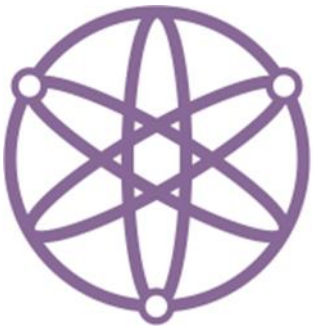


WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Ekonometria

Dokumentacja projektu

Gabriel Lichacz

Rzeszów, 2022

Spis treści

1. Wstęp

Dane zostały pobrane z portalu Federal Reserve Economic Data. Jako wartość y do modelu przyjmuję medianę cen sprzedaży domów w Stanach Zjednoczonych.

Median Sales Price of Houses Sold for the United States																	
Rok	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Kwota [$]	236550	243750	244950	229550	215650	222700	224900	244400	266225	285775	294150	305125	322425	325275	320250	336950	396800

tab. 1 Dane

Jako zmienne objaśniające wybieram:

- **Business Applications: Retail Trade in the United States** (liczba nowych numerów identyfikacyjnych pracodawcy w sektorze handlowym) – Zmienną wybrałem ze względu na korelowanie z liczbą powstających biznesów. Więcej miejsc pracy może oznaczać bardziej majętne społeczeństwo, co przekłada się na więcej nieruchomości.
- **Total Business Inventories** (całkowita wartość inwentarza firm) – Im wskaźnik wyższy, tym społeczeństwo jest bogatsze. Oznacza to, że inwentarze biznesów są więcej warte. Zmienną wybrałem z tych samych względów co poprzednią zmienną.
- **Natural Gas Consumption** (konsumpcja gazu ziemnego) – Gaz ziemny w nowoczesnym budownictwie stosowany jest do ogrzewania budynków. Im wyższe jego zużycie tym więcej takich budynków może istnieć.
- **Rail Passenger Miles** (ruch jednego pasażera na 1 milę) – Domy na ogół budowane są w pewnej odległości od centrów miast. Wysoki wskaźnik może wskazywać, że istnieje wiele domów (w założeniu poza głównymi aglomeracjami), a ich mieszkańcy muszą przemieszczać się choćby do pracy.
- **Homeownership Rate in the United States** (współczynnik posiadania nieruchomości) – Jest to liczba mieszkańców nieruchomości posiadanych na własność do wszystkich mieszkańców danego regionu. Im wyższy wskaźnik tym więcej osób posiada nieruchomość w której mieszka. Wysoka wartość mogłaby wskazywać na relatywnie niskie ceny rynkowe nieruchomości.
- **Homeowner Vacancy Rate in the United States** (współczynnik pustostanów i nieruchomości na sprzedaż do zamieszkałych nieruchomości) – Im więcej domów przeznaczonych jest na sprzedaż tym większa konkurencyjność. Klient jest w stanie wybrać dla niego odpowiedni i o niższej cenie.
- **Total Business Sales** (całkowita sprzedaż) – Można przyjąć, że im większe kwoty sprzedaży zostały osiągnięte, tym społeczeństwo jest zamożniejsze – tym samym stać je na posiadanie nieruchomości.
- **Passenger Transportation Services Index** (jak wiele pasażerów zostało przetransportowanych przez firmy transportowe) – Przypadek analogiczny do Rail Passenger Miles.
- **Total Construction Spending: Total Construction in the United States** (całkowite wydatki na budownictwo) – Im wyższe wydatki na budownictwo, tym więcej powstaje nowych nieruchomości. Większa liczba nieruchomości na rynku może spowodować spadek ich wartości.

- **All-Transactions House Price Index for the United States** (indeks cen transakcji kupna domów) – Bezpośrednio wpływający czynnik na medianę cen domów.
- **Consumer Price Index for All Urban Consumers: All Items in U.S. City Average** (indeks cen średnich zakupów w miastach) – Wskaźnik mówiący o zamożności społeczeństwa. Droższe zakupy oznaczają większą majętność.
- **University of Michigan: Inflation Expectation** (przewidywana inflacja) – Zbyt wysoka inflacja może wpływać negatywnie na wszystkie gałęzie rynku, co doprowadza do zubożenia społeczeństwa.
- **Producer Price Index by Industry: Concrete Block and Brick Manufacturing: Concrete Brick** (indeks produkcji cegieł) – są to podstawowe budulce.
- **Producer Price Index by Industry: Cement and Concrete Product Manufacturing** (indeks produkcji cementu i betonu) – są to podstawowe budulce.

