 38(1–2), s. 103–125.
Nelder J., Pawitan Y., Lee H. (2006), *Generalized linear models with random effects: unified analysis via H-likelihood*, Chapman & Hall/CRC, London.
Oesch D., Rodriguez Menes J. (2010), *Upgrading or polarization? Occupational change in Britain, Germany, Spain and Switzerland, 1990–2008*, „Socio-Economic Review", vol. 9(3), s. 503–531.
Oughton C., Landabaso M., Morgan K. (2002), *The Regional Innovation Paradox: Innovation Policy and Industrial Policy*, „Journal of Technology Transfer", vol. 27(1), s. 97–110.
Owczarczuk M. (2010), *Modele zmiennych ograniczonych*, [w:] M. Gruszczyński (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych*, Oficyna a Wolters Kluwer business, Warszawa, s. 193–215.
Parteka A. (2018), *Import Intensity of Production, Tasks and Wages: Micro-Level Evidence for Poland*, „Entrepreneurial Business And Economics Review", vol. 6(2), s. 71–89.
Pellegrino J.W., Hilton M.L. (2012), *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, The National Academic Press, Washington.
Pfirrmann O. (1994), *The geography of innovation in small and medium-sized firms in West Germany*, „Small Business Economics", vol. 6(1), s. 41–54.
Piątkowski M. (2004), *The impact of ICT on growth in transition economies*, Transformation, Integration and Globalization Economic Research Working Paper, no. 59.
Pinheiro J., Bates D. (1995), *Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model*, „Journal of Computational and Graphical Statistics", vol. 4(1), s. 12–35.
Plackett R. (1975), *The Analysis of Permutations*, „Applied Statistics", vol. 24(2), s. 193–202.
Pleasance P., Balmer N.J., Reimers S. (2011), *What Really Drives Advice Seeking Behaviour? Looking Beyond the Subject of Legal Disputes*, „Onati Socio-Legal Series", vol. 1(6), s. 1–21.
Pleasance P., Genn H., Balmer N.J., Buck A., O'Grady A. (2003), *Causes of Action: First Findings of the LSRC Periodic Survey*, „Journal of Law and Society", vol. 30(1), s. 11–30.
Pocztowski A., Pauli U. (2013), *Profesjonalizacja zarządzania zasobami ludzkimi w małych i średnich przedsiębiorstwach*, „Zarządzanie Zasobami Ludzkimi", nr 3–4, s. 9–22.
Podemski K., Ziółkowski M. (2007), *Czy Wielkopolska jest (społecznie) bliżej Europy?*, [w:] W. Molik, A. Sakson, T. Stryjkiewicz (red.), *Wielkopolska wobec wyzwań XXI wieku*, Centrum „Instytut Wielkopolski", Poznań, s. 29–41.
Porter M. (1998), *The Adam Smith Address: Location, Cluster, and the „New" Microeconomics of Competition*, „Business Economics", vol. 33(1), s. 1–7.
Porter M., Stern S. (2001), *Innovation: Location Matters*, „MIT Sloan Management Review", vol. 42, s. 28–36.

Preisert A., Schimanek T., Waszak M., Winiarska A. (2013), *Poradnictwo prawne i obywatelskie w Polsce. Stan obecny i wizje przyszłości*, Instytut Spraw Publicznych, Warszawa.

Psacharopoulos G., Ng Y.C. (1994), *Earnings and Education in Latin America*, „Education Economics", vol. 2(2), s. 187–207.

Rabe-Hesketh S., Skrondal A., Pickles A. (2005), *Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects*, „Journal of Econometrics", vol. 128(2), s. 301–323.

Raciborski J. (1997), *Polskie wybory*, Wydawnictwo Naukowe „Scholar", Warszawa.

Raudenbush S., Yang M., Yosef M. (2000), *Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation*, „Journal of Computational And Graphical Statistics", vol. 9(1), s. 141–157.

Rice N., Jones A. (1997), *Multilevel models and health economics*, „Health Economics", vol. 6(6), s. 561–575.

Rivers D., Vuong Q. (1988), *Limited information estimators and exogeneity tests for simultaneous probit models*, „Journal of Econometrics", vol. 39(3), s. 347–366.

Rodriguez G., Goldman N. (1995), *An Assessment of Estimation Procedures for Multilevel Models with Binary Responses*, „Journal of the Royal Statistical Society. Series A (Statistics in Society)", vol. 158(1), s. 73–89.

