AKADEMIA GÓRNICZO-HUTNICZA
IM.STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Analiza wybranych czynników wpływających na ceny domów, rok 2008.

Ekonometria

Imię i Nazwisko: Anna Pietryka

Kraków, rok akademicki 2020/2021

Spis treści

1. Wstęp i cel projektu

W 2008 roku w Stanach Zjednoczonych wybuchł kryzys gospodarczy. Jedną z przyczyn wybuchu kryzysu była sytuacja na rynku nieruchomości w USA. Podniesienie stóp procentowych spowodowało znaczne zwiększenie obciążeń odsetkowych, przy jednoczesnym zmniejszeniu atrakcyjności lokowania kapitału w nieruchomości. Banki, zajmując hipoteki i próbując sprzedać nieruchomości, przyspieszyły proces obniżki ich cen.

Poniższe badanie abstrahuje od czynników związanych z sytuacją na runku finansowym w tamtym okresie. Celem poniższej pracy jest wyłonienie czynników technicznych wpływających na wysokość ceny sprzedanych domów w 2008 roku w USA. Zostaną zatem rozważone parametry takie jak powierzchnia sprzedanego domu, ilość pokoi czy wiek budynku.

1.1 Hipotezy

Przed przystąpieniem do analizy sformułowano hipotezy dotyczące cen domów, które wynikają z powszechnych przekonań. W dalszej części pracy zostaną one zweryfikowane.

Hipotezy:

1. Im większa powierzchnia domu w stopach kwadratowych tym wyższa cena.
2. Im więcej lat upłynęło od ostatniego remontu domu tym niższa cena.
3. Im więcej lat upłynęło od wybudowania domu tym niższa cena.
4. Istoty wpływ na cenę sprzedaży domu ma powierzchnia garażu.
5. Obecność klimatyzacji centralnej w domu ma istotny wpływ na cenę.

2. Zagadnienia teoretyczne

2.1 Metoda najmniejszych kwadratów

Metoda najmniejszych kwadratów przyjmuje następujące złożenia:

- Zmienne objaśniające są nielosowe i nieskorelowane ze składnikiem losowym
- $r(X) = k + 1 \leq n$,
- wartość oczekiwana składnika losowego wynosi 0,
- $Var(\varepsilon) = D^2(\varepsilon) = \sigma^2 I, \sigma^2 < \infty$

Wówczas wektor oszacowań parametrów modelu (uzyskany przez minimalizację sumy kwadratów reszt) ma postać następującego iloczynu macierzy:

$$a = (XX^T)^{-1}X^T Y,$$

gdzie a jest wektorem oszacowań parametrów modelu, X jest macierzą zmiennych objaśniających, a Y wektorem wartości zmiennej objaśnianej. Jeżeli wszystkie wcześniej opisane założenia są spełnione to uzyskane estymatory są liniowe, nieobciążone, zgodne i najefektywniejsze w klasie estymatorów nieobciążonych[1].

[1] Wybrane metody estymacji i weryfikacji jednorównaniowych modeli regresji", Adam Szulc, wrzesień 2018, data dostępu: 27.05.2021

2.2 Istotność pojedynczej zmiennej objaśniającej – Test t-Studenta

Test t-Studenta został użyty do badania wpływu i-tej zmiennej objaśniającej na zmienną objaśnianą. Test przyjmuje następujący zestaw hipotez:

H_0 : $\alpha_i = 0$ (zmienna X_i jest nieistotna dla rozpatrywanego modelu)

H_1 : $\alpha_i \neq 0$ (zmienna X_i ma statystycznie istotny wpływ na zmienną objaśnianą)

Aby przeprowadzić test obliczana jest statystyka t :

$$t = \frac{|a_j|}{S_{a_j}},$$

gdzie a_j jest szacunkiem parametru α_j; S_{a_j} jest średnim błędem szacunku parametru α_j.

Statystyka ma rozkład t-Studenta o r = n – (k + 1) stopniach swobody. Jeżeli $|t| > t_{\alpha,r}$, gdzie $t_{\alpha,r}$ jest wartością odczytaną z tablic rozkładu t-Studenta na poziomie istotności równym α, to odrzucamy H_0 na rzecz H_1[2]. Podczas estymowania modelu zmienne będą uznawane jako istotne już na poziomie α = 10%. Test t-Studenta w programie Gretl wykonuje się automatycznie po estymacji modelu MNK. Wyświetlane są tylko wartości p-value odpowiednio przy każdej zmiennej objaśniającej.

2.3 Test Doornika-Hansena – rozkład reszt modelu

Sprawdzenie normalność rozkładu reszt modelu jest niezbędne do prawidłowej interpretacji współczynnika determinacji oraz jest jednym z założeń, które musi zostać spełnione, aby estymować model metodą MNK. Test Doornika-Hansena przyjmuje następujący zestaw hipotez:

H_0 : Dystrybuanta empiryczna posiada rozkład normalny.

H_1 : Dystrybuanta empiryczna nie posiada rozkład normalny.

Statystyka Doornika-Hansena ma następującą postać:

$$DH = z_1^2 + z_2^2$$

gdzie z_1^2 oraz z_2^2 są odpowiednio: transformowanym współczynnikiem skośności oraz transformowanym współczynnikiem kurtozy. Jeżeli składniki losowe mają rozkład normalny, wtedy statyka DH ma rozkład χ^2 o dwóch stopniach swobody. Jeżeli $DH > \chi_\alpha^2(2)$ to odrzucamy H_0, czyli reszty modelu nie mają rozkładu normlanego. Program Gretl oblicza wartość statystyki Doornika_Hansena wraz z odpowiadającej jej poziomem p-value[3].

2.4 Centralne Twierdzenie Graniczne

W przypadku, gdy na podstawie testu Doornika-Hansena odrzucona zostanie hipoteza mówiąca o normalności rozkładu reszt modelu, a próbka użyta do estymacji parametrów posiada dostatecznie dużo obserwacji, można zastosować Centralne Twierdzenie Graniczne.

[2] „Wybrane metody estymacji i weryfikacji jednorównaniowych modeli regresji”, Adam Szulc, wrzesień 2018, data dostępu: 27.05.2021

[3] „Weryfikacja modelu ekonometrycznego -teoria”, Marta Chylińska, data dostępu: 27.05.2021

Sformowanie twierdzenia:

Centralne Twierdzenie Graniczne to twierdzenie matematyczne mówiące, że jeśli X_i są niezależnymi zmiennymi losowymi o jednakowym rozkładzie, takiej samej wartości oczekiwanej μ oraz dodatniej i skończonej wariancji σ^2, to zmienna losowa postaci:

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}}$$

Zbiega według rozkładu do standardowego rozkładu normalnego, gdy n rośnie do nieskończoności[4].

2.5 Metoda regresji krokowej w tył

Metoda służąca do wyboru najlepszego zestawu zmiennych objaśniających do modelu ekonometrycznego. Modelem wyjściowym dla Metody Regresji Krokowej w Tył jest model składający się ze wszystkich potencjalnych zmiennych objaśniających. W kolejnych krokach obliczona zostaje wartość statystyki t-Studenta dla każdej zmienne objaśniającej:

- $t_j = \frac{|a_j|}{s_{a_j}}$,
- następnie obliczamy t_{min}= min ($|t_j|$) dla j = 0,1,…,n
- następnie odczytujemy z tablic wartość statystyki t-Studenta, na poziomie istotności równym α: $t^* := t_{n-(k+1),1-\frac{\alpha}{2}}$
- jeżeli $t^* > t_{min}$, to usuwamy z modelu zmienną realizującą minimum
- ponownie szacujemy model powtarzając wymienione kroki
- jeżeli $t^* \leq t_{min}$, to model przyjmujemy za ostateczny[5]

W programie Gretl decyzje o usunięci zmiennej z modelu podejmuje się na podstawie wartości p-valu obliczanej automatycznie dla testu t-Studenta. Kolejno z modelu są usuwane zmienne o największej wartości p-value, przekraczającej poziom istotności równy 10%.

2.6 Metoda Hellwiga

Metoda Hellwiga służy do wyboru optymalnego podzbioru zmiennych objaśniających. W obliczeniach wykorzystuje się współczynnik korelacji między poszczególnymi zmiennymi, w tym wektor współczynników korelacji między zmienną objaśnianą Y, a zmiennymi objaśniającymi X_1, X_2,…,X_n:

$$R_0 = \begin{matrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{matrix}$$

Macierz współczynników korelacji pomiędzy zmiennymi objaśniającymi X_1, X_2,…,X_n:

[4] „Centralne Twierdzenie Graniczne”, Wikipedia, data dostępu: 27.05.2021.

[5] „Specyfikacja i weryfikacja modelu liniowego, dobór zmiennych objaśniających” Barbara Jasiulis-Gołdyn, data dostępu: 27.05.2021

$$R = \begin{bmatrix} 1 & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{1n} & \cdots & 1 \end{bmatrix}$$

Metoda polega na znalezieniu kombinacji zmiennych objaśniających o największym integralnym wskaźniku pojemności informacyjnej. Dla każdej zmiennej zawartej w kombinacji definiuje się tzw. Indywidualną pojemność nośników informacji. Indywidulane wskaźniki pojemności informacyjnej dla rozpatrywanej kombinacji C_s są zdefiniowane następująco:

$$h_{sj} = \frac{r_j^2}{\sum_{i \in C_s} |r_{ij}|},$$

gdzie:

h_{sj} - indywidualna pojemność informacyjna zmiennej 'j' w C_s

r_j – wartość wektora korelacji R_0

r_{ij} – wartość z macierzy korelacji R

S – numer podzbioru zmiennych

j – numer zmiennej w kombinacji (j = 1,2…, m_k)

Po obliczeniu wartości indywidualnych pojemności nośników informacji oblicza się pojemności integralną kombinacji nośników informacji według wzoru:

$$H_s = \sum_{j \in C_s} h_{sj}$$

Wybiera się tę kombinację zmiennych objaśniających dla których integralna pojemność informacyjna jest największa[6].

2.7 Współczynnik determinacji R^2

Współczynnik determinacji jest miarą jakości dopasowanie modelu. Po zbudowaniu modelu ekonometrycznego pozwala zmierzyć w jakim stopniu model umożliwia objaśnienie zmienności zmiennej Y. Aby móc poprawnie interpretować współczynnik determinacji muszą być spełnione następujące założenia:

- w modelu musi występować wyraz wolny
- badane związki są liniowe
- rozkład reszt modelu jest normlany

Współczynnik determinacji jest obliczany w następujący sposób:

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \widehat{y_t})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

gdzie:

[6] „Zastosowanie metody Hellwiga do konstrukcji modelu ekonometrycznego dla stóp zwrotu z funduszy inwestycyjnych", Adam Kopiński, Dariusz Porębski, Lublin, 2014, data dostępu: 27.05.2021

y_t - wartości obserwowane

$\widehat{y_t}$ – wartości teoretyczne

$\bar{y}$ – wartość średnia

Współczynnik przyjmuje wartości od 0 do 1. Im wyższa wartość współczynnika tym lesze dopasowanie modelu do danych.

W praktyce do porównywania jakości dopasowania między dwoma modelami wykorzystuje się Skorygowany R^2. Współczynnik ten nie jest czuły na zmianę liczby zmiennych w modelu[7].

