
**AKADEMIA GÓRNICZO – HUTNICZA**
**im. Stanisława Staszica w Krakowie**

WYDZIAŁ ZARZĄDZANIA

# Ekonometria Danych Panelowych

Temat projektu:

*Analiza modeli ekonometrycznych dla danych panelowych opisujących koszt przelotów lotniczych w USA w latach 1970-1984.*

Imię i nazwisko autorów: Karol Kuciński, Mateusz Mulka

Kierunek: Informatyka i ekonometria II stopnień
Rok studiów: 2024, studia stacjonarne

# 1. Prezentacja i opis danych

➢ Opis wykorzystywanych danych i ich źródło.

Dane pochodzą z NYU Stern School of Business, z książki "Econometric Analysis, 5th Edition". Zestaw danych obejmuje dane przekrojowe, zawierające informacje o kosztach amerykańskich linii lotniczych, od 1970 do 1984 roku. Nasz dataset zawiera w sumie 90 obserwacji z 6 firm lotniczych.

1.1 Zmienne w danych panelowych

W tym zestawie danych panelowych używane są następujące zmienne:

- I = Linia lotnicza | Oznacza różne firmy lotnicze, a w tym przypadku mamy 6 linii lotniczych.

- T = Rok | Oznacza rok, dla którego obserwowane są dane przekrojowe. Okres T zawiera się od 1970 do 1984 roku, czyli przez 15 lat.

- PF = Cena paliwa w tysiącach $ | Indeks i dane cenowe pokazują globalną średnią cenę, jaką linie lotnicze płacą rafinerii za paliwo lotnicze w danym dniu. Cena paliwa jest zawsze ustalana na podstawie umowy negocjowanej między firmą lotniczą a dostawcą paliwa lotniczego. Zazwyczaj jest to wieloletnia umowa z warunkami określającymi cenę niezależnie od fluktuacji na rynku.

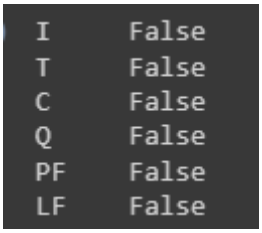
- LF = Wskaźnik wykorzystania floty, czyli średnie wykorzystanie zdolności przewozowej floty

To miara stopnia wykorzystania przewozowej zdolności linii lotniczej. Zależy głównie od pojemności miejsc w samolocie, trasy lotu, popytu itp.

- Q = Wydajność, w milach pasażerskich |  Mile pasażerskie to miara statystyczna branży transportowej, pokazująca liczbę mil, którą przelecieli pasażerowie, zazwyczaj jest to statystyka ruchu lotniczego. Q jest wyznaczane w jednostkach RPM  (revenue passenger mile czyli mile pasażerskie lub mile pasażersko-przychodowe), jest to jednostka miary specyficzna dla branży przewozowej. RPM jest obliczana poprzez pomnożenie liczby płacących pasażerów i przebytego przez nich dystansu.

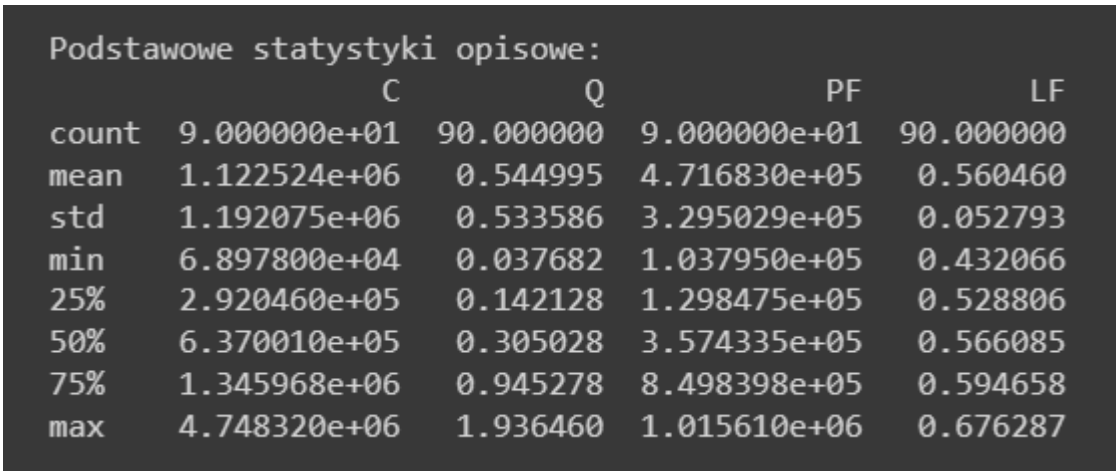
- C = Całkowity koszt, w tysiącach dolarów. | Zależy od różnych czynników, w tym ceny paliwa, wskaźnika wykorzystania floty, dzierżawy i amortyzacji, konserwacji samolotów, kosztów pracy oraz opłat za obsługę na lotnisku. Jest to też nasza zmienna objaśniana.

# Braki Danych



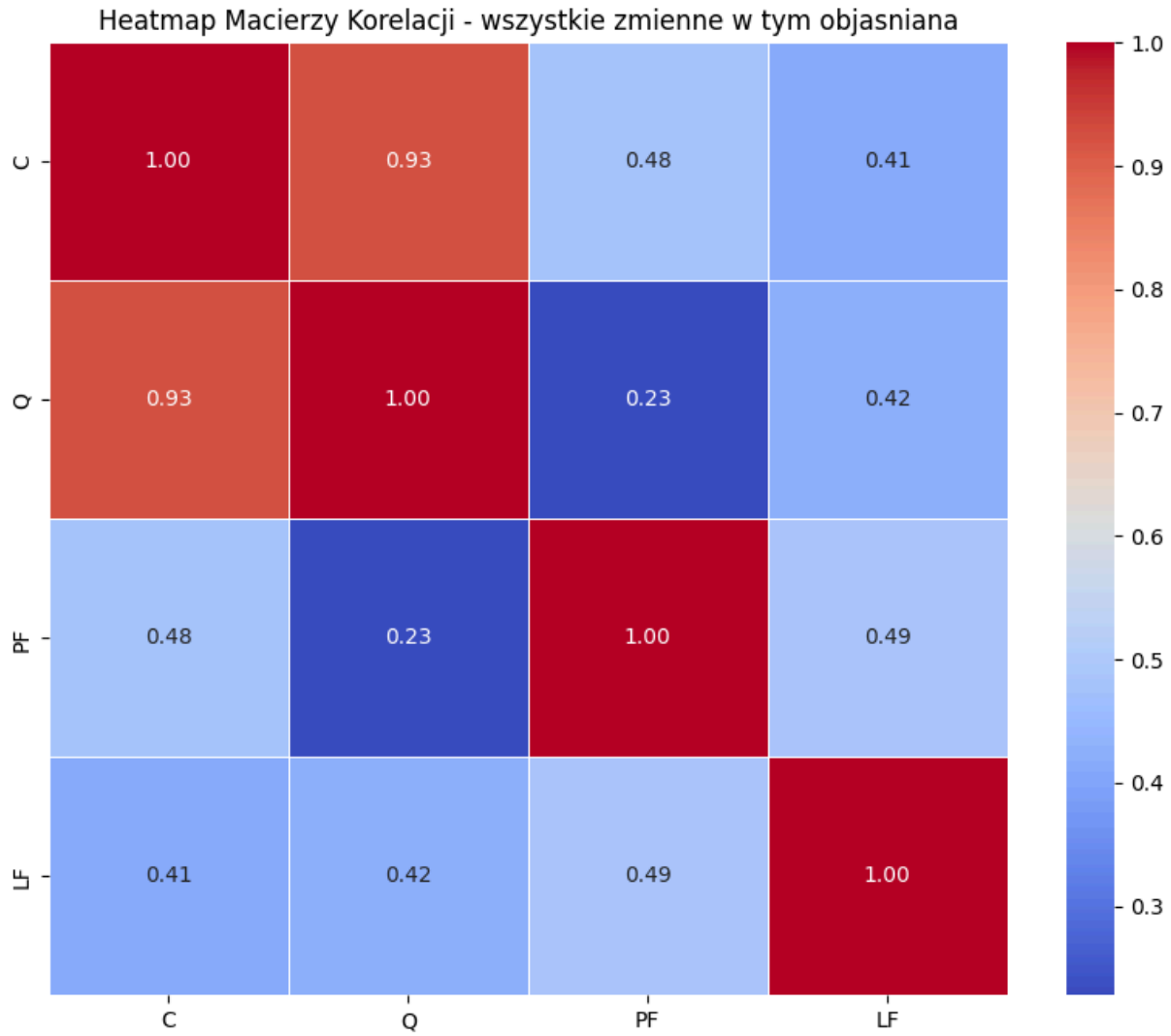

W początkowej fazie analizy sprawdzone zostały braki danych. Wybrany zbiór nie posiada braków danych na co wskazują wartość False dla każdej zmiennej.