Roszkowska S., Majchrowska A. (2014), *Premia z wykształcenia i doświadczenia zawodowego według płci w Polsce*, Instytut Ekonomiczny, Warszawa.

Rousseeuw P., Leroy A. (2005), *Robust Regression and Outlier Detection*, John Wiley & Sons, Hoboken.

Rousseeuw P., Yohai V. (1987), *Robust Regression by Means of S-estimators*, [w:] J. Franke, W. Hardle, D. Martin (red.), *Robust and Nonlinear Time Series Analysis*, Springer Verlag, Berlin, s. 256–272.

Rousseeuw P., Zomeren B. van (1990), *Unmasking Multivariate Outliers and Leverage Points*, „Journal of the American Statistical Association", vol. 85(411), s. 633–639.

Searle S., Casella G., McCulloch C. (1992), *Variance components*, Wiley, New York.

Sidor-Rządkowska M. (2015), *Kształtowanie nowoczesnych systemów ocen pracowników*, Wydawnictwo Wolters Kluwer, Warszawa.

Sinnewe E., Kortt M., Steen T. (2016), *Religion and earnings: evidence from Germany*, „International Journal of Social Economics", vol. 43(8), s. 841–855.

Skrondal A., Rabe-Hesketh S. (2003), *Multilevel logistic regression for polytomous data and rankings*, „Psychometrika", vol. 68(2), s. 267–287.

Solomon P., Cox D. (1992), *Nonlinear Component of Variance Models*, „Biometrika", vol. 79(1), s. 1–11.

Spiezia V. (2011), *Are ICT Users More Innovative?*, „OECD Journal: Economic Studies", no. 1, s. 1–21.

Springer F. (2016), *Miasto Archipelag*, Wydawnictwo „Karakter", Kraków.

Sternberg R., Arndt O. (2001), *The Firm or the Region: What Determines the Innovation Behavior of European Firms?*, „Economic Geography", vol. 77(4), s. 364–382.

Strawiński P., Broniatowska P., Majchrowska A. (2016), *Returns to Vocational Education. Evidence from Poland*, University of Warsaw, Faculty of Economic Sciences Working Paper, no. 16/2016.

Strawiński P., Majchrowska A., Broniatowska P. (2016), *Wage Returns to Different Education Levels. Evidence from Poland*, „Ekonomista", nr 1, s. 25–49.

Szczygielski K., Grabowski W. (2014), *Innovation strategies and productivity in the Polish services sector*, „Post-Communist Economies", vol. 26(1), s. 17–38.

Szczygielski K., Grabowski W., Woodward R. (2017), *Innovation and the growth of service companies: the variety of firm activities and industry effects*, „Industry and Innovation", vol. 24(3), s. 249–262.

Szczygielski K., Grabowski W., Pamukcu M., Tandogan V. (2017), *Does government support for private innovation matter? Firm-level evidence from two catching-up countries*, „Research Policy", vol. 46(1), s. 219–237.

Świadek A., Szajt M. (2018), *Nakłady na działalność innowacyjną a produkcja przemysłowa w Polsce w latach 2006–2015 – zróżnicowanie regionalne*, „Prace Komisji Geografii Przemysłu Polskiego Towarzystwa Geograficznego", vol. 32(3), s. 54–68.

Tanner M.A. (1993), *Tools for Statistical Inference*, Springer, New York.

Tokarski T. (2013), *Zróżnicowanie podstawowych zmiennych makroekonomicznych w powiatach*, [w:] M. Trojak (red.), *Regionalne zróżnicowanie rozwoju ekonomicznego Polski*, Wydawnictwo Uniwersytetu Jagiellońskiego, Kraków.

Veall M., Zimmermann K. (1992), *Pseudo-R2's in the ordinal probit model*, „The Journal of Mathematical Sociology", vol. 16(4), s. 333–342.

Verardi V., Croux C. (2009), *Robust Regression in Stata*, „Stata Journal", vol. 9(3), s. 439–453.

Vuong Q.H. (1989), *Likelihood ratio tests for model selection and non-nested hypotheses*, „Econometrica", vol. 57(2), s. 307–333.

Wajda K. (1990), *Uwarunkowania polskiej myśli politycznej i społecznej na ziemiach polskich pod panowaniem pruskim 1864–1914*, [w:] S. Kalembka (red.), *Studia z dziejów polskiej myśli politycznej II – Polska myśl polityczna w dzielnicy pruskiej w XIX w.*, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

Wedderburn R.W.M. (1974), *Quasi-likelihood functions, generalized linear models and the Gauss--Newton method*, „Biometrika", vol. 61(3), s. 439–447.