2.8 Współliniowość w modelu ekonometrycznym

Występowanie współliniowość w modelu może powodować zwiększenie wariancji estymatora MNK, a w efekcie zmniejszenie jego efektywności. Współliniowość może zostać zweryfikowana przy pomocy czynnika inflacji wariancji VIF (*ang. Variance Inflation Factor*). Dla każdej zmiennej objaśniającej konstruowany jest model x_j, w którym zmienną objaśnianą jest ona sama, a zmiennymi objaśniającymi pozostałe zmienne z wyjściowego zbioru regresorów, czyli:

$$\bigvee_{j \in J - \{j\}} x_j = \beta_0 + \beta_1 x_1 + \cdots + \beta_J x_J + \varepsilon$$

Dla każdego modelu jest obliczany współczynnik determinacji R^2, a następnie VIF_j:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Wartości VIF powyżej 10 wskazują na problem współliniowości[8].

2.9 Efekt katalizy

Możliwość otrzymania wysokiej wartości współczynnika determinacji, mimo, że charakter i rzeczywista siła powiązań zmiennych objaśniających nie uzasadniają tego wyniku nazywamy efektem katalizy. Efekt katalizy może mieć miejsce, gdy w modelu występują zmienne zwane katalizatorami. Dla regularnej pary korelacyjnej, zmienna X_i z pary (X_i, X_j) jest katalizatorem gdy:

$$r_{ij} < 0 \; lub \; r_{ij} > \frac{r_i}{r_j}$$

Badanie występowania efektu katalizy przeprowadza się przy pomocy badania miary nazywanej natężeniem efektu katalizy:

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

[7] „Wzór na R-kwadrat, współczynnik determinacji w modelu regresji liniowej”, Aukowiec.org, data dostępu: 27.05.2021

[8] „Weryfikacja liniowego modelu jednorównaniowego”, Jakub Mućk, data dostępu: 29.05.2021

gdzie H jest integralną pojemnością informacyjną zestawu zmiennych objaśniających modelu. W celu porównywania modeli określa się również miarę względnego natężenia efektu katalizy W_η:

$$W_\eta = \frac{\eta}{R^2} * 100\% \text{ [9]}.$$

2.10 Test F (Test Walda)

Test F posiada trzy zestawy hipotez i może być stosowany w trzech przypadkach. W pierwszym przypadku test może zostać wykorzystany do sprawdzania istotności współczynnika determinacji R^2. Przyjmuje się wtedy następujący zestaw hipotez:

H_0 : Współczynnik determinacji nie jest istotny.

H_1 : Współczynnik determinacji jest istotny.

W drugim przypadku test może zostać wykorzystany do zweryfikowania istotności wszystkich zmiennych objaśniających w modelu. Przyjmuje się wtedy następujący zestaw hipotez:

H_0: $\alpha_1 = \ \ldots \ = \alpha_k = 0$

H_1 : co najmniej jeden z α_j, j = k + 1, … k + m, jest różny od zera.

Przyjęcie hipotezy H_0 oznacza, że zmienne w modelu są nieistotne. W celu realizacji testu obliczamy następujące statystyki:

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

n – ilość zmiennych w modelu

k – ilość zmiennych objaśniających

Statystyka F ma rozkład F-Snedecora z r_1 = k oraz r_2 = n – (k + 1) stopniami swobody. Jeżeli F > F^* to odrzucamy H_0 na poziomie istotności równym α.

W trzecim przypadku test może zostać wykorzystany do badania istotności podzbioru zmiennych objaśniających. Odpowiada na pytanie jaki jest łączny efekt wprowadzenia do modelu dodatkowych zmiennych. Rozpatrujemy następujące modele:

Model $Y = \ \alpha_0 + \ \alpha_1 X_1 + \ \alpha_2 X_2 + \ \alpha_3 X_3 + \cdots + \ \alpha_k X_k \ + \ \varepsilon - model\ podstawowy$

Model $\quad Y = \ \alpha_0 + \ \alpha_1 X_1 + \ \alpha_2 X_2 + \ \alpha_3 X_3 + \cdots + \ \alpha_k X_k \ + \ \alpha_{k+1} X_{k+1} \ + \cdots + \alpha_{k+m} X_{k+m} \ + \ \varepsilon - \ model\ rozszerzony$

Przyjmuje się wtedy następujący zestaw hipotez:

H_0: $\alpha_{k+1} = \cdots = \alpha_{k+m} = 0$,

H_1 : co najmniej jeden z α_j, j = k + 1, … k + m, jest różny od zera.

[9] „Weryfikacja modelu”, data dostępu: 29.05.2021

Przyjęcie hipotezy H_0 oznacza, że rozszerzenie modelu jest niepotrzebne. W celu realizacji testu obliczamy następujące statystyki:

$$F = \frac{e^T e - r^T r}{r^T r} * \frac{n - (k+1) - m}{m}$$

e - wektor reszt modelu

n – ilość zmiennych w modelu

m -ilość dołożonych do modelu zmiennych

Statystyka F ma rozkład F-Snedecora z $r_1 = m$ oraz $r_2 = n - (k + 1) - m$ stopniami swobody. Jeżeli $F > F^*$ to odrzucamy H_0 na poziomie istotności równym α.

W programie Gretl przy estymacji modelu MNK wartość statystyki F-Snedecora oraz odpowiadająca jej wartość p-vale są obliczane automatycznie. Na tej podstawie możemy wnioskować o istotności współczynnika determinacji oraz istotności wszystkich zmiennych uwzględnionych w modelu. Weryfikacji podzbioru zmiennych objaśniających dokonuje się już samodzielnie.

2.11 Test Liczby Serii

Test Liczby Serii posiada dwa zestawy hipotez. Pierwszy zestaw hipotez służy do zweryfikowania czy zastosowany do analizy model regresji liniowej jest modelem o poprawnie dobranej postaci tzn. że rzeczywista regresja badanych zmiennych w populacji generalnej jest liniowa. Hipotezy przyjmują zatem postać:

$H_0 = E(Y|X = x) = \alpha x + \beta$

$H_1 = E(Y|X = x) \neq \alpha x + \beta$

Drugi zestaw hipotez pozwala określić czy próba została dobrana w sposób losowy. Serią nazywamy rząd wartości jednego znaku. Test przyjmuje następujący zestaw hipotez:

H_0: próba losowa,

H_1~H_0

Aby wykonać test postępujemy według punktów:

- Porządkujemy reszty w kolejności rosnącej jednej ze zmiennych objaśniających
- Zliczamy liczbę serii.
- Zliczamy liczbę reszt dodatnich i liczbę reszt ujemnych. (Gdybyśmy uzyskali resztę równą 0, ignorujemy ją w obliczeniach).
- Statystyką testową jest liczba serii:

  $S_1^* < S < S_2^*$, gdzie

  S – liczba serii
  S_1^* - wartość z rozkładu liczby serii dla liczby serii z symbolem „+”
  S_2^* - wartość z rozkładu liczby serii dla liczby serii z symbolem „-”

Jeżeli powyższa nierówność nie jest spełniona to odrzucamy hipotezę H_0 mówiącą o losowości próby[10].

2.12 Test Ramsey'a RESET

Test Ramsey'a RESET jest parametrycznym odpowiednikiem Testu Liczby Serii. Test służy do upewnienia się czy liniowa postać modelu została dobrze dobrana do zmiennej objaśnianej. W zależności od wybory testu i jego wariantu równanie pomocnicze przyjmuje różne postaci analityczne. W modelu mogą pojawić się dodatkowo kwadraty, sześciany zmiennej objaśnianej. Test posiada następujący zestaw hipotez:

H_0: Wybór postaci modelu jest prawidłowy,

H_1: Wybór postaci modelu nie jest prawidłowy.

W poniższym przykładzie przedstawiono model rozszerzony o kwadrat i sześcian zmiennej objaśnianej.

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_K x_{tK} + \xi_t; (t = 1, \ldots, T)$$

$$y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_K x_{tK} + \gamma_1 \hat{y}_t^{\,2} + \gamma_2 \hat{y}_t^{\,3} + \eta_t; (t = 1, \ldots, T)$$

H_0: $\gamma_1 + \gamma_2 = 0$ (Postać analityczna modelu jest poprawna)

H_1 : $\gamma_1 + \gamma_2 \neq 0$ (Postać analityczna modelu jest niepoprawna)

Statystyką testową jest statystyka F:

$F = \frac{\sum_{t=1}^{T} \xi_t^2 - \sum_{t=1}^{T} \eta_t^2}{2} * \frac{T-k-3}{\sum_{t=1}^{T} \eta_t^2}$, gdzie $F \sim F(2, T-k-3)$

Jeżeli $F^* \leq F$ to odrzucamy hipotezę zerową mówiącą o tym, że analityczna postać modelu jest dobrze dobrana[11].

2.13 Test Chowa

Szacując model regresji i wykorzystując go do prognozowania na pewnie przyszły okres zakładamy, że parametry są stałe w całym okresie estymacji oraz predykcji. Do weryfikacji hipotezy o stabilności parametrów wykorzystamy Test Chowa. Przeprowadzając test zakładamy, że mamy dwa niezależne zbiory danych o liczebności odpowiednio n_1 i n_2. Równania regresji mają postać:

$y = \alpha_1 + \beta_{11} x_1 + \beta_{12} x_2 + \ldots + \beta_{1k} x_k + \mu$ dla pierwszego zbioru danych,

$y = \alpha_2 + \beta_{21} x_1 + \beta_{22} x_2 + \ldots + \beta_{2k} x_k + \mu$ dla drugiego zbioru danych.

Test stabilności polega na weryfikacji hipotez:

$H_0 : \beta_{11} = \beta_{21}, \beta_{12} = \beta_{22}, \ldots, \beta_{1k} = \beta_{2k}, \alpha_1 = \alpha_2$

H_1: $Parametry\ modelu\ nie\ są\ stabilne$

[10] „Statystyka od podstaw" J.Jóźwiak, J.Podgórski, Warszawa 1998

[11] „Weryfikacja modelu ekonometrycznego – teoria", Marta Chylińska, da40,00ta dostępu: 27.05.2021

Statystyką wykorzystaną do testu jest statystyka F, która przyjmuje postać:

$$F = \frac{RSK-(RSK_I+RSK_{II})}{RSK_I+RSK_{II}} * \frac{n-2(k+1)}{k+1}$$, gdzie

$RSK_I \; i \; RSK_{II}$ są odpowiedni resztową sumą kwadratów dla n_1 i n_2.

Wówczas statystka F ma rozkład F-Snedecora o $r_1 = k + 1$ i $r_2 = n - 2(k + 1)$ stopniach swobody. Jeżeli $F^* < F$, to odrzucamy hipotezę zerową, mówiącą o stabilności parametrów modelu[12].

2.14 Test White'a

Test White'a również służy do padania homoskedastyczności wariancji składnika losowego. Do weryfikacji hipotez tworzymy nowy model regresji liniowej w którym zmienną objaśnianą jest kwadrat otrzymanych residuów, a zmiennymi objaśniającymi zmienne wyjściowe, ich kwadraty, oraz proste iloczyny drugiego rzędu, tzn. wektory zmiennych x_j, x_j^2 oraz $x_j x_l$ dla $j, l = 1, \dots, k$ gdzie$j \neq l$. Statystyka testowa wynosi:

$$LM_w = nR^2,$$

Gdzie n to wielkość próby, a R^2 to współczynnik dopasowania dodatkowego modelu. Zakładając brak homoskedastyczności w próbce statystyka W powinna mieć rozkład χ^2 o h stopniach swobody, gdzie h to ilość estymowanych współczynników w dodatkowym modelu[13].

