UNIWERSYTET EKONOMICZNY W POZNANIU

Wydział Informatyki
i Gospodarki Elektronicznej

Magdalena Szczepańska

Zastosowanie uogólnionego modelu selekcji Heckman'a do opisu przedsiębiorstw gospodarki nieformalnej

Usage of generalized Heckman's sample selection model to describe companies in the informal economy

Praca licencjacka

Promotor: dr Maciej Beręsewicz

Data złożenia:

Podpis promotora

Kierunek: Informatyka i ekonometria

Specjalność: Analityka gospodarcza

Poznań 2019

Spis treści

Wstęp **2**

1 Szara strefa i jej pomiar **3**

1.1 Definicje szarej strefy . . . 3

1.1.1 Perspektywy szarej strefy . . . 3

1.1.2 Pojęcie gospodarki nieformalnej w literaturze . . . 4

1.1.3 Statystyka publiczna . . . 6

1.1.4 Państwowa Inspekcja Pracy . . . 7

1.2 Powody istnienia gospodarki ukrytej . . . 8

1.3 Metody pomiaru szarej strefy . . . 10

1.4 Szara strefa w Polsce w świetle dostępnych źródeł danych . . . 12

1.4.1 Badanie Aktywności Ekonomicznej Ludności . . . 12

1.4.2 Badanie Kapitału Ludzkiego . . . 13

1.4.3 Państwowa Inspekcja Pracy . . . 15

1.5 Podsumowanie . . . 17

2 Uogólniony model doboru próby Heckman'a **18**

2.1 Problematyka prób nielosowych . . . 18

2.1.1 Losowy i nielosowy dobór próby . . . 18

2.1.2 Selekcja w próbie . . . 20

2.2 Model doboru próby Heckman'a . . . 20

2.2.1 Regresja dla zmiennych ograniczonych . . . 20

2.2.2 Klasyczny model Heckman'a . . . 22

2.2.3 Uogólniony model Heckman'a . . . 25

2.2.4 Model Heckman'a dla rozkładu Poisson'a . . . 25

2.2.5 Model selekcji dla rozkładu ujemnego dwumianowego 26
2.3 Wybrane implementacje w pakiecie R 28
2.3.1 Pakiet sampleSelection 28
2.3.2 Pakiet SemiParSampleSel 29
2.3.3 Podsumowanie 29

3 Model opisujący liczbę nielegalnie zatrudnionych 30
3.1 Dane Państwowej Inspekcji Pracy 30
3.2 Eksploracyjna analiza danych 31
3.2.1 Badanie rozkładu próby 31
3.2.2 Model bez uwzględnienia procesu selekcji 34
3.3 Estymacja modelu Heckman'a 37
3.3.1 Wyniki badania 37

Podsumowanie 44

Bibliografia 47

Spis tabel 48

Spis rysunków 49

A Spis programów 50
A.1 Skrypty wykorzystane do przygotowania danych 50
A.1.1 Skrypty wykorzystane do estymacji modeli 51

Spis programów 53

Wstęp

Rynek pracy zawsze był obszarem, w którym możemy zbadać wiele zależności oraz cech pracowników bądź pracodawców. Szczególnie gospodarka nieformalna daje możliwość wykorzystania wielu narzędzi statystycznych do m.in. oszacowania rozmiaru tego zjawiska. Oprócz określenia rozmiaru szarej strefy możemy również spróbować ustalić jakie cechy posiadają osoby, które pracują ”na czarno” bądź też przedsiębiorcy, którzy takich zatrudnień dokonują.

Praca ma na celu zbadanie przedsiębiorstw działających w gospodarce nieformalnej. Podmiotem zainteresowania jest określenie cech jakimi wyróżniają się firmy zatrudniające nielegalnie. Weryfikowane zmienne w badaniu to sekcje PKD oraz wielkość zatrudnienia badanych podmiotów. Dodatkowo wykonane będzie porównanie uzyskanych wyników z 2010 oraz 2016 roku.

Na potrzeby pracy licencjackiej wykorzystano dane pozyskane od Państwowej Inspekcji Pracy dotyczące skontrolowanych podmiotów oraz wyników tej kontroli. Dane zostały udostępnione na potrzeby prac badawczych dra Macieja Beręsewicza (Uniwersytet Ekonomiczny w Poznaniu) oraz dr Dagmary Nikulin (Politechnika Gdańska). Niniejsza praca jest jednym z etapów tego badania.

W pierwszym rozdziale odniesiono się do samego pojęcia szarej strefy, określając różne perspektywy, poznając definicje czy też metody pomiaru. Dodatkowo znalazły się też szacowane powody istnienia tego zjawiska. W ostatniej rozdziału pierwszego części przedstawione zostały badania, które wykonywane są po to by oceniać zachodzące zmiany na rynku pracy czy też poznawać kolejne zależności.

Drugi rozdział przybliża nam model jaki używany jest do badania różnych cech określonych zjawisk. Pierwsza część definiuje losowy i nielosowy dobór próby, który ma znaczenie w badaniach statystycznych. Pokazane też zostało zjawisko selekcji, czyli pojęcia związanego z nielosowymi obserwacjami w badaniach. Jako wstęp dla modelu selekcji opisano pojęcie regresji dla zmiennych ograniczonych. Następnie scharakteryzowano klasyczny i uogólniony model

Heckman'a. Na końcu tego rozdziału znajduje się opis przydatnych funkcji w języku R, które umożliwią wykonanie analizy oraz zapis implementacji modelu Heckman'a.

Ostania część pracy zawiera przede wszystkim wyniki uzyskane dla modeli zarówno z selekcją jak i bez w porównaniu dla 2010 i 2016 roku. Dodatkowo przybliżone zostaną dane wykorzystane w procesie estymacji oraz całkowity zapis implementacji w R.

Rozdział 1

Szara strefa i jej pomiar

1.1 Definicje szarej strefy

1.1.1 Perspektywy szarej strefy

Szara strefa to pojęcie stosowane, aby określić zjawisko działalności niezarejestrowanej, czyli takiej która jest odbywa się z wyłączeniem kontroli państwa. W literaturze jest to jedno z wielu występujących i używanych sformułowań opisujących tą sferę gospodarki (Jarocka, 2011).

Jednym z wielu określeń są "gospodarka nieformalna", "gospodarka drugiego obiegu" czy też "gospodarka ukryta". Obszerna terminologia to w większości zasługa organizacji międzynarodowych, które nieustannie od lat 80-tych XX wieku badają i szacują rozmiary zjawiska szarej strefy nadając przy tym nowe określenia. Mnogość terminów wynika również z perspektywy z jakiej badana jest gospodarka nieoficjalna. Wyróżnia się perspektywy: ekonomiczne, antropologiczne oraz socjologiczno-polityczne (Jarocka, 2011).

Gospodarka nieformalna w znaczeniu perspektywy ekonomicznej odnosi się do części produktu krajowego brutto, która dotyczy niezarejestrowanych towarów i usług szarej strefy. Błędne statystyki przyczyniają się do niepoprawnego odzwierciedlenia rzeczywistości, a co za tym idzie źle prowadzonej polityki społecznej i gospodarczej. Samo istnienie gospodarki ukrytej pokazuje, że oficjalnie funkcjonujący system gospodarczy może nie być odpowiednio przystosowany dla społeczeństwa i przy wzroście szarej strefy zmiany powinny być konieczne. Odnosząc się do polityki finansowej państwa, tak znacząca część gospodarki, od której nie pobierane są podatki stanowi obciążenie dla budżetu państwa (Jarocka, 2011).

Jeżeli chodzi o perspektywę antropologiczną dotyczy ona roli zawodowych wykonywanych

przez jednostki. Wszelkie przyczynianie się do istnienia szarej strefy w tym ujęciu jest traktowane jako zachowanie patologiczne. Szukając powodów takich zachowań badacze odnoszą się zarówno do przyczyn ekonomicznych jak i społecznych. Spojrzenie to w pewnym stopniu jest traktowane jako połączenie perspektywy ekonomicznej i społecznej (Jarocka, 2011).

Ostania perspektywa nawiązuje do zbyt restrykcyjnych obciążeń podatkowych i rozumiana jest jako konflikt pomiędzy strefą publiczną, a prywatną. Mnogość perspektyw przyczynia się do powstania wielu definicji pojęcia pracy nierejestrowanej (Jarocka, 2011).

1.1.2 Pojęcie gospodarki nieformalnej w literaturze

W literaturze nie ustalono jednej, poprawnej definicji szarej strefy. W zależności od tego pod jakim kątem badamy gospodarkę istnieje wiele różnych tłumaczeń tego zjawiska gospodarczego. Profesor Edgar L. Feige (Feige, 2007) uważa, że wszelkie działania związane z gospodarką i dochodami, które są dokonywane poza kontrolą rządu wliczają się w szarą strefę. Organizacja Współpracy Gospodarczej i Rozwoju (Organisation for Economic Co-operation and Development, 2001) uważa, że legalna działalność, która jest niezgłoszona władzom publicznym to właśnie ukryta strefa. Kolejną z wielu prób wyjaśnienia gospodarki równoległej jest następujący opis:

Według Europejskiego Systemu Rachunków Narodowych i Regionalnych (ESA 2010) gospodarka nieformalna jest złożona z trzech elementów (Łapiński, Peterlik & Wyżnikiewicz, 2014):

1. działalności nielegalnej, w przypadku której obie strony są dobrowolnymi partnerami transakcji gospodarczej,

2. działalności ukrytej, w przypadku której transakcje same w sobie nie są sprzeczne z prawem, ale nie są zgłaszane w celu uniknięcia urzędowej kontroli,

3. działalności określonej jako „nieformalna", zazwyczaj w sytuacji, gdy nie prowadzi się żadnych rejestrów.

Na pierwszy element składa się działanie które jest niezgodne z prawem, rozumiane już jako wykroczenie lub nawet przestępstwo. Typowymi przykładami są przemyt używek takich jak alkohol bądź papierosy, czy też poczynania takie jak produkcja leków i substancji, które nie są objęte nadzorem państwa. Oprócz wcześniej wymienionych przykładów bardzo często

zarejestrowani przedsiębiorcy biorą udział w operacjach, które są powiązane z nielegalnie działającymi podmiotami gospodarczymi (Pater, 2007).

Drugi element to przede wszystkim zatajane niektórych elementów prowadzenia przedsiębiorstwa przez legalnie działające podmioty gospodarcze. Głównym problemy z jakimi idzie się zmierzyć to pomniejszanie faktycznej liczby obrotów. Najczęściej to mikro- i małe firmy dopuszczają się takiej praktyki, co tłumaczą że ich potencjał gospodarczy nie jest wystarczający w stosunku do warunków prowadzenia biznesu (Pater, 2007).

Ostatni, trzeci element dotyczy aktywności gospodarczej często klasyfikowanej jako dorywcza bądź sezonowa. Pojedyncze podmioty, najczęściej osoby fizyczne, wykonują odpłatnie określone usługi. Pracownicy pracujący w tej części działalności określani są jako pracujący na czarno (Pater, 2007). Zazwyczaj są to:

1. bezrobotni (najczęściej osoby o zbyt niskich kwalifikacjach aby znaleźć zatrudnienie),
2. pracownicy pracujący legalnie, lecz dorabiający sobie w szarej strefie,
3. emeryci oraz renciści,
4. imigranci zarobkowi,
5. studenci i uczniowie.

Według badań Instytutu Pracy i Spraw Socjalnych (Pater, 2007) znacznie częściej praca w szarej strefie jest podejmowana regularnie. Najpopularniejsze były prace domowe, opieka nad dziećmi, usługi gastronomiczne, korepetycje czy prace budowlane. Głównymi pracodawcami były gospodarstwa domowe, rolne i ogrodnicze, dopiero na końcu firmy prywatne (Pater, 2007).

Powody pozyskiwania dochodów z nieformalnego źródła tak ja w przypadku innych badań były podobne. Typowo wymieniano brak wystarczającej ilości środków do życia, zbyt wysokie podatki, problemy ze znalezieniem pracy w sektorze gospodarki legalnej, wygodny godzinowo i zmianowy charakter pracy, zbyt niskie wykształcenie i niepokój związany z straceniem korzyści socjalnych w przypadku podjęcia legalnej pracy. W tabeli (1.1) ukazane są najpopularniejsze obszary gdzie szara strefa jest największa (Pater, 2007).

Tabela 1.1. Szacunki udziałów szarej gospodarki w tworzeniu PKB w latach 2008–2011 (w proc.) według sekcji PKD 2007

Sekcja PKD 2007	2008	2009	2010	2011
Przemysł	1,2	1,3	1,3	1,1
Budownictwo	2,1	2,2	2,2	2,4
Handel i naprawa pojazdów samochodowych, zakwaterowanie i gastronomia	5,1	5,7	5,8	5,7
Transport i gospodarka magazynowa	0,6	1,0	0,8	0,6
Obsługa rynku nieruchomości	1,1	1,3	1,1	1,3
Pozostałe sekcje	1,7	1,6	1,6	1,5
Szara gospodarka razem	11,8	13,1	12,8	12,6
PKB razem z szarą gospodarką	100,0	100,0	100,0	100,0

Źródło: Rachunki narodowe według sektorów i podsektorów instytucjonalnych 2008 – 2011, Główny Urząd Statystyczny, Warszawa, 2013. (Główny Urząd Statystyczny, 2013, s. 372)

1.1.3 Statystyka publiczna

Spośród zdefiniowanych pojęć Główny Urząd Statystyczny określa pracę nieformalną jako pracę wykonywaną bez „nawiązania stosunku pracy” (Główny Urząd Statystyczny, 2004). Oznacza to, że brak jakiegokolwiek rodzaju umowy, nieważne czy jest to praca wykonywana w gospodarstwach rolnych czy też firmach, zaliczana jest do pracy nierejestrowanej.

Badania Głównego Urzędu Statystycznego (GUS) określają nie tylko rozmiar pracy nieformalnej, ale również poszczególne udziały dla konkretnych sekcji PKD (1.1). znaczna większość gospodarki nieformalnej składa się z sektora usługowego, jedynym wyjątkiem jest przemysł. Prawie połowa zbadanego obszaru dotyczy handlu, zakwaterowania i gastronomii i stanowiła 5,7 % Produktu Krajowego Brutto (PKB). Wartość dodana, która została wytworzona głównie została przeznaczona na spożycie gospodarstw domowych oraz nakłady brutto na środki trwałe (Główny Urząd Statystyczny, 2004).

Jeśli chodzi o europejski urząd statystyczny (Eurostat) prace nieformalną definiuje w następujący sposób (Vanderseypen, Tchipeva, Peschner, Rennoy & Williams, 2013):

- Gospodarka nieobserwowana to obszar gospodarki obejmujący grupę działalności ekonomicznych, dla których istnieje największe prawdopodobieństwo, że będą nieobserwowane. Są to: 1) działalność w szarej strefie, 2) działalność nielegalna, 3) działalność sektora nieformalnego lub działalność gospodarstw domowych na własny użytek. Działalność może być także pominięta z powodu braków w sposobie zbierania statystycznych

danych podstawowych.

- Szara gospodarka obejmuje działania produkcyjne w sensie ekonomicznym, całkowicie legalne (pod względem spełniania norm i regulacji prawnych), ale ukrywane przed władzami publicznymi z następujących przyczyn: 1) aby uniknąć płacenia podatku dochodowego, podatku od wartości dodanej (VAT) i pozostałych podatków, 2) aby uniknąć płacenia składek na ubezpieczenie społeczne, 3) aby uniknąć stosowania wymogów prawa takich jak: płaca minimalna, maksymalny czas pracy, warunki bezpieczeństwa pracy, 4) aby uniknąć procedur administracyjnych takich jak wypełnianie kwestionariuszy statystycznych i innych formularzy.