# Analiza podstawowych statystyk.

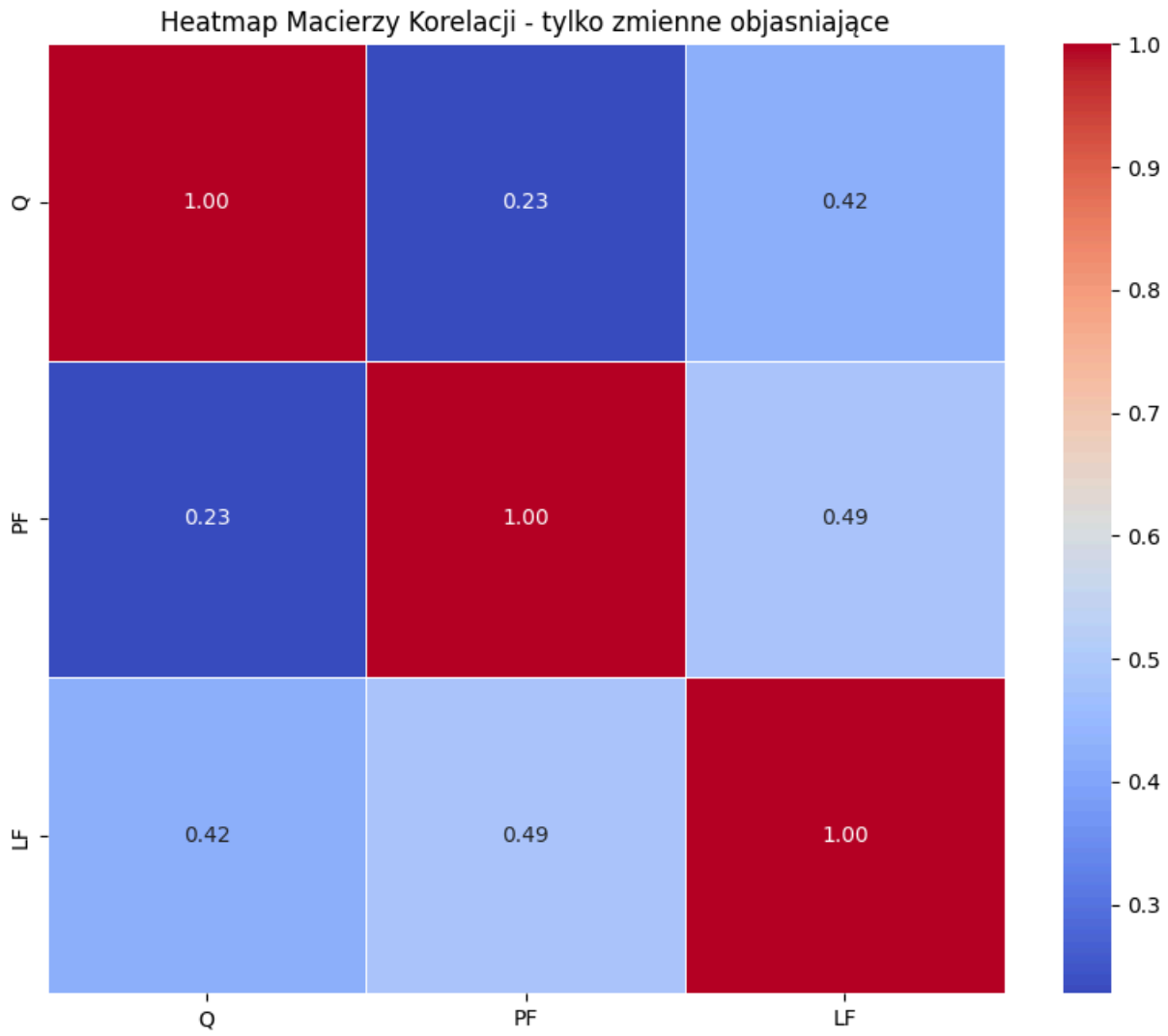



Średni całkowity koszt dla linii lotniczych(C) wynosi około 1,122,524 tysięcy dolarów. Jednakże, wartości te różnią się znacznie, co widoczne jest w dużym odchyleniu standardowym wynoszącym około 1,192,075 tysięcy dolarów. Najniższy koszt wynosił 68,978 tysięcy dolarów, a najwyższy aż 4,748,320 tysięcy dolarów.
Średnia wydajność(Q), wyrażona w milach pasażerskich, wynosi około 0.545. Jednakże, wartości są dosyć zróżnicowane, co sugeruje znaczną zmienność w osiąganej wydajności między różnymi liniami lotniczymi.
Średnia cena paliwa(PF) dla linii lotniczych to około 471,683 tysięcy dolarów. Wartości cen paliwa różnią się znacznie, co potwierdza wysokie odchylenie standardowe wynoszące około 329,5029 tysięcy. Najniższa cena paliwa wyniosła 103,795 tysięcy dolarów, a najwyższa 1,015,610 tysięcy dolarów.
Średni wskaźnik wykorzystania floty(LF) wynosi około 0.560, co wskazuje na średnie wykorzystanie przewozowej zdolności floty linii lotniczych. Wartości są stosunkowo zbliżone, z odchyleniem standardowym wynoszącym około 0.052793.

