AKADEMIA GÓRNICZO-HUTNICZA

im. Stanisława Staszica w Krakowie

Wydział Zarządzania

Badanie czynników wpływających na wynagrodzenie administratorów baz danych w USA w 2019 roku.

INFORMATYKA I EKONOMETRIA

Ekonometria

Bartłomiej Kalata

Prowadzący: mgr. Aneta Bech

Kraków, czerwiec 2019

Spis treści

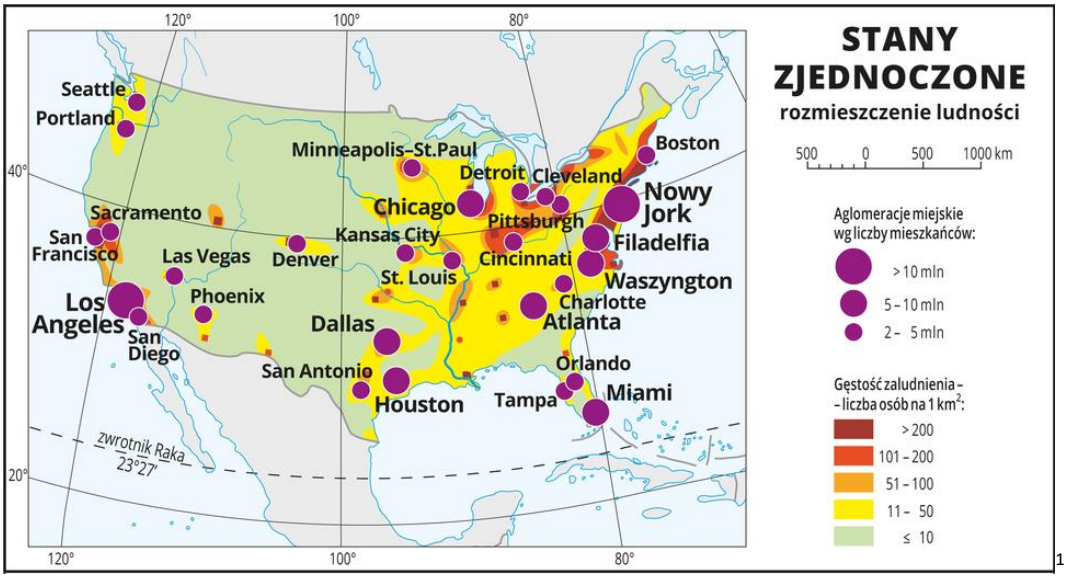
Wstęp

DBA DEVELOPER

Większość firm i instytucji nie mogłoby funkcjonować bez bazy danych. W dzisiejszych czasach niemal wszystkie bazy gromadzone są na dyskach w postaci zapisu cyfrowego. Coraz częściej poszukiwanym zawodem jest wykwalifikowany programista baz danych, który zaprojektuje stabilne i niezawodne bazy danych, zgodnie z potrzebami dostosowanymi do danej firmy. Jest on odpowiedzialny głównie za tworzenie, testowanie, ulepszanie i utrzymywanie nowych i istniejących baz danych, aby pomóc użytkownikom w efektywnym pozyskiwaniu danych. W ramach zespołów IT ściśle współpracuje z programistami w celu zapewnienia spójności systemu, jak również z administratorami i klientami w celu zapewnienia wsparcia technicznego i identyfikacji nowych wymagań. Kluczowe znaczenie na tym stanowisku mają umiejętności komunikacyjne i organizacyjne oraz podejście do rozwiązywania problemów. Jego zadania różnią się więc znacznie od typowego programisty.

Cel projektu

Celem tego projektu jest zbadanie poszczególnych czynników wchodzących w skład wynagrodzenia netto administratora baz danych w Stanach Zjednoczonych w 2019 roku. Kraj ten jest największą potęgą gospodarczą, dlatego więc zarobki sięgają tam najwyższych szczebli. Dane do projektu pochodzą ze strony www.brentozar.com. Są to wyniki corocznej ankiety zebrane od profesjonalnych developerów baz danych na przestrzeni całego kraju. Na podstawie badań naukowych nad determinantami zarobków wśród programistów/administratorów baz danych, wybrano 12 potencjalnych kandydatów do zmiennych objaśniających, które zostaną opisane w następnym rozdziale. Jednym z czynników determinujących zarobki są wielkości miejscowości. W Stanach Zjednoczonych możemy podzielić je na wielkie metropolie, miasta, oraz mniejsze miejscowości. Poniższa mapa przedstawia rozłożenie znaczących miast w USA.


[1]

Rysunek 1 Rozmieszczenie ludności USA

[1] Źródło: https://epodreczniki.pl/

W ankiecie podział na wielkość miejscowości odnosił się do 5 kategorii:

- 100K-299K (miasto)
- 300K-1M (duże miasto)
- 1M+ (metropolia)
- 20K-99K (duża miejscowość)
- <= 20,000 (miasteczko)

Podział wielkości zaludnienia użyty do modelu zostanie opisany w następnym rozdziale.

Odnosząc się do ankiet internetowych z amerykańskich stron takich jak: www.glassdoor.com / www.payscale.com / www.itcareerfinder.com jednym z ważniejszych czynników determinujących zarobki jest również region zamieszkania. Poniższa mapa przedstawia rozłożenie pasów przemysłowych w Stanach Zjednoczonych.

[2]

Rysunek 2 Rozmieszczenie okręgów przemysłowych USA

Do podziału na regiony zostały użyte kody pocztowe ameryki zgodnie z wzorem początkowych numerów które odnoszą się do poszczególnych stanów. Podział ten zostanie opisany w następnym rozdziale.

[2] Źródło: https://epodreczniki.pl/

Aby przeprowadzić dogłębną analizę determinantów zarobków, wskazówek udziela artykuły naukowy „Employment and Salaries of Recent Doctorates in Computer Sciences”[3], który odpowiada między innymi na takie na pytania:

1. Gdzie znajdują zatrudnienie współcześni ludzie z wyższym wykształceniem informatycznym?
2. Jaką znajdują pracę?
3. Jak dużo obowiązków zostaje im narzuconych?
4. Jakie są ich przeciętne startowe zarobki?
5. Jak długo muszą poszukiwać pracy?

W przeprowadzonym badaniu brało udział 123 ankietowanych. W ten sposób uzyskano następujące wyniki:

- Stwierdzono bardzo niskie bezrobocie, które jest zgodne z krajowymi danymi na temat nauk ścisłych i inżynierii (S&E).
- Wszyscy respondenci mieli pełne etaty. Jednak 28% stwierdziło, że pozycje były tymczasowe (tj. miały określoną datę końcową), a trzy czwarte z nich było na stanowiskach podoktorskich. Ponad połowa (53%) osób na stanowiskach tymczasowych stwierdziła, że odpowiednia stała praca nie jest dostępna.
- Pracownicy pracujący w przemyśle/biznesie mają najwyższe zarobki w dużych przedsiębiorstwach.

Na podstawie tych badań do wstępnej bazy zmiennych objaśniających zostało włączone wykształcenie, lata doświadczenia, certyfikaty oraz rodzaj zatrudnienia, rodzaj przedsiębiorstwa , czy zaangażowanie danego pracownika związane z jego mobilnością. Do jednych z mniej oczywistych charakterystyk z powodu coraz bardziej zacieranych granic różnicy pomiędzy zarobkami jest cecha płci. Jednak z uwagi na to, że większość developerów baz danych jest mężczyznami, zmienna ta również znalazła miejsce wśród tej bazy.

Badania nad zarobkami[4] przeprowadził również Departament Pracy USA „USUAL WEEKLY EARNINGS OF WAGE AND SALARY WORKERS FIRST QUARTER 2019”[5] który wykazał podobne zależności wykształcenia, regionu pracy, lat doświadczenia co powyższy artykuł. Dodatkowymi determinantami były również sezonowość pracy, grupy etniczne, różnica płci. Dzięki tym badaniom dokonano ostatecznego wyboru zmiennych objaśniających.

[3] „Employment and salaries of recent doctorates in computer science”-Maisel,Herbert,Gaddy,Catheriner-11.97r.