Welfe A. (2009), *Ekonometria. Metody i ich zastosowanie*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Welfe A., Karp P. (2017), *Makroekonometryczny miesięczny model gospodarki Polski WM-1*, „Gospodarka Narodowa", nr 290(4), s. 5–38.

Welfe W., Welfe A. (2004), *Ekonometria stosowana*, Polskie Wydawnictwo Ekonomiczne, Warszawa.

Welfe A., Majsterek M., Florczak W. (1994), *Model pętli inflacyjnej w gospodarce polskiej – analiza kointegracyjna*, „Przegląd Statystyczny", nr 3, s. 245–264.

Wetherill G.B., Glazerbook K.D. (1986), *Sequential Methods in Statistics*, Chapman & Hall, London.

White H. (1980), *A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity*, „Econometrica", vol. 48(4), s. 817–838.

Wickham J. (2011), *Low Skill Manufacturing Work: from skill biased change to technological Regression/Niedrig qualifizierte Industriearbeit: vom qualifikationsbeeinflussten Strukturwandel zur technologischen Regression*, „Arbeit", vol. 20(3), s. 224–238.

Wielicki T., Arendt Ł. (2010), *A knowledge-driven shift in perception of ICT implementation barriers: Comparative study of US and European SMEs*, „Journal of Information Science", vol. 36(2), s. 162–174.

Winczorek J. (2015), *Przegląd literatury na temat dostępności i korzystania z pomocy prawnej*, INPRIS – Instytut Prawa i Społeczeństwa, Warszawa.

Wiśniewski J. (2015), *Microeconometrics in business management*, Wiley & Sons, London.

Wu F. (1999), *Intrametropolitan FDI firm location in Guangzhou, China*, „The Annals of Regional Science", vol. 33(4), s. 535–555.

Yohai V. (1987), *High Breakdown-point and High Efficiency Estimates for Regression*, „The Annals of Statistics", no. 15, s. 642–665.

Zarycki T. (2015), *The electoral geography of Poland: between stable spatial structures and their changing interpretations*, „Erdkunde", vol. 69(2), s. 107–124.

Żukowski R. (2004), *Historical path dependence, institutional persistence, and transition to market economy*, „International Journal of Social Economics", vol. 31(10), s. 955–973.

Abstract

Multilevel Models – The Use Of Regional Data in Microeconomic and Sociological Research

In economic and social research more and more attention is devoted to the analysis of the relationships observed at the individual level. For example, the level of access to data reflecting decisions of enterprises is increasing due to waves of the Community Innovation Survey. Information. Data concerning the economic activity of citizens is available due to Labour Force Surveys. Wages of individual workers are available from the Structure of Earnings by Occupations database. Apart from data obtained from cyclical researches, there are databases based on individual research, in which enterprises, workers and households are units. The data may be publicly available or commercially purchased by research institutes. As a result, the possibility of using microeconometric methods is still increasing.

In microeconometric and sociological research the role of context is ignored very often. It is assumed that there are relationships among variables and the location of an individual does not have any impact on decision-making process. Eventual differences among units from different regions or industries are treated as fixed. It means that the role of these differences is measured on the basis of dummy variables in the econometric model. However, there are methods of analysis of random differences between units from different industries or regions. These methods are used during the estimation of the parameters of multilevel models.

The book is aimed at providing knowledge concerning multilevel models and their applications in microeconomic and sociological research. The theory concerning the estimation of the parameters if multilevel models are presented. The broad range of multilevel models is presented. Methods of estimation of the parameters of these models are described in details. Applications of these models are presented in microeconomic as well as sociological research. Research devoted to the analysis of determinants of wages, innovativeness, and categories from the sociology of law are presented. In these empirical research, random effects associated with the location of individuals are taken into account. Moreover, variables available at the regional level are also used as explanatory ones.

The advantage of using multilevel models in comparison with standard ones is proved. Estimates of parameters from the multilevel model with regional variables are compared with the estimates obtained on the basis of models without random effects and without regional variables. The goodness of fit of the full model is compared with the goodness of fit of alternative, reduced models. In each case it is shown that multilevel models provide better predictions than reduced ones. Moreover, the results of testing of the presence of random effects indicate that multilevel models are better than standard ones.

