

**Wydział Zarządzania**

***Badanie czynników wpływających***

***na średnie wynagrodzenie w podregionach***

***w Polsce w 2021r.***

Autor: *Arkadiusz Florek*

Kierunek studiów: *Informatyka i Ekonometria*

Przedmiot: Ekonometria

Prowadzący: dr Paweł Zając

Kraków, 2023r.

# Cel projektu

Wynagrodzenie to doskonały sposób na weryfikację stanu gospodarki. Poprawa perspektyw biznesowych znajduje odzwierciedlenie w wyższym poziome pensji. Jej wzrost jest wynikiem lepszej sytuacji na rynku pracy(lepsza pozycja negocjacyjna pracowników) oraz dobrej sytuacji finansowej firm[[1]](#footnote-1). Można stwierdzić, że w wynagrodzenia to bardzo ważny czynnik w aktualnych społeczeństwach, przez co zostały obiektem badań, jednak nie można badać wynagrodzenia każdego człowieka osobno, ponieważ zależy od zbyt wielu czynników, dlatego potrzebowano znaleźć uniwersalny wskaźnik pozwalający badać to zjawisko.

Obraz zawierający tekst, zrzut ekranu, linia, Czcionka

Opis wygenerowany automatycznieWskaźnikiem najczęściej używanym do opisu wynagrodzeń w Polsce jest przeciętne miesięczne wynagrodzenie brutto. W okresie od 2004 r. do 2012 r. przeciętne wynagrodzenie wzrosło z 2289,57 zł do 3521,67, czyli o 54%. Dynamika wzrostu była mocno zróżnicowana: w latach 2004–2006 nie przekraczała 5%, w latach 2007 i 2008 zwiększyła się, a w latach   
2009–2011 utrzymywała się na wyraźnie niższym poziomie, nieznacznie przekraczającym wskaźnik wzrostu cen towarów i usług konsumpcyjnych[[2]](#footnote-2).

Rysunek 1 Krzywa przedstawiająca wartość średniego wynagrodzenia brutto w latach 2012-2021

Tak jak widać na powyższym wykresie [rysunek 1], stworzonym na podstawie danych z GUS, od 2012 roku średnie wynagrodzenie brutto zachowało swoją tendencje wzrostową   
i nic nie wskazuje, że ma to ulec zmianie.

W projekcie chce zbadać wpływ wybranych czynników na średnie wynagrodzenie  
w podregionach w Polsce w 2021 r. W kolejnych częściach przedstawię bliżej zmienne objaśniające, które wziąłem pod uwagę oraz wybiorę właściwy model i zweryfikuje postawione hipotezy.

# Hipotezy badawcze

1. Wyższy stopień rejestrowanego bezrobocia skutkuje niższą liczbą poszkodowanych osób w wypadkach przy pracy.
2. Zwiększenie się liczby kotłowni przyczynia się istotnie do wyższej liczby ludzi ulegających wypadkom w pracy.
3. Czym większa liczba zawartych małżeństw tym większa liczba ludzi poszkodowanych w wypadkach przy pracy.
4. Im większe nakłady inwestycyjne w przedsiębiorstwach przypadające na 1 mieszkańca tym większa liczba osób ulegających wypadkom w pracy.

# Opis danych

Dane, które wykorzystuje w projekcie, pochodzą z zasobów Głównego Urzędu Statystycznego. Poniższy link przekieruje do strony internetowej GUS, z której zostały pobrane: https://bdl.stat.gov.pl/bdl/dane/podgrup/temat.

Dane przedstawiają jedną zmienną objaśnianą Y oraz X1…X4 zmiennych objaśniających dla podregionów(wartości odstające zostały usunięte) - 64 obserwacji - w  
2021r. Model ten można zapisać w następującej postaci:

Zmienna objaśniania:

* Y - Liczba poszkodowanych osób w wypadkach przy pracy [osoba]

Potencjalne zmienne objaśniające:

