AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

**WYDZIAŁ ZARZĄDZANIA**

# Zastosowanie regresji liniowej do badania wpływu wybranych czynników na liczbę nowych przedsiębiorstw w powiatach Polsce.

Autor: *Izabela Guła*

Kierunek: *Informatyka i ekonometria (stacjonarne), rok II*

Przedmiot: *Ekonometria*

Kraków, 2024

# Spis treści

# 1. Część I

## 1. Cel projektu

Celem niniejszej pracy jest przeprowadzenie analizy czynników wpływających na liczbę nowo zarejestrowanych przedsiębiorstw w systemie REGON w powiatach w Polsce. Projekt ma na celu określenie, które z wybranych zmiennych mają największy wpływ na poziom liczbę przedsiębiorstw i w jaki sposób ten wpływ się objawia. W pierwszej części projektu przedstawione zostaną zmienne, ich statystyki opisowe oraz zależności, które pozwolą wprowadzić do tematu pracy. Kolejnym krokiem będzie opracowanie modelu regresji, który pozwoli na prognozowanie liczby przedsiębiorstw w powiatach na podstawie wybranych zmiennych objaśniających. Pozwoli to na weryfikacje hipotez badawczych dotyczących zależności między zmiennymi objaśniającymi a liczbą nowych przedsiębiorstw. Ostatnim krokiem będzie analiza statystyczna i interpretacja wyników w celu wyciągnięcia praktycznych wniosków. Przeprowadzenie szczegółowych testów statystycznych pozwoli na ocenę jakości modelu i jego zastosowanie w rzeczywistych scenariuszach planowania.

## 2. Hipotezy badawcze

1. Większe bezrobocie przyczynia się do zwiększenia liczby nowych przedsiębiorstw. Wysokie bezrobocie może prowadzić do tzw. przymuszonej przedsiębiorczości, gdzie osoby niemogące znaleźć pracy decydują się na założenie własnych firm.
2. Większa liczba mieszkań oddanych do użytku przyczynia się do zwiększenia liczby nowych przedsiębiorstw. Nowe mieszkania mogą przyciągać więcej mieszkańców do danego obszaru, co zwiększa popyt na różne produkty i usługi, prowadząc do zakładania nowych przedsiębiorstw.
3. Wyższy współczynnik urbanizacji przyczynia się do zwiększenia liczby nowych przedsiębiorstw. Obszary miejskie często mają lepszą infrastrukturę, dostęp do zasobów i sieci biznesowych, co sprzyja zakładaniu nowych firm.
4. Wyższa gęstość zaludnienia przyczynia się do zwiększenia liczby nowych przedsiębiorstw. Wyższa gęstość zaludnienia oznacza większą koncentrację potencjalnych klientów na danym obszarze, co zwiększa popyt na produkty i usługi oferowane przez nowe przedsiębiorstwa.
5. Większa liczba rozwodów przyczynia się do zwiększenia liczby nowych przedsiębiorstw. Rozwody mogą prowadzić do zmiany sytuacji finansowej osób

zaangażowanych, co może skłonić niektóre osoby do zakładania własnych firm w celu zwiększenia dochodów.

6. Większa liczba osób poszkodowanych w wypadkach przy pracy przyczynia się do zmniejszenia liczby nowych przedsiębiorstw. Wypadki przy pracy mogą mieć negatywny wpływ na morale pracowników oraz efektywność pracy, co może utrudniać rozwój nowych firm.

## 3. Opis danych

Do badania empirycznego wykorzystany został zbiór danych utworzony przeze mnie z danych dostępnych z bazy danych Głównego Urzędu Statystycznego. Plik z danymi zawiera informacje z 2022 roku dla wszystkich 314 powiatów w Polsce bez miast na prawach powiatów. Miasta na prawach powiatów zostały nieuwzględnione w projekcie, gdyż miały bardzo duże wartości odstające, przez co zaburzały analizę. Wybrane zmienne to:

- Bezrobocie – określa stopę bezrobocia rejestrowanego. Jest to stosunek liczby bezrobotnych zarejestrowanych do liczby cywilnej ludności aktywnej zawodowo.
- Mieszkania – dane określające mieszkania oddane do użytkowania, liczba podana na 1000 mieszkańców.
- Urbanizacja – dane określające procentowy udział mieszkańców miast w ogólnej liczbie ludności.
- Przedsiębiorstwa – liczba określająca nowo zarejestrowane w rejestrze REGON podmioty gospodarki narodowej.
- Gęstość zaludnienia – dane określające gęstość zaludnienia, czyli średnią ilość mieszkańców przypadającą na kilometr kwadratowy na danym terenie.
- Rozwody – dane określające liczbę zawartych rozwodów w powiatach w 2022 roku.
- Wypadki – dane określające liczbę osób poszkodowanych w wypadkach przy pracy.

Dane zostały połączone w Excelu, a następnie eksportowane do Gretla, gdzie można rozpocząć część badawczą projektu.

## 4. Statystyki opisowe

| | Średnia | Mediana | Minimalna | Maksymalna |
|---|---|---|---|---|
| bezrobocie | 7,8719 | 7,1000 | 1,0000 | 24,600 |
| mieszkania | 4,9588 | 4,4000 | 0,90000 | 22,200 |
| urbanizacja | 39,767 | 39,800 | 0,0000 | 78,200 |
| rozwody | 114,44 | 95,000 | 14,000 | 534,00 |
| gestosczaludnien~ | 98,645 | 75,400 | 18,200 | 722,60 |
| przedsiebiorstwa | 660,50 | 529,00 | 108,00 | 5047,0 |
| wypadki | 112,81 | 91,000 | 7,0000 | 1122,0 |

| | Odch.stand. | Wsp. zmienności | Skośność | Kurtoza |
|---|---|---|---|---|
| bezrobocie | 4,0240 | 0,51119 | 0,92853 | 0,80509 |
| mieszkania | 2,9927 | 0,60351 | 2,0553 | 6,1325 |
| urbanizacja | 16,617 | 0,41785 | -0,094607 | -0,64367 |
| rozwody | 68,797 | 0,60115 | 1,7799 | 5,3086 |
| gestosczaludnien~ | 79,316 | 0,80405 | 3,2452 | 15,985 |
| przedsiebiorstwa | 524,26 | 0,79374 | 3,5466 | 19,782 |
| wypadki | 95,570 | 0,84716 | 4,4372 | 39,160 |

| | Percentyl 5% | Percentyl 95% | Zakres Q3-Q1 | Brakujące obs. |
|---|---|---|---|---|
| bezrobocie | 2,7000 | 15,730 | 5,3500 | 0 |
| mieszkania | 1,5700 | 11,520 | 2,7500 | 0 |
| urbanizacja | 11,540 | 66,600 | 24,550 | 0 |
| rozwody | 39,700 | 253,00 | 82,000 | 0 |
| gestosczaludnien~ | 31,840 | 244,12 | 60,850 | 0 |
| przedsiebiorstwa | 222,50 | 1505,2 | 421,50 | 0 |
| wypadki | 23,700 | 264,10 | 96,000 | 0 |

Tab. 1 Statystyki opisowe zmiennych

Interpretacja statystyk opisowych dla każdej ze zmiennych:

- Bezrobocie

Średnia wartość bezrobocia w Polsce w 2022 dla powiatów wynosiła 7,87. Wartość środkowa wyniosła 7,1. Powiat z najmniejszą stopą bezrobocia rejestrowanego miał wartość wynoszącą 1, a największa wartość wyniosła 24,6. Odchylenie standardowe wyniosło 4,02. Wartość współczynnika zmienności wyniosła 0,51 i oznacza silną zmienność. Skośność zmiennej sugeruje prawostronną skośność i wydłużony prawy ogon rozkładu. Kurtoza wynosi 0,8, oznacza rozkład platokurtyczny. Percentyl 5% wynosi 2,7, oznacza to, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych wartości 2,7. Percentyl 95% wynosi 15,730, oznacza to, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 15,730. Środkowe 50% danych w zestawie ma rozpiętość 5,35 jednostek. Zmienna nie ma brakujących wartości.

- Mieszkania

Średnia liczba mieszkań oddanych do użytkowania wyniosła 4,96. Wartością środkową jest liczba 4,4. Najmniejszą liczbę mieszkań zarejestrowano w powiecie gdzie wyniosła ona 0,9,

czyli 900 mieszkań. Natomiast największa liczba wynosi 22,2, czyli 22 200 mieszkań. Odchylenie od średniej wynosi 2,99 i można uznać to za małe odchylenie od średniej. Współczynnik zmienności wynosi 0,6 i oznacza silną zmienność. Współczynnik skośności wynosi 2,05 i oznacza prawostronną skośność i wydłużony prawy ogon rozkładu. Kurtoza wynosiła 2 i oznacza rozkład platokurtyczny. Percentyl 5% wynosi 1,6, oznacza to, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych 1,6. Percentyl 95% wynosi 11,52, oznacza to, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 11,52. Środkowe 50% danych w zestawie ma rozpiętość 2,75 jednostek. Zmienna nie posiada brakujących wartości.