The book consists of five chapters. In the first chapter, methods of estimation of parameters used in the case of individual data are shortly presented. At the beginning models assuming continuous dependent variable are presented and described. The classical linear regression model, methods of estimation in the case of heteroscedasticity of error term as well as robust methods of estimation are presented. In the next sub-chapters, methods of estimation and inference in the case of dichotomous, polychotomous are presented. Moreover, estimation and inference are presented in the case of count data models and rank regression. Cases of more than two dependent discrete variables are considered in sub-chapters 1.7 and 1.9–1.11. Methods of estimation of the parameters for biprobit multivariate probit and endogenous probit model are presented.

In the second chapter regional data used in economic and sociological research are presented in detail. Firstly, the administrative, statistical and historical division of Poland is described in sub-chapter 2.1. In the next sub-chapter, historical and cultural differences among Polish regions are presented. In subchapter 2.3 sources of data used in regional analyses for Poland are described. Apart from the local data bank, which consists of many socioeconomic variables observable at different levels of the NUTS classification, regional innovation scoreboard is presented as an important source of data about innovativeness of regions. Moreover, other sources of regional data are presented in the sub-chapter 2.3.

Chapter 3 is devoted to standard linear multilevel models. Equation presenting the econometric model based on individual data with regional and industry variables is presented. Next standard notion for the linear multilevel model is provided. Estimation of the parameters of this model and prediction of random effects is presented in sub-chapter 3.3 and 3.4. In sub-chapter 3.5 linear multilevel model is applied for the analysis of factors influencing wages in Polish enterprises. Apart from individual variables, regional categories as well as random effects associated with membership of workers in specific regions, industries, professional groups, are taken into account. It turns out that using a linear multilevel model with cross-effects is justified since there are significant random differences among individuals

from different groups. The econometric model including random effects and regional variables outperforms competitive ones.

The fourth chapter of this book is devoted to generalized linear multilevel models. In the sub-chapter 4.1 and 4.2 basic equation for this model is given and the likelihood function is presented. Two popular methods of estimation of the parameters of multilevel models are presented in sub-chapters 4.3 and 4.4. In the sub-chapter 4.3, methods of estimation using approximations are described. In the sub-chapter 4.4, simulating methods of estimation of parameters of generalized linear multilevel models are presented. In the sub-chapter 4.5, proposals of the methods of estimation of the parameters of the multilevel multivariate probit models as well as multilevel Heckman models are provided. Multilevel discrete choice model is used in order to identify the relationships between using ICT, innovativeness, and productivity in Polish enterprises. Moreover, random effects associated with the location of enterprises as well as categories from the regional innovation scoreboard are used in the final specification. Results of the estimates indicate that using information and communication technologies in advanced economic processes stimulates innovativeness and productivity in Polish enterprises. However using ICT complementarities and introducing organizational change turned out to have a moderating impact on the relationship between using ICT, innovativeness, and productivity. Moreover, it turned out that there exist region-specific differences in these relationships. Enterprises located in the same region cooperate, change opinions, use the same sources of information. Therefore companies located in a similar part of the country have more similar attitudes than firms from different parts of Poland. It turned out that random effects are significant. When categories from the Regional Innovation Scoreboard were included in the final specification, the goodness of fit was higher. The results indicate that when the relationship between using ICT, innovativeness, and productivity for the Polish economy is analyzed, contextual factors should be taken into account.

The fifth chapter is devoted to the analysis of multilevel multinomial logit models as well as multinomial rank regression models. Notions for these models are presented in the sub-chapter 5.2. In the sub-chapter 5.3 empirical investigation devoted to the choice of the method of reaction to a legal problem is presented. The multilevel multinomial logit model is applied. Moreover, the Heckit-type estimator is proposed. Results of the empirical investigation indicate that apart from individual socioeconomic variables, contextual factors affect the probability of experiencing legal problems and a method of reaction to them. It turns out that membership of an individual to the historical region has a very strong impact on his/her decision in the case of legal problems.

Spis rysunków

Rysunek 1. Rodzaje obserwacji nietypowych w modelu regresji 38

Rysunek 2. Struktura hierarchiczna związana z podziałem działalności wykonywanych przez firmy na sekcje 119

Rysunek 3. Struktura hierarchiczna związana z województwami, w których zlokalizowane są firmy 121

Rysunek 4. Hierarchiczna struktura danych ze względu na przynależność pracowników do grup zawodowych oraz ze względu na przynależność tych grup do grup ze względu na poziom kwalifikacji 122

Rysunek 5. Hierarchiczna struktura danych ze względu na przynależność pracowników do grup zawodowych oraz grup ze względu na przynależność tych grup do grup zadaniowych 122