* X1 - Stopa bezrobocia, obliczono jako stosunek liczby bezrobotnych do liczby cywilnej ludności aktywnej zawodowo(bez osób odbywających czynną służbę wojskową itd.) [%]
* X2 - Liczba kotłowni znajdujących się w gospodarce mieszkaniowej i komunalnej[szt.]
* X3 - Liczba zawartych małżeństw
* X4 - Nakłady inwestycyjne w przedsiębiorstwach na 1 mieszkańca [zł]

# Statystki opisowe

Tabela 1 Statystyki opisowe

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Zmienna****Statystyka** | **Y** | **X1** | **X2** | **X3** | **X4** |
| **Średnia** | 915,95 | 6,5159 | 473,57 | 2198,2 | 4137,5 |
| **Mediana** | 824 | 6,5 | 451 | 2106 | 3846 |
| **Minimalna** | 298 | 1,6 | 187 | 763 | 1573 |
| **Maksymalna** | 1806 | 12,5 | 831 | 3906 | 7665 |
| **Odch.stand** | 410,84 | 2,7062 | 154,6 | 698,74 | 1349,2 |
| **Wsp.zmienności** | 0,44854 | 0,41533 | 0,32646 | 0,31787 | 0,3261 |
| **Skośność** | 0,50319 | 0,15355 | 0,31963 | 0,261 | 0,64832 |
| **Kurtoza** | -0,82455 | -0,55719 | -0,41083 | -0,53654 | -0,11129 |

Źródło: Własne opracowanie na podstawie wyników programu Gretl

W tabeli 1 zostały przedstawione podstawowe statystyki zmiennych, które zostały wykorzystane w badaniu.

* Zmienna Y – Wartość minimalna wynosi 298, a wartość maksymalna 1806.  
  W zmiennej objaśnianej Y średnia jest większa od mediany, co wskazuje na wystąpienie asymetrii prawostronnej, potwierdza to współczynnik skośności wynoszący 0.5. Typowa wielkość Y różni się od wartości przeciętnej średnio 410.84.
* Zmienna X1 – Minimalna wartość stopy bezrobocia wynosi 1.6, a wartość maksymalna 6.5. Współczynnik skośności wynosi 0.15. Typowa wielkość X6 różni się od wartości przeciętnej średnio o 2.71.
* Zmienna X2 – Minimalna wartość wynosi 187, a wartość maksymalna 831. W zmiennej objaśniającej X2 średnia jest większa od mediany, co wskazuje na wystąpienie asymetrii prawostronnej, potwierdza to współczynnik skośności wynoszący 0.31. Typowa wielkość X2 różni się od wartości przeciętnej średnio o 154.6.
* Zmienna X3 - Minimalna liczba małżeństw wynosi 763, a wartość maksymalna 3906.

Współczynnik skośności wynosi 0.26. Typowa wielkość X2 różni się od wartości przeciętnej średnio o 698.74.

* Zmienna X4 – Minimalna ilość nakładów inwestycyjnych w przedsiębiorstwach jest równa 1573, a wartość maksymalna 7665. W zmiennej objaśniającej X3 średnia jest większa od mediany, co wskazuje na wystąpienie asymetrii prawostronnej, potwierdza to współczynnik skośności wynoszący 0.65. Typowa wielkość X3 różni się od wartości przeciętnej średnio o 1349.2.

Obraz zawierający linia, wzór

Opis wygenerowany automatycznie

Rysunek 2 Wykres rozrzutu

Na wykresie można zauważyć ciekawe zależności:

* Y i X1 odznaczają się ujemną korelacją, aczkolwiek jest dużo obserwacji odstających, które mogą zaburzać tę zależność
* Y i X3 wykazują dość mocną dodatnią korelację, punkty skupiają się wzdłuż szybko rosnącej linii prostej
* Y i X4 odznaczają się dodatnią korelacją
* Wykresy gęstości Y, X3,X4 cechują się podobnym rozkładem(prawostronna asymetria), co może sugerować korelację między nimi

