Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Michał Sękowski

Determinanty zarobków netto w Polsce

Praca zaliczeniowa na ćwiczenia
z Ekonometrii prowadzone
przez dr inż. Janusza Gajdę

Warszawa, luty 2021

Abstrakt

Poniższa praca bada determinanty zarobków netto Polaków dla danych z 2017 roku. Celem jest zweryfikowanie prawdziwości hipotez znalezionych w literaturze, a także hipotez autora. Dzięki zbudowaniu modelu ekonometrycznego metodą regresji liniowej otrzymano oszacowania parametrów mających istotny wpływ na wysokość zarobków. Do takich należy zaliczyć płeć, doświadczenie, poziom wykształcenia, stan cywilny, znajomość języka obcego, miejsce zamieszkania, obejmowane kierownicze stanowisko, fakt pracy zagranicą oraz interakcję pomiędzy płcią i małżeństwem oraz między płcią i stażem.

Słowa kluczowe: analiza ekonometryczna, model regresji liniowej, determinanty dochodów, luka płacowa, marriage premium, wykształcenie

Spis treści

Wstęp

Determinanty dochodów to interesujący temat i obiekt badań wielu analiz ekonometrycznych. Klasyczna już funkcja zarobków Mincera – badająca zależność między dochodem a edukacją i stażem – okazała się użyteczna również przy badaniu bardziej skomplikowanych zagadnień. Szybko zauważono, że modyfikacja tej funkcji o inne zmienne pozwala zbadać takie zjawiska jak luka płacowa między płciami, efekt premii wynikającej ze związku małżeńskiego, wpływ wykształcenia rodziców itp. Badanie przy użyciu modelu ekonometrycznego ma niepodważalną zaletę – pozwala zbadać wpływ pewnych czynników ceteris paribus. Jest to szczególnie ważne przy badaniu np. wspomnianej luki płacowej, ponieważ interesuje nas różnica w zarobkach wynikająca z samej różnicy płci, przy zachowaniu innych parametrów na tym samym poziomie. Przy oszacowaniu dochodów istotny jest również szereg innych czynników, takich jak wielkość miejscowości zamieszkania, znajomość języka obcego etc.

Praca ma następującą strukturę:
Pierwszy rozdział poświęcony jest przeglądowi istniejącej literatury dotyczącej ekonometrycznej analizy determinantów dochodów.
Drugi rozdział przedstawia treść hipotez, które zostaną zweryfikowane.
Trzeci rozdział stanowi opis danych i zmiennych, które posłużą przy konstrukcji modelu.
Czwarty rozdział pokazuje konstrukcję modelu i sprawdzenie spełnienia założeń KMRL; oraz weryfikację hipotez w oparciu o ostateczną wersję modelu.
Praca kończy się powtórzeniem najważniejszych wniosków.

1. Przegląd literatury

Próby wskazania determinantów zarobków podejmowali się teoretycy jeszcze kilka wieków temu. Prawdziwy postęp w tej kwestii dokonał się jednak w XX wieku, wraz z przyjęciem przez badaczy bardziej empirycznego podejścia, równolegle z postępem narzędzi ekonometrycznych. Publikacje i odkrycia Jacoba Mincera – naukowca urodzonego w Polsce, nazywanego czasem ojcem współczesnej ekonomiki pracy – w pewnym sensie stanowiły fundament dla dalszych prac nad zbadaniem determinantów dochodów.

Artykuł J. Mincera z 1958 roku "Investment in Human Capital and Personal Income Distribution" można podsumować w ten sposób:
Edukacja stanowi kosztowną inwestycję w kapitał ludzki. Osoba decydująca się na to, opóźnia swoje wejście na rynek pracy i skraca ilość lat jaką łącznie w życiu przepracuje. Aby więc decyzja o dalszej edukacji lub natychmiastowym wejściu na rynek pracy sprowadzała się jedynie do preferencji danej jednostki, wartość teraźniejsza (PV) życiowych zarobków przy obu opcjach musi być sobie ekwiwalentna. W związku z tym, w zawodach, do których wymaga się odpowiednio dłuższego treningu/procesu edukacyjnego, zarobki będą odpowiednio wyższe. Z kolei różnice w zarobkach wewnątrz tego samego zawodu da się wyjaśnić, rozszerzając pojęcie ludzkiego kapitału o wiek. Wraz z wiekiem rośnie nasze doświadczenie, ale też zmieniają się możliwości naszego organizmu. Dlatego też zależność pomiędzy wiekiem a zarobkami nie będzie prostą zależnością liniową, a raczej krzywą. Ważnym wnioskiem pojawiającym się w artykule jest to, że w zawodach wymagających dłuższego treningu, wzrost produktywności związany ze wzrostem doświadczenia jest większy, a spadek związany ze starzeniem się mniejszy, w porównaniu do zawodów, przy których nie potrzeba długiej edukacji.
W 1974 roku, w artykule „Schooling, Experience, and Earnings. Human Behavior & Social Institutions" Jacob Mincer przedstawił model, nazywany funkcją zarobków Mincera:

$$\ln w = f(s,x) = \ln w_0 + \rho s + \beta_1 x + \beta_2 x^2$$

W której logarytm z płacy to funkcja logarytmu stałej (oczekiwanej wypłaty dla osoby bez edukacji i doświadczenia) oraz skorygowane o odpowiednie parametry ilość lat edukacji s i ilość lat potencjalnego doświadczenia na rynku pracy x. Uzasadnienie dla tych zmiennych jest takie samo jak w pierwszym przytoczonym artykule. Zmienna x pojawia się dwa razy, raz w pierwszej potędze i raz w drugiej, ze względu na wspomnianą wcześniej nieliniowość zjawiska. Rosnący początkowo bonus związany z doświadczeniem, w pewnym momencie zostaje zneutralizowany przez rosnące obciążenie związane z wiekiem. Mimo pewnych problemów (funkcja nie uwzględnia jakości edukacji, w rezultacie stopień korelacji pomiędzy dochodami a latami edukacji był niewielki) model ten w znacznym stopniu ukierunkował przyszłe badania na temat czynników wpływających na zarobki.

W późniejszych latach powstało wiele opracowań badających aktualność funkcji Mincera, a także jej dokładność. W artykule "Estimating the return to investments in education: how useful is the standard Mincer equation?" szwedzcy naukowcy sprawdzają na ile model Mincera

sprawdza się do tłumaczenia danych z ich kraju. Jak pokazują, sprowadzenie wpływu edukacji do prostej funkcji log-liniowej może prowadzić do błędnego wnioskowania i należałoby rozróżnić pojedyncze etapy edukacji. Udowadniają to na empirycznym przykładzie, gdzie dla danych z lat 1968-1981 wartość parametru przy latach edukacji malała. Po głębszej analizie, okazało się, że stopa zwrotu z edukacji licealnej była cały czas taka sama, jedynie spadła wartość studiowaniu na College'u. Przy prostej funkcji Mincera odkrycie takiej zależności byłoby niemożliwe. Kolejną kwestią było pokazanie, że założenie Mincera - o tym, że decydując się na edukację skracamy łączną ilość lat pracy – nie zgadza się z uzyskanymi obserwacjami, a to poważnie wpływa na oszacowanie parametru przy zmiennej o latach edukacji. Jak pokazują dane dla Szwecji, osoby podejmujące się zawodów wymagających dłuższej edukacji, zwykle również w późniejszym wieku przechodzą na emeryturę. Spowodowane jest to między innymi tym, że tego typu zawody są pod względem wysiłku fizycznego mniej wymagające.