2.15 Test Bruscha-Pagana

Test służy do weryfikacji stałości wariancji składnika losowego. Załóżmy że, $V(\mu_i) = \sigma_i^2$. Jeżeli istnieją zmienne $z_1, z_2, z_3, \dots., z_r$, które mają wpływ na wariancję składnika losowego w modelu oraz jeśli $\sigma_i^2 = f(\alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \dots + \alpha_r z_{ri})$, to test Breuscha-Pagana polega na weryfikacji hipotezy:

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0.$

$H_1: reszty \; są \; heteroskedastyczne$

Funkcja $f(.)$ może być funkcją dowolnego rodzaju. Niech:

$$\hat{\sigma}^2 = \sum \frac{\hat{\mu_i}^2}{n},$$

S_0 = regresyjna suma kwadratów z regresji $\hat{\mu_i}^2$ względem $z_1, z_2, z_3, \dots., z_r$, wówczas $\lambda = \frac{S_0}{2\,\sigma^4}$ ma asymptotyczny rozkład χ^2 z r stopniami swobody. Jeżeli $\lambda > \chi^2_{tab}$ to odrzucamy hipotezę zerową mówiącą o stałości wariancji składnika losowego[14].

2.16 Ważona MNK

Ważona Metoda Najmniejszych Kwadratów jest wykorzystywany w przypadku heteroskedastyczności składnika losowego. W przypadku heteroskedastyczności składnika losowego, macierz wariancjikowariancji jest następująca:

[12] „Ekonometria", G.S.Maddal, Wydawnictwo Naukowe PWN, Warszawa 2006, str. 209
[13] „Ekonometria", dr.Marcin Pitera, data dostępu: 31.05.2021
[14] „Ekonometria", G.S.Maddal, Wydawnictwo Naukowe PWN, Warszawa 2006

$$Var(\varepsilon) = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix},$$

a samą wariancję składnika losowego możemy zapisać jako:

$$\sigma_i^2 = \omega_i \sigma^2,$$

wtedy macierz wariancji-kowariancji można zapisać:

$$Var(\varepsilon) = \sigma^2 \Omega = \sigma^2 \begin{bmatrix} \omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_n \end{bmatrix}.$$

Stosując estymator $\beta^{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$

możemy zauważyć, że nieznana macierz Ω^{-1} to macierz zawierająca elementy ω_i na przekątnej (weights):

$$\Omega^{-1} = \begin{bmatrix} \frac{1}{\omega_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\omega_n} \end{bmatrix}.$$

Kluczowe jest zdefiniowanie macierz P, a więc macierzy transformacji GLS. Pamiętając, że $\Omega^{-1} = P^T P$ łatwo pokazać, że taką macierzą dla estymatora WLS można zapisać następująco:

$$P = \begin{bmatrix} \frac{1}{\sqrt{\omega_1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sqrt{\omega_n}} \end{bmatrix}.$$

Powyższe własności (macierz P) implikują następującą transformację danych:

$$y_{i*} = \frac{y_i}{\sqrt{\omega_i}},$$

a następnie możliwość zastosowania estymatora OLS dla następującego modelu:

$$y_* = X_* \beta + \eta_i.$$

W aplikacjach empirycznych zazwyczaj nie znamy wag, tj. $\sqrt{\omega_i}$. Trzeba wyznaczyć je samodzielnie[15].

2.17 Błędy prognozy

Błędy prognozy służą do oceny trafności prognozy. Pozwalają określić o ile zbudowany model się myli. Wyróżniamy następujące błędy prognozy:

$ME = \frac{1}{S} \sum_{t=1}^{S} (y_\Gamma - y_\Gamma^P)$ – pomaga ocenić przeciętne obciążenie prognozy

[15] „Metody Ekonometryczne, Estymator GLS (UMNK)", Jakub Mućk, data dostępu: 31.05.2021

$MAE = \frac{1}{S}\sum_{t=1}^{S}|y_\Gamma - y_\Gamma^P|$ - podaje, w jednostkach bezwzględnych, o ile średnio prognoza różni się od wartości rzeczywistej.

$RMSE = \sqrt{\frac{1}{S}\sum_{t=1}^{S}(y_\Gamma - y_\Gamma^P)^2}$ - podaje, w jednostkach bezwzględnych, o ile średnio prognoza różni się od wartości rzeczywistej, ale jest bardziej czuły na wartości skrajne.

$MAPE = \frac{1}{S}\sum_{t=1}^{S}\left|\frac{y_\Gamma - y_\Gamma^P}{y_\Gamma}\right| * 100$ - podaje o ile procent średnio prognoza różni się od wartości rzeczywistej[16]

3. Analiza danych

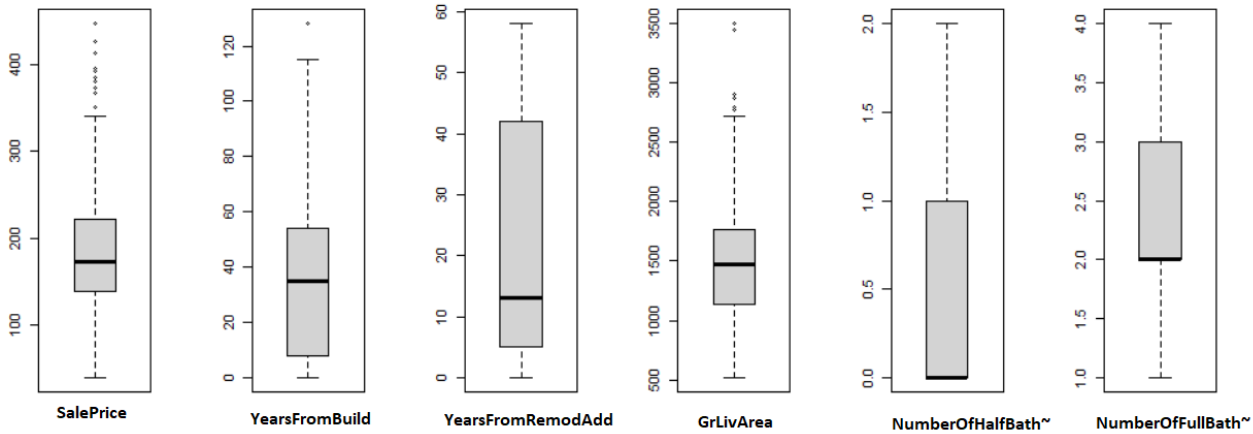
3.1 Opis danych i statystyki opisowe

Zbiór danych składa się ze zmiennej objaśnianej: *SalePrice* - ceny domów, oraz z 12 zmiennych objaśniających w tym jednej zmiennej binarnej: *YearsFromBuild* (ilość lat, która upłynęła od wybudowania domu), *YearsFromRemodAll* (ilość lat, która upłynęła od remontu domu), *CentralAir* – zmienna binarna (1 – występuje klimatyzacja, 0 – brak klimatyzacji), *GrLivArea* (powierzchnia domu w stopach kwadratowych), *NumberOfHalfBathroom* (ilość łazienek, w których znajduje się tylko toaleta), *NumberOfFullBathroom* (ilość pełnych łazienek w domu), *NumberOfKitchen* (ilość kuchni), *TotNumberOfRooms* (ilość wszystkich pokoi w domu), *NuberOfFireplace* (ilość kominków w domu), *GarageArea* (powierzchnia garażu w stopach kwadratowych).

Dane pochodzą ze strony Kaggle.com, zbioru „House Price Dataset" z pliku o nazwie „train.csv". Z pliku został wybrany wyżej opisany zbiór zmiennych objaśniających.

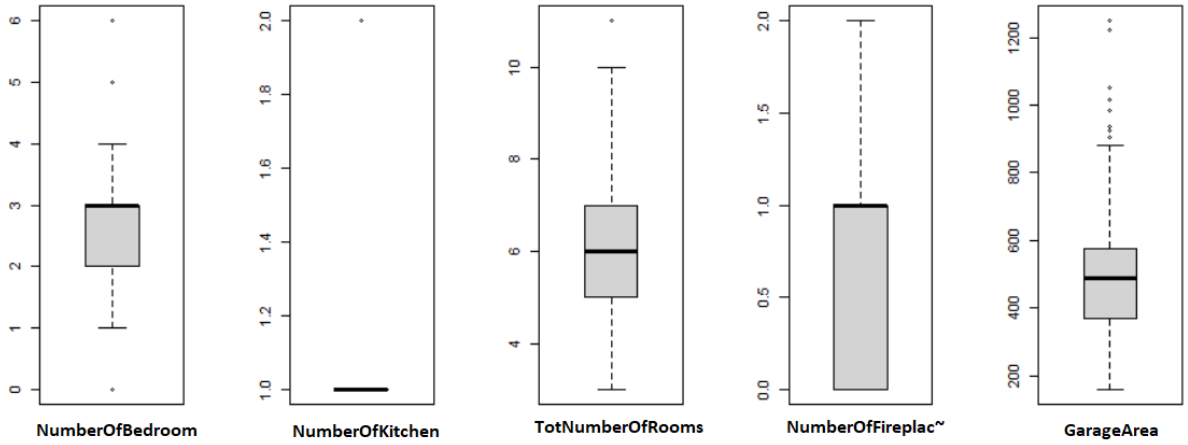
Braki danych i wartości odstające

W danych nie zaobserwowano wartości brakujących. Na zestawieniu wykresów pudełkowych poszczególnych zmiennych można zaobserwować wartości odstające.



Rysunek 1

[16] „Metody prognozowania: Jakość prognoz", Dr inż. Sebastian Skoczypiec, data dostępu: 27.05.2021



Rysunek 2

Zostało usunięte 5% największych wartości z *SalePrice*. Nadal pozostały wartości odstające, ale nie zostały usunięte ze względu zmniejszający się współczynnik zmienności, a także obawę przed utrata nadmiernej ilości informacji. Dodatkowo zostały usunięte dwie najbardziej ostające wartości z *LivGridArea*.

3.2 Podział zbioru na testowy i uczący

Zbiór danych składający się z 270 obserwacji został podzielony na zbiór testowy i zbiór uczący w stosunku: 90% obserwacji zbiór uczący, 10% obserwacji zbiór testowy.

3.3 Opis zmiennych

Zmienna objaśniana

SalePrice - wysokość ceny sprzedaży domu w USA w roku 2008 w tysiącach dolarów amerykańskich.

Statystyka	Wartość statystyki
Średnia	182,25
Mediana	175,00
Minimalna	324,00
Maksymalna	53,803
Odch. standardowe	0,63309
Skośność	0,63309
Kurtoza	0,050197

Tabela 1

Średnia cena sprzedaży domu 2008 roku w USA wynosiła 182,25 tys. dolarów z odchylaniem standardowym równym 0,63,31 tys. dolarów. Współczynnik skośności jest dodatni co wskazuje na to, że większa cześć obserwacji przyjmuje wartości wyższe od średniej. Kurtoza poniżej trzech mówi o silnej koncentracji danych wokół średniej.