2. Odrzucenie danych

2.1. Wstępne odrzucenie

Na początku obliczam współczynnik zmienności i współczynnik korelacji Pearsona. Przyjmuję jako wartość krytyczną współczynnika zmienności 10%, a współczynnika korelacji Pearsona 70%.

```
fs_table <- data.frame(c(), c())
for (i in 1:ncol(dane_x)) {
  fs_table[1,i] <- sd(dane_x[,i])/mean(dane_x[,i])
  fs_table[2,i] <- cor(dane_x[,i], dane_y,  method = "pearson")
}
row.names(fs_table) <- c("wsp_zmienn", "pearson_korr")
colnames(fs_table) <- colnames(dane_x)
```

W ten sposób odrzucam zmienne objaśniające, które nie spełniają powyższych kryteriów:

- Rail Passenger Miles – zbyt mała korelacja
- Homeownership Rate in the United States – zbyt mała korelacja i zmienność
- Passenger Transportation Services Index – zbyt mała korelacja
- Homeownership Rate for the United States – zbyt mała korelacja i zmienność
- Consumer Price Index for All Urban Consumers: All Items in U.S. City Average – za mały współczynnik zmienności
- University of Michigan: Inflation Expectation – zbyt mała korelacja

Wybieram do dalszej analizy:

- Retail Trade in the United States
- Total Business Inventories
- Natural Gas Consumption
- Homeowner Vacancy Rate in the United States
- Total Business Sales
- Total Construction Spending: Total Construction in the United States
- All-Transactions House Price Index for the United States
- Producer Price Index by Industry: Concrete Block and Brick Manufacturing: Concrete Brick
- Producer Price Index by Industry: Cement and Concrete Product Manufacturing

Skrypt odrzucający zmienne:

```
dane_w <- data.frame(c(1:17))
licznik <- 1
for (i in 1:ncol(fs_table)) {
  if (abs(fs_table[1,i]) > 0.10 && abs(fs_table[2,i]) > 0.70) {
    dane_w[,licznik] <- dane_x[,i]
    colnames(dane_w)[licznik] <- colnames(dane_x)[i]
    licznik <- licznik + 1
  } else {
    cat("odrzucam",  colnames(dane_x)[i], "\n")
  }
}
```

2.2. Metoda Hellwiga

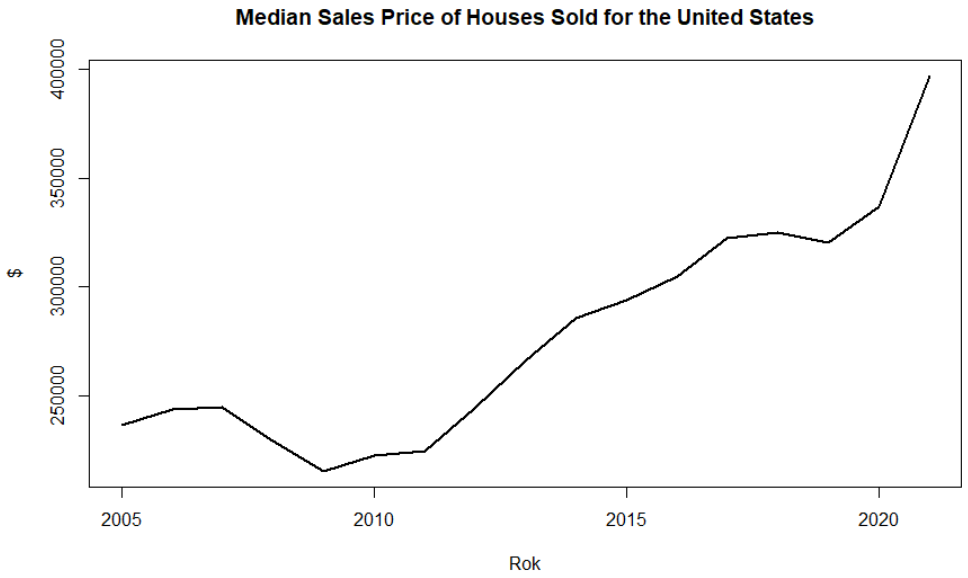
Stosuję metodę Hellwiga do dalszego odrzucenia zbędnych zmiennych. Skrypt zwraca 15 najlepszych kombinacji zmiennych do wybrania. Sprawdzam trzy pierwsze.

