Ekonometria

Jakub Wojna

Grudzień 2024

1 Parametry modelu

Zmienne objaśniające, na których opieram dalszą analizę, pochodzą z strony Banku Danych Lokalnych, oraz z Eurostat. Strony te są bogatym oraz wiarogodnym źródłem informacji na temat danych naszego kraju.

Zdecydowałem się wybrać do modelowania parametry które mogą wpływać, przynajmniej moim zdaniem, na poziom przestępczości. Szczegółowy opis parametrów załączam poniżej:

- średnia ludności na 1 km^2
- średnia liczba ludności w tysiącach
- średnia liczba ludności w tysiącach mężczyźni
- średnia ludność w tysiącach kobiety
- średni wskaźnik urbanizacji w %
- średnia liczba bezrobotnych osób
- średnia liczba bezrobotnych mężczyzn
- średnia liczba bezrobotnych kobiet
- Średni dochód budżetu powiatów na mieszkańca
- średnie dochody budżetów powiatu

2 Główne hipotezy badawcze

- **Wpływ poziomu urbanizacji na przestępczość:** Wysoki poziom urbanizacji ma istotny wpływ na poziom przestępczości. Hipoteza zakłada, że im wyższy poziom urbanizacji w danym regionie, tym wyższy poziom przestępczości. Zwiększenie koncentracji ludzkiej i rozwoju miast może prowadzić do trudności w utrzymaniu porządku publicznego, co sprzyja wzrostowi przestępczości.
- **Wpływ średniej liczby ludności na przestępczość:** Istnieje przypuszczenie, że średnia liczba ludności w danym regionie może wpływać na poziom przestępczości. Im wyższa średnia liczba ludności, tym kolokwialnie większe ryzyko, że dwie osoby spotkają się w celach rabunkuc czy napaści. Zwiększona liczba mieszkańców może prowadzić również do większej anonimowości, trudności w monitorowaniu działań społecznych i zatem zwiększać ryzyko przestępczości.
- **Wpływ bezrobocia na poziom przestępczości:** Hipoteza zakłada, że wyższy poziom bezrobocia w danym regionie wpływa na wzrost przestępczości. Bezrobocie może prowadzić do trudności ekonomicznych, poczucia frustracji i braku perspektyw wśród osób, co może skutkować poszukiwanie nielegalnych źródeł dochodu lub wyładowywaniem negatywnych emocji poprzez działania przestępcze.

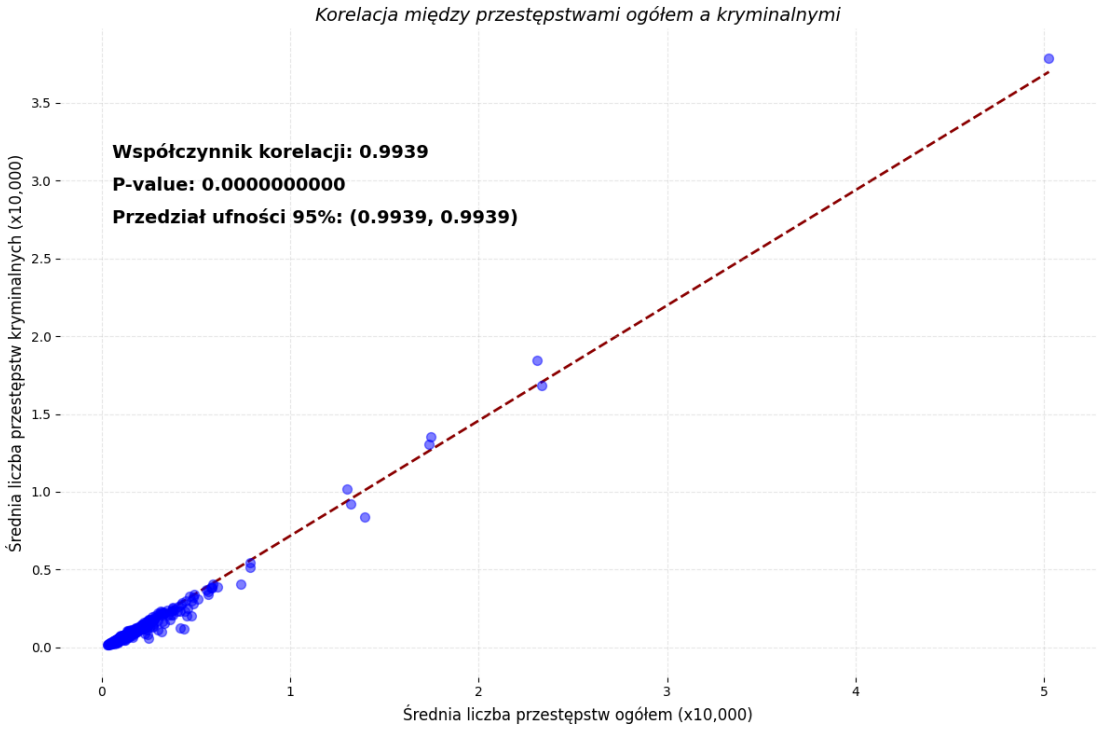
3 Dane wejściowe

Dane, na których opieram analizę, pochodzą z oficjalnej strony Banku Danych Lokalnych. Strona ta jest źródłem danych statystycznych dotyczących Polski, udostępnianych przez Główny Urząd Statystyczny.

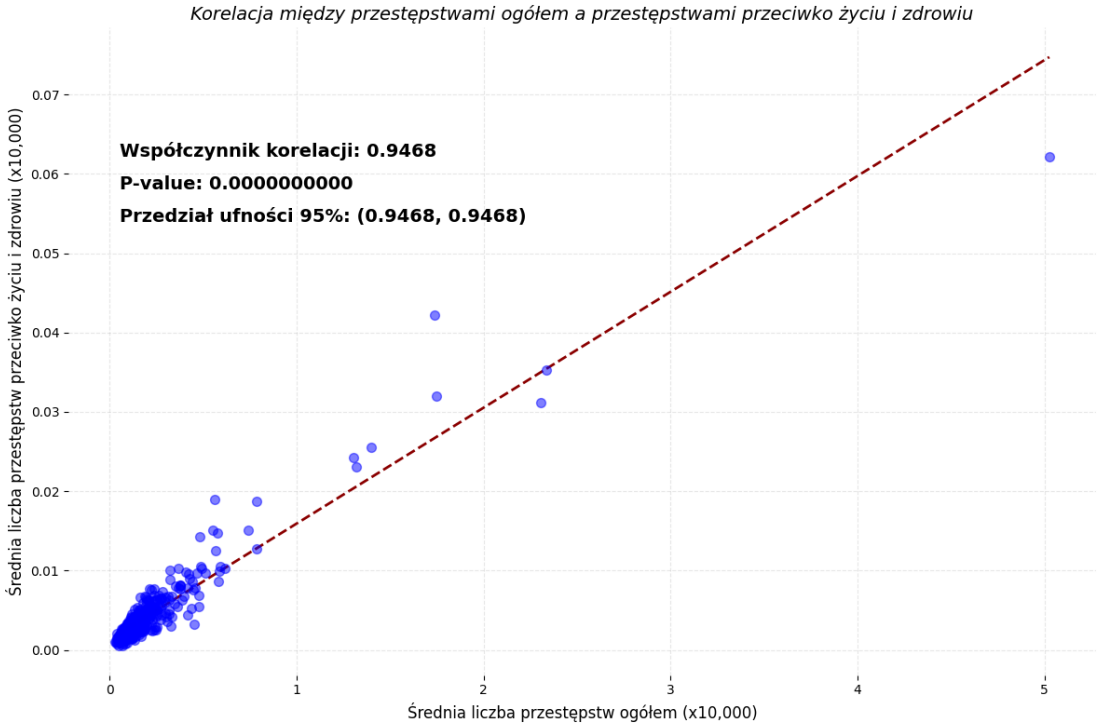
Zdecydowałem się wybrać do modelowania statystyki przestępstw w Polsce w latach 2013–2023, z podziałem na kategorie. Celem uzyskania danych przekrojowych obliczyłem średnią arytmetyczną dla każdej podkategorii danych z ww. okresów. Szczegółowy opis kategorii przestępczych załączam poniżej:

- Średnia liczba przestępstw ogółem
- Średnia liczba przestępstw o charakterze kryminalnym
- Średnia liczba przestępstw o charakterze gospodarczym
- Średnia liczba przestępstw przeciwko bezpieczeństwu powszechnemu i bezpieczeństwu w komunikacji (drogowe)
- Średnia liczba przestępstw przeciwko życiu i zdrowiu
- Średnia liczba przestępstw przeciwko mieniu
- Średnia liczba przestępstw przeciwko wolności, wolności sumienia, wolności seksualnej i obyczajności razem
- Średnia liczba przestępstw przeciwko rodzinie i opiece
- Średnia liczba przestępstw przeciwko bezpieczeństwu powszechnemu i bezpieczeństwu w komunikacji razem

Pierwszym, co zostało przeze mnie odnotowane, to fakt, że dane o przestępstwach wydawały się być ze sobą mocno skorelowane. Analizujac współczynnik korelacji Perasona otrzymujemy:

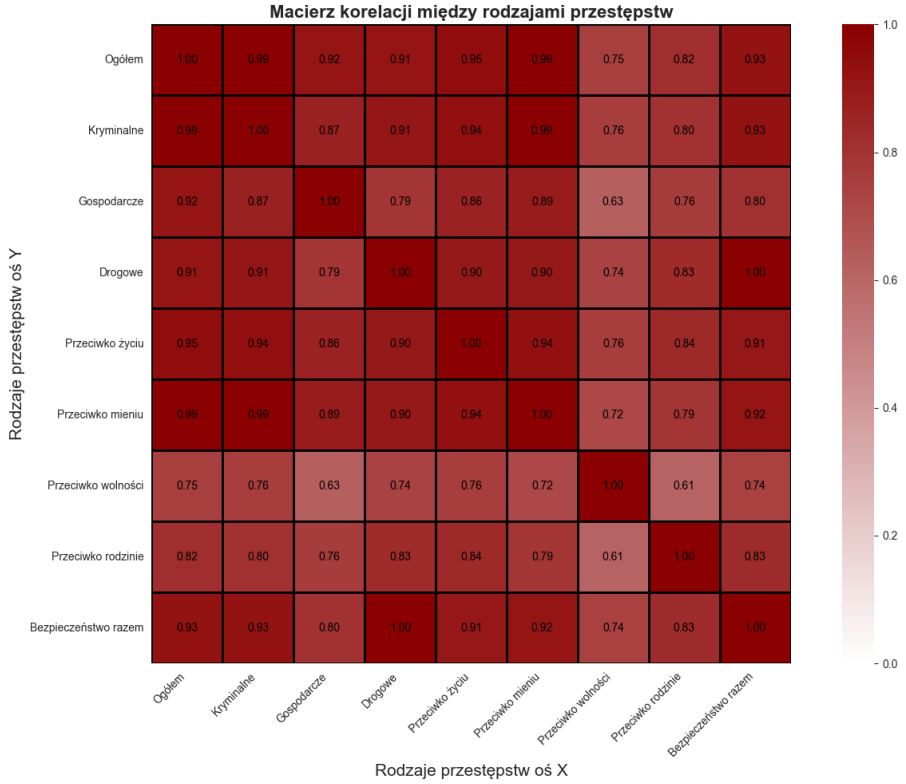


Rysunek 1: Pierwsza regresja



Rysunek 2: Kolejna regresja

Na rysunku widać wyraźną tendencję koncentracji moich zmiennych wokół linii regresji. Poniżej zamieszczam macierz korelacji dla dowolnych dwóch permutacji tych zmiennych.