# Analiza zależności pomiędzy zmiennymi.

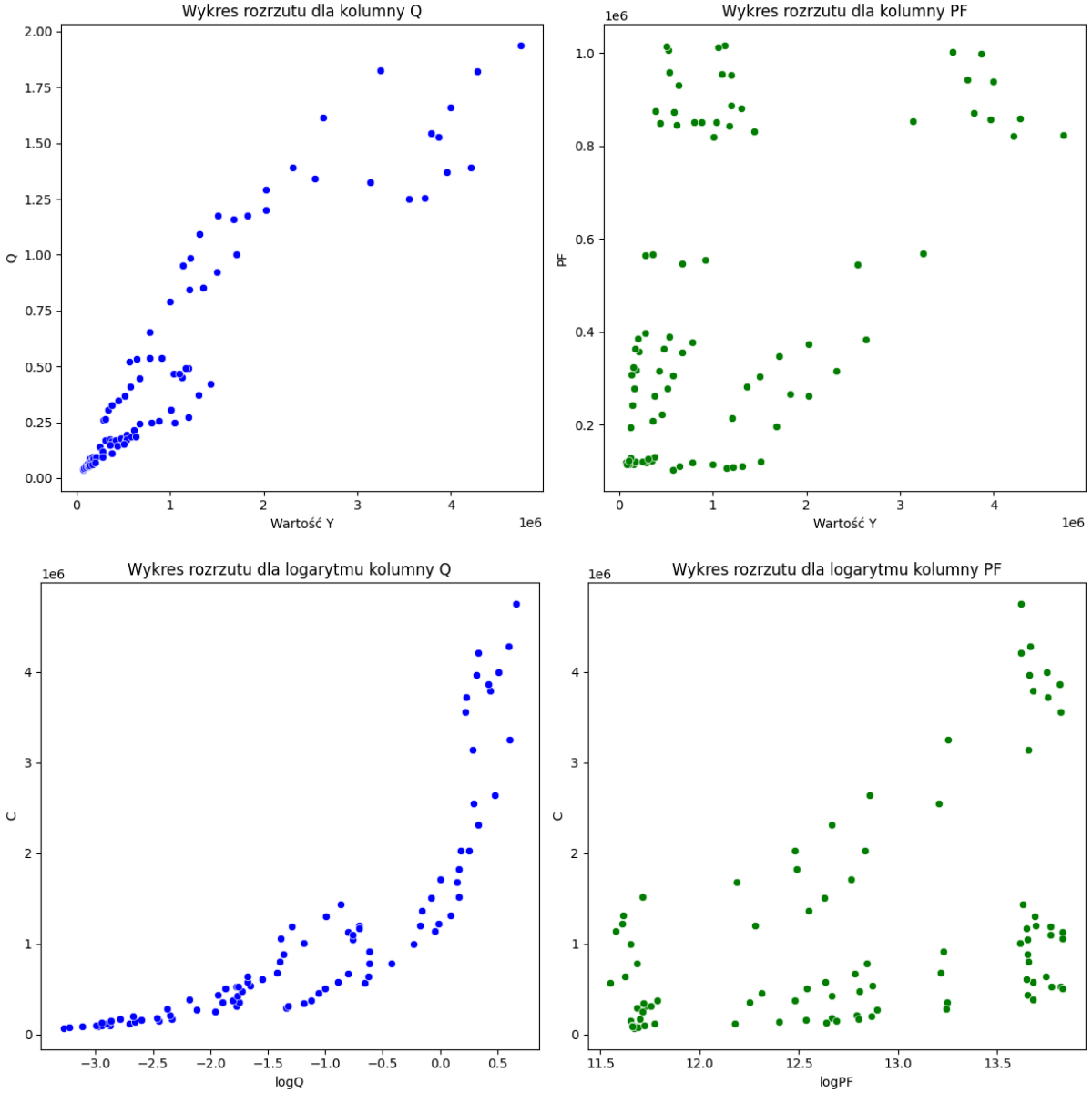



Na podstawie analizy macierzy korelacji między zmiennymi objaśniającymi oraz zmienną objaśniającą istnieje bardzo silna korelacja między całkowitym kosztem a wydajnością, wynosząca 0.93. Oznacza to, że linie lotnicze o wyższych kosztach mają tendencję do osiągania wyższej wydajności. Może to być zaskakujące, ponieważ można by się spodziewać, że niższe koszty przekładają się na lepszą wydajność, ale ta korelacja sugeruje odwrotność w analizowanym zestawie danych. Istnieje umiarkowana korelacja między całkowitym kosztem a ceną paliwa, wynosząca 0.48. Oznacza to, że linie lotnicze o wyższych kosztach są bardziej skłonne do płacenia wyższych cen za paliwo. Istnieje umiarkowana korelacja między całkowitym kosztem a wskaźnikiem wykorzystania floty, wynosząca 0.41. Oznacza to, że linie lotnicze o wyższych kosztach mają tendencję do wyższego wykorzystania swojej floty. Istnieje słaba korelacja między wydajnością a ceną paliwa, wynosząca 0.23. Oznacza to, że linie lotnicze o wyższej wydajności niekoniecznie płacą wyższe ceny za paliwo. Istnieje umiarkowana korelacja między wydajnością a wskaźnikiem wykorzystania floty, wynosząca 0.42. Oznacza to, że linie lotnicze o wyższej wydajności mają tendencję do wyższego wykorzystania swojej floty. Istnieje umiarkowana pozytywna korelacja między ceną paliwa a wskaźnikiem wykorzystania floty, wynosząca 0.49. Oznacza to, że linie lotnicze płacą wyższe ceny za paliwo mają tendencję do wyższego wykorzystania swojej floty.




Istnieje słaba korelacja między wydajnością (Q) a ceną paliwa (PF), wynosząca 0.23, co sugeruje, że linie lotnicze o wyższej wydajności niekoniecznie płacą wyższe ceny za paliwo. Umiarkowana korelacja między wydajnością a wskaźnikiem wykorzystania floty (LF), wynosząca 0.42, wskazuje, że linie lotnicze o wyższej wydajności mają tendencję do wyższego wykorzystania swojej floty. Istnieje również umiarkowana korelacja między ceną paliwa a wskaźnikiem wykorzystania floty, wynosząca 0.49, co sugeruje, że linie lotnicze płacą wyższe ceny za paliwo starają się efektywniej wykorzystać dostępną flotę. Analiza korelacji wskazuje na pewne zależności między analizowanymi zmiennymi, które mogą być istotne dla zrozumienia dynamiki branży lotniczej.

Na podstawie wartości z poszczególnych kolumn zastanawialiśmy się, czy nie powinniśmy zlogarytmować kolumn Q oraz PF. W celu weryfikacji tego pomysłu, stworzyliśmy wykresy rozrzutu dla tych kolumn w wersji oryginalnej oraz zlogarytmowanej, a następnie stworzyliśmy model regresji łącznej dla obu przypadków.




Wnioski wyciągnięte z przeprowadzonej analizy znajdują się w dalszej części sprawozdania. Na tym etapie możemy jednak określić, na podstawie kształtu wykresów, że logarytmowanie kolumny Q może być dobrym pomysłem.

# 2. Wybór postaci modelu

## Model Regresji Łącznej

W celu analizy wcześniej postawionej tezy stworzyliśmy modele dla oryginalnych kolumn oraz utworzonych wcześniej logarytmów wybranych kolumn.

Model dla podstawowych danych :

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | C | R-squared: | 0.946 |
| Model: | OLS | Adj. R-squared: | 0.944 |
| Method: | Least Squares | F-statistic: | 503.1 |
| Date: | Wed, 24 Jan 2024 | Prob (F-statistic): | 2.10e-54 |
| Time: | 17:01:02 | Log-Likelihood: | -1255.0 |
| No. Observations: | 90 | AIC: | 2518. |
| Df Residuals: | 86 | BIC: | 2528. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.159e+06 | 3.61e+05 | 3.213 | 0.002 | 4.42e+05 | 1.88e+06 |
| PF | 1.2253 | 0.104 | 11.814 | 0.000 | 1.019 | 1.432 |
| LF | -3.066e+06 | 6.96e+05 | -4.403 | 0.000 | -4.45e+06 | -1.68e+06 |
| Q | 2.026e+06 | 6.18e+04 | 32.781 | 0.000 | 1.9e+06 | 2.15e+06 |

| | | | |
|---|---|---|---|
| Omnibus: | 0.754 | Durbin-Watson: | 0.434 |
| Prob(Omnibus): | 0.686 | Jarque-Bera (JB): | 0.766 |
| Skew: | 0.211 | Prob(JB): | 0.682 |
| Kurtosis: | 2.839 | Cond. No. | 1.51e+07 |

Model dla tych gdzie PF oraz Q są w logarytmach

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | C | R-squared: | 0.747 |
| Model: | OLS | Adj. R-squared: | 0.738 |
| Method: | Least Squares | F-statistic: | 84.67 |
| Date: | Wed, 24 Jan 2024 | Prob (F-statistic): | 1.38e-25 |
| Time: | 17:44:56 | Log-Likelihood: | -1324.6 |
| No. Observations: | 90 | AIC: | 2657. |
| Df Residuals: | 86 | BIC: | 2667. |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -1.979e+06 | 1.12e+06 | -1.764 | 0.081 | -4.21e+06 | 2.52e+05 |
| logPF | 5.266e+05 | 9.94e+04 | 5.299 | 0.000 | 3.29e+05 | 7.24e+05 |
| LF | -4.695e+06 | 1.69e+06 | -2.778 | 0.007 | -8.05e+06 | -1.34e+06 |
| logQ | 8.447e+05 | 6.49e+04 | 13.021 | 0.000 | 7.16e+05 | 9.74e+05 |

| | | | |
|---|---|---|---|
| Omnibus: | 12.093 | Durbin-Watson: | 0.440 |
| Prob(Omnibus): | 0.002 | Jarque-Bera (JB): | 12.951 |
| Skew: | 0.914 | Prob(JB): | 0.00154 |
| Kurtosis: | 3.340 | Cond. No. | 347. |

Jak możemy zauważyć współczynnik determinacji wyszedł zdecydowanie lepszy dla modelu z oryginalnymi danymi. Test Jarque-Bera ponownie wskazuje na to, że pierwszy model jest lepszy, gdyż w przypadku drugiego modelu nasze reszty nie tworzą rozkładu normalnego, a w pierwszym modelu ich rozkład jest znacznie bardziej zbliżony do rozkładu normalnego. Z tego powodu odrzucamy wcześniejszy pomysł o stosowaniu logarytmów w modelach. Trzeba jednak pamiętać, że testy te wykazały autokorelację w resztach. Przechodzimy do analizy pierwszego modelu regresji łącznej

**Interpretacja Współczynników Regresji:**

**Wszystkie** współczynniki są bardzo istotne statystycznie i posiadają p-value bliskie 0.

- const: Wartość : 1,158559e+06 czyli 1 158 559 Jest to współczynnik dla stałej.

- PF: Wartość : 1,2253

Dla każdej jednostki wzrostu w cenach paliwa, oczekujemy wzrostu kosztów o 1.2253 jednostki.

