AKADEMIA GÓRNICZO-HUTNICZA Im. Stanisława Staszica w Krakowie
Wydział Zarządzania, Kierunek Informatyka i Ekonometria

Łukasz Pyrek

Badanie wpływu wybranych czynników na wyniki egzaminów uczniów 8 klasy w stanie Massachusetts

Czerwiec 2023

Spis treści

CEL PROJEKTU

Celem projektu jest wyznaczenie najważniejszych determinant wyników egzaminów 8-klasistów. W tym celu następuje próba dopasowania modelu ekonometrycznego za pomocą klasycznej metody najmniejszych kwadratów.

1. Opis danych

Dane zawierają średnie wyniki dla poszczególnych dystryktów publicznych szkół podstawowych w Massachusetts w 1998 roku. Wyniki testu pochodzą z testu Massachusetts Comprehensive Assessment System (MCAS), przeprowadzonego wiosną 1998 roku w publicznych szkołach w Massachusetts. Test jest sponsorowany przez Departament Edukacji Massachusetts i jest obowiązkowy dla wszystkich szkół publicznych. Dane analizowane tutaj dotyczą ogólnego wyniku całkowitego, który jest sumą wyników z części testu z języka angielskiego, matematyki i nauk przyrodniczych.

Dane dotyczące stosunku uczniów do nauczycieli, procenta uczniów otrzymujących dofinansowane lunch, a także procenta uczniów, którzy wciąż uczą się angielskiego, są średnimi dla każdego dystryktu szkół podstawowych z roku szkolnego 1997-1998 i zostały uzyskane od Departamentu Edukacji Massachusetts. Dane dotyczące średniego dochodu dystryktu pochodzą z Narodowego Spisu Powszechnego z 1990 roku.

Próbki zawierające braki danych zostały usunięte przed analizą w celu zapewnienia czystości i spójności danych.

Serie danych użytych w modelu
Zmienna objaśniana:
totsc8: Wynik ósmoklasistów (matematyka+język angielski+nauki przyrodnicze)

Zmienne objaśniające:
regday: Wydatki na ucznia, zwykłe
speced: Procent uczniów ze specjalnymi potrzebami edukacyjnymi
lnchpct: Procent osób uprawnionych do lunchu w cenie obniżonej lub bezpłatnego
percap: Dochód na osobę (roczny, w tys. dolarów)

Dane pochodzą z podręcznika do ekonometrii:
Stock and Watson, *Introduction to Econometrics.*

Udostępnione na oficjalnej stronie programu Gretl:
https://gretl.sourceforge.net/gretl_data.html

2. Statystyki opisowe

Należy mieć na uwadze, że pojedyncza obserwacja w tym zestawie danych jest średnią wartością z danego dystryktu.

Tabela 1. Statystyki opisowe zmiennej totsc8

Średnia	Mediana	Minimalna	Maksymalna
698,41	698	641	747
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
21,053	0,030144	-0,19802	-0,091792
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
661	731,9	27	0

Średni wynik egzaminu 8-klasistów w stanie wynosił 698,41. Odchylenie standardowe wynosi 21, co wskazuje na niewielkie rozproszenie się danych w okól średniej. Skośność i kurtoza jest bliska zeru, oraz mediana jest równa średniej więc można podejrzewać, że te dane pochodzą z rozkładu normalnego.

Tabela 2. Statystyki opisowe zmiennej regday

Średnia	Mediana	Minimalna	Maksymalna
4709,7	4525	3023	8759
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
867,32	0,18416	1,4688	3,3974
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
3693,5	6368,3	892,75	0

Średnie wydatki na ucznia wynosiły 4709. Odchylenie standardowe wynosi 867. Rozkład zmiennej jest prawostronnie skośny. 95% percentyl wynosi 6368, maksymalna osiągana wartość 8759. Wskazuje to na występowanie nie wielkiej ilości dystryktów które średnio wydają znacząco więcej na ucznia. Współczynnik zmienności wynosi 18% co oznacza niską zmienność.

Tabela 3. Statystyki opisowe zmiennej speced

Średnia	Mediana	Minimalna	Maksymalna
16,053	15,55	10,4	26
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
3,2651	0,2034	0,60885	0,18229
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
11,3	22,075	4,275	0

Zmienna ta jest wyrażona w procentach.

Średnio 16% uczniów posiada specjalne potrzeby edukacyjne. Obserwacje są rozrzucone wokół średniej o około 3 punkty procentowe. Współczynnik zmienności wynosi 20% co oznacza niską zmienność.

Tabela 4. Statystyki opisowe zmiennej lnchpct

Średnia	Mediana	Minimalna	Maksymalna
16,057	11,2	0,4	76,2
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
15,951	0,99343	1,8422	3,2532
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
2,01	53,87	15,9	0

Zmienna ta jest wyrażona w procentach.

Średnio 16% uczniów jest uprawnionych do lunchu w cenie obniżonej lub bezpłatnego. Obserwacje są rozrzucone wokół średniej o około 16 punktów procentowych. Współczynnik zmienności wynosi 100% co oznacza bardzo wysoką zmienność. Rozkład zmiennej jest prawostronnie skośny oraz mediana jest znacząco mniejsza od średniej. Wskazuje to na większą ilość obserwacji mniejszych od średniej oraz skoncentrowanie wartości odstających w prawym ogonie.