Rysunek 6. Rozszerzony (o wykorzystanie TIiK) model CDM 175

Spis tabel

Tabela 1. Rozkłady składnika losowego dla różnych modeli dwumianowych ... 41
Tabela 2. Efekty krańcowe dla różnych rozkładów składnika losowego ... 43
Tabela 3. Tablica trafności predykcji ... 46
Tabela 4. Województwa w Polsce i ich stolice ... 74
Tabela 5. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część pierwsza ... 85
Tabela 6. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część druga ... 86
Tabela 7. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część trzecia ... 87
Tabela 8. Kategorie i grupy zmiennych dostępnych w Banku Danych Lokalnych. Część czwarta ... 88
Tabela 9. Zmienne wykorzystywane do mierzenia innowacyjności na poziomie regionalnym dostępne w Regional Innovation Scoreboard 2017 ... 89
Tabela 10. Kody zawodów jednocyfrowych ... 116
Tabela 11. Przyporządkowanie trzycyfrowych grup zawodów do grup zadaniowych ... 117
Tabela 12. Zmienne objaśniające wpływające na wysokość wynagrodzeń pracowników ... 119
Tabela 13. Oszacowania parametrów modelu wielopoziomowego wyjaśniającego kształtowanie się wynagrodzeń w polskiej gospodarce ... 124
Tabela 14. Weryfikacja hipotez o równości „stóp zwrotu" dla osób o różnych poziomach wykształcenia ... 124
Tabela 15. Efekty losowe dla województw w poszczególnych latach. Odchylenia standardowe zawarte są w nawiasach ... 126
Tabela 16. Efekty losowe dla sekcji PKD w poszczególnych latach ... 127
Tabela 17. Średnie efekty losowe dla grup ze względu na poziom kwalifikacji w poszczególnych latach (w nawiasach podane są odchylenia standardowe) ... 130
Tabela 18. Średnie efekty losowe dla trzycyfrowych grup zawodowych ... 132
Tabela 19. Predykcje efektów losowych dla grup zadaniowych w poszczególnych latach (w nawiasach podane są odchylenia standardowe) ... 135
Tabela 20. Średnie błędy szacunku dla czterech porównywanych modeli ... 137
Tabela 21. Rodzaje wielopoziomowych uogólnionych modeli liniowych ... 139
Tabela 22. Definicje zmiennych binarnych związanych z wykorzystanie technologii informatycznych i telekomunikacyjnych ... 178
Tabela 23. Empiryczny rozkład dla zmiennych binarnych zdefiniowanych w tabeli 22 ... 179
Tabela 24. Empiryczny rozkład zmiennej ilustrującej fakt dokonania inwestycji w rozwój TIiK przez firmy ... 180

Tabela 25. Empiryczny rozkład zmiennej wskazującej, czy firma posiada wyodrębniony wydział B+R ... 181
Tabela 26. Empiryczny rozkład zmiennych binarnych związanych z wprowadzeniem określonych typów innowacji ... 182
Tabela 27. Podstawowe statystyki opisowe dla zmiennej *PRODUKTYWNOSC* ... 183
Tabela 28. Zmienne egzogeniczne rozważane w badaniu empirycznym ... 184
Tabela 29. Zmienne ilustrujące innowacyjność otoczenia rozważane w badaniu empirycznym ... 186
Tabela 30. Wyniki estymacji parametrów wielopoziomowego, wielorównaniowego modelu probitowego (w nawiasach podano średnie błędy szacunku) ... 191
Tabela 31. Średnie efekty losowe dla poszczególnych województw ... 196
Tabela 32. Wartości oczekiwane oraz odchylenia standardowe efektów losowych dla poszczególnych województw ... 197
Tabela 33. Oszacowania parametrów wielopoziomowego, wielorównaniowego modelu probitowego wyjaśniającego skłonność do wprowadzania innowacji ... 199
Tabela 34. Średnie efekty losowe dla poszczególnych województw ... 203
Tabela 35. Oszacowania parametrów modelu wyjaśniającego produktywność w polskich przedsiębiorstwach wykorzystujących technologie informatyczne i telekomunikacyjne[a)] ... 205
Tabela 36. Predykcje efektów losowych dla liniowego modelu wielopoziomowego ... 206
Tabela 37. Oszacowania efektów netto wskazujących na wpływ egzogenicznych zmiennych binarnych na prawdopodobieństwo, iż dana zmienna zależna przyjmuje wartość 1 ... 210
Tabela 38. Spadek wartości predykcyjnej modelu po usunięciu określonych grup zmiennych ... 211
Tabela 39. Potencjalne determinanty sposobu reakcji wobec wystąpienia problemu prawnego ... 222
Tabela 40. Wyniki estymacji parametrów wielopoziomowego modelu logitowego wyjaśniającego skłonność do doświadczania problemu prawnego ... 226
Tabela 41. Oszacowania efektów losowych dla poszczególnych województw ... 229
Tabela 42. Oszacowania efektów losowych dla poszczególnych powiatów ... 230
Tabela 43. Oszacowania parametrów dla efektów stałych wielopoziomowego, wielomianowego modelu logitowego ... 233
Tabela 44. Oszacowania efektów losowych dla poszczególnych regionów historycznych ... 236