Obraz zawierający tekst, zrzut ekranu, kwadrat, Prostokąt

Opis wygenerowany automatycznie

Rysunek 3 Macierz korelacji

Wzajemne korelacje zmiennych objaśniających oraz korelacje pomiędzy zmiennymi objaśniającymi X, a zmienną objaśnianą Y zostały przedstawione na rysunku 3. Wiemy, że  
z założeń budowania modelu, zmienne powinny być silnie skorelowane ze zmienną objaśnianą, natomiast słabo skorelowane między sobą. Wiele z tych zależności zostało wykryte i opisane przy okazji analizowania wykresu rozrzutu [rysunek 2], dlatego omówię te zależności, które były nieoczywiste i nie zostały rozważone, jak na przykład:

* X2 jest słabo skorelowane ze zmienną objaśnianą Y i najprawdopodobniej zmienna ta zostanie wyeliminowana na następnych etapach estymacji
* Najwyższą korelacją na poziomie 0.7 ze zmienną objaśnianą Y odznacza się zmienna X3, czyli liczba zawartych małżeństw
* Potencjalne zmienne objaśniające nie wykazują korelacji między sobą, co sugeruje, że nie wystąpi problem współliniowości predyktorów

# Wstępna Analiza modelu

Analiza regresji jest jedną z najczęściej stosowanych technik do analizy danych wieloczynnikowych. Jej szeroka atrakcyjność i użyteczność wynika z konceptualnie logicznego procesu użycia równania do wyrażenia związku między zmienną będącą przedmiotem zainteresowania (odpowiedzią) a zestawem powiązanych zmiennych predykcyjnych. Analiza regresji jest również interesująca ze strony teoretycznej ze względu na elegancką matematykę i dobrze rozwiniętą teorię statystyczną. Skuteczne wykorzystanie regresji wymaga zarówno zastosowania teorii, jak i praktycznych rozwiązań problemów, które zwykle pojawiają się, gdy technika ta jest stosowana z danymi ze świata rzeczywistego[[3]](#footnote-3).

W celu przeanalizowania modelu i oszacowania parametrów wykorzystam metodę najmniejszych kwadratów(KMNK). Ogólną postać modelu można zapisać następująco:

Ideą KMNK jest znalezienie takich wartości nieznanego wektora parametrów α, który minimalizuje sumę̨ kwadratów reszt, czyli różnic pomiędzy wartościami obserwowanymi,   
a teoretycznymi. Po oszacowaniu parametrów modelu ekonometrycznego:

Po próbach zmiany zmiennej X2 na logarytmy i kwadraty tych zmiennych model nie wykazywał poprawy w parametrach, dlatego usunąłem zmienną i stworzyłem model z trzema zmiennymi objaśniającymi.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Zmienna******Statystyka*** | **Współczynnik** | **Błąd standardowy** | **t-studenta** | **Wartość p** |
| **const** | −73,4994 | 180,406 | −0,4074 | 0,6852 |
| **X1** | −23,8889 | 12,4426 | −1,920 | 0,0597 |
| **X3** | 0,384993 | 0,048188 | 7,989 | 5,70e-11 |
| **X4** | 0,072221 | 0,024455 | 2,953 | 0,0045 |

Tabela 2 Model nr 1

Źródło: Własne opracowanie na podstawie wyników programu Gretl

Z tabeli 2, można odczytać, że największą wartość p-value ma X1 i wynosi ona 0.0597(wartość mniejsza niż poziom istotności o wartości 5%), co oznacza, że można wnioskować, iż zmienna X6 jest statystycznie nieistotna i jest kandydatką do wyeliminowana w modelu, co potwierdza statystyka t-studenta wynosząca -1.190(wartość krytyczna jest równa około 2), więc można uznać, że zmienna ma statystycznie nieistotny wpływ na zmienną objaśnianą Y, czyli jest nieistotna dla modelu.

