Model Regresji w R Studio

Igor Gryzło 222492

Regression Models in Business Analytics

GBS - 2311

Model Regresji cen mieszkań we Wrocławiu

1. Uzasadnienie wyboru modelu i wyjaśnienie jego elementów

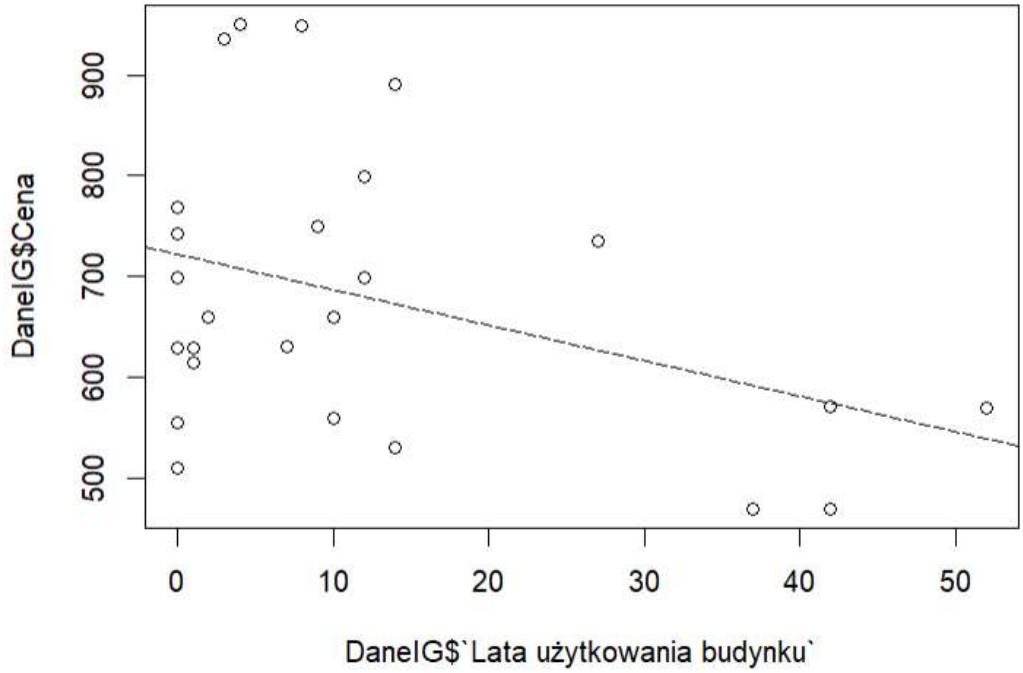
Wybrałem dane opisujące ceny mieszkań we Wrocławiu z powodu mojego zainteresowania rynkiem nieruchomości, od dłuższego czasu śledzę trendy zachodzące na rynku mieszkań. W moich danych posiadam 3 zmienne:

- „Lata użytkowania budynku" – określają ile lat minęło od wybudowania budynku
- „Odległość od Centrum" – określa drogę od rynku głównego we Wrocławiu wyrażoną w kilometrach.
- „Metraż" - określa wielkość mieszkania wyrażoną w metrach kwadratowych.

Wybrałem te dane, ponieważ takie dane są najbardziej istotnymi kryteriami wyboru mieszkania, co znacząco odbija się na jego cenie – Lata użytkowania określają stopień zużycia budynku i infrastruktury, odległość od centrum wpływa na dojazd (czas) do najpotrzebniejszych miejsc w życiu codziennym a metraż wpływa na komfort użytkowania.
Źródło: https://www.otodom.pl
Przygotowania danych polegało na odrzuceniu tych w których nie znałem dokładniej lokalizacji lub data budowy budynku nie była wyszczególniona w ogłoszeniu.