Sprowadzenie kwestii zarobków do funkcji determinantów pozwoliło na zbadanie wielu innych nurtujących zagadnień. Jednym z nich jest luka płacowa pomiędzy płciami. To, że kobiety średnio zarabiają mniej od mężczyzn nie jest jeszcze wystarczającą przesłanką na uznanie płacowej dyskryminacji płciowej. Dopiero oszacowanie ekonometrycznego modelu pozwala zbadać, czy przy zachowaniu zasady ceteris paribus, zarobki kobiety różnią się w istotny sposób od zarobków mężczyzny.
Artykuł "What Determines Our Wage: The Econometric Analysis of Male-Female Wage Gap" dowodzi, że nieuzasadniona luka płacowa istnieje nadal, a jej wartość jest znacząca. Dodatkowo badanie pokazuje, że wpływ edukacji oraz doświadczenia na zarobki różni się w zależności od płci, co więcej - są to zależności nieliniowe. Za szczególnie interesujące uznałem właśnie wpływ stażu na różnice w zarobkach. Zgodnie z estymacją, mężczyźnie z 10 letnim stażem dodatkowy rok przyniesie wzrost zarobków w wysokości 1,6%. Dla takiej samej sytuacji w przypadku kobiety – 1,9%. W przypadku 20 letniego doświadczenia dodatkowy rok przyniesie mężczyźnie 1% - wzrost wynagrodzenia, a kobiecie 0,9%. Luka płacowa będzie się więc zmieniać w zależności od długości stażu. Badaczom udało się udowodnić jeszcze jeden ważny fenomen tj. wpływ małżeństwa na zarobki. Obliczono, że mężczyzna będący w takim związku zarabia o 9,4% od mężczyzn niebędących z związku (obojętnie czy singli, rozwiedzionych czy wdowców). Taki efekt zachodzi również w analogicznej sytuacji w przypadku kobiet, jednak jego siła jest znacznie słabsza – wzrost zarobków wynosi odpowiednio 4,9%. Zjawisko to nazywane jest 'marriage

premium' i istnieje wiele literatury poświęconej próbie wyjaśnienia go; efekt można uzasadnić np. teorią specjalizacji wewnątrz gospodarstwa domowego. Badanie wskazało również na istotny wpływ miejsca zamieszkania – osoba mieszkająca w Londynie może liczyć na zarobki wyższe o 28,5% w porównaniu do osoby mieszkającej w innym miejscu w UK, pomimo tych samych kwalifikacji.

"The effect of parents' education and earnings upon the education and earnings of their children" przy pomocy analizy regresji liniowej bada w jaki sposób zarobki i edukacja rodziców przekłada się na zarobki i edukację ich dzieci. Badania w tym obszarze są dość trudne, ze względu na brak możliwości zmierzenia zdolności, które są potencjalnie dziedziczone. Autor zasugerował podejście, w którym reszty z modelu opisującego zależność między płacami a edukacją mają w sobie informację o zdolnościach jednostek do przyswajania wiedzy i zarabiania pieniędzy. Dzięki temu, udało mu się dojść do następujących wniosków:
Rodzice z wysokimi pensjami mają lepiej wyedukowane dzieci. To z kolei przekłada się na wyższe zarobki dzieci, jednak w wyniku niewielkiego przełożenia pensji rodziców na edukację dzieci (niska elastyczność), efekt ten jest dość słaby. Oszacowano, że gdy rodzice zarabiają 10% więcej niż średnia, ich dziecko będzie zarabiać jedynie 1% więcej od średniej.
Rodzice z wyższą edukacją również mają lepiej wyedukowane dzieci, jednak w tym przypadku efekt jest jeszcze słabszy. Ostatecznie, jeśli proces edukacji rodziców trwał 10% dłużej, można spodziewać się większych zarobków u dziecka o 0,4%. Oba efekty są więc małe, jednak statystycznie istotne.

W publikacji „The wage premium from foreign language skills" Jacek Liwiński bada wysokość premii wynikającą ze znajomości języka obcego, na przykładzie polskiego rynku pracy. Udowodnione zostaje, że znajomość języka obcego w istotny sposób wiąże się z wyższymi zarobkami, jednak efekt ten, w zależności od języka, różni się znacznie. Zaawansowana znajomość języka hiszpańskiego zapewnia aż 32 procentową premię; co prawdopodobnie związane jest dużą różnicą pomiędzy popytem a ilością osób płynnie posługujących się tym językiem. Najbardziej popularne języki obce – j. angielski i j. niemiecki, są na tyle powszechne, że premia związana z ich znajomością jest niższa – odpowiednio 11 i 12%. Autor dodaje, że znajomość języka obcego nie tylko może wiązać się z większą produktywnością, ale także być sygnałem dla potencjalnego pracodawcy odnośnie wyższych zdolności – co czyni aplikację bardziej atrakcyjną, nawet jeśli w danej pracy język obcy nie jest potrzebny.

W artykule „Wpływ wykształcenia na rozkład zarobków w Polsce w latach 1988–2013” Henryk Domański dokonał oszacowania parametrów mających wpływ na zarobki, dla polskiego rynku (użyty przez niego model również był log-liniowy), dla 9 różnych punktów czasowych. Co interesujące, według jego estymacji, różnica w zarobkach pomiędzy osobą o wykształceniu podstawowym a średnim, okazała się statystycznie nieistotna. Dopiero fakt uzyskania wykształcenia wyższego powodował istotną różnicę – w zestawieniu z grupą referencyjną tj. osobami z wykształceniem niepełnym podstawowym, absolwenci szkół wyższych mogli liczyć na zarobki większe o 35,1%. Na zarobki w sposób istotny wpływała także płeć, jednak dane dla tego parametru dla różnych punktów czasowych różniły się znacznie. Przykładowo z estymacji dla danych z 2010 roku wynika, że mężczyźni, przy zachowaniu pozostałych parametrów na tym samym poziomie, zarabiają 24% więcej od kobiet. Z kolei dla danych z 2013 roku wartość różnicy wynosi aż 43%. Innymi istotnymi zmiennymi są: zmienna wskazująca czy respondent obejmuje stanowisko kierownicze oraz wielkość miejscowości (w jednym roku istotna okazała się również pozycja zawodowa ojca). Wnioski końcowe artykułu stanowią niejako potwierdzenie wniosków Mincera - wysokie wykształcenie otwiera drzwi do wysoko położonych pozycji zawodowych, gdzie obietnica wyższych zarobków zostaje z dużym prawdopodobieństwem zrealizowana.

2. Hipotezy

H1: Im wyższe osiągnięte wykształcenie, tym wyższe zarobki.

H2: Istnieje luka płacowa pomiędzy płciami.

H3: Osoby w związku małżeńskim zarabiają więcej od pozostałych osób, efekt ten zachodzi zarówno u mężczyzn jak i kobiet, z tym, że u kobiet jest słabszy.

H4: Fakt posiadania przez rodziców wykształcenia wyższego, będzie pozytywnie i w istotny sposób wpływał na dochody.

H5: Powyższą hipotezę można rozszerzyć o stanowisko, że wpływ wykształcenia rodziców będzie różny w zależności od płci. Na mężczyzn silniejszy wpływ będzie mieć wykształcenie ojca i vice versa.

H6: Na zarobki wpływ będzie mieć wielkość miejscowości zamieszkania. Większa miejscowość oznacza więcej ofert pracy i większą szansę na znalezienie pracy zgodnej z umiejętnościami.

H8: Znajomość kolejnych języków wpłynie na wielkość zarobków. W czasach, gdy znajomość jednego języka (angielskiego) jest standardem, znajomość drugiego, trzeciego języka obcego jest z pewnością czymś co w oczach pracodawcy pozwoliłoby się wyróżnić.