- Gęstość zaludnienia

W 2022 roku średnio w Polsce w powiatach mieszkało 98,64 osób na kilometr kwadratowy. Wartość środkowa wyniosła 75,4. Najmniejszą wartością okazuje się powiat, w którym mieszkało tylko 18 osób na kilometr kwadratowy, a największą wartość zarejestrowano dla powiatu, którego gęstość zaludnienia wyniosła 722,6. Odchylenie od średniej wyniosło 79,32. Współczynnik zmienności osiąga wysokie wartości i wskazuje na silna zmienność. Skośność zmiennej sugeruje asymetrię prawostronną i wydłużony prawy ogon rozkładu. Wartość współczynnika kurtozy równa 15,98 wskazuje na rozkład leptokurtyczny i większe skupienie danych wokół średniej. Percentyl 5% wynosi 31,8, oznacza to, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych wartości 31,8. Percentyl 95% wynosi 244,12 oznacza to, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 244,12. Środkowe 50% danych w zestawie ma rozpiętość 60,85 jednostek. Zmienna nie ma brakujących wartości.

- Urbanizacja

W 2022 roku średni współczynnik urbanizacji wynosił 39,76 i oznacza, że 39% mieszkańców powiatów mieszka w miastach. Wartość środkowa wynosiła 39,8. Powiaty posiadające współczynnik urbanizacji równy zero to wsie. Natomiast największa wartość wynosząca 78,2 to powiaty, w których 78% osób mieszka w mieście. Odchylenie standardowe wyniosło 16,6. Współczynnik zmienności wynosi 0,41 i oznacza średnią zmienność. Współczynnik skośności wynosi -0,09 i oznacza lewostronną skośność i wydłużony lewy ogon rozkładu. Kurtoza wynosiła -0,64, oznacza to rozkład platokurtyczny i mniejsze skupienie wartości wokół średniej. Percentyl 5% wynosi 11,54, oznacza to, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych 11,54. Percentyl 95% wynosi 66,6, oznacza to, że 95%

wszystkich wyników w zestawie danych jest mniejszych lub równych 66,6. Środkowe 50% danych w zestawie ma rozpiętość 24,55 jednostek. Zmienna nie ma brakujących wartości.

- Rozwody

Średnia liczba rozwodów w Polsce wyniosła 114,44. Wartością środkową jest liczba 95. Powiat posiadający najmniejszą liczbę rozwodów posiadał ich zaledwie 14. Natomiast powiat z największą liczbą rozwodów posiadał ich 534. Odchylenie standardowe wynosi 68,79. Współczynnik zmienności wynosi 0,60, co oznacza umiarkowaną zmienność. Współczynnik skośności wynosi 1,78, co wskazuje na prawostronną skośność i wydłużony prawy ogon rozkładu. Kurtoza wynosi 5,31, co oznacza rozkład leptokurtyczny. Percentyl 5% wynosi 39,7, co oznacza, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych 39,7. Percentyl 95% wynosi 253, co oznacza, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 253. Środkowe 50% danych w zestawie ma rozpiętość 82 jednostek. Zmienna nie ma brakujących wartości.

- Przedsiębiorstwa

Średnia liczba nowo zarejestrowanych w rejestrze REGON podmiotów w powiatach wynosiła 660,5. Wartością środkową była liczba 529. Powiat posiadający najmniejszą liczbę nowych przedsiębiorstw posiadał ich zaledwie 108, natomiast powiat z największą liczbą posiadał ich aż 5047. Odchylenie standardowe wynosi 524,26. Współczynnik zmienności wynosi 0,79, co oznacza silną zmienność. Współczynnik skośności wynosi 3,54, co oznacza prawostronną skośność i wydłużony prawy ogon rozkładu. Kurtoza wynosiła 19,78, co oznacza rozkład leptokurtyczny. Percentyl 5% wynosi 222,5, co oznacza, że 5% wszystkich wyników w zestawie danych jest mniejszych lub równych 222,5. Percentyl 95% wynosi 1505,2, co oznacza, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 1505,2. Środkowe 50% danych w zestawie ma rozpiętość 421,5 jednostek. Zmienna nie ma brakujących wartości.

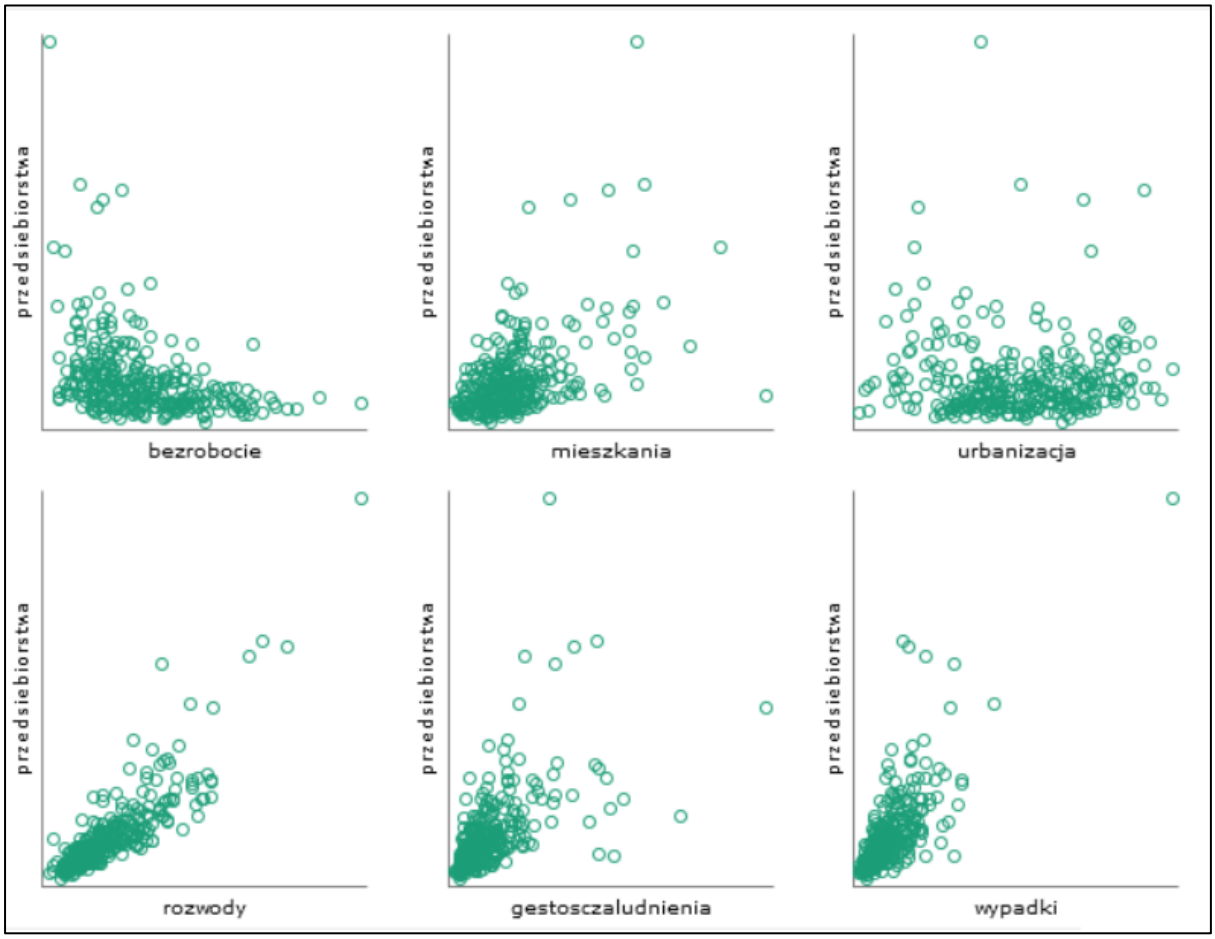
- Wypadki

Średnia liczba osób poszkodowanych przy wypadkach przy pracy wynosiła 112,81. Wartością środkową była liczba 91. Powiat posiadający najmniejszą liczbę osób poszkodowanych posiadał ich 7, natomiast największa liczba w powiecie z największą liczbą wynosiła aż 1122. Odchylenie standardowe wynosi 95,57. Współczynnik zmienności wynosi 0,84, co oznacza silną zmienność. Współczynnik skośności wynosi 4,43, co oznacza prawostronną skośność i wydłużony prawy ogon rozkładu. Kurtoza wynosiła 39,16, co oznacza rozkład leptokurtyczny. Percentyl 5% wynosi 23,7, co oznacza, że 5% wszystkich wyników w zestawie danych jest

mniejszych lub równych 23,7. Percentyl 95% wynosi 264,10, co oznacza, że 95% wszystkich wyników w zestawie danych jest mniejszych lub równych 264,10. Środkowe 50% danych w zestawie ma rozpiętość 96,000 jednostek. Zmienna nie ma brakujących wartości.

## 5. Wykresy zależności

Wykresy zależności są narzędziem wizualizacyjnym używanym do zobrazowania związku między dwoma zmiennymi. Pozwalają na identyfikację punktów danych, które znacząco odbiegają od reszty danych i mogą wpłynąć na wyniki analizy. Wykresy zależności mogą pokazać, czy zmienne wpływają na siebie, czy przeciwnie, zmienne zmieniają się w przeciwnych kierunkach, czy też nie ma między nimi związku.