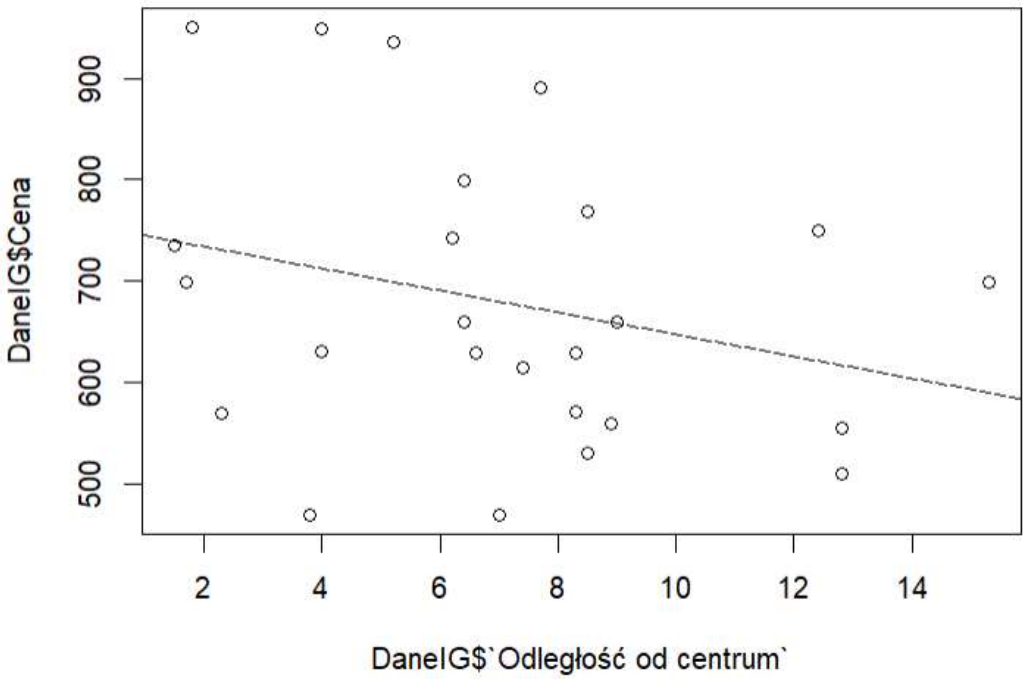
2. Analiza dopasowania modelu

Wykres zależności Ceny i okresu użytkowania budynku:



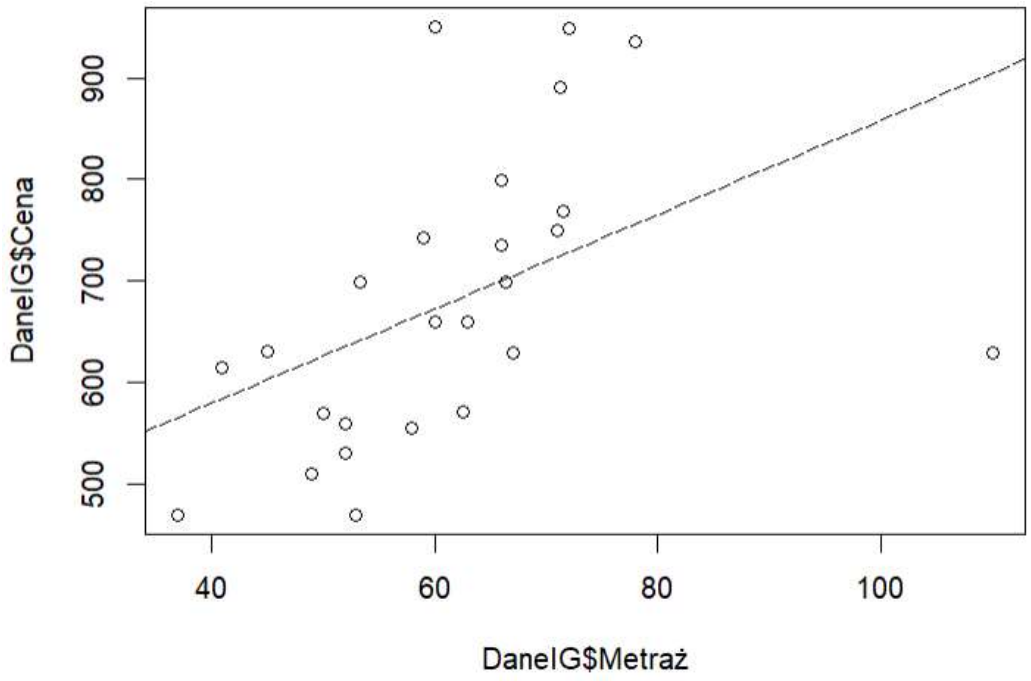
Dla tej zmiennej występuję zależność liniowa ujemna, oznacza to, że wraz ze wzrostem lat użytkowania cena mieszkania spada.