H9: Województwo ma istotny wpływ na zarobki. Prawdziwość tej hipotezy może wydać się oczywista – różnice w średnim dochodzie pomiędzy województwami to temat wielu raportów Głównego Urzędu Statystycznego. Jednak dopiero oszacowanie modelu ekonometrycznego pozwoli stwierdzić, czy rzeczywiście, przy zachowaniu zasady ceteris paribus, na zarobki wpłynie sam fakt zamieszkania w danym województwie a nie w innym.

H10: Na zarobki w istotny sposób będzie wpływać staż, a jego efekt będzie różny w zależności od płci.

H11: Stanowisko kierownicze będzie się wiązać z wyższymi zarobkami.

3. Opis danych i zmiennych

Wszystkie dane użyte do konstrukcji modelu pochodzą z Bilansu Kapitału Ludzkiego – ankiety przeprowadzonej na osobach dorosłych w wieku 18-69 (w 2017 r.) i mają charakter przekrojowy.

Próba początkowo liczyła 4057 obserwacji. Następnie dokonano przeglądu i selekcji danych:

W celu zachowania większej porównywalności, w arkuszu pozostawiono jedynie ludzi aktualnie zatrudnionych na podstawie umowy o pracę. Usunięto również obserwacje dotyczące jednostek pracujących na dwa etaty jednocześnie. W dalszej kolejności przyjrzano się brakom danych dla zmiennej objaśnianej tj. miesięcznym zarobkom netto. Ankieta była przeprowadzona w taki sposób, że jeśli respondent nie chciał/nie był w stanie udzielić dokładnej odpowiedzi na temat jego przeciętnych miesięcznych zarobków netto, ankieter prosił go, aby wskazał chociaż przedział zarobkowy, w którym znajdować się będzie odpowiednia wartość. Ze względu na liczność takich sytuacji (wśród osób, które odpowiedziały na pytanie o zarobkach, 1/6 odpowiedzi polegała na wskazaniu przedziału) postanowiono ich nie usuwać, a uwzględnić średnią z przedziału. Innymi słowy, przykładowo, jeśli ktoś nie podał dokładnej wartości, ale wskazał, że jego miesięczne zarobki znajdują się w przedziale 3000-3500 złotych, przypisywano mu dochód netto na poziomie 3250. Takie uproszczenie wiąże się z pewnym ryzykiem, jednak pozwala zachować liczebność próby. Problem pojawił się przy osobach, które zaznaczyły górny przedział „8 tys. i więcej";

ich zarobki mogły wynosić 9 tys. równie dobrze jak 30tys., a próba arbitralnego ustalenia wysokości ich zarobków byłaby wątpliwa. Z tego względu, ze zbioru danych usunięto takie obserwacje (łącznie 16 przypadków).

Następnie usunięto braki danych występujące przy pozostałych zmiennych branych pod uwagę, zmniejszając ostateczną próbkę do 1232 osób.

<u>Zmienna objaśniana</u>

- Dochód – miesięczne zarobki netto wyrażone w złotówkach, ankietowani byli proszeni o wskazanie ich przeciętnego poziomu z przestrzeni ostatnich 12 miesięcy

<u>Zmienne objaśniające, rozważane w trakcie konstrukcji modelu</u>

- Wiek – zmienna ilościowa ciągła, przyjmująca wartości całkowite z przedziału [19;69]

- Staż – zmienna ilościowa ciągła, wskazująca na łączną ilość przepracowanych lat

- Płeć – zmienna jakościowa binarna, przyjmująca wartości odpowiednio 0 dla mężczyzn i 1 dla kobiet

- Wykształcenie – zmienna jakościowa wskazująca na najwyższy poziom edukacji osiągnięty przez respondenta. Przyjmuje jeden z 7 poziomów:

1	Gimnazjalne i poniżej
2	zasadnicze zawodowe
3	średnie ogólnokształcące
4	średnie zawodowe (liceum, technikum)
5	średnie (szkoła polic, inna nie wyższa)
6	wyższe (licencjat/inż.)
7	wyższe (mgr)

  Przy budowie modelu została ona rozkodowana na 7 zmiennych binarnych. Aby uniknąć współliniowości, wykształcenie gimnazjalne lub niższe zostało przyjęte jako poziom bazowy i usunięte ze zmiennych.

- Miejsce – zmienna jakościowa wskazująca na wielkość miejscowości zamieszkania respondenta, mierzonej poprzez liczbę mieszkańców. W modelowania użyta została w rozkodowanej postaci kilku zmiennych binarnych:

Wieś (teren niezurbanizowany)	
Miasto 20k	miasto o populacji nie przekraczającej 20 tysięcy mieszkańców
Miasto 100k	miasto o populacji zawierającej się w przedziale 20-100 tysięcy mieszkańców
Duże miasto	miasto o populacji większej niż 100 tysięcy, z wyłączeniem Warszawy

Warszawa specjalny poziom utworzony tylko dla tego miasta

Gdzie zmienna „wieś” została przyjęta za poziom bazowy i usunięta.

- Województwo – zmienna jakościowa określająca województwo zamieszkania ankietowanego. Do modelu została rozkodowana na 16 zmiennych binarnych dla każdego województwa, po czym woj. Mazowieckie zostało przyjęte jako poziom referencyjny.

- Kierownik – zmienna jakościowa binarna, przyjmująca wartość 1 dla osób, które w pracy mają swoich podwładnych i 0 w p.p.

- Zagranica – zmienna jakościowa binarna, przyjmująca wartość 1 dla osób, które przynajmniej część ostatniego roku przepracowały zagranicą i 0 w p.p.

- Małżeństwo – zmienna jakościowa binarna, przyjmująca wartość 1 dla osób w związku małżeńskim i 0 w p.p.

- Język obcy – grupa trzech zmiennych jakościowych binarnych opisujących znajomość języków obcych:

    - J obcy – zmienna binarna przyjmująca wartość 1, jeśli respondent zna przynajmniej jeden język obcy i 0 w p.p.
    - obcy2 – zmienna binarna przyjmująca wartość 1, jeśli respondent zna przynajmniej dwa języki obce i 0 w p.p.
    - obcy3 – zmienna binarna przyjmująca wartość 1, jeśli respondent zna przynajmniej trzy języki obce i 0 w p.p.

  Zmienne zostały zakodowane w taki sposób, że jeśli przykładowo ktoś zna trzy języki obce, wartość przy każdej ze zmiennych dot. języka będzie równa 1. Dzięki temu współczynniki przy zmiennych obcy2 i obcy3 będą bezpośrednio wskazywać wielkość premii wynikającej ze znajomości kolejnego języka.

- W_matka1 – zmienna jakościowa binarna, przyjmująca wartość 1, jeśli (w momencie, kiedy respondent miał 14 lat) jego matka posiadała wykształcenie wyższe i 0 w p.p.

- W_ojciec1 – zmienna jakościowa binarna, przyjmująca wartość 1, jeśli (w momencie, kiedy respondent miał 14 lat) jego ojciec posiadał wykształcenie wyższe i 0 w p.p.