- Działalność nielegalna obejmuje: 1) produkcję wyrobów i usług, których sprzedaż, rozprowadzanie lub posiadanie są zabronione przez prawo; 2) działalność produkcyjną, która jest zwykle legalna, lecz staje się nielegalna gdy jest wykonywana przez producentów nie mających do tego prawa, na przykład praktyka medyczna bez licencji.

W pierwszej części definicji mamy rozgraniczenie na działalność w szarej strefie i działalność nielegalną. Czasami te pojęcia występują zamiennie, jednak są pewne różnice w ich znaczeniu. Jeśli chodzi o postępowanie nielegalne to odnosi się ono do nie przestrzegania norm, bezpieczeństwa czy ochrony zdrowia w miejscu pracy. Natomiast działalność w szarej strefie odnosi się do działalności kryminalnej, czyli łamania przepisów administracyjnych (Vanderseypen i in., 2013).

1.1.4 Państwowa Inspekcja Pracy

Państwowa Inspekcja Pracy (PIP) rozróżnia działalność ukrytą bądź częściowo ukrytą (PIP, 2016). W obecnej ustawie Państwowej Inspekcji Pracy nie ma konkretnych definicji pojęć takich jak legalne zatrudnienie (Drabek, 2012). Jednak na podstawie kontroli widoczne są pewne praktyki wśród przedsiębiorców, najczęściej dotyczą one księgowości, gdzie zgłaszane do opodatkowania są wybrane zlecenia. Zarejestrowana i częściowo legalnie prowadzona firma jest już częścią gospodarki nieformalnej. Na podstawie dokonywanych kontroli można określić jakie działania definiują szarą strefę. Oprócz zafałszowań w księgowości problematyczne są również zatrudnienia gdzie nie występuje umowa lub pracownik nie został zgłoszony do ubezpieczenia społecznego. Możliwe jest też nie opłacanie składek społecznych. Innym przykładem jest praca

zadeklarowana fałszywie, czyli próba unikania podatków i składek poprzez zaniżanie zobowiązań wobec państwa. Według Państwowej Inspekcji Pracy brak pisemnej umowy o pracę czy też potwierdzenie zawartej umowy po za terminem jest najbardziej szkodliwe dla pracowników (PIP, 2016).

1.2 Powody istnienia gospodarki ukrytej

Nie tylko rozmiar i obszar obejmujący gospodarkę nieformalną jest problematyczny do zbadania. Oprócz istnienia problemu ważne jest to by poznać jego podłoże, dlatego istnieje szereg badań, które oprócz samego zjawiska badają również przyczyny gospodarki ukrytej.

Głównymi motywami, które są winne istnieniu zjawiska szarej strefy to mała ilość lub całkowity brak ofert pracy w konkretnym sektorze oraz próba unikania płacenia podatków (Kubiczek, 2010).

Często osoby z brakiem kwalifikacji zawodowych czy słabym wykształceniem mają trudności aby znaleźć formalne zatrudnienie, tak więc szukają pracy w innych sektorach, które są dostępne w obszarze ukrytej gospodarki. Możliwe, że mimo chęci pracowania po "dobrej" stronie, brak innych możliwości zmusza takie osoby do przyjęcia jakiegokolwiek stanowiska, które daje poczucie finansowej stabilności (Fundowicz, Łapiński & Wyżnikiewicz, 2018).

Popularnie nazywana praca "na czarno" ma szereg wad. Często takie zatrudnienie oznacza gorsze zarobki, brak domagania się jakichkolwiek praw pracownika oraz brak ubezpieczenia. Natomiast patrząc na status osoby bezrobotnej, a tak naprawdę pracującej w szarej strefie można zrozumieć pewne korzyści jakie dana jednostka uzyskuje takim postępowaniem. Widniejąc w rejestrze w urzędzie pracy jako bezrobotny ma możliwość korzystania z bezpłatnej opieki zdrowotnej, która dotyczy również rodziny konkretnej osoby (Fundowicz i in., 2018).

Natomiast odnosząc się do przedsiębiorstw próbujących działać w szarej strefie, często podejście jest własnowolne. Doświadczenie pokazuje, że wśród najczęstszych powodów takiego działania są próby unikania zbyt wysokich podatków czy też obciążeń na rzecz ZUS. Zaoszczędzone pieniądze przyczyniają się do uzyskania wyższych dochodów niż byłoby to możliwe przy legalnie prowadzonej działalności. Kolejne argumenty odnoszą się do rynku gdzie dany podmiot funkcjonuje. Przedsiębiorcy starają się zminimalizować wszelkie obciążenia finansowe, tak by zdołać zrównać się z konkurencją (Fundowicz i in., 2018).

Takie postępowanie prowadzi do znacznego obciążenia równocześnie gospodarki oraz spo-

łeczeństwa. Duża liczba zarejestrowanych bezrobotnych wymaga zaplecza finansowego, które jest zapewiane jednostkom pozostającym bez pracy. Koszty ubezpieczenia zdrowotnego czy przede wszystkim nienależące się zasiłki dla bezrobotnych pracujących w nielegalny sposób utrudniają wzmacnianie dochodu narodowego. Zmniejszający się budżet państwa wymusza zwiększenie stawek podatkowych oraz obniżenie jakości dóbr publicznych. Co więcej dalsze konsekwencje związane z wydatkami publicznymi wywierają wpływ na rozwój przedsiębiorczości. Dochody ludności maleją i zjawisko fiskalizmu przyczynia się do wzrostu problemu stymulacji wśród sektora prywatnego. Wraz ze wzrostem redystrybucji dochodowej wielkość szarej strefy się zwiększa, taki mechanizm pokazuje absurd tego zjawiska Fundowicz i in. (2018).

Tabela 1.2. Opinie Polaków na temat przyczyn podejmowania pracy nierejestrowanej w 2010r. (w %)

Przyczyna	Ogółem	Kobiety	Mężczyźni
Brak możliwości znalezienia pracy	53,1	54,8	51,3
Niewystarczające dochody	44,7	43,7	43,9
Pracodawca oferuje wyższe wynagrodzenie bez rejestrowania umowy o pracę	24,0	21,9	26,4
Wysoka składka ubezpieczeniowa	17,5	15,5	19,7
Wysokie podatki zniechęcające do rejestrowania dochodów	14,8	13,1	16,6
Możliwość utraty niektórych świadczeń przy podjęciu pracy niezarejestrowanej	7,2	7,8	6,6
Sytuacja rodzinna lub życiowa	6,4	7,9	4,7
Niechęć wiązania się na stałe z miejscem pracy	1,1	1,0	1,1

Źródło: Konsumenci i gospodarstwa domowe na nieformalnym rynku pracy w Polsce. Konsumpcja i rozwój, Mróz, B. (2012). (Mróz, 2012, s. 28) .

W 2010 roku zbadano przyczyny, dla których respondenci podejmowali pracę w szarej strefie (1.2). Najczęściej podawanym powodem był brak możliwości znalezienia pracy, zarówno wśród kobiet jak i mężczyzn. Związanie na stałe z miejscem pracy okazuje się ważne, ponieważ była to najrzadziej określana przyczyna nielegalnego zatrudnienia. Często też znacząca okazała się wysokość dochodów. Nielegalna praca pozwala uzyskać je na wyższym poziomie (Mróz, 2012).

Istnieje również teza mówiąca o pewnych pozytywach związanych z gospodarką nieformalną. Nawiązuje ona do zmian jakie mogą zajść w instytucjach publicznych czy też na samym rynku pracy. Jednak należy mieć na uwadze, że powiększanie się szarej strefy nie jest bez znaczenia dla PKB i pewne z pozoru zyski nie są w stanie zrekompensować negatywnych konsekwencji (Gołębiowski, 2007) .

Samo zjawisko gospodarki ukrytej ma pewne odwzorowanie na zachowanie społeczeństwa.

Powoduje nie tylko spadek moralności podatkowej, ale ogólne poczucie wśród obywateli na przyzwolenie do czynności, które nie są zgodne z prawem. Problematyczne jest z pewnością to iż rozmiar konsekwencji nie jest możliwie do określenia (Gołębiowski, 2007).

1.3 Metody pomiaru szarej strefy

Mimo charakteru gospodarki nieformalnej ekonomiści nieustannie starają się oszacować rozmiar szarej strefy. Momentem przełomowym okazała się druga połowa lat siedemdziesiątych. Wtedy została opublikowana praca Gutmanna (Gutmann, 1977), która zapoczątkowała szereg badań nad tematem gospodarki nieformalnej. Obecnie weryfikowaniem tej strefy gospodarki zajmują się urzędy statystyczne, ale również ośrodki naukowe, które często wykazują, że rozmiar nielegalnej sfery jest większy niż donoszą urzędy. Różnica wynika z charakteru obu środowisk. Urzędy bazują na potwierdzonych danych z pewnych źródeł, natomiast ośrodki badawcze same oceniają i wybierają dane (Jarocka, 2011).

Oficjalnie istnieje rozgraniczenie na metody bezpośrednie jak i pośrednie w sposobie kolekcjonowania informacji. Metoda bezpośrednia nawiązuje do pozyskiwania informacji u samego źródła, czyli podmiotów, które są zatrudnione w szarej strefie. Polega na przeprowadzaniu ankiet, analizie deklaracji podatkowych, obserwacji podmiotów czy też wywiadzie. Natomiast metoda pośrednia, zwana inaczej wskaźnikową, skupia się na wielkościach i wskaźnikach ekonomicznych i próbuje znaleźć pewne zależności wskazujące na nielegalne postępowanie w badanym obszarze (Jarocka, 2011).

Zaletą metody bezpośredniej jest to, że dzięki bliskiemu kontaktowi z jednostkami budującymi szara strefę istnieje możliwość zebrania szczegółowych informacji o wewnętrznej strukturze i motywach napędzających funkcjonowanie gospodarki ukrytej. Jednak istnieje również wiele wad, które posiada omawiana procedura. Przede wszystkim badanie odbywa się na pewnej próbie dobranych gospodarstw i sprawdzany jest stopień zależności z nieformalną gospodarką. Oczywiste wydaje się iż osoby uczestniczące w analizie mają świadomość konsekwencji wynikających z nielegalnych postępowań, dlatego warto mieć na uwadze, że przedstawiona rzeczywistość będzie przekłamana. Kolejnym wyzwaniem jest tylko częściowy obraz jaki badacze uzyskują po przeprowadzeniu pomiarów. Utrudnia to uzyskanie pełnego obrazu faktycznego stanu gospodarki nielegalnej (Jarocka, 2011).

Badania podatkowe są również jedną z metody badania gospodarki nieformalnej. Odnosząc

się do rejestrów służb podatkowych dokonuje się szczegółowych kontroli składanych zeznań podatkowych. Znaczącą wadą tego rozwiązania jest to, że brane pod uwagę są tylko osoby, które składają zeznania, co może powodować mylny obraz badania. Przypisanie typu dochodu osobom, które są określane jako pracujące, może też powodować, że metoda ta jest narażona na nieprecyzyjność pomiaru (Jarocka, 2011).

Metody pośrednie bazujące na wskaźnikach są dość proste w zastosowaniu. Przykładowym sposobem jest porównanie liczby zarejestrowanych bezrobotnych z szacowaną liczbą bezrobotnych według definicji bezrobotnego podawanej przez Międzynarodową Organizację Pracy (Jarocka, 2011).

Kolejną metodą należącą do metod pośrednich jest przyglądanie się wydatkom gospodarstw domowych. Badanie polega na porównaniu wydatków do oficjalnych statystyk dochodów i oszczędności. W momencie kiedy wynik znacząco odchyla się w stronę dokonywanych wydatków, można podejrzewać o pozyskiwanie zarobku z nielegalnego źródła. Rezultaty, które udaje się pozyskać na podstawie badania na próbie losowej uznawane są jako wyniki, które mogły być uzyskane podczas badania populacji. Taki sposób badania rachunków społecznych jest z pewnością obciążony błędem. Przyglądając się wydatkom nie mamy pewności czy jednostki korzystają z oszczędności, czy zakupy są dokonywane na nielegalnych rynkach i które z zasobów pochodzą z rejestrowanego czy nierejestrowanego źródła (Jarocka, 2011).

Makroekonomiczna analiza rynku pracy ocenia różnicę pomiędzy poziomem oficjalnej i faktycznej aktywności ekonomicznej. Faktyczne zatrudnienie jest szacowane przy pomocy szeregu danych empirycznych z wcześniejszych lat. Oprócz stosownych obliczeń w kolejnym etapie należy zagregować. Znaczącą wada jest dość odważna hipoteza, mówiąca że różnica pomiędzy faktycznym i oficjalnym zatrudnieniem jest skutkiem działania gospodarki nieformalnej. Powody takich wyników niekoniecznie muszą być związane z szarą strefą. Chociażby nieustanne zmiany w polityce społecznej czy też całkiem popularne, szczególnie wśród młodych ludzi, pracowanie na więcej niż jednym etacie (Jarocka, 2011).

Metoda pośrednia zawiera również podgrupę metod monetarnych. Skupia się ona na rzeczywistym i nominalnym popycie na pieniądz. Słabą stroną tego sposobu badania jest założenie, że wszystkie transakcje odbywające się w szarej strefie są dokonywane tylko i wyłącznie za pomocą gotówki. Badacze widząc wzrost popytu na banknoty utożsamiają to z wzrostem gospodarki nieoficjalnej (Jarocka, 2011).

Podmioty badające szarą strefę są świadome niedoskonałości jakie posiadają wyżej opi-

sane metody, dlatego postanowiono wykorzystać metody ekonometryczne. Badaniu głównie podlegały powody kryjące się za wielkością popytu na pieniądz gotówkowy i odnalezieniem istniejących zależności między popytem na pieniądz gotówkowy i opodatkowaniem dochodów (Jarocka, 2011).

Ekonometryczne metody często odnoszą się do zużycia materiałów bądź surowców. Popularna metoda Kaufmanna polega na ustaleniu zależności pomiędzy zużyciem energii elektrycznej i produkcji, która została wytworzona. Przyrównując do roku bazowego produkcję względem energii jesteśmy w stanie oszacować produkcję w roku badanym. Wynikająca różnica z teoretycznych i empirycznych wartości jest traktowana jako dowód na istnienie szarej strefy (Jarocka, 2011). Tabela 1.3 podsumowująca sposoby badania gospodarki nieformalnej.

Każda z przedstawionych metod 1.3 ma swoje plusy oraz minusy. Najczęstszą wadą wśród poznanych sposobów pomiaru jest ich arbitralność czy też subiektywizm. Gospodarka nieformalna jest na tyle trudnym do zbadania pojęciem, że intuicja badającego może okazać się przydatna. Natomiast łatwość pozyskania informacji okazuje się być największym atutem większości pokazanych metod.

1.4 Szara strefa w Polsce w świetle dostępnych źródeł danych

1.4.1 Badanie Aktywności Ekonomicznej Ludności

Główny Urząd Statystyczny co kwartał dokonuje Badania Aktywności Ekonomicznej Ludności (BAEL), dzięki któremu możliwe jest poznanie struktury pracujących. Dodatkowo pozyskane dane pozwalają obserwować zachodzące zmiany oraz umożliwiają dokonywać porównań danych międzynarodowych. Zbierane dane są przyporządkowane do poszczególnych modułów np. osoby niepełnosprawne na rynku pracy, czas i rozkład pracy czy też praca w nietypowych formach zatrudnienia. Badaniu podlegają gospodarstwa domowe i osoby w wieku powyżej 15 roku życia. Istnieje rozgraniczenie na trzy podstawowe zbiorowości badanej populacji - pracujący, bezrobotni (grupa aktywnych zawodowo) oraz bierni zawodowo (Główny Urząd Statystyczny, 2019).