Rysunek 3: Macierz korelacji moich zmiennych

Na *Rysunku 1* oraz *Rysunku 2* widać wyraźną tendencję koncentracji moich zmiennych wokół linii regresji. Macierz korelacji dla dowolnych dwóch permutacji zmiennych, przedstawiona na *Rysunku 3*, wskazuje na wysoką, bardzo bliską wartości 1 korelację między parami rozłącznych zmiennych

Przyjrzyjmy sie jak wyglada test Persona korelacji od strony matematycznej. Dla pary zmiennych losowych (X, Y) wzór na ρ jest następujący:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y},$$

gdzie $\text{cov}(X, Y)$ oznacza kowariancję zmiennych X i Y, a σ_X oraz σ_Y to odchylenia standardowe odpowiednio zmiennych X i Y.

Mamy:

$$0 \leq \mathbb{E}\left(\frac{X - \mathbb{E}X}{\sigma_X} - \frac{Y - \mathbb{E}Y}{\sigma_Y}\right)^2 = \frac{\mathbb{E}(X - \mathbb{E}X)^2}{\sigma_X^2} + \frac{\mathbb{E}(Y - \mathbb{E}Y)^2}{\sigma_Y^2} - \frac{2\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\sigma_X \sigma_Y}.$$

Ponieważ, w naszym przypadku:

$$\text{Corr}(X, Y) \approx 1,$$

to:

$$0 \leq 2 - 2\text{Corr}(X, Y) \approx 0.$$

To implikuje, że:

$$\frac{X - \mathbb{E}X}{\sigma_X} - \frac{Y - \mathbb{E}Y}{\sigma_Y} = 0 \quad \text{p.n.},$$

Można przeorganizować powyższą tożsamość, aby uzyskać:

$$Y = aX + b \quad \text{p.n},$$

gdzie:

$$a = \frac{\sigma_Y}{\sigma_X}, \quad b = \mathbb{E}Y - \frac{\sigma_Y}{\sigma_X}\mathbb{E}X.$$

Wynika z tego, że dla zmiennych silnie skorelowanych każdą z nich można wyrazić jako kombinację liniową pozostałych, z dokładnością do stałej

Dla przejrzystości i w związku z wysoką korelacją moich zmiennych, w dalszej części pracy założę, że wpływ poszczególnych parametrów na różne rodzaje przestępczości jest analogiczny, jak wpływ tych parametrów na jedną wybraną zmienną. Analizując macierz korelacji, zauważam, że najwyższe wartości osiąga zmienna „średnia przestępstw ogółem", stąd też będzie ona moim punktem odniesienia.

4 Model I - OLS

OLS Regression(Podsumowanie)

Variable	Coef.	Std. Err.	z	P>|z|
const	-0.0902	0.237	-0.381	0.704
średnia ludności na 1 km²	6.88×10^{-6}	1.39×10^{-5}	0.497	0.619
średnia liczba ludności w tys.	-2.6773	2.100	-1.275	0.202
średnia liczba ludności mężczyźni	2.6666	2.098	1.271	0.204
średnia liczba ludności kobiety	2.6919	2.104	1.279	0.201
urbanizacja (%)	0.0011	0.000	2.288	0.022
bezrobotni	-9.95×10^{-6}	5.57×10^{-6}	-1.787	0.074
bezrobotni mężczyźni	-3.79×10^{-5}	1.84×10^{-5}	-2.059	0.040
bezrobotni kobiety	2.79×10^{-5}	1.65×10^{-5}	1.692	0.091
dochód na mieszkańca	8.15×10^{-9}	4.72×10^{-8}	0.173	0.863
dochód budżetów powiatu	7.21×10^{-12}	5.58×10^{-10}	0.013	0.990

Tabela 1: OLS Regression Results (HC3 Robust Errors)

Model Summary

- **R-squared:** 0.961
- **Adj. R-squared:** 0.960
- **F-statistic:** 18.42
- **Prob (F-statistic):** 4.80×10^{-21}
- **Log-Likelihood:** 475.29
- **AIC:** -930.6
- **BIC:** -891.2
- **Observations:** 380
- **Df Residuals:** 370
- **Df Model:** 9
- **Std. Errors:** HC3 Robust
- **Multicollinearity:** Detected

Jak widać, wyniki zwykłej regresji najmniejszych kwadratów nie są zadowalające. Z jednej strony, wysoka wartość R^2 sugeruje dobre dopasowanie modelu. Z drugiej strony, brak heteroskedastyczności reszt oraz dość wysokie Log-podobieństwo budzą wątpliwości co do normalności i rozkładu reszt w moim modelu.

Interpretacja wyników modelu regresji OLS

Ogólna jakość modelu

- **R-squared (R^2): 0.961**
  Model wyjaśnia 96,1% zmienności zmiennej zależnej (*średnia liczba przestępstw ogółem*). Wysoka wartość R^2 implikuje dobre dopasowanie modelu do danych.

- **Adj. R-squared: 0.960**
  Skorygowana wartość R^2, uwzględniająca liczbę predyktorów i liczbę obserwacji, jest bardzo zbliżona do R^2, co sugeruje, że dodatkowe zmienne nie są nadmiarowe.

- **F-statistic (18.42, $p < 0.001$)**
  Wynik testu F-statystyki wskazuje, że model jako całość jest statystycznie istotny, co oznacza, że przynajmniej jeden predyktor istotnie wpływa na zmienną zależną.

- **Log-Likelihood: 475.29**
  Wartość logarytmu wiarygodności (Log-Likelihood) wskazuje, jak dobrze model opisuje dane. Wyższa wartość sugeruje lepsze dopasowanie modelu do danych, przy założeniu porównywalnej liczby parametrów.

- **AIC: -930.6**
  Akaike Information Criterion- generalnie niższa wartość AIC implikuje wyższą jakość modelu.. Wartość -930.6 świadczy o bardzo dobrym dopasowaniu modelu.

- **BIC: -891.2**
  Bayesian information criterion - tutaj jw. niższa wartość BIC implikuje wyższą jakość modelu. Wynik -891.2 wskazuje na silne dopasowanie modelu do danych przy danej liczbie parametrów.

Interpretacja zmiennych istostnych statystycznycznie (p-value 10%)[1]

- **Wskaźnik urbanizacji (%):** 0.0011 ($p = 0.022$)
  Wzrost wskaźnika urbanizacji o 1% wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 0.0011. Może to wynikać z większej koncentracji ludności w obszarach zurbanizowanych.

- **Bezrobotni (ogółem):** -9.95×10^{-6} ($p = 0.074$)
  Nieistotny statystycznie, jednak wskazuje na niewielką, negatywną korelację między liczbą bezrobotnych a przestępczością.

- **Bezrobotni mężczyźni:** -3.79×10^{-5} ($p = 0.040$)
  Statystycznie istotny. Wzrost liczby bezrobotnych mężczyzn o 1 wiąże się z istotnym spadkiem liczby przestępstw. Wynik ten może być nielogiczny i sugerować problem z danymi lub współliniowość.

- **Bezrobotni kobiety:** 2.79×10^{-5} ($p = 0.091$)
  Wynik nieistotny statystycznie, ale sugeruje niewielką tendencję do wzrostu liczby przestępstw wraz ze wzrostem liczby bezrobotnych kobiet.

[1] Poniżej przedstawiam interpretację współczynników przy założeniu stałości pozostałych zmiennych (ceteris paribus).

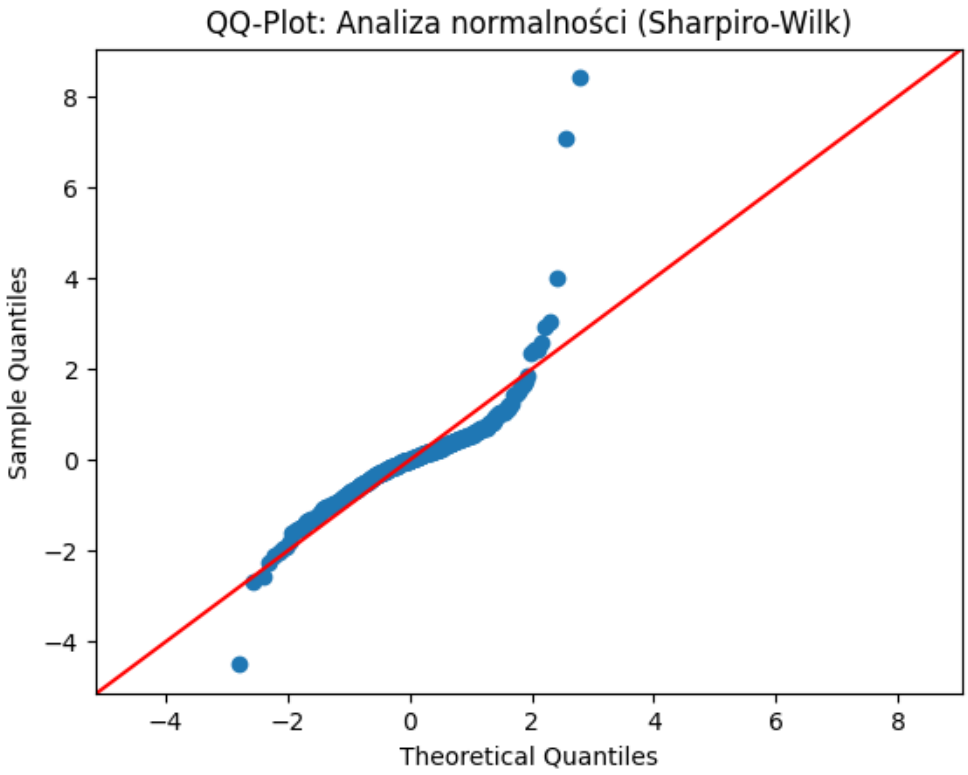
Zmienne nieisotne statystycznie (p-value 10%)