	Retail Trade in the United States	Total Business Inventories	Natural Gas Consumption	Homeowner Vacancy Rate in the United States	Total Business Sales	Total Construction Spending: Total Construction in the United States	All-Transactions House Price Index for the United States	Producer Price Index by Industry: Concrete Block and Brick Manufacturing: Concrete Brick	Producer Price Index by Industry: Cement and Concrete Product Manufacturing	hellwigP
1	0.0000000	0.0000000	0.0000000	0.2439407	0.2484068	0.2352260	0.0000000	0.0000000	0.2443388	0.9719123
2	0.0000000	0.0000000	0.0000000	0.3214631	0.3385701	0.3093474	0.0000000	0.0000000	0.0000000	0.9693806
3	0.0000000	0.2020721	0.0000000	0.1987945	0.2029966	0.1834073	0.1818076	0.0000000	0.0000000	0.9690780
4	0.0000000	0.2428342	0.0000000	0.2424792	0.2445905	0.2389792	0.0000000	0.0000000	0.0000000	0.9688832
5	0.0000000	0.1648729	0.0000000	0.1661433	0.1667153	0.1545364	0.1518039	0.0000000	0.1622101	0.9662819
6	0.0000000	0.2434089	0.0000000	0.2437315	0.2417632	0.0000000	0.2366473	0.0000000	0.0000000	0.9655508
7	0.0000000	0.1910371	0.0000000	0.1955935	0.1937788	0.1921937	0.0000000	0.0000000	0.1919503	0.9645535
8	0.1752095	0.2010044	0.0000000	0.1961293	0.1996736	0.1919188	0.0000000	0.0000000	0.0000000	0.9639356
9	0.0000000	0.0000000	0.0000000	0.1997757	0.2056183	0.1811885	0.1772546	0.0000000	0.1981044	0.9619415
10	0.0000000	0.0000000	0.0000000	0.3236676	0.3331767	0.0000000	0.3048913	0.0000000	0.0000000	0.9617357
11	0.0000000	0.0000000	0.0000000	0.2452081	0.2454911	0.0000000	0.2289910	0.0000000	0.2410734	0.9607636
12	0.0000000	0.1643342	0.1519484	0.1650457	0.1662255	0.1574910	0.1555830	0.0000000	0.0000000	0.9606279
13	0.0000000	0.0000000	0.0000000	0.2445489	0.2490814	0.2392648	0.0000000	0.2275328	0.0000000	0.9604279
14	0.1452819	0.1641614	0.0000000	0.1642776	0.1644675	0.1605354	0.0000000	0.0000000	0.1613685	0.9600922
15	0.0000000	0.1657455	0.0000000	0.1664252	0.1670189	0.1562694	0.1524670	0.1514177	0.0000000	0.9593438

tab. 2 Wyniki analizy metodą Hellwiga

```
func_hellwig <- function(dane) {
  N <- length(dane) # oblicza liczbe zmiennych objasniajacych
  M <- 2^N-1 # oblicza ilosc kombinacji 0-1
  zm_obj <- dane[,1:N] # macierz zmiennych objasniajacych
  r <- cor(zm_obj) # macierz korelacji miedzy zmiennymi
  r <- as.matrix(abs(r)) # wartosci bezwzgledne i macierz
  R <- cor(zm_obj, dane_y) # wektor korelacji Y z kazda ze zmiennych
  R <- as.vector(R)
  tab <- as.matrix(expand.grid(rep(list(0:1), N)))[-1,] # macierz 0-1 kombinacji
  colnames(tab) <- colnames(dane_w)[1:N]
  wyniki <- matrix(0,M,N) # macierz 0 na wyniki czastkowe pojemnosci
  colnames(wyniki) <- colnames(dane_w)[1:N]

  for (i in 1:M) {
    for (j in 1:N) {
      if (tab[i,j] != 0) {
        wyniki[i,j] <- (R[j]^2)/(tab[i,]%*%(as.vector(r[,j])))}}}
  maks <- which.max(rowSums(wyniki))

  wynikiS <- cbind(wyniki,0)
  wynikiS[,(N+1)] <- rowSums(wyniki)
  nazwy <- colnames(as.data.frame(wyniki))
  colnames(wynikiS) <- c(nazwy, "hellwigP")
  ind <- order(wynikiS[,(N+1)], decreasing = TRUE)[1:15] # zwroci 15 najlepszych
  najlepsze15 <- wynikiS[ind,]

  return(as.data.frame(najlepsze15))
}
hellwig_wynik <- func_hellwig(dane_w)
```

3. Budowa modelu



rys. 3-1 Szereg czasowy – zmienna y

Sprawdzam 3 scenariusze doboru zmiennych otrzymanych metodą Hellwiga.

3.1. Wskaźniki w podsumowaniu modelu

3.1.1. Residuals

Wartości residuals powinny być niewielkie – to reszty między wartościami modelu a wartościami rzeczywistymi.

3.1.2. Coefficients

Estimate – wagi dla poszczególnych zmiennych.

PR(>|t|) – wartość p (im mniejsza tym lepiej), np. p-value=0.65, tzn. 65% szans, że dana wartość modelu jest błędna.