W badaniu BAEL aspekt pracy nierejestrowanej jest również badany. Ze względu na trudną specyfikę zjawiska, dokonywane badanie nie określa dokładnych wyników. Według źródeł liczba nielegalnie zatrudnionych jest szacunkową wartością, która określa dolną granicę osób pracujących w szarej strefie. Oznacza to, że wyniki badania wskazują na minimalną przypusz-

Tabela 1.3. Metody pomiaru szarej strefy - zalety i wady

Metody	Plusy	Minusy
Bezpośrednie		
Badania gospodarstw domowych	Informacje zaczerpnięte bezpośrdnio u źródła, od podmiotów uczestniczących w gospodarce nieformalnej	Subiektywizm i ograniczona wiarygodność respondentów
Badania podatkowe	Łatwość dotarcia do materiałów źródłowych (zeznania podatkowe)	Ograniczona przydatność (nie obejmują podmiotów niezarejestrowanych)
Badania ankietowe rynku pracy	Możliwość uzyskania szerokiego zakresu informacji od podmiotów, funkcjonujących na nieoficjalnym rynku pracy	Problematyczna wiarygodność odpowiedzi respondentów, zwłaszcza na pytania dotyczące ich uczestnictwa w gospodarce ukrytej
Bezpośrednie badania rynków cząstkowych	Możliwość uzyskania szczegółowych informacji o konkretnych formach i przejawach gospodarki nieoficjalnej	Fragmentaryczność badań, trudność z uogólnieniem wyników
Pośrednie		
Analiza rozbieżności między wydatkami a statystyką dochodów	Relatywna łatwość stosowania	Niebezpieczeństwo automatyzmu i interpretacyjnego (traktowanie wszelkich rozbieżności jako rezultatu działalności w szarej strefie)
Makroekonomiczne analizy rynku pracy	Możliwość lepszego rozpoznania faktycznych relacji między popytem a podażą na rynku pracy(z uwzględnieniem zatrudnienia nierejestrowanego)	Silne uzależnienie wyników od sformułowanych hipotez wyjściowych
Metody monetarne	Łatwość pozyskania niezbędnych informacji, możliwość dokonywania porównań międzyokresowych	Arbitralność i subiektywizm przyjętych założeń, pomijanie wpływu innych czynników wpływających na podaż pieniądza
Metody ekonometryczne	Przejrzystość i elegancja formalna	Subiektywizm w doborze wskaźników, silne uzależnienie wyników od przyjmowanych założeń i hipotez wyjściowych
Metody oparte na bilansach zużycia niektórych materiałów i surowców	Relatywna łatwość stosowania i pozyskania niezbędnych informacji	Zbyt silne uzależnienie od jednej zmiennej objaśniającej, arbitralność i subiektywizm przyjętych założeń
Metody wieloczynnikowe	Uwzględnienie wpływu wielu czynników na gospodarkę nieoficjalną	Umowność i arbitralność przyjmowanych założeń
Pośrednie analityczne metody cząstkowe	Różnorodność wykorzystywanych źródeł informacji	Fragmentarność badań, trudności z uogólnianiem i syntezą wyników badań

Źródło: Jarocka, M. S. (2011). Analiza wybranych metod bezpośrednich i pośrednich służących o badania szarej strefy. (Jarocka, 2011, s. 42)

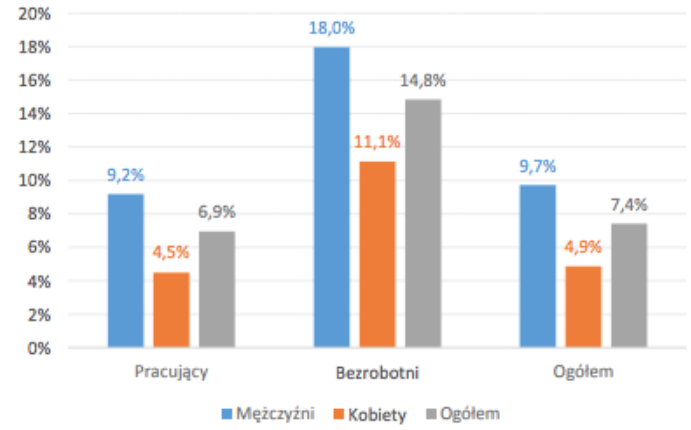
czaną wartość nielegalnych pracowników, jednak ich liczba w rzeczywistości może być większa (Główny Urząd Statystyczny, 2019).

1.4.2 Badanie Kapitału Ludzkiego

Oprócz Badania Aktywności Ekonomicznej Ludności istnieje wiele innych badań, które nadzorują sytuację na rynku pracy, jednak wiele z nich, ze względu na obszerny zakres badawczy, sku-

pia się na konkretnych zjawiskach. Kolejnym przykładem jest Bilans Kapitału Ludzkiego (BKL), który jest dokonywany przez Polską Agencję Rozwoju Przedsiębiorstwa wraz ze współpracy z Uniwersytetem Jagiellońskim w Krakowie. Celem badania jest porównanie strony podażowej i popytowej rynku pracy, tak by móc zweryfikować na ile potrzeby obydwu stron są zaspokajane. Przedsiębiorcy wypełniają ankietę zawierającą szereg pytań dotyczących pracowników oraz branży w jakiej działają. W formularzu znajduje się pytanie dotyczące nowych stanowisk pracy (Z3.1) i jedna z możliwych odpowiedzi świadczy o nieformalnym zatrudnianiu pracownika (PARP, 2017).

Rysunek 1.1. Odsetek osób wykonujących pracę w ramach umowy nieformalnej wśród aktywnych zawodowo



Minimalna liczebność (bezrobotne kobiety) = 77.

Źródło: Badanie kapitału ludzkiego, Polską Agencja Rozwoju Przedsiębiorstwa. (PARP, 2017, s. 17)

Odnosząc się do badań z 2017 roku z osób aktywnych zawodowo 7,4 % przyznało się do pracy nieformalnej. W grupie przebadanych najczęściej mężczyźni oraz osoby, które są zarejestrowane jako bezrobotne deklarowały się do pracy w szarej strefie. Powodem dominacji płci męskiej były z pewnością sektory w jakich najczęściej przedsiębiorcy deklarowali brak pracowników. Należało do nich przemysł i górnictwo (tutaj aż 34% przedsiębiorców wskazywało na brak odpowiedniej kadry pracowniczej) oraz budownictwo i transport (w tym dziale natomiast było 32% takich zgłoszeń) (PARP, 2017).

Natomiast kobiety znacznie częściej występowały w sektorach takich jak gastronomia, handel czy praca opiekuńcza. Jeśli chodzi o jednostki da których podejmowana była praca nieformalna to ogólnie dominowała praca dla firm czy też dla kogoś z rodziny bądź znajomych (PARP, 2017).

1.4.3 Państwowa Inspekcja Pracy

Bieżące kontrole Państwowej Inspekcji Pracy (PIP) pozwalają zweryfikować stan legalności zatrudnień obywateli. W 2017 roku dokonano 23,6 tys. kontroli, gdzie badano sektory gospodarki takie jak handel, budownictwo czy przetwórstwo przemysłowe. Jeżeli chodzi o podmioty, które zostały sprawdzane przez PIP wybierane są one na podstawie wniosków czy skarg z urzędów administracji publicznej. Skargi pochodzą głównie od powiatowych urzędów pracy, zakładu ubezpieczeń społecznych (ZUS) czy też urzędów skarbowych (US) (Państwowa Inspekcja Pracy, 2018). Zwracano również uwagę na konkretne sektory i kontroli podlegały:

1. działalności, w których wykryto nieprawidłowości w poprzednich latach
2. podmioty, które w szczególności potrzebują pracowników sezonowych (dotyczy to głównie branży gastronomicznej czy hotelarskiej)
3. przedsiębiorstwa działające w branży leśnej bądź z nią związane
4. podmioty poruszające się w specyficznych branżach (mogą dotyczyć konkretnego regionu)

Kontrolą objęto 176 tys. osób. Tabela (1.2) pokazuje liczby osób nielegalnie zatrudnionych.

Rysunek 1.2. Nielegalne zatrudnienie lub nielegalna inna praca zarobkowa - wg województw

Nielegalne zatrudnienie lub nielegalna inna praca zarobkowa – wg województw

Województwo	Liczba osób nielegalnie zatrudnionych	% w danym województwie
Śląskie	2 866	21,8
Małopolskie	2 236	26,0
Wielkopolskie	1 659	8,6
Łódzkie	1 349	14,7
Podkarpackie	1 286	12,6
Świętokrzyskie	798	11,7
Kujawsko-Pomorskie	590	4,5
Dolnośląskie	545	3,8
Podlaskie	485	4,3
Warmińsko-Mazurskie	483	11,6
Zachodniopomorskie	444	4,9
Pomorskie	345	2,7
Lubelskie	277	1,3
Mazowieckie	218	2,7
Opolskie	178	1,8
Lubuskie	61	1,3

Źródło: Sprawozdanie z działalności Państwowej Inspekcji Pracy w 2016 roku. (Państwowa Inspekcja Pracy, 2018, s. 101)

Nielegalne zatrudnienie dotyczyło 13,8 tys. obywateli. Większość z nich (12,8 tys.) nie została zgłoszona do ubezpieczenia społecznego. Pozostała część nie posiadała żadnego pisemnego zabezpieczenia o wykonywanej pracy ani określonych warunków zatrudnienia (Państwowa Inspekcja Pracy, 2018).

Według sekcji PKD najczęściej przepisy naruszono w strefie gospodarki magazynowej i transportu, tuż za tą częścią była działalność profesjonalna, a następnie informacja i komunikacja (Państwowa Inspekcja Pracy, 2018).

Jeżeli chodzi o wielkości przedsiębiorstw to wśród mikroprzedsiębiorstw (do 9 pracowników) najczęściej nie przestrzegano legalności zatrudnień. Zazwyczaj prowadzona działalność jest nieuporządkowana i nie prowadzi obsługi księgowej. Kolejne były małe przedsiębiorstwa i ostanie, prawie po równo, średnie i duże przedsiębiorstwa (Państwowa Inspekcja Pracy, 2018).

W odniesieniu do lat poprzednich zauważalna jest tendencja spadkowa. W 2017 r. zaniedbanie obowiązku powiadomienia powiatowego urzędu pracy (PUP) o zmianach w strefie zatrudnień nie dokonało 778 osób, przy czym w 2016 r było to 816 osób, a w 2015 r. 1361 osób (Państwowa Inspekcja Pracy, 2018).

Państwowa Inspekcja Pracy w sprawozdaniu z działalności (Państwowa Inspekcja Pracy, 2018) przytacza najistotniejsze przyczyny podejmowania pracy w gospodarce nieformalnej. Najczęściej wymieniana redukcja kosztów przez przedsiębiorców (niepłacenie składek ubezpieczeń społecznych, koniecznych opłat wynikających z zatrudnienia pracownika czy też unikanie zaliczek podatkowych od dochodu), gdzie obie strony żyskują", ponieważ pracodawca uzyskuje wyższe dochody przez co pracownik otrzymuje większe wynagrodzenie netto. Znacząca jest również sytuacja osobista pracowników nielegalnie zatrudnionych. Wykazanie faktycznego dochodu w przypadku osób korzystających z pomocy socjalnej mogłoby utracić możliwość korzystania z nich (Państwowa Inspekcja Pracy, 2018).

Innym wspomnianą przyczyną jest nieprzychylny system i mało dostosowane przepisy do rzeczywistości. Państwowa Inspekcja Pracy wskazuje, że problemem mogą być komunikacja z podmiotami współpracującymi(np. ZUS czy PUP), ze względu na brak baz online (Państwowa Inspekcja Pracy, 2018).

Rozwiązania jakie podaje PIP opierają się głównie na zmianach legislacyjnych. Przykładowo pracodawca powinien być zobowiązany do zgłoszenia pracownika do Zakładu Ubezpieczeń Społecznych przed dopuszczeniem jej do pracy oraz pisemnego potwierdzenia warunków umowy cywilno-prawnej (Państwowa Inspekcja Pracy, 2018).

Zwiększenie skuteczności kontroli Państwowej Inspekcji Pracy z pewnością również pomogłoby w zwalczaniu nielegalnych postępowań wśród przedsiębiorców. Jednym z pomysłów są zmiany związane z regulowaniem sankcji. Jej wysokość powinna być powiązana z ilością osób, które są zatrudnione nielegalnie. Ze względu na erę cyfryzacji PIP dąży do zapewnienia inspektorom rejestrów w wersji online (np. informacji o ubezpieczeniach społecznych i składkach od ZUS'u czy publicznych służb zatrudnienia) (Państwowa Inspekcja Pracy, 2018).

1.5 Podsumowanie

Szara strefa jest badana nie tylko, aby poznać przyczyny jej istnienia, ale też ze względu na wyzwania jakie stawia przed badaczami. Biorąc po uwagę dostępność źródeł możemy w ten sposób ograniczyć koszta analizy i wykorzystywać dostępne już dane. Gotowe rozwiązania mogą okazać się szczególnie przydatne dla przedsiębiorców, którzy korzystają dostępne informacje na potrzeby swoich przedsiębiorstw.

Jednak ze względu na trudność w badaniu gospodarki nieformalnej oraz dostępność danych konieczne w dalszej analizie jest wzięcie pod uwagę sposobu w jaki zebrane obserwacje są tworzone. Często proste narzędzia statystyczne mogą okazać się niewystarczające, aby zbadać lepiej szacowane zjawisko. Wtedy pomocne okazują się bardziej złożone modele mikroekonometryczne, pomagające w poprawny sposób zbadać interesujące nas zjawisko.

Rozdział 2

Uogólniony model doboru próby Heckman'a

2.1 Problematyka prób nielosowych

2.1.1 Losowy i nielosowy dobór próby

Zaczynając badanie jednym z początkowych etapów jest projektowanie badania częściowego, a dokładniej sposobu doboru próby. Wśród dostępnych metod mamy metody probabilistyczne, czyli losowe, bądź nieprobabilistyczne czyli nielosowe. Wybór ten uzależniony jest od wielu czynników. Przede wszystkim brane są pod uwagę potrzeby wnioskowania statystycznego, informacje bądź dostęp do nich o populacji, czy też budżet i czas na realizację przedsięwzięcia. Określenie techniki wiąże się z ustaleniem liczebności próby, sposobu pozyskania operatu losowania czy też włączenia do próby posiadanych już informacji (a priori) (Szreder & Krzykowski, 2005).

Zbierając wszelkie informacje na temat zbiorowości i wzbogacając wiedzę o nich, możemy dzięki temu przyczyniać się do rozwoju technik nielosowych. Wiedzę a priori, czyli wstępną oraz wnioski z próby to główne źródło informacji, które jest nieocenione przy badaniu próbkowym (Szreder & Krzykowski, 2005).

Przewagą próby losowej jest możliwość stosowania zasad rachunku prawdopodobieństwa, które nie mają zastosowania tam gdzie nie wystąpił mechanizm losowy czy też próba losowa (Szreder & Krzykowski, 2005).

Podstawą wnioskowania statystycznego są wielkości próby, wyniki estymacji oraz hipotezy testów, których źródłem jest właśnie rachunek prawdopodobieństwa (Szreder & Krzykowski, 2005).

Istotne jest również określenie liczebności próby. Nie jest to łatwe zadanie jeśli nie mamy ustalonych kryteriów. Najczęściej określa się dwa podstawowe założenia, mówiące o poziomie ufności oraz o średnim błędzie, który jest różnicą między oceną z próby, a jej wartością w populacji. Bez założenia, że przy selekcji próby posługujemy się techniką probabilistyczną nie byłoby możliwe abyśmy mogli spełnić kryteria. Inne sposoby wyboru nie były możliwe, jeżeli chcielibyśmy mieć próbę z określonym poziomem istotności czy też średnim błędem (Szreder, 2010).

Dobór próby nielosowej nawet obecnie jest uznawany za konieczność. Kiedy nie mamy operatu losowania, czyli bazy umożliwiającej wylosowanie próby do badania, lub jest on nie najlepszej jakości próba nielosowa może być pomocna. Innym aspektem, który może być jednocześnie motywem wyboru nielosowego doboru próby jak i korzyścią to niskie koszty jakie generują techniki nielosowe oraz oszczędność czasu. Dzięki temu dostajemy w dość szybkim czasie wyniki ukazujące pewną część populacji, dlatego warto mieć na uwadze jaki jest cel badania i czy losowy dobór próby jest potrzebny (Szreder, 2010).