Zmienne objaśniające

- *YearsFromBulid* – zmienna określająca ile lat upłynęło od wybudowania domu, do roku sprzedaży.

Statystyka	Wartość statystyki
Średnia	35,177
Mediana	35,00
Minimalna	0,000
Maksymalna	108,00
Odch. standardowe	27,63
Skośność	0,55210
Kurtoza	-0,58068

Tabela 2

Przeciętny wiek domu w chwili sprzedaży wnosi 35 lat z odchylaniem standardowym równym 27 lat. Wartość minimalna równa 0 sugeruje, że niektóre domy zostały sprzedane bezpośrednio po wybudowaniu. Współczynnik skośności jest dodatni, co wskazuje na to, że większa cześć obserwacji znajduje się powyżej średniej. Kurtoza poniżej 3 wskazuje na słabą koncentrację danych wokół średniej.

- *YearsFromRemodAdd* - zmienna określająca ile lat upłynęło od ostatniego remontu domu, do roku sprzedaży.

Statystyka	Wartość statystyki
Średnia	22,757
Mediana	14,00
Minimalna	0,00
Maksymalna	58,00
Odch. standardowe	19,811
Skośność	0,55643
Kurtoza	-1,2161

Tabela 3

Przeciętna ilość lat jaka upłynęła od ostatniego remontu domu do czasu sprzedaży wynosiła około 23 lata z odchylaniem standardowym równym w przybliżeniu 20 lat. Wartość minimalna równa zero mówi o tym, że dom został wyremontowany w roku sprzedaży. Dodatni współczynnik skośności świadczy o tym, że większość obserwacji z próbki przyjmuje wartości wyższe od wartości średniej. Kurtoza poniżej 3 wskazuje na słabą koncentrację danych wokół średniej.

- *GrLivArea* – powierzchnia domu w stopach kwadratowych

Statystyka	Wartość statystyki
Średnia	1538,1
Mediana	1501,00
Minimalna	904,00
Maksymalna	2898,00
Odch. standardowe	418,96
Skośność	0,84893
Kurtoza	0,68604

Tabela 4

Przeciętna powierzchnia sprzedanych domów była równa 1538,1 stopy kwadratowej z odchyleniem standardowym równym 418,96 stopy kwadratowej. Współczynnik skośności jest dodatni co wskazuje na to, że częściej powierzchnia domu była większa od średniej. Kurtoza poniżej 3 wskazuje na słabą koncentrację danych wokół średniej.

- *NumberOfHalfBathroom* - ilość łazienek w których znajduje się tylko toaleta

Statystyka	Wartość statystyki
Średnia	0,52675
Mediana	0,00
Minimalna	0,00
Maksymalna	2,00
Odch. standardowe	0,58414
Skośność	0,58155
Kurtoza	-0,61813

Tabela 5

Mediana równa zero wskazuje, że w 50% przypadków w sprzedanych domach nie było łazienek składających się tylko z toalety. Średnia na poziomie 0,5 z odchyleniem standardowym równym 0,6 wskazuję, że przeciętnie ilość toalet w sprzedanych domach wahała się pomiędzy zero, a jeden.

- *NumberOfFullBathroom* – ilość pełnych łazienek w sprzedanym domu

Statystyka	Wartość statystyki
Średnia	2,0576
Mediana	2,00
Minimalna	1,00
Maksymalna	4,00

Odch. standardowe	0,67785
Skośność	0,089540
Kurtoza	-0,40115

Tabela 6

Przeciętnie w sprzedanych domach znajdowały się dwie łazienki, minimalnie występowała jedna łazienka, a maksymalnie cztery.

- *NubmberOfBedroom* – ilość sypialni w domu

Statystyka	Wartość statystyki
Średnia	2,9630
Mediana	3,00
Minimalna	0,00
Maksymalna	6,00
Odch. standardowe	0,7785
Skośność	-0,30446
Kurtoza	1,9778

Tabela 7

Zarówno wartość średniej, jak i mediany wskazuje na to, że najczęściej w sprzedanych domach były trzy sypialnie. Zdarzały się jednak przypadki, w których sypialni nie było. Współczynnik skośności jest ujemny co świadczy o tym, że częściej w sprzedanych domach było mniej niż 3 sypialnie.

- *NumberOfKitchen* – ilość kuchni w sprzedanym domu

Statystyka	Wartość statystyki
Średnia	1,0288
Mediana	1,00
Minimalna	1,00
Maksymalna	2,00
Precentyl 95%	1,00
Skośność	5,6342
Kurtoza	29,744

Tabela 8

Na podstawie wartości średniej, mediany i 95% precentyla możemy stwierdzić, że w prawie wszystkich sprzedanych domach była jedna kuchnia. Zmienne maksymalnie przyjmują wartość dwa, ale dla niewielkiej ilości przypadków. Wartość kurtozy świadczy o silnej koncentracji danych wokół wartości średniej.

- *TotNumberOfRooms* – liczba wszystkich pokoi w domu.

Statystyka	Wartość statystyki
Średnia	6,5967
Mediana	6,00
Minimalna	4,00
Maksymalna	11,00
Precentyl 95%	9,00
Skośność	0,61154
Kurtoza	0,45519

Tabela 9

Przeciętnie liczba pokoi w domu była równa sześć. Minimalnie dom miał cztery pokoje, a maksymalnie jedenaście. Na podstawie 95% precentyl można zauważyć, że tylko 5% obserwacji przyjmowała wartości większe, bądź równe dziewięć. Dodatni współczynnik skośności wskazuje na to, że więcej obserwacji przyjmuje wartości powyżej średniej. Kurtoza poniżej trzech wskazuje na słabą koncentrację danych wokół wartości średniej.

- *NumberOfFireplace* – ilość kominków w domu

Statystyka	Wartość statystyki
Średnia	0,65021
Mediana	1,00
Minimalna	0,00
Maksymalna	2,00
Precentyl 95%	2,00
Skośność	0,53270
Kurtoza	-0,72481

Tabela 10

Mediana wskazuje, że w 50% przypadków sprzedanych domów występował nie więcej niż jeden kominek. Maksymalna liczba kominków w sprzedanych domach była równa 2. Dodatni współczynnik skośności sugeruje, że większość obserwacji przyjmuje wartości powyżej średniej. Kurtoza poniżej trzech świadczy o słabej koncentracji danych wokół wartości przeciętnej.

- *GarageArea* – powierzchnia garażu w stopach do kwadratu

Statystyka	Wartość statystyki
Średnia	487,04
Mediana	490,00

Minimalna	200,00
Maksymalna	936,00
Odch.standardowe	155,13
Skośność	0,45992
Kurtoza	0,10236

Tabela 11

Przeciętnie powierzchnia garażu wynosi 487,04 stóp kwadratowych z odchyleniem standardowym równym 155,13 stóp kwadratowych. Wartość minimalna wynosi 200 stóp kwadratowych, co świadczy o tym, że w próbce nie ma domu bez garażu. Współczynnik skośności jest dodatni co świadczy o tym, że częściej powierzchnia garażu jest większa od wartości średniej. Kurtoza poniżej trzech świadczy o słabej koncentracji danych wokół średniej.

Zmienne binarne i kategoryczne

- *CentralAir* – zmienna informująca o występowaniu klimatyzacji w domu lub nie

1	0
Występuje	Nie występuje

Tabela 12

Statystyka	Wartość statystyki
Średnia	0,96708
Mediana	1,00

Tabela 13

Wartość średniej jak i mediany świadczy o tym, że w większości domów występowała klimatyzacja.

Współczynnik zmienności

Współczynnik zmienności jest stosowany w statystyce jako parametry mierzący zróżnicowanie cechy. Do budowy modelu ekonometrycznego nie wykorzystuje się zmiennych, których współczynnik zmienności przyjmuje wartości mniejsze od 10%. Zmienne ze zbyt niskim współczynnikiem zmienności uważa się za quasi-stałe, czyli takie, które nie wnoszą istotnych informacji do modelu.

Zmienna	Wsp. zmienności
SalePrice	0,29521
YearsFromBuild	0,78547
YearsFromRemodAdd	0,87052
CentralAir	0,18489
GrLivArea	0,27239
NumberOfHalfBath~	1,109
NumberOfFullBath~	0,32943
NumberOfBedroom	0,26274
NumberOfKitchen	0,16291
TotNumberOfRooms	0,21844
NumberOfFireplac~	1,0239

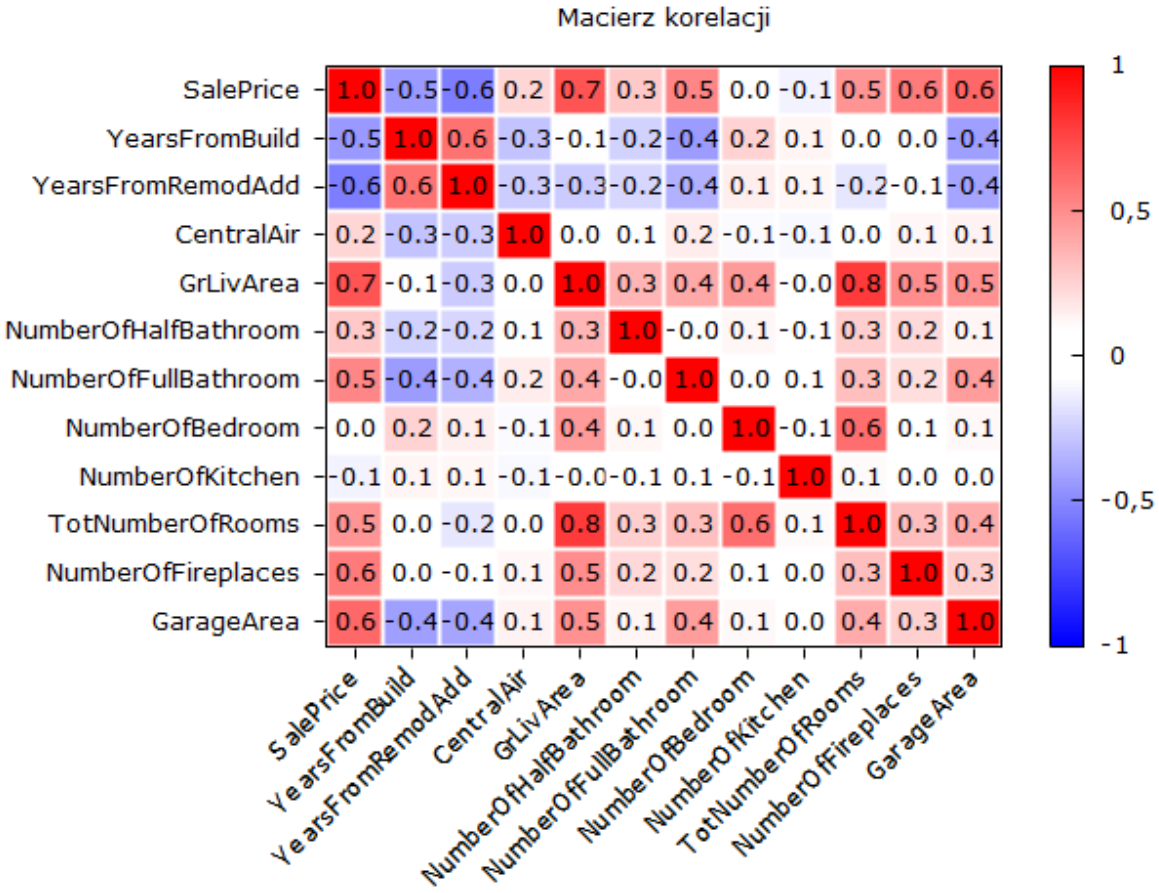
GarageArea	0,31852

Tabela 14

Na podstawie tabeli *Tabela 14* można zauważyć, że współczynniki zmienności dla wszystkich zmiennych objaśniających przyjmują wartość powyżej 10%. Na tym etapie analizy wszystkie zmienną mogą zostać wykorzystane do estymacji modelu.