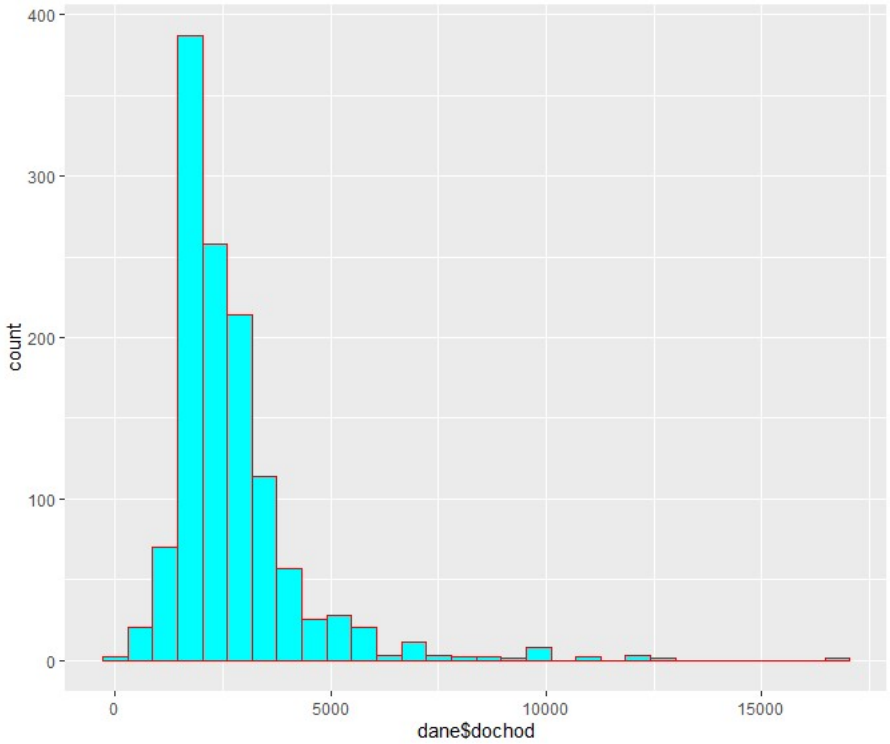
Poza zmiennymi, przy konstrukcji modelu rozważane są 4 interakcje pomiędzy nimi:

- Płeć X w_matka1
- Płeć X w_ojciec1
- Płeć X małżeństwo
- Płeć X staż

Analiza zmiennych i ustalenie formy funkcyjnej

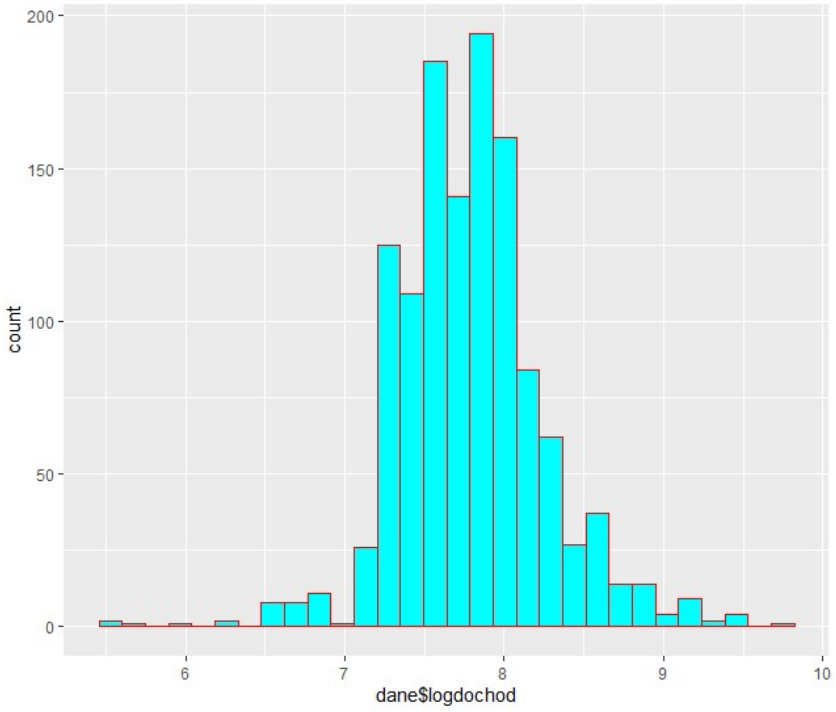
Zmienne takie jak PKB per capita, zarobki, bardzo często mają rozkład zbliżony do log normalnego. Aby sprawdzić czy tak się dzieje w przypadku badanej próby, wygenerowano histogram zmiennej dochody.

Rysunek 1 - Histogram zmiennej dochód



Zgodnie z przypuszczeniami, wydruk histogramu sugeruje potrzebę zastosowania logarytmu zmiennej objaśnianej, w celu lepszego dopasowania.

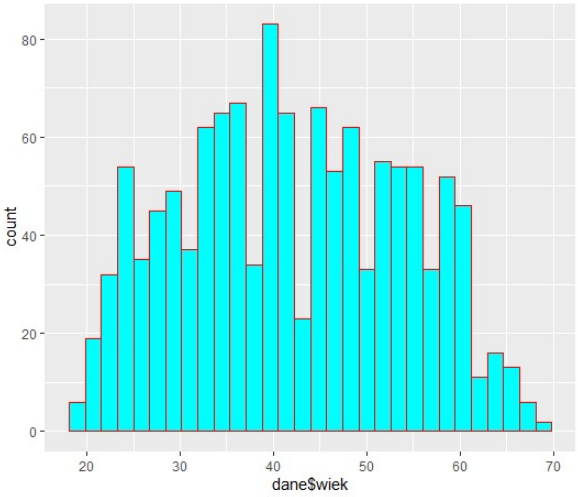
Rysunek 2 - Histogram logarytmu zmiennej dochód



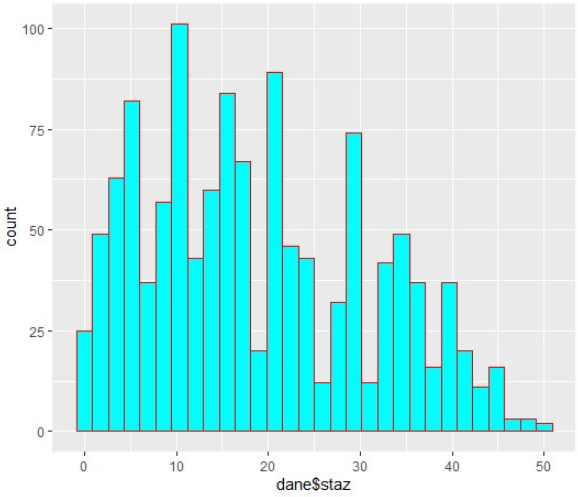
Po zlogarytmowaniu rozkład nadal nie może zostać uznany za normalny, chociaż histogram jest teraz zdecydowanie bardziej zbliżony do krzywej Gaussa niż w pierwotnej postaci. Z tego względu już od pierwszej iteracji modelu badana będzie zależność pomiędzy logarytmem dochodu a zmiennymi objaśniającymi.

Wśród wymienionych zmiennych objaśniających pojawiają się dwie ciągłe, dla których również wygenerowano histogramy.