Pojęcie klasycznego prawdopodobieństwa pozwala nam wyobrazić sobie na ile możliwe jest pewne zdarzenie. Poszukując odpowiednika dla prawdopodobieństwa w nielosowych próbach T. Bayes, L.J. Savage oraz B. de Finetti (Szreder, 1994) sformułowali pojęcie subiektywnego prawdopodobieństwa. Odnosi się ono do prawdziwości osądu zdarzenia, a konkretniej do stopnia pewności że dany osąd jest prawdziwy. Jak wynika z nazwy prawdopodobieństwo o może być różne w zależności od osoby. Najczęściej ten rodzaj prawdopodobieństwa stosowany jest do zdarzeń rzadko powtarzających się bądź zdarzeń jednostkowych. Na obecną chwilę metoda ta zyskuje coraz większą popularność. Obecnie zdobycie informacji na temat populacji nie jest trudne, dlatego teraz kluczowe jest zbadanie jak tą wiedze wykorzystać (Szreder, 2010).

Inne spostrzeżenie pokazuje, że nielosowy dobór próby jest potrzebny, ponieważ gdyby mechanizm losowania był w pełni dopracowany, pojęcie to by nie było konieczne. Największą obawą przed używaniem technik nielosowego doboru próby jest trudność w estymacji błędów. W technikach nielosowych nie ma zdefiniowanego żadnego składnika błędu, co powoduje niepewność odnośnie wniosków, które powstają przy realizacji badania. Jednak jednostki odpowiedzialne za badania znalazły sposób wykorzystując wiedzę oraz doświadczenie badaczy czy nawet ekspertów do określenia błędu. Kiedy jednak populacja jest trudna wciąż do zbadania najczęściej ośrodki przeprowadzające badanie rezygnują z wnioskowań i pozostają tylko przy opisie statystycznym. (Szreder, 2010)

2.1.2 Selekcja w próbie

Chcąc zbadać problemy mikroekonomiczne dotyczące płac, szarej strefy czy odnosząc się do innych dziedzin jak na przykład psychologia chcemy poznać skutek i przyczynę danych zjawisk. W przypadku tego typu problemów możemy wykorzystać próby losowe (np. reprezentatywne badania częściowe) lub próby nielosowe (np. dobór celowy, rejestry administracyjne pokrywające określone subpopulacje). Oznacza to, że wartość naszej zmiennej zależnej jest uzależniona od innej zmiennej, która dokonuje selekcji wyboru elementów do naszej próby (Wołodźko, 2015).

Kiedy podjęliśmy decyzję odnośnie wyboru próby i to jakie czynniki chcemy zbadać możemy rozpoczynać dalszą część badania. Próbę możemy opisać jako grupę jednostek wybraną przez badającego. W literaturze ekonometrycznej używa się w tym celu pojęcia selekcja, które jest bezpośrednim tłumaczeniem z j. angielskiego *selection*. Jest to ogólne pojęcie odnoszące się zarówno do metod probabilistycznych, jak i nieprobabilistycznych (Toomet, Henningsen i in., 2008).

Często aby zrozumieć istotę modelu z selekcją podawanym przykładem jest właśnie badanie szarej strefy. Problem jej zbadania odnosi się właśnie do takiego modelu. Znamy liczbę bezrobotnych, ale tylko tą część, która widnieje zarejestrowana. Nie znając wartości cech, które badamy u osób które nie są oficjalnie bezrobotne, ale też nie pracują, ciężko jest nam określić dlaczego dana osoba nie rejestruje swojego bezrobocia. Innym przykładem może być udział w ankiecie. Jest wiele czynników, dla których niektórzy biorą udział w różnego rodzaju badaniach czy ankietach. Przedmiotem badania byłoby określenie od czego zależy czy ktoś weźmie udział bądź nie w ankiecie (Strawiński, 2007). W przypadku gdy jednostki same decydują o uczestnictwie w określonym badaniu lub korzystaniu z określonych usług czy serwisów (np. społecznościowych) wtedy mamy do czynienia z autoselekcją lub samoselekcją.

2.2 Model doboru próby Heckman'a

2.2.1 Regresja dla zmiennych ograniczonych

Badanie pewnych wzorców zachowań wśród jednostek jest obecnie bardzo popularne w przeróżnych działalnościach związanych ze sprzedażą. Firmy interesuje co motywuje ludzi do podejmowania określonych decyzji. Istnieją modele pomagające bliżej poznać charakterystykę

naszych wyborów. Przykładem może być model tobitowy, którego nazwa pochodzi od nazwiska twórcy, Jamesa Tobina (Amemiya, 1984). Początkowo nazywany był modelem ograniczonej zmiennej zależnej (ang. truncated regression), jednak ze względu na podobną charakterystykę do modelu probitowego, model ostatecznie nazywany jest tobitowym. Tobin badał wydatki na dobra w gospodarstwach domowych przy pomocy regresji. Wydatki nie mogły być ujemne, dlatego narzucono pewne ograniczenia. Założeniem modelu jest to, że znacząca część wartości zmiennej objaśnianej jest równa pewnej stałej, przy czym pozostałe wartości są większe od tej stałej (Gruszczyński, 2010).

Innym przykładem może być przeznaczanie datków na działalności charytatywne, znaczna część ludzi nie przekazuje żadnych pieniędzy. Chcąc zbadać co wpływa na przekazanie pieniędzy na rzecz fundacji musimy skonstruować model. Mamy więc zmienną, która określa czy pieniądze zostały wpłacone czy nie. Interesują nas tylko zjawiska wpłaty datków tak by móc zbadać jakie cechy określają osoby wpłacające datki. Obserwacje, dla których wpłata była zerowa są cenzurowane, ieważ nie dadzą nam informacji, którą chcemy zbadać (Gruszczyński, 2010).

W modelu tobitowym występuje stała, która określa czy rozwiązaniem jest rozwiązanie brzegowe czy jednak wynikiem jest cenzurowanie zmiennej. Stała jest wartością progową, która określa przedział dla zmiennej decyzyjnej. Przykładem mogą być wydatki na wakacje wśród decydentów. Nie możemy wydawać mniej niż zero na wakacje, dlatego wartość stałej będzie równa 0 (czyli osoby, które nie wydają pieniędzy na wakacje), jednak istnieją przypadki kiedy ludzie wydają pieniądze na wakacje (i są to różne kwoty) stąd powstaje nam przedział $< 0, +\infty)$ wydatków na wakacje (Gruszczyński, 2010).

Ocenzurowanie wiąże się z postacią modelu zaproowanego przez Tobina. Oznacza to, że niemożliwe jest poznanie wartości mniejszych od naszej stałej. To co nas powinno najbardziej interesować to jakie jest prawdopodobieństwo przyjęcia wartości brzegowej przez naszą zmienną oraz ewentualnie jaka będzie jej wartość oczekiwana jeśli wartość będzie wyższa niż brzegowa (Gruszczyński, 2010).

Istnieją dwa pojęcia odnoszące się do zmiennych ograniczonych.

- **Zmienna ocenzurowana** jest to zmienna, która jest obserwowana na określonym przedziale. Krańce tego przedziału to wartości brzegowe. Przykładowo zmienna spoza danego przedziału może oznaczać chęć podjęcia pewnej decyzji, natomiast zmienna w przedziale jest podjętą decyzją (Gruszczyński, 2010).

- **Zmienna ucięta** różni się od zmiennej ocenzurowanej tym, że obserwator nie zna ob-

serwacji poza obecnymi w danym przedziale. Ucinamy zbiór jeszcze przed samym losowaniem, co oznacza, że istnieje część obserwacji, która nie została zaobserwowana Gruszczyński (2010).

Model Tobitowy należy do grupy modeli regresji uciętej. Regresja ucięta dotyczy wartości zmiennych cenzurowanych. Badacz najczęściej ich nie zna i próba w takim przypadku dotyczy tylko obserwacji nieocenzurowanych. Co za tym idzie, usuwamy pewną część populacji i to uniemożliwia nam korzystania z metody metody najmniejszych kwadratów (im ograniczenie jest większe, tym obciążenie MNK jest większe). Istnieją tak więc inne sposoby na szacowanie modeli regresji np. model selekcji próby (Gruszczyński, 2010).

2.2.2 Klasyczny model Heckman'a

W grupie modeli powiązanych z regresją uciętą jest model selekcji próby Heckman'a. Model ten zawiera równianie regresji, które ustala jakie obserwacje znajdują się w próbie. Tworząc modele selekcji, zakładamy że kluczową rolę odgrywa użyteczność. Klasyczny model selekcji próby ma następującą postać (Toomet, Henningsen i in., 2008):

$$\begin{cases} y_i^{S*} & = \beta^{S'} x_i^S + \varepsilon_i^S, \\ y_i^{O*} & = \beta^{O'} x_i^O + \varepsilon_i^O, \end{cases} \tag{2.1}$$

gdzie y_i^{S*} jest zmienną latentną dla równania selekcji (*selection*), y_i^{O*} jest to zmienna latentna (dotyczy równania wynikowego(*outcome*)) i x_i^S, x_i^O zmienne, które tworzą równania wynikowe i selekcji dla modelu. x_i^S, x_i^O mogą być sobie równe, jednak nie muszą.

Podstawowym założeniem obserwacji zmiennych są następujące założenia

$$y_i^S = \begin{cases} 0 & \text{jeżeli } y_i^{S*} < 0, \\ 1 & \text{w innym przypadku.} \end{cases} \tag{2.2}$$

$$y_i^O = \begin{cases} 0 & \text{jeżeli } y_i^S = 0, \\ y_i^{O*} & \text{w innym przypadku.} \end{cases} \tag{2.3}$$

Zmienne naszego równania (2.1) są brane pod uwagę wtedy i tylko wtedy kiedy zmienna y_i^S jest dodatnia. Zmienna latentna jest powiązana z naszą zmienną w równaniu selekcji. Przykładowo zmienna y_i^S przyjmuje wartość 1 gdy badane przedsiębiorstwo zatrudnia nielegalnie

pracowników, w przeciwnym razie 0. W takim przypadku y_i^{O*} określa skłonność przedsiębiorców do nielegalnych zatrudnień (Toomet, Henningsen i in., 2008).

Parametry β to wektor szukanych parametrów modelu selekcji. Mają one interpretacją taką samą jak przy regresji liniowej, czyli wzrost zmiennej objaśniającej o jedną jednostkę powoduje odpowiednią zmianę zmiennej objaśnianej. Równanie wynikowe jest równaniem regresji, jednak różni się ono od tradycyjnego równania tym, że zmienne są ocenzurowane. Wartość naszej zmiennej niezależnej jest wynikiem dla próby, która została poddana selekcji. Równanie selekcji, uznawane za skutek uboczny modelowania, wskazuje obciążenie próby. Obciążenie próby przy nielosowej selekcji może być spowodowane występowaniem jednego z wielu czynników. Przykładem może być samoselekcja, kiedy to decydent określa czy znajduje się w konkretnej grupie (na przykład pracuje lub nie pracuje). Równanie wynikowe (2.1) możemy szacować przy pomocy metody najmniejszych kwadratów, tylko wtedy gdy składniki losowe ε_i^S i ε_i^O są nieskorelowane. Jednak kiedy zachodzi korelacja pomiędzy nimi to im wyższa wartość tym wpływ selekcji jest silniejszy (Toomet, Henningsen i in., 2008).

Dokonując selekcji modelu Heckman'a najważniejsze założenie, które musi być spełnione dotyczy składników losowych. Muszą mieć one rozkład normalny, co pozwoli wyznaczyć wartość oczekiwaną zmiennej objaśnianej równania (2.1). Zapis (2.4) oznacza dwywymiarowy rozkład reszt, przy czym ρ to współczynnik korelacji reszt, a σ to ich odchylenie standardowe (Toomet, Henningsen i in., 2008).

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix} \right). \tag{2.4}$$

Jeśli chodzi o zależność między badaną cechą, a zmiennymi objaśniającymi w populacji zakłada się zależność liniową. Z faktu, iż model Heckman'a nie dotyczy całej próbki, uwzględniamy selekcję odnoszącą się do składnika losowego. Wtedy wartość oczekiwana ma następującą postać:

$$\begin{aligned} &\mathrm{E}\left[y^O | x^O = x_i^O, x^S = x_i^S, y^S = 1\right] = \\ &= \beta^{O\prime} x_i^O + \mathrm{E}\left[\varepsilon^O | \varepsilon^S \geq -\beta^{S\prime} x_i^S\right]. \end{aligned} \tag{2.5}$$

Mając taką postać zależności, możemy podstawić ją do równania wynikowego (2.1) i otrzy-

mujemy:

$$y_i^O = \beta^{O'} x_i^O + \mathrm{E}\left[\varepsilon^O | \varepsilon^S \geq -\beta^{S'} x_i^S\right] + \eta_i \equiv \beta^{O'} x_i^O + \rho\sigma\lambda\left(\beta^{S'} x_i^S\right) + \eta_i. \tag{2.6}$$

Równanie (2.6) jest odpowiedzialne za określenie nielosowego doboru do próby. W tym przypadku użycie metody najmniejszych kwadratów dla równania wynikowego byłoby nieodpowiednie, ieważ pominęlibyśmy część wartości oczekiwanej $\rho\sigma\lambda$, określane predyktorem (Gruszczyński, 2010). Istotny predyktor świadczy o nielosowym charakterze doboru do próby. Jeśli współczynnik korelacji ρ jest równy 0, wtedy wyrażenie wynosi 0, jednak im większe ρ tym wartość oczekiwana jest wyższa, ale też oznacza, że wpływ selekcji jest wyższy. Gdzie $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ określana jako iloraz Mills'a, gdzie $\phi(\cdot), \Phi(\cdot)$ są funkcjami standardowego rozkładu normalnego i skumulowaną funkcją standardowego rozkładu normalnego(Toomet, Henningsen i in., 2008).

Chcąc estymować model Heckman'a mamy do wyboru dwustopniową procedurę bądź metodę największej wiarygodności.

Metoda dwustopniowa zawiera następujące kroki:

1. Szacowanie modelu probitowego dla równania selekcji. Wyznaczenie odwrotnego ilorazu Mills'a.

2. Szacowanie modelu

$$y_i^O = \beta^{O'} x_i^O + \rho\sigma\lambda\left(\beta^{S'} x_i^S\right), \tag{2.7}$$

   korzystając z metody najmniejszych kwadratów.

Multiplikator $\rho\sigma$ może być szacowany za pomocą metody najmniejszych kwadratów ($\hat{\beta}^{\lambda}$). Nieznany predyktor $\lambda(\cdot)$ zastępujemy estymatorem z oszacowanej regresji probitowej z kroku pierwszego (Gruszczyński, 2010).

Estymator wariancji ε_i^O ma następującą postać

$$\hat{\sigma}^2 = \frac{\hat{\eta}'\hat{\eta}}{n^O} + \frac{\sum_i \hat{\delta}_i}{n^O}\hat{\beta}^{\lambda}, \tag{2.8}$$

gdzie $\hat{\eta}'$ jest wektorem reszt z estymacji najmniejszych kwadratów równania (2.6), n^O jest liczbą obserwacji w tej estymacji, oraz $\hat{\delta}_i = \hat{\lambda}_i\left(\hat{\lambda}_i + \hat{\beta}^{S'} x_i^S\right)$. Estymator korelacji pomiędzy ε_i^O i ε_i^s możemy zapisać jako $\hat{\rho} = \hat{\beta}^{\lambda}/\hat{\sigma}$. Wartości $\hat{\rho}$ mogą wykraczać poza przedział $[-1, 1]$.