- **Stała (*const*):** -0.0902 $(p = 0.704)$
  Wartość stałej nie jest istotna statystycznie, co oznacza, że trudno wyciągnąć wnioski dotyczące średniej liczby przestępstw przy zerowych wartościach wszystkich parametrów.
- **Średni dochód budżetu powiatów na mieszkańca:** 8.15×10^{-9} $(p = 0.863)$
  Zmienna nieistotna statystycznie, ponadto zupełnie nieistotna praktycznie (zerowy rząd wiekości).
- **Średnia ludności na 1 km²:** 6.88×10^{-6} $(p = 0.619)$
  Wzrost średniej gęstości ludności o 1 jednostkę powoduje niewielki, nieistotny praktycznie wzrost liczby przestępstw, pomimo tego zmienna sama w sobie jest istotna statystycznie.
- **Średnie dochody budżetów powiatu:** 7.21×10^{-12} $(p = 0.990)$
  Również nieistotna statystycznie, jak i praktycznie.
- **Średnia liczba ludności w tysiącach:**
  - **Ogółem:** -2.6773 $(p = 0.202)$
    Wzrost liczby ludności o 1000 osób wiąże się z nieistotnym statystycznie spadkiem średniej liczby przestępstw.
  - **Mężczyźni:** 2.6666 $(p = 0.204)$
    Wzrost liczby mężczyzn o 1000 osób wiąże się z nieistotnym statystycznie wzrostem średniej liczby przestępstw.
  - **Kobiety:** 2.6919 $(p = 0.201)$
    Wzrost liczby kobiet o 1000 osób wiąże się z podobnym nieistotnym statystycznie wzrostem średniej liczby przestępstw.

5 Co działa, a co nie? - KMRL

Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - Statystyka LM: 64.97
  - Wartość p (LM): 4.11×10^{-10}
  - Statystyka F: 7.61
  - Wartość p (F): 4.59×10^{-11}
- **Test Goldfelda-Quandta:**
  - Statystyka F: 1.09
  - Wartość p: 0.290
- **Test Shapiro-Wilka:**
  - Statystyka: 0.9587
  - Wartość p: 7.60×10^{-9}
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 28.61
  - Wartość p: 2.83×10^{-12}
- **Test Jacques-Bera:**



Rysunek 4: Test S-Wilka

  - Statystyka: 223.80
  - Wartość p: 2.53×10^{-49}

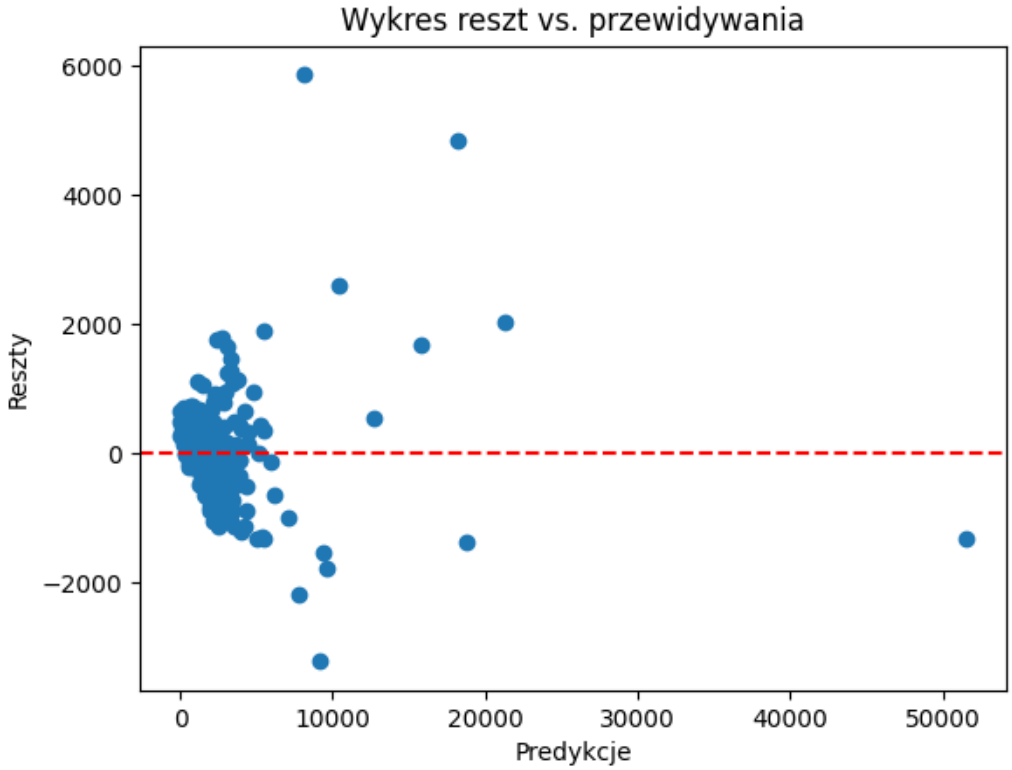
- **Test Durbin-Watsona:**
  - Statystyka: 1.6035

Interpretacja wyników

- **Test Breuscha-Pagana**: Heteroskedastyczność
  - Wartość p (LM: 4.11×10^{-10}, F: 4.59×10^{-11}) jest znacznie mniejsza niż typowy poziom istotności ($\alpha = 0.05$).
  - **Interpretacja**: Odrzucam hipotezę zerową o braku heteroskedastyczności. Wskazuje to na obecność heteroskedastyczności w modelu, co narusza założenia klasycznej regresji liniowej.
- **Test Shapiro-Wilka**: Normalność rozkładu reszt w modelu
  - Wartość p (7.60×10^{-9}) - znacznie mniejsza niż $\alpha = 0.05$.
  - **Interpretacja**: Odrzucam hipotezę zerową o normalności reszt. Wskazuje to na istotne odstępstwa od normalności.
- **Test RESET (forma funkcyjna)**: Poprawność specyfikacji modelu
  - Wartość p (2.83×10^{-12}) - znacznie mniejsza niż $\alpha = 0.05$.
  - **Interpretacja**: Odrzucam hipotezę zerową o poprawnej specyfikacji modelu. Wyniki sugerują, że model może być niepoprawnie wyspecyfikowany. Zapewne problem leży w niewłaściwej forma funkcji.
- **Test Jacques-Bera**: Normalność rozkładu reszt

- Wartość p (2.53×10^{-49}) - znacznie mniejsza niż $\alpha = 0.05$.
- **Interpretacja**: Odrzucam hipotezę zerową o normalności reszt. Wyniki wskazują na istotne odstępstwa od normalności, zgodne z wynikami testu Shapiro-Wilka.

- **Test Durbina-Watsona**: Autokorelacja reszt
  - Statystyka (1.6035) sugeruje możliwość umiarkowanej dodatniej autokorelacji, ponieważ wynik odbiega od wartości 2, oznaczającej brak autokorelacji.
  - **Interpretacja**: Wynik sugeruje konieczność dalszej analizy autokorelacji, ponieważ sam w sobie jest zbyt blisko wartości 2, aby był rozstrzygający.



Rysunek 5: Multikorelacja reszt

Ocena spełnienia założeń KMRL

- **Założenie o liniowości modelu:**
  - To oczywiście występuje.
- **Założenie o braku heteroskedastyczności:**
  - Wynik testu Breuscha-Pagana ($p < 0.05$) wskazuje na obecność heteroskedastyczności.
- **Założenie o normalności rozkładu reszt:**
  - Wyniki testów Shapiro-Wilka ($p < 0.05$) i Jacques-Bera ($p < 0.05$) jednoznacznie wskazują na brak normalności reszt.
- **Założenie o braku autokorelacji reszt:**
  - Statystyka Durbina-Watsona (1.6035) sugeruje umiarkowaną dodatnią autokorelację - potencjalne naruszenie.

- **Założenie o poprawnej specyfikacji modelu:**
  - Wynik testu RESET ($p < 0.05$) wskazuje na błędną specyfikację modelu - niewłaściwa forma funkcji lub błędy w strukturze modelu.

Jak mogę to poprawić

Na podstawie wyników testów mogę spróbować moderować poszczególne rezultaty:

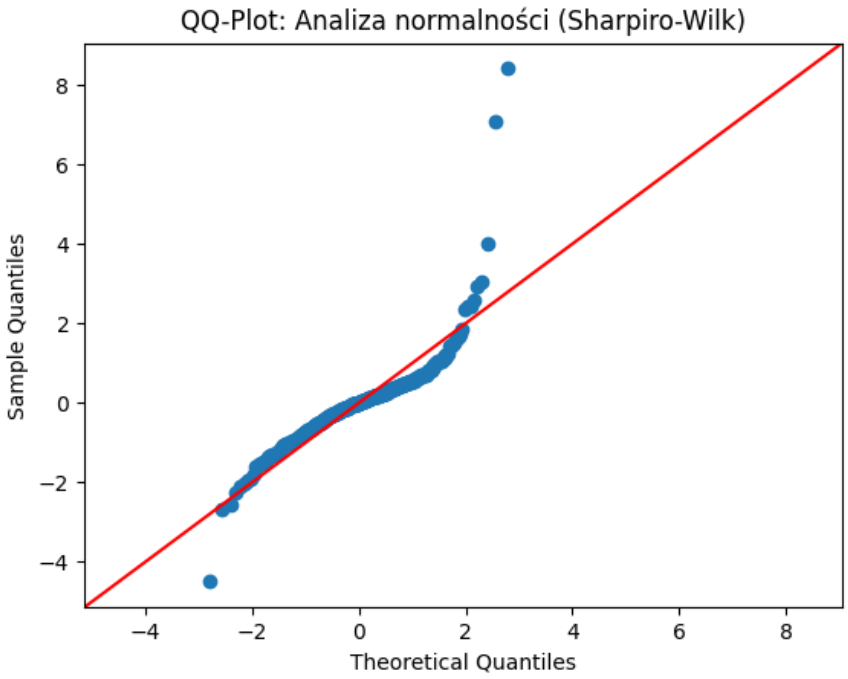
1. **Rozwiązywanie heteroskedastyczności:**
   - Zastosowanie heteroskedastyczności-odpornej estymacji wariancji np. macierz White'a.
   - Rozważenie przekształceń zmiennych
2. **Poprawa normalności reszt:**
   - Sprawdzenie danych pod kątem wartości odstających i wyrzucenie ich.
   - Rozważenie przekształceń zmiennych zależnych lub niezależnych.
3. **Zarządzanie autokorelacją reszt:**
   - Użycie modelu GLM
4. **Poprawa specyfikacji modelu:**
   - Weryfikacja, czy wszystkie istotne zmienne zostały uwzględnione w modelu.
   - Rozważenie nieliniowych relacji (np. interakcje, kwadratowe zmienne).