Rysunek 3a - Histogram zmiennej wiek



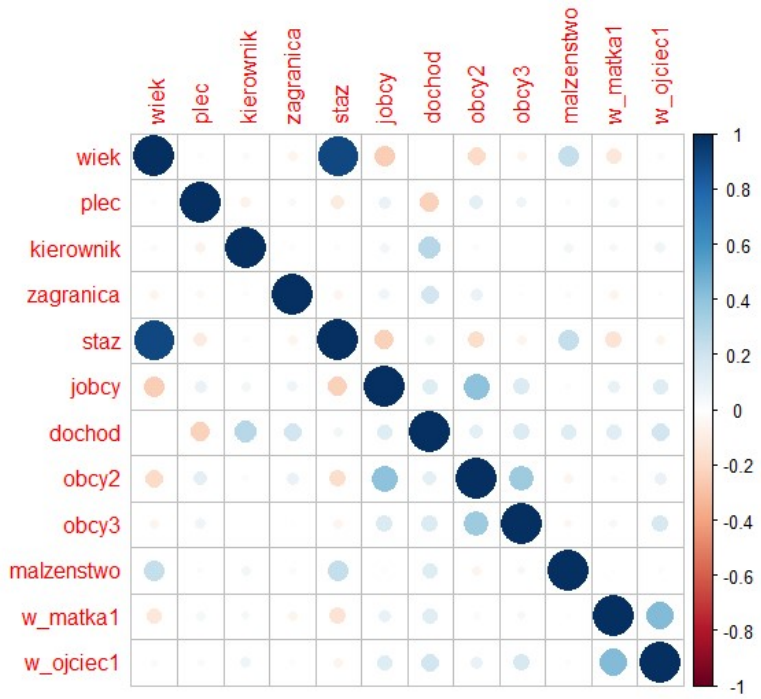
Rysunek 3b – Histogram zmiennej staż



Jednak ich rozkłady nie sugerują potrzeby modyfikacji; poza tym utrudniłoby to interpretację parametrów.

Następnie przy pomocy wykresów sprawdzono korelację pomiędzy wybranymi zmiennymi i ewentualne zależności liniowe.

Tabela 1 Siła korelacji pomiędzy wybranymi zmiennymi



Powyższy graf sugeruje pewien problem – zmienne wiek i staż są ze sobą bardzo silnie skorelowane dodatnio i obie oddziałują na zmienną dochód w tym samym kierunku (zgodnie z intuicją).

Oznacza to, że w przypadku wzrostu dochodu, nie będzie wiadomo której ze zmiennych należy przypisać odpowiedzialność za zmianę. Przeprowadzono więc serię testów na korelację metodą Pearsona:

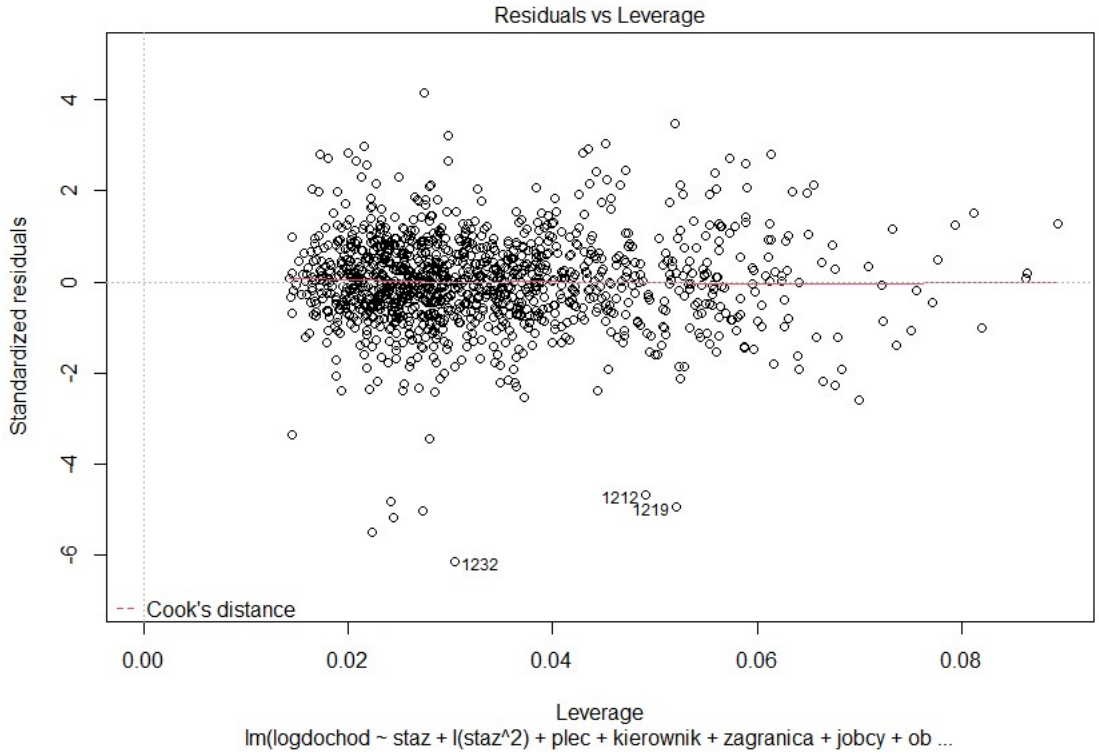
Korelacja pomiędzy wiekiem a stażem: 0.907579; p-value < 2.2e-16
Korelacja pomiędzy logarytmem dochodu a stażem: 0.08327909; p-value = 0.003442
Korelacja pomiędzy logarytmem dochodu a wiekiem: 0.004448806; p-value = 0.876

Hipoteza zerowa w tym przypadku mówi o tym, że nie ma korelacji pomiędzy wskazanymi zmiennymi. Na podstawie p-value możemy więc powiedzieć, że zależność pomiędzy wiekiem a logarytmem dochodu jest nieistotna statystycznie. Taki wynik analizy jest również bardziej rozsądny z intuicyjnego punktu widzenia – pracodawca może zapłacić komuś więcej z racji tego, że ma większe doświadczenie, a nie ze względu na sam wiek. Dlatego też, zmienna wiek została odrzucona z dalszej procedury modelowania.

Korelacje pomiędzy pozostałymi zmiennymi widocznymi na grafie są niewielkie, ewentualne problemy ze współliniowością zostaną rozpatrzone przy użyciu funkcji VIF.

Po przeprowadzeniu wstępnej analizy danych, przy pomocy wykresu standaryzowanych reszt względem dźwigni sprawdzono bazę danych pod kątem ewentualnych obserwacji odstających.

Rysunek 4 Wykres standaryzowanych reszt względem dźwigni



Wykres wskazuje na brak obserwacji, które zaburzałyby wnioskowanie, (tj. np. ‘bad levarage point’) chociaż kilka z nich jest dość zastanawiających. Obserwacje przy których program wyświetlił numer, łączą bardzo niskie zadeklarowane dochody, rzędu kilkuset złotych (np. numer 1219 to 750 złotych). Trudno stwierdzić, czy wynika to z pomyłki, czy np. dana osoba w ciągu ostatniego roku pracowała na część etatu. Z tego względu nic nie usuwano.

4. Weryfikacja hipotez

Do modelowania zastosowano procedurę od ogółu do szczegółu, rozpoczynając od modelu zawierającego wszystkie potencjalnie istotne zmienne i interakcje, a następnie usuwając te, które okazały się statystycznie nieistotne.
Pierwsza, ogólna postać modelu, ma postać:

$$\ln(dochod) = \beta 0 + \beta 1staz + \beta 2staz^2 + \beta 3plec + \beta 4kierownik + \beta 5zagranica + \beta 6jobcy + \beta 7obcy2 + \beta 8obcy3 + \beta 9w\_matka1 + \beta 10w\_ojciec1 + \beta 11malzenstwo + \beta 12miejsce + \beta 13wojewodztwo + \beta 14wyksztalcenie + \beta 15plec\ X\ w\_matka1 + \beta 16plec\ X\ w\_ojciec1 + \beta 17plec\ X\ malzenstwo + \beta 18plec\ X\ staz$$

Przy czym zmienne miejsce, wojewodztwo i wyksztalcenie w programie są rozkodowane w sposób wspomniany w poprzednim rozdziale.

Wynik regresji wskazuje, że model jest statystycznie istotny i wyjaśnia około 37% całkowitej zmienności zmiennej objaśnianej. Bardzo wiele zmiennych okazało się statystycznie nieistotnych, jednak przed przystąpieniem do modyfikacji, model zostanie zbadany pod kątem spełnienie założeń KMRL.