- LF: Wartość : -3.066e+06 czyli - 3 066 000

Dla każdej jednostki wzrostu w wskaźniku wykorzystania floty, oczekujemy spadku kosztów o -3.066e+06 jednostek.

- Q: Wartość: 2.026e+06 czyli 2 026 000

Dla każdej jednostki wzrostu w wydajności (mila pasażerska), oczekujemy wzrostu kosztów o 2.026e+06 jednostek.

**Miary Dopasowania Modelu:**

- R-squared: Wartość: 0,946

Około 94,6% zmienności w zmiennej objaśnianej (`C`) jest wyjaśniane przez model.

- Adj. R-squared: Wartość: 0,944

Skorygowany R-kwadrat uwzględnia liczbę zmiennych niezależnych i obserwacji.

- F-statistic: Wartość: 503,1 , p-value bliskie 0

Test ogólnej istotności modelu. Wartość statystyki F wynosząca 503,1 przy p-value 2.10e-54 jest istotna statystycznie, wskazuje na to że co najmniej jedna z niezależnych zmiennych w modelu ma istotny wpływ na zmienną zależną. Odrzucamy hipotezę zerową mówiącą o tym, że wszystkie współczynniki regresji są równe zeru.

**Testy Autokorelacji i Homoskedastyczności:**

Analiza wyników testów autokorelacji (Durbin-Watson) i homoskedastyczności (Jarque-Bera) dostarcza informacji o dwóch ważnych aspektach modelu regresji.

Test autokorelacji (Durbin-Watson):

- Statystyka Durbina-Watsona (DW) wynosi 0.434.

- Zakres wartości statystyki Durbin-Watson to 0-4, gdzie wartości bliskie 2 sugerują brak autokorelacji (korelacji między kolejnymi obserwacjami).

- Wartość DW poniżej 2 sugeruje obecność autokorelacji pozytywnej, a wartość powyżej 2 sugeruje obecność autokorelacji negatywnej.

W tym przypadku wartość DW jest bardzo niska (0.434), co sugeruje obecność silnej autokorelacji dodatniej w resztach modelu. Taka sytuacja może zniekształcać wyniki estymacji modelu i sugeruje, że model może nie być odpowiedni.

Test homoskedastyczności (Jarque-Bera):

- Statystyka Jarque-Bera wynosi 0.766, a p-value to 0.682.

- Test Jarque-Bera ocenia, czy residua są homoskedastyczne, czyli czy wariancja błędów jest stała.

Wartość p-value wynosząca 0.682 jest duża, co sugeruje, że nie ma podstaw do odrzucenia hipotezy zerowej o homoskedastyczności. To oznacza, że błędy nie wykazują znaczących odchyleń od homoskedastyczności.

Wnioski:

- Model wykazuje problem z autokorelacją, co może sugerować, że pewne wzorce czasowe w danych nie zostały uwzględnione lub że inne zmienne wpływające na zmienną zależną mogą być pominięte.

- Homoskedastyczność, na podstawie testu Jarque-Bera, nie stanowi problemu.

W przypadku problemu z autokorelacją, może być konieczne dostosowanie modelu lub uwzględnienie dodatkowych zmiennych w celu poprawy jego trafności.

# Model z efektami losowymi

```
Model 6: Estymacja Losowe efekty (GLS), z wykorzystaniem 90 obserwacji
Włączono 6 jednostek danych przekrojowych
Szereg czasowy długości = 15
Zmienna zależna (Y): C

             współczynnik   błąd standardowy     z      wartość p
  ----------------------------------------------------------------
  const       1,07429e+06    377468             2,846   0,0044    ***
  PF          1,12359             0,103441     10,86    1,75e-027 ***
  LF         -3,08499e+06    725680            -4,251   2,13e-05  ***
  Q           2,28859e+06    109494            20,90    5,18e-097 ***

Średn.aryt.zm.zależnej   1122524   Odch.stand.zm.zależnej   1192075
Suma kwadratów reszt    8,47e+12   Błąd standardowy reszt  311936,9
Logarytm wiarygodności -1264,729   Kryt. inform. Akaike'a  2537,458
Kryt. bayes. Schwarza   2547,457   Kryt. Hannana-Quinna    2541,490
Autokorel.reszt - rho1  0,674675   Stat. Durbina-Watsona   0,640462

'Between' wariancji = 1,15372e+010
'Within' wariancji = 4,42777e+010
theta wykorzystuje quasi-demeaning = 0,548635
corr(y,yhat)^2 = 0,943221

Joint test on named regressors -
  Asymptotyczna statystyka testu: Chi-kwadrat(3) = 883,501
  z wartością p = 3,35609e-191

Test Breuscha-Pagana na -
  Hipoteza zerowa: Wariancja błędu w jednostce = 0
  Asymptotyczna statystyka testu: Chi-kwadrat(1) = 0,613087
  z wartością p = 0,433628

Test Hausmana -
  Hipoteza zerowa: Estymator UMNK (GLS) jest zgodny
  Asymptotyczna statystyka testu: Chi-kwadrat(3) = 53,8045
  z wartością p = 1,23513e-11
```

**Interpretacja współczynników:**

Wszystkie wybrane zmienne (PF, LF,Q) są bardzo istotne statystycznie. Współczynnik dla zmiennej PF (Cena paliwa w tysiącach $) wynosi 1,12359. Oznacza to, że dla jednostkowego wzrostu ceny paliwa, oczekiwana wartość kosztu rośnie o tę wartość. Współczynnik dla zmiennej LF (Wskaźnik wykorzystania floty) wynosi -3,08499e+06. Oznacza to, że dla jednostkowego wzrostu wskaźnika wykorzystania floty, oczekiwana wartość kosztu maleje o tę wartość. Współczynnik dla zmiennej Q (Wydajność, w milach pasażerskich, liczba indeksowa) wynosi 2,28859e+06. Oznacza to, że dla jednostkowego wzrostu wydajności, oczekiwana wartość kosztu rośnie o tę wartość.

**Błędy modelu:**

Odchylenie standardowe zmiennej zależnej wynosi 1192075. Wartość odchylenia standardowego na tak wysokim poziomie sugeruje, że dane w tej zmiennej mają tendencję do rozrzutu na znaczną odległość od średniej. Oznacza to, że punkty danych są rozproszone wokół średniej z dużą zmiennością.
Błąd standardowy reszt wynosi 311936,9. Im niższa wartość błędu standardowego reszt, tym lepiej model dopasowuje się do danych. Oznacza to, że różnice między rzeczywistymi a przewidywanymi wartościami są na ogół mniejsze.

**Test Autokorelacji:**

Statystyka Durbina-Watsona wynosi 0,64 co sugeruje obecność silnej autokorelacji dodatniej w resztach modelu. Taka sytuacja może zniekształcać wyniki estymacji modelu i sugeruje, że model może nie być odpowiedni.

# Model z efektami ustalonymi jednokierunkowy

```
Model 8: Estymacja Ustalone efekty, z wykorzystaniem 90 obserwacji
Włączono 6 jednostek danych przekrojowych
Szereg czasowy długości = 15
Zmienna zależna (Y): C

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const      1,07730e+06   310799              3,466     0,0008    ***
  PF         0,773071           0,0973190      7,944     9,70e-012 ***
  LF        -3,79737e+06   613773            -6,187     2,37e-08  ***
  Q          3,31902e+06   171354             19,37     3,86e-032 ***

Średn.aryt.zm.zależnej   1122524   Odch.stand.zm.zależnej   1192075
Suma kwadratów reszt    3,59e+12   Błąd standardowy reszt  210422,8
LSDV R-kwadrat          0,971642   Within R-kwadrat        0,929366
LSDV F(8, 81)           346,9188   Wartość p dla testu F   2,53e-59
Logarytm wiarygodności -1226,082   Kryt. inform. Akaike'a  2470,164
Kryt. bayes. Schwarza   2492,662   Kryt. Hannana-Quinna    2479,236
Autokorel.reszt - rho1  0,674675   Stat. Durbina-Watsona   0,640462

Joint test on named regressors -
  Statystyka testu: F(3, 81) = 355,254
  z wartością p = P(F(3, 81) > 355,254) = 1,69684e-46

Test na zróżnicowanie wyrazu wolnego w grupach -
  Hipoteza zerowa: grupy posiadają wspólny wyraz wolny
  Statystyka testu: F(5, 81) = 14,5952
  z wartością p = P(F(5, 81) > 14,5952) = 3,46743e-10
```

**Interpretacja współczynników:**

**Wszystkie** współczynniki są bardzo istotnie statystycznie i posiadają p-value bliskie zero.

- const Wartość: 1,07730e+06 , jeśli wszystkie zmienne objaśniające będą równe 0, oczekuje się że zmienna zależna będzie wynosić 1,07730e+06

- PF Wartość: 0,773071
  Interpretacja: Dla jednostkowego wzrostu zmiennej niezależnej PF, oczekujemy wzrostu zmiennej zależnej o 0,773071 jednostki, przy założeniu, że pozostałe zmienne pozostają stałe.