Druga metoda pozwala to konstrukcja estymatora funkcji największej wiarygodności. Korzystając własności dwumianowego rozkładu normalnego funkcje wiarygodności możemy zapisać jako (Toomet, Henningsen i in., 2008):

$$\ell = \sum_{\{i:y_i^S=0\}} \log\Phi\left(-\beta^{S\prime}x_i^S\right) + \\ + \sum_{\{i:y_i^S=1\}} \left[\log\Phi\left(\frac{\beta^{S\prime}x_i^S + \frac{\rho}{\sigma}\left(y_i^O - \beta^{O\prime}x_i^O\right)}{\sqrt{1-\rho^2}}\right) - \frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}\frac{\left(y_i^O - \beta^{O\prime}x_i^O\right)^2}{\sigma^2}\right]. \tag{2.9}$$

Według źródeł oryginalny artykuł przedstawiający model Heckman'a sugerował wybór metody dwustopniowej ze względu na to że była tańsza w porównaniu do metody największej wiarygodności. Jednak w obecnych czasach nie ma to już znaczenia, ale metoda dwustopniowa wciąż ma przewagę nad metodą największej wiarygodności. Przede wszystkim pozwala ona na pewne uogólnienia oraz przy bardzo wysokim skorelowaniu reszt algorytm liczący estymatory największej wiarygodności może się nie pokryć. Oznacza to, że gdy ρ = 1 metoda dwustopniowa wyznaczy parametry, natomiast MNW może się nie sprawdzić (Toomet, Henningsen i in., 2008).

2.2.3 Uogólniony model Heckman'a

Podstawowy model selekcji Heckman'a (2.6) dotyczył równań przy których głównym założeniem był rozkład normalny reszt, zmienna objaśniana była zmienną ciągłą, a regresja opisująca równania była regresją logistyczną. W przypadku kiedy mamy zmienne licznikowe (ang. *count data*) możemy wykorzystać uogólniony model Heckman'a. Zmienne wtedy mogą mieć rozkład Poisson'a, dwumianowy czy też dwumianowy ujemny i dokonanie estymacji jest możliwe.

2.2.4 Model Heckman'a dla rozkładu Poisson'a

W tej sekcji postaramy się wyznaczyć model selekcji Heckman'a dla zmiennej o rozkładzie Poisson'a. Funkcja dla tego rozkładu ma następującą postać (Greene, 1995):

$$\begin{aligned} \text{Prob}\left(y_i = j\right) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, j = 0, 1, \\ \lambda_i &= e^{\beta' x_i}, \end{aligned} \tag{2.10}$$

gdzie wartość oczekiwana (średnia) to λ_i. Dla ułatwienia zapisu uznamy, że $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ będzie określane jako M_i.

W przypadku rozkładu Poissona wartość oczekiwana ma następującą postać (Greene, 1995):

$$E\left[y_i^O|\mathbf{x}_i^O, y_i^s = 1\right] = e^{\beta'\mathbf{x}_i^O + \theta M_i}, \tag{2.11}$$

Wartości z równania wynikowego, czyli dla y_i^O i x_i^O są obserwowane tylko wtedy, kiedy y_i^s wynosi 1. Dwukrokowa estymacja zaczyna się od dopasowania modelu probitowego dzięki funkcji największej wiarygodności, następnie wyznaczenie M_i i dopasowanie modelu Poisson'a z jego zwiększoną wartością oczekiwaną (Greene, 1995).

Aby dokonać estymacji modelu musimy dokonać dwukrokwej procedury. W przypadku rozkładów warunkowych funkcja największej wiarygodności będzie odwróconym hesjanem macierzy kowariancji estymatora największego prawdopodobieństwa. Mechanizm selekcji ma następującą postać (Greene, 1995):

$$\ln E\left[y_i^O|\mathbf{x}_i^O, \varepsilon_i\right] = \beta'\mathbf{x}_i^O + \varepsilon_i. \tag{2.12}$$

Łączymy normalność ε_i i ε_s i otrzymujemy

$$\begin{aligned} E\left[y_i^O|\mathbf{x}_i^O, y_i^s = 1\right] &= E\left[y_i^O|\mathbf{x}_i^O, \varepsilon^{S''} > -\beta^{S''}\mathbf{x}_i^s\right] \\ &= e^{\beta^{O''}\mathbf{x}_i^O + \sigma^2/2}\left[\frac{\Phi\left(\theta + \beta^{S''}\mathbf{x}_i^s\right)}{\Phi\left(\beta^{S''}\mathbf{x}_i^s\right)}\right] \\ &= e^{\beta^{O''}\mathbf{x}_{iO}}\left[\Psi(\theta, \tau_i)\right] \\ &= \lambda_i\Psi_i, \\ &\text{gdzie } \tau_i = \beta^{S''}\mathbf{x}_i^s. \end{aligned} \tag{2.13}$$

W wyniku przekształceń w równaniu 2.13 otrzymaliśmy funkcję dzięki której możemy dokonać estymacji metodą najmniejszych kwadratów (Greene, 1995). Oszacowania można również dokonać z wykorzystaniem metody największej wiarygodności.

2.2.5 Model selekcji dla rozkładu ujemnego dwumianowego

W tym podrozdziale przedstawiono uogólniony model selekcji Heckman'a w przypadku zmiennych z rozkładem dwumianowym ujemnym. Rozkład ten jest połączeniem funkcji Gamma Γ oraz rozkładu Poisson'a. Dotyczy on najczęściej danych licznikowych (*ang. count data*) z dużą

zmiennością (*ang. overdispersion*) (Zeileis, Kleiber & Jackman, 2008). Funkcja ma następującą postć (Greene, 1994):

$$\begin{gathered} p\left(y_i\right)=\frac{\Gamma(y_i+\theta)}{\Gamma(\theta) y!} u_i^0\left(1-u_i\right)^{y_i}, \theta>0, y_i=0,1, \ldots, \\ u_i=\frac{\theta}{\theta+\lambda_i}, \end{gathered} \tag{2.14}$$

gdzie Γ to funkcja Gamma, θ to liczba sukcesów.

$$E\left[y_i\right]=\lambda_i, \tag{2.15}$$

$$\operatorname{Var}\left[y_i\right]=\lambda_i\left[1+(1/\theta)\lambda_j\right]=\lambda_i\left(1+\alpha\lambda_i\right). \tag{2.16}$$

Model rozkładu ujemnego dwumianowego został sformułowany z dużą zmiennością (Greene, 1994):

$$\frac{\operatorname{Var}\left[y_i\right]}{E\left[y_i\right]}=1+\alpha E\left[y_i\right]>1. \tag{2.17}$$

Wyłączając nieobserwowaną heterogeniczność otrzymano:

$$E\left[y_i|\varepsilon_i\right]=\lambda_i\varepsilon_i. \tag{2.18}$$

gdzie ε_i to rozkład o rozkładzie gamama ze średnią 1 i wariancją α (Greene, 1994).

$$g(\varepsilon_i)=\frac{\theta^0}{\Gamma(\theta)} e^{-\theta_a} \varepsilon_1^{0.1}>0, \theta=\frac{1}{\alpha}, \tag{2.19}$$

$$f(y_{i\varepsilon_i})=\frac{e^{-exp(\beta' x_i+\varepsilon_i)}[exp(\beta' x_i+\varepsilon_i)]^{y_i}}{y!}. \tag{2.20}$$

Funkcja rozkładu brzegowego ma postać

$$f\left(y_i\right)=\int_0^{\infty} f(y_{i-\varepsilon_i})(g(\varepsilon_i))(d_{\varepsilon_i}). \tag{2.21}$$

Natomiast funkcja największej wiarygodności

$$\ell(y_i, u_i)=\sum_{i=1}^N \ln\left(\Gamma\left(\theta_i+y_i\right)\right)-\sum_{i=1}^N \ln\left(\theta_i!\right)-N \ln(\Gamma(y_i))+\sum_{i=1}^N \theta_i \ln(u_i)+N y_i \ln(1-u_i). \tag{2.22}$$

W równaniach (2.23),(2.24) znajdują się pochodne pierwszego i drugiego stopnia. Hesjan

jest zawsze ujemnie określony, co pozwala w metodzie Newton'a w wygodny sposób wyznaczyć parametry β dla funkcji największej wiarygodności. Model ten dotyczy nieliniowej regresji, tak więc β może zostać wyznaczona przez nieliniową metodę najmniejszych kwadratów (Greene, 1994).

$$\frac{\partial \ell(y_i, u_i)}{\partial u_i} = \left[\sum_{i=1}^{N} \theta_i \frac{1}{u_i}\right] - N y_i \frac{1}{1-u_i} = 0, \tag{2.23}$$

$$\frac{\partial \ell(y_i, u_i)}{\partial y_i} = \left[\sum_{i=1}^{N} \psi(\theta_i + y_i)\right] - N\psi(y_i) + N\ln(1-u_i) = 0. \tag{2.24}$$

Model Heckman'a ma zastosowanie dla zmiennych o różnych rozkładach. Dane ze zmienną licznikową (*ang. count data*) również mogą być użyte w estymacji, co zostanie przedstawione w rozdziale 3 (Greene, 1994).

2.3 Wybrane implementacje w pakiecie R

2.3.1 Pakiet sampleSelection

Możliwość przeprowadzenia estymacji modelu selekcji próby Heckman'a jest dostępny w kilku oprogramowaniach służących do analiz statystycznych. Między innymi takie badanie możemy wykonać w SPSS'ie, Stacie czy R. W tej części pracy przyjrzymy się dokładniej pakietowi SampleSelection, który jest wykorzystywany w R (Toomet, Henningsen i in., 2008).

W pakiecie SampleSelection mamy wiele funkcji umożliwiających poznanie naszego modelu. Możemy ustalić współczynniki modelu poprzez funkcję *coef.selection* oraz dzięki *coef.summary.selection* ich podstawowe własności takie jak błędy czy p-wartość. Oprócz tego istnieje funkcja *fitted.selection*, która wyznacza wartości dopasowane do naszego modelu selekcji. Reszty modelu natomiast zbadamy przy pomocy *residuals.probit*. Jeśli chcielibyśmy wyznaczyć odwrotny iloczyn Mills'a istnieje funkcja *invMillsRatio*, której argumentami jest estymowany model i binarna zmienna określająca czy wszystkie obserwacje mają znaleźć się w wskaźniku. Kolejną przydatną funkcją jest *heckitVcov*, której wynikiem jest macierz kowariancji (Toomet, Henningsen i in., 2008).

Oprócz wyznaczania podstawowych parametrów modelu istnieje też funkcja umożliwiająca estymację. *Selection* zawiera wiele argumentów, które pozwalają lepiej określić szacowany model. Możemy przeprowadzać estymację nie tylko dla modelu z jednym równaniem, ale też

z dwoma (argument *outcome*). *Weights* to wektor, który określa wagi (kiedy nie mamy wag argument przyjmuje wartość NULL). Kolejnym argumentem jest *method*. Pozwala określić w jaki sposób model ma być estymowany. Wartość ”m1” oznacza metodę największej wiarygodności , ”2step” odnosi się do dwuetapowej procedury estymacji oraz ”model.frame” zwraca tylko wszystkie zmienne użyte do szacowania (Toomet, Henningsen i in., 2008). .

2.3.2 Pakiet SemiParSampleSel

Innym pakietem, który możemy wykorzystać jest *SemiPaSampleSel*. Został on stworzony w 2017 roku. Dotyczy on szacowania modelu selekcji dla zmiennych o rozkładzie innym niż normalny. Posiada on wiele analogicznych funkcji do pakietu *SampleSelection*. Początkowo możemy oszacować model selekcji dla zmiennych o rozkładach Poisson'a, ujemnym dwumianowym, geometrycznym, logarytmicznym i wiele więcej. *summary.SemiParSampleSel* zwraca podsumowanie estymowanego modelu, czyli wartości parametrów i ich błędy standardowe szacunku, poziom istotności czy też wartość współczynnika korelacji. *logLik.SemiParSampleSel* pozwala poznać wartość estymatora największej wiarygodności obliczonego modelu oraz stopnie swobody. Chcąc zwizualizować dopasowanie predyktorów naszego modelu możemy użyć funkcji *plot.SemiParSampleSel*. Argumentami tej funkcji jest model wyznaczony przy pomocy *SemiParSampleSel* oraz równanie na podstawie, którego wykres ma zostać stworzonych, określający dopasowanie modelu. Dodatkowo dzięki *predict.SemiParSampleSel* możemy uzyskać predyktory dla nowego zestawu wartości lub dla estymowanego już modelu (Wojtys, Marra & Radice, 2016).

2.3.3 Podsumowanie

Model selekcji Heckman'a staje się coraz bardziej popularny i zyskuje coraz szersze zastosowanie. Prosta implementacja modelu we współczesnych oprogramowaniach statystycznych pozwala wykorzystać możliwości jakie daje sam model. Nawiązując do szarej strefy można zbadać cechy przedsiębiorców, którzy dokonują nielegalnych zatrudnień, bądź też samych zatrudnionych. Takie badanie pozwoli porównać wyniki pochodzące ze statystyk publicznych i umożliwi wyciągnąć wnioski oraz pokaże nieco inne spojrzenie na gospodarkę nieformalną. W rozdziale 3 dokonane będzie szacowanie charakterystyki przedsiębiorstw, które poprzez kontrole Państwowej Inspekcji Pracy wykazały nielegalność zatrudnionych pracowników.

Rozdział 3

Model opisujący liczbę nielegalnie zatrudnionych

3.1 Dane Państwowej Inspekcji Pracy

W tej części pracy dokonamy estymacji modelu Heckman'a dla danych z badań Państwowej Inspekcji Pracy. Badane były przedsiębiorstwa dokonujące nielegalnych zatrudnień lub takie które mogłyby to robić . Zbiór dotyczy lat 2010 do 2016, jednak model będzie szacowany dla roku 2016. Dane zostały pozyskane na potrzeby badań przeprowadzonych przez dra Macieja Beręsewicza (Uniwersytet Ekonomiczny w Poznaniu) i dr Dagmary Nikulin (Politechnika Gdańska) i udostępnione na potrzeby niniejszej pracy.

Dane zostały zebrane w trakcie kontroli Państwowej Inspekcji Pracy. W 2016 roku zostało sprawdzonych 27,5 tysięcy przedsiębiorstw, a w 2010 24,6 tysięcy. Badane firmy zostały wybierane na podstawie kilku czynników. Skupiano się głownie na przedsiębiorstwach, gdzie w poprzednich latach wykryto nieprawidłowości oraz te które działające głównie sezonowo. Następne kryterium dotyczyło podmiotów działających w sekcjach specyficznych dla wybranego regionu.

W badanym zbiorze występuje 14 zmiennych opisanych następująco:

- `rok` – rok badania (od 2010 do 2016),
- `woj` – województwo,
- `pow` – kod powiat,
- `pow_kod` – kod powiatu (określane przez Główny Urząd Statystyczny),

- pkd – kody Polskiej Klasyfikacji Działalności,
- wielk – wielkość badanego przedsiębiorstwa (określona w przedziałach),
- liczba_spr – liczba zbadanych przedsiębiorstw,
- liczba_niel – liczba nielegalnie zatrudnionych,
- sekcja – sekcja PKD,
- kontrole – liczna przeprowadzonych kontroli,
- skargi – liczba skarg,
- skargi_bez_um – skargi na przedsiębiorstwa, które zatrudniały bez umów,
- bez_ubezp_n – liczba zatrudnionych bez ubezpieczenia,
- bez_umowy_n – liczba zatrudnionych bez umowy,
- selekcja – zmienna wskazująca czy występowało nielegalne zatrudnienie (przyjmuje wartość 1 jeśli zatrudniano nielegalnie i 0 w przeciwnym wypadku).

Celem analizy będzie zbadanie charakterystyki przedsiębiorstw, które wspierają szara strefę. Założeniem jest sprawdzenie, czy firmy zatrudniające nielegalnie zatrudniają coraz więcej, czy raczej utrzymują zatrudnienia na stałym poziomie. Dodatkowo sprawdzony zostanie ewentualny wzrost bądź spadek pomiędzy 2010 a 2016 rokiem.