[4] https://www.bls.gov/bls/wages.htm

[5] https://www.bls.gov/cps/earnings.htm#education

Opis zmiennych

1. Zmienna objaśniana ilościowa
    - Y: Roczne wynagrodzenie administratora baz danych netto w dolarach US ($), bez opodatkowania.
2. Zmienne objaśniające:
    a. Ilościowe:
        - X1: Ilość przepracowanych średnio godzin tygodniowo. (l. godzin/tydzień)
        - X2: Liczba lat przepracowanych na jednym stanowisku. (Lata doświadczenia na danym stanowisku).(l. lata)
        - X3: Liczba dni w których pracownik dodatkowo pracował zdalnie po godzinach pracy. (l. dni/tydzień) Zmienna ta określa zarządzanie bazą danych zdalnie, oraz mobilność pracownika.
    b. Binarne:
        - X4: Wykształcenie wyższe informatyczne. (1- „Tak” / 0 - „Nie”).

Zmienna ta powstała z 5 różnych odpowiedzi respondentów, tj. Czy posiadają oni stopień licencjat/inżynier/magister/doktor oraz czy ich wykształcenie jest powiązane z informatyką. Zmienna ta odpowiada na pytanie posiadania jakiegokolwiek stopnia naukowego związanego z informatycznym wykształceniem. Wyjaśnienie powstania tej zmiennej zostanie opisane szerzej podczas sprawdzenia współczynnika determinacji R^2.

- X5: Płeć (1- „Mężczyzna” / 0 - „Kobieta”)
- X6: Ważne certyfikaty informatyczne. (1- „Tak” /0- „Nie”) -
  Posiadanie tzw. „Industrial Certificates” o kierunku IT, wzmacnia CV, zachęca do wyższych pensji i pomaga w utrzymaniu pracy.
- X7: Przedsiębiorstwo prywatne/państwowe. (1-„Prywatne”/ 0 – „Państwowe”)

c. Kategoryczne:

Do zmiennych kategorycznych zostały przypisane warunki demograficzno-geograficzne tj. zaludnienie w miejscowości, w której znajduje się biuro, jak również podział USA na 4 regiony. Podział ten został dokonany na podstawie dodatkowych danych w ankiecie tzn. na podstawie kodu pocztowego, gdzie najliczniejszą grupą wyników były pochodzące z północno-wschodniej części kraju. Oto podział kraju na regiony wg. Mapy przedstawionej na wstępie:

- Północno-Wschodnia część, w której znajduje się tak zwany pas przemysłowy, oraz największe miasta i metropolie.
- Południowy Wschód tzw. „Sun Belt”
- Centralna Część USA
- Zachodnie Wybrzeże gdzie znajdują się również największe ośrodki przemysłowe takie jak np. „Dolina Krzemowa” gdzie znajdują się potęgi światowego IT.

Jako kategorię, do której zostaną porównane wnioski wybrano wielkość miasta od 100 tys. do 1 miliona mieszkańców, oraz region posiadający początek kodu pocztowego od 00 do 20, tj. region nad wschodnim wybrzeżem. Zawiera on takie miejscowości jak: New York, Boston, Washington, Philadelphia, New Jersey.

- X8: Wielkość miejscowości, w której znajduje się biuro. (powyżej 1 miliona mieszkańców) (1-„Tak”/0-„Nie”).
- X9: Wielkość miejscowości, w której znajduje się biuro. (poniżej 100 tys. mieszkańców) (1-„Tak”/0-„Nie”).
- X10: Początek kodu od 21 do 39 tj. południowo-wschodnia część kraju
- X11: Początek kodu od 55 do 89 tj. centralna część kraju
- X12: Początek kodu od 90 do 99 tj. zachodnie wybrzeże

Podstawowe Statystyki

Przygotowanie danych do modelowania

Wczytane dane do pliku zostały wyfiltrowane zgodnie z regułą 3 sigm, która wskazuje te obserwacje które nie mieszczą się w przedziale 3*odchylenie standardowe od średniej, w celu poszukiwania outlierów. Funkcja, wykonywująca ten filtr jest dostępna w skrypcie. Nie wykazała ona żadnych wartości odstających. Dane tak obrobione zostały poddane wstępnym obliczeniom podstawowych statystyk.

Statystyki

Podstawowe statystyki zostały wykonane za pomocą funkcji „describe", dostępnej w pakiecie „psych".

	ŚREDNIA	ODCHYLENIE STANDARDOWE	MEDIANA	MIN	MAX	SKOŚNOŚĆ	KURTOZA
Y	105785,48	26496,99	104000	60000	190000	0,55	-0,15
X1	43,77	6,80	40	20	100	2,25	12,47
X2	8,82	6,91	7	1	38	1,07	0,66
X3	1,32	1,78	1	0	5	1,19	-0,07
X4	0,53	0,5	1	0	1	-0,11	-1,99
X5	0,9	0,3	1	0	1	-2,68	5,17
X6	0,46	0,5	0	0	1	0,15	-1,98
X7	0,81	0,39	1	0	1	-1,57	0,46
X8	0,37	0,48	0	0	1	0,56	-1,69
X9	0,11	0,32	0	0	1	2,46	4,05
X10	0,14	0,34	0	0	1	2,11	2,44
X11	0,28	0,45	0	0	1	0,98	-1,04
X12	0,12	0,32	0	0	1	2,37	3,65

Tabela 1 Podstawowe statystyki opisowe

Y: Roczne wynagrodzenie administratora baz danych netto w dolarach US ($), bez opodatkowania.

Średnia a mediana: średnia tych danych jest różna od mediany o 1785,48$ co oznacza, że w tych danych zauważyliśmy istotne zmienne odstające wpływające na tak dużą różnicę. Odchylenie standardowe wynosi 26496,99. Wysoka wartość odchylenia standardowego świadczy o dużym rozproszeniu wyników wokół średniej. Skośność wynosi 0,55 (0 to symetryczny) co wskazuje na asymetrię prawostronną tzn. więcej danych z lewej strony, dłuższy prawy ogon wykresu danych. Kurtoza wynosi -0,15 czym możemy wnioskować małe zagęszczenie danych wokół średniej.

X1: Ilość przepracowanych średnio godzin tygodniowo. (l. godzin/tydzień)

Średnia a mediana: średnia tych danych jest różna od mediany o 3.77h co oznacza, że w tych danych zauważyliśmy istotne zmienne odstające wpływające na tak dużą różnicę. Odchylenie standardowe wynosi 6.80. Wysoka wartość odchylenia standardowego świadczy o dużym rozproszeniu wyników wokół średniej. Skośność wynosi 2.25 (0 to symetryczny) co wskazuje na asymetrię prawostronną tzn. więcej danych z lewej strony, dłuższy prawy ogon wykresu danych. Kurtoza wynosi 12,47> 0 czym możemy wnioskować duże zagęszczenie danych wokół średniej.

X2: Liczba lat przepracowanych na jednym stanowisku. (Lata doświadczenia na danym stanowisku).(l. lata)

Średnia a mediana: średnia tych danych jest różna od mediany o 1,82 lat co oznacza, że w tych danych zauważyliśmy istotne zmienne odstające wpływające na tak dużą różnicę. Odchylenie standardowe wynosi 6.91. Wysoka wartość odchylenia standardowego świadczy o dużym rozproszeniu wyników wokół średniej. Skośność wynosi 1,07 (0 to symetryczny) co wskazuje na asymetrię prawostronną tzn. więcej danych z lewej strony, dłuższy prawy ogon wykresu danych. Kurtoza wynosi 0.66 > 0 czym możemy wnioskować średnie zagęszczenie danych wokół średniej.

X3: Liczba dni przepracowanych w domu zdalnie. (l. dni/tydzień)

Średnia a mediana: średnia tych danych jest różna od mediany o 0.31 dnia co oznacza, że w tych danych nie zauważyliśmy istotne zmienne odstające wpływające na tak dużą różnicę. Odchylenie standardowe wynosi 1.78. Niska wartość odchylenia standardowego świadczy o małym rozproszeniu wyników wokół średniej. Skośność wynosi 1.19 (0 to symetryczny) co wskazuje na asymetrię prawostronną tzn. więcej danych z lewej strony, dłuższy prawy ogon wykresu danych. Kurtoza wynosi -0.07 < 0 czym możemy wnioskować małe zagęszczenie danych wokół średniej.