Diagnostyka modelu

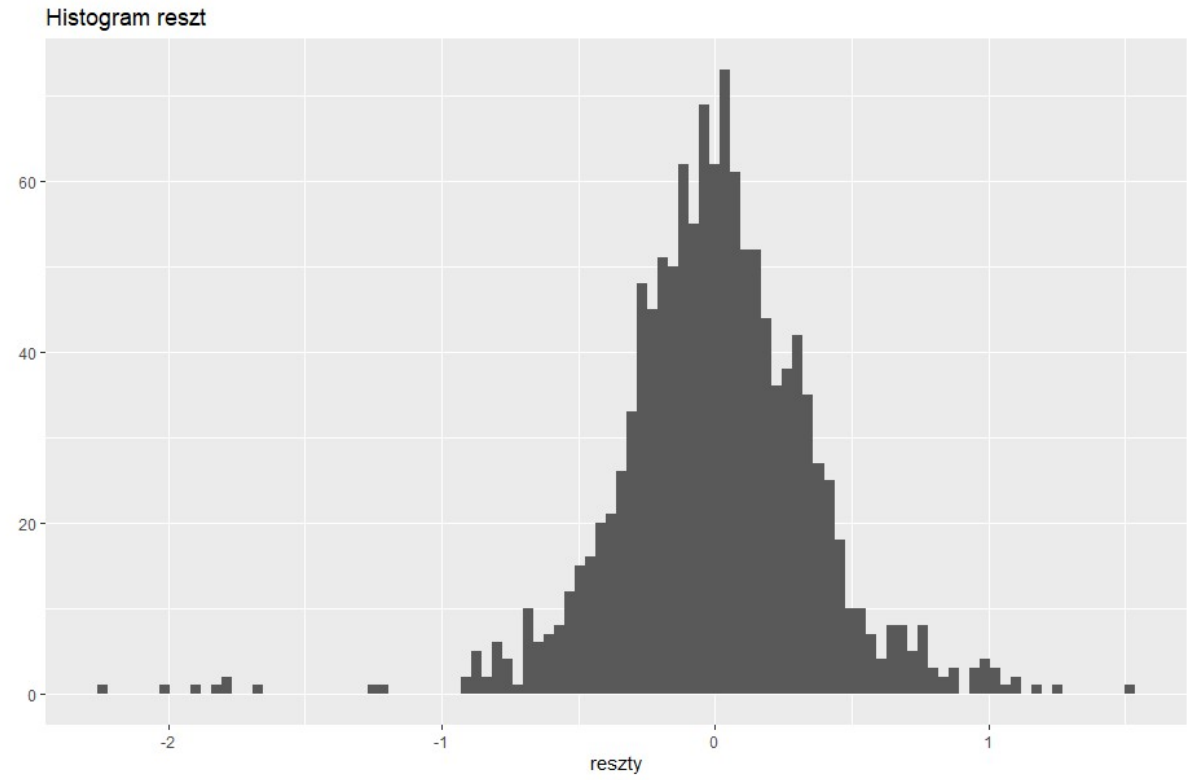
- Poprawność formy funkcyjnej

Użyty został test RESET zarówno w formie „fitted" jak i „regressor". W obu przypadkach p-value było większe od 0,05; co na przyjętym 5%-poziomie istotności oznacza brak podstaw do odrzucenia hipotezy zerowej o poprawności formy funkcyjnej. (Inaczej stało się w przypadku użycia niezlogarytmowanego dochodu, wtedy test wskazuje na niepoprawną formę funkcyjną, co potwierdza słuszność podjętej wcześniej decyzji.)

- Normalność reszt

Wygenerowano wykres:

Rysunek 5 Histogram reszt



Który wskazuje na pewne oddalenie się od kształtu krzywej Gaussa – tego rozkładu nie można uznać za normalny. Wnioski wynikające z wykresu potwierdza również test Jarque-Berra. P-Value na poziomie bliskim 0 każe odrzucić hipotezę zerową o normalności rozkładu reszt. Konsekwencje braku spełnienia tego założenia są poważne – taka sytuacja uniemożliwia rzetelną ocenę istotności zmiennych, ponieważ p-value dla każdej z nich jest zaburzone.

Ponieważ jednak badana próbka jest relatywnie duża (1232 obserwacje), taka sytuacja nie stanowi problemu – dla dużych prób rozkłady statystyk służących do testowania są zbliżone do standardowych rozkładów. Dlatego też uzyskiwane wyniki odnośnie istotności zmiennych będą uznawane za wiarygodne.

- Homoscedastyczność

Dla przeprowadzonego testu na homoscedastyczność (stałość wariancji składnika losowego) Breuscha-Pagana uzyskano p-value na poziomie 0.0004232. Na przyjętym poziomie istotności odrzucono więc hipotezę zerową na korzyść alternatywnej o istnieniu heteroscedastyczności. Konsekwencje w tym przypadku są podobne jak w poprzednim podpunkcie – estymatory są co prawda nie obciążone, jednak nie można wierzyć w p-value które pojawia się przy zmiennych (macierz wariancji-kowariancji jest nieprawidłowa).
W związku z tym, aby uniknąć błędów we wnioskowaniu statystycznym, w dalszej części użyto ‘odpornej’ macierzy White’a.

- Autokorelacja

Ze względu na przekrojowy charakter próbki, uznano, że nie występuje autokorelacja reszt.

- Wartość oczekiwana błędu losowego

Aby spełnione były wszystkie założenia KMRL, E(**ε**) powinna wynosić 0. Niestety, w przypadku opracowywanego modelu pojawia się problem – zgodnie z intuicją, a także licznymi badaniami, na zarobki wpływają także (a być może przede wszystkim) umiejętności i inteligencja. Uwzględnienie takich zmiennych jest niemożliwe, ponieważ nie istnieje ogólnopolski rejestr dla ilorazu inteligencji, umiejętności są trudne do zmierzenia, a użycie danych pokroju jak respondent ocenia swoje umiejętności w danej dziedzinie także nie byłoby dobrym rozwiązaniem ze względu na wysoką subiektywność. Taka sytuacja nie stanowi problemu, gdy zmienne pominięte są nieskorelowane ze zmiennymi występującymi w modelu. W tym jednak przypadku zmienna o inteligencji byłaby prawdopodobnie dość silnie skorelowana ze zmienną dot. poziomu wykształcenia; w rezultacie będziemy tutaj mieli do czynienia z obciążeniem estymatora.

Na koniec sprawdzono jeszcze współliniowość zmiennych przy pomocy funkcji VIF. W literaturze przyjęło się, że VIF>10 świadczy o silnej, niedokładnej współliniowości. Najwyższą wartość osiągnięto dla zmiennej staż (20.782210) oraz staż^2 (16.529271) co jest jednak

normalną rzeczą w przypadku umieszczenia w modelu zmiennej będącej przekształceniem drugiej zmiennej. VIF dla pozostałych zmiennych jest niski, w związku z tym nie usunięto żadnej zmiennej.

Po sprawdzeniu założeń KMRL, przystąpiono do procedury ‘general to specific’ z poprawką na heteroscedastyczność w postaci używania macierzy odpornej White’a.