3.2 Eksploracyjna analiza danych

3.2.1 Badanie rozkładu próby

Pierwszym krokiem jest dopasowanie rozkładu do próby. Jest to potrzebne, aby określić z jakiego typu modelu należy skorzystać. Klasyczny model dotyczył zmiennej objaśnianej o rozkładzie normalnym, natomiast uogólniony model pozwala na estymacje zmiennej z innym rozkładem. W przypadku naszego modelu badamy liczbę nielegalnych zatrudnień. Zmienna będzie badana pod względem dopasowania do modelu Posisson'a bądź modelu ujemnego dwumianowego (nazywanego inaczej rozkładem Pascal'a).

Tabela 3.1. Statystyki opisowe dla liczby nielegalnie zatrudnionych w 2010 i 2016 r.

Rok	Minimum	Q1	Mediana	Średnia	Q3	Maximum
2010	1.0	1.0	2.0	2.9	3.0	116.0
2016	1.0	1.0	2.0	2.7	3.0	106.0

Źródło: Opracowanie własne na podstawie danych z Państwowej Inspekcji Pracy

Przed sprawdzaniem rozkładu zmiennej, zbadamy podstawowe statystyki dla nielegalnych zatrudnień dla 2010 roku oraz dla 2016 roku.

Jak widzimy w 2010 r. nie tylko średnia była większa, ale również wartość maksymalna. Jednak dla tworzenia modelu tak wysokie wartości można usunąć z naszego zbioru, ponieważ mogą one wpływać na wynik. Wartości odstające (ang. *outlier*) są to obserwacje które zdarzają się w zbiorze, jednak znacząco różnią się od reszty.

W języku R istnieje pakiet *vcd*, który posiada kilka funkcji umożliwiających sprawdzenie dopasowania rozkładu. Między innymi funkcja *goodfit* opierająca się na teście χ^2 dopasowania, pozwala sprawdzić poziom dobranych zmiennych dyskretnych do rozkładu (Meyer, Zeileis, Hornik, Meyer & KernSmooth, 2007). Układ hipotez tego testu wygląda następująco:

- H_0: Nie ma znaczącej różnicy pomiędzy wartością obserwowaną a oczekiwaną rozkładu referencyjnego,
- H_1: Jest znacząca różnica pomiędzy wartością obserwowaną a oczekiwaną rozkładu referencyjnego.

Testujemy zmienną wskazującą liczbę nielegalnych zatrudnień z 2016 roku. Otrzymane wyniki podane w tabeli 3.2 pozwolą ustalić, która z hipotez jest prawdziwa. Wartość dla rozkładu dwumianowego ujemnego jest dużo niższa niż dla rozkładu Poisson'a, co świadczy o dużo lepszym dopasowaniu do danych empirycznych.

Tabela 3.2. Test χ^2 dopasowania rozkładu Poissona i ujemnego dwumianowego

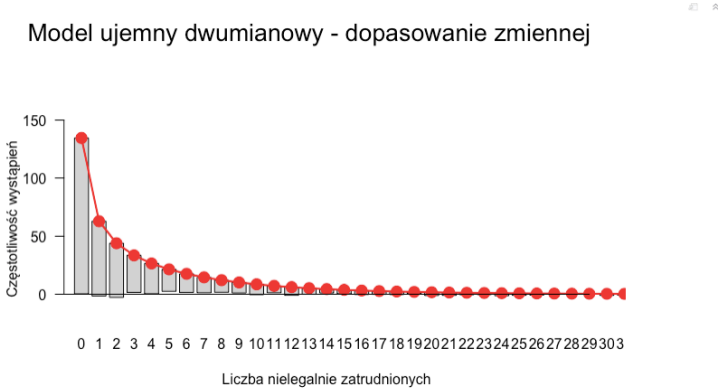
	Rozkład Poisson'a	Rozkład Pascal'a
χ^2	31788.91	468.35
Liczba stopni swobody	38	37
p-value	0	4.19e-76

Źródło: Opracowanie własne na podstawie danych z Państwowej Inspekcji Pracy

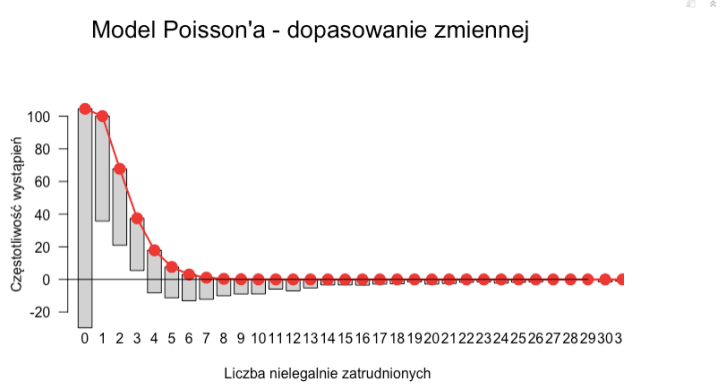
Oprócz wartości liczbowych możemy też zobaczyć jak wartości empiryczne pasują do wartości teoretycznych. Przy użyciu funkcji `rootogram` z pakietu `vcd` tworzymy wykres, który

przedstawia dane empiryczne oraz pokazuje wartości teoretyczne badanego rozkładu. Patrząc na wykresy (3.1, 3.2), zauważalne jest to, że rozkład ujemny dwumianowy jest bardziej dopasowany do danych. Natomiast rozkład Poisson'a odbiega od wartości empirycznych, widoczne jest to w początkowej fazie wykresu, gdzie wartości i częstości nie są w stanie dopasować się do teoretycznej linii rozkładu.

Rysunek 3.1. Wykres porównujący wartości teoretyczne i empiryczne dla rozkładu ujemnego dwumianowego



Rysunek 3.2. Wykres porównujący wartości teoretyczne i empiryczne dla rozkładu Poisson'a



Chcąc poznać charakterystykę przedsiębiorstw, które zatrudniają nielegalnie pracowników, musimy mieć na uwadze nielosowy charakter próby wynikający z tego w jaki sposób PIP dobiera przedsiębiorstwa do badania. Oznacza to, że w badaniu znajdują się obserwacje, których liczba nielegalnie zatrudnionych pracowników jest niezerowa.

Procedura badawcza przedstawiona w tym rozdziale składać się będzie z następujących kroków:

1. oszacowania modelu objaśniającego liczbę zatrudnionych nielegalnie bez uwzględnienia procesu selekcji,

2. oszacowaniu modelu objaśniającego liczbę zatrudnionych nielegalnie z uwzględnieniem procesu selekcji,

3. porównaniu obu modelu.

Dla obydwu modeli założony zostanie rozkład Poissona oraz ujemny dwumianowy celem wybrania najlepszego modelu.

3.2.2 Model bez uwzględnienia procesu selekcji

W tej części oszacujemy model nie uwzględniając selekcji. Dokonamy estymacji parametrów tylko dla równania wynikowego. Bierzemy pod uwagę modele dla rozkładu Poisson'a i ujemnego dwumianowego dla danych z 2016 roku. Estymowany model będzie miał następującą postać

$$y_i = g(\beta' X_i), \tag{3.1}$$

gdzie y_i określa liczbę nielegalnie zatrudnionych, X_i jest wektorem zmiennych objaśnianych zawierających informacje o sekcji oraz wielkości, a funkcja g to funkcja łącząca wykorzystywana w uogólnionym modelu liniowym. Do oszacowania modeli wykorzystano funkcję `glm` i `glm.nb` z pakietu R.

Przy pomocy funkcji `summary` w R dostajemy tabele 3.3, w której porównujemy dwa modele z rozważanymi rozkładami. Tabela 3.3 przedstawia zmienne, które są istotne na różnym poziomie oraz wartości oszacowanych parametrów. W nawiasach znajdują się błędy standardowe szacunku. Oznaczają one bardzo małą część odchyleń dopasowanego modelu, stąd im mniejsza wartość tym lepsze dopasowanie. Wartości błędów są porównywalnie małe dla obu rozkładów. Lepsze dopasowanie wykazuje rozkład Poisson'a, jednak miara błędów standardowych sprawdza się przy mało zróżnicowanych próbach, dlatego warto sprawdzić też statystyki opisowe dla reszt obu modeli.

Porównując przedstawione modele w większości znaki parametrów są takie same. W tej kwestii różnią się wartości dla sekcji dot. działalności profesjonalnej, naukowej i technicznej (sekcja M) oraz edukacji (sekcja P). Zmienna będąca istotna statystycznie na poziomie 0,1 dla obu modeli to sekcja opieki zdrowotnej i pomocy społecznej (sekcja Q). Dodatkowo model ujemny dwumianowy zawiera istotne zmienne odnoszące się do sekcji dostawy wody, gospo-

Tabela 3.3. Porównanie parametrów modelu bez selekcji dla rozkładu Poisson'a oraz ujemnego dwumianowego

	Zmienna zależna:	
	Liczba nielegalnych zatrudnień	
	Poisson	*Ujemny dwumianowy*
	(1)	(2)
sekcja C (przemysł)	−0.048 (0.057)	−0.028 (0.080)
sekcja E (ścieki i odpady)	−0.274*** (0.098)	−0.236* (0.133)
sekcja F (budownictwo)	0.054 (0.058)	0.072 (0.080)
sekcja G (handel)	−0.090 (0.057)	−0.078 (0.079)
sekcja H (transport)	0.016 (0.062)	0.024 (0.086)
sekcja I (gastronomia i zakwaterowanie)	0.059 (0.059)	0.070 (0.082)
sekcja J (informacja i komunikacja)	−0.108 (0.076)	−0.090 (0.105)
sekcja K (finanse i ubezpieczenia)	−0.032 (0.087)	−0.005 (0.120)
sekcja L (rynek nieruchomości)	−0.095 (0.082)	−0.073 (0.114)
sekcja M (nauka i technika)	−0.013 (0.065)	0.006 (0.090)
sekcja N (usługi administrowania)	0.359*** (0.060)	0.366*** (0.086)
sekcja P (edukacja)	−0.005 (0.070)	0.037 (0.099)
sekcja Q (pomoc społeczna)	0.122* (0.066)	0.161* (0.094)
sekcja R (kultura i rozrywka)	0.208** (0.081)	0.210* (0.118)
sekcja S (pozostałe usługi)	−0.085 (0.066)	−0.070 (0.091)
wielk 10-49	0.476*** (0.014)	0.475*** (0.019)
wielk 50-249	0.645*** (0.025)	0.644*** (0.038)
wielk 250+	0.657*** (0.039)	0.634*** (0.062)
Stała	0.764*** (0.056)	0.749*** (0.078)
Liczba obserwacji	9,299	9,299
Logarytm funkcji wiaryg.	−21,202.060	−18,663.570
θ		3.019*** (0.081)
AIC	42,442.130	37,365.140
Uwaga:		*p<0.1; **p<0.05; ***p<0.01

Źródło: Opracowanie własne.

darowania ściekami i odpadami oraz działalności związanej z rekultywacją (sekcja E) oraz działalność związana z kulturą, rozrywką i rekreacją (sekcja R).

Szczególnie wysokie wartości parametrów pojawiają się dla zmiennych odnoszących się do wielkości przedsiębiorstw. W przypadku rozkładu Poisson'a im większa liczba zatrudnionych pracowników tym więcej nielegalnych zatrudnień. Dla przedsiębiorstw powyżej 250 pracujących nielegalnie zatrudnionych jest o 66% więcej w zestawieniu z nielegalnymi pracownikami z małych przedsiębiorstw. Natomiast niewielka różnica wśród parametrów rozkładu dwu-

mianowego ujemnego dla zmiennej `wielk50_249` świadczy, że firmy mające zatrudnienie w przedziale od 50 do 249 zatrudniają o 64,5% więcej pracowników szarej strefy niż firmy zatrudniające do 9 pracowników.

Zarówno dla modelu Poisson'a oraz ujemnego dwumianowego zmienna odnosząca się do działalności w zakresie usług administrowania jest o ponad 35% większa od liczby nielegalnych pracowników w sekcji A (Rolnictwo). Największy spadek liczby nielegalnie zatrudnionych w odniesieniu do sektora rolnictwa i leśnictwa jest dla zmiennej dotyczącej gospodarowania ściekami i odpadami. Dla rozkładu Poisson'a zmienna jest mniejsza na poziomie 27%, a dla rozkładu ujemnego dwumianowego 24%. .

Dodatkowo tabela zawiera wartość kryterium informacyjnego Akaike (AIC). Pomoże nam ono wskazać, który z modeli jest lepiej dopasowany do danych. Im niższa wartość tego kryterium tym informacje wskazane przez model będą bardziej adekwatne. Wartość kryterium AIC jest mniejsza dla ujemnego modelu dwumianowego (Pascal'a), co świadczy, że model ten jest lepszy.

Dla statystyk reszt obu modeli możemy przeanalizować reszty umieszczone w Tabeli 3.4. Model Poisson'a ma reszty z znacząco większymi wartościami, szczególnie widać różnicę przy wartości maksymalnej (max). Jest ona prawie dwa razy większa, dlatego też mamy podstawy aby uznać, że model ujemny dwumianowy lepiej opisuje naszą zmienną.

Tabela 3.4. Statystyki opisowe reszt

	Model Poisson'a	Model ujemny dwumianowy
Min	-2.51	-1.67
1Q	-0.94	-0.71
Mediana	-0.74	-0.56
3Q	0.38	0.25
Średnia	-0.17	-0.18
Max	24.66	12.45

Źródło: Opracowanie własne.

Podsumowując na podstawie oszacowanego modelu nieuwzględniającego selekcji najlepszy jest ten zakładający rozkład ujemny dwumianowy.

3.3 Estymacja modelu Heckman'a

Tak samo jak w przypadku modelu bez selekcji dokonamy wyboru obserwacji, w których obecne jest zatrudnienie niezgodne z przepisami. Estymację wykonamy dla zarówno dla modelu z rozkładem Poisson'a i ujemnym dwumianowym. W tym celu stworzymy model Heckman'a gdzie występują dwa równania – selekcji i wynikowe. Układ równań ma następującą postać:

$$\begin{cases} \text{selekcja} = \text{sekcja} + \text{wielkość} + \text{województwo} + \text{skargi}, \\ \text{liczba nielegalnie zatrudnionych} = \text{sekcja} + \text{wielkość}. \end{cases} \tag{3.2}$$

W modelu (3.3) zmienna `selekcja` została zdefiniowana następująco

$$\text{selekcja} = \begin{cases} 1, & \text{jeżeli firma zatrudniała co najmniej jedną osobę nielegalnie}, \\ 0, & \text{w przeciwnym przypadku}. \end{cases} \tag{3.3}$$

Do wyjaśnienia selekcji dobrano zmienne `sekcja`, `wielkość`, `woj` oraz `skargi`, gdzie `woj` i `skargi` mają charakter zmiennych instrumentalnych. Natomiast do wyjaśnienia liczby nielegalnie zatrudnionych wykorzystano zmienną `sekcja` (PKD) oraz `wielkość` (firmy).

W dalszej części oszacowano dwa modele dla 2010 i 2016 roku celem porównania czy związki między badanymi zmiennymi mają charakter stały.

3.3.1 Wyniki badania

W tym podrozdziale przedstawione zostaną wyniki estymacji modelu selekcji dla danych z roku 2010 oraz 2016. Do szacowania wybrano model nawiązujący do ujemnego rozkładu dwumianowego. W tabeli 3.7 i 3.6 przedstawiono wartości kryteriów AIC oraz BIC dla oszacowanych obu rozkładów. Tabele nawiązują do 2016 roku, natomiast dla 3.5 i 3.8 wyniki odnoszą się do obserwacji z 2010 roku.

Tabela 3.5. Kryterium AIC dla modeli z 2010 roku

	df	AIC
R. Poisson'a	55.00	61906.03
R. ujemny dwumianowy	56.00	59982.99

Źródło: Opracowanie własne.