6 Korekta OLS - transformacja zmiennej

Zmienna zależna podzielona przez 10000

Tabela 2: OLS Regression Results

Dependent Variable:	średnia liczba przestępstw ogółem
R-squared:	**0.961**
Adjusted R-squared:	**0.960**
F-statistic:	**18.42**
Prob (F-statistic):	**4.80e-21**
Model:	OLS
Method:	Least Squares
Log-Likelihood:	475.29
No. Observations:	380
Df Residuals:	370
Df Model:	9
AIC:	**-930.6**
BIC:	**-891.2**
Covariance Type:	**HC3**
Omnibus:	**260.451**
Durbin-Watson:	**1.545**
Jarque-Bera (JB):	**6963.706**
Prob(JB):	**0.00**
Skew:	2.452
Kurtosis:	23.390
Cond. No.:	**3.21e+20**
Smallest eigenvalue:	**5.04** $\times 10^{-21}$



Rysunek 6: Test Sharpio-Wilka

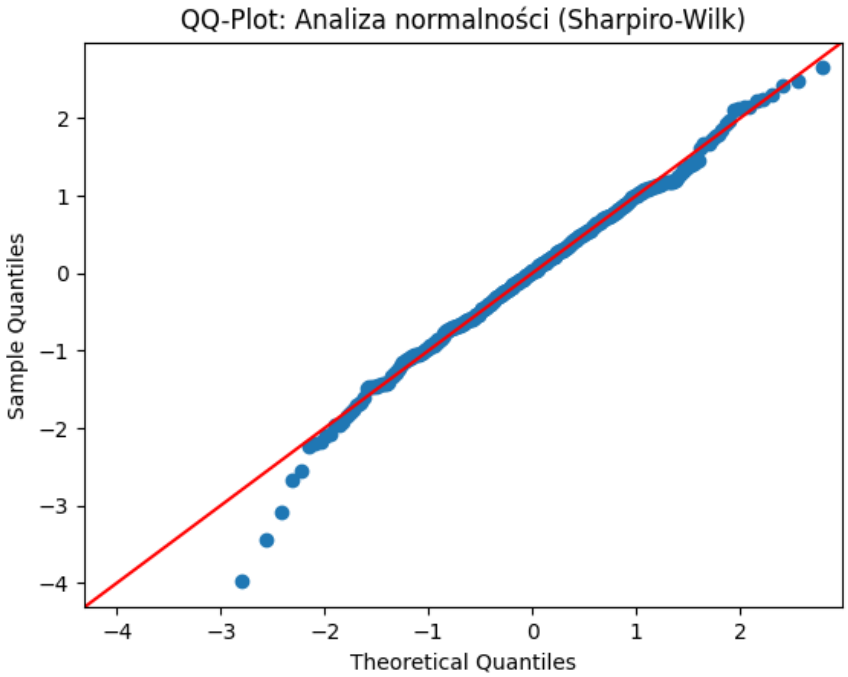
Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - Statystyka LM: 92.94
  - Wartość p (LM): 1.39×10^{-15}
  - Statystyka F: 11.95
  - Wartość p (F): 6.53×10^{-18}
- **Test Shapiro-Wilka:**
  - Statystyka: 0.8039
  - Wartość p: 4.00×10^{-21}
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 53.64
  - Wartość p: 3.67×10^{-21}
- **Test Jacques-Bera:**
  - Statystyka: 6403.21
  - Wartość p: 0.00
- **Test Durbin-Watsona:**
  - Statystyka: 1.5815

Zmienna zależna podzielona przez 1000 i logarytmowana

Tabela 3: OLS Regression Results

Dependent Variable:	średnia liczba przestępstw ogółem
R-squared:	**0.845**
Adjusted R-squared:	**0.841**
F-statistic:	**41.30**
Prob (F-statistic):	**7.63e-43**
Model:	OLS
Method:	Least Squares
Log-Likelihood:	-69.553
No. Observations:	380
Df Residuals:	370
Df Model:	9
AIC:	**159.1**
BIC:	**198.5**
Covariance Type:	**HC3**
Omnibus:	**9.894**
Durbin-Watson:	**1.595**
Jarque-Bera (JB):	**11.747**
Prob(JB):	**0.00281**
Skew:	-0.273
Kurtosis:	3.666
Cond. No.:	**3.21e+20**
Smallest eigenvalue:	**5.04** $\times 10^{-21}$



Rysunek 7: Test Sharpio-Wilka

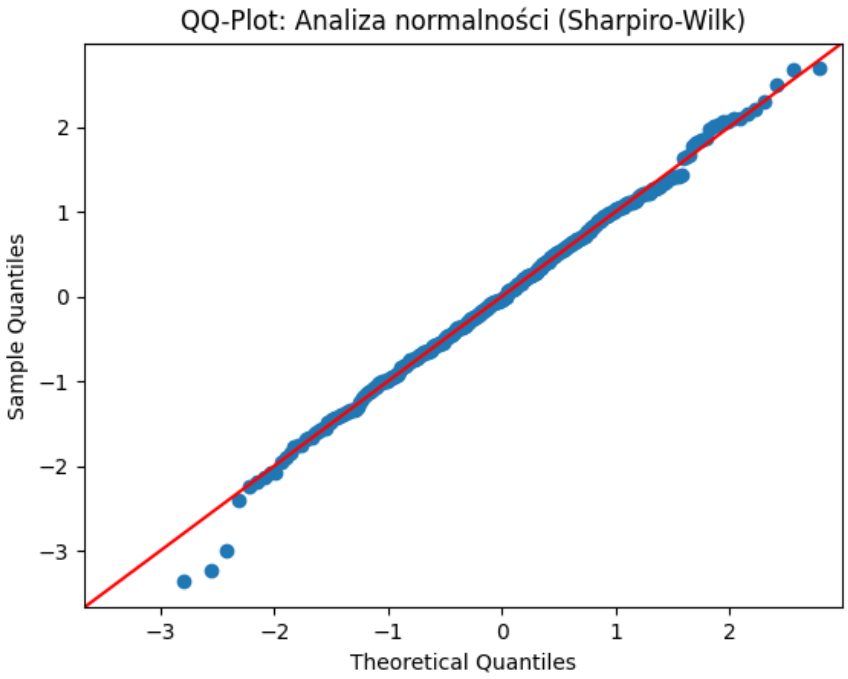
Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - Statystyka LM: 22.61
  - Wartość p (LM): 0.0123
  - Statystyka F: 2.33
  - Wartość p (F): 0.0112
- **Test Shapiro-Wilka:**
  - Statystyka: 0.9965
  - Wartość p: 0.57465
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 43.05
  - Wartość p: 1.60×10^{-17}
- **Test Jacques-Bera:**
  - Statystyka: 1.31
  - Wartość p: 0.5185
- **Test Durbin-Watsona:**
  - Statystyka: 1.6604

Zmienna zależna tylko logarytmowana

Tabela 4: OLS Regression Results

Dependent Variable:	średnia liczba przestępstw ogółem
R-squared:	**0.845**
Adjusted R-squared:	**0.841**
F-statistic:	**41.30**
Prob (F-statistic):	**7.63e-43**
Model:	OLS
Method:	Least Squares
Log-Likelihood:	-69.553
No. Observations:	380
Df Residuals:	370
Df Model:	9
AIC:	**159.1**
BIC:	**198.5**
Covariance Type:	**HC3**
Omnibus:	**9.894**
Durbin-Watson:	**1.595**
Jarque-Bera (JB):	**11.747**
Prob(JB):	**0.00281**
Skew:	-0.273
Kurtosis:	3.666
Cond. No.:	**3.21e+20**
Smallest eigenvalue:	**5.04** $\times 10^{-21}$



Rysunek 8: Test Sharpio-Wilka

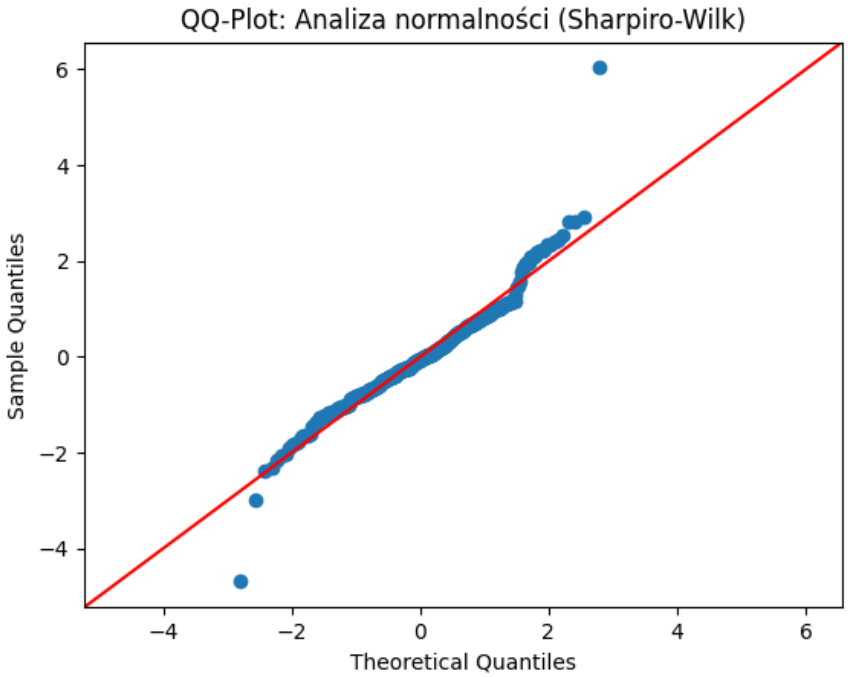
Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - LM statistic: 22.61
  - LM p-value: 1.23×10^{-2}
  - F statistic: 2.33
  - F p-value: 1.12×10^{-2}
- **Test Shapiro-Wilka:**
  - Statistic: 0.9965
  - p-value: 5.75×10^{-1}
- **Test RESET (forma funkcyjna):**
  - F statistic: 43.05
  - p-value: 1.60×10^{-17}
- **Test Jacques-Bera:**
  - Statistic: 1.31
  - p-value: 5.18×10^{-1}
- **Test Durbina-Watsona:**
  - Statistic: 1.66