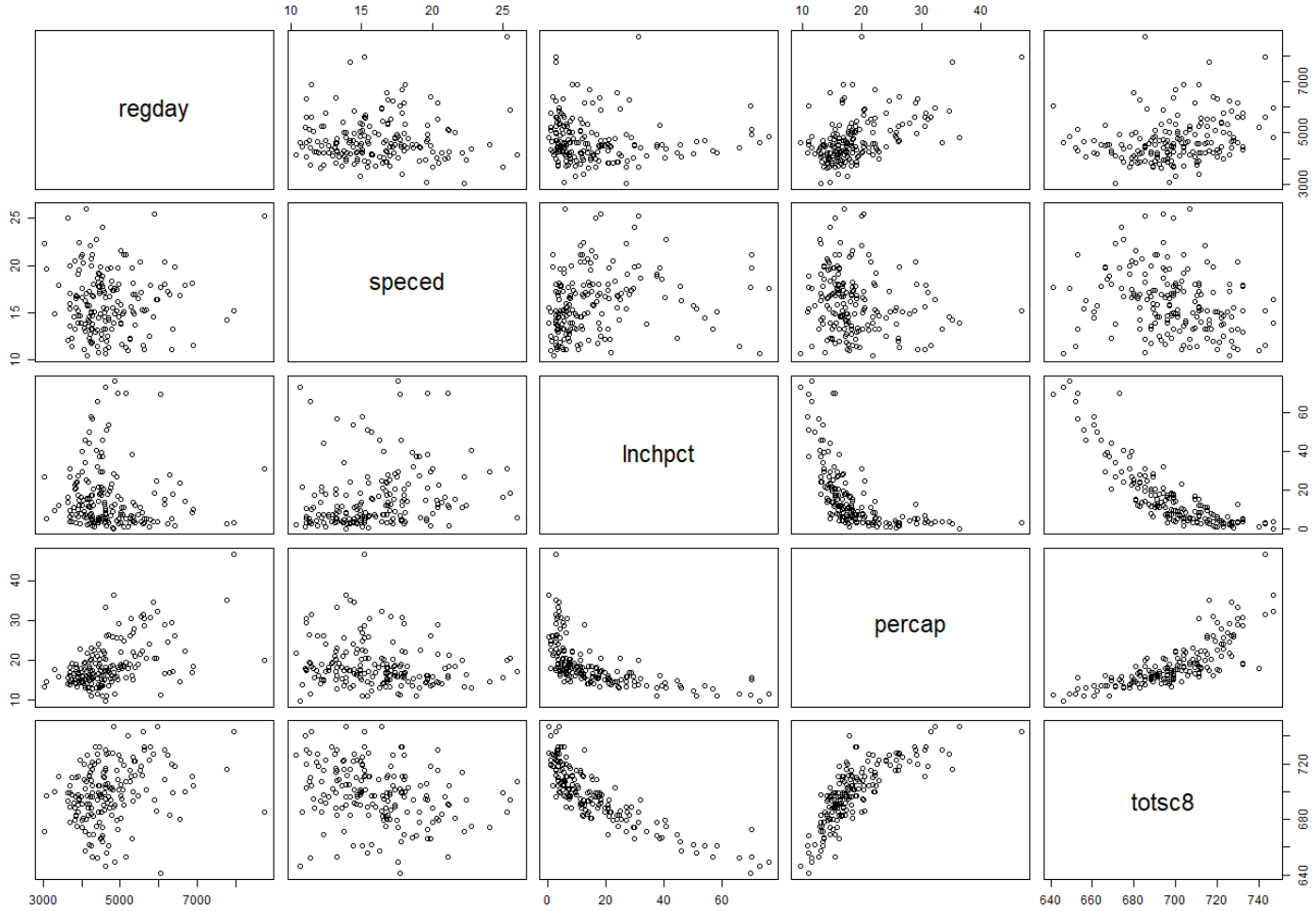
Tabela 5. Statystyki opisowe zmiennej percap

Średnia	Mediana	Minimalna	Maksymalna
18,739	17,313	9,686	46,855
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
5,6191	0,29986	1,6498	3,6834
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
12,712	30,713	5,2183	0

Dane te pochodzą z 1990 roku, natomiast pozostałe z 1998 r.

Średni dochód na osobę wynosi 18,7tys. dolarów rocznie. Obserwacje są rozrzucone wokół średniej o około 5,6tys. Współczynnik zmienności wynosi 30% co oznacza przeciętną zmienność. Rozkład zmiennej jest prawostronnie skośny oraz wartość mediany jest niewiele mniejsza od średniej wskazuje to na skoncentrowanie wartości odstających w prawym ogonie.

3. Wykres zależności zmiennych



Wykres 1. Wykres zależności zmiennych

Analizując wykresy, można zaobserwować silną korelację między zmiennymi *lnchpct* oraz *percap* a zmienną objaśnianą. Wydaje się, że to głównie te zmienne będą odgrywały kluczową rolę w modelu. Jednakże, ważne jest zauważenie, że te zmienne nie wykazują liniowej zależności. W rzeczywistości, rozrzut danych dla tych zmiennych formuje krzywą. Wobec tego, możliwym krokiem do podjęcia może być transformacja tych zmiennych. Przekształcenia mogą pomóc w uwzględnieniu nieliniowości w zależności między tymi zmiennymi a zmienną objaśnianą.

Możemy też się spodziewać ewentualnego problemu ze współliniowością zmiennych „lnchpct" i „percap"

Natomiast, na podstawie analizy wykresów pozostałych zmiennych, nie wydają się one wykazywać silnej zależności z zmienną objaśnianą.

4. Analiza korelacji

regday	speced	lnchpct	percap	totsc8	
	Corr: 0.030	Corr: -0.071	Corr: 0.518***	Corr: 0.260***	regday
Corr: 0.030		Corr: 0.202**	Corr: -0.171*	Corr: -0.263***	speced
Corr: -0.071	Corr: 0.202**		Corr: -0.574***	Corr: -0.834***	lnchpct
Corr: 0.518***	Corr: -0.171*	Corr: -0.574***		Corr: 0.777***	percap
Corr: 0.260***	Corr: -0.263***	Corr: -0.834***	Corr: 0.777***		totsc8

Tabela 6. Macierz korelacji zmiennych

Tabela również informuje o wyniku testu na istotność statystyczną korelacji z następującym zestawem hipotez:

$$H_0: corr = 0$$
$$H_1: corr \neq 0$$

Gwiazdki przy wartości korelacji oznaczają przedział w jakim znajduje się p-value policzone dla testu dla danej korelacji:

*** p-value < 0.001
** p-value < 0.01
* p-value < 0.05

UWAGA: Wszystkie testy przeprowadzone w tym projekcie zakładają poziom istotności równy 5%.

Na początku skupmy się na korelacji zmiennej objaśnianej *totsc8* ze zmiennymi objaśniającymi. Jak można było wnioskować na podstawie wykresu rozrzutu zmiennych, największa korelacja występuje ze zmiennymi *lnchpct* i *percap,* jest to pożądany efekt. Głownie te zmienne będą objaśniały *totsc8*. Pozostałe zmienne posiadają istotną statystycznie wartość korelacji rzędu 0.26 (na moduł). Jest to stosunkowo niewielka wartość natomiast nie oznacza to, że zmienne te nie mają wpływu na zmienną objaśnianą.

Natomiast jeżeli chodzi o korelacje pomiędzy zmiennymi objaśniającymi, chcemy, aby była ona jak najmniejsza. Jedynie zmienna *percap* cechuje się znaczącą wartością współczynnika korelacji z pozostałymi zmiennymi. Wynosi ona kolejno 0.518 ze zmienną *regday* oraz -0.574 ze zmienną *lnchpct*. Oznacza to, że możemy spodziewać się problemów ze współliniowością, która będzie testowana w dalszej części analizy.

5. Model ściśle liniowy

Szacujemy następujący model liniowy:

$$totsc8 = \beta_0 + \beta_1\, regday + \beta_2\, speced + \beta_3\, lnchpct + \beta_4\, percap$$

Równanie 1. Model 1

Do estymacji parametrów wykorzystamy Klasyczną Metodę Najmniejszych Kwadratów (KMNK).

Model 1: Estymacja KMNK, wykorzystane obserwacje 1-180
Zmienna zależna (Y): totsc8

	współczynnik	błąd standardowy	t-Studenta	wartość p	
const	687,710	5,01395	137,2	4,80e-180	***
regday	-0,000201686	0,000931250	-0,2166	0,8288	
speced	-0,472067	0,205051	-2,302	0,0225	**
lnchpct	-0,745203	0,0528171	-14,11	1,12e-030	***
percap	1,66470	0,175300	9,496	1,64e-017	***

Średn.aryt.zm.zależnej	698,4111	Odch.stand.zm.zależnej	21,05268
Suma kwadratów reszt	13258,53	Błąd standardowy reszt	8,704197
Wsp. determ. R-kwadrat	0,832880	Skorygowany R-kwadrat	0,829060
F(4, 175)	218,0385	Wartość p dla testu F	7,65e-67
Logarytm wiarygodności	-642,3585	Kryt. inform. Akaike'a	1294,717
Kryt. bayes. Schwarza	1310,682	Kryt. Hannana-Quinna	1301,190

Tabela 7. Estymacja Modelu 1

*UWAGA: * - p-value < 0.1; ** - p-value < 0.05; *** - p-value < 0.01*
Na wstępie należy zbadać poprawność modelu. Po analizie okazuje się, że model posiada następujące wady:

Wady modelu

Brak normalności rozkładu reszt.