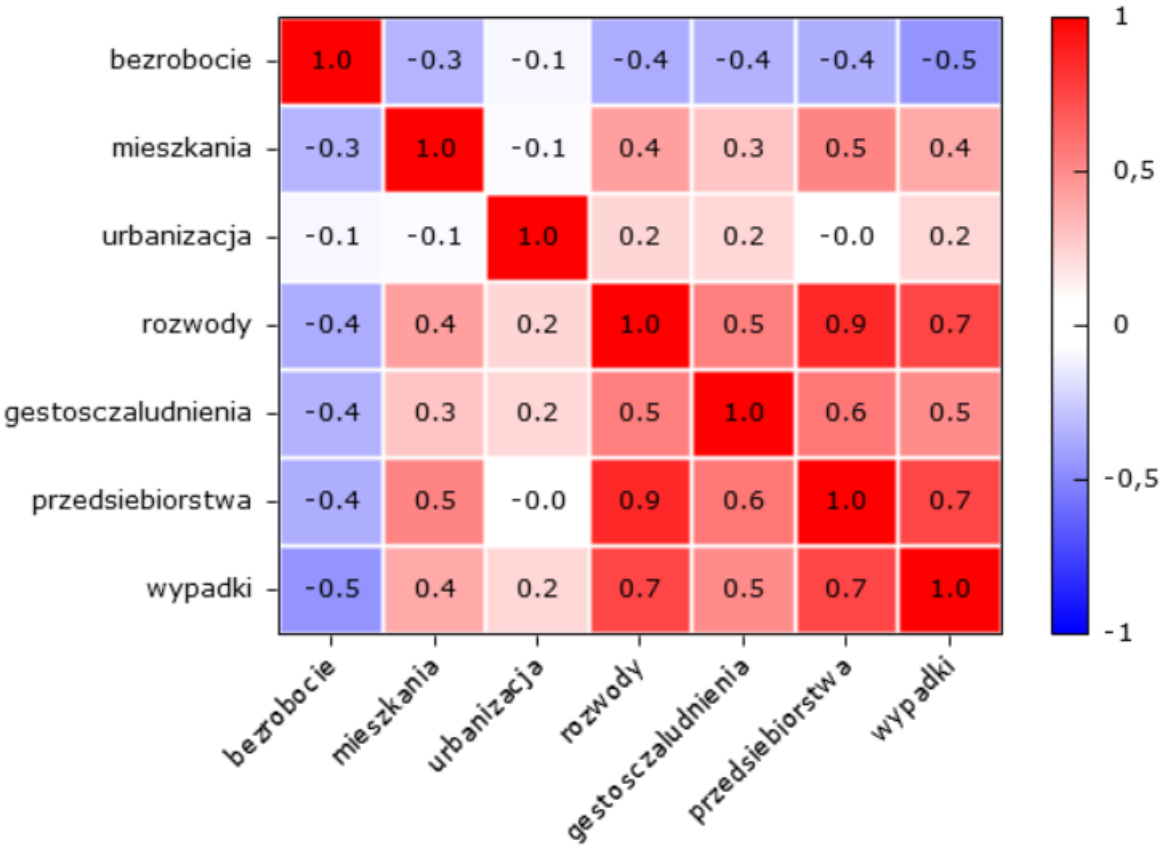



Rys. 1 Wykresy zależności zmiennych

Wykres pierwszy pokazuje zależność między liczbom nowych przedsiębiorstw w powiatach a stopą bezrobocia. Większość danych skupiona jest dla niskiej liczby nowych przedsiębiorstw oraz niskimi i średnimi wartościami dla stopy bezrobocia. Na wykresie można zauważyć

wartości odstające od normy. Drugi wykres przedstawia zależność między liczbą przedsiębiorstw a liczbą mieszkań oddanych do użytku. Większość obserwacji kumuluje się dla niskiej liczby przedsiębiorstw i niskiej liczby mieszkań. Istnieją wartości odstające, jednak można zauważyć, że dane mogą kształtować linię trendu, gdzie wraz ze wzrostem liczby przedsiębiorstw, rośnie liczba mieszkań oddanych do użytku. W przypadku trzeciego wykresu dla liczby przedsiębiorstw i współczynnika urbanizacji obserwacje są najbardziej rozrzucone. Ciężko zaobserwować trend, gdyż większość obserwacji znajduje się dla niskiej ilości przedsiębiorstw i różnej wartości stopy bezrobocia. Wykres czwarty, piąty oraz szósty są do siebie bardzo podobne. Dla każdego z tych wykresów można zobaczyć wyraźnie zaznaczony trend, oznaczający wzrost liczby nowych przedsiębiorstw wraz ze wzrostem zmiennej objaśniającej.

## 6. Macierz korelacji




Rys. 2 Macierz korelacji zmiennych

Macierz korelacji służy do analizy związku między wszystkimi parami zmiennych w zestawie danych. Korelacja przyjmuje wartości od -1 do 1. Korelacja równa 1 lub -1 oznacza doskonałą

zależność między zmiennymi. Korelacja bliska zeru oznacza brak lub bardzo słabą zależność między zmiennymi. Macierz korelacji w graficznej formie ułatwia jej interpretacje i zrozumienie. Powyższa macierz pozwala powiedzieć, że istnieje bardzo silna zależność (1-0,7) między zmiennymi: przedsiębiorstwa a rozwody, wypadki a rozwody oraz wypadki a rozwody. Umiarkowaną korelację (0,7 – 04) posiadają pary zmiennych: wypadki a bezrobocie, bezrobocie a przedsiębiorstwa, bezrobocie a gęstość zaludnienia, bezrobocie a rozwody, mieszkania a wypadki, mieszkania a przedsiębiorstwa, gęstość zaludnienia a rozwody, przedsiębiorstwa a gęstość zaludnienia, wypadki a gęstość zaludnienia. Oznacza to, że zmienne w jakiś sposób na siebie wpływają. Pozostałe pary zmiennych oznaczają słabą korelację, czyli małą zależność między zmiennymi. Korelacja między zmienną objaśnianą a zmiennymi objaśniającymi, mają bardzo różne wartości.

# 2. Część II

## 7. Model ściśle liniowy

Model 25: Estymacja KMNK, wykorzystane obserwacje 1-313
Zmienna zależna (Y): przedsiebiorstwa

| | współczynnik | błąd standardowy | t-Studenta | wartość p | |
|---|---|---|---|---|---|
| const | 6,17518 | 56,9669 | 0,1084 | 0,9137 | |
| bezrobocie | 5,69660 | 3,56099 | 1,600 | 0,1107 | |
| mieszkania | 21,5314 | 4,77927 | 4,505 | 9,45e-06 | *** |
| urbanizacja | -6,83915 | 0,786699 | -8,693 | 2,18e-016 | *** |
| rozwody | 4,59787 | 0,287145 | 16,01 | 2,44e-042 | *** |
| gestosczaludnien~ | 0,993804 | 0,190322 | 5,222 | 3,28e-07 | *** |
| wypadki | 1,33374 | 0,204416 | 6,525 | 2,82e-010 | *** |

| | | | |
|---|---|---|---|
| Średn.aryt.zm.zależnej | 660,4952 | Odch.stand.zm.zależnej | 524,2582 |
| Suma kwadratów reszt | 14440283 | Błąd standardowy reszt | 217,2337 |
| Wsp. determ. R-kwadrat | 0,831604 | Skorygowany R-kwadrat | 0,828303 |
| F(6, 306) | 251,8584 | Wartość p dla testu F | 3,5e-115 |
| Logarytm wiarygodności | -2124,833 | Kryt. inform. Akaike'a | 4263,666 |
| Kryt. bayes. Schwarza | 4289,889 | Kryt. Hannana-Quinna | 4274,145 |

Rys. 3 Model KMNK

Szacowanie parametrów regresji metodą najmniejszych kwadratów, opisuje zależność między liczbom nowych przedsiębiorstw a wcześniej przedstawionymi zmiennymi objaśniającymi. Powyższy model posiada kilka wad, które mogą wpływać na jego wiarygodność i interpretację wyników. Pierwszą wadą jest mniejsza istotność statystyczna zmiennej bezrobocie oraz stałej,

wskazuje na to najniższa wartość p pośród zmiennych. Sugeruje to, że wymienione zmienne nie mają istotnego wpływu na liczbę przedsiębiorstw w analizowanym modelu. W tym momencie model nie ma innych większych wad.