Macierz korelacji

Zmienne użyte do budowy modelu ekonometrycznego powinny charakteryzować się wysokim współczynnikiem korelacji pomiędzy zmiennymi objaśnianymi, a zmienną objaśnianą, a niskim współczynnikiem korelacji pomiędzy zmiennymi objaśniającymi. Na podstawie analizy macierzy korelacji można dokonać wstępnych założeń co do zmiennych objaśniających, które mogą znaleźć się w ostatecznym modelu.



Rysunek 3

Na podstawie macierzy korelacji możemy zaobserwować następujące zależności:

- Najsłabsze korelacje pomiędzy zmienną objaśnianą występują pomiędzy ilością sypialni i ilością kuchni. Najsilniej zmienna *SalePrice* jest skorelowana z powierzchnią domu w stopach kwadratowych, ilością kominków w domu, powierzchnia garażu i latami, które upłynęły od ostatniego remontu.
- Ujemna zależność występuje pomiędzy zmienną *SalePrice* i zmiennymi *YearsFromBulid* i *YearsFromRemodeAdd*. Oznacza to, że im starszy dom lub więcej lat upłynęło od remontu to tym niższa cena sprzedaży domu.

- Silna korelacja występuje pomiędzy zmiennymi objaśniającymi *LiveGrArea* i *TotNumberOfRooms*. Jest to zależność dość oczywista, ale w modelu zapewne będzie występowała jedna z tych wartości.

3.4 Dobór zmiennych do modelu

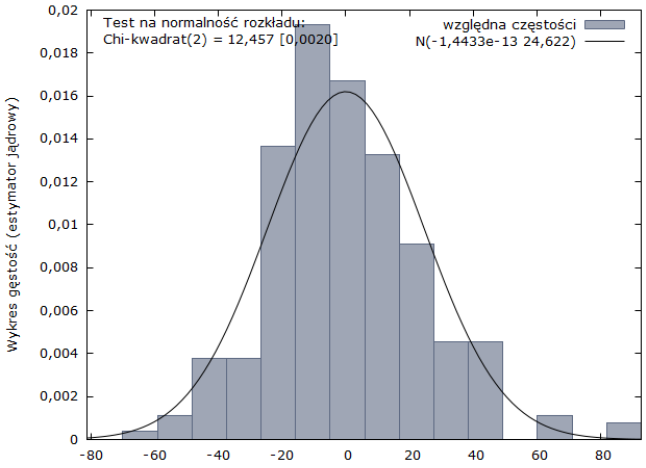
Pierwszy model zostanie wyestymowanych za pomocą regresji krokowej w tył. Metoda polega na wyjściu od modelu ze wszystkimi potencjalnymi zmiennymi objaśniającymi, a następnie z modelu będą eliminowane zmienne nieistotne.

Model wyjściowy M_RKwT:

```
Model 1: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                106,069         17,2294          6,156     3,27e-09  ***
  YearsFromBuild        -0,260260       0,0865968      -3,005     0,0029    ***
  YearsFromRemodAdd     -0,500512       0,106949       -4,680     4,89e-06  ***
  CentralAir            14,8759         9,53102         1,561     0,1199
  GrLivArea              0,0587763      0,00762980      7,704     3,87e-013 ***
  NumberOfHalfBath~     -2,91093        3,21411        -0,9057    0,3661
  NumberOfFullBath~      6,03394        3,05653         1,974     0,0496    **
  NumberOfBedroom      -10,1544         2,83943        -3,576     0,0004    ***
  NumberOfKitchen      -34,6045         9,98701        -3,465     0,0006    ***
  TotNumberOfRooms       0,0768515      2,10383         0,03653   0,9709
  NumberOfFireplac~     21,5355         2,88622         7,461     1,73e-012 ***
  GarageArea             0,0661810      0,0134210       4,931     1,56e-06  ***

Średn.aryt.zm.zależnej  182,2509   Odch.stand.zm.zależnej  53,80273
Suma kwadratów reszt    140043,1   Błąd standardowy reszt  24,62209
Wsp. determ. R-kwadrat  0,800089   Skorygowany R-kwadrat   0,790569
F(11, 231)              84,04652   Wartość p dla testu F   2,65e-74
Logarytm wiarygodności -1117,134   Kryt. inform. Akaike'a  2258,269
Kryt. bayes. Schwarza   2300,185   Kryt. Hannana-Quinna    2275,152
```

Rysunek 4

Weryfikacja normalności rozkładu reszt, w celu sprawdzenia, czy założenia, metody MNK są poprawne, a także czy interpretacja Testu t-Studenta oraz współczynnika determinacji będzie poprawna.



Rysunek 5

Na podstawie wartości p-value hipoteza zerowa zostaje odrzucana, rozkład reszt nie jest rozkładem normlanym. Ponieważ próbka użyta do estymacji jest odpowiednio duża na podstawie Centralnego Twierdzenia Granicznego można założyć normalność rozkładu reszt.

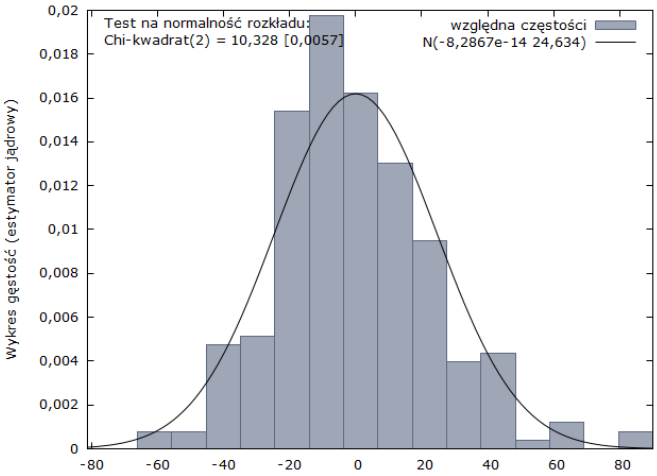
W kolejnych krokach zostaną wyeliminowane zmienne o najmniejszej istotności. Schemat eliminacji kolejnych zmiennych nie został udokumentowany z uwagi na dużą liczbę kroków, którą trzeba było wykonać do otrzymania ostatecznej postaci modelu.

Ostateczny model uzyskany z regresji krokowej w tył M_RKwT:

```
Model 4: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                121,393         13,1249          9,249    1,47e-017 ***
  YearsFromBuild        -0,253906       0,0809523      -3,136    0,0019    ***
  YearsFromRemodAdd     -0,517360       0,105482       -4,905    1,75e-06  ***
  GrLivArea              0,0560252      0,00586169      9,558    1,74e-018 ***
  NumberOfFullBath~      7,09406        2,88950         2,455    0,0148    **
  NumberOfBedroom      -10,0897         2,51917        -4,005    8,33e-05  ***
  NumberOfKitchen      -35,7190         9,72046        -3,675    0,0003    ***
  NumberOfFireplac~     21,9559         2,83182         7,753    2,73e-013 ***
  GarageArea             0,0680858      0,0132315       5,146    5,64e-07  ***

Średn.aryt.zm.zależnej  182,2509   Odch.stand.zm.zależnej  53,80273
Suma kwadratów reszt    142000,2   Błąd standardowy reszt  24,63409
Wsp. determ. R-kwadrat  0,797295   Skorygowany R-kwadrat   0,790365
F(8, 234)               115,0482   Wartość p dla testu F   1,15e-76
Logarytm wiarygodności -1118,821   Kryt. inform. Akaike'a  2255,641
Kryt. bayes. Schwarza   2287,079   Kryt. Hannana-Quinna    2268,304
```

Rysunek 6

Weryfikacja normalności rozkładu reszt modelu M_RKwT:



Rysunek 7

Na podstawie wartości p-value hipoteza zerowa zostaje odrzucana, rozkład reszt nie jest rozkładem normlanym. Ponownie można powołać się na Centralne Twierdzenie Graniczne i założyć normalność rozkładu reszt.

Dla wyżej wyestymowanego modelu współczynnik determinacji R^2 przyjmuje wartość równą 79,73%, a skorygowany R^2 wartość równą 79,04%. Wartości te nie różnią się znacznie od siebie. Wartość p dla testu F wskazuje, że wartość współczynnika determinacji jest istotna.

3.5 Dobór zmiennych objaśniających metodą Hellwiga

Do wyboru optymalnego podzbioru został wykorzystany skrypt z ćwiczeń laboratoryjnych. Według metody Hellwiga zbiorem zmiennych objaśniających o najwyższym integralnym wskaźniku pojemności informacyjnej jest zbiór składający się z następujących zmiennych:

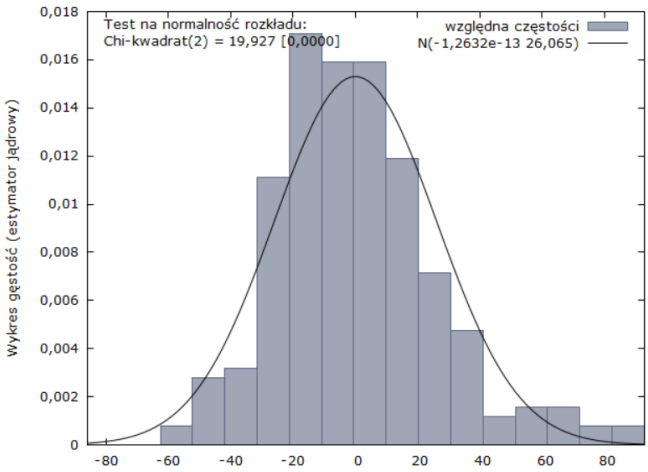
- YearsFromBuild
- YearsFromRemodAdd
- GrLivArea
- NumberOfFireplaces
- GarageArea

Model M_Hellwig:

```
Model 5: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                88,1202         8,93874          9,858    1,96e-019 ***
  YearsFromBuild       -0,394763       0,0798580       -4,943    1,45e-06  ***
  YearsFromRemodAdd    -0,626157       0,109378        -5,725    3,12e-08  ***
  GrLivArea             0,0482173      0,00520527       9,263    1,24e-017 ***
  NumberOfFireplac~    24,8673         2,90501          8,560    1,44e-015 ***
  GarageArea            0,0655680      0,0138002        4,751    3,51e-06  ***

Średn.aryt.zm.zależnej  182,2509   Odch.stand.zm.zależnej  53,80273
Suma kwadratów reszt    161017,7   Błąd standardowy reszt  26,06529
Wsp. determ. R-kwadrat  0,770147   Skorygowany R-kwadrat   0,765298
F(5, 237)               158,8191   Wartość p dla testu F   1,44e-73
Logarytm wiarygodności -1134,091   Kryt. inform. Akaike'a  2280,183
Kryt. bayes. Schwarza   2301,141   Kryt. Hannana-Quinna    2288,625
```

Rysunek 8

Weryfikacja normalności rozkładu reszt modelu M_Hellwig:



Rysunek 9

Na podstawie wartości p-value hipoteza zerowa zostaje odrzucana, rozkład reszt nie jest rozkładem normlanym. Ponownie można powołać się na Centralne Twierdzenie Graniczne i założyć normalność rozkładu reszt.