Zmienna zależna pierwiastek kwadratowy

Tabela 5: OLS Regression Results

Dependent Variable:	średnia liczba przestępstw ogółem
R-squared:	**0.928**
Adjusted R-squared:	**0.926**
F-statistic:	**33.13**
Prob (F-statistic):	**1.09e-35**
Model:	OLS
Method:	Least Squares
Log-Likelihood:	-1187.8
No. Observations:	380
Df Residuals:	370
Df Model:	9
AIC:	**2396.0**
BIC:	**2435.0**
Covariance Type:	**HC3**
Omnibus:	**65.290**
Durbin-Watson:	**1.508**
Jarque-Bera (JB):	**339.921**
Prob(JB):	**1.54e-74**
Skew:	0.589
Kurtosis:	7.481
Cond. No.:	**3.21e+20**
Smallest eigenvalue:	**5.04** $\times 10^{-21}$



Rysunek 9: Test Sharpio-Wilka

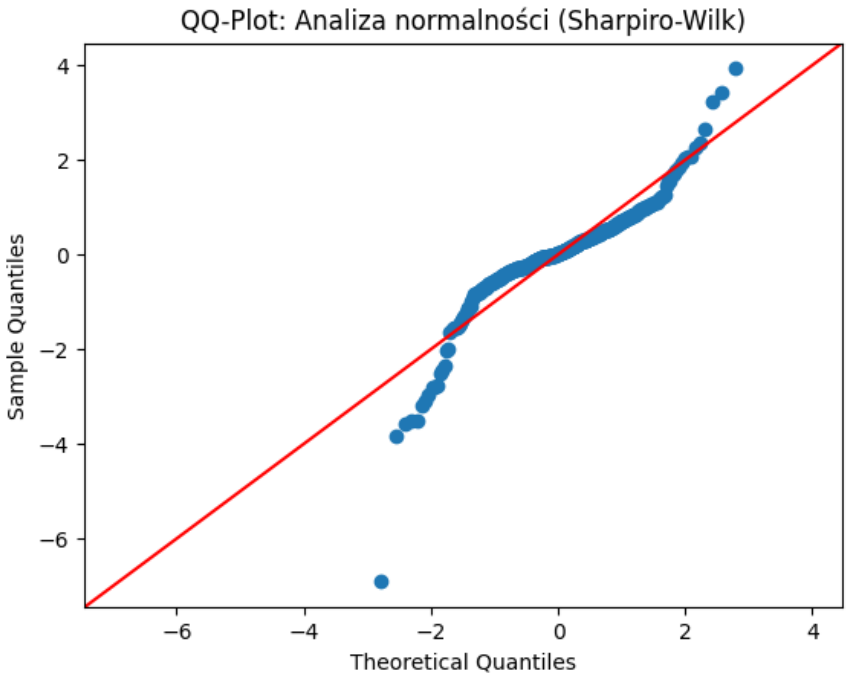
Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - LM statistic: 64.97
  - LM p-value: 4.11×10^{-10}
  - F statistic: 7.61
  - F p-value: 4.59×10^{-11}
- **Test Shapiro-Wilka:**
  - Statistic: 0.9587
  - p-value: 7.60×10^{-9}
- **Test RESET (forma funkcyjna):**
  - F statistic: 28.61
  - p-value: 2.83×10^{-12}
- **Test Jacques-Bera:**
  - Statistic: 223.80
  - p-value: 2.53×10^{-49}
- **Test Durbina-Watsona:**
  - Statistic: 1.60

Zmienna zależna podniesiona do kwadratu

Tabela 6: OLS Regression Results

Dependent Variable:	średnia liczba przestępstw ogółem
R-squared:	**0.987**
Adjusted R-squared:	**0.987**
F-statistic:	**15.44**
Prob (F-statistic):	**9.41e-18**
Model:	OLS
Method:	Least Squares
Log-Likelihood:	-6828.1
No. Observations:	380
Df Residuals:	370
Df Model:	9
AIC:	**1.368e+04**
BIC:	**1.372e+04**
Covariance Type:	**HC3**
Omnibus:	**260.451**
Durbin-Watson:	**1.545**
Jarque-Bera (JB):	**6963.706**
Prob(JB):	**0.00**
Skew:	2.452
Kurtosis:	23.390
Cond. No.:	**3.21e+20**
Smallest eigenvalue:	**5.04** $\times 10^{-21}$



Rysunek 10: Test Sharpio-Wilka

Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - Statystyka LM: 140.70
  - Wartość p (LM): 3.03×10^{-25}
  - Statystyka F: 21.69
  - Wartość p (F): 8.78×10^{-32}
- **Test Shapiro-Wilka:**
  - Statystyka: 0.8635
  - Wartość p: 9.47×10^{-18}
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 588.45
  - Wartość p: 3.20×10^{-115}
- **Test Jacques-Bera:**
  - Statystyka: 1263.83
  - Wartość p: 3.65×10^{-275}
- **Test Durbin-Watsona:**
  - Statystyka: 1.7034

Wnioski

Niezależnie od przeprowadzenia jednorodnych transformacji na zmiennej, każdy z modeli napotyka identyczne problemy jak model wyjsciowy. Parametry modelu wyjściowego z względem parametrów modeli transformowanych nie różnią się praktycznie w ogóle, ponadto nie przechodzą tych samych testów diagnostycznych. Z racji tego, iż nie chciałbym powtarzać analogicznego komentarza, odnosnik do oceny poszczególnych parametrów zamieszczam tutaj.

Pomimo braku spełnienia KMRL warto odnotować jeszcze jedną kwestię: **w każdym z prezentowanych przykładów najmniejsza z wartości własnych macierzy wariancji jest bardzo mała (rzędu stałej $\times 10^{-21}$).**

Wartość własna macierzy projektowania odpowiada za miarę kierunku w przestrzeni zmiennych objaśniających. Jeśli wartość własna jest bardzo mała, oznacza to, że jedna lub więcej zmiennych objaśniających jest bliska bycia liniowo zależną (lub silnie skorelowaną) z innymi zmiennymi w modelu. To oznacza, że w danych istnieje silna multikolinearność. Zjawisko to prowadzi do problemów w estymacji parametrów modelu, takich jak:

- Niestabilne oszacowania współczynników regresji: Może to prowadzić do dużych i niestabilnych oszacowań współczynników regresji, które mogą się zmieniać w zależności od drobnych zmian w danych.
- Silna kolinearność: Sprawia, że testy statystyczne na współczynniki mogą wskazywać na brak istotności, nawet jeśli zmienne są faktycznie istotne.
- Wysoka korelacja między zmiennymi sprawia, że założenia klasycznego twierdzenia Gaussa-Markova nie są spełnione.

7 Model II - WLS

WLS Regression (Podsumowanie)

Variable	Coef.	Std. Err.	t	P>|t|
const	-1097.7092	76.654	-14.320	0.000
średnia ludności na 1 km^2	0.0080	0.032	0.250	0.803
średnia liczba ludności w tysiącach	-26310.0	1775.806	-14.814	0.000
średnia liczba ludności w tysiącach mężczyźni	26200.0	1776.489	14.748	0.000
średnia liczba ludności w tysiącach kobiety	26460.0	1774.751	14.906	0.000
średni wskaźnik urbanizacji w %	11.0356	0.201	54.817	0.000
średnia liczba bezrobotnych osób	-0.0971	0.002	-47.447	0.000
średnia liczba bezrobotnych mężczyzn	-0.3629	0.011	-34.099	0.000
średnia liczba bezrobotnych kobiet	0.2658	0.011	23.655	0.000
średni dochód budżetu powiatów na mieszkańca	1.0×10^{-4}	1.59×10^{-5}	7.521	0.000
średnie dochody budżetów powiatu	-3.255×10^{-8}	1.31×10^{-7}	-0.248	0.804

Tabela 7: WLS Regression Results

Model Summary

- **R-squared:** 0.999
- **Adj. R-squared:** 0.999
- **F-statistic:** 5.831×10^4
- **Prob (F-statistic):** 0.000
- **Log-Likelihood:** -2602.7
- **AIC:** 5225.0
- **BIC:** 5265.0
- **Observations:** 380
- **Df Residuals:** 370
- **Df Model:** 9
- **Covariance Type:** Nonrobust

Procedura dostosowania wag na podstawie oszacowania wariancji błędu

Aby dostosować wagi na podstawie oszacowania wariancji błędu dla każdej obserwacji skorzystałem z poniższego przepisu:

1. **Pierwsze dopasowanie modelu OLS:** Dopasowałem model OLS, aby uzyskać reszty.
2. **Szacowanie wariancji błędu:** Obliczyłem kwadraty reszt jako estymację wariancji błędu.
3. **Utworzenie wag:** Wagi zdefiniowałem jako odwrotność oszacowanej wariancji błędu. .
4. **Dopasowanie modelu WMNK:** Dopasowałem model WLS z użyciem obliczonych wag.

W wyniku tej procedury otrzymujemy estymatory o mniejszych wariancjach, co może poprawić dokładność szacunków w obecności heteroskedastyczności.

W tym wpadku dane, na pierwszy rzut oka, wygladają bardziej obiecująco. Praktycznie wszystkie p-value są mniejsze niż 5%. Co prawda wątpliwości budzą ponadprzeciętnie wysokie współczynniki przy parametrach, sprawdzę więc jakość tych oszacowań w toku dalszych badań statystycznych.

Warto zauważyć, że w każdym z rozpatrywanych modeli zastosowano heteroskedastyczną korekcję dla macierzy wariancji (HC3 Robust), co powinno wykluczyć problemy związane z heteroskedastycznością.

Interpretacja wyników modelu regresji WLS

Ogólna jakość modelu

- **R-squared (R^2): 0.999**
  Model wyjaśnia 99,9% zmienności zmiennej zależnej (*średnia liczba przestępstw ogółem*). Tak wysoka wartość R^2 świadczy idealnym dopasowaniu modelu do danych.
- **Adj. R-squared: 0.999**
  Skorygowana wartość R^2, uwzględniająca liczbę predyktorów i liczbę obserwacji, jest praktycznie identyczna z R^2, co sugeruje, że w modelu nie ma nadmiarowych zmiennych.