Jednym z założeń poprawnej interpretacji współczynnika R^2 oraz testów statystycznych analizujących model jest normalność rozkładu reszt.

```
Rozkład częstości dla uhat, obserwacje 1-180
liczba przedziałów = 13, średnia = 5,36855e-014, odch.std. = 8,7042

      Przedziały       średnia   liczba   częstość  skumulowana

          < -17,836   -19,900        2      1,11%    1,11%
   -17,836 - -13,708  -15,772        5      2,78%    3,89%
   -13,708 - -9,5793  -11,643       10      5,56%    9,44% *
   -9,5793 - -5,4509  -7,5151       32     17,78%   27,22% ******
   -5,4509 - -1,3226  -3,3868       38     21,11%   48,33% *******
   -1,3226 -  2,8057   0,74154      33     18,33%   66,67% ******
    2,8057 -  6,9340   4,8699       23     12,78%   79,44% ****
    6,9340 -  11,062   8,9982       16      8,89%   88,33% ***
    11,062 -  15,191   13,127       14      7,78%   96,11% **
    15,191 -  19,319   17,255        3      1,67%   97,78%
    19,319 -  23,447   21,383        2      1,11%   98,89%
    23,447 -  27,576   25,511        1      0,56%   99,44%
         >=   27,576   29,640        1      0,56%  100,00%

Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.Test Doornika-
Hansena (1994) - transformowana skośność i kurtoza.:
Chi-kwadrat(2) = 10,268 z wartością p 0,00589
```

Tabela 8. Test Doornika-Hansena na normalność rozkładu reszt modelu 1

Brak poprawnej postaci funkcyjnej.

Test RESET Ramseya wykazał, że istnieje inna postać funkcyjna która lepiej opisuje model.

```
Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 1-180
Zmienna zależna (Y): totsc8

              współczynnik    błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const     -60037,7          24501,5             -2,450     0,0153   **
  regday         0,0267565        0,0107663        2,485     0,0139   **
  speced        62,7267          25,2912           2,480     0,0141   **
  lnchpct       98,7289          39,8910           2,475     0,0143   **
  percap      -220,757           88,9982          -2,480     0,0141   **
  yhat^2         0,190866         0,0775501        2,461     0,0148   **
  yhat^3        -9,08388e-05      3,74673e-05     -2,424     0,0164   **

Statystyka testu: F = 3,710002,
z wartością p = P(F(2,173) > 3,71) = 0,0264
```

Tabela 9. Test RESET Ramseya dla modelu 1

6. Próba poprawy modelu

a. Redukcja ilości zmiennych – Metoda Helwiga i metoda krokowo-wsteczna

W celu poprawy modelu zredukuje ilość zmiennych w modelu (wyeliminuje nieistotne).

```
Najlepsza kombinacja:
regday speced lnchpct percap
  0       0      1      1

Integralna pojemność informacyjna: 0,82481
```

Tabela 10. Dobór zmiennych metodą Helwiga

```
Sekwencyjna eliminacja nieistotnych zmiennych przy dwustronnym
obszarze krytycznym, alfa = 0,05

Wyeliminowano nieistotną zmienną: regday (wartość p = 0,829)

Test porównawczy z Modelem 1

Hipoteza zerowa: parametr regresji jest równy zero dla regday
Statystyka testu: F(1, 175) = 0,0469049, wartość p 0,828791
Pominięcie zmiennych poprawiło 3 z 3 kryteriów informacyjnych (AIC,
BIC, HQC).
```

Tabela 11. Dobór zmiennych metodą krokowo-wsteczną

Dobór zmiennych objaśniających

Metoda Helwiga: *lnchpct, percap*
Metoda krokowo wsteczna: *lnchpct, percap, speced*

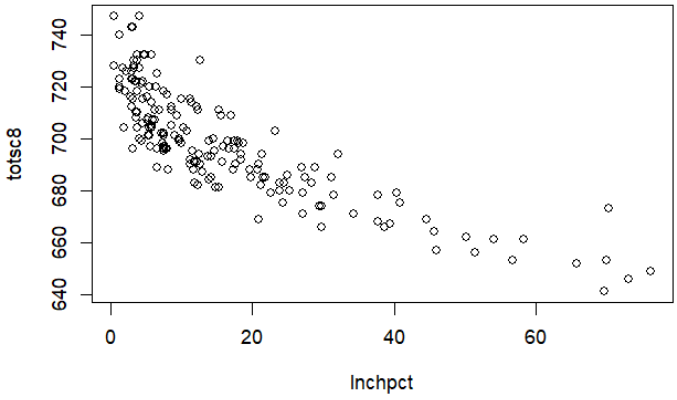
Modele z powyższymi kombinacjami zmiennych objaśniających nadal posiadają te same wady co model wyjściowy, natomiast redukcja zmiennych poprawiła kryteria informacyjne, prostotę modelu oraz nieznacznie zmniejszyła współczynnik determinacji R^2 (pomijalne wartości rzędu 0.005).

Różnica kryteriów informacyjnych oraz wsp. R^2 pomiędzy kombinacją zmiennych (*lnchpct, percap) a (lnchpct, percap, speced)* jest nieznacząca.

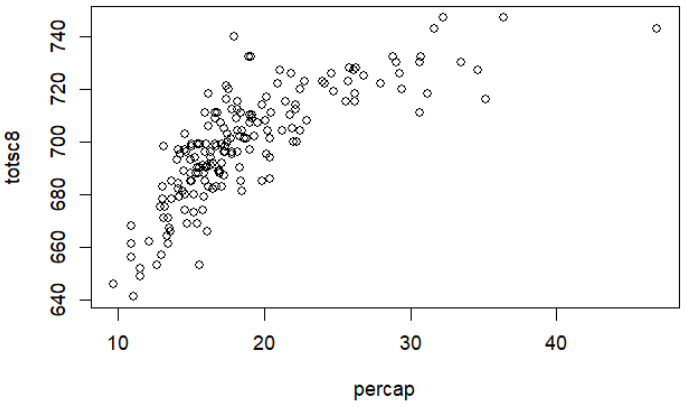
Z powodu następnych transformacji oraz ułatwienia prób poprawy modelu decyduję się na następujący dobór zmiennych: *lnchpct, percap*

b. Zmiana postaci funkcyjnej modelu

Wyniki Testu RESET Ramseya wskazują, że istnieje inna postać funkcyjna modelu która lepiej pasuje do danych. Wynika to z tego, że występuje nieliniowa zależność między zmiennymi objaśniającymi a zmienną objaśnianą.