Dla modelu M_Hellwig współczynnik determinacji R^2 przyjmuje wartość równą 77,01%, a skorygowany R^2 wartość równą 76,53%. Wartości te nie różnią się znacznie od siebie. Wartość p dla testu F wskazuje, że wartość współczynnika determinacji jest istotna.

3.6 Porównanie modeli

Współczynnik	Wartość współczynnika	
	M_Hellwig	**M_RKwT**
Wsp. Determinacji R^2	0,770147	0,797295
Skorygowany R^2	0,765298	0,790365

Kryt. bayes. Schwarza	2301,141	2287,079
Kryt. inform. Akaike'a	2280,183	2255,641
Kryt. Hannana-Quinna	2288,625	2268,304

Tabela 15

Zarówno współczynnik determinacji jak i kryteria informacyjne wskazują na lepsze dopasowanie modelu M_RkwT, czyli modelu dla którego zmienne objaśniające zostały wybrane Metodą Krokową. Ponieważ różnice w wartości współczynników nie są duże dalsze badanie będzie przebiegało na dwóch modelach równocześnie.

3.7 Weryfikacja czy model jest koincydentny

Sprawdzenie koincydencji modelu jest konieczne do prawidłowej interpretacji współczynnika determinacji. Analizie poddany jest znak parametru strukturalnego i współczynnik korelacji zmiennej objaśnianej i zmiennej objaśniającej. Jeżeli dla każdej zmiennej objaśniającej w modelu spełniony jest warunek:

$$sgn(r_i) = sgn(a_i)$$

to model jest koincydentny[17].

Model M_RKwT:

Zmienna	Znak oszacowania parametru	Znak współczynnika korelacji
YearsFromBuild	-	-
YearsFromRemodAdd	-	-
GrLivArea	+	+
NumberOfFullBath	+	+
NumberOfBedroom	-	-
NumberOfKitchen	-	-
NumberOfFireplace	+	+
GarageArea	+	+

Tabela 16

Wszystkie znaki parametrów w tabeli *Tabela 16* są zgodne. Model M_RKwT jest koincydentny.

Model M_Hellwig:

Tabela 17

Zmienna	Znak oszacowania parametru	Znak współczynnika korelacji
YearsFromBuild	-	-
YearsFromRemodAdd	-	-
GrLivArea	+	+
NumberOfFireplace	+	+
GarageArea	+	+

Wszystkie znaki parametrów w tabeli *Tabela 17* są zgodne. Model M_Hellwig jest koincydentny.

[17] „Podstawy Ekonometrii z elementami algebry liniowej”, Eligiusz W.Nowakowski, data dostępu: 28.05.2021

3.8 Sprawdzanie występowania współliniowości w modelu

W programie Gretl została użyta następująca funkcja:

ols SalePrice const 2 3 5 7 8 9 11 12

vif

Weryfikacja współliniowości w modelu M_RKwT:

```
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

      YearsFromBuild    1,995
   YearsFromRemodAdd    1,741
          GrLivArea    2,405
NumberOfFullBathroom    1,530
     NumberOfBedroom    1,534
     NumberOfKitchen    1,059
  NumberOfFireplaces    1,418
          GarageArea    1,680
```

Rysunek 10

Żadna z powyższych wartości nie jest większa od 10, co oznacza, że w modelu M_RKwT nie występuje zjawisko współliniowości.

Weryfikacja współliniowości w modelu M_Hellwig:

```
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

    YearsFromBuild    1,734
 YearsFromRemodAdd    1,672
         GrLivArea    1,694
NumberOfFireplaces    1,332
        GarageArea    1,632
```

Rysunek 11

Żadna z powyższych wartości nie jest większa od 10, co oznacza, że w modelu M_Hellwig nie występuje zjawisko współliniowości.

3.9 Występowanie katalizatorów

Kolejnym etapem weryfikacji jednorównaniowego modelu ekonometrycznego jest sprawdzenie czy w modelu występuje efekt katalizy. Efekt katalizy występuje, gdy w modelu są uwzględnione zmienne silnie ze sobą skorelowane. Występowanie katalizatorów w modelu może zaburzać rzeczywistą wartość współczynnika determinacji R^2[18]. Sprawdzenie występowania zjawiska katalizy zostanie przeprowadzone w programie Gretl za pomocą skryptu z zajęć laboratoryjnych.

[18] „Weryfikacja modelu”, data dostępu: 29.05.2021

Katalizatory w modelu M_RKwT :

```
KATALIZATOR:
NumberOfBedroom
W PARZE:
NumberOfBedroom YearsFromBuild
KATALIZATOR:
NumberOfBedroom
W PARZE:
NumberOfBedroom YearsFromRemodAdd
KATALIZATOR:
NumberOfBedroom
W PARZE:
NumberOfBedroom GrLivArea
KATALIZATOR:
NumberOfKitchen
W PARZE:
NumberOfKitchen NumberOfFullBathroom
KATALIZATOR:
YearsFromBuild
W PARZE:
YearsFromBuild NumberOfFireplaces
KATALIZATOR:
NumberOfKitchen
W PARZE:
NumberOfKitchen NumberOfFireplaces
KATALIZATOR:
NumberOfBedroom
W PARZE:
NumberOfBedroom GarageArea
KATALIZATOR:
NumberOfKitchen
W PARZE:
NumberOfKitchen GarageArea
```

Rysunek 12

Na listingu powyżej można zauważyć, że w modelu występują katalizatory. Model z katalizatorami nie może zostać użyty w dalszych rozważaniach. Z modelu zostaną usunięte zmienne będące katalizatorami: *NumberOfBedroom*, *NumberOfKitchen*, *YearsFromBuild*. Poniżej znajduje się model M_RKwT po usunięciu katalizatorów. Rozkład reszt modelu nie jest rozkładem normalnym. Ponieważ próbka jest odpowiednio duża można założyć normalność rozkładu reszt na podstawie Centralnego Twierdzenia Granicznego.

```
Model 9: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                66,3019         9,53180          6,956    3,39e-011 ***
  YearsFromRemodAdd    -0,839146       0,0978822       -8,573    1,32e-015 ***
  GrLivArea             0,0397068      0,00533966       7,436    1,89e-012 ***
  NumberOfFullBath~    10,0251         2,97101          3,374    0,0009    ***
  NumberOfFireplac~    24,2221         2,97908          8,131    2,38e-014 ***
  GarageArea            0,0771904      0,0137491        5,614    5,50e-08  ***

Średn.aryt.zm.zależnej  182,2509   Odch.stand.zm.zależnej  53,80273
Suma kwadratów reszt    169477,7   Błąd standardowy reszt  26,74127
Wsp. determ. R-kwadrat  0,758071   Skorygowany R-kwadrat   0,752967
F(5, 237)               148,5250   Wartość p dla testu F   6,06e-71
Logarytm wiarygodności -1140,313   Kryt. inform. Akaike'a  2292,626
Kryt. bayes. Schwarza   2313,584   Kryt. Hannana-Quinna    2301,068
```

Rysunek 13

```
? natezenie_efektu_katalizy
0,011660052
```

Rysunek 14

Z listingu powyżej wynika, że w modelu nie występuje już zjawisko katalizy. Natężenie efektu katalizy wynosi 0,012. Model M_RKwT na tym etapie badania trzeba poddać ponownej weryfikacji, ponieważ w wyniku usuwania katalizatorów zmieniła się postać modelu. Na rysunku *Rysunek 13* możemy zauważyć, że wszystkie zmienne objaśniające w modelu są istotne. Wartość skorygowanego współczynnika R^2 wynosi 75,3 %. Wartość p-value dla testu F mówi o istotności współczynnika determinacji. Wysoka wartość współczynnika świadczy o dobrym dopasowaniu modelu do danych.

Sprawdzanie czy model M_RKwT jest koincydentny:

Zmienna	Znak oszacowania parametru	Znak współczynnika korelacji
YearsFromRemodAdd	-	-
GrLivArea	+	+
NumberOfFullBath	+	+
NumberOfFireplace	+	+
GarageArea	+	+

Tabela 18

Wszystkie znaki parametrów w *Tabeli 18* są zgodne. Model M_RKwT jest koincydentny.

Weryfikacja współliniowości w modelu M_RKwT:

```
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

   YearsFromRemodAdd    1,272
           GrLivArea    1,694
NumberOfFullBathroom    1,373
  NumberOfFireplaces    1,331
         GarageArea     1,540
```

Rysunek 15

Żadna z powyższych wartości VIF_j nie jest większa od 10, co świadczy o braku występowania zjawiska współliniowości w modelu.

Katalizatory w modelu M_Hellwig:

```
KATALIZATOR:
YearsFromBuild
W PARZE:
YearsFromBuild NumberOfFireplaces
```

Rysunek 16

Na listingu powyżej widać, że w modelu M_Hellwig występuje katalizator *YearsFromBuild*. Poniżej znajduje się model M_Hellwig po usunięciu katalizatora. Rozkład reszt modelu nie jest rozkładem normalnym. Ponieważ próbka jest odpowiednio duża można założyć normalność rozkładu reszt na podstawie Centralnego Twierdzenia Granicznego.

```
Model 11: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice

                    współczynnik  błąd standardowy  t-Studenta  wartość p
  ---------------------------------------------------------------------------
  const                77,7078         9,10470          8,535    1,67e-015
  YearsFromRemodAdd    -0,908499       0,0977658       -9,293    9,94e-018
  GrLivArea             0,0432059      0,00535108       8,074    3,38e-014
  NumberOfFireplac~    24,3869         3,04298          8,014    4,97e-014
  GarageArea            0,0880954      0,0136524        6,453    6,09e-010

Średn.aryt.zm.zależnej  182,2509   Odch.stand.zm.zależnej  53,80273
Suma kwadratów reszt    177619,7   Błąd standardowy reszt  27,31852
Wsp. determ. R-kwadrat  0,746448   Skorygowany R-kwadrat   0,742187
F(4, 238)               175,1658   Wartość p dla testu F   1,09e-69
Logarytm wiarygodności -1146,014   Kryt. inform. Akaike'a  2302,028
Kryt. bayes. Schwarza   2319,494   Kryt. Hannana-Quinna    2309,063
```

Rysunek 17

Ponowne sprawdzenie czy w modelu M_Hellwig występują katalizatory:

```
? natezenie_efektu_katalizy
0,0016486516
```

Rysunek 18

Z listingu powyżej wynika, że w modelu nie występuje już zjawisko katalizy. Natężenie efektu katalizy wynosi 0,0016. Model M_Hellwig na tym etapie badania trzeba poddać ponownej weryfikacji. Na *Rysunku 17* możemy zauważyć, że wszystkie zmienne objaśniające w modelu są istotne. Wartość skorygowanego współczynnika R^2 wynosi 74,22 %. Wartość p-value dla testu F mówi o istotności współczynnika determinacji. Wysoka wartość współczynnika świadczy o dobrym dopasowaniu modelu do danych.