- **F-statistic:** 5.831×10^4**,** $p < 0.001$
  Bardzo wysoka wartość F-statystyki oraz jej istotność ($p < 0.001$) wskazują, że model jako całość jest statystycznie istotny. Oznacza to, że przynajmniej jedna z uwzględnionych zmiennych niezależnych istotnie wpływa na zmienną zależną.

- **Log-Likelihood: -2602.7**
  Mimo ujemnej wartości, co może być typowe dla dużych modeli z dużą liczbą obserwacji, wartość ta jest zgodna z ogólnym wysokim dopasowaniem modelu.

- **AIC: 5225.0**
  Wartość 5265.0 wskazuje na dobre dopasowanie modelu przy uwzględnieniu liczby parametrów., ale należy ją interpretować w kontekście porównania z innymi testami statystycznymi.

- **BIC: 5265.0**
  Wartość 5265.0 wskazuje na dobre dopasowanie modelu przy uwzględnieniu liczby parametrów.

Interpretacja zmiennych istotnych statystycznie (p-value 10%)[2]

- **Średnia liczba ludności w tysiącach:** -2.631×10^4 $(p = 0.000)$
  Wzrost liczby ludności o 1000 osób wiąże się ze statystycznie istotnym spadkiem liczby przestępstw o 26,310. Może to wskazywać na większą rozproszenie ludności, co obniża intensywność przestępczości.

- **Średnia liczba ludności w tysiącach mężczyzn:** 2.62×10^4 $(p = 0.000)$
  Wzrost liczby mężczyzn o 1000 osób wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 26,200. Możliwe, że większa liczba mężczyzn może prowadzić do większej liczby incydentów przestępczych.

- **Średnia ludność w tysiącach kobiet:** 2.646×10^4 $(p = 0.000)$
  Wzrost liczby kobiet o 1000 osób wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 26,460. Jest to podejrzane, jak bardzo ta liczba jest podobna do analogicznego parametru u mężczyzn. Wymaga to dalszej analizy.

- **Średni wskaźnik urbanizacji w %:** 11.0356 $(p = 0.000)$
  Wzrost wskaźnika urbanizacji o 1% wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 11.036. Może to wynikać z większej koncentracji ludności w obszarach zurbanizowanych, gdzie często występuje wyższa przestępczość.

- **Średnia liczba bezrobotnych osób:** -0.0971 $(p = 0.000)$
  Wzrost liczby bezrobotnych o 1 osobę wiąże się ze statystycznie istotnym spadkiem liczby przestępstw o 0.0971. Wynik ten może wskazywać na zjawisko, w którym wzrost bezrobocia prowadzi do zmniejszenia przestępczości. Moim zdaniem nie jest to logiczne, gdyż historycznie własnie wysokie bezrobocie było silnie skorelowane z przestępczością.

- **Średnia liczba bezrobotnych mężczyzn:** -0.3629 $(p = 0.000)$
  Wzrost liczby bezrobotnych mężczyzn o 1 osobę wiąże się ze statystycznie istotnym spadkiem liczby przestępstw o 0.3629. Jest to również mało intuicyjne.

- **Średnia liczba bezrobotnych kobiet:** 0.2658 $(p = 0.000)$
  Wzrost liczby bezrobotnych kobiet o 1 osobę wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 0.2658, jak wyżej.

- **Średni dochód budżetu powiatów na mieszkańca:** 0.0001 $(p = 0.000)$
  Wzrost dochodu budżetu powiatów na mieszkańca o 1 jednostkę wiąże się ze statystycznie istotnym wzrostem liczby przestępstw o 0.0001. Zmienna jest istotna statystycznie, ale praktycznie nie ma znaczenia.

[2]Poniżej przedstawiam interpretację współczynników przy założeniu stałości pozostałych zmiennych (ceteris paribus).

Zmienne nieistotne statystycznie (p-value 10%)

- **Średnia ludności na 1 km²:** 0.0080 ($p = 0.803$)
  Wzrost średniej ludności na 1 km² o 1 jednostkę nie ma statystycznie, ani praktycznie wpływu na liczbę przestępstw, co sugeruje, że gęstość ludności na obszarze nie jest istotnym czynnikiem w wyjaśnianiu przestępczości.
- **Średnie dochody budżetów powiatu:** -3.255×10^{-8} ($p = 0.804$)
  Zmienna ta nie jest istotna statystycznie, a jej wpływ na liczbę przestępstw jest znikomy. Praktycznie nie ma żadnego znaczenia, co sugeruje brak wpływu dochodów powiatu na przestępczość.

Wyniki testów statystycznych

- **Test Breuscha-Pagana:**
  - Statystyka LM: 87.02
  - Wartość p (LM): 2.08×10^{-14}
  - Statystyka F: 12.21
  - Wartość p (F): 6.02×10^{-17}
- **Test Shapiro-Wilka:**
  - Statystyka: 0.7917
  - Wartość p: 1.03×10^{-21}
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 62.14
  - Wartość p: 5.61×10^{-24}
- **Test Jacques-Bera:**
  - Statystyka: 7596.95
  - Wartość p: 0.00×10^{0}
- **Test Durbin-Watsona:**
  - Statystyka: 1.5522

Wnioski - KMRL

- **Założenie o liniowości modelu:**
  Założenie to jest spełnione.
- **Założenie o braku heteroskedastyczności:**
  Test Breuscha-Pagana ($p < 0.05$) wskazuje na obecność heteroskedastyczności.
- **Założenie o normalności rozkładu reszt:**
  Testy Shapiro-Wilka ($p < 0.05$) i Jacques-Bera ($p < 0.05$) jednoznacznie wskazują na brak normalności reszt.
- **Założenie o braku autokorelacji reszt:**
  Statystyka Durbina-Watsona (1.6035) sugeruje dodatnią autokorelację reszt.
- **Założenie o poprawnej specyfikacji modelu:**
  Test RESET ($p < 0.05$) wskazuje na błędną specyfikację modelu - niewłaściwa forma funkcji.

8 Model III - Generalized linear model

Występuje brak heteroskedastyczności dla OLS i WLS, sprobuję więc zastosować GLM z różnymi rodzinami rozkładów. Jest szansa że moje parametry pochodzą od innych rozkładów, stąd wynika brak heteroskedastyczności.

OLS - Binomial

Nie będę znów dodawał statystyk poszczeólnych parametrów. Zajmuje to wiele miejsca a nie ma znaczenia, o ile model nie spełnia założeń. Zajmę się samymi założeniami. Na podstawie wyników testów oraz współczynników VIF przeanalizowałem, czy model spełnia założenia twierdzenia Gaussa-Markova.

Współliniowość (Multikolinearność)

- Współczynniki VIF dla niektórych zmiennych są ekstremalnie wysokie:
  - **„średnia liczba ludności w tysiącach”**: 5.97×10^9,
  - **„średnia liczba ludności w tysiącach mężczyźni”**: 1.28×10^9,
  - **„średnia ludność w tysiącach kobiety”**: 1.72×10^9,
  - **„średnia liczba bezrobotnych osób”, „mężczyzn”, „kobiet”**: *nieskończone*.
- **Interpretacja**: Wysokie wartości VIF wskazują na bardzo silną korelację między zmiennymi, co oznacza problem współliniowości. Może to prowadzić do niestabilności oszacowań parametrów modelu.

Homoskedastyczność (Równa wariancja reszt)

- Analiza reszt deviance:
  - Średnia reszt deviance: 390.533 (powinna być bliska 0),
  - Odchylenie standardowe reszt deviance: 197.184 (wysoka wartość sugeruje heteroskedastyczność).
- **Interpretacja**: Nierównomierność reszt wskazuje na możliwą heteroskedastyczność.

Autokorelacja reszt

- Statystyka Durbin-Watson: 1.545.
- **Interpretacja**: Wartość bliska 2 wskazuje na brak autokorelacji. Wynik 1.545 sugeruje możliwą autokorelację.

Podsumowanie

Model **nie spełnia założeń twierdzenia Gaussa-Markova** z następujących powodów:

1. **Silna współliniowość** między zmiennymi niezależnymi.
2. **Heteroskedastyczność** w resztach.
3. **Autokorelacja** reszt.

OLS - Gamma

Współliniowość (Multikolinearność) - Druga część

- Współczynniki VIF dla zmiennych:
  - **„średnia ludności na 1 km²”**: 3.384401,
  - **„średnia liczba ludności w tysiącach”**: 5.974002×10^9,
  - **„średnia liczba ludności w tysiącach mężczyźni”**: 1.282514×10^9,
  - **„średnia liczba bezrobotnych osób”**: ∞,
  - **„średnia liczba bezrobotnych mężczyzn”**: ∞,
  - **„średnia liczba bezrobotnych kobiet”**: ∞,
  - **„średni dochód budżetu powiatów na mieszkańca”**: 2.321866,
- **Interpretacja**: Wysokie wartości VIF, w tym ∞, wskazują na problem współliniowości. Zmienne takie jak liczba ludności, bezrobocie i dochody mają bardzo silną korelację, co może prowadzić do niestabilności modelu.

Homoskedastyczność (Równa wariancja reszt)

- Analiza reszt:
  - Średnia reszt: 0 jest okej
  - Odchylenie standardowe reszt: 687.450 (wysoka wartość sugeruje heteroskedastyczność).
- **Interpretacja**: Duże odchylenie standardowe reszt wskazuje na heteroskedastyczność.

Autokorelacja reszt

- Statystyka Durbin-Watson: 1.545.
- **Interpretacja**: Wartość bliska 2 sugeruje brak autokorelacji. Jednak wynik 1.545 może wskazywać na obecność słabej autokorelacji.