Końcowa wersja modelu

Przy 13stej iteracji model uzyskał swoją ostateczną formę, wszystkie zmienne na poziomie istotności 0.05 są statystycznie istotne, a statystyka R^2 wyniosła 0,358 – oznacza to, że model tłumaczy 35,8% całkowitej zmienności logarytmu dochodów, co biorąc pod uwagę wielkość próby i złożoność zjawiska, jest rezultatem dość zadowalającym.
Otrzymany model ma postać (tutaj już w pełni rozkodowaną):

$$\ln(\text{dochod}) = \beta 0 + \beta 1\text{staz} + \beta 2\text{staz}^2 + \beta 3\text{plec} + \beta 4\text{kierownik} + \beta 5\text{zagranica} + \beta 6\text{jobcy} + \beta 7\text{malzenstwo} + \beta 8\text{warszawa} + \beta 9\text{lubelskie} + \beta 10\text{podkarpackie} + \beta 11\text{srednieogolnoksztalcace} + \beta 12\text{sredniezawodowe} + \beta 13\text{sredniepolicealne} + \beta 14\text{wyzszelic} + \beta 15\text{wyzszemgr} + \beta 16\text{plec X malzenstwo} + \beta 17\text{plec X staz}$$

Przy czym w zmiennych binarnych za poziom odniesienia przyjęto mieszkańca wsi województwa mazowieckiego, o wykształceniu gimnazjalnym lub niższym.

Poniżej umieszczono tabelę przedstawiającą porównanie modelu podstawowego i modelu ostatecznego. Obok każdego umieszczono również oszacowania przy użyciu macierzy White’a, ze skorygowanymi błędami standardowymi i p-value.

	Dependent variable:			
	logdochod		logdochod	
	OLS	coefficient test	OLS	coefficient test
	(1)	(2)	(3)	(4)
staz	0.011***	0.011***	0.011***	0.011***
	(0.004)	(0.004)	(0.004)	(0.004)
I(staz2)	-0.0002***	-0.0002***	-0.0002***	-0.0002***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
plec	-0.348***	-0.348***	-0.350***	-0.350***
	(0.049)	(0.049)	(0.048)	(0.048)
kierownik	0.193***	0.193***	0.190***	0.190***
	(0.024)	(0.026)	(0.024)	(0.026)
zagranica	0.508***	0.508***	0.511***	0.511***
	(0.065)	(0.104)	(0.064)	(0.103)
jobcy	0.073**	0.073***	0.079***	0.079***
	(0.031)	(0.028)	(0.029)	(0.027)
obcy2	-0.008	-0.008		
	(0.026)	(0.025)		
obcy3	0.069	0.069		
	(0.044)	(0.057)		
w_matka1	0.144**	0.144*		
	(0.070)	(0.086)		
w_ojciec1	-0.029	-0.029		
	(0.069)	(0.102)		
malzenstwo	0.181***	0.181***	0.185***	0.185***
	(0.038)	(0.036)	(0.037)	(0.036)
miasto20k	0.009	0.009		
	(0.030)	(0.029)		
miasto100k	-0.009	-0.009		
	(0.030)	(0.028)		
duzemiasto	-0.012	-0.012		
	(0.032)	(0.029)		
warszawa	0.168**	0.168*	0.270***	0.270***
	(0.083)	(0.092)	(0.071)	(0.077)
dolnoslaskie	-0.031	-0.031		
	(0.058)	(0.059)		
lubelskie	-0.183***	-0.183***	-0.123**	-0.123**
	(0.069)	(0.065)	(0.057)	(0.051)
lubuskie	-0.052	-0.052		
	(0.075)	(0.059)		
lodzkie	-0.117**	-0.117**		
	(0.053)	(0.050)		
malopolskie	-0.065	-0.065		
	(0.055)	(0.059)		
opolskie	-0.161*	-0.161**		
	(0.087)	(0.078)		
podkarpackie	-0.165***	-0.165***	-0.106**	-0.106***
	(0.058)	(0.053)	(0.043)	(0.036)
podlaskie	-0.041	-0.041		
	(0.072)	(0.065)		
pomorskie	-0.028	-0.028		
	(0.061)	(0.055)		
slaskie	-0.024	-0.024		
	(0.052)	(0.053)		
swietokrzyskie	-0.102	-0.102		
	(0.063)	(0.063)		
wielkopolskie	-0.116**	-0.116*		
	(0.054)	(0.060)		
zachodniopomorskie	-0.012	-0.012		
	(0.061)	(0.056)		
kujawskopomorskie	-0.078	-0.078		
	(0.062)	(0.067)		
warminskomazurskie	-0.146**	-0.146**		
	(0.070)	(0.068)		

zasadniczezawodowe	-0.001	-0.001		
	(0.059)	(0.047)		
srednieogolnoksztalcace	0.137**	0.137***	0.136***	0.136***
	(0.068)	(0.053)	(0.044)	(0.038)
sredniezawodowe	0.125**	0.125***	0.125***	0.125***
	(0.060)	(0.048)	(0.031)	(0.031)
sredniepolicealne	0.156**	0.156***	0.156***	0.156***
	(0.067)	(0.051)	(0.042)	(0.035)
wyzszelic	0.278***	0.278***	0.282***	0.282***
	(0.071)	(0.059)	(0.048)	(0.048)
wyzszemgr	0.421***	0.421***	0.443***	0.443***
	(0.062)	(0.050)	(0.033)	(0.032)
I(plec * w_matka1)	-0.113	-0.113		
	(0.088)	(0.095)		
I(plec * w_ojciec1)	0.120	0.120		
	(0.087)	(0.112)		
I(plec * malzenstwo)	-0.098**	-0.098**	-0.106**	-0.106**
	(0.049)	(0.048)	(0.048)	(0.047)
I(plec * staz)	0.005***	0.005***	0.006***	0.006***
	(0.002)	(0.002)	(0.002)	(0.002)
Constant	7.489***	7.489***	7.433***	7.433***
	(0.082)	(0.075)	(0.051)	(0.056)
Observations	1,232		1,232	
R2	0.372		0.358	
Adjusted R2	0.351		0.349	
Residual Std. Error	0.370 (df = 1191)		0.370 (df = 1214)	
F Statistic	17.662*** (df = 40; 1191)		39.850*** (df = 17; 1214)	
Note:			*p<0.1; **p<0.05; ***p<0.01	

Oszacowany model ma charakter log-liniowy, a więc mamy tutaj do czynienia z semi-elastycznością. Aby dokładnie i poprawnie zinterpretować wielkość parametrów, należy posłużyć się wzorem

$$(e^{\beta} - 1) * 100\%$$

Hipoteza dotycząca zarobków wzrastających wraz z poziomem wykształcenia zostaje potwierdzona – oszacowania parametrów przy kolejnych poziomach stopniowo się zwiększają. W porównaniu do osoby z wykształceniem gimnazjalnym lub niższym – absolwentów studiów I stopnia może liczyć na wynagrodzenia większe o 32,58%, natomiast absolwent studiów II stopnia – 55,73%. Jedynie wykształcenie zasadnicze zawodowe okazało się statystycznie nie różnić od poziomu bazowego.

Luka płacowa pomiędzy płciami według tego oszacowania jest istotna i przy zerowym stażu, wynosi 29,53% (zarobki kobiety są o tyle% niższe). Jednocześnie, interakcja pomiędzy płcią a stażem również okazała się statystycznie istotna a wartość jej parametru sprawia, że różnica w

wynagrodzeniu pomiędzy mężczyznami a kobietami będzie się zmniejszać wraz ze stażem – dla 20 letniego stażu, kobieta będzie zarabiać już 17,5% mniej.

Wyniki regresji potwierdziły również hipotezę trzecią, dot. ‘marriage premium’. Mężczyzna w związku małżeńskim może spodziewać się dochodów większych o 20,3%. U kobiet to odpowiednio +10,2%; a więc rezultaty co do zasady zgadzają się z literaturą.