## 8. Redukcja ilości zmiennych

W celu poprawienia jakości modelu nastąpi redukcja ilości zmiennych za pomocą metody Hellwiga oraz metody krokowej wstecznej, w celu znalezienia najlepszego modelu. Jako pierwszą posłużę się metodą Hellwiga, jej celem jest znalezienie takiego zestawu zmiennych, który maksymalizuje wartość pojemności integralnej, będącego miarą jakości modelu. Za pomocą kodu zamieszczonego w skrypcie do projektu, obliczona zostaje pojemność integralna i wskazane zostają najlepsze zmienne objaśniające według metody Hellwiga.

```
? H_max
0,74831776
? najlepszalista
mieszkania rozwody gestosczaludnienia wypadki
```

Rys. 4 Metoda Hellwiga

Wyniki wskazują, że najlepszym zbiorem zmiennych objaśniających są zmienne: mieszkania, rozwody, gęstość zaludnienia i wypadki. W celu sprawdzenia stworzony zostaje model z tymi zmiennymi.

```
Model 27: Estymacja KMNK, wykorzystane obserwacje 1-313
Zmienna zależna (Y): przedsiebiorstwa

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                -181,875         30,9237          -5,881    1,06e-08  ***
  mieszkania             29,3479         5,10776          5,746    2,20e-08  ***
  rozwody                 4,37142        0,319096        13,70     1,10e-033 ***
  gestosczaludnien~       0,745459       0,208902         3,568    0,0004    ***
  wypadki                 1,09063        0,219637         4,966    1,14e-06  ***

Średn.aryt.zm.zależnej  660,4952   Odch.stand.zm.zależnej  524,2582
Suma kwadratów reszt    18157226   Błąd standardowy reszt  242,8004
Wsp. determ. R-kwadrat  0,788259   Skorygowany R-kwadrat   0,785509
F(4, 308)               286,6523   Wartość p dla testu F   1,8e-102
Logarytm wiarygodności -2160,679   Kryt. inform. Akaike'a  4331,357
Kryt. bayes. Schwarza   4350,088   Kryt. Hannana-Quinna    4338,843
```

Rys. 5 Model po zastosowaniu metody Hellwiga

Pierwsze co można zauważyć, to że wszystkie zmienne w modelu są istotne, a model nie posiada problemu współliniowości. Negatywną informacją jest spadek współczynnika determinacji oraz wzrost kryterium informacyjnego.

W celu próby poprawienia jakości modelu, wykorzystana zostanie druga metoda, metoda krokowa – wsteczna. Metoda ta polega na skonstruowaniu modelu zawierającego wszystkie potencjalne zmienne objaśniające, a następnie na stopniowym eliminowaniu zmiennych tak, aby utrzymać model z najwyższą wartością współczynnika determinacji przy zachowaniu istotności parametrów. W celu upewnienia się, którą zmienną usunąć, sprawdzony zostanie współczynnik VIF, odpowiedzialny za ocenę współliniowości. Jako pierwszą usunięta zostanie zmienna bezrobocie, gdyż posiada ona największą wartość p.

Model 31: Estymacja KMNK, wykorzystane obserwacje 1-313
Zmienna zależna (Y): przedsiebiorstwa

| | współczynnik | błąd standardowy | t-Studenta | wartość p | |
|---|---|---|---|---|---|
| const | 70,9747 | 40,1570 | 1,767 | 0,0781 | * |
| mieszkania | 20,0749 | 4,70364 | 4,268 | 2,63e-05 | *** |
| urbanizacja | -6,86599 | 0,788515 | -8,707 | 1,95e-016 | *** |
| rozwody | 4,61749 | 0,287611 | 16,05 | 1,56e-042 | *** |
| gestosczaludnien~ | 0,950327 | 0,188850 | 5,032 | 8,27e-07 | *** |
| wypadki | 1,24845 | 0,197840 | 6,310 | 9,71e-010 | *** |

| | | | |
|---|---|---|---|
| Średn.aryt.zm.zależnej | 660,4952 | Odch.stand.zm.zależnej | 524,2582 |
| Suma kwadratów reszt | 14561048 | Błąd standardowy reszt | 217,7846 |
| Wsp. determ. R-kwadrat | 0,830196 | Skorygowany R-kwadrat | 0,827431 |
| F(5, 307) | 300,1937 | Wartość p dla testu F | 6,9e-116 |
| Logarytm wiarygodności | -2126,136 | Kryt. inform. Akaike'a | 4264,272 |
| Kryt. bayes. Schwarza | 4286,750 | Kryt. Hannana-Quinna | 4273,255 |

Rys. 6 Metoda krokowa wsteczna

Powstały model poprawił istotność stałej. Współczynnik determinacji nie zmienił się, tak samo, jak kryterium informacyjne. W tym momencie każda zmienna jest istotna statystycznie, a model nie ma problemu współliniowości.

Modele wybrane metodą Hellwiga oraz metodą krokową wsteczną różnią się od siebie. Model wybrany metoda Hellwiga posiada mniejszy współczynnik determinacji oraz wyższe kryterium informacyjne, co sprawia, że model wybrany metodą krokową wsteczną może być uznawany za lepszy. Niestety wadą obu modeli jest heteroskedastyczność i brak normalności rozkładu reszt.

# 3. Część III

## 9. Wybór ostatecznej postaci modelu

W celu znalezienia lepszego modelu nastąpi transformacja zmiennych za pomocą logarytmów. Dla każdej zmiennej zostanie obliczony logarytm i stworzony nowy model. W przypadku logarytmów ponownie zostały zastosowane obie metody, które pomogą wybrać model. Metoda Helliwiga wybrała logarytmy zmiennych: mieszkania, rozwody, gęstość zaludnienia i wypadki. Wszystkie zmienne w modelu są istotne, model ma normalność rozkładu reszt, jednak nadal jest heteoskedastyczny. W przypadku model wybranego metodą wsteczną krokową posiada on jeszcze dodatkową zmienną urbanizacja i, mimo że ma normalny rozkład reszt, posiada on heteroskadastyczność. W celu znalezienia modelu, którego wszystkie zmienne będą miały normalny rozkład reszt, homoskedastyczność oraz istotność zmiennych. Wykorzystany zostanie kod z metody Helwiga, który pozwoli znaleźć zestaw zmiennych, dla którego wartość p będzie na tyle duża, że pozwoli na znalezienie modelu, który będzie miał normalny rozkład reszt oraz będzie homoskedastyczny. Wykorzystanie przekształceń na modelu na logarytmy ma wiele zalet, model może być dostosowany do różnych sytuacji, poprawiając przy tym postać modelu.[1]

```
Model 16: Estymacja KMNK, wykorzystane obserwacje 2-313 (n = 312)
Zmienna zależna (Y): l_przedsiebiorstwa

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const           4,12909       0,160043         25,80     5,80e-079 ***
  l_mieszkania    0,254842      0,0412685         6,175    2,08e-09  ***
  l_urbanizacja  -0,198040      0,0376566        -5,259    2,71e-07  ***
  l_wypadki       0,560789      0,0309553        18,12     1,92e-050 ***

Średn.aryt.zm.zależnej  6,295086   Odch.stand.zm.zależnej  0,602678
Suma kwadratów reszt    40,29725   Błąd standardowy reszt  0,361712
Wsp. determ. R-kwadrat  0,643266   Skorygowany R-kwadrat   0,639792
F(3, 308)               185,1298   Wartość p dla testu F   1,30e-68
Logarytm wiarygodności -123,4205   Kryt. inform. Akaike'a  254,8410
Kryt. bayes. Schwarza   269,8130   Kryt. Hannana-Quinna    260,8249
```

Rys. 7 Ostateczna postać modelu.

[1] Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach,* Michigan 2012, s. 334-335

Powstały model ma trzy zmienne objaśniające mieszkania, urbanizacja, wypadki (wszystkie w postaci z logarytmu), a każda ze zmiennych jest istotna. Jedynymi problemami, które powstały w przypadku postaci tego modelu jest współczynnik determinacji, który jednak nadal ma zadowalającą wartość, gdyż 64% zmiennych objaśnianych wyjaśnia zmienną objaśniającą. Drugim problemem jest wzrost kryterium informacyjnego. Pomimo tych kwestii, model w końcu ma normalny rozkład reszt i jest homoskedastyczny. Taka postać modelu jest zadowalająca, a aspekt homoskedastyczności pomoże w dalszej analizie powyższego modelu.

# 4. Część IV - Pełen opis i testowanie własności modelu

## 10. Współczynnik determinacji

Współczynnik determinacji ocenia, jak dobrze funkcja regresji dopasowuje się do danych empirycznych. Wskazuje on, jaka część zmienności zmiennej objaśnianej Y została wyjaśniona przez zmienne objaśniające X. Współczynnik determinacji przyjmuje wartości w przedziale od 0 do 1, gdzie wartość bliższa 1 oznacza lepsze dopasowanie funkcji regresji do danych. W przypadku analizowanego modelu wartość współczynnika determinacji jest równa 0,64. Oznacza to, że 64% zmienności zmiennej objaśnianej jest wyjaśnione przez zmienne objaśniające (l_mieszkania, l_urbanizacja, l_wypadki) oraz stałą w modelu regresji. W kontekście modelu oznacza to, że około 64% zmienności w liczbie nowo zarejestrowanych przedsiębiorstw w powiatach jest wyjaśnione przez zmienne: liczba mieszkań oddanych do użytkowania, stopień urbanizacji w powiatach oraz liczba osób poszkodowanych w czasie pracy. Do weryfikacji istotności współczynnika determinacji posłuży uogólniony test Walda, gdzie do obliczeń wykorzystywany jest współczynnik determinacji.