R^2 – żeby model był dopasowany dobrze, wartość ta powinna być bliska liczbie 1.

3.1.3. Performance Measures

Residual standard error – odchylenie standardowe reszt, im jest mniejsze tym model lepiej oddaje zjawisko opisywane.

Multiple R-squared – R^2 powinno być bliskie liczbie 1.

F-statistic – test czy choć jedna ze zmiennych jest zmienną znaczącą. Wartość p powinna być mniejsza od 0.05.

3.2. Zestawy zmiennych

3.2.1. Zestaw nr 1

Zmienne użyte w modelu:

- Homeowner Vacancy Rate in the United States
- Total Business Sales
- Total Construction Spending: Total Construction in the United States
- Producer Price Index by Industry: Cement and Concrete Product Manufacturing

```
Residuals:
   Min     1Q Median     3Q    Max
-16184  -7370   2216   5773  10707

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.019e+04  4.820e+04   0.834  0.42061
dane_h1[, 1] -2.010e+04  8.164e+03  -2.462  0.02994 *
dane_h1[, 2]  1.197e-01  3.326e-02   3.601  0.00364 **
dane_h1[, 3]  7.391e-02  1.723e-02   4.290  0.00105 **
dane_h1[, 4]  2.835e+02  3.110e+02   0.911  0.38008
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9195 on 12 degrees of freedom
Multiple R-squared:  0.9758,    Adjusted R-squared:  0.9678
F-statistic: 121.2 on 4 and 12 DF,  p-value: 1.364e-09
```

rys. 3-2 Scenariusz nr 1

Wartość p dla zmiennej *Producer Price Index by Industry: Cement and Concrete Product Manufacturing* wynosi aż 0.38. Jest to powyżej normy i zaburza dokładność modelu. Optymalnym krokiem byłoby zbudowanie modelu bez tej zmiennej.

3.2.2. Zestaw nr 2

Zmienne użyte w modelu:

- Homeowner Vacancy Rate in the United States
- Total Business Sales
- Total Construction Spending: Total Construction in the United States

```
Residuals:
   Min     1Q Median     3Q    Max
-14926  -7016    701   7982  11350

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.867e+04  4.698e+04   1.036 0.319140
dane_h2[, 1] -2.077e+04  8.078e+03  -2.571 0.023263 *
dane_h2[, 2]  1.415e-01  2.301e-02   6.148  3.5e-05 ***
dane_h2[, 3]  8.062e-02  1.548e-02   5.210 0.000168 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9135 on 13 degrees of freedom
Multiple R-squared:  0.9742,    Adjusted R-squared:  0.9682
F-statistic: 163.4 on 3 and 13 DF,  p-value: 1.435e-10
```

rys. 3-3 Scenariusz nr 2

Model ten to w zasadzie model nr 1 z odrzuconą czwartą zmienną. Wszystkie wskaźniki są w porządku. Reszty są mniejsze niż w przypadku pozostałych modeli, wartości p są bardzo małe, R^2 bliskie 1, a współczynnik F wskazuje, że co najmniej jedna zmienna jest znacząca. Zmienną o najwyższej wadze jest *Total Business*, drugą *Sales Total Construction Spending: Total Construction in the United State*, a *Homeowner Vacancy Rate in the United States* jest zmienną o najniższej wadze. Model został wybrany do dalszej analizy.

3.2.3. Zestaw nr 3

Zmienne użyte w modelu:

- Total Business Inventories
- Homeowner Vacancy Rate in the United States
- Total Business Sales
- Total Construction Spending: Total Construction in the United States
- All-Transactions House Price Index for the United States