Sprawdzanie czy model M_Hellwig jest koincydentny:

Zmienna	Znak oszacowania parametru	Znak współczynnika korelacji
YearsFromRemodAdd	-	-
GrLivArea	+	+
NumberOfFireplace	+	+
GarageArea	+	+

Tabela 19

Wszystkie znaki parametrów w tabeli *Tabela 19* są zgodne. Model M_Hellwig jest koincydentny.

Weryfikacja współliniowości w modelu M_RKwT:

```
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

 YearsFromRemodAdd    1,216
         GrLivArea    1,630
NumberOfFireplaces    1,331
        GarageArea    1,454
```

Rysunek 19

Żadna z powyższych wartości VIF_j nie jest większa od 10, co świadczy o braku występowania zjawiska współliniowości w modelu.

3.10 Test Walda dla zmiennej *NumberOfFullBathroom*

Podzbiory rozważanych modeli różnią się tylko występowaniem jednej zmiennej objaśniającej NumberOfFullBathroom. Testem Walda zweryfikujemy, czy wprowadzenie do modelu tej zmiennej objaśniającej jest istotne.

```
Test porównawczy z Modelem 5

  Hipoteza zerowa: parametr regresji jest równy zero dla NumberOfFullBathroom
  Statystyka testu: F(1, 237) = 11,386, wartość p 0,000864411
  Dodanie zmiennych poprawiło 3 z 3 kryteriów informacyjnych (AIC, BIC, HQC).
```

Rysunek 20

Na podstawie powyższego listingu i wartości p-vale dla testu Walda stwierdzamy, że wprowadzenie dodatkowej zmiennej do modelu jest istotne i poprawa wszystkie kryteria informacyjne, a także wartość skorygowanego R^2.

3.11 Wybór modelu

Do tej pory analiza była prowadzona na dwóch modelach równocześnie. Dzięki temu został wybrany optymalny podzbiór zmienny objaśniających. Optymalny podzbiór cechuje się istotnością wszystkich zmiennych objaśniających, brakiem występowania współliniowości, a także brakiem katalizatorów. Dodatkowo modele są koincydentne.

Współczynnik	Wartość współczynnika	
	M_Hellwig	M_RKwT
Wsp. Determinacji R^2	0,746448	0,758071
Skorygowany R^2	0,742187	0,752967
Kryt. bayes. Schwarza	2319,494	2313,584
Kryt. inform. Akaike'a	2302,028	2292,626
Kryt. Hannana-Quinna	2309,063	2301,068

Tabela 20

W powyższej tabeli zestawiono współczynniki determinacji i kryteria informacyjne dotyczące dwóch rozważanych modeli. Można zauważyć, że Skorygowany R^2 jest wyższy, a kryteria informacyjne dla modelu M_RKwT są niższe. Test Walda potwierdza przypuszczenie, że wprowadzenie dodatkowej zmiennej objaśniającej *NumberOfFullBathroom* jest istotne. Na tej podstawie można wnioskować, że model, do którego zmienne zostały wybrane metodą krokową wstecz jest lepiej dopasowanych do danych.

3.12 Weryfikacja statystyczna modelu

Do weryfikacji czy próba użyta do estymacji parametrów modelu jednorównaniowego została dobrana w sposób losowy zostanie wykorzystany Test Liczby Serii.

```
Test serii

Liczba serii (R) dla zmiennej 'ResztyM_RKwT' = 129
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(122,5, 7,77817),
test z-score = 0,835672, przy dwustronym obszarze krytycznym p = 0,40334
```

Rysunek 21

Na podstawie wartości p-value nie ma podstaw do odrzucenia hipotezy zerowej, mówiącej o losowości próbki.

3.13 Badanie liniowości modelu

Do weryfikacji czy analityczna postać modelu została dobrana prawidłowo służy test Ramsey'a RESET i jego nieparametryczny odpowiednik Test Liczby Serii. Jak pokazano na rysunku *Rysunek 21* Test Liczby Serii i obliczona dla niego wartość p-value nie dają podstaw do odrzucenia hipotezy zerowej. Test Ramsey'a RESET jest jednak silniejszym testem i zostanie użyty do potwierdzenia wyników Testu Serii.

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 5,565067,
z wartością p = P(F(2,235) > 5,56507) = 0,00435

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 10,037327,
z wartością p = P(F(1,236) > 10,0373) = 0,00174

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 9,162564,
z wartością p = P(F(1,236) > 9,16256) = 0,00274
```

Rysunek 22

Test Ramsey'a RESET we wszystkich przypadkach daje podstawy do odrzucenia hipotezy zerowej, mówiącej o tym, że dobrany model liniowy jest odpowiedni do estymacji danego zjawiska. Ponieważ postać modelu nie jest optymalna dokonamy transformacji zmiennych, aby uzyskać odpowiednią postać modelu.

W modelu dokonano transformacji zmiennej objaśnianej :

$$Y_{nowy} = \ln(Y)$$

Test Ramseya' RESET dla modelu z nową zmienną objaśnianą:

```
Test RESET na specyfikację (kwadrat i sześcian zmiennej)
Statystyka testu: F = 0,438287,
z wartością p = P(F(2,235) > 0,438287) = 0,646

Test RESET na specyfikację (tylko kwadrat zmiennej)
Statystyka testu: F = 0,846412,
z wartością p = P(F(1,236) > 0,846412) = 0,359

Test RESET na specyfikację (tylko sześcian zmiennej)
Statystyka testu: F = 0,838267,
z wartością p = P(F(1,236) > 0,838267) = 0,361
```

Rysunek 23

Na podstawie powyższego listingu można zauważyć, że wszystkie wartości p-value są wyższe od 5%. Nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o poprawnie dobranej postaci modelu. Poniżej wyestymowany model ze zmienną objaśnianą SalePrice_ln.

Model M_lnY:

```
Model 14: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): SalePrice_ln

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                4,54900       0,0532557        85,42     5,09e-180 ***
  YearsFromRemodAdd   -0,00521316    0,000546884      -9,532    1,93e-018 ***
  GrLivArea            0,000195353   2,98336e-05       6,548    3,58e-010 ***
  NumberOfFullBath~    0,0716390     0,0165995         4,316    2,34e-05  ***
  NumberOfFireplac~    0,129964      0,0166446         7,808    1,86e-013 ***
  GarageArea           0,000408744   7,68184e-05       5,321    2,39e-07  ***

Średn.aryt.zm.zależnej  5,161820   Odch.stand.zm.zależnej  0,300405
Suma kwadratów reszt    5,290482   Błąd standardowy reszt  0,149408
Wsp. determ. R-kwadrat  0,757749   Skorygowany R-kwadrat   0,752638
F(5, 237)               148,2648   Wartość p dla testu F   7,08e-71
Logarytm wiarygodności  120,1969   Kryt. inform. Akaike'a -228,3938
Kryt. bayes. Schwarza   -207,4355   Kryt. Hannana-Quinna   -219,9520
```

Rysunek 24

Rozkład reszt modelu M_lnY nie jest rozkładem normalnym. Ponieważ próbka jest odpowiednio duża można założyć normalność rozkładu reszt na podstawie Centralnego Twierdzenia granicznego. Skorygowany R^2 wynosi 75,26%. Wartość p-value dla testu F świadczy o tym, że współczynnik determinacji jest istotny. Ponieważ postać modelu uległa zmianie konieczna jest ponowna weryfikacja założeń MNK.

Sprawdzanie czy model M_lnY jest koincydentny:

Zmienna	Znak oszacowania parametru	Znak współczynnika korelacji
YearsFromRemodAdd	-	-
GrLivArea	+	+
NumberOfFullBath	+	+
NumberOfFireplace	+	+
GarageArea	+	+

Tabela 21

Wszystkie znaki parametrów w tabeli *Tabela 21* są zgodne. Model M_lnY jest koincydentny.

Weryfikacja czy w modelu występuje zjawisko współliniowości:

```
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

   YearsFromRemodAdd    1,272
             GrLivArea  1,694
NumberOfFullBathroom    1,373
  NumberOfFireplaces    1,331
          GarageArea    1,540
```

Rysunek 25

Żadna z powyższych wartości VIF_j nie jest większa od 10, co świadczy o braku występowania zjawiska współliniowości w modelu.

Katalizatory w modelu M_lnY :

```
? natezenie_efektu_katalizy
0,0088337023
```

Rysunek 26

Z listingu powyżej wynika, że w modelu nie występuje już zjawisko katalizy. Natężenie efektu katalizy wynosi 0,0088.

Test Serii dla modelu M_lnY:

```
Test serii

Liczba serii (R) dla zmiennej 'resztyM_lnY' = 128
Test niezależności oparty na liczbie dodatnich i ujemnych serii.
Hipoteza zerowa: próba jest losowa, dla R odpowiednio N(122,5, 7,77817),
test z-score = 0,707107, przy dwustronym obszarze krytycznym p = 0,4795
```

Rysunek 27

Na podstawie wartości p-value nie ma podstaw do odrzucenia hipotezy zerowej, mówiącej o losowości próbki.

3.14 Weryfikacja stabilności postaci funkcyjnej modelu

Model jest budowany na podstawie próby, ale docelowo opisuje zmienność całej populacji. W związku z tym weryfikacja stabilności postaci funkcyjnej modelu jest konieczna przez wykorzystaniem modelu do prognozowania.

Do wykonania testu Chowa podzbiór został podzielony na połowę ponieważ w modelu nie występuje zmienna binarna. Wyniki testu:

```
Test Chowa na zmiany strukturalne przy podziale próby w obserwacji 122
  F(6, 231) = 1,39738 z wartością p 0,2165
```

Rysunek 28

Na podstawie wartości p-value dla Testu Chowa nie ma podstaw do odrzucenia hipotezy zerowej mówiącej o stabilności parametrów modelu.

3.15Weryfikacja stałości wariancji składnika losowego

Weryfikacja stałości wariancji składnika losowego jest jednym z założeń MNK. Jeżeli wariancja składnika ε nie jest stała to estymatory parametrów modelu nie są efektywne. Do weryfikacji tego założenia w programie Gretl można użyć trzech testów: testu Breuscha-Pagana, White'a i Koenkera.

Test White'a dla modelu M_lnY :

```
Statystyka testu: TR^2 = 19,272232,
z wartością p = P(Chi-kwadrat(20) > 19,272232) = 0,504199
```

Rysunek 29

Na podstawie statystyki p-valeu z rysunku *Rysynk 29* nie ma podstaw do podrzucenia hipotezy zerowej, mówiącej o stałości wariancji składnika losowego.

Test Breuscha-Pagana dla modelu M_lnY :

```
Test Breuscha-Pagana na heteroskedastyczność
Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): standaryzowane uhat^2

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                1,42237        1,09133          1,303      0,1937
  YearsFromRemodAdd    0,0122807      0,0112069        1,096      0,2743
  GrLivArea            0,000442174    0,000611358      0,7233     0,4702
  NumberOfFullBath~   -0,402914       0,340163        -1,184      0,2374
  NumberOfFireplac~   -0,0644821      0,341086        -0,1890     0,8502
  GarageArea          -0,00104918     0,00157419      -0,6665     0,5057

  Wyjaśniona suma kwadr. = 54,3026

Statystyka testu: LM = 27,151289,
z wartością p = P(Chi-kwadrat(5) > 27,151289) = 0,000053
```

Rysunek 30

Na podstawie statystyki p-valeu z rysunku *Rysunek 30* odrzucamy hipotezę zerową mówiącą o stałości wariancji składnika losowego. W modelu występuje zatem heteroskedastyczność składnika losowego.

3.16Korekta heteroskedastyczności

Korekta heteroskedastyczności została wykonana za pomocą Ważonej Metody Najmniejszych Kwadratów. W analizowanym modelu występują dwie potencjalne zmienne, które mogą powodować heteroskedastyczność. Są to NumberOfFullBathroom i YearsFromRemodeAdd.