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - (k + 1)}{k}$$

gdzie:

$R^2$– współczynnik determinacji

$k$ – liczba zmiennych objaśniających

$n$ – liczba obserwacji

$H_0$: $R^2$ jest nieistotny.

$H_1$: $R^2$ jest istotny.

```
F(3, 308): prawostronny obszar krytyczny dla 185,731 = 9,41726e-069
(lewostronny obszar krytyczny: 1)
```

Rys. 8 Test istotności współczynnika determinacji.

Wynik testu daje podstawy do odrzucenia hipotezy $H_0$ i wniosek o istotności współczynnika determinacji.

## 11.Efekt katalizy

Efekt katalizy oznacza sytuację, w której współczynnik determinacji może osiągnąć wysoką wartość, mimo że rzeczywisty charakter oraz siła związku między zmiennymi objaśniającymi a zmienną objaśnianą nie uzasadnia takiego wyniku. Innymi słowy, informacja przekazywana przez współczynnik determinacji może być mylna lub nieadekwatna do faktycznych zależności występujących w danych. Efekt katalizy można zmierzyć natężeniem zjawiska katalizy.

$$\eta = R^2 - H$$

gdzie:

$R^2$ - współczynnik determinacji

$H$ - integralna pojemność informacyjna zestawu zmiennych objaśniających.

Współczynnik $\eta$ należy do przedziału [0,1]. Jeśli jest znacząco różny od zera, wskazuje to na występowanie efektu katalizy.

$H_0$: $R^2$ jest wynikiem przypadkowym i nie odzwierciedla rzeczywistego związku między zmiennymi objaśniającymi a zmienną objaśnianą, czyli $\eta = 0$.

$H_1$: $R^2$ jest wynikiem istniejącego zjawiska katalizy, co oznacza, że rzeczywisty związek między zmiennymi objaśniającymi a zmienną objaśnianą jest słabszy niż sugeruje $R^2$, czyli $\eta > 0$.

```
Wygenerowano skalar ni = 0,132706
```

Rys. 9 Efekt katalizy

Współczynnik $\eta$ wynosi 0,13 i oznacza, że istnieją podstawy do odrzucenia hipotezy zerowej. Jest to istotna informacja dotycząca występowania efektu katalizy w analizowanym modelu regresji. Oznacza to, że część wysokiej wartości współczynnika determinacji $R^2$ może być

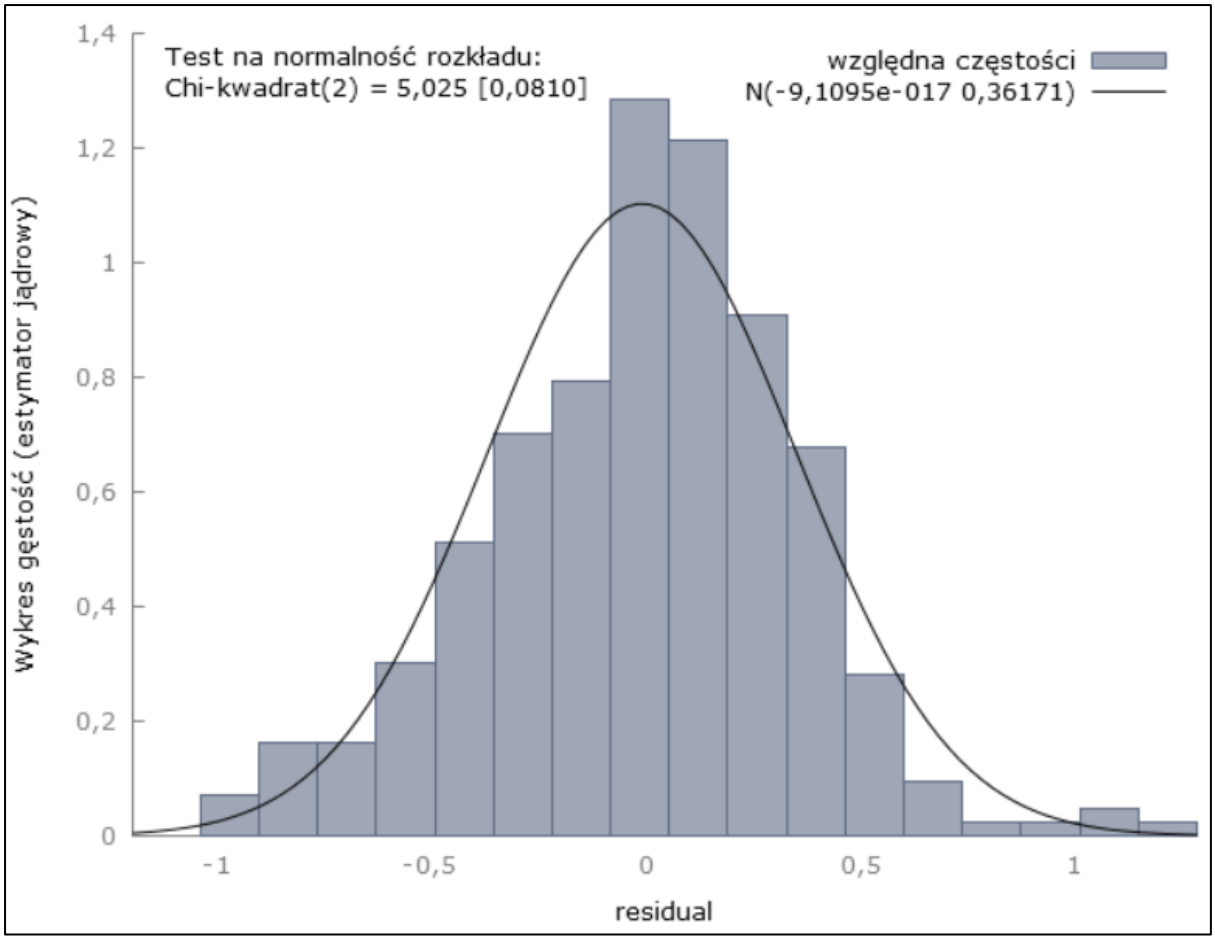
wynikiem katalizy. Informacja dostarczana przez współczynnik determinacji może być nieprawdziwa lub nieodpowiednia w odniesieniu do rzeczywistych relacji obecnych w danych.

## 12.Normalność rozkładu składnika losowego

Normalność składnika losowego w modelu oznacza, że reszty mają rozkład, który jest symetryczny i zbliżony do normalnego. To założenie jest istotne, ponieważ umożliwia poprawne użycie wielu klasycznych metod statystycznych, takich jak testy istotności, przedziały ufności oraz prognozowanie. Brak normalności składnika losowego może prowadzić do błędnych wniosków statystycznych i wymaga zastosowania odpowiednich metod diagnostycznych lub alternatywnych podejść statystycznych. Przy użyciu Gretla, można łatwo zbadać normalność rozkładu składnika losowego za pomocą testu Doornika-Hansena, test weryfikuje hipotezę zerową o istnieniu normalnego rozkładu reszt, wobec hipotezy alternatywnej, że składnik modelu nie ma rozkładu normalnego. Wyniki podawane są w postaci histogramu ze statystyką testową oraz wartością p.

$H_0$: Składnik losowy ma rozkład normalny.

$H_1$: Składnik losowy nie ma rozkładu normalnego.




Rys. 10 Test na normalność rozkładu reszt

Wartość p przeprowadzonego testu jest większa niż założony poziom istotności ($\alpha = 0{,}05$). Wynik sugeruje, że nie ma wystarczających podstaw do odrzucenia hipotezy zerowej o normalności składnika losowego. Dlatego można stwierdzić, że składnik losowy w analizowanym modelu ma rozkład losowy.

## 13.Istotność zmiennych

Istotność zmiennych objaśniających w modelu ekonometrycznym informuje, czy efekt danej zmiennej na zmienną objaśnianą jest wystarczająco duży, aby uznać go za istotny statystycznie. W celu sprawdzenia istotności można przeprowadzić testy każdej ze zmiennej z osobna lub test istotności wszystkich zmiennych. Poniższy zapis hipotez nie jest dla testu, łącznej istotności parametrów, jednak w celu łatwiejszego zapisania testu istotności każdej z hipotez.

$H_0$: $\beta_1 = 0$ , $\beta_2 = 0$, $\beta_3 = 0$

$H_1$: $\beta_1 \neq 0$ , $\beta_2 \neq 0$, $\beta_3 \neq 0$

```
T oblicozne dla parametru przy zmiennej l_mieszkania:  6,17523
T oblicozne dla parametru przy zmiennej l_urbanizacja:  -5,2591
T oblicozne dla parametru przy zmiennej l_wypadki:  18,1161
Wartość krytyczna:  1,9677
```

Rys. 11 Test istotności parametrów

Wartość statystyki na moduł jest w każdym przypadku większa niż wartość krytyczna. Wynik taki daje podstawy do odrzucenia hipotezy zerowej o nieistotności parametrów modelu. Oznacza to, że każdy parametr w modelu jest istotny.