Dla reszty zmiennych objaśniających w modelu, czyli X3,X4 wartość p-value jest mniejsza od wartości poziomu istotności, więc są to zmienne statystycznie istotne, co potwierdza statystyka t-studenta, która jest większa od wartości krytycznej, więc wnioskuje, że zmienne mają statystycznie istotny wpływ na zmienną objaśnianą Y.

|  |  |  |  |
| --- | --- | --- | --- |
| **Średn.aryt.zm.zależnej** | 915,9524 | **Odch.stand. zm.zależnej** | 410,8409 |
| **Suma kwadratów reszt** | 3742475 | **Błąd standardowy reszt** | 251,8567 |
| **Wsp.determ.R2** | 0,642382 | **Skorygowany R2** | 0,624198 |
| **F(5,59)** | 35,32678 | **Wartość p dla testu F** | 3,36e-13 |
| **Logarytm wiarygodności** | −435,6450 | **Kryt.inform.Akaike’a** | 879,29 |
| **Kryt. Bayes. Schwarza** | 887,8625 | **Kryt. Hannana - Quinna** | 882,6616 |

Tabela 3 Statystyki modelu 1

Źródło: Własne opracowanie na podstawie wyników programu Gretl

Muszę zredukować ilość zmiennych, przy użyciu metody Hellwiga oraz metody krokowej-wstecznej.

# Metoda krokowa-wsteczna

Metoda wsteczna została wprowadzona na początku lat 60-tych(Marill & Green, 1963). Jest to jedno z głównych podejść regresji krokowej. W statystyce metoda wsteczna jest metodą dopasowywania modeli regresji, w której wybór zmiennych predykcyjnych regresji odbywa się za pomocą automatycznej procedury. Obejmuje ona zaczynanie od wszystkich zmiennych kandydujących, testując usunięcie każdej zmiennej przy użyciu wybranego kryterium dopasowania modelu, usunięcie zmiennej (jeśli w ogóle), której utrata powoduje najbardziej nieistotne statystycznie pogorszenie dopasowania modelu i powtórzenie tej procedury do momentu, gdy nie będzie można usunąć kolejnych zmiennych bez statystycznie znaczącej utraty dopasowania. Celem selekcji jest zredukowanie zestawu zmiennych predykcyjnych do tych, które są niezbędne. Zasadniczo selekcja pomaga określić poziom ważności każdej zmiennej predykcyjnej. Pomaga również w ocenie efektów po statystycznym wyeliminowaniu innych zmiennych predykcyjnych[[4]](#footnote-4).

Chcąc zweryfikować, czy zbiór zmiennych można odrzucić przy opisie zmiennej objaśnianej można posłużyć się testem na odrzucenie pominiętych zmiennych, wykorzystując statystykę F. Test ten weryfikuje zasadność usunięcia z modelu wybranej zmiennej lub zmiennych[[5]](#footnote-5).

H0 : Parametry dla wskazanych zmiennych są równe 0.

H1 : Parametry dla wskazanych zmiennych nie są równe 0.

W modelu w pierwszej kolejności wykonuje test pominiętych zmiennych dla zmiennej X1.

Dla X1 wartość statystyki F(1,59) = 3.686 z wartością p = 0.0597. W takim razie nie ma podstaw do odrzucenia hipotezy zerowej, co oznacza usunięcie zmiennej X1 z modelu.

Tabela 4 Model nr. 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Zmienna******Statystyka*** | **Współczynnik** | **Błąd standardowy** | **t-studenta** | **Wartość p** |
| **const** | −308,867 | 135,284 | −2,283 | 0,026 |
| **X3** | 0,407942 | 0,047716 | 8,549 | 5,70e-12 |
| **X4** | 0,079294 | 0,024711 | 3,209 | 0,0021 |

Źródło: Własne opracowanie na podstawie wyników programu Gretl

|  |  |  |  |
| --- | --- | --- | --- |
| **Średn.aryt.zm.zależnej** | 915,9524 | **Odch.stand. zm.zależnej** | 410,8409 |
| **Suma kwadratów reszt** | 3976291 | **Błąd standardowy reszt** | 257,4326 |
| **Wsp.determ.R2** | 0,620039 | **Skorygowany R2** | 0,607374 |
| **F(2,60)** | 48,95546 | **Wartość p dla testu F** | 2,47e-13 |
| **Logarytm wiarygodności** | −437,5540 | **Kryt.inform.Akaike’a** | 881,108 |
| **Kryt. Bayes. Schwarza** | 887,5374 | **Kryt. Hannana - Quinna** | 883,6367 |

Tabela 5 Statystyki modelu 2

Źródło: Własne opracowanie na podstawie wyników programu Gretl

Pominięcie zmiennej poprawiło jedno kryterium informacyjne modelu. Porównując tabele 5 z tabelą 3 można zauważyć, że R2 i skorygowane R2 uległo tylko niewielkiemu pogorszeniu w stosunku do poprzedniego modelu.