Wykres 2. Wykres rozrzutu zmiennych totsc8 i lnchpct



Wykres 3. Wykres rozrzutu zmiennych totsc8 i percap

W celu wyjaśnienia nieliniowej zależności należy zmienić równanie modelu poprzez np. transformacje zmiennych. Okazuje się, że samo logarytmowanie lub podniesienie zmiennych do kwadratu nie rozwiązuje problemu. Decyduję się więc na dodanie dodatkowych zmiennych do równania. Po kilku próbach dopasowania odpowiedniego modelu decyduje się na następujący:

$$totsc8 = \beta_0 + \beta_1\, percap + \beta_2\, lnchpct + \beta_3\, lnchpct\_sq$$

Równanie 2. Model 2

gdzie *lnchpct_sq = lnchpct*^{2}

```
Model 2: Estymacja KMNK, wykorzystane obserwacje 1-180
Zmienna zależna (Y): totsc8

             współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       687,116          3,92794         174,9      2,64e-199 ***
  percap        1,46244        0,151675          9,642    6,26e-018 ***
  lnchpct      -1,21656        0,144740         -8,405    1,40e-014 ***
  lnchpct_sq    0,00670445     0,00201732        3,323    0,0011    ***

Średn.aryt.zm.zależnej  698,4111   Odch.stand.zm.zależnej  21,05268
Suma kwadratów reszt    12868,62   Błąd standardowy reszt  8,550859
Wsp. determ. R-kwadrat  0,837795   Skorygowany R-kwadrat   0,835030
F(3, 176)               303,0157   Wartość p dla testu F   2,98e-69
Logarytm wiarygodności -639,6721   Kryt. inform. Akaike'a  1287,344
Kryt. bayes. Schwarza   1300,116   Kryt. Hannana-Quinna    1292,523
```

Tabela 12. Estymacja modelu 2

Współczynnik determinacji R^2 dla Modelu 2 jest o około 0.05 większy od R^2 dla modelu 1 oraz w przybliżeniu 0.1 większy od współczynnika dla liniowego modelu dla tych samych zmiennych: $totsc8 = \beta_0 + \beta_1\, percap + \beta_2\, lnchpct$.

Wszystkie z 3 kryteriów informacyjnych się poprawiły (AIC, BIC, HQC) w porównaniu do tych dwóch modeli.

Test RESET Ramseya dla Modelu 2:

```
Statystyka testu: F = 0,340487,
z wartością p = P(F(2,174) > 0,340487) = 0,712
```

Tabela 13. Test RESET Ramseya dla Modelu 2

Test nie odrzuca hipotezy zerowej o poprawności funkcyjnej modelu.

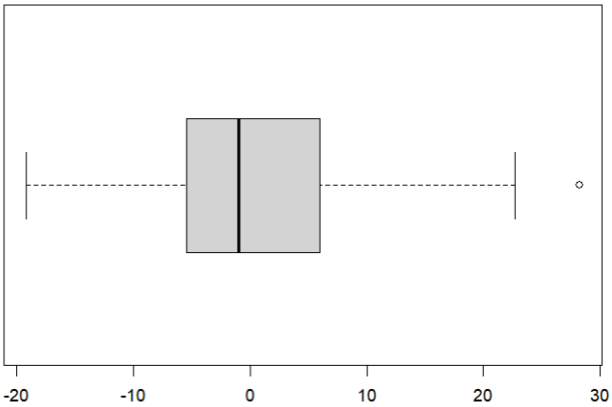
c. Wartości odstające

Model 2 nadal nie spełnia założenia o normalności rozkładu reszt:

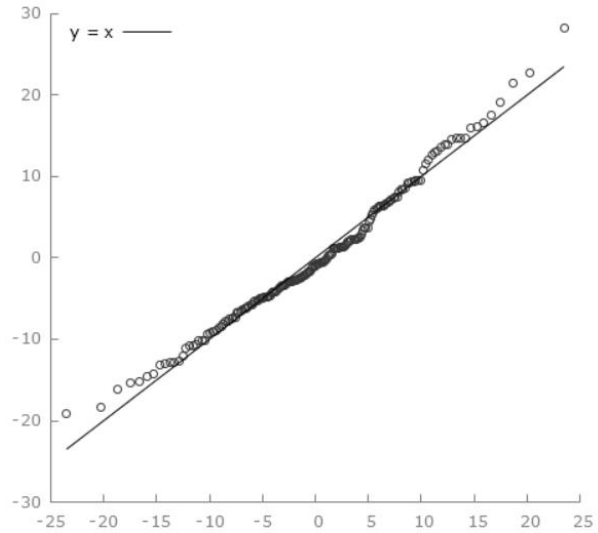
```
Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.Test Doornika-
Hansena (1994) - transformowana skośność i kurtoza.:
Chi-kwadrat(2) = 6,667 z wartością p 0,03568
```

Tabela 14. Test Doornika-Hansena na normalność rozkładu reszt modelu 2

Przeanalizujmy wykres pudełkowy oraz wykres kwantylowy (Q-Q) reszt:



Wykres 4. Wykres pudełkowy reszt modelu 2



Wykres 5. Wykres Q-Q reszt modelu 2

uhat	totsc8	percap	lnchpct	lnchpct_sq
28,22274	740	17,937	1,3	1,69

Tabela 15. Wartości obserwacji z największą resztą modelu 2.

Jak widać jedna obserwacja odstaje, jest to największa wartość reszt modelu.

Spróbujmy usunąć tę obserwacje i teraz przetestować normalność rozkładu reszt.

```
Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.Test Doornika-
            Hansena (1994) - transformowana skośność i kurtoza.:
                Chi-kwadrat(2) = 4,380 z wartością p 0,11193
```

Tabela 16. Test Doornika-Hansena dla modelu 2 po usunięciu wartości odstającej

Po usunięciu jednej obserwacji, test nie odrzuca hipotezy zerowej dotyczącej normalności rozkładu.

Ostatecznie udało się skonstruować model, który uwzględnia nieliniową zależność danych oraz spełnia założenia MNK i założenie o normalności rozkładu reszt. Dzięki temu przeprowadzona dalej analiza będzie mogła zostać poprawnie interpretowana.