X4: Wykształcenie wyższe informatyczne. (1- „Tak" / 0 - „Nie").

Porównanie średniej z medianą w tym przypadku nie ma sensu, ponieważ jest to zmienna binarna. Średnia wynosi 0.53 co świadczy o tym, iż w próbce została zawarta podobna ilość osób posiadających certyfikaty oraz nieposiadających. Odchylenie standardowe wynosi 0.5– w porównaniu do średniej (0,53) ma to sens, w związku z faktem, że jest to zmienna binarna.

X5: Płeć (1- „Mężczyzna" / 0 - „Kobieta")

Porównanie średniej z medianą w tym przypadku nie ma sensu, ponieważ jest to zmienna binarna. Średnia wynosi 0,9 co świadczy o tym, iż w próbce wystąpiło więcej mężczyzn niż kobiet. Odchylenie standardowe wynosi 0.3, co oznacza że jest bardzo małe odchylenie od średniej.

X6: Ważne certyfikaty informatyczne. (1- „Tak" /0- „Nie")

Kolejny przykład w którym porównanie średniej z medianą nie ma sensu, ponieważ jest to zmienna binarna. Średnia wynosi 0.46 co świadczy o tym, iż w próbce mniej więcej połowa ankietowanych posiada ważny certyfikat. Odchylenie standardowe wynosi 0,5, o czym świadczy małe odchylenie wynikó• od średniej.

X7: Przedsiębiorstwo prywatne/państwowe. (1-„Prywatne"/ 0 – „Państwowe")

Dla tej zmiennej binarnej średnia wynosi 0.81 co świadczy o tym, iż w próbce więcej respondentów pracuje w prywatnych firmach. Odchylenie wynoszące 0,39 dowodzi że większość wyników jest jednostronna.

X8: Wielkość miejscowości, w której znajduje się biuro. (powyżej 1milion mieszkańców) (1-„Tak"/0- „Nie").

Dla tej zmiennej kategorycznej, przy założeniu że pracownik pracuje w mieście o wielkości od 100 tys. do 1 miliona mieszkańców średnia wynosi 0.37 co wskazuje że pracowników z metropolii jest mniej.

X9: Wielkość miejscowości, w której znajduje się biuro. (poniżej 100 tys. mieszkańców) (1-„Tak”/0-„Nie”).

Dla kolejnej zmiennej kategorycznej, przy założeniu że pracownik pracuje w mieście o wielkości od 100 tys. do 1 miliona mieszkańców średnia wynosi 0.11 co wskazuje że pracowników z małych miast jest o wiele mniej.

X10: Początek kodu od 21 do 39 tj. południowo-wschodnia część kraju

X11: Początek kodu od 55 do 89 tj. centralna część kraju

X12: Początek kodu od 90 do 99 tj. zachodnie wybrzeże

Zmienne kategoryczne X10 oraz X12 posiadają bardzo podobne statystyki świądczące o tym że w tych częściach kraju znajdowała się bardzo niewielka ilość ankietowanych.

Zmienna X11 posiada bardziej znaczące dane tj. średnią na poziomie 0.28, co może być spowodowane tym że jest to największy z obranych obszarów. Na tej podstawie stwierdzono, że zmienna do której zmienne te zostaną porównane ma najbardziej znaczące statystki.

Współczynnik zmienności V

Współczynnik zmienności sprawdza czy w obserwacjach następuje dostatecznie duża zmienność, aby użyć je w modelu.

Poniższe zmienne objaśniające ilościowe:

Y: Roczne wynagrodzenie administratora baz danych netto w dolarach US ($), bez opodatkowania.

$$V_Y \approx 25\%$$

X1: Ilość przepracowanych średnio godzin tygodniowo. (l. godzin/tydzień)

$$V_{X1} \approx 16\%$$

X2: Liczba lat przepracowanych na jednym stanowisku. (Lata doświadczenia na danym stanowisku).(l. lata)

$$V_{X2} \approx 78\%$$

X3: Liczba dni przepracowanych w domu zdalnie. (l. dni/tydzień)

$$V_Y \approx 135\%$$

Posiadają współczynniki zmienności V powyżej 15% wobec czego możemy stwierdzić, że mają dosyć dużą zmienność, aby użyć je w modelu. Zmienne binarne nie potrzebują interpretacji tego współczynnika, abyśmy mogli ich użyć w modelu.

Macierz korelacji z zmienną objaśnianą Y

X1	X2	X3	X4
0,2	0,36	0,13	-0,02
X5	X6	X7	X8
0,04	0,08	0,15	0,26
X9	X10	X11	X12
-0,13	0,05	0,00	0,07

Tabela 2 Korelacja zmiennych z Y

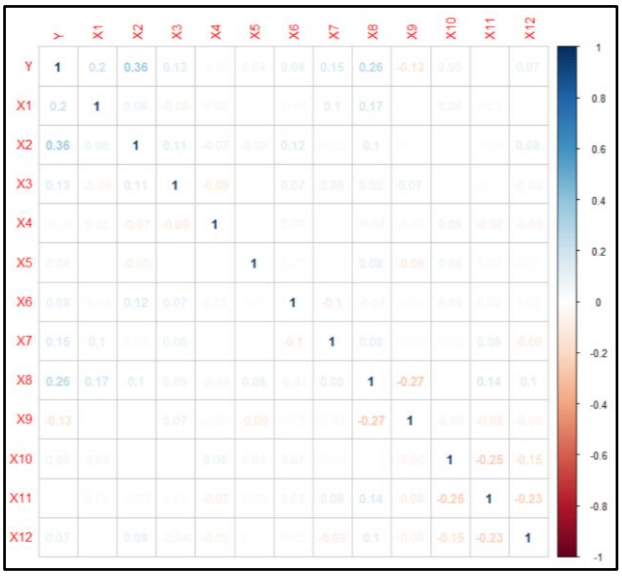
Zmienna oznaczająca liczbę lat doświadczenia (**X2**) posiada największą korelację (0.36), co oznacza, że to czy pracownik posiada dużą liczbę lat doświadczenia ma największe znaczenie dla jego zarobków. Kolejnymi ważnymi zmiennymi są : liczba przepracowanych godzin tygodniowo (**X1**) (0.2), oraz biuro znajdujące się w metropolii (**X8)**(0.27). Oprócz zmiennych **X3,X7** oraz **X9** gdzie wartości znajdują się w przedziale (0.2-0.1), wartości korelacji są znikome co świadczy o słabym powiązaniu ich ze zmienną **Y.**

W modelu MNK spodziewać się więc można występowania właśnie tych zmiennych.

Macierz korelacji zmiennych objaśniających

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Y	1												
X1	0.2	1											
X2	0.36	0.06	1										
X3	0.13	-0.06	0.11	1									
X4	-0.02	0.03	-0.07	-0.09	1								
X5	0.04	0	-0.05	-0.01	-0.01	1							
X6	0.08	-0.03	0.12	0.07	0.03	0.01	1						
X7	0.15	0.1	0.02	0.06	0	-0.01	-0.1	1					
X8	0.26	0.17	0.1	0.05	-0.04	0.08	-0.04	0.08	1				
X9	-0.13	-0.01	-0.01	0.07	-0.03	-0.09	0.01	-0.02	-0.27	1			
X10	0.05	0.04	0	0	0.06	0.04	0.04	-0.02	0	-0.06	1		
X11	0	0.02	-0.02	0.01	-0.07	0.02	0.03	0.08	0.14	-0.08	-0.25	1	
X12	0.07	0	0.08	-0.04	-0.05	0.01	0.02	-0.09	0.1	-0.06	-0.15	-0.23	1