- LF Wartość: -3,79737e+06
  Interpretacja: Dla jednostkowego wzrostu zmiennej niezależnej LF, oczekujemy spadku zmiennej zależnej o -3,79737e+06 jednostek, przy założeniu, że pozostałe zmienne pozostają stałe.

- Q Wartość: 3,31902e+06
  Interpretacja: Dla jednostkowego wzrostu zmiennej niezależnej Q, oczekujemy wzrostu zmiennej zależnej o 3,31902e+06 jednostek, przy założeniu, że pozostałe zmienne pozostają stałe.

**Miary dopasowania**

R-kwadrat w tym modelu wynosi w przybliżeniu 0,97 co oznacza, że około 97% zmienności zmiennej zależnej jest wyjaśniane przez nasz model.

R-kwadrat w wariancji wewnątrzgrupowej wynosi 0,92

**Błędy modelu**

Odchylenie standardowe zmiennej zależnej wynosi 1192075

Błąd standardowy reszt wynosi 210422,8

**Testy**

Statystyka DW wynosi 0,640462 co sugeruje obecność silnej autokorelacji dodatniej w resztach modelu. Mamy tutaj zatem sytuację podobną do modelu z efektami losowymi.

# Model z efektami ustalonymi dwukierunkowy

```
Model 9: Estymacja Ustalone efekty, z wykorzystaniem 90 obserwacji
Włączono 6 jednostek danych przekrojowych
Szereg czasowy długości = 15
Zmienna zależna (Y): C

              współczynnik     błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const       525466             607872              0,8644    0,3904
  PF               1,51505            2,37734        0,6373    0,5261
  LF              -2,92656e+06        1,07580e+06   -2,720     0,0083    ***
  Q                3,41606e+06  203851              16,76      1,18e-025 ***
  dt_2          14184,5           129022              0,1099    0,9128
  dt_3          -1844,24          138494             -0,01332   0,9894
  dt_4         -80091,6           142710             -0,5612    0,5765
  dt_5         -50350,2           291449             -0,1728    0,8634
  dt_6        -129266             428106             -0,3019    0,7636
  dt_7        -210664             488980             -0,4308    0,6680
  dt_8        -279957             594675             -0,4708    0,6393
  dt_9        -349468             687311             -0,5085    0,6128
  dt_10       -595606                  1,08827e+06   -0,5473    0,5860
  dt_11       -711163                  1,78655e+06   -0,3981    0,6918
  dt_12       -729430                  2,14110e+06   -0,3407    0,7344
  dt_13       -685593                  1,99425e+06   -0,3438    0,7321
  dt_14       -677961                  1,81114e+06   -0,3743    0,7093
  dt_15       -599055                  1,72299e+06   -0,3477    0,7292

Średn.aryt.zm.zależnej   1122524   Odch.stand.zm.zależnej   1192075
Suma kwadratów reszt     3,31e+12  Błąd standardowy reszt   222354,9
LSDV R-kwadrat           0,973808  Within R-kwadrat         0,934761
LSDV F(22, 67)           113,2281  Wartość p dla testu F    1,72e-44
Logarytm wiarygodności -1222,507   Kryt. inform. Akaike'a   2491,014
Kryt. bayes. Schwarza    2548,509  Kryt. Hannana-Quinna     2514,199
Autokorel.reszt - rho1   0,675445  Stat. Durbina-Watsona    0,638508
```

```
Joint test on named regressors -
  Statystyka testu: F(3, 67) = 97,5698
  z wartością p = P(F(3, 67) > 97,5698) = 2,11638e-24

Test na zróżnicowanie wyrazu wolnego w grupach -
  Hipoteza zerowa: grupy posiadają wspólny wyraz wolny
  Statystyka testu: F(5, 67) = 11,4613
  z wartością p = P(F(5, 67) > 11,4613) = 5,16597e-08

Test Walda na łączną istotność zmiennych 0-1 jednostek czasu -
  Hipoteza zerowa: No time effects
  Asymptotyczna statystyka testu: Chi-kwadrat(14) = 5,53993
  z wartością p = 0,976791
```

**Interpretacja współczynników:**

Na podstawie stworzonego modelu możemy wywnioskować że jedynie zmienne Współczynnik dla zmiennej PF wynosi 1,51505. Oznacza to, że dla jednostkowego wzrostu ceny paliwa, oczekiwana wartość zmiennej zależnej rośnie o tę wartość. Współczynnik dla zmiennej LF wynosi -2,92656e+06. Współczynnik dla zmiennej Q wynosi 3,41606e+06.
Zmienne dt_2 do dt_15 są to współczynniki dla zmiennych czasowych od 2 do 15. Każda z tych zmiennych reprezentuje rok od 2 do 15. Wartości tych współczynników wskazują na wpływ każdego kolejnego roku na zmienną zależną.

**Miary dopasowania modelu:**

R-kwadrat dla LSDV (Least Squares Dummy Variable) wynosi 0,9738, co oznacza, że model wyjaśnia 97,38% zmienności zależnej zmiennej. Analizując wartości testu F możemy wywnioskować że przynajmniej jedna zmienna jest istotna.

**Błędy modelu**

Błąd standardowy reszt wynosi 222354,9.
Odchylenie standardowe zależnej zmiennej wynosi 1192075

**Testy**

Statystyka DB jest na poziomie 0,63 co podobnie jak w poprzednich model wskazuje na wysokie prawdopodobieństwo występowania autokorelacji w resztach modelu. Test F na współczynnik jest istotny co oznacza że co najmniej jedna zmienna niezależna różni się istotnie od innych w grupach.
Kryterium informacyjne Akaike'a oraz kryterium Hannana-Quinna mają bardzo wysokie wartość. Obydwa kryteria są miarami oceny modeli, gdzie niższe wartości wskazują na lepsze dopasowanie do danych. Wysokie wartości tych kryteriów może być problematyczne ze względu na duże prawdopodobieństwo że model gorzej tłumaczy zmienność w danych.

# 3. Wybór modelu

Wybór najlepszej postaci modelu jest kluczowym elementem analizy. Postanowiono wykonać test porównujący model regresji łącznej z jednokierunkowym modelem efektów ustalonych. W tym celu przeprowadzono test na zróżnicowanie wyrazu wolnego w grupach.

**Test na zróżnicowanie wyrazu wolnego w grupach**

```
Test na zróżnicowanie wyrazu wolnego w grupach -
  Hipoteza zerowa: grupy posiadają wspólny wyraz wolny
  Statystyka testu: F(5, 67) = 11,4613
  z wartością p = P(F(5, 67) > 11,4613) = 5,16597e-08
```

```
Test na zróżnicowanie wyrazu wolnego w grupach -
  Hipoteza zerowa: grupy posiadają wspólny wyraz wolny
  Statystyka testu: F(5, 81) = 14,5952
  z wartością p = P(F(5, 81) > 14,5952) = 3,46743e-10
```

H0: Nie ma istotnej różnicy w zróżnicowaniu wyrazu wolnego między grupami, gdy stosowany jest model regresji łącznej w porównaniu do jednokierunkowego modelu efektów ustalonych (alfy są sobie równe i są stałe, lepszy jest model regresji łącznej).

H1: Istnieje istotna różnica w zróżnicowaniu wyrazu wolnego między grupami (jedna z alf jest różna, lepszy jest model efektów ustalonych).

Wartość statystyki F dla o odpowiednio 5 oraz 81 stopniach swobody wyniosła 14,5952. Wartość p-value wyniosła 3,46e-10. Na poziomie istotności 0,05 możemy odrzucić hipotezę zerową i przyjąć że dla wybranego zbioru danych lepszy jest **model efektów ustalonych.**

**Test Hausmana**

```
Test Hausmana -
  Hipoteza zerowa: Estymator UMNK (GLS) jest zgodny
  Asymptotyczna statystyka testu: Chi-kwadrat(3) = 53,8045
  z wartością p = 1,23513e-11
```

H0: Preferowanym modelem jest model efektów losowych.
H1: Preferowanym modelem jest model efektów ustalonych.