7. Testowanie własności modelu

Finalnie następujący wyestymowany model będzie poddany analizie:

$$totsc8 = 685{,}222 + 1{,}51645\ percap - 1{,}14542\ lnchpct + 0{,}0059\ lnchpct\_sq$$

Równanie 3. Wyestymowany Model 3

```
Model 3: Estymacja KMNK, wykorzystane obserwacje 1-179
Zmienna zależna (Y): totsc8

             współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       685,222          3,85085         177,9      1,09e-199 ***
  percap        1,51645        0,148011         10,25     1,36e-019 ***
  lnchpct      -1,14542        0,141956         -8,069    1,09e-013 ***
  lnchpct_sq    0,00587719     0,00197218        2,980    0,0033    ***

Średn.aryt.zm.zależnej  698,1788   Odch.stand.zm.zależnej  20,87904
Suma kwadratów reszt    12048,58   Błąd standardowy reszt  8,297532
Wsp. determ. R-kwadrat  0,844727   Skorygowany R-kwadrat   0,842065
F(3, 175)               317,3498   Wartość p dla testu F   1,62e-70
Logarytm wiarygodności -630,7238   Kryt. inform. Akaike'a  1269,448
Kryt. bayes. Schwarza   1282,197   Kryt. Hannana-Quinna    1274,617
```

Tabela 17. Estymacja Modelu 3

a. Współczynnik determinacji

Współczynnik determinacji R^2 wynosi 0.844, natomiast skorygowany R^2 0.842, oznacza to, że model wyjaśnia około 84% zmienności, oraz nie jest przeparametryzowany.

b. Efekt katalizy

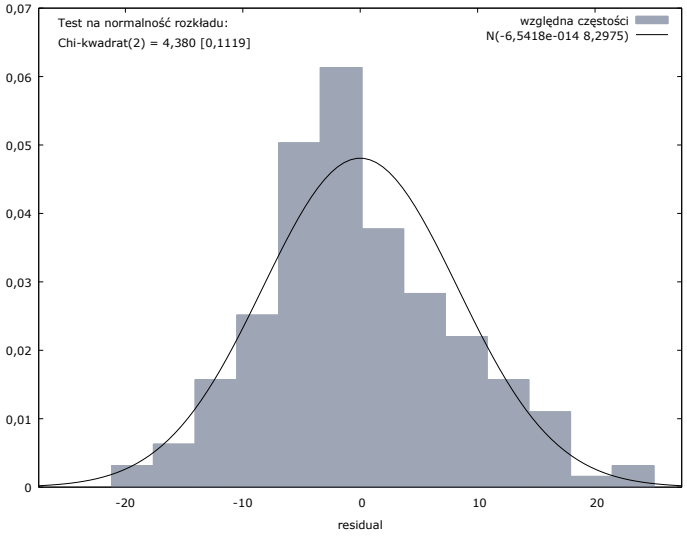
Ważnym aspektem jest zbadanie występowania efektu katalizy, czyli możliwości występowania pary zmiennych która powoduje zawyżenie współczynnika determinacji, pomimo tego, że charakter i siła powiązań zmiennych objaśniających i zmiennej objaśnianej nie uzasadniają takiego wyniku.

```
Zmienna "lnchpct_sq" jest katalizatorem ze zmienną "lnchpct"
Natężenie: 0,047972
Względne natężenie: 5,679044%
```

Tabela 18. Efekt katalizy dla Modelu 3

Efekt katalizy jest nieznacząco różny od zera więc można stwierdzić, że efekt katalizy jest nieistotny lub że nie występuje.

c. Normalność rozkładu składnika losowego.



Wykres 4. Histogram wraz estymowanym wykresem gęstośći rozkładu reszt modelu 3.

Rozkład częstości dla residual, obserwacje 1-179
liczba przedziałów = 13, średnia = -6,54176e-014, odch.std. = 8,29753

Przedziały	średnia	liczba	częstość	skumulowana	
< -17,617	-19,393	2	1,12%	1,12%	
-17,617 - -14,064	-15,840	4	2,23%	3,35%	
-14,064 - -10,510	-12,287	10	5,59%	8,94%	**
-10,510 - -6,9570	-8,7336	16	8,94%	17,88%	***
-6,9570 - -3,4037	-5,1803	32	17,88%	35,75%	******
-3,4037 - 0,14957	-1,6271	39	21,79%	57,54%	*******
0,14957 - 3,7029	1,9262	24	13,41%	70,95%	****
3,7029 - 7,2561	5,4795	18	10,06%	81,01%	***
7,2561 - 10,809	9,0328	14	7,82%	88,83%	**
10,809 - 14,363	12,586	10	5,59%	94,41%	**
14,363 - 17,916	16,139	7	3,91%	98,32%	*
17,916 - 21,469	19,693	1	0,56%	98,88%	
>= 21,469	23,246	2	1,12%	100,00%	

Hipoteza zerowa: dystrybuanta empiryczna posiada rozkład normalny.Test Doornika-Hansena (1994) - transformowana skośność i kurtoza.:
Chi-kwadrat(2) = 4,380 z wartością p 0,11193

Tabela 19. Test Doornika-Hansena dla reszt modelu 3.

Testy nie odrzucają hipotezy zerowej o normalności rozkładu składnika losowego.

Normalność nie jest wymaganą właściwością składnika losowego, ale umożliwia korzystanie z testów statystycznych weryfikujących pozostałe własności składnika losowego, dlatego dążyliśmy do uzyskania tej własności.

d. Istotność zmiennych.

Test t-studenta:

$$H_0: \alpha_j = 0$$
$$H_1: \alpha_j \neq 0$$

Test F:

$$H_0: \alpha_1 = \cdots = \alpha_k = 0$$
$$H_1: conajmniej\ jeden\ parametr\ jest\ różny\ od\ zera$$

	t-Studenta	wartość p	
const	177,9	1,09e-199	***
percap	10,25	1,36e-019	***
lnchpct	-8,069	1,09e-013	***
lnchpct_sq	2,980	0,0033	***
Wartość p dla testu F		1,62e-70	

Tabela 20. Wyniki testów t-studenta i testu F na istotność zmiennych.

Testy odrzucają hipoteze zerową o nieistotności zmiennych.

Zmienne modelu cechują się duża istotnością, świadczy o tym bardzo niskie p-value zarówno dla testów t-studenta dla istotności pojedyńczych zmiennych jak i testu F dla istotności całego podzbioru zmiennych.

e. Testy dodanych (pominiętych zmiennych).

```
Hipoteza zerowa: parametry regresji dla wskazanych zmiennych są równe zero
  regday, speced
Statystyka testu: F(2, 173) = 0,918396, wartość p 0,401096
Dodanie zmiennych poprawiło 0 z 3 kryteriów informacyjnych (AIC, BIC, HQC).
```

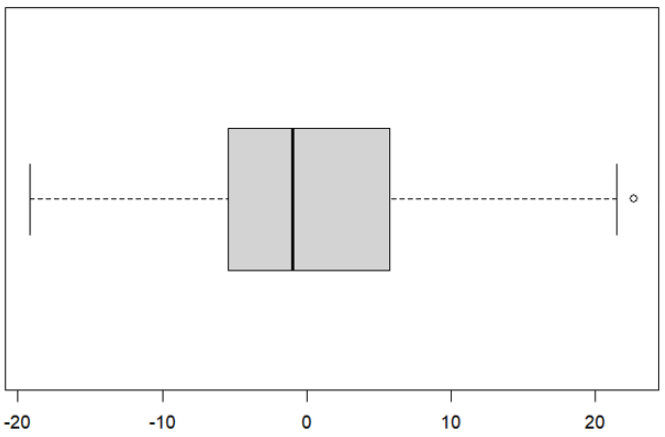
Tabela 21. Wyniki testu dodanych zmiennych dla modelu 3.

Test nieodrzuca hipotezy zerowej o nieistotności parametrów regday i speced

Dobór zmiennych jest odpowiedni. Zmienne nieuwzględnione w modelu są nieistotne.

f. Obserwacje odstające.

Uprzednio usunęliśmy najbardziej odstającą obserwację, aby poprawić właściwości modelu.