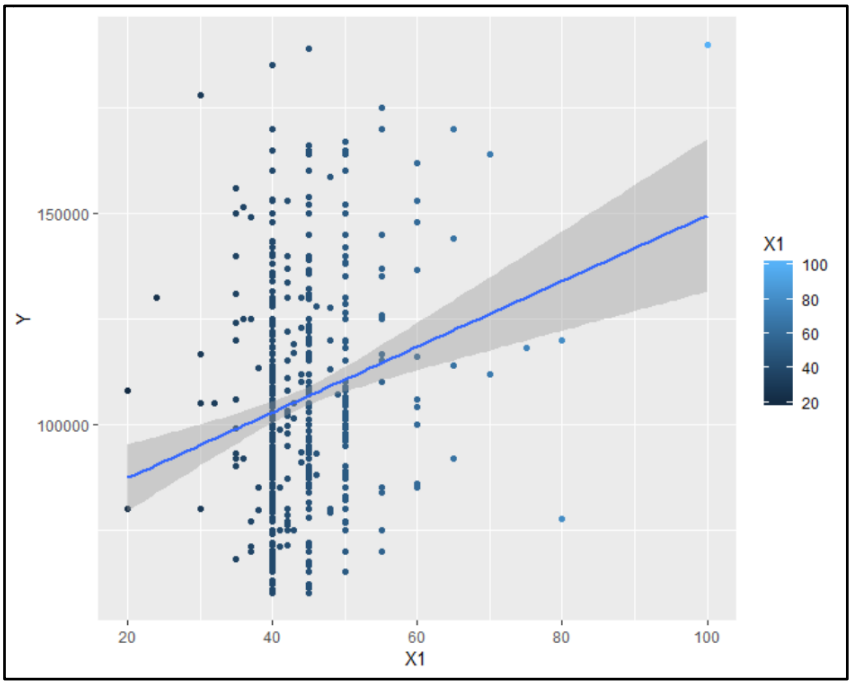
Tabela 3 Korelacje zmiennych



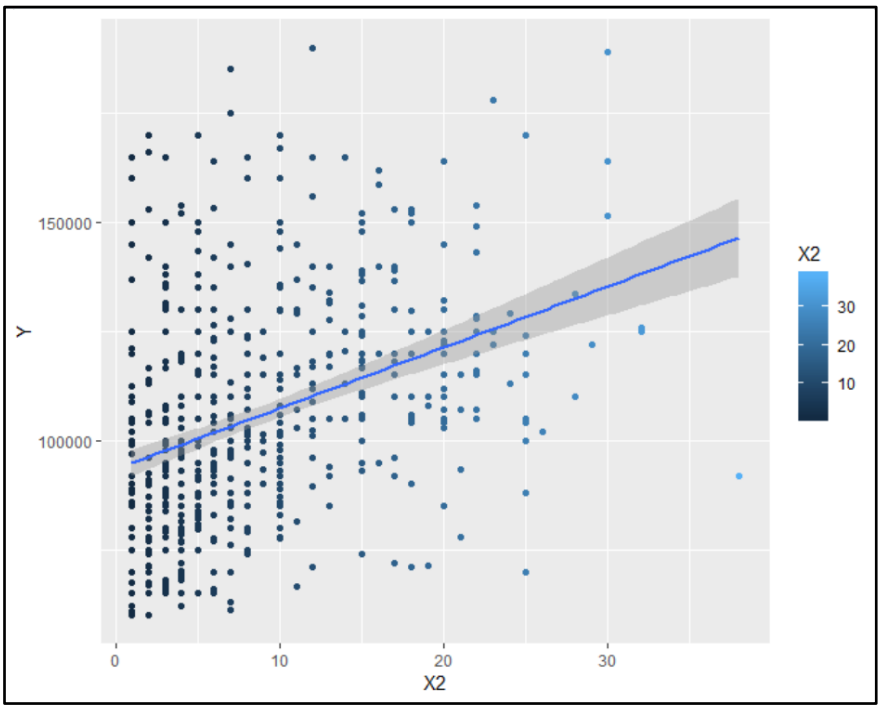
Rysunek 3 Wykres numeryczny korelacji z Y

Wykresy zależności zmiennych X do zmiennej Y

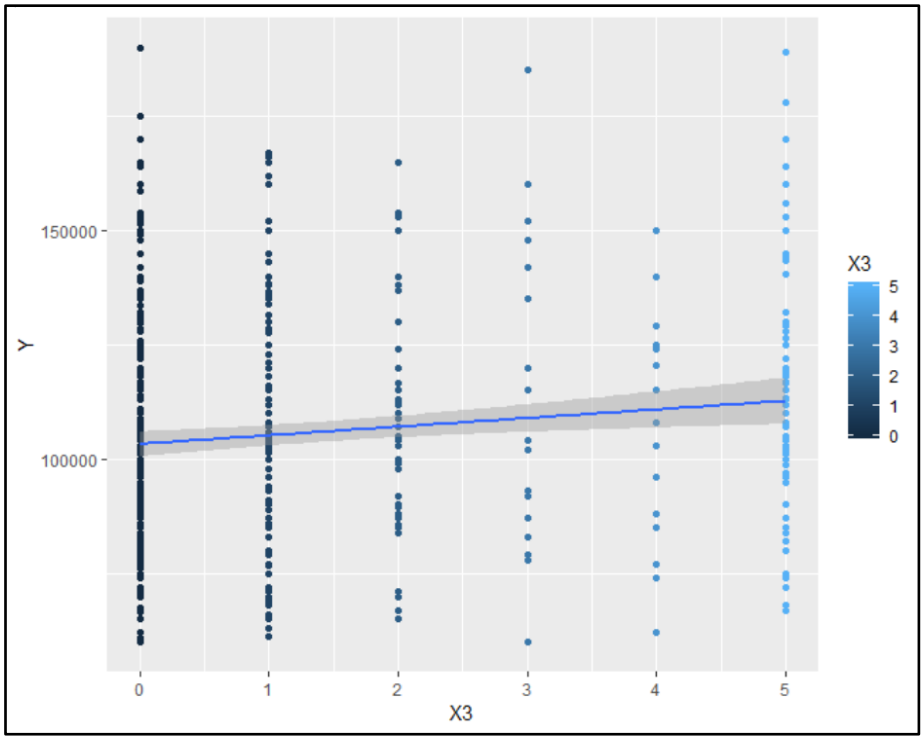
Poniżej zostały przedstawione wykresy zależności od Y Zmiennych ilościowych X1, X2, X3. Pozostałe zmienne to zmienne binarne, lub kategoryczne zmienione na binarne, których wykresy nie wskazują na żadną zależność. Na wykresach tych zauważamy dodatnią zależność zmiennych X w stosunku do zmiennej Y.



Rysunek 4 Zarobki wzgl. Liczby przepracowanych godzin.



Rysunek 5 Zarobki względem lat doświadczenia na stanowisku.



Rysunek 6 Zarobki względem liczby przepracowanych dodatkowo dni po godzinach pracy.

Model MNK

Klasyczna metoda najmniejszych kwadratów to określenie na wszystkie warunki stosowalności MNK do szacowania wektora α w modelu

$$Y = X\alpha + \epsilon$$

Równanie 1

Przedstawiony poniżej model został wyestymowany metodą prób i błędów, ponieważ przy pierwszej próbie współczynnik determinacji wynosił zaledwie 8%. Po pierwszej modyfikacji zmiennych (zamiast 5 zmiennych odnoszących się osobno do poziomu wyższego wykształcenia, oraz do nawiązania do informatyki połączone zostały one w jedną binarną **X4**, co spowodowało podniesienie współczynnika o 7%. Kolejnym krokiem transformacji zmiennych było dodanie zmiennych kategorycznych odnośnie wielkości miasta, oraz podziału kraju na regiony. Ta ostateczna konwersja spowodowała podniesienie współczynnika o kolejne 10%. Na tym etapie zakończono dodawanie i konwertowanie zmiennych objaśniających.

```
Call:
lm(formula = Y ~ ., data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-44219 -16558  -3208  13443  84489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54287.0     7634.9   7.110 3.67e-12 ***
X1             558.2      149.7   3.730 0.000212 ***
X2            1218.8      147.2   8.281 9.61e-16 ***
X3            1306.4      569.4   2.294 0.022152 *
X4             480.8     2013.8   0.239 0.811396
X5            3269.0     3348.3   0.976 0.329353
X6            3301.5     2027.8   1.628 0.104080
X7            8177.5     2576.5   3.174 0.001590 **
X8            9619.1     2227.1   4.319 1.86e-05 ***
X9           -6025.0     3314.9  -1.818 0.069683 .
X10           2665.3     3078.5   0.866 0.386990
X11          -1087.4     2430.5  -0.447 0.654775
X12           2925.3     3312.2   0.883 0.377522
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23360 on 542 degrees of freedom
Multiple R-squared:  0.2394,    Adjusted R-squared:  0.2226
F-statistic: 14.22 on 12 and 542 DF,  p-value: < 2.2e-16
```