Wpływ wykształcenia rodziców w tym modelu okazał się zupełnie nie istotny – zarówno H4 jak i H5 należy odrzucić. Być może związane jest to ze zbytnim uproszczeniem zmiennej (zmienna binarna przyjmująca 1, gdy rodzic ma wykształcenie wyższe i 0 w p.p.); a być może podział należało przeprowadzić na zasadzie posiada maturę/nie posiada matury.

W przypadku zmiennych dotyczących wielkości miejscowości – większość z nich okazała się statystycznie nie odróżnialna od poziomu bazowego (wsi). Jedynie zmienna Warszawa pozostała w modelu – mieszkańcy tego miasta zarabiają o 31% więcej w porównaniu do reszty.

Znajomość języka obcego zwiększa oczekiwane zarobki o 8,22%; natomiast nie zaobserwowano statystycznie istotnej premii wynikającej ze znajomości drugiego i trzeciego języka obcego. Te oszacowania są dużym uproszczeniem – w modelu nie rozróżniano poszczególnych języków obcych, a także poziomu umiejętności, co z dużym prawdopodobieństwem miałoby istotny wpływ.

Zarobki w większości województw okazały się statystycznie nie różnić od poziomu referencyjnego (woj. Mazowieckiego). Wyróżniło się tutaj jedynie województwo Lubelskie i Podkarpackie – mieszkańcy tych województw, przy zachowaniu zasady ceteris paribus, będą zarabiać odpowiednio 11,6% i 10% mniej. Zgadza się to z danymi empirycznymi tzn. są to jedne z najbiedniejszych województw; jednak warto zauważyć, że różnica w wynagrodzeniu jest znacznie mniejsza, niż gdyby zwyczajnie porównać przeciętne płace w wymienionych województwach w roku 2017.

Zgodnie z funkcją Mincera, wpływ stażu na płacę okazał się ważny i nieliniowy. Z tego powodu, aby policzyć wpływ dodatkowego roku doświadczenia na zarobki, konieczne jest policzenie pochodnej po stażu. Dzięki takiej operacji otrzymujemy, że np. z punktu widzenia mężczyzny z 10 letnim stażem, dodatkowy rok przełoży się na wzrost zarobków o 7,5%.

Głębsza analiza tej zależności wskazuje, że w przypadku pozostałych zmiennych na stałym poziomie, mężczyzna osiągnie maksimum zarobków mając około 28 lat doświadczenia na rynku pracy. Po przekroczeniu 55 lat stażu jego zarobki są gorsze niż gdyby nie miał stażu w ogóle. Wyjaśnienie tego zjawiska, zgodnie z publikacjami Mincera sprowadza się do tego, że w pewnym momencie wiedza nagromadzona przez lata doświadczenia nie jest w stanie zrekompensować skutków starzenia się. W przypadku kobiet, do obliczeń należy jeszcze uwzględnić pojawiającą się w modelu interakcję pomiędzy płcią i stażem. Po takiej poprawce uzyskujemy, że ceteris paribus kobieta może liczyć na maksymalne dochody po 40 latach obecności na rynku pracy. Jest to estymacja nieco zaskakująca i wynika prawdopodobnie z tego, że zależność pomiędzy interakcją płećXstaż a dochodem jest także nieliniowa, czego nie uwzględniono w modelu.

Osoby obejmujące stanowiska kierownicze (mające pod sobą podwładnych) mogą liczyć na zarobki wyższe o 21% w porównaniu do innych pracowników.

Najwyższe w całym modelu oszacowanie parametru znalazło się przy zmiennej ‘zagranica’ – osoby, które przynajmniej część roku przepracowały za granicą, miały przeciętne roczne dochody netto wyższe aż o 67% w porównaniu, do osób które cały rok pracowały w kraju. Jest to zgodne z danymi empirycznymi – w 2018 roku Polacy pracujący za granicą zarobili w przeliczeniu średnio 8,6 tys. złotych w skali miesiąca.

Zakończenie

Oszacowany model pozwolił zweryfikować szereg pojawiających się w literaturze hipotez. Estymacja pokazuje, że rzeczywiście mamy do czynienia w Polsce z nieusprawiedliwioną luką płacową (dyskryminacją płciową). Analiza pokazała również obecność zjawiska ‘marriage premium’ – wyższych zarobków wśród osób będących w związku małżeńskim; a także niepodważalny wpływ osiągniętego wykształcenia. Dla dwóch województw zarobki okazały się istotnie niższe, co jest szczególnie interesujące ze względu na fakt, że na ogół takiej zmiennej nie bierze się pod uwagę. W kwestii wpływu języka obcego, wykształcenia rodziców i luki płacowej zmieniającej się wraz ze stażem - model okazał się prawdopodobnie zbyt prosty

na wychwycenie zależności; jest to zdecydowanie element warty do wzięcia pod uwagę przy kolejnych, bardziej zaawansowanych badaniach. Ogółem, do wyników powyższej estymacji należy podchodzić z dozą ostrożności, ponieważ niezwykle ważna zmienna związana z umiejętnościami i inteligencją jest w modelu pominięta; potencjalnie obciążając uzyskane estymatory.

Bibliografia

Mincer, J. (1958). Investment in Human Capital and Personal Income Distribution. Journal of Political Economy, 66(4), 281–302.

Mincer, J.(1974). Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2. 1

Björklund, A., & Kjellström, C. (2002). Estimating the return to investments in education: how useful is the standard Mincer equation? Economics of Education Review, 21(3), 195–210.

Eli, S. (2009). What Determines Our Wage : The Econometric Analysis of Male-Female Wage Gap.

Beenstock, M. (2005). The Effect of Parents' Education and Earnings upon the Education and Earnings of their Children.

Liwiński, J.(2019). The wage premium from foreign language skills. *Empirica* 46**,** 691–711

Domański, H. (2018). Wpływ wykształcenia na rozkład zarobków w Polsce w latach 1988–2013

PARP. Baza danych z badania ludności w wieku 18-69 - Bilans Kapitału Ludzkiego 2017 (XLSX) Dostęp 13 lutego 2021.

https://www.parp.gov.pl/images/publications/Bilans_Kapitalu_Ludzkiego_2017/Baza-danych-z-badania-ludnoci-w-wieku-18-69---Bilans-Kapitau-Ludzkiego-2017.xlsx

PARP. Kwestionariusz BKL - badanie ludności 2017 Dostęp 13 lutego 2021. https://www.parp.gov.pl/images/publications/Bilans_Kapitalu_Ludzkiego_2017/Kwestionariusz-BKL---badanie-ludnoci-2017.pdf

GUS. Obwieszczenie w sprawie wysokości przeciętnego miesięcznego wynagrodzenia brutto w gospodarce narodowej w województwach w 2017 roku. Dostęp 17 lutego. https://stat.gov.pl/sygnalne/komunikaty-i-obwieszczenia/lista-komunikatow-i-obwieszczen/obwieszczenie-w-sprawie-wysokosci-przecietnego-miesiecznego-wynagrodzenia-brutto-w-gospodarce-narodowej-w-wojewodztwach-w-2017-roku,295,4.html

Woźniak Rafał. Ekonometria dla IiE i MSEMat W07. On-line. Dostęp 15 luty 2021. https://moodle.wne.uw.edu.pl/pluginfile.php/71511/mod_resource/content/1/EKIiE_W07_2020.pdf

Woźniak Rafał. Ekonometria dla IiE i MSEMat W08. On-line. Dostęp 15 luty 2021. http://coin.wne.uw.edu.pl/rwozniak/pliki/EKIiE_WYK/EKIiE_W08_2020.pdf