Wykres 5. Wykres pudełkowy reszt modelu 3

Analizując wykres pudełkowy można stwierdzić, że nie występują znaczące wartości odstające.

g. Test liczby serii

Estymacja modelu za pomocą KMNK jest równoważne z założeniem o liniowej zależności zmiennej objaśnianej od zmiennej objaśniającej. Weryfikacja tego założenia jest niezbędna do prawidłowej interpretacji współczynnika determinacji.

$$H_0: postać\ modelu\ jest\ dobrze\ dobrana;\ model\ jest\ liniowy$$
$$H_1: \sim H_1$$

Gretl dla testu serii podaje hipoteze zerową jako „próba jest losowa", jest to równoważne z powyższą hipotezą.

```
Liczba serii (R) dla zmiennej 'e' = 94
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(90,5, 6,67083),
test z-score = 0,524672, przy dwustronym obszarze krytycznym p = 0,599811
```

Tabela 22. Test serii dla modelu 3

Nie odrzucamy hipotezy zerowej o liniowości modelu. Model jest poprawny.

h. Test RESET

Podobnej informacji co test serii dostarcza test RESET Ramseya. Test ten upewnia nas czy wybrana postać modelu jest dobrze dobrana do opisu zmienności danej zmiennej objaśnianej, a dokładnie stosowany jest w celu sprawdzenia, czy to liniowa postać modelu (względem funkcji kwadratowej lub sześciennej) jest najlepszym możliwym do wybrania modelem.

Wcześniej postaraliśmy się, aby ten model był poprawny funkcyjnie.

$$H_0: postać\ funkcyjna\ modelu\ jest\ dobrze\ dobrana$$
$$H_1: \sim H_0$$

Pomocnicze równanie regresji dla testu specyfikacji RESET
Estymacja KMNK, wykorzystane obserwacje 1-179
Zmienna zależna (Y): totsc8

	współczynnik	błąd standardowy	t-Studenta	wartość p
const	−26907,0	50911,0	−0,5285	0,5978
percap	−89,4549	171,398	−0,5219	0,6024
lnchpct	67,8299	129,623	0,5233	0,6014
lnchpct_sq	−0,348066	0,667855	−0,5212	0,6029
yhat^2	0,0884059	0,160356	0,5513	0,5821
yhat^3	−4,32731e-05	7,58073e-05	−0,5708	0,5689

Statystyka testu: F = 0,671662,
z wartością p = P(F(2,173) > 0,671662) = 0,512

Tabela 23. Test RESET dla modelu 3

Test nie odrzuca hipotezy zerowej o poprawności funkcyjnej modelu. Model jest dobrze dobrany.

i. Testowanie heteroskedastyczności

Występowanie heteroskedastyczności składnika losowego w modelu wiąże się z niespełnieniem założeń MNK. Przeprowadzimy test White'a oraz Breuscha-Pagana które mają ten sam zestaw hipotez:

```
Test White'a na heteroskedastyczność reszt (zmienność wariancji resztowej)
Estymacja KMNK, wykorzystane obserwacje 1-179
Zmienna zależna (Y): uhat^2
Z powodu ścisłej współliniowości pominięto zmienną: sq_lnchpct

                współczynnik   błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const          188,253         196,269           0,9592     0,3388
  percap         -10,1814         13,0454          -0,7805     0,4362
  lnchpct         -2,71846        16,3315          -0,1665     0,8680
  lnchpct_sq      -0,262047        0,557705        -0,4699     0,6391
  sq_percap        0,188989        0,210261         0,8988     0,3700
  X2_X3            0,323576        0,592216         0,5464     0,5855
  X2_X4            0,000281238     0,00872054       0,03225    0,9743
  X3_X4            0,00611857      0,00983963       0,6218     0,5349
  sq_lnchpct_sq   -3,97087e-05     6,63651e-05     -0,5983     0,5504

  Wsp. determ. R-kwadrat = 0,042000

Statystyka testu: TR^2 = 7,517951,
z wartością p = P(Chi-kwadrat(8) > 7,517951) = 0,481914
```

Tabela 25. Test White'a na heteroskedastyczność reszt dla modelu 3

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-179
Zmienna zależna (Y): standaryzowane uhat^2 (odporna wariancja Koenkera)

              współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const       -4,01402         43,0797         -0,09318    0,9259
  percap       0,893598         1,65581         0,5397     0,5901
  lnchpct     -1,08717          1,58808        -0,6846     0,4945
  lnchpct_sq   0,00936574       0,0220629       0,4245     0,6717

  Wyjaśniona suma kwadr. = 28559,1

Statystyka testu: LM = 3,327208,
z wartością p = P(Chi-kwadrat(3) > 3,327208) = 0,343874
```

Tabela 24. Test Breuscha-Pagana na heteroskedastyczność reszt modelu 3

$$H_0: \sigma_i^2 = \sigma^2; homoskedastyczność\ składnika\ losowego$$
$$H_1: \sigma_i^2 \neq \sigma^2; heteroskedastyczność\ składnika\ losowego$$

Oba testy nie odrzucają hipotezy zerowej o homoskedastyczności reszt. Reszty nie są heteroskedastyczne.

j. Test Chowa

Test Chowa pozwala na statystyczną identyfikację zmiany strukturalnej parametrów, czyli zbadanie stabilności parametrów. Punktem wyjścia jest wybór punktu załamania strukturalnego. W praktyce jest to punkt względem, którego dzielimy dane na dwie podpróbki, a następnie estymujemy 2 modele dla każdej z podpróbki i badamy czy parametry tych modeli istotnie się róznią:

$$H_0: Wszystkie\ parametry\ modelu\ są\ takie\ same; stabilność\ parametrów$$
$$H_1: Conajmniej\ jeden\ parametr\ się\ rózni;\ brak\ stabilności\ parametrów$$

Obserwując nasze dane ciężko się doszukać jakiegoś specficznego punktu załamania strukturalnego, więc decyduję się na podzielenie danych na dwie równe podpróbki.

```
Pomocnicze równanie regresji dla testu Chowa
Estymacja KMNK, wykorzystane obserwacje 1-179
Zmienna zależna (Y): totsc8

                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------
  const          694,474          6,48465        107,1      1,17e-158 ***
  percap           1,32503        0,174108         7,610    1,76e-012 ***
  lnchpct         -2,63665        1,48197         -1,779    0,0770    *
  lnchpct_sq       0,0865762      0,123451         0,7013   0,4841
  splitdum       -21,1337        10,6749          -1,980    0,0493    **
  sd_percap        0,771775       0,427949         1,803    0,0731    *
  sd_lnchpct       1,65813        1,50277          1,103    0,2714
  sd_lnchpct_sq   -0,0821749      0,123488        -0,6654   0,5067

Średn.aryt.zm.zależnej  698,1788   Odch.stand.zm.zależnej  20,87904
Suma kwadratów reszt    11720,72   Błąd standardowy reszt  8,279023
Wsp. determ. R-kwadrat  0,848953   Skorygowany R-kwadrat   0,842769
F(7, 171)               137,2992   Wartość p dla testu F   9,31e-67
Logarytm wiarygodności -628,2546   Kryt. inform. Akaike'a  1272,509
Kryt. bayes. Schwarza   1298,008   Kryt. Hannana-Quinna    1282,849

Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 90
  F(4, 171) = 1,19584 z wartością p 0,3145
```

Tabela 26. Test Chowa na stabilność parametrów dla modelu 3

Test nie odrzuca hipotezy zerowej o stabilnośći parametrów. Parametry modelu są stabilne.

k. Współliniowość

Problem współlinowości zmiennych objaśniających może powodować zawyżenie błedów standardowych współczynników. W celu badania wspóliniowości skorzystamy ze współczynnika VIF (Variance inflation factor).