Rysunek 7 Model MNK

W modelu MNK z wszystkimi zmiennymi objaśniającymi największą istotność sprawdzoną testem t-Studenta mają **X1, X2, X7, X8**. Na tej podstawie można stwierdzić ze reszta przypuszczeń z poprzedniego podrozdziału potwierdziły się, ponieważ są to zmienne posiadające jedne z największych korelacji z zmienną **Y**. Model ten opisuje rzeczywistość w około 24%. Skorygowany wsp. determinacji wynosi 0.22, jednak nie został on brany pod uwagę ze względu na to, że liczba obserwacji znacznie przewyższa liczbę zmiennych. Wyraz wolny wskazany został za istotny statystycznie na poziomie istotności poniżej 0,1%.

Dobór zmiennych objaśniających

Metoda Hellwiga

Opis metody:

M_k- zestaw zmiennych w kombinacji k-tej (istnieje 2p-1 ilości kombinacji gdzie p jest liczbą zmiennych)
r_j – korelacja pomiędzy Y i X_j
r_{ij} – korelacja pomiędzy X_i i X_j

Wtedy:

$$h_j = \sum_{j \in m_k} \left(\frac{r^2}{\sum_{i \in mk} |r_{ij}|} \right)$$

Równanie 2

Oraz:

$$H = \sum_{j=1} h_{ij}$$

Równanie 3

Należy wybrać ten zestaw zmiennych o największej pojemności integralnej. Metoda Hellwiga przeprowadzona została przez funkcję stworzoną na zajęciach. Po przeprowadzeniu algorytmu doboru zmiennych objaśniających metodą Hellwiga najlepszą kombinacją zmiennych okazały się:

"X1" "X2" "X7" "X8" "X9" "X10"

O optymalnej pojemności integralnej 0.2151. Oto podsumowanie modelu:

```
Call:
lm(formula = Y ~ X1 + X2 + X7 + X8 + X9 + X10, data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-44854 -16650  -3698  13685  85503

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  61750.1     6723.6   9.184  < 2e-16 ***
X1             524.5      149.7   3.504 0.000496 ***
X2            1285.0      145.0   8.861  < 2e-16 ***
X7            7773.2     2553.6   3.044 0.002447 **
X8           10013.5     2192.1   4.568 6.09e-06 ***
X9           -5609.3     3291.6  -1.704 0.088919 .
X10           3039.7     2904.8   1.046 0.295822
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23460 on 548 degrees of freedom
Multiple R-squared:  0.2246,    Adjusted R-squared:  0.2161
F-statistic: 26.46 on 6 and 548 DF,  p-value: < 2.2e-16
```

Rysunek 8 Model dobrany metodą Hellwiga

Zauważono, że w tej konfiguracji zmienne X10 oraz X9 są najmniej istotnymi zmiennymi. Podsumowanie tego modelu, zostenie opisane po przeprowadzeniu metody krokowej doboru zmiennych.

Metoda krokowa wsteczna

Polega na odrzucaniu zmiennych z modelu podstawowego, przy jednoczesnym sprawdzeniu warunku normalności reszt. W rzeczywistości posługuje się ona istotnością zmiennych przy wyborze zmiennej do odrzucenia. W języku R oparta jest na kryterium informacyjnym AIC.

```
Start:  AIC=11178.21
Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 +
    X12

       Df  Sum of Sq        RSS   AIC
- X4    1 3.1110e+07 2.9586e+11 11176
- X11   1 1.0925e+08 2.9594e+11 11176
- X10   1 4.0913e+08 2.9624e+11 11177
- X12   1 4.2575e+08 2.9625e+11 11177
- X5    1 5.2024e+08 2.9635e+11 11177
<none>               2.9583e+11 11178
- X6    1 1.4468e+09 2.9727e+11 11179
- X9    1 1.8031e+09 2.9763e+11 11180
- X3    1 2.8731e+09 2.9870e+11 11182
- X7    1 5.4981e+09 3.0133e+11 11186
- X1    1 7.5917e+09 3.0342e+11 11190
- X8    1 1.0182e+10 3.0601e+11 11195
- X2    1 3.7426e+10 3.3325e+11 11242
```

Krok 1 Usunięcie zmiennej X4

```
Step:  AIC=11176.27
Y ~ X1 + X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12

       Df  Sum of Sq        RSS   AIC
- X11   1 1.1867e+08 2.9598e+11 11174
- X12   1 4.1397e+08 2.9627e+11 11175
- X10   1 4.1556e+08 2.9627e+11 11175
- X5    1 5.1601e+08 2.9637e+11 11175
<none>               2.9586e+11 11176
- X6    1 1.4705e+09 2.9733e+11 11177
- X9    1 1.8263e+09 2.9768e+11 11178
- X3    1 2.8433e+09 2.9870e+11 11180
- X7    1 5.5067e+09 3.0136e+11 11184
- X1    1 7.6260e+09 3.0348e+11 11188
- X8    1 1.0161e+10 3.0602e+11 11193
- X2    1 3.7432e+10 3.3329e+11 11240
```

Krok 2 Usunięcie zmiennej X11

```
Step:  AIC=11174.49
Y ~ X1 + X2 + X3 + X5 + X6 + X7 + X8 + X9 + X10 + )

       Df  Sum of Sq        RSS   AIC
- X5    1 5.0821e+08 2.9649e+11 11173
- X12   1 6.0372e+08 2.9658e+11 11174
- X10   1 6.1918e+08 2.9660e+11 11174
<none>               2.9598e+11 11174
- X6    1 1.4252e+09 2.9740e+11 11175
- X9    1 1.7654e+09 2.9774e+11 11176
- X3    1 2.8474e+09 2.9882e+11 11178
- X7    1 5.4501e+09 3.0143e+11 11183
- X1    1 7.6185e+09 3.0360e+11 11187
- X8    1 1.0056e+10 3.0603e+11 11191
- X2    1 3.7540e+10 3.3352e+11 11239
```

Krok 3 Usunięcie zmiennej X5

```
Step:  AIC=11173.44
Y ~ X1 + X2 + X3 + X6 + X7 + X8 + X9 + X10 + X1

       Df  Sum of Sq        RSS   AIC
- X12   1 6.1149e+08 2.9710e+11 11173
- X10   1 6.6581e+08 2.9715e+11 11173
<none>               2.9649e+11 11173
- X6    1 1.4483e+09 2.9793e+11 11174
- X9    1 1.9128e+09 2.9840e+11 11175
- X3    1 2.8455e+09 2.9933e+11 11177
- X7    1 5.4179e+09 3.0190e+11 11182
- X1    1 7.6019e+09 3.0409e+11 11186
- X8    1 1.0367e+10 3.0685e+11 11190
- X2    1 3.7152e+10 3.3364e+11 11237
```

Krok 4 Usunięcie zmiennej X12

```
Step:  AIC=11172.59
Y ~ X1 + X2 + X3 + X6 + X7 + X8 + X9 + X10

       Df  Sum of Sq        RSS   AIC
- X10   1 4.9838e+08 2.9760e+11 11172
<none>               2.9710e+11 11173
- X6    1 1.4785e+09 2.9858e+11 11173
- X9    1 2.0018e+09 2.9910e+11 11174
- X3    1 2.7332e+09 2.9983e+11 11176
- X7    1 5.1068e+09 3.0220e+11 11180
- X1    1 7.5581e+09 3.0465e+11 11184
- X8    1 1.0898e+10 3.0799e+11 11191
- X2    1 3.8046e+10 3.3514e+11 11238
```

Krok 5 Usunięcie zmiennej X10

```
Step:  AIC=11171.52
Y ~ X1 + X2 + X3 + X6 + X7 + X8 + X9

       Df  Sum of Sq        RSS   AIC
<none>               2.9760e+11 11172
- X6    1 1.5493e+09 2.9914e+11 11172
- X9    1 2.1359e+09 2.9973e+11 11174
- X3    1 2.7568e+09 3.0035e+11 11175
- X7    1 5.0452e+09 3.0264e+11 11179
- X1    1 7.7392e+09 3.0533e+11 11184
- X8    1 1.0819e+10 3.0841e+11 11189
- X2    1 3.7959e+10 3.3555e+11 11236
```

Krok 6 Model ostateczny

W związku z tym, że próbka posiada dużą liczbę obserwacji zgodnie z twierdzeniem CTG możemy założyć normalność rozkładu reszt. Na tej podstawie możemy założyć że w metodzie tej nie ma potrzeby sprawdzania normalności rozkładu reszt przy każdej estymacji.