Pozostałe zmienne objaśniające w modelu są istotne.

# Metoda Hellwiga

Metoda Hellwiga została opracowana przez Zdzisława Hellwiga w 1969 roku. Metoda ta pozwala na dobór zmiennych objaśniających do badania ekonometrycznego. Zmienne wybrane do liniowego modelu ekonometrycznego powinny być silnie skorelowane ze zmienną objaśnianą i ewentualnie słabo skorelowane ze zmienną zależną . Idea stojąca za tą metodą opiera się na kryterium liczbowym, które pozwala wybrać najlepszą kombinację zmiennych spośród potencjalnych kombinacji zmiennych objaśniających branych pod uwagę[[6]](#footnote-6).

Przy użyciu kodu własnego autorstwa w programie Gretl otrzymałem największą integralną pojemność informacyjną równą 0.59528 dla zmiennych objaśniających: {X3,X4}.

Ostateczny model otrzymany przy pomocy dwóch metoda jest taki sam.

# Wybór ostatecznej postaci modelu

Po próbach transformacji modelu korzystając z logarytmów, odwrotności i kwadratów zmiennych nie udało się stworzyć lepszego modelu niż ten opisany i wybrany metodą krokową-wsteczną oraz metodą Hellwiga we wcześniejszych rozdziałach.

# Opis i testowanie własności modelu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Zmienna******Statystyka*** | **Współczynnik** | **Błąd standardowy** | **t-studenta** | **Wartość p** |
| **const** | −308,867 | 135,284 | −2,283 | 0,026 |
| **X3** | 0,407942 | 4,77e-02 | 8,549 | 5,70e-12 |
| **X4** | 0,079294 | 0,024711 | 3,209 | 0,0021 |

Tabela 6 Finalna forma modelu

Źródło: Własne opracowanie na podstawie wyników programu Gretl

Tabela 7 Statystyki finalnego modelu

|  |  |  |  |
| --- | --- | --- | --- |
| **Średn.aryt.zm.zależnej** | 915,9524 | **Odch.stand. zm.zależnej** | 410,8409 |
| **Suma kwadratów reszt** | 3976291 | **Błąd standardowy reszt** | 257,4326 |
| **Wsp.determ.R2** | 0,620039 | **Skorygowany R2** | 0,607374 |
| **F(2,60)** | 48,95546 | **Wartość p dla testu F** | 2,47e-13 |
| **Logarytm wiarygodności** | −437,5540 | **Kryt.inform.Akaike’a** | 881,108 |
| **Kryt. Bayes. Schwarza** | 887,5374 | **Kryt. Hannana - Quinna** | 883,6367 |

Źródło: Własne opracowanie na podstawie wyników programu Gretl

## Współczynnik determinacji

Współczynnik determinacji wynosi 0.620039, co oznacza, że model w 62% wyjaśnia zmienność liczby osób poszkodowanych w wypadkach przy pracy. Biorąc pod uwagę ilość zmiennych i ilość obserwacji, model wyjaśnia zmienność średniego wynagrodzenia brutto w 60.74%. Różnica między tymi wskaźnikami nie jest duża, czyli jest wysokie zaufanie do interpretacji R2.

## Efekt katalizy

Przy wykorzystaniu kodu własnego autorstwa w programie Gretl, mogę wnioskować, że efekt katalizy nie występuje dla danych zmiennych objaśniających. Natężenie efektu katalizy dla tych zmiennych wynosi: 0.024756.

## Normalność rozkładu składnika losowego

Jednym z warunków poprawności modelu ekonometrycznego jest rozkład normalny składnika losowego. Istnieje wiele testów statystycznych, które pozwalają zweryfikować ten warunek. Posłużę się testem Doornika-Hansena, który wykonam w Gretlu i sprawdzę czy składnik losowy pochodzi z rozkładu normalnego.