Od Redakcji

Wojciech Grabowski ukończył w 2005 roku jednolite studia magisterskie w Szkole Głównej Handlowej w Warszawie na kierunku metody ilościowe i systemy informacyjne. W tym samym roku rozpoczął pracę na stanowisku asystenta na Wydziale Ekonomiczno-Socjologicznym Uniwersytetu Łódzkiego w Katedrze Modeli i Prognoz Ekonometrycznych. Stopień doktora uzyskał w 2011 roku. Jego rozprawa doktorska, przygotowana pod kierunkiem naukowym prof. Aleksandra Welfe, Członka Korespondenta Polskiej Akademii Nauk, była poświęcona testowaniu stacjonarności i analizie kointegracji w modelach zmiennych wielomianowych. Szczególny nacisk położony został na zagadnienia testowania stopnia zintegrowania zmiennych nieobserwowalnych związanych ze zmiennymi dwumianowymi i wielomianowymi kategorii uporządkowanych oraz testowania restrykcji w modelach dwumianowych i wielomianowych kategorii uporządkowanych zawierających niestacjonarne regresory. Od 2012 roku jest zatrudniony na Uniwersytecie Łódzkim jako adiunkt w Katedrze Modeli i Prognoz Ekonometrycznych. Jest również zatrudniony na stanowisku profesora wizytującego w Katedrze Ekonomii Matematycznej na Uniwersytecie Ekonomicznym we Wrocławiu.

Jego zainteresowania naukowe koncentrują się na zastosowaniach metod mikroekonometrycznych w badaniach ekonomicznych i socjologicznych. Efekty pracy autora zostały opublikowane w języku polskim oraz angielskim w ponad 50 publikacjach obejmujących: monografie, rozdziały w monografiach, artykuły w czasopismach, między innymi takich jak: „Research Policy", „Industry and Innovation", „Economic Modelling", „Emerging Markets Finance and Trade", „Economics of Transition", „Applied Economics Letters", „Prague Economic Papers", „Czech Journal of Economics and Finance", „Post-Communist Economies", „Eastern European Economics".

Wojciech Grabowski jest kierownikiem grantu Narodowego Centrum Nauki pt. „Podatność rynków giełdowych krajów grupy wyszehradzkiej na niestabilności zewnętrzne i wewnętrzne" (2015/19/D/HS4/03354). Realizowany projekt obejmuje zagadnienia związane z wrażliwością giełd krajów Europy Środkowo-Wschodniej na zachowania innych rynków giełdowych w okresach stabilności oraz kryzysu, a także szacowanie indeksów transmisji zmienności między rynkami. Wojciech Grabowski uczestniczy i uczestniczył także w innych grantach Narodowego Centrum Nauki oraz grantach krajowych i międzynarodowych jako wykonawca, m.in.

w grancie międzynarodowym pt. „SMEthod – segmenting SMEs for better innovation support" finansowanym przez HORYZONT 2020.

W 2017 roku Wojciech Grabowski uzyskał nagrodę Rektora Uniwersytetu Łódzkiego drugiego stopnia za cykl publikacji pt. *Zastosowanie metod mikroekonometrycznych w badaniach mikroekonomicznych, makroekonomicznych i socjologicznych*. W 2012 roku otrzymał nagrodę im. Profesora Władysława Welfe za najlepsze wystąpienie na Warsztatach Doktorskich z Zakresu Statystyki i Ekonometrii.

Jest także cenionym dydaktykiem. W 2015 roku przyznano mu nagrodę Najlepszego Wykładowcy Instytutu Ekonometrii. Regularnie wykłada na Wydziale Ekonomiczno-Socjologicznym Uniwersytetu Łódzkiego przedmioty z zakresu teorii ekonometrii i jej zastosowań oraz promuje studentów w ramach prac licencjackich.