Metoda ta tworzy model o zmiennych:

X1 X2 X3 X6 X7 X8 X9

Oto podsumowanie modelu metody krokowej:

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X6 + X7 + X8 + X9, data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-45790 -16183  -2565  13709  84346

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  58001.3     6827.7   8.495  < 2e-16 ***
X1             562.8      149.2   3.772  0.00018 ***
X2            1219.4      146.0   8.353 5.50e-16 ***
X3            1273.8      565.9   2.251  0.02478 *
X6            3402.8     2016.4   1.688  0.09207 .
X7            7784.1     2556.2   3.045  0.00244 **
X8            9750.5     2186.5   4.459 9.98e-06 ***
X9           -6497.7     3279.3  -1.981  0.04805 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23320 on 547 degrees of freedom
Multiple R-squared:  0.2349,    Adjusted R-squared:  0.2251
F-statistic: 23.99 on 7 and 547 DF,  p-value: < 2.2e-16
```

Zauważono w nim zmienne o małej istotności takie jak X3, X6 czy X9.

Wybór modelu

Do wyboru modelu po doborze zmiennych metodą krokową oraz Hellwiga posłużono się następującymi kryteriami:

- Współczynnik determinacji pozwala zmierzyć, w jakim stopniu model umożliwia objaśnienie zmienności zmiennej Y.
- Kryteria informacyjne Aikaike'a oraz Bayesa-Shwarza, których ideą jest dostarczenie miary stanowiącej równowagę między stopą dopasowania a oszczędną specyfikacją modelu, czyli jego prostotą.

	Metoda Krokowa	Metoda Hellwiga
Wsp. Determinacji	23,49%	22,46%
Kryt. Akaike	12748,54	12753,93
Kryt. Bayesa-Shwarza	12787,41	12788,48

Ponieważ korelacje pomiędzy zmiennymi nie przekraczają poziomu 0,2 oraz w powyższym porównaniu wszystkie kryteria są na podobnym poziomie, model krokowy wydaje się być nieznacznie lepszym modelem. Z powodu małej różnicy pomiędzy danymi zmiennymi (które okazują się być najistotniejsze) oraz na postawie tych kryteriów wybrany został model dobrany metodą krokową.

Postać analityczną takiego wstępnego modelu można zapisać jako:

$$Y = 58001{,}3 + 562{,}8 * X1 + 1219{,}4 * X2 + 1273{,}8 * X3 + 3402{,}8 * X6 + 7784{,}1 * X7 + 9750{,}5 * X8 - 6497{,}7 * X9$$

Diagnostyka wybranego modelu

Ocena istotności

Ocena istotności parametrów strukturalnych ma na celu zbadanie, czy zmienne objaśniające w istotny sposób wpływają na ukształtowanie się zmiennej objaśnianej Y. Hipotezę zerową oznaczającą nieistotność zmiennej, odrzucamy, jeśli prawdopodobieństwo zaobserwowania wartości większej lub równej od tj jest mniejsze niż ustalony poziom istotności wynoszący 0.05. Innymi słowy, jeśli prawdopodobieństwo, że wyznaczona na podstawie danych empirycznych wartość statystyki testowej jest bardzo małe, wnioskujemy że hipoteza jest fałszywa.

$$H0: \beta_j = 0$$

$$H1\ \beta_j \neq 0$$

W tabeli zostały przedstawione wartości p-value dla testu t-Studenta dla oceny istotności każdej zmiennej z modelu stworzonego metodą krokową.

Const	X1	X2	X3	X6	X7	X8	X9
2e-16	0,00018	5,50e-16	0,02478	0,09207	0,00244	9,98e-06	0,04805

Oznacza to że tylko niektóre z tych zmiennych są istotne. W modelu tym zmienne X3,X6,X9 nie wykazały istotności w teście t-Studenta. Powołując się na twierdzenie CTG oznaczające normalność i losowość reszt

w modelu o dużej próbie obserwacji, zmienne ta zostaną usunięte. Poniżej przedstawione zostaną modele z usuniętymi kolejno nieistotnymi zmiennymi.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X7 + X8 + X9, data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-43552 -16779  -3048  13709  82736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59778.6     6757.3   8.846  < 2e-16 ***
X1             559.9      149.5   3.746 0.000199 ***
X2            1248.2      145.2   8.594  < 2e-16 ***
X3            1335.8      565.7   2.362 0.018546 *
X7            7346.1     2547.2   2.884 0.004082 **
X8            9602.3     2188.4   4.388 1.37e-05 ***
X9           -6504.2     3284.9  -1.980 0.048197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23360 on 548 degrees of freedom
Multiple R-squared:  0.2309,    Adjusted R-squared:  0.2225
F-statistic: 27.42 on 6 and 548 DF,  p-value: < 2.2e-16
```

Etap 1 Usunięcie zmiennej X9

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X7 + X8, data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-46685 -16754  -2756  13434  84010

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  59344.9     6771.7   8.764  < 2e-16 ***
X1             545.9      149.7   3.647  0.00029 ***
X2            1247.6      145.6   8.568  < 2e-16 ***
X3            1237.1      564.9   2.190  0.02897 *
X7            7362.7     2554.0   2.883  0.00410 **
X8           10799.0     2108.9   5.121 4.22e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23430 on 549 degrees of freedom
Multiple R-squared:  0.2254,    Adjusted R-squared:  0.2184
F-statistic: 31.95 on 5 and 549 DF,  p-value: < 2.2e-16
```

Etap 2 Usunięcie zmiennej X3

```
Call:
lm(formula = Y ~ X1 + X2 + X7 + X8, data = dane)

Residuals:
   Min     1Q Median     3Q    Max
-49251 -16718  -3652  13892  86075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  61440.5     6726.9   9.134  < 2e-16 ***
X1             520.3      149.7   3.475 0.000552 ***
X2            1281.4      145.3   8.819  < 2e-16 ***
X7            7701.3     2558.1   3.011 0.002727 **
X8           11025.8     2113.6   5.217 2.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23510 on 550 degrees of freedom
Multiple R-squared:  0.2186,    Adjusted R-squared:  0.213
F-statistic: 38.48 on 4 and 550 DF,  p-value: < 2.2e-16
```

Etap 3 Finalny model z istotnymi zmiennymi

Za pomocą testu Walda możemy zweryfikować też hipotezę o łącznej istotności wszystkich zmiennych objaśniających w modelu. Dla testu statystycznego F dla łącznej istotności p-value wyniosło 2,2e-16 co daje wynik, który potwierdza łączną istotność wszystkich zmiennych.

Model po przetestowaniu istotności można przedstawić w postaci analitycznej:

$$Y = 61440{,}5 + 520{,}3 * X1 + 1281{,}4 * X2 + 7701{,}3 * X7 + 11025{,}8 * X8$$

Efekt katalizy

Współczynnik determinacji jest miarą dopasowania modelu ekonometrycznego do danych empirycznych, lecz informacja, jaką niesie o modelu, może być fałszywa, jeśli w modelu występują zmienne, które nazywamy katalizatorami.

ZAŁOŻENIE: (R,R0)- regularna para korelacyjna. Zmienna Xi z pary zmiennych $(Xi, Xj), i < j$, jest katalizatorem jeżeli:

$$r_{ij} < 0 \text{ lub } r_{ij} > r_i/r_j$$

Zmienną Xi należy wtedy wyeliminować z modelu i ponownie oszacować parametry strukturalne modelu. Na podstawie skryptu z zajęć został sprawdzony efekt katalizy w wybranym modelu. Nie wykazał on żadnych katalizatorów w powyższym modelu.

Test Breush-Pagan na heteroskedastyczność

Heteroskedastyczność składnika losowego modelu jest odstępstwem od klasycznych założeń MNK. Oznacza to, że choć składniki losowe są wzajemnie nieskorelowane, to mają różne wariancje. W konsekwencji otrzymujemy estymator parametrów modelu, który pozostaje estymatorem nieobciążonym, liniowym, i zgodnym, ale nie jest estymatorem najefektywniejszym w klasie estymatorów liniowych i nieobciążonych. Zespół hipotez przyjmuje zatem postać:

$$H0: \sigma_i^2 = const$$

$$H1: \sigma_i^2 \neq const$$