H0 : Składnik losowy ma rozkład normalny

Obraz zawierający tekst, diagram, Wykres, zrzut ekranu

Opis wygenerowany automatycznie H1 : Składnik losowy nie posiada rozkładu normalnego

Rysunek 4 Test normalności rozkładu reszt

Na podstawie rysunku 4 wnioskuje, że wartość p-value w teście wynosi 0.3067, co jest większe od przyjętego poziomu istotności, który wynosi 5%. Z tego powodu nie ma podstaw do odrzucenia hipotezę zerowej ­- składnik losowy ma rozkład normalny.

## Istotność zmiennych

H0 :

H1 :

Wszystkie zmienne objaśniające oraz wyraz wolny mają p-value mniejsze od poziomu istotności, co oznacza, że są statystycznie istotne oraz ich wartość t-studenta jest większa od wartości krytycznej, więc wnioskuje, że mają one statystycznie istotny wpływ na zmienną objaśnianą Y.

## Obserwacje odstające

### Tabela 8 Obserwacje odstające

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Statystyka******Indeks*** | **Reszty** | **Leverage** | **Influence** | **DFFITS** |
| **10** | -408,56 | 0,144\* | -68,763 | -0,716 |
| **14** | -326,94 | 0,116\* | -42,892 | -0,493 |
| **27** | 147,71 | 0,119\* | 19,956 | 0,224 |
| **30** | 149,30 | 0,101\* | 16,805 | 0,204 |
| **41** | -8,54 | 0,113\* | -1,0869 | -0,012 |

Źródło: Własne opracowanie na podstawie wyników programu Gretl

Podregion 30(podregion miasto Wrocław) jest obserwacją nietypową, odstającą od reszty, ze względu, że w samym mieście Wrocław na rynku popularne są sektory takie jak przemysł, budownictwo, przetwórstwo przemysłowe[[7]](#footnote-7), które są sektorami szczególnie narażonymi na urazy oraz wypadki podczas pracy. Pominięcie obserwacji dźwigniowej dla podregionu Miasto Wrocław, mimo jego nietypowej natury, może prowadzić do zniekształcenia ogólnego obrazu dotyczącego wypadków przy pracy w regionie. Warto zachować tę obserwację, aby lepiej zrozumieć i monitorować specyficzne wyzwania związane z bezpieczeństwem pracy w tym obszarze.

Podregion 10(podregion rybnicki) jest obserwacją nietypową, ponieważ jest to region składający się głównie z populacji w wieku produkcyjnym, co przekłada się na to, że jest statystyczna większa szansa wypadku przy większej liczbie osób pracujących oraz jest to region skupiający się na pracy fizycznej, co przekłada się na większą liczbę wypadków.

Podregion 14(podregion Gorzowski) jest obserwacją odstającą od reszty, ze względu na to, że rejon skupiający się na rolnictwie, który nie jest narażony na wypadki przy pracy oraz w tym regionie znajduje się Gorzów Wielkopolski, który jest bardzo szybko rozwijającym się miastem zatrudniającym głównie w sektorach słabo narażonych na urazy oraz wypadki.

## Test serii

1. Mariusz Nyk*, Macroeconomic conditions of pay in Polish economy in 1990-2007,* Łódź 2009, s. 29 [↑](#footnote-ref-1)
2. Wojciech Stefan Zgliczyński, *Wynagrodzenia w Polsce,* Warszawa 2013, s. 1 [↑](#footnote-ref-2)
3. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining, *Introduction to Linear Regression Analysis 5th Edition*, New Jersey 2012, s.15 [↑](#footnote-ref-3)
4. Changxia Shao, Boehringer Ingelheim, *Backward Selection-a way to final model,* Philadelphia 2019, s.1 [↑](#footnote-ref-4)
5. https://mat.ug.edu.pl/~olanowel/pliki/e10.pdf [↑](#footnote-ref-5)
6. Gałecka Malgorzata, Smolny Katarzyna, *Evaluation of theater activity using Hellwig’s method,* Białystok 2018, s.41 [↑](#footnote-ref-6)
7. https://www.gowork.pl/poradnik/17/pracodawcy/najwieksi-pracodawcy-we-wroclawiu/ [↑](#footnote-ref-7)