W przypadku testu na łączną istotność to został on już przeprowadzony przy badaniu istotności współczynnika determinacji, gdyż jest to ten sam uogólniony test Walda. W tym przypadku również nastąpi odrzucenie hipotezy zerowej i wniosek o łącznej istotności wszystkich parametrów modelu.

## 14. Test dodanych i pominiętych zmiennych

Test pominiętych zmiennych jest używany w analizie regresji do oceny istotności pominięcia pewnych zmiennych objaśniających w modelu. Hipoteza zerowa tego testu zakłada, że parametry regresji dla wskazanych zmiennych są równe zero, co oznacza, że te zmienne nie są istotne statystycznie w wyjaśnianiu zmienności zmiennej zależnej.

W celu stworzenia właściwego modelu nastąpiło wiele przekształceń. Jako pierwotny model został przyjęty model wybrany metodą krokową wsteczną, gdzie później wszystkie zmienne poza stała zostały pominięte na rzecz znalezienia modelu, który nie będzie posiadał heteroskedastyczności. Dlatego przeprowadzony zostanie test z modelu wybranego metodą krokową wsteczną, gdzie pominięte zostaną wszystkie zmienne, poza stałą.

```
Hipoteza zerowa: parametry regresji dla wskazanych zmiennych są równe zero
  mieszkania, urbanizacja, rozwody, gestosczaludnienia, wypadki
Statystyka testu: F(5, 307) = 300,194, wartość p 6,87691e-116
Pominięcie zmiennych poprawiło 0 z 3 kryteriów informacyjnych (AIC, BIC, HQC).
```

Rys. 12 Test pominiętych zmiennych

Z przeprowadzonego testu wynika niska wartość p (< 0,05) i oznacza, że możemy odrzucić hipotezę zerową. Zmienne te mają istotny wpływ na zmienną zależną. Test informuje również o braku poprawy kryteriów informacyjnych. Jednak jak zostało już wcześniej wspomniane, zmienne te musiały zostać usunięte, ze względu na heteroskedastyczność.

Test dodanych zmiennych jest używany w analizie regresji do oceny istotności dodania nowych zmiennych objaśniających do istniejącego modelu regresyjnego. Zakłada się, że te dodatkowe zmienne mogą poprawić dopasowanie modelu i zwiększyć jego zdolność do wyjaśnienia zmienności zmiennej zależnej. Do testu dodanych zmiennych zastosuje logarytmy ze zmiennych: mieszkania, wypadki i urbanizacja. Niestety przy próbie dodania zmiennej l_urbanizacja widnieje informacja o brakujących danych, dlatego przeprowadzony zostanie test dla tylko dwóch wymienionych wcześniej zmiennych.

```
Hipoteza zerowa: parametry regresji dla wskazanych zmiennych są równe zero
  l_mieszkania, l_wypadki
Statystyka testu: F(2, 305) = 6,27884, wartość p 0,00212703
Dodanie zmiennych poprawiło 3 z 3 kryteriów informacyjnych (AIC, BIC, HQC).
```

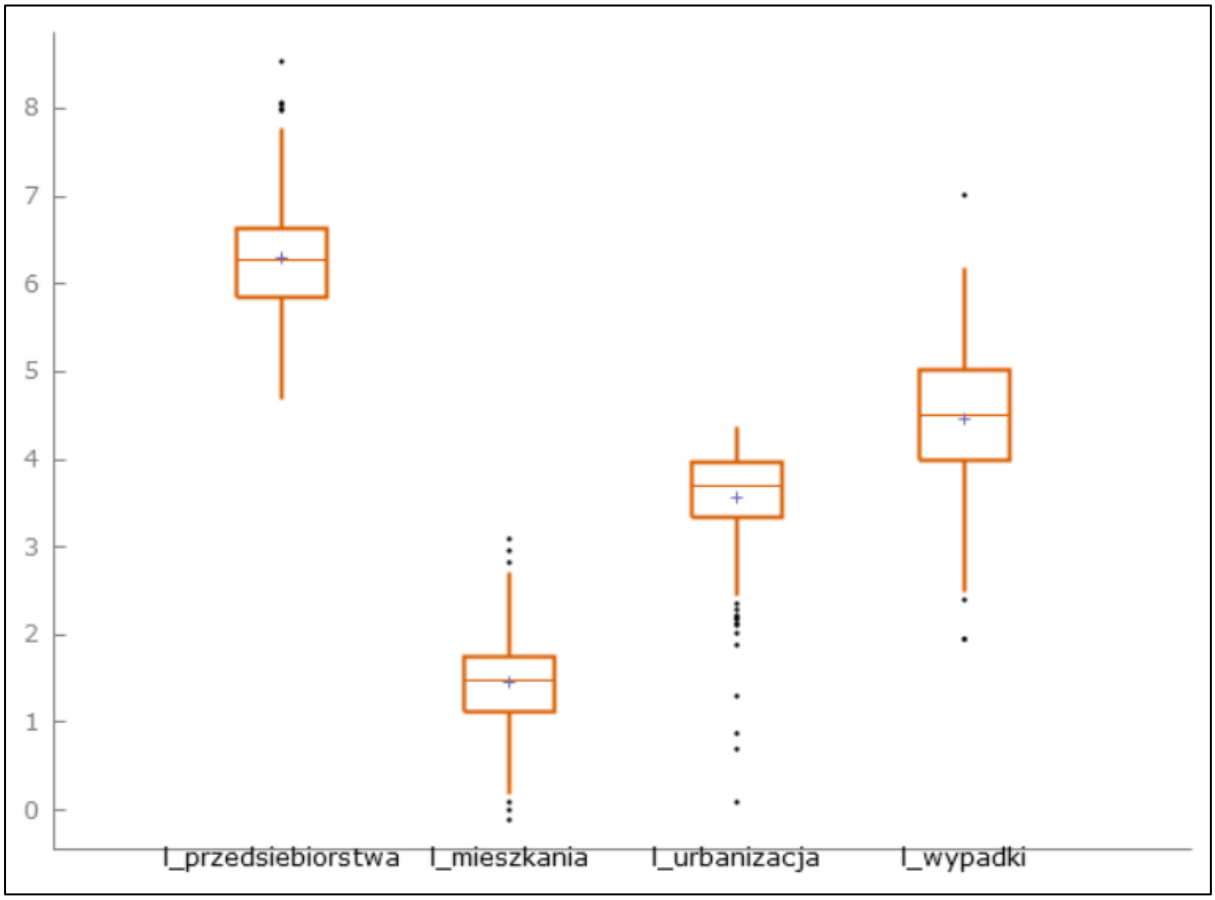
Rys. 13 Test dodanych zmiennych

Wartość p jest mniejsza niż założony poziom istotności, wskazuje to na odrzucenie hipotezy zerowej i fakt, że dodanie zmiennych ma istotny wpływ na zmienną zależną i poprawia jakość modelu. Można również zauważyć, że poprzez dodanie zmiennych poprawiły się wszystkie kryteria informacyjne.

## 15.Odstające obserwacje

Wartości odstające to dane, które znacznie odbiegają od pozostałych obserwacji w zbiorze danych. Odstające obserwacje mogą mieć znaczny wpływ na wynik analizy regresji czy prognozowanie. Ważne jest ich odpowiednie zidentyfikowanie i podjęcie decyzji co do dalszego postępowania z nimi. W celu sprawdzenia, czy w zbiorze zmiennych występują wartości odstające, wykorzystane zostaną wykresy pudełkowe. Wykresy pudełkowe przedstawiają rozkład zmiennych oraz występowanie elementów takich jak wartość środkowa, rozproszenie danych oraz obecność wartości odstających. Centralna część wykresu nazywana „pudełkiem” rozciąga się od pierwszego do trzeciego kwartylu, wewnątrz pudełka znajduje się linia, które przedstawia medianę zestawu danych. Linie rozciągające się od pudełka do najmniejszej i największej wartości w obrębie 1.5-krotności IQR od Q1 i Q3. Z uwagi na posługiwanie się logarytmami ze zmiennych, każda z nich została przedstawiona na jednym wykresie, który przedstawia wykres pudełkowy dla każdej ze zmiennych. Przyglądając się wartościom odstającym, można zauważyć, że posiada je każda z analizowanych zmiennych. W przypadku logarytmu ze zmiennej przedsiębiorstwa zauważalne są wartości odstające w

górnej części wykresu. Oznacza to, że w zbiorze danych istnieją powiaty, które miały znacznie większą liczbę nowo zarejestrowanych przedsiębiorstw niż inne analizowane powiaty.