```
        studentized Breusch-Pagan test

data:  hml
BP = 6.9289, df = 4, p-value = 0.1397
```

P-value w teście okazało się być powyżej wartości potwierdzającej homoskedastyczność wobec czego nie zostaną przeprowadzone transformacje modelu w celu usunięcia tejże wady.

Koincydencja

Wybrany model ma własność koincydencji, jeżeli zachodzi warunek $sgn\ (ri) = sgn\ (ai)$ dla $i = 1,2,\dots,k$. Tylko wtedy parametry strukturalne mają oceny sensowne ze względu na znak. Jeżeli dla pewnego i, $sgn\ (ri) \neq sgn\ (ai)$, to model nie ma własności koincydencji. Ocena ai nie jest sensowna ze względu na znak. Zmienną Xi należy wtedy wyeliminować z modelu i ponownie oszacować parametry strukturalne modelu.

	Znak przy współczynniku „a”	Znak przy korelacji z zmienną Y
X1	+	+
X2	+	+
X7	+	+
X8	+	+

Warunek został spełniony, zatem model nasz posiada własność koincydencji. Oznacza to, że wraz ze wzrostem poziomu / lub występowania poniższych zmiennych:

X2: Liczba lat przepracowanych na jednym stanowisku. (Lata doświadczenia na danym stanowisku).(l. lata)

X7: Przedsiębiorstwo prywatne/państwowe. (1-„Prywatne”/ 0 – „Państwowe”)

X8: Wielkość miejscowości, w której znajduje się biuro. (powyżej 1milion mieszkańców) (1-„Tak”/0-„Nie”).

Wzrastają zarobki wśród administratorów baz danych.

Współliniowość

Współliniowość jest cechą zbioru danych statystycznych, wykorzystywanych do szacowania parametrów modelu ekonometrycznego, pojawiającą się, kiedy zmienne objaśniające są ze sobą silnie skorelowane. Do jej oszacowania wykorzystujemy współczynnik VIF- (Variance Inflation Factor). Współczynnik VIF gdy nie wyniósł powyżej wartości 10, oznacza że nie występuje zjawisko wpółliniowości. W modelu nie występuje więc zjawisko współliniowości.

X1	X2	X7	X8
1,039268	1.011648	1.015540	1.040897

Normalność rozkładu składnika resztowego

Jednym z warunków poprawności modelu ekonometrycznego i stosowania go w praktyce jest rozkład normalny reszt, przy czym istnieje wiele testów statystycznych pozwalających na weryfikacje spełnienia tego warunku. Podczas weryfikacji modelu, zostały użyte dwa spośród nich: Jarque-Bera oraz Shapiro Wilka.

Hipotezy:

H0- składnik resztowy posiada rozkład normalny,

H1- składnik resztowy nie posiada rozkładu normalnego.

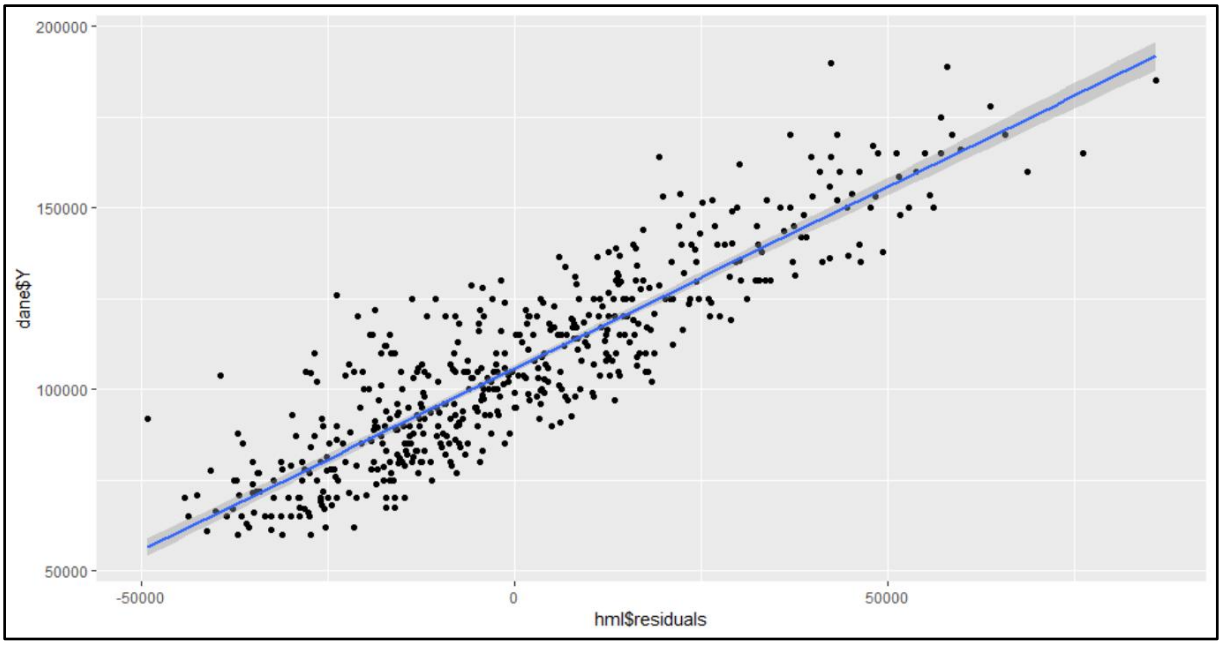
```
        Shapiro-Wilk normality test

data:  hml$residuals
W = 0.97071, p-value = 4.393e-09
```

```
        Jarque-Bera Normality Test

data:  hml$residuals
JB = 39.183, p-value = 3.101e-09
alternative hypothesis: greater
```

W próbach nieskończonych lub dostatecznie dużych, założenie o normalności nie jest zbyt restrykcyjne, ponieważ z Centralnego Twierdzenia Granicznego (CTG) wiemy, że suma n niezależnych zmiennych losowych o jednakowych, aczkolwiek dowolnych rozkładach, zbiega według rozkładu do rozkładu normalnego. W praktyce twierdzenie to interpretuje się następująco: dla dostatecznie dużej liczby zmiennych losowych o rozkładzie ze średnią μ i wariancją σ^2 ich suma ma rozkład $N(n\mu, n\sigma 2)$. Wobec tego, niezależnie od rozkładu składnika losowego asymptotyczne rozkłady statystyk testowych będą poprawne o ile zmienne są niezależne (co będzie sprawdzone w następnym kroku weryfikacji). Poniższy wykres przedstawia zależność reszt od zmiennej objaśnianej Y. Stwierdzono w modelu rozkład normalny składnika resztowego.



Rysunek 9 Wykres reszt w zależności od zmiennej Y

Test liczby serii

Sformułowanie jednorównaniowego modelu ekonometrycznego jest równoważne z przyjęciem założenia o liniowej zależności zmiennej objaśnianej od zmiennej objaśniającej. Weryfikacja tego założenia jest niezbędna do prawidłowej interpretacji współczynnika determinacji. Test serii to nieparametryczny test losowości próby.

H0: dobór jednostek do próby jest losowy; model jest liniowy.

H1: dobór jednostek do próby nie jest losowy; model jest nieliniowy.