Podsumowanie

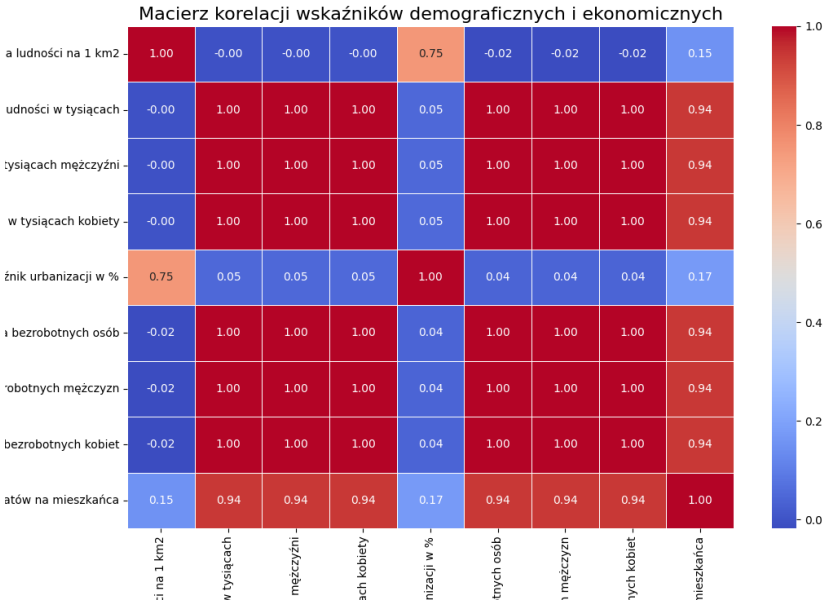
Model **nie spełnia założeń twierdzenia Gaussa-Markova** z następujących powodów:

1. **Silna współliniowość** między zmiennymi niezależnymi (wysokie VIF i ∞).
2. **Problem heteroskedastyczności** - Durbin-Watson

Każdy inny OLS (dla innych rozkładów)

Problem wystepuje dla każdego rozkładu ten sam. VIF wybucha, odchylenie standardowe reszt jest ogromne.

9 Autokorelacja macierzy parametrów

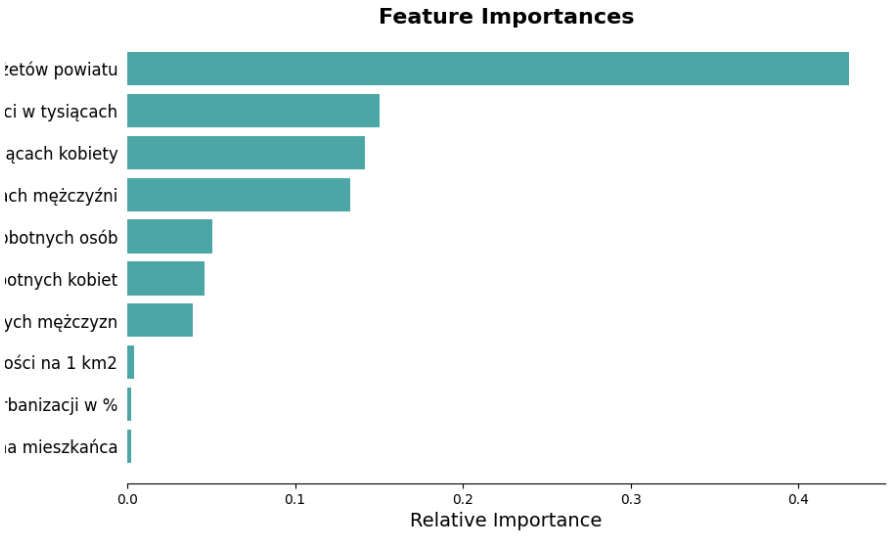


Rysunek 11: Macierz korelacji parametrów

Wszystkie zmienne ktore na macierzy korelacji sa czerwone sa ze soba absolutnie skorelowane.

10 Algorytm: Random Forrest

Chcialbym usunąc z mojej bazy parametrów te ktorę są mocno skorelowane, z drugiej strony chciałbym zostawić te, które są statystycznie istotne.

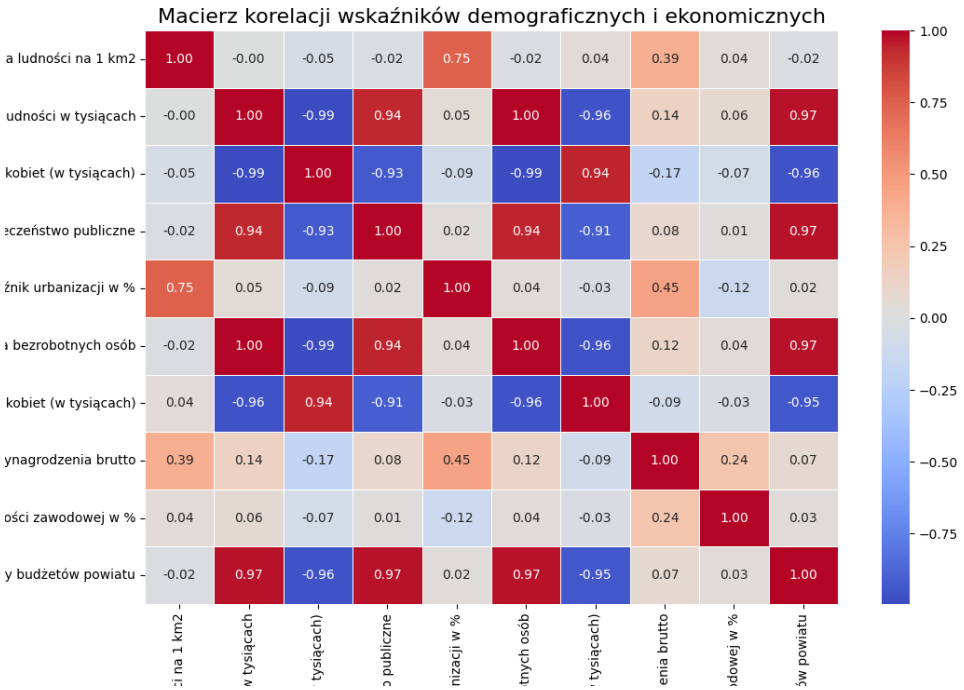


Rysunek 12: Random Forrest - które parametry mam zostawić

Dalsze kroki i poprawione dane

- **Zostawione zmienne ze względu na ich znaczenie lub niską kolinearność:**
  - **średnie dochody budżetu powiatu:** Wyniki Random Forest wskazują na najwyższą istotność tej zmiennej w modelu, dlatego pozostaje w analizie.
  - **średnia liczba ludności w tysiącach:** Również istotna w wynikach Random Forest, więc pozostaje w modelu.
  - **średnia liczba bezrobotnych osób:** jak wyżej
  - **wskaźnik urbanizacji:** Słabo skorelowana z innymi zmiennymi, co czyni ją wartościową do pozostawienia w modelu.
  - **liczba ludności na 1 km²:** Słabo skorelowana z innymi zmiennymi, dlatego może pozostać.
- **Zmienne przekształcone w nowe wskaźniki:**
  - **średnia liczba ludności w tysiącach mężczyźni** oraz **średnia liczba ludności w tysiącach kobiety:**
  - średnia przewaga liczby mężczyzn nad liczbą kobiet (w tysiącach)
  - **średnia liczba bezrobotnych mężczyzn** oraz **średnia liczba bezrobotnych kobiet:**
  - średnia przewaga liczby bezrobotnych mężczyzn nad liczbą bezrobotnych kobiet (w tysiącach)
- **Dodane zmienne:**
  - **średnie miesięczne wynagrodzenia brutto**
  - **średni poziom aktywności zawodowej**
  - **średnie roczne wydatki na bezpieczeństwo publiczne**

11 Macierz korelacji dl poprawionych danych



Rysunek 13: Sytuacja widocznie się poprawiła

12 OLS dla poprawionych danych

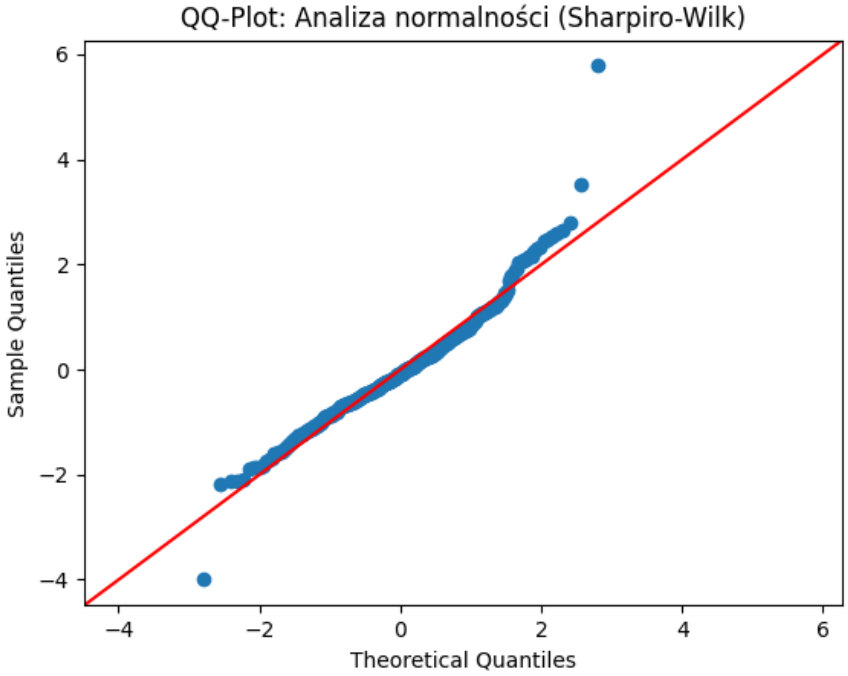
Variable	Coef.	Std. Err.	z	P>|z|
const	529.1341	1823.510	0.290	0.772
średnia ludności na 1 km^2	0.1485	0.244	0.607	0.544
średnia liczba ludności w tys.	20.9790	18.709	1.121	0.262
średnia przewaga mężczyzn nad kobietami (tys.)	-116.0967	76.174	-1.524	0.127
wydatki na bezpieczeństwo (roczne)	4.13×10^{-7}	1.66×10^{-6}	0.249	0.803
urbanizacja (%)	8.8707	2.351	3.773	0.000
bezrobotni	-0.1713	0.071	-2.404	0.016
bezrobotni mężczyźni - kobiety (tys.)	-0.2897	0.164	-1.766	0.077
miesięczne wynagrodzenie brutto	0.1848	0.213	0.867	0.386
aktywność zawodowa (%)	-30.5388	20.112	-1.518	0.129
dochody budżetów powiatów	1.05×10^{-7}	3.58×10^{-6}	0.029	0.977

Tabela 8: OLS Regression Results (HC3 Robust Errors)

Model Summary

- **R-squared:** 0.962
- **Adj. R-squared:** 0.961
- **F-statistic:** 12.02
- **Prob (F-statistic):** 2.79×10^{-15}
- **Log-Likelihood:** -3021.5
- **AIC:** 6065.0
- **BIC:** 6108.0
- **Observations:** 380
- **Df Residuals:** 369
- **Df Model:** 10
- **Std. Errors:** HC3 Robust
- **Multicollinearity:** Detected