```
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

      percap    1,798
     lnchpct   13,266
  lnchpct_sq   10,901

VIF(j) = 1/(1 - R(j)^2), gdzie R(j) jest współczynnikiem korelacji wielorakiej
pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.
```

Tabela 27. Ocena wspólliniowości VIF dla zmiennych modelu 3

Wartości współczynnika powyżej 10 oznaczają wysoką współliniowość danych. Występuje ona pomiędzy zmienną *lnchpct* oraz *lnchpct_sq,* jest to tak zwana wspóliniowość strukturalna (spowodowana jest użyciem dodatkowej zmiennej, która została stworzona na podstawie już istniejącej).

Możemy usunąc problem współliniowości. Gdy wycentrujemy zmienną lnchpct, czyli odejmiemy od niej jej średnią, problem współliniowości będzie umiarkowany, a model będzie wyglądał następująco:

```
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

        percap    1,798
   lnchpct_cen    4,727
lnchpct_cen_sq    3,395

VIF(j) = 1/(1 - R(j)^2), gdzie R(j) jest współczynnikiem korelacji wielorakiej
pomiędzy zmienną 'j' a pozostałymi zmiennymi niezależnymi modelu.
```

Tabela 28. Ocena spółliniowości VIF dla wycentrowanych zmiennych

```
                współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const          668,267         2,68242         249,1      3,67e-225 ***
  percap           1,51645       0,148011         10,25     1,36e-019 ***
  lnchpct_cen     -0,955714      0,0847374       -11,28     1,60e-022 ***
  lnchpct_cen_sq   0,00587719    0,00197218        2,980    0,0033    ***