```
        Runs Test

data:  hml$residuals
statistic = 0.68039, runs = 286, n1 = 277, n2 = 277, n = 554, p-value = 0.4963
alternative hypothesis: nonrandomness
```

Ponieważ p-value w teście liczby serii jest większa od poziomu istotności $\alpha = 0.05$, została przyjęta hipoteza zerowa. Stwierdzono że model liniowy jest dobrze dobrany oraz że próbka została losowo dobrana.

Test Ramsey'a RESET

Test RESET jest bardzo ogólnym testem poprawności konstrukcji model. Wiele funkcji nieliniowych można przybliżyć za pomocą wielomianów. Jeśli zatem dopasowanie do zbioru zmiennych objaśniających ich wyższych potęg znacząco poprawi dopasowanie modelu, wskazuje to na złe dobranie jego początkowej postaci funkcyjnej. Został więc on zastosowany w celu sprawdzenia, czy to liniowa postać modelu (względem funkcji kwadratowej, lub sześciennej) jest najlepszym możliwym do wybrania modelem.

Hipotezy:

H0- $\beta_{k+1} + \beta_{k+2} = 0$ – opisuje to sytuację w której specyfikacja modelu jest poprawna, tzn. model został poprawnie dobrany.

H1- Model nie jest poprawnie dobrany, istnieje lepsza postać wielomianowa.

```
        RESET test

data:  hml
RESET = 2.0247, df1 = 2, df2 = 548, p-value = 0.133
```

Ponieważ p-value w teście Ramsey'a jest większa od poziomu istotności α, oznacza to że model został poprawnie dobrany.

Test Chowa

Szereg danych może często zawierać strukturalną przerwę. Aby ją przetestować, wykorzystany zostanie test Chowa. Wykorzystuje on test F w celu określenia, czy pojedyncza regresja jest bardziej wydajna niż dwie oddzielne regresje obejmujące podział danych na dwie podpróbki.
Hipotezy:

H0 : $\beta_1 = \beta_2 = \cdots = \beta_n$- parametry są takie same w podpróbach

H1 : $\beta_1 \neq \beta_2 \neq \cdots \neq \beta_n$- parametry różnią się w podpróbach.

```
        M-fluctuation test

data:  hml
f(efp) = 0.96784, p-value = 0.8391
```

Jak możemy zauważyć p-value jest większe niż α, została więc przyjęta hipoteza zerowa o stabilności parametrów.

Autokorelacja test Durbina-Watsona

Zjawisko autokorelacji jest konsekwencją niespełnienia założenia czwartego MNK dla liniowego modelu ekonometrycznego. W takim przypadku estymator a wektora parametrów α jest mało efektywny (tzn. wariancje estymatorów są stosunkowo duże).

Założenia dla testu Durbina-Watsona:

- model ekonometryczny posiada wyraz wolny,
- składnik losowy ma rozkład normalny,
- w modelu nie występuje opóźniona zmienna objaśniana jako zmienna

Hipotezy:

$$H0: \rho = 0$$

$$H1: \rho > 0$$

```
        Durbin-Watson test

data:  hml
DW = 2.0203, p-value = 0.595
alternative hypothesis: true autocorrelation is greater than 0
```

Oznacza to brak podstaw do odrzucenie hipotezy H0 co oznacza brak występowania autokorelacji w modelu.

Prognoza

Prognoza punktowa ex post

Rzeczywiste wartości	Prognozowane wartości	Błąd bezwzględny	Błąd względny
102000	128475,3	-26475,3	-0,25956157
80000	101565	-21565	-0,26956238
93000	108695,1	-15695,1	-0,16876409
170000	133016,4	36983,57	0,217550412
75000	93317,3	-18317,3	-0,24423067
80000	102769	-22769	-0,28461263
152000	108707,7	43292,26	0,2848175
75000	101603,7	-26603,7	-0,354716
105000	133016,4	-28016,4	-0,26682314
140321	111231,9	29089,09	0,207303896
121000	102261,8	18738,19	0,154861074
120000	131812,4	-11812,4	-0,09843675
123000	117664,7	5335,29	0,043376341
89000	88660,46	339,54	0,003815056
135000	107972,2	27027,8	0,200205926
107000	102769	4230,99	0,039541963
78052	98443,07	-20391,1	-0,26124981
91000	96349,12	-5349,12	-0,05878154
88000	88660,46	-660,46	-0,00750523
80000	88595,71	-8595,71	-0,10744638

Tabela 4 Prognozowane wartości

Błąd prognozy (predykcji) ex post

Błąd MAPE, wskazuje na procentowy błąd prognozy EX POST i jest on obliczany wzorem:

$$MAPE = \frac{1}{n} * \sum \left| \frac{y_t - y_t^p}{y_t} \right| * 100\%$$

Błąd dla tej prognozy wynosi 17%. Oznacza to że prognoza myli się średnio od wartości rzeczywistej właśnie o 17%.

Podsumowanie/Wnioski

Interpretacja parametrów modelu

$$Y = 61440{,}5 + 520{,}3 * X1 + 1281{,}4 * X2 + 7701{,}3 * X7 + 11025{,}8 * X8$$

Współczynnik α_0 wynosi 61440,5, co oznacza że minimalnie pracownik który pracuje w średniej wielkości miejscowości, o minimalnym doświadczeniu, w państwowym przedsiębiorstwie, minimalną liczbę godzin tygodniowo, otrzyma 61440$ rocznej wypłaty netto.

Współczynnik α_1 przy zmiennej X1 wynosi 520.3.Oznacza to, że gdy pracownik będzie pracował o godzinę dłużej tygodniowo, **ceteris paribus,** to jego zarobki wzrosną o 520$/rocznie.

Współczynnik α_2 przy zmiennej X2 wynosi 1281,4 . Oznacza to, że gdy pracownik będzie miał o jeden rok więcej doświadczenia na danym stanowisku, **ceteris paribus,** to jego zarobki wzrosną o 1281,4$/rocznie.

Współczynnik α_3 przy zmiennej X7 wynosi 7701,3. Oznacza to, że gdy pracownik będzie pracował w prywatnym przedsiębiorstwie, **ceteris paribus,** to jego zarobki wzrosną o 7701,3$/rocznie.

Współczynnik α_4 przy zmiennej X8 wynosi 11025,8. Oznacza to, że biuro w którym pracuje administrator baz danych będzie znajdować się w metropolii, **ceteris paribus,** to jego zarobki wzrosną o 111025,8$/rocznie.

Podsumowanie

Na podstawie przeprowadzonego badania wykazano najważniejsze determinanty zarobkowe, odnoszące się do wyników ankiety. Model przewiduje zarobek na podstawie 4 najważniejszych cech pracownika z dokładnością 83%. Okazały się nimi: liczba przepracowanych tygodniowo godzin, lata doświadczenia, rodzaj przedsiębiorstwa, oraz położenie biura. Wynagrodzenie średnio netto wg. www.payscale.com wynosi około 71000$ rocznie. Można więc przyjąć że model który został tutaj przedstawiony ukazuje zarobki na odpowiednim poziomie.

Spis tabel / rysunków

Bibliografia

- „Employment and salaries of recent doctorates in computer science”- Maisel,Herbert,Gaddy,Catheriner
- “Ekonometria I Badania Operacyjne”- M.Gruszczyński, T.Kuszewski, M.Podgórska
- „Ekonometria”- prof. Henryk Gurgul- Wydział Zarzadzania Samodzielna Pracownia Zastosowań Matematyki w Ekonomii
- Ekonometria – G. S. Maddala
- https://www.bls.gov/bls/wages.htm - „USUAL WEEKLY EARNINGS OF WAGE AND SALARY WORKERS FIRST QUARTER 2019 ”badania nad zarobkami Departamentu Pracy USA. 01.2019r.
- https://www.brentozar.com/ Ankieta Zarobkowa z której pobrano dane
- Przykładowe ankiety zarobkowe: www.glassdoor.com / www.payscale.com / www.itcareerfinder.com
- Źródło map: https://epodreczniki.pl/