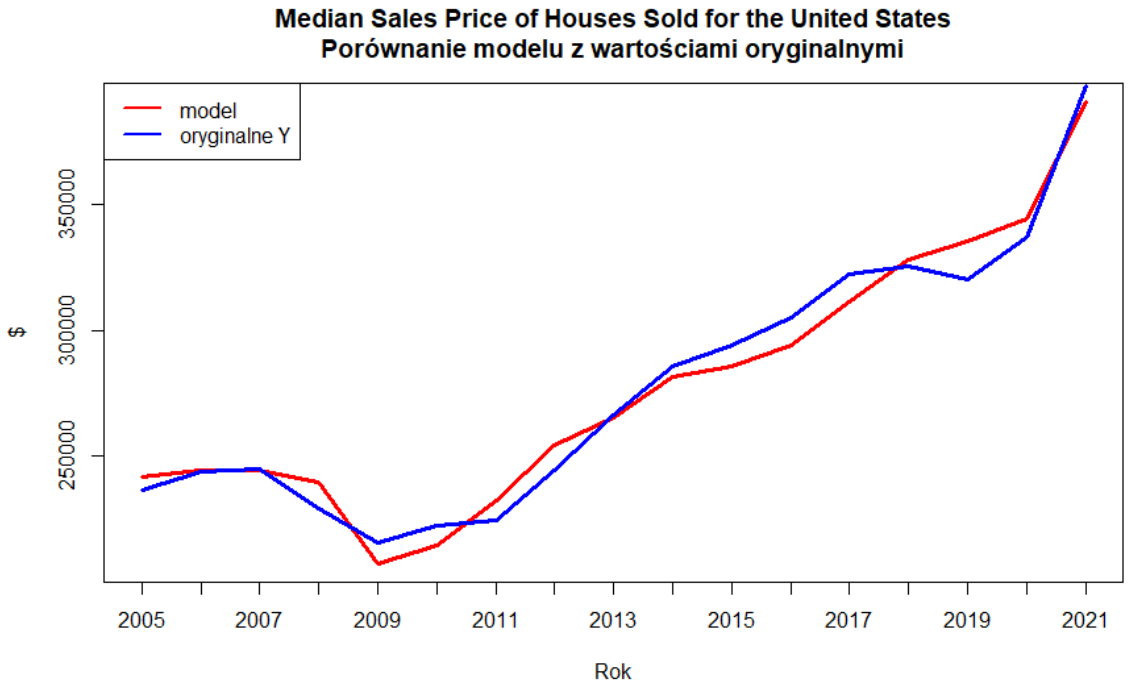
```
Residuals:
     Min       1Q   Median       3Q      Max
-15861.2  -7278.5    933.4   7591.4  10929.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.695e+04  5.186e+04   0.905   0.3846
dane_h3[, 1]  9.936e-03  3.875e-02   0.256   0.8023
dane_h3[, 2] -2.028e+04  8.976e+03  -2.260   0.0451 *
dane_h3[, 3]  1.292e-01  5.733e-02   2.254   0.0455 *
dane_h3[, 4]  7.807e-02  4.462e-02   1.750   0.1080
dane_h3[, 5]  6.367e+00  1.755e+02   0.036   0.9717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9898 on 11 degrees of freedom
Multiple R-squared:  0.9743,    Adjusted R-squared:  0.9627
F-statistic: 83.53 on 5 and 11 DF,  p-value: 2.272e-08
```

rys. 3-4 Scenariusz nr 3

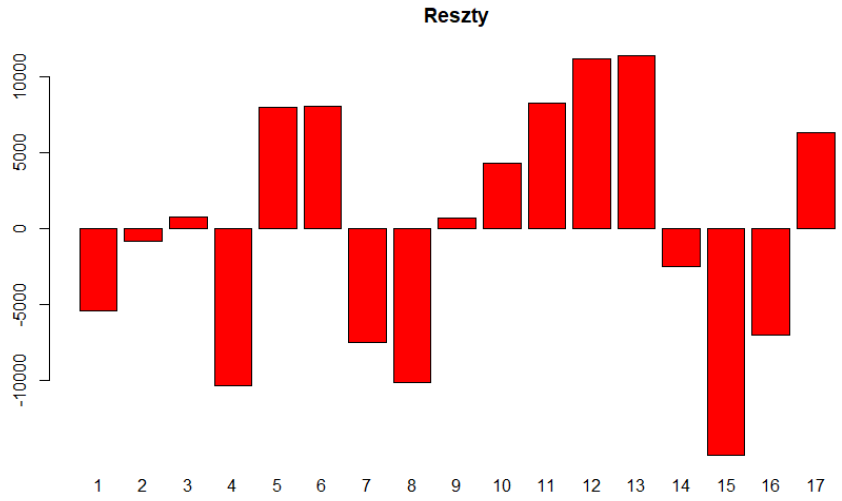
Model zwraca niskie reszty, przynajmniej jedna zmienna jest znacząca. Wartości p dla poszczególnych zmiennych jednak są zbyt wysokie co mówi o małym znaczeniu niektórych z nich. Model orzucam.

3.3. Wybrany model

Model dość dobrze oddaje kształt krzywej zmiennej badanej. W kilku miejscach następuje przeszacowanie lub niedoszacowanie, lecz nie jest ono wielkie.