Tabela 3.6. Kryterium BIC dla modeli z 2010 roku

	df	BIC
R. Poisson'a	55.00	62352.03
R. ujemny dwumianowy	56.00	60437.10

Źródło: Opracowanie własne.

Tabela 3.7. Kryterium AIC dla modeli z 2016 roku

	df	AIC
R. Poisson'a	55.00	72182.80
R. ujemny dwumianowy	56.00	70384.06

Źródło: Opracowanie własne.

Tabela 3.8. Kryterium BIC dla modeli z 2016 roku

	df	BIC
R. Poisson'a	55.00	72634.65
R. ujemny dwumianowy	56.00	70844.12

Wartości kryteriów wskazują na lepsze dopasowanie modelu oszacowanego przy pomocy rozkładu ujemnego dwumianowego. Estymacja zostanie przedstawiona dla właśnie tego modelu.

Do estymacji użyto funkcji z pakietu GRJM. Zapis w języku R znajduje się w sekcji dotyczącej implementacji A.5. Wyniki równania selekcji dla modelu dla 2010 roku oraz 2016 znajdują się w tabeli 3.9.

Uzyskane wyniki 3.9 otrzymano dzięki funkcji `summary` w R. Odnoszą się one do równania selekcji, którego interpretacja jest taka sama jak przy parametrach uogólnionego modelu liniowego. W przypadku obu modeli wśród obserwacji wyłączono te sekcje, których wartości odróżniały się najbardziej w porównaniu do reszty. Dla roku 2010 oraz 2016 były to sekcje dotyczące górnictwa i wydobywania (sekcja B), wytwarzania energii elektrycznej, gazu i innych (sekcja D) oraz administracja publiczna i ubezpieczenia społeczne (sekcja O). Dodatkowo w 2010 roku z zebranych danych odrzucono obserwacje dla gospodarstw domowych oraz gospodarstw, które produkują własne wyroby i świadczą usługi (sekcja T).

Zmienna objaśniana w równaniu selekcji przyjmuje wartości 1 bądź 0. Ekonomiczna interpretacja oznacza decyzję danego przedsiębiorstwa aby zatrudniać kogoś nielegalnie. W kontekście danych z PIP oznacza to, że dla tej firmy stwierdzono nielegalne zatrudnienie. Dla modelu pierwszego 3.9 wartość równania selekcji najbardziej istotne zmienne odnoszą się do woje-

Tabela 3.9. Wartości parametrów dla równania selekcji z modelu Heckman'a dla 2010 i 2016 roku

	Zmienna objaśniana:	
	selekcja	
	(2010)	(2016)
Wyraz wolny	−0.838*** (0.067)	−0.825*** (0.069)
sekcja C (przemysł)	−0.005 (0.066)	0.022 (0.068)
sekcja E (ścieki i odpady)	0.238* (0.113)	−0.064 (0.111)
sekcja F (budownictwo)	0.082 (0.065)	0.081 (0.068)
sekcja G (handel)	0.087 (0.063)	0.110 (0.066)
sekcja H (transport)	0.378*** (0.075)	0.289*** (0.075)
sekcja I (gastronomia i zakwaterowanie)	0.335*** (0.069)	0.264*** (0.071)
sekcja J (informacja i komunikacja)	0.066 (0.119)	0.268** (0.095)
sekcja K (finanse i ubezpieczenia)	0.324** (0.120)	0.197· (0.107)
sekcja L (rynek nieruchomości)	0.148 (0.111)	0.152 (0.100)
sekcja M (nauka i technika)	0.213* (0.086)	0.099 (0.077)
sekcja N (usługi administrowania)	0.208** (0.079)	0.261*** (0.077)
sekcja P (edukacja)	−0.029 (0.093)	−0.003 (0.087)
sekcja Q (pomoc społeczna)	0.380*** (0.097)	0.181* (0.085)
sekcja R (kultura i rozrywka)	0.036 (0.111)	−0.061 (0.105)
sekcja S (pozostałe usługi)	0.156 (0.080)	0.044 (0.076)
wielk 10-49	0.157*** (0.019)	0.214*** (0.019)
wielk 50-249	0.059 (0.034)	0.122** (0.039)
wielk 250+	−0.176** (0.061)	0.267*** (0.073)
woj KUJAWSKO-POMORSKIE	0.120** (0.039)	-0.013 (0.042)
woj LUBELSKIE	0.130** (0.042)	0.033*** (0.041)
woj LUBUSKIE	−0.362*** (0.065)	0.243*** (0.051)
woj ŁÓDZKIE	0.067 (0.036)	0.311*** (0.039)
woj MAŁOPOLSKIE	0.385*** (0.039)	0.565*** (0.041)
woj MAZOWIECKIE	−0.115** (0.044)	0.002 (0.046)
woj OPOLSKIE	0.096* (0.049)	0.042 (0.046)
woj PODKARPACKIE	0.162*** (0.040)	0.206*** (0.039)
woj PODLASKIE	-0.014 (0.046)	−0.160*** (0.044)
woj POMORSKIE	0.059 (0.040)	−0.082* (0.039)
woj ŚLĄSKIE	0.379*** (0.033)	0.575*** (0.037)
woj ŚWIĘTOKRZYSKIE	0.308*** (0.050)	0.417*** (0.047)
woj WARMIŃSKO - MAZURSKIE	−0.044 (0.052)	0.095* (0.048)
woj WIELKOPOLSKIE	0.420*** (0.034)	0.477*** (0.037)
woj ZACHODNIOPOMORSKIE	0.490*** (0.045)	0.549*** (0.047)
Liczba skarg	0.076*** (0.011)	0.088*** (0.009)
Liczba obserwacji	24,566	27,318
Log funkcji wiarygodności	−14,673.620	−16,761.010
Uwagi:		*p<0.1; **p<0.05; ***p<0.01

wództwa opolskiego. Skargi okazały się istotne w doborze selekcji. Natomiast firmy zatrudniające od 50 do 249 osób nie mają istotnego wpływu na opis zjawiska selekcji.

Większość sekcji PKD okazało się istotnych dla roku 2010 i 2016. W pierwszym modelu między innymi działalność w zakresie usług administrowania, działalność naukowa i techniczna, ubezpieczeniowa i finansowa oraz sekcja dostawy wody i gospodarowania ściekami i odpadami. W modelu selekcji szacowanym dla rozkładu ujemnego dwumianowego z 2016 roku istotne okazały się sekcje transportu i gospodarki magazynowej, działalności w strefie gastronomii, informacji i komunikacji oraz opieki społecznej czy administrowania.

W odniesieniu się do sekcji A (rolnictwo i leśnictwo) mniejsza szansa na występowanie zjawiska zatrudnienia w szarej strefie w 2016 roku dotyczy obszarów edukacji, gospodarowania ściekami oraz kultury i rekreacji. Największa wartość parametru to sekcja H, czyli w przypadku transportu i gospodarki magazynowej możemy się spodziewać zatrudnień ”na czarno” o 28,9% więcej w porównaniu do rolnictwa.Również odnosząc się do sekcji A działalność gastronomii i zakwaterowania była większa o 26,4%, pomoc społeczna o 18,1% oraz usługi administrowania o 26,1%.

Dla 2010 roku wartości parametrów są podobne. Tym razem przemysł i edukacja okazały się mniejsze pod względem nielegalnych zatrudnień w porównaniu do sekcji leśnictwa i rolnictwa. Spośród sekcji, w których praca ”na czarno” występowała częściej niż w sekcji A to działalność dotycząca nauki i techniki (o 21,3 %), transport (o 37,8 %) czy też gastronomia i zakwaterowanie (o 33,5 %).

Podobnie jak we wszystkich wcześniejszych modelach wielkość zatrudnienia ma znaczenie. W firmach powyżej 250 zatrudnionych szara strefa może pojawiać się rzadziej, niż w porównaniu do firm zatrudniających do 9 pracowników (o około 18 %). Oprócz województwa warmińsko-mazurskiego i podlaskiego również w mazowieckim i małopolskim istnieje mniejsze prawdopodobieństwo nielegalnych zatrudnień w odniesieniu do województwa dolnośląskiego. Najwyższa szansa istnieje w województwie małopolskim i śląskim (w porównaniu do dolnośląskiego).

Istotność statystyczna parametrów dla obu równań jest bardzo podobna. Natomiast w 2016 roku więcej zmiennych ma dodatni parametr, co oznacza, że zmniejszyła się liczba sekcji PKD gdzie szansa na nielegalne zatrudnienie maleje. Zjawisko spadku liczby sekcji gdzie nielegalne zatrudnienie spadło może wiązać się ze spadkiem stopy bezrobocia, która w ostatnich latach maleje. Możliwe jest również wytłumaczenie wzrostu pracujących cudzoziemców.

Tabela 3.10. Wartości parametrów dla równania liczby nielegalnie zatrudnionych z modelu Heckman'a dla 2010 i 2016 roku

	Zmienna zależna:	
	liczba nielegalnie zatrudnionych	
	(2010)	(2016)
Wyraz wolny	1.709^{***} (0.080)	$1{,}315^{***}$ (0.086)
sekcja C (przemysł)	-0.125 (0.080)	-0.070 (0.069)
sekcja E (ścieki i odpady)	-0.237^{*} (0.074)	-0.230^{*} (0.114)
sekcja F (budownictwo)	0.097 (0.115)	-0.478 (0.069)
sekcja G (handel)	-0.269^{***} (0.074)	-0.136^{*} (0.067)
sekcja H (transport)	-0.327^{***} (0.082)	-0.136 (0.076)
sekcja I (gastronomia i zakwaterowanie)	-0.310^{***} (0.078)	-0.044 (0.072)
sekcja J (informacja i komunikacja)	-0.132 (0.129)	-0.180^{*} (0.091)
sekcja K (finanse i ubezpieczenia)	-0.324^{**} (0.125)	-0.083 (0.105)
sekcja L (rynek nieruchomości)	-0.463^{***} (0.127)	-0.132 (0.098)
sekcja M (nauka i technika)	-0.284^{**} (0.093)	-0.075 (0.078)
sekcja N (usługi administrowania)	-0.020 (0.085)	0.133 (0.076)
sekcja P (edukacja)	-0.288^{*} (0.121)	-0.050 (0.086)
sekcja Q (pomoc społeczna)	-0.380^{***} (0.100)	-0.096 (0.083)
sekcja R (kultura i rozrywka)	-0.195 (0.121)	0.032 (0.114)
sekcja S (pozostałe usługi)	-0.257^{**} (0.089)	-0.124 (0.078)
wielk 10-49	0.413^{***} (0.020)	0.355^{***} (0.018)
wielk 50-249	0.501^{***} (0.037)	0.479^{***} (0.034)
wielk 250+	0.688^{**} (0.064)	0.249^{***} (0.060)
Liczba obserwacji	7,460	9,299
Log funkcji wiarygodności	$-5{,}994.842$	$-6{,}930.284$
Parametr θ	-0.698	-0.535
Uwagi:		*p$<$0.1; **p$<$0.05; ***p$<$0.01

W tabeli 3.10 znajdują się wartości parametrów dla równania wynikowego z 2010 i 2016 roku. Zarówno w 2010 i 2016 roku sekcja gospodarowania odpadami okazała się istotna. Oprócz tego w 2010 roku wpływ na liczbę nielegalnych zatrudnień ma również działalność finansowa i ubezpieczeniowa, edukacja czy też pozostała działalność usługowa. Jednak wszystkie wartości parametrów sekcji PKD mają wartości ujemne. Świadczy to o mniejszej liczbie pracowników zatrudnionych nielegalnie w porównaniu do sekcji rolnictwa i leśnictwa. Spadek możemy zaobserwować dla zmiennej dotyczącej działalności administracyjnej (sekcja N), w odniesieniu do rolnictwa liczba nielegalnie zatrudnionych spadnie o 0,02%. Nawiązując do sekcji A i porównując ją do największej różnicy wśród zbadanych sektorów PKD to działalność rynku nieruchomości (sekcja L) wykazuje o 46% mniej nielegalnie zatrudnionych.

Również jak w poprzednich szacowaniach wielkość przedsiębiorstw oceniana w liczbie zatrudnionych ma znaczenie. W szczególności gospodarka nieformalna częściej spotykana jest wśród dużych firm (powyżej 250 pracowników). W odniesieniu do małych firm (do 9 pracujących) firmy zatrudniające powyżej 250 osób zatrudniają nielegalnie o 68% więcej. Najmniejszy wzrost istnieje dla przedsiębiorstw małych (zatrudniających od 10 do 49 pracowników), wynosi około 41%.

Podobnie jak w równaniu wynikowym dla modelu z 2010 roku większość wartości parametrów jest ujemna. Sekcje R i N mają dodatnie parametry. Oznacza to, że w porównaniu z rolnictwem (sekcja A) dla działalności związanej z administrowaniem i sektorem kultury i rozrywki możemy się spodziewać wzrostu liczby zatrudnień ”na czarno” odpowiednio o 13 % i 3 %. Jeśli chodzi o największy spadek liczby zatrudnień to w porównaniu do sektora rolnictwa w 2016 roku zachodził on w obszarze budownictwa (mniej o 48 %) .

Natomiast w przedsiębiorstwach zatrudniających pomiędzy 50 a 250 pracowników szara strefa może pojawiać się częściej. Liczba nielegalnych zatrudnień dla firm zatrudniających pomiędzy 50 a 249 pracowników, w odniesieniu do małych przedsiębiorstw, jest większa o 48 %.

Wartość parametru θ dla obu modeli ma wartości ujemne. Sama wartość to szacowany parametr zależności, oznacza to siłę połączenia równania wynikowego i selekcji. Ujemna wartość świadczy, że błędy w równaniach są ujemnie skorelowane (Wojtys i in., 2016). Oznacza to, że jeżeli dana firma zatrudnia pracowników nielegalnie to zwykle zatrudnia ich niewielką liczbę.

Podsumowanie

Celem pracy było zbadanie i zapoznanie się z charakterystyką przedsiębiorstw, które dokonywały zatrudnień nieformalnych. Zbadano również to jakie firmy decydują się na działanie w gospodarce nieformalnej. Dodatkowo zweryfikowano jak zjawisko pracy w szarej strefie zmieniło się na przestrzeni 2010 i 2016 roku.

Dane, które zostały wykorzystane do przeprowadzonego badania pochodzą z Państwowej Inspekcji Pracy, co świadczy o ich obiektywnym charakterze. Zbiór dotyczy obserwacji dla lat od 2010 do 2016 roku. Zgodnie z wiedzą autorki oraz Promotora pracy, dra Macieja Beręsewicza jest to jedno z pierwszych badań, do których te dane zostały wykorzystane.

W analizie wykorzystano uogólniony model Heckman'a który składał się z dwóch równań. Pierwsze z nich, równanie selekcji, odnosiło się do podjęcia decyzji badanej firmy odnośnie zatrudnienia nielegalnie. Natomiast równanie wynikowe określało charakterystykę przedsiębiorstw, które już działały w gospodarce nieformalnej.

W początkowym etapie analizy dokonano szacowania modelu bez części odnoszącej się do selekcji. Wyniki, które uzyskano mają wyższe parametry w porównaniu do wartości dla modelu z selekcją. Dla sekcji działalności w zakresie administrowania w modelu bez selekcji liczba nielegalnie zatrudnionych była o 36% większa niż w sektorze wykorzystanym do porównania (sekcja rolnictwa i leśnictwa). Natomiast dla modelu z selekcją to porównanie wskazywało na 13%.

W przeprowadzonym badaniu równanie wynikowe sugeruje, że pozostałe sekcje nie różnią się istotnie w porównaniu z sekcją leśnictwa i rolnictwa. W 2010 roku najczęściej na nielegalne zatrudnienia decydowali się przedsiębiorcy z województwa małopolskiego z sekcji transportu, bądź pomocy społecznej. W 2016 roku również większość firm rozpoczynających nielegalne działania pochodziło z województwa małopolskiego bądź śląskiego z sekcji transportu.