Rys. 14 Wykresy pudełkowe

Logarytm ze zmiennej mieszkania również ma wartości odstające, jednak w tym przypadku są one w górnej, jak i dolnej części wykresu. Jest to informacja, że w analizowanym zbiorze danych istnieją powiaty, w których liczba mieszkań oddana do użytkowania jest bardzo mała i odbiega od normy oraz istnieją powiaty, które zarejestrowały o wiele wyższą liczbę mieszkań oddanych do użytkowania. Trzecią zmienną jest logarytm ze zmiennej dotyczącej wskaźnika urbanizacji. W tym przypadku liczba wartości odstających jest największa, są one tylko w dolnej części wykresu, a ich rozproszenie jest największe. Oznacza to, że wiele powiatów charakteryzuje się znacznie niższym współczynnikiem urbanizacji niż reszta. Może to wynikać z faktu, że większość tych powiatów może być głównie obszarami wiejskimi lub rolniczymi, gdzie współczynnik urbanizacji jest niższy. Ostatnią z analizowanych zmiennych jest logarytm dla zmiennej wypadki. Charakteryzuje się ona jedną wysoką wartością odstającą oraz dwoma niskimi wartościami odstającymi. Są to powiaty, w których liczba osób poszkodowanych w wypadkach przy pracy jest znacznie wyższa oraz znacznie niższa niż w pozostałych powiatach.

## 16.Test liczby serii

Nieparametryczny test serii stosuje się do sprawdzenia, czy wyniki spełniają założenie losowości próby. Polega na rozróżnieniu ciągu reszt uzyskanych po oszacowaniu modelu na dwie grupy: reszty dodatnie i reszty ujemne. Wyznaczonym elementom przypisuje się odpowiednie symbole w zależności od znaku reszty. W danym teście weryfikacji podlegają hipotezy:

$H_0$: Postać modelu jest dobrze dobrana.

$H_1$: Postać modelu jest źle dobrana.

Do wykonania testu serii w programie Gretl należy najpierw zapisać reszty estymowanego modelu, a następnie przeprowadzić test nieparametryczny – test serii.

```
Test serii

Liczba serii (R) dla zmiennej 'e' = 146
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(157, 8,8176),
test z-score = -1,24751, przy dwustronym obszarze krytycznym p = 0,212212
```

Rys. 15 Test liczby serii – test poprawności postaci modeli

Wartość p przeprowadzonego testu jest większa niż przyjęte 5%, zatem nie ma podstaw do odrzucenia $H_0$. Występuje liniowa zależność zmiennej objaśnianej oraz zmiennych objaśniających, a postać modelu jest dobrze dobrana.

## 17.Test RESET

Test RESET Ramseya dla reszt modelu, inaczej nazywany testem stabilności postaci analitycznej modelu. Często używany jest w celu wykrywania ogólnych problemów z niedopasowaniem modelu. Do zastosowania testu RESET, należy zdecydować, ile funkcji wartości dopasowanych uwzględnić w rozszerzonej regresji. Nie ma jednoznacznej odpowiedzi na to pytanie, ale okazało się, że kwadraty i sześciany są użyteczne w większości zastosowań.[2] Test ten weryfikuje, czy dobrana liniowa postać modelu jest najlepszą z możliwych.

[2] Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach,* Michigan 2012, s. 334-335

$H_0$: Wybór postaci analitycznej modelu jest prawidłowy.

$H_1$: Wybór postaci analitycznej modelu jest nie prawidłowy.

```
Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 2-313 (n = 312)
Zmienna zależna (Y): l_przedsiebiorstwa

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          -11,8761        14,6565         -0,8103     0,4184
  l_mieszkania    -2,05875        1,85173        -1,112      0,2671
  l_urbanizacja    1,59819        1,43755         1,112      0,2671
  l_wypadki       -4,48395        4,06428        -1,103      0,2708
  yhat^2           1,24263        1,13852         1,091      0,2759
  yhat^3          -0,0554022      0,0593592      -0,9333     0,3514

Statystyka testu: F = 5,197424,
z wartością p = P(F(2,306) > 5,19742) = 0,00603
```

Rys. 16 Test RESET

Wartość p – value jest mniejsza od 5%. Wynik ten wskazuje na odrzucenie hipotezy zerowej i stwierdzenie, że wybór postaci analitycznej modelu jest nieprawidłowy. Istnieje kilka kroków, które można teraz podjąć w przypadku odrzucenia hipotezy zerowej dla testu RESET. Pierwszym krokiem jest pozostawienie takiego wyniku i dalsza analiza. Jak można zauważyć, test liczby serii wykazał, że postać modelu jest poprawna, dla upewnienia najlepszym rozwiązaniem będzie, przeprowadzenie testu Chowa, który sprawdza stabilność parametrów modelu i następnie będzie można decydować o innych krokach.

## 18. Test Chowa

Test Chowa jest używany do sprawdzania, czy występują istotne różnice między dwoma różnymi zestawami danych lub pomiędzy dwoma różnymi podgrupami danych w kontekście analizy regresji. Test ten bada, czy parametry regresji zmieniają się istotnie w wyniku podziału danych. Dlatego dla danych zawierających 312 obserwacji, do przeprowadzenia testu, dane zostały podzielone w 156 obserwacji.

$H_0$: Parametry modelu są stabilne.

$H_1$: Parametry modelu nie są stabilne.

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 156
  F(4, 304) = 3,91331 z wartością p 0,0041
```

Rys. 17 Test Chowa

Niestety wartość p  jest mniejsza niż założony poziom istotności. Wynik ten wskazuje na odrzucenie hipotezy zerowej i stwierdzenie, że parametry modelu nie są stabilne.

Według literatury najlepszym sposobem na stwierdzenie poprawności analitycznej w teście RESET jest zmiana zmiennych na logarytmy, niestety badane zmienne są już logarytmami.[3] Innym sposobem jest oszacowanie nieliniowego modelu danych lub powrót do samego początku i przeprowadzenie regresji zawierającej inne zmienne. W tym przypadku zmiana postaci modelu na tym etapie projektu byłaby uciążliwa, gdyż trzeba byłoby zacząć od nowa. Dlatego biorąc pod uwagę wnioski z testu RESET i Chowa, będzie postępować dalsza analiza tego samego modelu.

## 19. Testowanie heteroskedastyczności

Heteroskedastyczność to sytuacja, gdy wariancje zmiennych losowych nie są stałe w całym zakresie wartości niezależnych zmiennych objaśniających. Innymi słowy, występuje heteroskedastyczność, gdy wariancja błędów modelu zmienia się w zależności od poziomu wartości jednej lub więcej zmiennych objaśniających. W ogólności heteroskedastyczność oznacza, że jedno z założeń metody najmniejszych kwadratów zostało naruszone. Estymator będzie nadal liniowy i nieobciążony, jednak nie będzie najbardziej efektywny. Co więcej, błędy standardowe będą nieprawidłowe, podobnie jak przedziały ufności i testy hipotez, które bazują na tych błędach standardowych. Zjawiskiem pożądanym i przeciwnym do heteroskedastyczności jest homoskedastyczność, czyli sytuacja gdzie wariancja składnika losowego jest stała. W Gretlu istnieją wbudowane testy, które pozwalają sprawdzić heteroskedastyczność reszt, jest to test White'a oraz test Breuscha-Pagana.

$H_0$: Heteroskedastyczność reszt nie występuje.

$H_1$: Heteroskedastyczność reszt występuje.

[3] https://www.cambridge.org/features/economics/brooks/downloads/Answers%20to%20end%20of%20chapter%20questions/Chapter4_solutions.doc, data dostępu: 20.06.2024

```
Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej) -
  Hipoteza zerowa: heteroskedastyczność reszt nie występuje
  Statystyka testu: LM = 11,9293
  z wartością p = P(Chi-kwadrat(9) > 11,9293) = 0,217324

Test Breuscha-Pagana na heteroskedastyczność -
  Hipoteza zerowa: heteroskedastyczność reszt nie występuje
  Statystyka testu: LM = 10,6653
  z wartością p = P(Chi-kwadrat(3) > 10,6653) = 0,0136806
```

Rys. 18 Testy na heteroskedastyczność

W przypadku testu White'a wartość p przeprowadzonego testu jest większa niż przyjęte 5%, zatem nie ma podstaw do odrzucenia $H_0$ i daje to podstawy, że model jest homoskedastyczny. Natomiast Test Breuscha – Pagana wskazuje, że wartość p jest mniejsza niż poziom istotności wynoszący 0,05, natomiast jest większa niż poziom istotności 0,01. Możliwe jest stwierdzenie, że na poziomie istotności 0,05, ale nie na poziomie istotności 0,01, istnieją różnice w wariancji resztowej. Oznacza to, że heteroskedastyczność może występować, ale nie jest ona na tyle silna, aby ją wykryć na bardziej rygorystycznym poziomie istotności 0,01. Biorąc pod uwagę, że homoskedastyczność wyszła w teście White'a i w teście Bresucha – Pagana na poziomie istotności równym 0,01, stwierdzone zostaje, że model nie posiada problemu heteroskedastyczności.

## 20. Współliniowość

Współliniowość w kontekście analizy regresji odnosi się do sytuacji, gdy jedna z niezależnych zmiennych objaśniających w modelu regresyjnym jest silnie skorelowana z inną lub innymi zmiennymi objaśniającymi. Współliniowość utrudnia interpretację modelu oraz może prowadzić do problemów z dokładnością oszacowań i wniosków statystycznych.[4] W Gretlu można bardzo łatwo zbadać problem współliniowości, za pomocą komendy *vif*.