Asymptotyczna statystyka testy Chi-kwadrat o trzech stopniach swobody wynosi 53,8 z wartością p-value równą 1,23513e-11. Wartość p-value jest bardzo mała i na poziomie istotność równym 0,05 możemy odrzucić hipotezę zerową i przyjąć że lepszym modelem niż model efektów losowych jest **model efektów ustalonych**.

**Test Breuscha-Pagana**

```
Test Breuscha-Pagana na -
  Hipoteza zerowa: Wariancja błędu w jednostce = 0
  Asymptotyczna statystyka testu: Chi-kwadrat(1) = 0,613087
  z wartością p = 0,433628
```

H0: Preferowanym modelem jest model regresji łącznej
H1: Preferowanym modelem jest model efektów losowych

Asymptotyczna statystyka testu Chi-kwadrat o jednym stopniu swobody wynosi 0,613. Wartość p-value wynosi 0,433 i jest większę od 0,05 dlatego na podstawie tego testu podjęta zostaje decyzja o nie odrzuceniu hipotezy zerowej to znaczy lepszy od modelu efektów ustalonych jest model **regresji łącznej.**

**Test Walda**

```
Test Walda na łączną istotność zmiennych 0-1 jednostek czasu -
  Hipoteza zerowa: No time effects
  Asymptotyczna statystyka testu: Chi-kwadrat(14) = 5,53993
  z wartością p = 0,976791
```

H0: Preferowanym modelem jest model jednokierunkowy.
H1: Preferowanym modelem jest model dwukierunkowy.

Asymptotyczna statystyka testy Chi-kwadrat o czternastu stopniach swobody wynosi 5,53993 z wartością p-value równą 0,97 . Wartość p-value jest bardzo wysokie co oznacza że nie ma powodu aby odrzucić hipotezę zerową więc przyjmujemy ją za prawdziwą. Preferowanym modelem jest **model jednokierunkowy**.

**Wnioski**

Analizując wyniki przeprowadzonych testów, możemy stwierdzić z prawdopodobieństwem 95%, że model efektów ustalonych jest lepszy od modelu regresji łącznej, który jest natomiast lepszy od modelu efektów losowych. Po przeanalizowaniu modelu efektów ustalonych dochodzimy do wniosku, że jego preferowaną formą jest model jednokierunkowy. Możemy zatem stwierdzić, że w tym przypadku najlepiej radzi sobie **jednokierunkowy model efektów ustalonych.**

# 4. Ostateczny dobór zmiennych.

W wybranym modelu jednokierunkowym efektów ustalonych wszystkie zmienne są istotne dlatego zdecydowano się aby w kolejnym etapie analizy pozostawić wszystkie zmienne.

# 5. Weryfikacja modelu

## Testowanie autokorelacji składnika losowego.

**Test Durbina-Watsona**

```
Stat. Durbina-Watsona = 0,640462

Hl: dodatnia autokorelacja składnika losowego
   wartość p = 3,49609e-013
Hl: ujemna autokorelacja składnika losowego
   wartość p = 1
```

H0: Brak autokorelacji reszt pierwszego rzędu w danych.
H1: Występuje autokorelacji reszt pierwszego rzędu w danych

Autokorelacja oznacza, że błędy modelu są ze sobą skorelowane, co może wpływać na niepoprawność statystyk testowych i wyników modelu. Statystyka Durbin-Watsona przyjmuje wartości między 0 a 4, gdzie wartość bliska 2 sugeruje brak autokorelacji natomiast wartość bliska 0 lub 4 sugeruje silną autokorelację. W naszym przypadku statystyka Durbin-Watsona wynosi 0,640462, co jest znacznie mniejsze niż 2. Oznacza to, że istnieje dodatnia autokorelacja składnika losowego w resztach modelu jednokierunkowego z efektami ustalonymi.
Dla H1: Dodatnia autokorelacja składnika losowego, wartość p = 3,49609e-013. Wartość p bardzo bliska zeru, co sugeruje silne dowody na odrzucenie hipotezy zerowej na rzecz hipotezy alternatywnej. Istnieje istotna dodatnia autokorelacja składnika losowego.
Dla H1: Ujemna autokorelacja składnika losowego, wartość p = 1. Wartość p wynosi 1, co sugeruje brak istotności statystycznej dla tej hipotezy alternatywnej. Nie ma istotnej ujemnej autokorelacji składnika losowego.
Podsumowując, wyniki wskazują na **obecność dodatniej autokorelacji składnika losowego** w resztach modelu jednokierunkowego z efektami ustalonymi.

**Test Wooldridge'a**

```
First differenced equation (dependent, d_y:

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------
  d_Q         1,59111e+06   250068              6,363     0,0014   ***
  d_PF        0,909289           0,325527       2,793     0,0383   **
  d_LF       -1,46188e+06   670965             -2,179     0,0812   *

  n = 84, R-squared = 0,5838

Autoregresyjne reszty (zm. zależna, reszty):

            współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------
  uhat(-1)     0,509873        0,166931        3,054      0,0283   **

  n = 78, R-squared = 0,2265

Test Wooldridge na autokorelację dla danych panelowych -
  Hipoteza zerowa: Brak autokorelacji rzędu pierwszego (rho = -.05)
  Statystyka testu: F(1, 5) = 36,5984
  z wartością p = P(F(1, 5) > 36,5984) = 0,00177923
```

H0: Brak autokorelacji w danych.
H1: Występuje autokorelacji w danych

Statystyka Wooldridge'a jest statystyką F o stopniach swobody równych 1 i 5. P-value jest równe 0,00177923, a zatem jest niższa od 0,05. Mamy zatem podstawy do odrzucenia hipotezy zerowej, a zatem w naszych danych występuje autokorelacja.

**Próba eliminacji autokorelacji**

W celu usunięcia autokorelacji przeprowadziliśmy przekształcenie Princa-Winsona.
Oto nowo powstały model :

```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                      C   R-squared:                        0.9294
Estimator:                   PanelOLS   R-squared (Between):              0.4449
No. Observations:                  90   R-squared (Within):               0.9294
Date:                Thu, Jan 25 2024   R-squared (Overall):              0.6394
Time:                        16:06:04   Log-likelihood                   -1226.1
Cov. Estimator:       Driscoll-Kraay

                                        F-statistic:                      355.25
Entities:                           6   P-value                           0.0000
Avg Obs:                       15.000   Distribution:                    F(3,81)
Min Obs:                       15.000
Max Obs:                       15.000   F-statistic (robust):             1798.5
                                        P-value                           0.0000
Time periods:                      15   Distribution:                    F(3,81)
Avg Obs:                       6.0000
Min Obs:                       6.0000
Max Obs:                       6.0000

                            Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
const       1.077e+06  3.923e+05     2.7462     0.0074   2.968e+05   1.858e+06
PF             0.7731     0.1331     5.8072     0.0000      0.5082      1.0379
LF         -3.797e+06  8.478e+05    -4.4789     0.0000  -5.484e+06   -2.11e+06
Q           3.319e+06  4.068e+05     8.1592     0.0000    2.51e+06   4.128e+06
================================================================================

F-test for Poolability: 14.595
P-value: 0.0000
Distribution: F(5,81)

Included effects: Entity
```

Przeprowadzamy teraz ponownie test Durbina-Watsona

```
Test Durbina-Watsona:
Durbin-Watson Statistic: 0.6932877606358621
```

Tym razem statystyka tego testu wynosi w przybliżeniu 0,69. Jest to wynik nieznacznie lepszy od wcześniejszego, chociaż w dalszym stopniu wynik statystyki jest daleki od 2, niestety nie udało nam się jeszcze bardziej zniwelować autokorelacji.