Średn.aryt.zm.zależnej  698,1788   Odch.stand.zm.zależnej  20,87904
Suma kwadratów reszt    12048,58   Błąd standardowy reszt  8,297532
Wsp. determ. R-kwadrat  0,844727   Skorygowany R-kwadrat   0,842065
F(3, 175)               317,3498   Wartość p dla testu F   1,62e-70
Logarytm wiarygodności -630,7238   Kryt. inform. Akaike'a  1269,448
```

Tabela 29. Model z wycentrowanymi zmiennymi

Analizując model dla wycentrowanych zmiennych można zauważyć, że p-value dla *lnchpct_sq* pozostaje takie samo, a dla pozostałych zmiennych dalej są bliskie zeru. Znaki przy wyestymowanych parametrach również pozostają bez zmian. Problem współliniowości też nie wpływa na dopasowanie modelu, prognozowane wartości czy też przedziały ufności. Z tego powodu ignoruje problem współliniowości i uznaje go za nieistotny. Pozostaje przy wyjściowym modelu ze zmiennymi nie wycentrowanymi.

l. Koincydencja

Koincydencja występuje, gdy znak przy współczynniku korelacji zmiennej objaśnianej ze zmienną objaśniającą jest taki sam, co znak przy wyestymowanym parametrze przy tej zmiennej objaśniającej, tj.:

$$sgn\left(r(x_j, y)\right) = sgn(\beta_j)$$

Równanie 4. Warunek koincydencji

totsc8	percap	lnchpct	lnchpct_sq	
1,0000	0,7868	-0,8347	-0,7095	totsc8

Tabela 30. Współczynniki korelacji pomiędzy zmienną objaśnianą a zmiennymi objaśniającymi

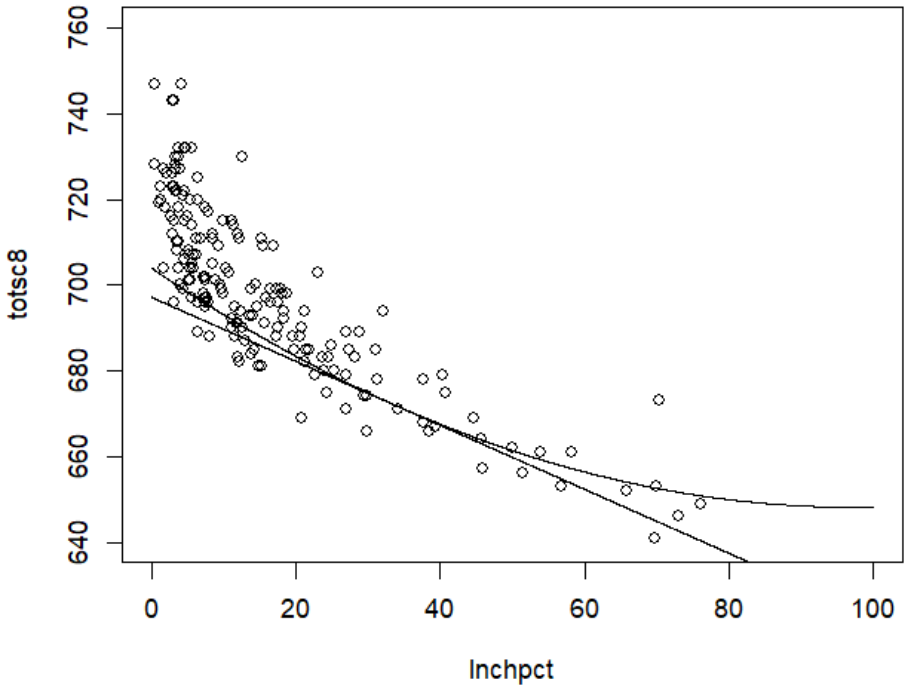
	współczynnik
const	685,222
percap	1,51645
lnchpct	-1,14542
lnchpct_sq	0,00587719

Tabela 31. Współczynniki parametrów modelu 3

Jak widać zmienna *lnchpct_sq* nie jest koincydentna, jednak ma ona na celu wyjaśnienie nieliniowej zależności między zmiennymi, widać to dobrze na wykresie rozrzutu zmiennej lnchpct i totsc8 na którą zostały nałożone 2 modele, badany kwadratowy oraz ściśle liniowy bez zmiennej lnchpct_sq, gdzie percap jest stałe i przyjmuje wartość średnią.

$$totsc8 = 685{,}222 + 1{,}51645\,\overline{percap} - 1{,}14542\,lnchpct + 0{,}0059\,lnchpct\_sq$$

$$totsc8 = 678{,}424 + 1{,}51645\,\overline{percap} - 0{,}746954\,lnchpct$$



Wykres 6. Wykres rozrzutu zmiennych totsc8 i lnchpct z dwoma modelami regresji

Jak widać na wykresie dla naszego badanego modelu wraz ze wzrostem wartości zmiennej *lnchpct* zmienna *totsc8* maleje, więc zgadza się to ze współczynnikiem korelacji. Można stwierdzić, że model jest koincydentny.

m. Interpretacja parametrów modelu

$$totsc8 = 685{,}222 + 1{,}51645\ percap - 1{,}14542\ lnchpct + 0{,}0059\ lnchpct\_sq$$

- ***1,51645 percap*** – wzrost rocznych dochodów na osobę o 1000 dolarów powoduje wzrost wyniku testów ośmioklasistów o około 1,5 punktów, ceteris paribus;
- $\mathbf{-1{,}14542\ lnchpct + 0{,}0059\ lnchpct\_sq}$ – ciężej interpretować nieliniowy wpływ zmiennej lnchpct na totsc8.

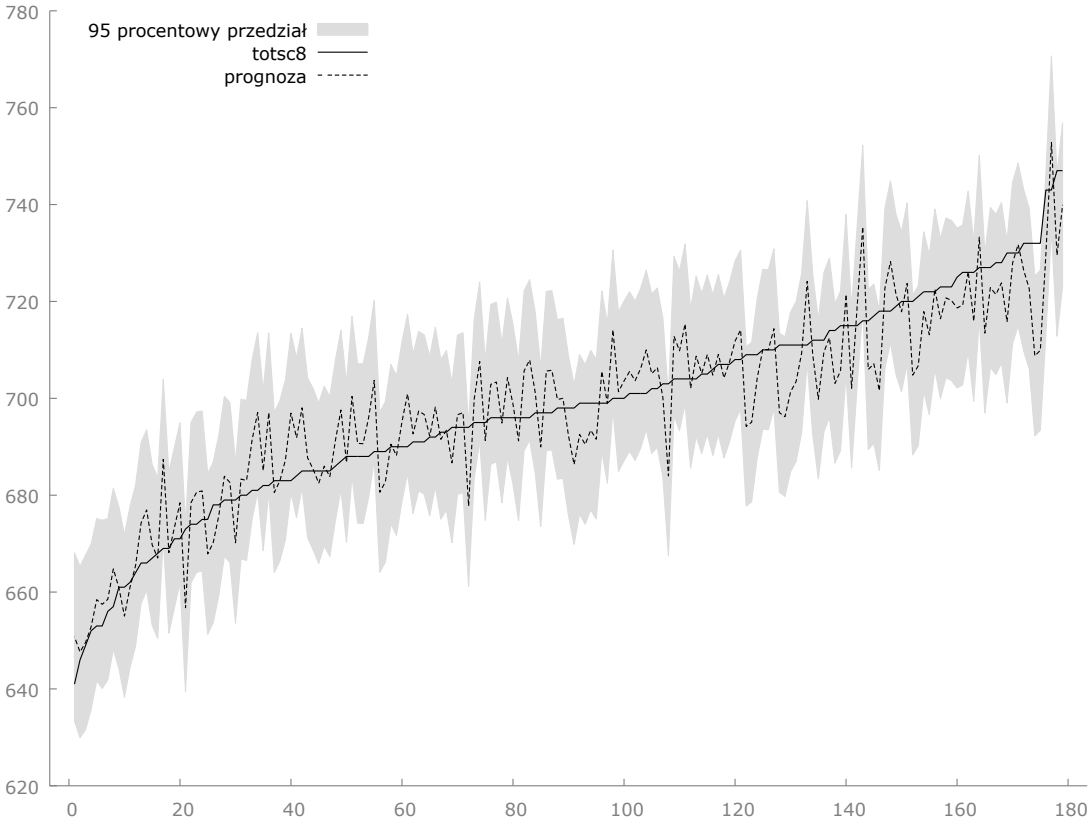
  Wzrost procenta osób uprawnionych do lunchu w cenie obniżonej lub bezpłatnego o jeden punkt procentowy powoduje zmiane wyniku testów ośmioklasistów o około $-1{,}14542 + 2 * 0{,}0059\ lnchpct$, ceteris paribus.

  Jak widać wartość jaką wzrost zmiennej o jeden punkt procentowy wpływa na zmienną objaśnianą zależy od wartości zmiennej którą przyjmuje. Wraz z wiekszą wartością zmiennej lnchpct ma ona mniejszy wpływ na zmienną *totsc8*, widać to wyraźnie na poprzednim wykresie.

Warto również zwrócić uwage dla interpretacji, gdy zmienna *lnchpct* przyjmuje wartość średnią:

Wzrost procenta osób uprawnionych do lunchu w cenie obniżonej lub bezpłatnego o jeden punkt procentowy powoduje zaniżenie wyników testów ośmioklasistów o średnio 1 punkt ceteris paribus.

n. Predykcja wraz z 95% przedziałem ufności



Wykres 7. Prognoza punktowa wraz z 95% przedziałem ufności

Średni błąd predykcji	ME =	-6,2877e-014
Pierwiastek błędu średniokwadr.	RMSE =	8,2043
Średni błąd absolutny	MAE =	6,5028
Średni błąd procentowy	MPE =	-0,01368
Średni absolutny błąd procentowy	MAPE =	0,92747
Współczynnik Theila (w procentach)	U1 =	0,0058731
Udział obciążoności predykcji	UM =	0
Udział niedost. elastyczności	UR =	0
Udział niezgodności kierunku	UD =	1

Tabela 32. Błędy prognoz

Model średnio myli się o około 6.5 punktów. Przeprowadzmy również predykcje punktową dla wartości średnich:

Prognoza punktowa	696,691
Wariancja prognozy	69,4831
Błąd prognozy	8,33565
95% przedział ufności	<680,239; 713,142>

8. Podsumowanie

Celem projektu było wyznaczenie najważniejszych determinant wyników egzaminów uczniów 8 klasy w stanie Massachusetts oraz stworzenie modelu wyjaśniający te zależności.

Udało się stworzyć poprawny wielomianowy model liniowy względem parametrów, który spełnia wszystkie założenia poprawności modelu, posiadający wysoki współczynnik R^2 oraz niskimi błędami predykcji.

Wyniki pokazały, że główne zmienne objaśniające, tj. "lnchpct" (procent osób uprawnionych do lunchu w cenie obniżonej lub bezpłatnego) oraz "percap" (dochód na osobę), wykazywały silną korelację z wynikami egzaminów.

Pozostałe zmienne objaśniające, takie jak "regday" (wydatki na ucznia) i "speced" (procent uczniów ze specjalnymi potrzebami edukacyjnymi), wykazywały słabszą korelację z wynikami egzaminów.

Bibliografia

1. Skrypt do przedmiotu Ekonometria I, M. Rubaszek *et al.,* Szkoła Główna Handlowa w Warszawie, https://web.sgh.waw.pl/~mrubas/Econometrics/pdf/EI_TallPL.pdf
2. Reducing Structural Multicollinearity STAT 501, Eberly College of Science, https://online.stat.psu.edu/stat501/lesson/12/12.6, [dostęp: 30.06.2023]
3. Nonlinear relationships Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/stats2/l61.pdfm [dostęp: 30.06.2023]
4. WARTOŚCI RESZTOWE W PROCESIE REGRESJI, D. Ampuła, Wojskowy Instytut Techniczny Uzbrojenia