rys. 3-5 Porównanie modelu ze zmienną y



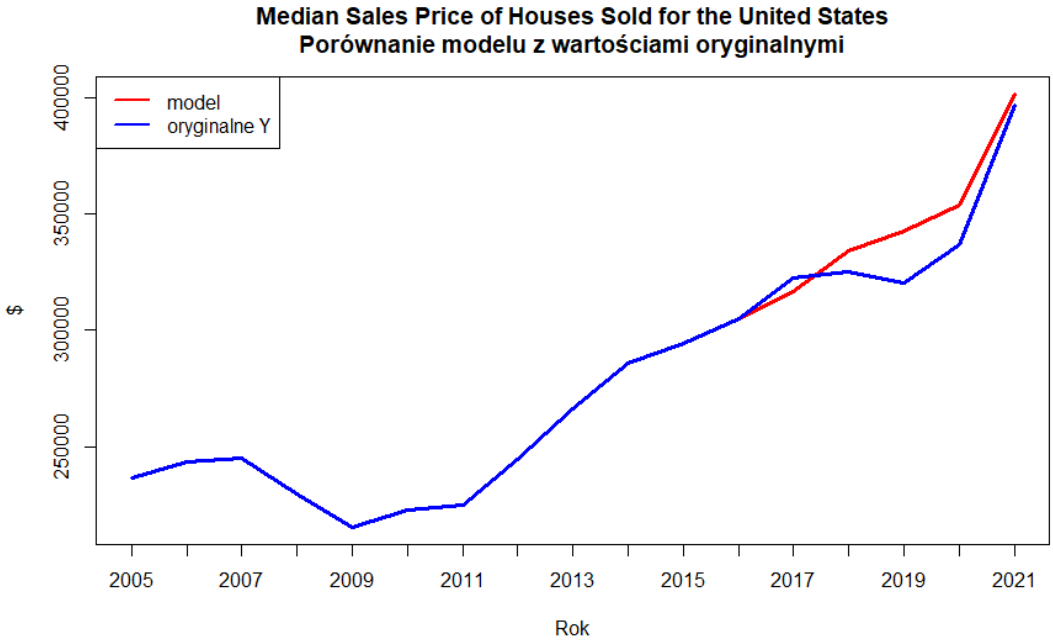
rys. 3-6 Reszty modelu dla poszczególnych lat

4. Test modelu

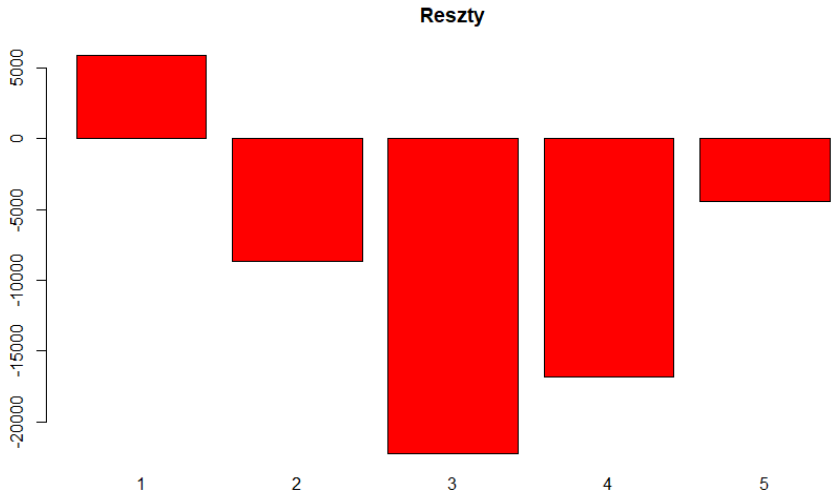
Zbiór danych wyjściowych oraz wejściowych zostaje podzielony na zbiory testowe i uczące. Następnie dla zbiorów uczących zostaje stworzony wybrany wyżej model. Test napisany jest jako funkcja przyjmująca długości zbioru uczącego i testowego. Dzięki temu można zaobserwować zmianę dokładności modelu w zależności od tego ile danych przyjmie.

4.1.1. Zbiór uczący to 70% danych oryginalnych

Model całkiem dobrze oddał kształt badanego szeregu czasowego, nie licząc spadku cen w latach 2018-2020. W tym miejscu nastąpiło przeszacowanie – widać niewielki spadek, lecz jest on zbyt mały. Końcowy wzrost jest zobrazowany bardzo dobrze co widać po wielkości reszt.