13 Niestety syuacja dalej nie uległa poprawie - dalsze testy



- **Test Breuscha-Pagana:**
  - Statystyka LM: 140.6961
  - Wartość p (LM): 3.03×10^{-25}
  - Statystyka F: 21.6950
  - Wartość p (F): 8.78×10^{-32}
- **Test Shapiro-Wilka:**
  - Statystyka: 0.8635
  - Wartość p: 9.47×10^{-18}
- **Test RESET (forma funkcyjna):**
  - Statystyka F: 588.4461
  - Wartość p: 3.20×10^{-115}
- **Test Jacques-Bera:**
  - Statystyka: 1263.8332
  - Wartość p: 3.65×10^{-275}
- **Test Durbin-Watsona:**
  - Statystyka: 1.7034

Testy Breuscha-Pagana dla heteroskedastyczności każdej zmiennej:

- **Średnia ludności na 1 km²:** p-value = 0.0000
- **Średnia liczba ludności w tysiącach:** p-value = 0.0000
- **Średnia przewaga liczby mężczyzn nad liczbą kobiet (w tysiącach):** p-value = 0.0000
- **Średnie roczne wydatki na bezpieczeństwo publiczne:** p-value = 0.5454
- **Średni wskaźnik urbanizacji w %:** p-value = 0.0000
- **Średnia liczba bezrobotnych osób:** p-value = 0.0000
- **Średnia przewaga liczby bezrobotnych mężczyzn nad liczbą bezrobotnych kobiet (w tysiącach):** p-value = 0.8641
- **Średnie miesięczne wynagrodzenia brutto:** p-value = 0.0000
- **Średni poziom aktywności zawodowej w %:** p-value = 0.0014
- **Średnie dochody budżetów powiatu:** p-value = 0.0000

Współczynniki VIF (wielokolinearność):

Variable	VIF
const	521.361542
Średnia ludności na 1 km^2	3.074709
Średnia liczba ludności w tysiącach	43.370412
Średnia przewaga liczby mężczyzn nad liczbą kobiet (w tysiącach)	35.751616
Średnie roczne wydatki na bezpieczeństwo publiczne	1.022381
Średni wskaźnik urbanizacji w %	3.234747
Średnia liczba bezrobotnych osób	5.185496
Średnia przewaga liczby bezrobotnych mężczyzn nad liczbą bezrobotnych kobiet (w tysiącach)	1.318869
Średnie miesięczne wynagrodzenia brutto	1.535240
Średni poziom aktywności zawodowej w %	1.724428
Średnie dochody budżetów powiatu	52.076162

Wnioski

1. Testy Breuscha-Pagana dla heteroskedastyczności

Niskie wartości p ($p < 0.05$) sugerują obecność heteroskedastyczności.

- Zmienna **średnie roczne wydatki na bezpieczeństwo publiczne** oraz **średnia przewaga liczby bezrobotnych mężczyzn nad liczbą bezrobotnych kobiet (w tysiącach)** mają wysokie wartości p ($p > 0.05$), co oznacza brak heteroskedastyczności.
- Wszystkie inne zmienne mają niskie wartości p ($p < 0.05$), co wskazuje na obecność heteroskedastyczności.

Model nie spełnia założenia jednorodności wariancji błędów.

2. Współczynniki VIF (wielokolinearność)

Wartości powyżej 10 sugerują istotne problemy:

- **średnia liczba ludności w tysiącach** ($VIF = 43.37$)
- **średnia przewaga liczby mężczyzn nad liczbą kobiet** ($VIF = 35.75$)
- **średnie dochody budżetów powiatu** ($VIF = 52.08$)
- **const** ($VIF = 521.36$) — to wynik stałej w modelu.

Wielokolinearność jest poważnym problemem w tym modelu, zwłaszcza dla wyżej wymienionych zmiennych. Mógłbym spróbować:

- Usunąć zmienne z wysokim VIF. Móbłym naprawic problem ze wspoliniowoscia, poprzez usunienie danych z wysokim VIF ale osłabiłoby isotność wyników w moim modelu.
- Rozważenie odrzucenia mniej istotnych zmiennych na podstawie analizy wartości p w wynikach regresji.

3. Test Shapiro-Wilka dla normalności reszt

Wartość $p < 0.05$ w teście Shapiro-Wilka wskazuje na odrzucenie hipotezy zerowej o normalności reszt. Wynik: p-value $= 4.0010 \times 10^{-21}$, co oznacza, że reszty nie są normalnie rozłożone.

Ocena spełnienia założeń KMRL

- **Założenie o liniowości modelu:**
  - Występuje.
- **Założenie o braku heteroskedastyczności:**
  - Wynik testu Breuscha-Pagana ($p < 0.05$) wskazuje na obecność heteroskedastyczności.
- **Założenie o normalności rozkładu reszt:**
  - Wyniki testów Shapiro-Wilka ($p < 0.05$) i Jacques-Bera ($p < 0.05$) jednoznacznie wskazują na brak normalności reszt.
- **Założenie o braku autokorelacji reszt:**
  - Statystyka Durbina-Watsona (1.6035) sugeruje umiarkowaną dodatnią autokorelację - potencjalne naruszenie.
- **Założenie o poprawnej specyfikacji modelu:**
  - Wynik testu RESET ($p < 0.05$) wskazuje na błędną specyfikację modelu - niewłaściwa forma funkcji lub błędy w strukturze modelu.

14 Podsumowanie

Zastosowałem w modelowaniu moich danych wiele metod, między innymi:

- metodę najmniejszych kwadratów,
- metodę ważonych najmniejszych kwadratów,
- metodę *generalized linear model* z różnymi rodzinami zmiennych,
- transpozycje zmiennych do OLS i WLS,
- metodę *Random Forest* połączoną z dodaniem nowych danych i usunięciem problematycznych.

Niestety, moje modele w dalszym ciągu nie spełnia założeń KMRL.

Wnioski

- Obecne problemy z heteroskedastycznością, autokorelacją reszt i normalnością rozkładu wskazują na potrzebę dalszych prac nad transformacjami danych oraz doborem odpowiednich metod modelowania.
- Wyniki wskazują na istotne współliniowości w danych, co wpływa na stabilność oszacowań współczynników regresji.
- Problemy z poprawną specyfikacją modelu sugerują konieczność bardziej zaawansowanej analizy, np. wykorzystania nieliniowych modeli regresji lub analizy przyczynowości.
- Problemy z formą funkcyjną sugerują konieczność analizy wzorców w resztach i przetestowania alternatywnych form zmiennych, takich jak transformacje logarytmiczne, wielomiany, czy interakcje między zmiennymi. Wdrożenie elastycznych metod, np. regresji splajnów lub modeli nieliniowych, może również poprawić wyniki.
- Mimo problemów, zastosowane metody pozwoliły na zidentyfikowanie kluczowych zmiennych wpływających na modelowaną zmienną zależną, co daje solidną podstawę do dalszych prac.

Plany dalszych badań

- Eksperymenty z dodatkowymi zmiennymi objaśniającymi, np. danymi panelowymi, mogą poprawić jakość modelu.
- Zastosowanie metod analizy klastrów w celu grupowania danych o podobnych cechach i identyfikacji różnic między klastrami, co może pomóc w lepszym dopasowaniu modelu do specyficznych podgrup danych.
- Zastosowanie metod analizy wielowymiarowej, takich jak PCA, w celu redukcji współliniowości między zmiennymi.
- Rozważenie jeszcze wielu modyfikacji liniowej obecnych zmiennych, np. przez zastosowanie funkcji potęgowych lub logarytmów o róznych podstawach.

15 Tabela typu publikacyjnego OLS, WLS, WHITE

Tabela 9:

	OLS	White (HC3)	WLS
const	-901.7564***	-901.7564	-1097.7092***
	(316.5795)	(2369.6706)	(76.6541)
średnia ludności na 1 km2	0.0689	0.0689	0.0080
	(0.1024)	(0.1387)	(0.0321)
średnia liczba ludności w tysiącach	-26772.7172	-26772.7172	-26306.5454***
	(23182.1073)	(20998.5732)	(1775.8058)
średnia liczba ludności w tysiącach mężczyźni	26666.3350	26666.3350	26198.9817***
	(23182.9288)	(20977.5565)	(1776.4886)
średnia ludność w tysiącach kobiety	26919.1856	26919.1856	26455.1214***
	(23181.3639)	(21042.7257)	(1774.7508)
średni wskaźnik urbanizacji w %	10.7711***	10.7711**	11.0356***
	(2.3121)	(4.7087)	(0.2013)
średnia liczba bezrobotnych osób	-0.0995***	-0.0995*	-0.0971***
	(0.0188)	(0.0557)	(0.0020)
średnia liczba bezrobotnych mężczyzn	-0.3789***	-0.3789**	-0.3629***
	(0.1302)	(0.1840)	(0.0106)
średnia liczba bezrobotnych kobiet	0.2794**	0.2794*	0.2658***
	(0.1307)	(0.1651)	(0.0112)
średni dochód budżetu powiatów na mieszkańca	0.0001	0.0001	0.0001***
	(0.0001)	(0.0005)	(0.0000)
średnie dochody budżetów powiatu	0.0000	0.0000	-0.0000
	(0.0000)	(0.0000)	(0.0000)
R-squared	0.9614	0.9614	0.9993
R-squared Adj.	0.9605	0.9605	0.9993
R-squared	0.9614	0.9614	0.9993
Observations	380	380	380

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

16 Tabela typu publikacyjnego - ciekawe GLM'y

Tabela 10:

	GLM (Gaussian)	GLM (Binomial)
const	-0.0000	-3959478087211936.0000***
	(0.0101)	(3442612.0849)
x1	-0.0000	14676237.5573***
	(0.0000)	(0.3900)
x2	0.0126	-763653432983969.7500***
	(0.0186)	(6333284.3096)
x3	-910.4001	-10007440313882038272.0000***
	(782.2491)	(266085255068.8257)
x4	420.1324	4592038731862589440.0000***
	(362.4461)	(123287547524.3075)
x5	491.4862	5426745595176462336.0000***
	(419.9917)	(142861915870.6449)
x6	0.0835***	1320141617179005.5000***
	(0.0178)	(6052678.1639)
x7	-0.0388***	-570268382696197.8750***
	(0.0068)	(2329601.9308)
x8	-0.1568***	-2681880253553903.0000***
	(0.0475)	(16144300.0299)
x9	0.0836*	1616645835745699.5000***
	(0.0473)	(16101098.5560)
x10	0.0199	-964158500138222.5000***
	(0.0154)	(5245738.4517)
x11	0.0209	1112941958032670 8.0000***
	(0.0775)	(26376755.0095)
Observations	380	380

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01