```
Model 28: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): kwadratyResztM_lnY

                     współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const              0,0127232      0,00649880        1,958      0,0514   *
  YearsFromRemodAdd  0,000397602    0,000215583       1,844      0,0664   *

Średn.aryt.zm.zależnej  0,021772   Odch.stand.zm.zależnej  0,066767
Suma kwadratów reszt    1,063788   Błąd standardowy reszt  0,066438
Wsp. determ. R-kwadrat  0,013918   Skorygowany R-kwadrat   0,009826
F(1, 241)               3,401471   Wartość p dla testu F   0,066366
Logarytm wiarygodności  315,0918   Kryt. inform. Akaike'a -626,1837
Kryt. bayes. Schwarza  -619,1975   Kryt. Hannana-Quinna   -623,3697
```

Rysunek 31

```
Model 29: Estymacja KMNK, wykorzystane obserwacje 1-243
Zmienna zależna (Y): kwadratyResztM_lnY

                      współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                 0,0460757      0,0136435         3,377     0,0009   ***
  NumberOfFullBath~    -0,0118118      0,00629909       -1,875     0,0620   *

Średn.aryt.zm.zależnej  0,021772   Odch.stand.zm.zależnej  0,066767
Suma kwadratów reszt    1,063289   Błąd standardowy reszt  0,066423
Wsp. determ. R-kwadrat  0,014380   Skorygowany R-kwadrat   0,010291
F(1, 241)               3,516237   Wartość p dla testu F   0,061980
Logarytm wiarygodności  315,1489   Kryt. inform. Akaike'a -626,2977
Kryt. bayes. Schwarza  -619,3116   Kryt. Hannana-Quinna   -623,4838
```

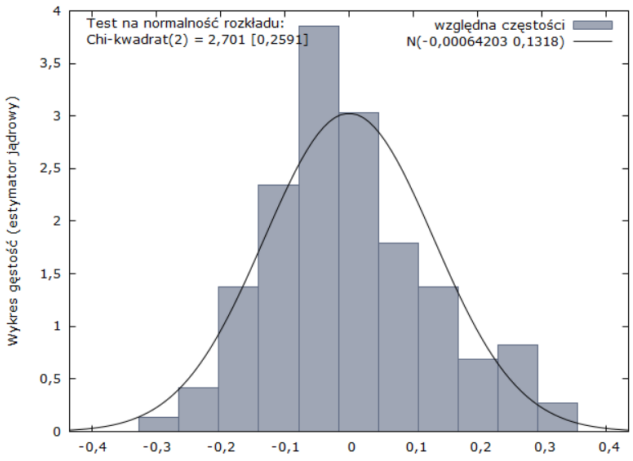
Rysunek 32

Za wagę została przyjęta $w_i = sqrt(NumberOfFullBathroom)$, ponieważ model z wskazaną wagą wyestymowany za pomocą WMNK ma najwyższy Skorygowany współczynnik determinacji i najniższe kryteria informacyjne. Poniżej znajduje się model M_WMNK po korekcie heteroskedastyczności:

```
Model 21: Estymacja WLS, wykorzystane obserwacje 1-243 (n = 117)
Liczba pominiętych niekompletnych obserwacji: 126
Zmienna zależna (Y): SalePrice_ln
Zmienna jako waga: wagi

                      współczynnik  błąd standardowy  t-Studenta  wartość p
  ------------------------------------------------------------------------
  const                 4,58210        0,0662413         69,17     4,10e-093 ***
  YearsFromRemodAdd    -0,00524749     0,000807472       -6,499    2,38e-09  ***
  GrLivArea             0,000188007    3,95989e-05        4,748    6,18e-06  ***
  NumberOfFullBath~     0,0364086      0,0209666          1,737    0,0852    *
  NumberOfFireplac~     0,140645       0,0205275          6,852    4,28e-010 ***
  GarageArea            0,000506274    0,000103025        4,914    3,10e-06  ***

Podstawowe statystki dla ważonych danych:

Suma kwadratów reszt    1,855987   Błąd standardowy reszt  0,129308
Wsp. determ. R-kwadrat  0,792992   Skorygowany R-kwadrat   0,783667
F(5, 111)               85,04205   Wartość p dla testu F   2,49e-36
Logarytm wiarygodności  76,39400   Kryt. inform. Akaike'a -140,7880
Kryt. bayes. Schwarza  -124,2150   Kryt. Hannana-Quinna   -134,0595

Podstawowe statystyki dla oryginalnych danych:

Średn.aryt.zm.zależnej  5,242181   Odch.stand.zm.zależnej  0,283977
Suma kwadratów reszt    1,928319   Błąd standardowy reszt  0,131804
```

Rysunek 33

Weryfikacja normalności rozkładu reszt:



Rysunek 34

Na podstawie wartości p-vale testu Doornika-Hansena reszty modelu mają rozkład normalny. Skorygowany współczynnik determinacji R^2wynosi 78,37%. Jest to najwyższy współczynnik R^2 jaki udało się uzyskać podczas całego badania. Również kryteria informacyjne dla modelu M_WMNK przyjmują wartość najmniejszą. Na podstawie wartości p-value dla testu F możemy stwierdzić, że współczynnik determinacja jest istotny. Model w 78,37% obrazuje rzeczywistość.

3.17 Ostateczna postać modelu

Ostateczna postać modelu kształtuje się następująco:

$$lnY = 4{,}582 - 0{,}005X_3 + 0{,}0002X_5 + 0{,}036X_7 + 0{,}141X_{11} + 0{,}0005X_{12}$$

Interpretacja:

- Wzrost ilości lat, które upłynęły od ostatniego remontu domu o jednostkę, ceteris pribus, przekłada się na spadek ceny o 0,005%
- Zmiana powierzchni domu o jedną stopę kwadratową, ceteris pribus, przegada się na zmianę ceny domu o 0,0002%
- Zmiana ilość pełnych łazienek o jednostkę w domu, ceteris pribus, przekłada się na zmianę ceny o 0,036%
- Zmiana ilości kominków w domu o jednostkę, ceteris pribus, przekłada się na zmianę ceny o 0,141%
- Zmiana powierzchni garażu o jedną stopę kwadratową, ceteris pribus, przekłada się na zmianę ceny domu o 0,0005%.

3.18 Prognoza przedziałowa

```
? scalar down = SalePrice_predict_meanl_wartosc  - critical(t, $df,0.025) * \
  blad_prognozy
Zamieniono skalar down = 179,843
? scalar up = SalePrice_predict_meanl_wartosc  + critical(t, $df,0.025) * \
  blad_prognozy
Zamieniono skalar up = 180,359
```

Rysunek 35

Na postawie listingu powyżej można stwierdzić, że na poziomie ufności równym 95% cena sprzedaży domu będzie mieściła się w przedziel od 179,843 tysiące dolarów do 180,359 tysięcy dolarów amerykańskich.

3.19 Błędy prognozy

ME = -21,0707 – prognoza jest przeszacowana i przeciętnie zawyża cenę sprzedaży domu o 21,071tys. dolarów

MAE = 24,4335 – prognoza ceny domów średnio różni się od wartości rzeczywistej o 24,4334 tys. dolarów

RMSE = 31,1153 – zdecydowanie większa wartość współczynnika od MAE oznacza, że w zbiorze występuje dużo wartości skrajnych

MAPE = 0,236219 – prognoza średnio różni się od wartości rzeczywistej o 24%

4. Podsumowanie

Badanie wykazało, że nie wszystkie zmienne wybrane do estymacji modelu mają w rzeczywistości istotny wpływ na cenę sprzedaży domu. Większość wybranych zmiennych objaśniających była silnie skorelowana ze zmienną objaśnianą. Z macierzy korelacji można wywnioskować, że im większa powierzchnia domu w stopach kwadratowych tym wyższa cena sprzedaż domu. Cenę sprzedaży domu obniża ilość lat, które upłynęły od wybudowania oraz od ostatniego remontu domu. Istotny wpływ na cenę domu ma również powierzchnia garażu. Występowanie klimatyzacji centralnej w istotny sposób nie zmienia ceny sprzedaży.

Analizują postać modelu i wartości oszacowań współczynników modelu można wysnuć wnioski, że cena sprzedaży domu w bardzo niewielkim stopniu zależy od wybranych parametrów technicznych domów. Zapewne decydujący wpływ na cenę nieruchomości ma sytuacja na rynku finansowym, na rynku nieruchomości oraz wysokość oprocentowania kredytów hipotecznych. To aktualne nastroje na rynku determinują zachowania inwestorów i w naturalny sposób kształtują ceny.

5. Spis tabel

6. Spis rysunków

7. Bibliografia :

1. „Ekonometria”, G.S.Maddal, Wydawnictwo Naukowe PWN, Warszawa 2006
2. „Statystyka od podstaw” J.Jóźwiak, J.Podgórski, Warszawa 1998
3. „Weryfikacja modelu”, źródło: chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 29.05.2021
4. „Podstawy Ekonometrii z elementami algebry liniowej”, Eligiusz W.Nowakowski, źródło: https://wszechnicapolska.edu.pl/dokumenty/wydawnictwo/2011-E-W-Nowakowski-Podstawy-ekonometrii-z-elementami-algebry-liniowej.pdf, data dostępu: 28.05.2021
5. „Wybrane metody estymacji i weryfikacji jednorównaniowych modeli regresji”, Adam Szulc, wrzesień 2018, źródło: chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 27.05.2021
6. „Weryfikacja modelu ekonometrycznego -teoria”, Marta Chylińska, źródło: chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 27.05.2021
7. „Centralne Twierdzenie Graniczne”, Wikipedia, źródło: https://pl.wikipedia.org/wiki/Centralne_twierdzenie_graniczne, data dostępu: 27.05.2021.
8. „Specyfikacja i weryfikacja modelu liniowego, dobór zmiennych objaśniających” Barbara Jasiulis-Gołdyn, data dostępu: 27.05.2021
9. „Zastosowanie metody Hellwiga do konstrukcji modelu ekonometrycznego dla stóp zwrotu z funduszy inwestycyjnych”, Adam Kopiński, Dariusz Porębski, Lublin, 2014, data dostępu: 27.05.2021„Wzór na R-kwadrat, współczynnik determinacji w modelu regresji liniowej”, Aukowiec.org, data dostępu: 27.05.2021
10. „Weryfikacja liniowego modelu jednorównaniowego”, Jakub Mućk, data dostępu: 29.05.2021
11. „Weryfikacja modelu”, źródło: chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 29.05.2021
12. „Ekonometria”, chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 29.05.2021
13. „Metody Ekonometryczne, Estymator GLS (UMNK)”, Jakub Mućk, źródło: chrome-extension://mhjfbmdgcfjbbpaeojofohoefgiehjai/index.html, data dostępu: 31.05.2021