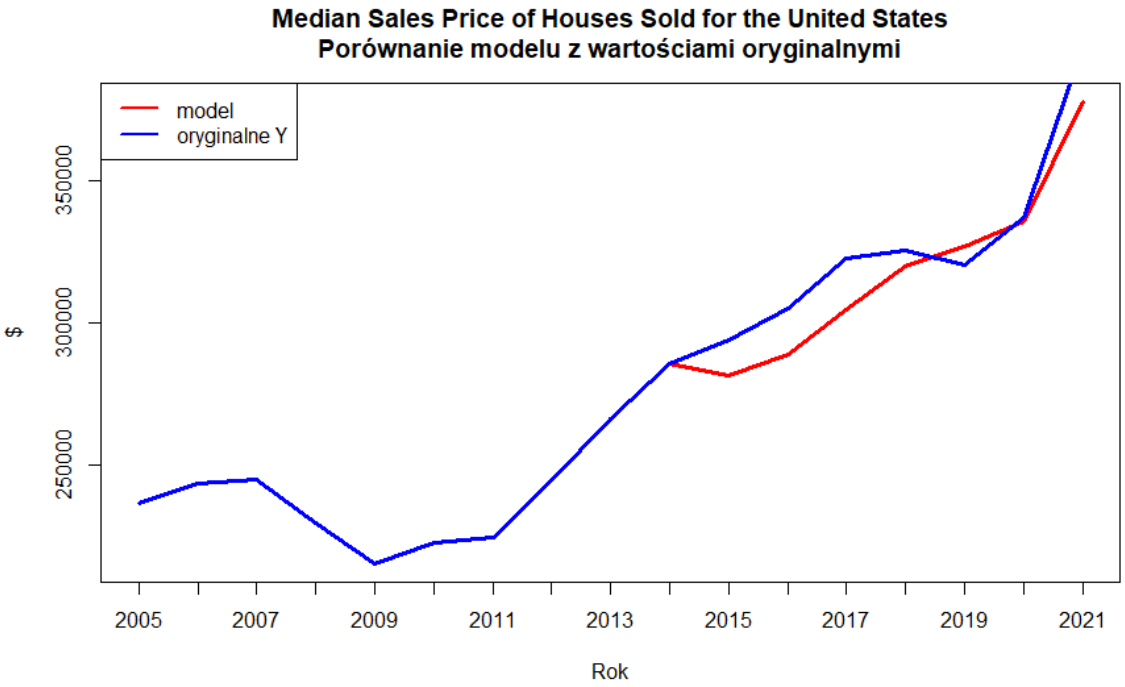
rys. 4-1 Model dla 70% danych uczących



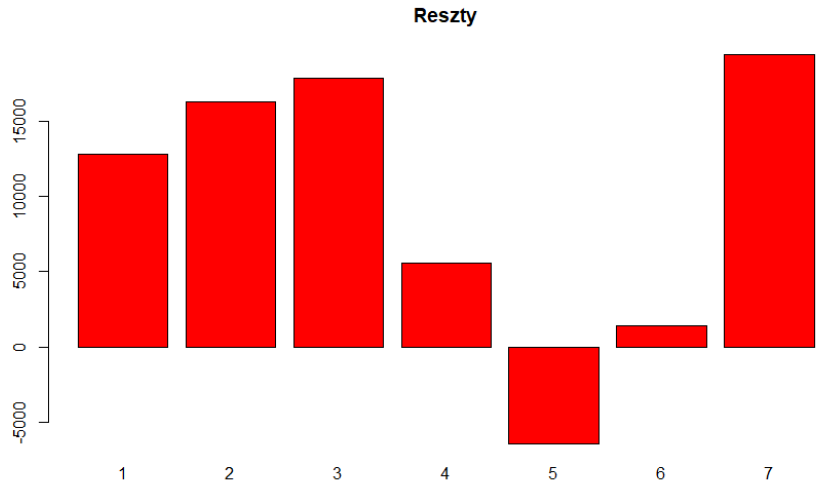
rys. 4-2 Reszty modelu dla 70% danych uczących

4.1.2. Zbiór uczący to 60% danych oryginalnych

Model nie poradził sobie źle – ceny w latach 2018-2020 przewidziane zostały zadowalająco. Nie udało się uchwycić dość stałego wzrostu w latach 2014-2018 – wartości modelu od początku dość gwałtownie spadły. Również w roku 2021 nastąpiło niedoszacowanie.