# Badanie heteroskedastyczności składnika losowego.

**Test Walda**

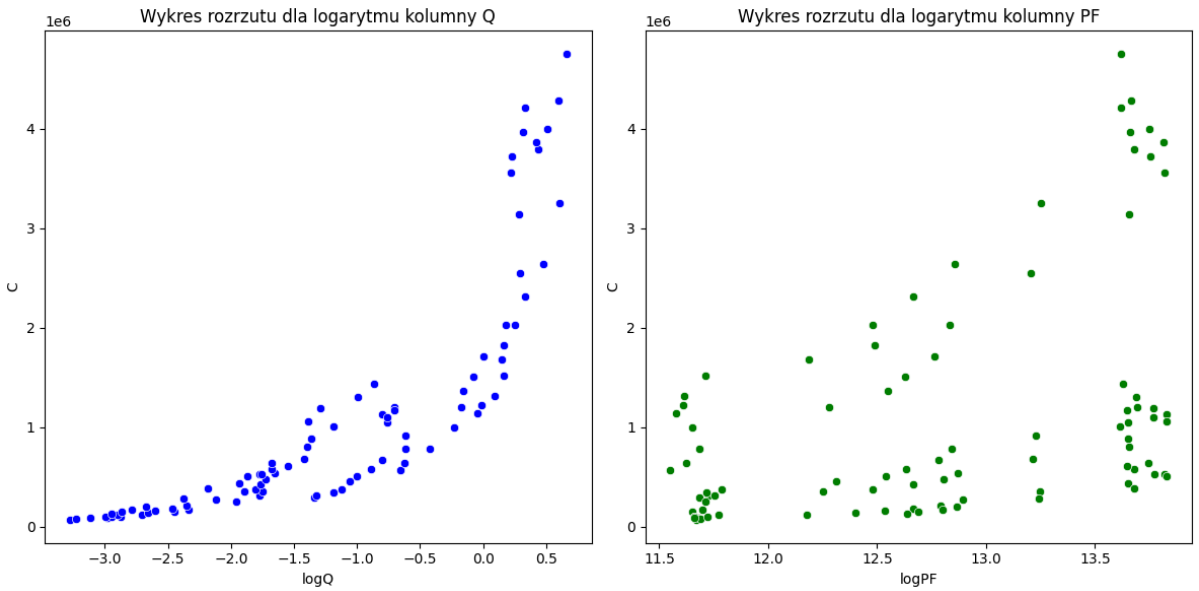
```
Dystrybuanta testu Walda na heteroskedastyczność:
 Chi-kwadrat(6) = 47,0078, z wartością p = 1,86445e-008

_Panelowa wariancja resztowa = 3,985e+010

 unit     variance
    18,24195e+010 (T = 15)
    26,53630e+010 (T = 15)
    32,55006e+010 (T = 15)
    41,95998e+010 (T = 15)
    52,12337e+010 (T = 15)
    62,49833e+010 (T = 15)
```

H0: Brak heteroskedastyczności

H1: Wariancja błędów nie jest stała dla co najmniej jednej zmiennej objaśniającej, co oznacza obecność heteroskedastyczności.

W przypadku tego testu statystyka chi-kwadrat ma 6 stopni swobody, a p-value jest liczbą bliską 0. Możemy więc odrzucić hipotezę zerową o stałej wariancji błędów. Oznacza to występowanie heteroskedastyczności.

**Próba eliminacji heteroskedastyczności**



Na podstawie wykresu, który wcześniej pojawił się w pierwszej części sprawozdania, podejmujemy ponowną próbę logarytmowania kolumn. Tym razem wybieramy jedynie kolumnę Q i przekształcamy ją na logarytm jej wartości.

```
Dystrybuanta testu Walda na heteroskedastyczność:
 Chi-kwadrat(6) = 56,4316, z wartością p = 2,38088e-010

_Panelowa wariancja resztowa = 2,20929e+011

 unit     variance
    13,87834e+011  (T = 15)
    24,93763e+011  (T = 15)
    39,89703e+010  (T = 15)
    47,54333e+010  (T = 15)
    51,47721e+011  (T = 15)
    61,21849e+011  (T = 15)
```

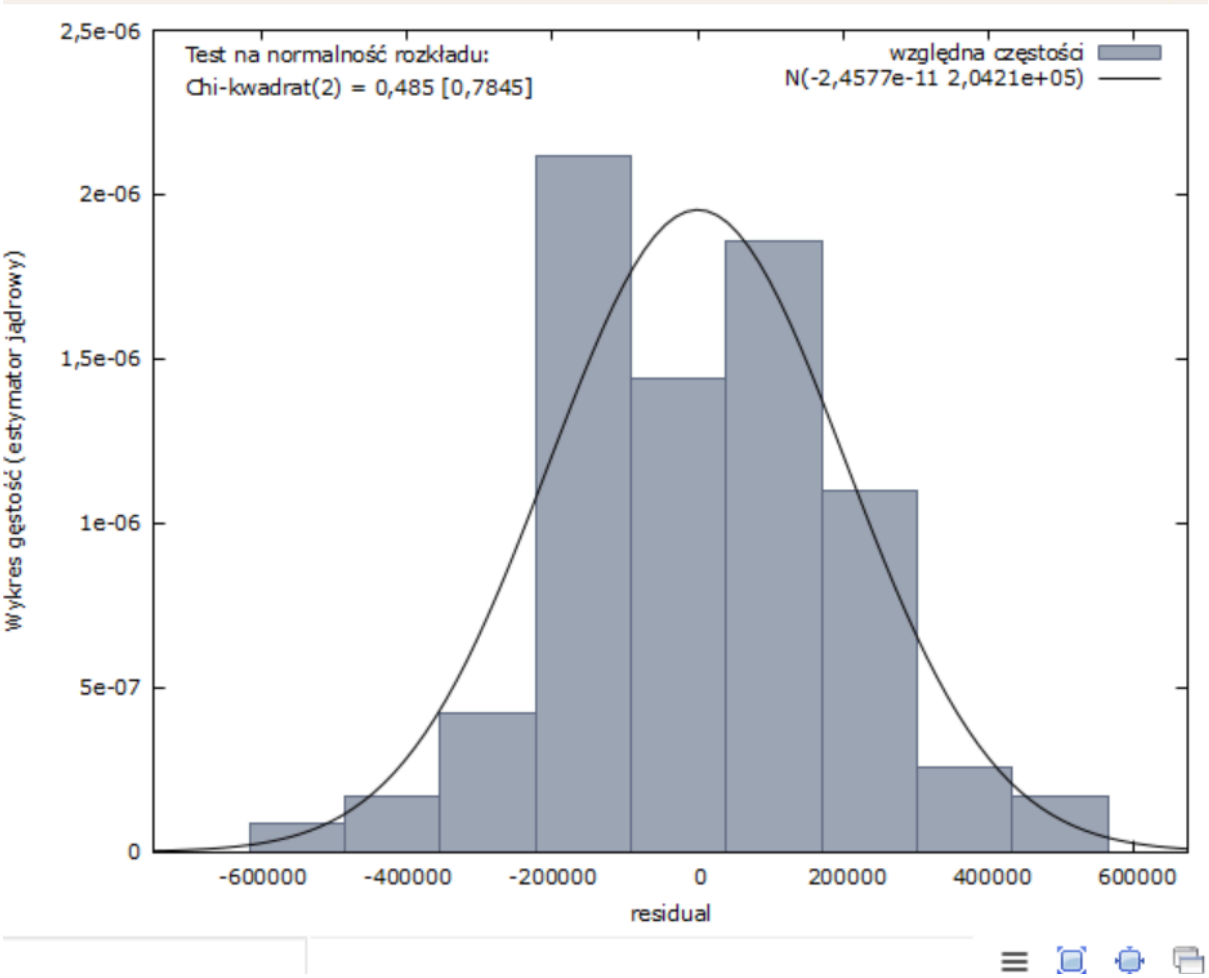
Jak widzimy wyniki się pogorszyły, dlatego zostajemy przy zmiennych bez logarytmów.

## Badanie normalności rozkładu reszt.

W przypadku występowania autokorelacji składnika losowego lub heteroskedastyczności należy podjąć próbę eliminacji tego problemu.