```
Ocena współliniowości VIF(j) - czynnik rozdęcia wariancji
VIF (Variance Inflation Factors) - minimalna możliwa wartość = 1.0
Wartości > 10.0 mogą wskazywać na problem współliniowości - rozdęcia wariancji

 l_mieszkania     1,226
l_urbanizacja     1,146
    l_wypadki     1,301
```

Rys. 19 Badanie współliniowości

[4] G.S Maddala, *Ekonometria,* Warszawa 2008, s 185.

Wartości czynnika rozdęcia wariancji są mniejsze niż 10 i oznaczają, że podane zmienne nie posiadają problemu współliniowości, czyli zmienne objaśniające są dość niezależne od siebie. Współliniowość między nimi nie jest na tyle silna, aby miała istotny wpływ na precyzję oszacowań współczynników regresji.

## 21.Koincydencja

Model ekonometryczny posiada własność koincydencji, jeśli dla każdej zmiennej objaśniającej znak współczynnika stojącego przy zmiennej w modelu jest równy znakowi współczynnika korelacji ze zmienną objaśnianą.[5] Zestawienie współczynników dla szacowanego modelu zamieszczono poniżej.

| | Współczynnik | Korelacja |
|---|---|---|
| L_mieszkania | 0,254842 | 0,51 |
| L_urbanizacja | −0,198040 | -0,03 |
| L_wypadki | 0,560789 | 0,74 |

Tab. 2 Koincydencja

Widać zatem, że w modelu zachodzi zjawisko koincydencji, ponieważ spełniony jest warunek, że dla każdej zmiennej znak współczynnika jest równy znakowi współczynnika korelacji.

## 22.Interpretacja parametrów modelu

Estymowany model ma następującą postać:

$$y = 4{,}129 + 0{,}255x_1 - 0{,}198x_2 + 0{,}561x_3$$

gdzie:

$y =$ logarytm z liczby nowych zarejestrowanych przedsiębiorstw.

$x_1 =$ logarytm z liczby mieszkań oddanych do użytku.

$x_2 =$ logarytm ze współczynnika urbanizacji.

$x_3 =$ logarytm z liczby osób poszkodowanych w wypadkach podczas pracy.

[5] Eligiusz W. Nowakowski, *Podstawy ekonometrii z elementami algebry liniowej*, Warszawa 2011, s. 61-62.

Interpretacja każdego z parametrów (zakładając, że inne zmienne pozostają stałe):

- Wzrost zmiennej mieszkania o jeden procent zwiększa liczbę nowych przedsiębiorstw o 0,255 jednostki.
- Wzrost zmiennej współczynnik urbanizacji o jeden procent zmniejsza liczbę nowych przedsiębiorstw o 0,198 jednostki.
- Wzrost zmiennej wypadki o jeden procent zwiększa liczbę nowych przedsiębiorstw o 0,561 jednostki.

### 23.Predykcja wraz z 95% przedziałem ufności.

W kontekście prognozowania w ekonometrii prognoza punktowa oraz przedział ufności są kluczowymi pojęciami, które pomagają określić zakres pewności dotyczący przewidywanej wartości. Prognoza punktowa to pojedyncza liczba, która wskazuje na wartość przewidywaną przez model. Natomiast 95% przedział ufności to zakres wartości, w którym z określonym prawdopodobieństwem znajduje się rzeczywista wartość zmiennej zależnej.

```
Prognoza punktowa dla modelu wynosi 6,29288
Dla 95% przedziału ufności wynoszącego (5,58, 7,00576)
```

Rys. 20 Predykcja z przedziałem ufności.

W przypadku estymowanego modelu wartość prognozy punktowej wynosi 6,29, oznacza to, że na podstawie danych i zbudowanego modelu ekonometrycznego przewidujesz, że wartość zmiennej objaśniającej wyniesie około 6,29. W przypadku, gdy zmienna jest logarytmem, oznacza to, że prognoza przewiduje, że liczba nowo zarejestrowanych przedsiębiorstw wyniesie około 539 dla przedziału od 265,07 do 1096,63.

## 5. Część V – Podsumowanie

Początkowe założenie analizy miało na celu zbadanie liczby nowych przedsiębiorstw rejestrowanych w systemie REGON w polskich powiatach przez następujące zmienne: „bezrobocie”, „mieszkania”, „urbanizacja”, „wypadki”, „rozwody”, „gęstość zaludnienia”.

Zastosowanie metody Hellwiga wykazało, że największą pojemność informacyjną posiadają zmienne: „mieszkania”, „wypadki”, „rozwody”, „gęstość zaludnienia”. W wyniku metody krokowej wstecz uzyskano model ze zmiennymi: „bezrobocie”, „mieszkania”, „urbanizacja”, „wypadki”, „rozwody”, „gęstość zaludnienia”, wszystkie istotne statystycznie. Niestety każdy powstałych modeli wykazywał heteroskedastyczność. W celu znalezienia modelu, który będzie miał stałą wariancję reszt, przekształcono zmienne objaśniane i objaśniającą na logarytmy i wybrano zmienne, posiadające największą wartość p. Powstały model ma rozkład normalny reszt, homoskedastyczność, a jego zmienne to: logarytm ze zmiennej „mieszkania”, logarytm ze zmiennej „urbanizacja” oraz logarytm ze zmiennej „wypadki”.

Współczynnik determinacji wykazuje, że model wyjaśnia około 64% zmienności zmiennej objaśnianej "przedsiębiorstwa", a przeprowadzony test wykazał istotność współczynnika determinacji. Badanie efektu katalizy wykazało, że rzeczywisty związek między zmiennymi objaśniającymi a zmienną objaśnianą jest słabszy niż sugeruje współczynnik determinacji. Test normalności reszt wykazał, że reszty mają rozkład normalny. Badanie istotności zmiennej „l_mieszaknia”, „l_urbanizacja”, „l_wypadki” pokazało, że zmienne są istotne i wpływają na liczbę przedsiębiorstw. Test pominiętych zmiennych wykazał, że usunięte zmienne mogły mieć istotny wpływ na zmienną zależną, jednak taka transformacja była konieczna, w celu uniknięcia heteroskedastyczności. Test dodanych zmiennych wykazał, że dodanie logarytmów zmiennej "mieszania" i "wypadki", było dobrym krokiem, gdyż zmienne mają istotny wpływ na liczbę przedsiębiorstw w powiatach oraz poprawiły jakość modelu. Wykresy pudełkowe wykazały jakie wartości odstające posiada każda ze zmiennych. Test liczby serii wykazał, że postać modelu została dobrze dobrana. Test RESET wykazał, że wybór postaci analitycznej jest nieprawidłowy. Niestety test Chowa również wykazał problemy z parametrami modelu, gdyż została potwierdzona hipoteza o niestabilności parametrów modelu. Zmienne w modelu zostały przekształcone na logarytmy, jedynym wyjściem była dalsza transformacja modelu lub powrót do początku, zmiana zmiennych i ponowna analiza regresji. Testy na heteroskedastyczność, wykazały. że ona nie istnieje i spełnione jest założenie MNK o stałej wariancji składnika losowego. Test współliniowości wykazał brak istotnej współliniowości między zmiennymi. Koincydencja została potwierdzona. Predykcja z 95% przedziałem ufności wykazała, liczba nowo zarejestrowanych przedsiębiorstw wyniesie około 539.

Potwierdzona została hipoteza, gdzie większa liczba mieszkań oddanych do użytku przyczynia się do wzrostu liczby nowych przedsiębiorstw w Polskich powiatach. Reszta hipotez się nie potwierdziła, a pozostałych nie udało się zbadać.

Ostateczny model regresji liniowej jest dobrze dopasowany do danych, wyjaśnia dużą część zmienności zmiennej zależnej i spełnia kluczowe założenia metody najmniejszych kwadratów, takie jak brak różnorodności wariancji reszt oraz brak współliniowości między zmiennymi niezależnymi. Jednakże odrzucenie hipotezy zerowej w przypadku testu Chowa i RESET, sugerują problemy z postacią analityczną modelu i niestabilnością parametrów. Sytuacja taka jest sygnałem, że model może mieć istotne problemy analityczne, które należy dokładnie zbadać i rozwiązać, aby zapewnić rzetelność i wiarygodność wyników analizy regresji.

Podsumowując, choć ostateczny model regresji liniowej jest dobrze dopasowany i wyjaśniający, ale należy wziąć pod uwagę wyniki testu Chowa i RESET jako potencjalne ograniczenie, które wymaga dalszej analizy i ewentualnych korekt.

# 6. Bibliografia

William H. Greene, *ECONOMETRIC ANALYSIS,* New Jersey 2003, s. 122.

Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach,* Michigan 2012, s. 334-335

https://www.cambridge.org/features/economics/brooks/downloads/Answers%20to%20end%20of%20chapter%20questions/Chapter4_solutions.doc, data dostępu: 20.06.2024

Eligiusz W. Nowakowski, *Podstawy ekonometrii z elementami algebry liniowej*, Warszawa 2011, s. 61-62.

G.S Maddala, Ekonometria, Warszawa 2008.

# 7. Spis rysunków

# 8. Spis tabel