Natomiast dla 2010 roku firmy działające najczęściej w szarej strefie pochodziły z sektora leśnictwa i rolnictwa oraz budownictwa. W 2016 roku obszar budownictwa był tym, który ten

odniósł największy spadek odnośnie sekcji A.

W wyniku badania otrzymano wartość θ, która wskazuje, że jeśli firma decyduje się na nielegalne zatrudnienie, zazwyczaj zatrudnia niewielką liczbę osób. Wśród otrzymanych wyników spośród wszystkich modeli, zauważalne jest to, że większe przedsiębiorstwa częściej postanawiają zatrudnić nielegalnie.

Ujemne wartości parametrów dla zmiennych odnoszących się do sekcji PKD wskazują na znaczącą przewagę sektora leśnictwa i rolnictwa. Jednak parametr θ zmalał na przełomie 6 lat. Powodem może być malejąca stopa bezrobocia, bądź większe zatrudnienie przybywających cudzoziemców.

Wyniki, które zostały przedstawione mogłyby udoskonalić działania związane z redukcją pracy nielegalnej. Lepsze poznanie charakterystyki firm, mogłoby zwiększyć skuteczność organów kontrolujących to zjawisko/ Spadek nielegalnych zatrudnień przyczyniłby się do rozwoju gospodarczego jak i wpłynąłby korzystnie na społeczeństwo, szczególnie tą cześć, która pracuje w szarej strefie. Wykorzystane dane mogą jednak z pewnością posłużyć do innych analiz, które ukażą nowe perspektywy rynku pracy.

Bibliografia

Łapiński, K., Peterlik, M. & Wyżnikiewicz, B. (2014). Szara strefa w polskiej gospodarce. *IBnGR, Warszawa*.

Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics*, *24*(1-2), 3–61.

Drabek, A. (2012). *Nielegalne zatrudnienie w prawie polskim*. Wolters Kluwer.

Feige, E. L. (2007). *The underground economies: Tax evasion and information distortion*. Cambridge University Press.

Fundowicz, J., Łapiński, K. & Wyżnikiewicz, B. (2018). Szara strefa 2018. *Instytut Prognoz i Analiz Gospodarczych, Warszawa*.

Główny Urząd Statystyczny. (2004). Praca nierejestrowana - Główny Urząd Statystyczny. https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/3896,pojecie.html/.

Główny Urząd Statystyczny. (2013). Rachunki narodowe według sektorów i podsektorów instytucjonalnych 2008–2011. Warszawa.

Główny Urząd Statystyczny. (2019). Praca nierejestrowana w Polsce w 2017 r., 13–14.

Giampiero, M. & Rosalba, R. (2019). The GJRM package, 28–32.

Gołębiowski, G. (2007). Zjawisko szarej strefy z uwzględnieniem gospodarki polskiej. *Współczesna Ekonomia*, *1*(1), 17–28.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *NYU working paper*, 4–7.

Greene, W. H. (1995). Sample selection in the Poisson regression model. *NYU Working Paper*.

Gruszczyński, M. (2010). *Mikroekonometria: modele i metody analizy danych indywidualnych*. Oficyna a Wolters Kluwer business.

Gutmann, P. M. (1977). The subterranean economy. *Financial Analysts Journal*, *33*(6), 26–27.

Halekoh, U., Højsgaard, S., Yan, J. i in. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software*, *15*(2), 1–11.

Jarocka, M. S. (2011). Analiza wybranych metod bezpośrednich i pośrednich służących o badania szarej strefy. *Zeszyty Naukowe PWSZ w Płocku. Nauki Ekonomiczne*, 31–43.

Kubiczek, A. (2010). Gospodarka nieformalna jako wyraz zawodności państwa. *Ekonomia i Prawo*, *6*, 361–370.

Meyer, D., Zeileis, A., Hornik, K., Meyer, M. D. & KernSmooth, S. (2007). The vcd package. *Retrieved October*, *3*, 2007.

Mróz, B. (2012). Konsumenci i gospodarstwa domowe na nieformalnym rynku pracy w Polsce. *Konsumpcja i rozwój*, (1 (2)), 24–35.

Organisation for Economic Co-operation and Development. (2001). *Underground Production*. https://stats.oecd.org/glossary/detail.asp?ID=2789.

Państwowa Inspekcja Pracy. (2018). Sprawozdanie z działalności Państwowej Inspekcji Pracy w 2016 roku, 99–132.

PARP. (2017). Bilans kapitału ludzkiego.

Pater, K. (2007). Przyczyny pracy nierejestrowanej, jej skala, charakter i skutki społeczne. *Ministerstwo Pracy i Polityki Społecznej, Warszawa*.

PIP. (2016). Skutki ”szarej strefy” dla budżetu państwa. https://www.pip.gov.pl/pl/rzecznik-prasowy/komunikaty-biezace/73016,skutki-szarej-strefy-dla-budzetu-panstwa-.html.

Strawiński, P. (2007). Przyczynowość, selekcja i endogeniczne oddziaływanie. *Przegląd Statystyczny*, *4*, 49–61.

Szreder, M. (1994). *Informacje a priori w klasycznej i bayesowskiej estymacji modeli regrecji*. Wydawn. Uniwersytetu Gdańskiego.

Szreder, M. (2010). Losowe i nielosowe próby w badaniach statystycznych. *Przegląd Statystyczny*, *57*(4), 168–174.

Szreder, M. & Krzykowski, G. (2005). Znaczenie informacji spoza próby w badaniach statystycznych. *Prace i Materiały Wydziału Zarządzania Uniwersytetu Gdańskiego*, (1), 157–168.

Toomet, O., Henningsen, A. i in. (2008). Sample selection models in R: Package sampleSelection. *Journal of statistical software*, *27*(7), 1–23.

Vanderseypen, G., Tchipeva, T., Peschner, J., Rennoy, P. & Williams, C. C. (2013). Undeclared work: recent developments. *Employment and Social Developments in Europe*, *2013*, 231–274.

Walker, A. (2015). openxlsx: Read, Write and Edit XLSX Files. R package version 3.0. 0.

Wickham, H., Francois, R., Henry, L. & Müller, K. (2015). dplyr: A Grammar of Data Manipulation. R package version 0.4. 3. *R Found. Stat. Comput., Vienna. https://CRAN. R-project. org/package= dplyr*.

Wołodźko, T. (2015). *Modele cech ukrytych w badaniach edukacyjnych, psychologii i socjologii: teoria i zastosowania*. Instytut Badań Edukacyjnych.

Wojtys, M., Marra, G. & Radice, R. (2016). Copula regression spline sample selection models: the R Package SemiParSampleSel. *Journal of Statistical Software*, *71*(6).

Zeileis, A., Kleiber, C. & Jackman, S. (2008). Regression models for count data in R. *Journal of statistical software*, *27*(8), 5.

Spis tabel

1.1 Szacunki udziałów szarej gospodarki w tworzeniu PKB w latach 2008–2011 (w proc.) według sekcji PKD 2007 6

1.2 Opinie Polaków na temat przyczyn podejmowania pracy nierejestrowanej w 2010r. (w %) 9

1.3 Metody pomiaru szarej strefy - zalety i wady 13

3.1 Statystyki opisowe dla liczby nielegalnie zatrudnionych w 2010 i 2016 r. 32

3.2 Test χ^2 dopasowania rozkładu Poissona i ujemnego dwumianowego 32

3.3 Porównanie parametrów modelu bez selekcji dla rozkładu Poisson'a oraz ujemnego dwumianowego 35

3.4 Statystyki opisowe reszt 36

3.5 Kryterium AIC dla modeli z 2010 roku 37

3.6 Kryterium BIC dla modeli z 2010 roku 38

3.7 Kryterium AIC dla modeli z 2016 roku 38

3.8 Kryterium BIC dla modeli z 2016 roku 38

3.9 Wartości parametrów dla równania selekcji z modelu Heckman'a dla 2010 i 2016 roku 39

3.10 Wartości parametrów dla równania liczby nielegalnie zatrudnionych z modelu Heckman'a dla 2010 i 2016 roku 41

Spis rysunków

1.1 Odsetek osób wykonujących pracę w ramach umowy nieformalnej wśród aktywnych zawodowo 14
1.2 Nielegalne zatrudnienie lub nielegalna inna praca zarobkowa - wg województw 15

3.1 Wykres porównujący wartości teoretyczne i empiryczne dla rozkładu ujemnego dwumianowego 33
3.2 Wykres porównujący wartości teoretyczne i empiryczne dla rozkładu Poisson'a 33

Dodatek A

Spis programów

A.1 Skrypty wykorzystane do przygotowania danych

W tej części dokonamy implementacji modelu selekcji Heckman'a w pakiecie R. Jest to pełen zapis przeprowadzonego badania. Uzyskane wynik zostaną zinterpretowane w dalszej części rozdziału.

Zaczynamy od wczytania potrzebnych pakietów, które są odpowiedzialne za niektóre funkcje.

```
library(sampleSelection)
library(haven)
library(tidyverse)
library(SemiParSampleSel)
library(openxlsx)
library(vcd)
install.packages(gjrm)
library(texreg)
library(xtable)
```

Program A.1. Wczytywanie pakietów

Następnie wczytujemy potrzebne dane w formacie .xlsx za pomocą funkcji read.xlsx z pakietu `openxlsx` (Walker, 2015). Wybieramy dane dla 2016 i 2010 roku, dokonujemy kilku czyszczeń naszych danych, usuwamy sekcje, które miały najmniejszą częstotliwość wystąpień, dokonujemy przeskalowania skarg. Dodajemy nową zmienną *selekcja*, która przyjmuje wartość 1 kiedy występuje nielegalne zatrudnienie, w przeciwnym wypadku 0. Funkcje umożliwiające przygotowanie danych pochodzą z pakietu dplyr (Wickham, Francois, Henry & Müller, 2015).

```

## Wczytanie danych
data <- read.xlsx("dane-jednostkowe-razem.xlsx")
head(data)

```

```
## Zapisanie danych w postaci data.frame
data <- as.data.frame(data)

##Sprawdzenie czestotliwosci wystepowania zmiennej okreslajacej sekcje PKD
data %>%
  filter(rok==2010) %>%
  group_by(sekcja) %>%
  count()

data %>%
  filter(rok==2016) %>%
  group_by(sekcja) %>%
  count()

## Utworzenie zbiorów dla 2010 i 2016 roku

data_2010 <- data %>%
  filter(rok == 2010) %>%
  mutate(sel = ifelse(liczba_niel == 0, 0, 1)) %>%
  filter(!sekcja %in% c(NA, "T", "O", "D", "B"),
         liczba_niel<=100) %>%
  mutate(skargi = scale(skargi),
         wielk = factor(wielk, c("do 9", "10-49", "50-249","250+")))

data_2016 <- data %>%
  filter(rok == 2016) %>%
  mutate(sel = ifelse(liczba_niel == 0, 0, 1)) %>%
  filter(!sekcja %in% c(NA, "B","D","O")) %>%
  mutate(skargi = scale(skargi),
         wielk = factor(wielk, c("do 9", "10-49", "50-249","250+")))
```

Program A.2. Załadowanie danych do środowiska R

A.1.1 Skrypty wykorzystane do estymacji modeli

Dokonujemy sprawdzenia dopasowania naszych danych do rozkładu Poisson'a oraz ujemnego dwumianowego. Przy użyciu funkcji `goodfit` sprawdzono dopasowanie danych empirycznych do danych teoretycznych. Wyniki dodatkowo zostaną przedstawione na wykresie dzięki funkcji `rootogram` z pakietu `vcd` (Meyer i in., 2007).

```

## Oszacowanie dopasowania
poisson_gf <-goodfit(data_2016$liczba_niel, type = "poisson")
nb_gf <- goodfit(data_2016$liczba_niel, type = "nbinom")

 ## Wizualizacja
rootogram(poisson_gf, main="Model Poisson'a - dopasowanie zmiennej",
  xlab="Liczba nielegalnie zatrudnionych", ylab="śćCzestotliwo wystapien",
     xlim=c(0, 30), ylim = c(-20,100))

rootogram(nb_gf, main="Model ujemny dwumianowy - dopasowanie zmiennej",
  xlab="Liczba nielegalnie zatrudnionych", ylab="Czestotliwosc ąwystpien",
     xlim=c(0, 30), ylim = c(-20,150))
```

Program A.3. Dopasowanie rozkładów

Kolejnym krokiem jest zbudowanie modelu bez selekcji. Do zapisu modelu wykorzystano funkcję `glm`, która jest przeznaczona dla uogólnionych modeli liniowych z pakietu o takiej sa-

mej nazwie (Halekoh, Højsgaard, Yan i in., 2006). Dodatkowo wyznaczono wartość kryteriów informacyjnych dla estymowanych modeli.

```
##Stworzenie modelu dla rozkladu Poisson'a
m0.1 <-glm(liczba_niel ~  sekcja + wielk,
             data = data_2016,
           subset = selekcja == 1,
         family = poisson())

##Stworzenie modelu dla rozkladu ujemnego dwumianowego
m0.2 <-glm.nb(liczba_niel ~  sekcja + wielk,
             data = data_2016,
           subset = selekcja == 1)

##Podsumowanie uzyskanych parametrów
summary(m0.1)
summary(m0.2)

##Wyznaczenie reszt oraz ich statystyk
reszty_01 <- summary(residuals.glm(m0.1))
reszty_02 <- summary(residuals.glm(m0.2))

## Dokonanie porównania modeli wzgledem kryteriów informacyjnych
aic0 <- AIC(m0.2, m0.1)
bic0 <- BIC(m0.2, m0.1)
```

Program A.4. Model bez selekcji - estymacja parametrów

Po skonstruowaniu modelu bez selekcji przechodzimy do modelu z selekcją. W tym przypadku zbadamy modele dla 2010 i 2016 roku dla obu rozkładów. Dla rozkładu Poisson'a użyto funkcji z pakietu SemiParSampleSel (Toomet, Henningsen i in., 2008). Model dla rozkładu ujemnego dwumianowego został zbudowany wykorzystując pakiet GJRM (Giampiero & Rosalba, 2019).

```

##Model selekcji dla rozkladu Poissona - rok 2010

m2_10 <- SemiParSampleSel(
  formula = list(selekcja ~ sekcja + wielk + skargi + woj,
                 liczba_niel ~ sekcja + wielk),
  data = data_2010,
  margin = c("probit","P")
)

##Model selekcji dla rozkladu Poissona - rok 2016

m2_16 <- SemiParSampleSel(
  formula = list(selekcja ~ sekcja + wielk + woj + skargi,
                 liczba_niel ~ sekcja + wielk),
  data = data_2016,
  margin = c("probit","P")
)


##Model dla rozkladu ujemnego dwumianowego dla 2010 roku

m3_10_2 <- GJRM::gjrm(
  formula = list(selekcja ~ sekcja + wielk + woj + skargi,
                 liczba_niel ~ sekcja + wielk),
  data = data_2016,
```

```
  margin = c("probit","NBI"),
  Model = "BSS"
)

##Model dla rozkladu ujemnego dwumianowego dla 2016 roku
m3_16_2 <- GJRM::gjrm(
  formula = list(selekcja ~ sekcja + wielk + woj + skargi,
                 liczba_niel ~ sekcja + wielk),
  data = data_2016,
  margin = c("probit","NBI"),
  Model = "BSS"
)

##Podsumowanie oszacowanych modeli
summary(m3_10)
summary(m3_16)

##Porównanie kryteriów AIC i~BIC
aic1 <- AIC(m2_10, m3_10_2)
bic1 <- BIC(m2_10, m3_10_2)

## oraz dla 2016 roku
aic1 <- AIC(m2_16, m3_16_2)
bic1 <- BIC(m2_16, m3_16_2)
```

Program A.5. Konstrukcja modeli z selekcją