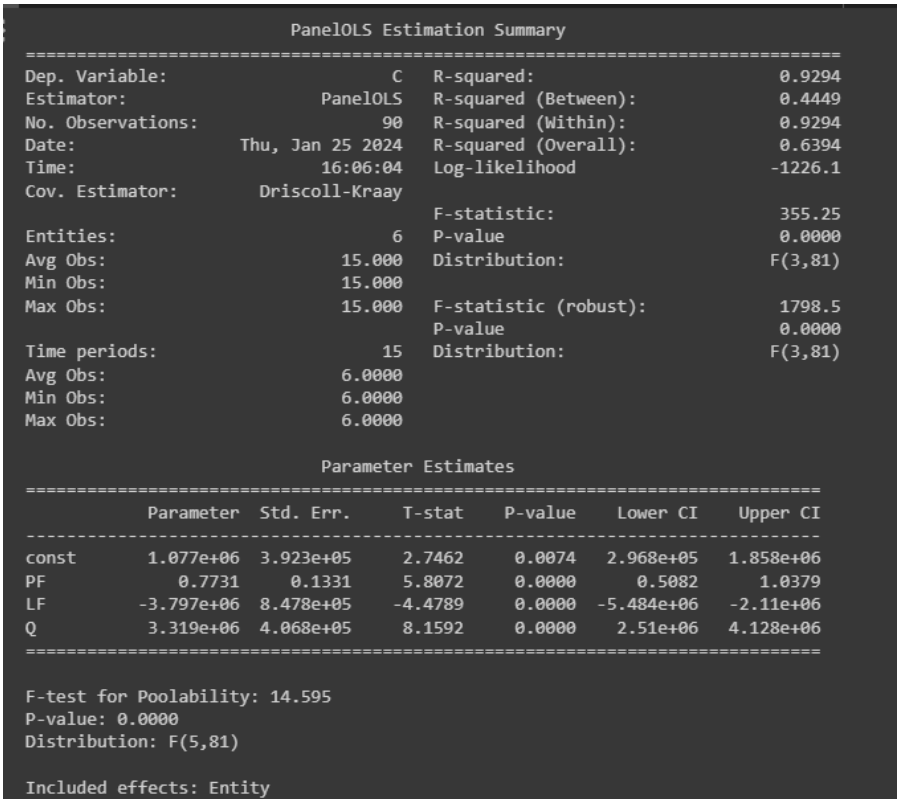
```
Rozkład częstości dla residual, obserwacje 1-90
liczba przedziałów = 9, średnia = -2,45766e-011, odch.std. = 204214

     Przedziały       średnia   liczba   częstość  skumulowana

          < -4,861e+005 -5,518e+005      1      1,11%    1,11%
-4,861e+005 - -3,547e+005 -4,204e+005      2      2,22%    3,33%
-3,547e+005 - -2,233e+005 -2,890e+005      5      5,56%    8,89% *
-2,233e+005 - -9,194e+004 -1,576e+005     25     27,78%   36,67% **********
-9,194e+004 -  3,945e+004 -2,624e+004     17     18,89%   55,56% ******
 3,945e+004 -  1,708e+005  1,051e+005     22     24,44%   80,00% ********
 1,708e+005 -  3,022e+005  2,365e+005     13     14,44%   94,44% *****
 3,022e+005 -  4,336e+005  3,679e+005      3      3,33%   97,78% *
          >=  4,336e+005  4,993e+005      2      2,22%  100,00%

Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.Test Doornika-Hansena (1994) - transformowana skośność i kurtoza.:
Chi-kwadrat(2) = 0,485 z wartością p 0,78454
```

Po przeprowadzeniu tego testu uzyskaliśmy p-value równe 0,78. Przy takim wyniku nie mamy statystycznych dowodów na odrzucenie hipotezy zerowej, przyjmujemy zatem, że dystrybuanta empiryczna posiada rozkład normalny.




Jak widzimy na histogramie, dane zostały podzielone na 9 przedziałów, na podstawie testu określiliśmy, że reszty mają rozkład normalny, jednak jak widzimy na załączonym obrazku, nie jest on idealny.

# 5. Prezentacja ostatecznej postaci modelu i jego ocena.



```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                      C   R-squared:                        0.9294
Estimator:                   PanelOLS   R-squared (Between):              0.4449
No. Observations:                  90   R-squared (Within):               0.9294
Date:                Thu, Jan 25 2024   R-squared (Overall):              0.6394
Time:                        16:06:04   Log-likelihood                   -1226.1
Cov. Estimator:       Driscoll-Kraay

                                        F-statistic:                      355.25
Entities:                           6   P-value                           0.0000
Avg Obs:                       15.000   Distribution:                   F(3,81)
Min Obs:                       15.000
Max Obs:                       15.000   F-statistic (robust):             1798.5
                                        P-value                           0.0000
Time periods:                      15   Distribution:                   F(3,81)
Avg Obs:                       6.0000
Min Obs:                       6.0000
Max Obs:                       6.0000

                            Parameter Estimates
================================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
--------------------------------------------------------------------------------
const       1.077e+06  3.923e+05     2.7462     0.0074   2.968e+05   1.858e+06
PF             0.7731     0.1331     5.8072     0.0000      0.5082      1.0379
LF         -3.797e+06  8.478e+05    -4.4789     0.0000  -5.484e+06   -2.11e+06
Q           3.319e+06  4.068e+05     8.1592     0.0000    2.51e+06   4.128e+06
================================================================================

F-test for Poolability: 14.595
P-value: 0.0000
Distribution: F(5,81)

Included effects: Entity
```

Ocena istotności zmiennych i współczynnika determinacji

W finalnie dobranym modelu wszystkie nasze zmienne są bardzo mocno istotne statystycznie, a model posiada R-kwadrat na poziomie 0,929 . Oznacza to, że aż około 93% zmienności zmiennej objaśnianej jest opisywana przez nasz model, jest to zadowalający wynik.

Interpretacja parametrów modelu może być utrudniona w przypadku dużych autokorelacji reszt w danych. Autokorelacja oznacza, że wartości reszt są skorelowane ze sobą w czasie, co może wprowadzać zakłócenia do estymacji parametrów modelu. W takich przypadkach, interpretowanie efektów poszczególnych zmiennych może być trudniejsze, ponieważ reszty wprowadzają dodatkową wariancję do modelu.

Parametry

Oto interpretacje parametrów naszego modelu:

1. Stała (const):

- Interpretacja: Gdy wszystkie inne zmienne niezależne są równe zeru, przewidywana wartość zmiennej zależnej wynosi 1.077e+06 (1 077 000).

2. PF:

- Interpretacja: Przy wzroście jednostkowym zmiennej niezależnej PF, przewidywana wartość zmiennej zależnej rośnie o 0.7731.

3. LF:

- Interpretacja: Przy wzroście jednostkowym zmiennej niezależnej LF, przewidywana wartość zmiennej zależnej maleje o 3.797e+06.

4. Q: - Interpretacja: Przy wzroście jednostkowym zmiennej niezależnej Q, przewidywana wartość zmiennej zależnej rośnie o 3.319e+06.

Każdy współczynnik ma swoją własną interpretację która ocenia jak zmiana danej zmiennej wpływa na przewidywaną wartość zmiennej zależnej, pod warunkiem, że wszystkie inne zmienne pozostają stałe. T-statistics i p-values pomagają ocenić istotność statystyczną każdego współczynnika. Interval ufności dostarcza zakresu możliwych wartości dla każdego współczynnika.

# Podsumowanie

W analizie danych przedstawiono zmienne panelowe, takie jak linia lotnicza (I), rok (T), cena paliwa (PF), wskaźnik wykorzystania floty (LF), wydajność (Q) i całkowity koszt (C). Analizując statystyki podstawowe, zauważono zróżnicowanie kosztów lotniczych oraz istotne różnice w cenach paliwa, wydajności i wskaźniku wykorzystania floty między badanymi liniami lotniczymi.

Analiza korelacji między wszystkimi zmiennymi wykazała, że istnieje silna zależność między całkowitym kosztem a wydajnością, co może być zaskakujące. Dodatkowo, istnieją umiarkowane zależności między kosztem a ceną paliwa oraz wskaźnikiem wykorzystania floty.

W kolejnym etapie analizowane były wpływ zlogarytmowania niektórych zmiennych, szczególnie Q i PF. W wyniku analizy zdecydowano się na nie wprowadzanie do modelu zmiennych zlogarytmowanych.

Następnie, przeprowadzono analizę różnych modeli takich jak: regresji łącznej, efektów losowych i efektów ustalonych. Testy porównawcze wykazały, że preferowanym modelem jest jednokierunkowy model efektów ustalonych. Testy takie jak Hausmana, Breuscha-Pagana i Walda potwierdziły tę decyzję, a kryteria informacyjne Akaike'a i Hannana-Quinna również wspierały ten wybór.

Weryfikacja modelu wykazała istotną dodatnią autokorelację składnika losowego w resztach, co może wpływać na brak trafności wyników. Wnioski z analizy wskazują, że mimo preferencji dla modelu efektów ustalonych, konieczne mogą być dalsze dostosowania w celu poprawy trafności modelu.