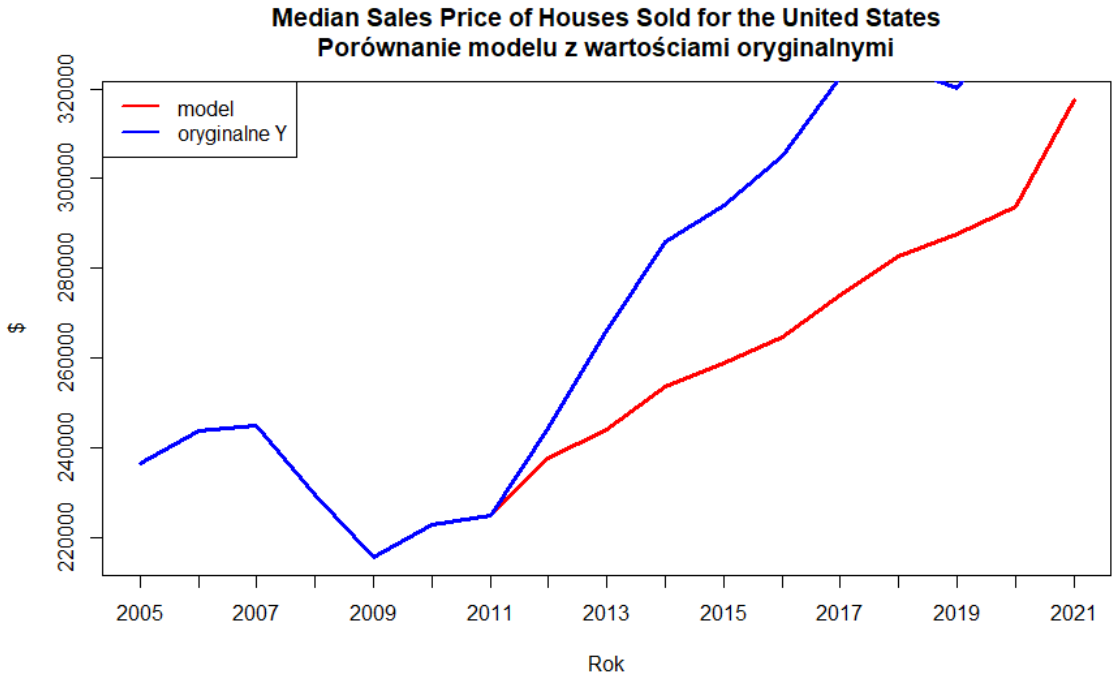
rys. 4-3 Model dla 60% danych uczących



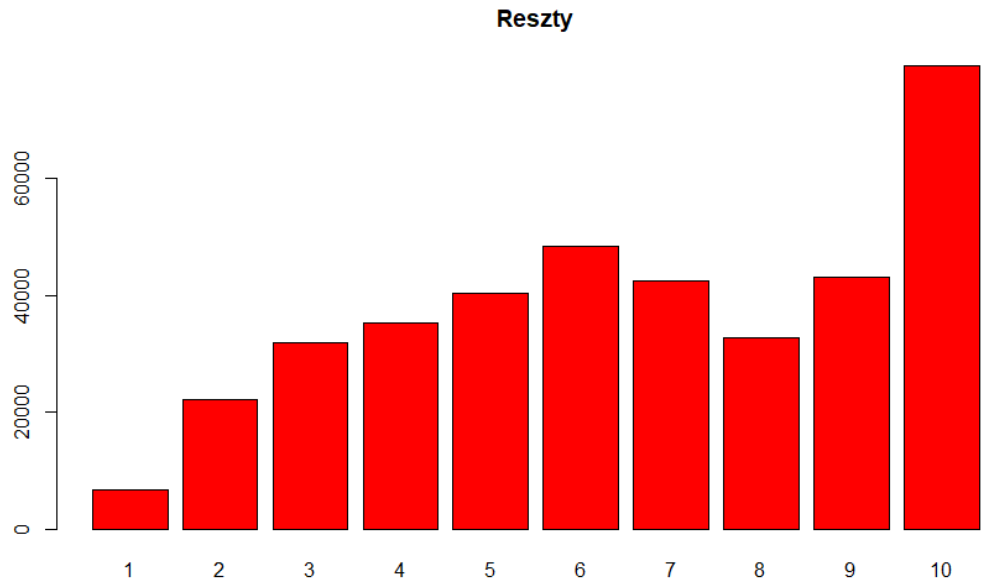
rys. 4-4 Reszty modelu dla 60% danych uczących

4.1.3. Zbiór uczący to 40% danych oryginalnych

Jak widać 40% danych to zbyt mało żeby stworzony model był w stanie odpowiednio przewidzieć zmianę cen w kolejnych latach. W każdym miejscu modelu nastąpiło niedoszacowanie, a reszty są bardzo duże.



rys. 4-5 Model dla 40% danych uczących



rys. 4-6 Reszty modelu dla 40% danych uczących

5. Podsumowanie

Stworzony model regresji liniowej całkiem dobrze oddaje kształt krzywej zmiennej wyjściowej. Reszty nie są duże poza nielicznymi przeszacowaniami. Podczas testów model najlepiej poradził sobie przyjmując 70% danych wejściowych jako dane uczące.

6. Spis ilustracji

7. Spis tabel

8. Źródła

[1] https://fred.stlouisfed.org/ dostęp 20.05.2022

[2] https://web.sgh.waw.pl/~mrubas/Econometrics/pdf/EI_TallPL.pdf dostęp 20.05.2022

[3] Brunon R. Górecki „Ekonometria podstawy teorii i praktyki” Warszawa 2010