Wykres zależności Ceni i odległości od centrum



Dla tej zmiennej występuje również zależność liniowa ujemna, oznacza to, że wraz ze wzrostem odległości od centrum cena mieszkania maleje.

Wykres zależności Ceny i Metrażu



Dla tej zmiennej występuje zależność liniowa dodatnia, oznacza to, że wraz ze wzrostem metrażu cena mieszkania również się zwiększa.
Z racji tego, że w tym modelu wszystkie zmienne są dobrze dopasowane (p-value poniżej 0,05) Nie zdecydowałem się na dodawanie dodatkowych zmiennych i transformowanie zmiennych.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                642.759    127.515   5.041 5.45e-05 ***
`Lata użytkowania budynku`  -4.225      1.706  -2.477   0.0219 *
`Odległość od centrum`     -19.194      6.790  -2.827   0.0101 *
Metraż                       3.643      1.665   2.188   0.0401 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analiza macierzy korelacji:

- Cena-Lata użytkowania - umiarkowana korelacja ujemna
- Cena-Odległość od centrum – słaba korelacja ujemna
- Cena-Metraż – umiarkowana korelacja dodatnia

	Cena	Lata użytkowania budynku	Odległość od centrum	Metraż
Cena	1.0000000	-0.3760838	-0.27323138	0.46898638
Lata użytkowania budynku	-0.3760838	1.0000000	-0.40023681	-0.32884441
Odległość od centrum	-0.2732314	-0.4002368	1.00000000	0.09412635
Metraż	0.4689864	-0.3288444	0.09412635	1.00000000

Ocena występowania problemu współliniowości – Powyższy wykres dodatkowo pozawala nam ocenić współliniowość, w tym celu musimy się skupić na korelacji między zmiennymi objaśniającymi. Jeśli któraś z tych wartości wynosiłaby powyżej 0,7 lub poniżej -0,7 moglibyśmy zaobserwować taki problem. W przypadku tych danych on nie występuje.

3. **Porównanie Modeli:**
   - Ocena sensowności znaków oszacowanych parametrów – Korzystając z poniższego screena można zauważyć, że wszystkie znaki i wartości parametrów są prawidłowe. (Metraż mieszkań zwiększa ich cenę a większa odległość od centrum i wiekowość budynku ją obniża) Oznacza to, że według tej metody wszystkie modele są równie precyzyjne.

```
#Model ze wszystkimi zmiennymi
coef(ModelX)
             (Intercept) `Lata użytkowania budynku`      `Odległość od centrum`     Metraż
               642.758761                 -4.225125                -19.193814   3.642558
#Model bez zmiennej "Metraż"
Model1 <- lm(Cena~`Lata użytkowania budynku`+`Odległość od centrum`,data=DaneIG)
coef(Model1)
             (Intercept) `Lata użytkowania budynku`      `Odległość od centrum`
               885.441735                 -5.416314                -19.837301
#Model bez zmiennej "Odległość od Centrum"
Model2 <- lm(Cena~`Lata użytkowania budynku`+`Metraż`,data=DaneIG)
coef(Model2)
             (Intercept) `Lata użytkowania budynku`                     Metraż
               471.249917                 -2.330938                3.846432
#Model bez zmiennej "Lata użytkowania budynku"
Model3 <- lm(Cena~`Odległość od centrum`+`Metraż`,data=DaneIG)
coef(Model3)
        (Intercept) `Odległość od centrum`                  Metraż
          463.41734                -12.58891                4.95853
```

   - Z kolejnych screenów (pod spodem) wynika, że Pierwotny model jest najlepiej dopasowany. Ponieważ P-value mają prawie najniższe wartości – oznacza to, że dane bardzo dobrze opisują model. Dodatkowo w 1 modelu wartości Ve i R^2 są najwyższe co oznacza, że model najlepiej opisuje dane. Podsumowując wybieram model nr. 1 ze wszystkimi zmiennymi.

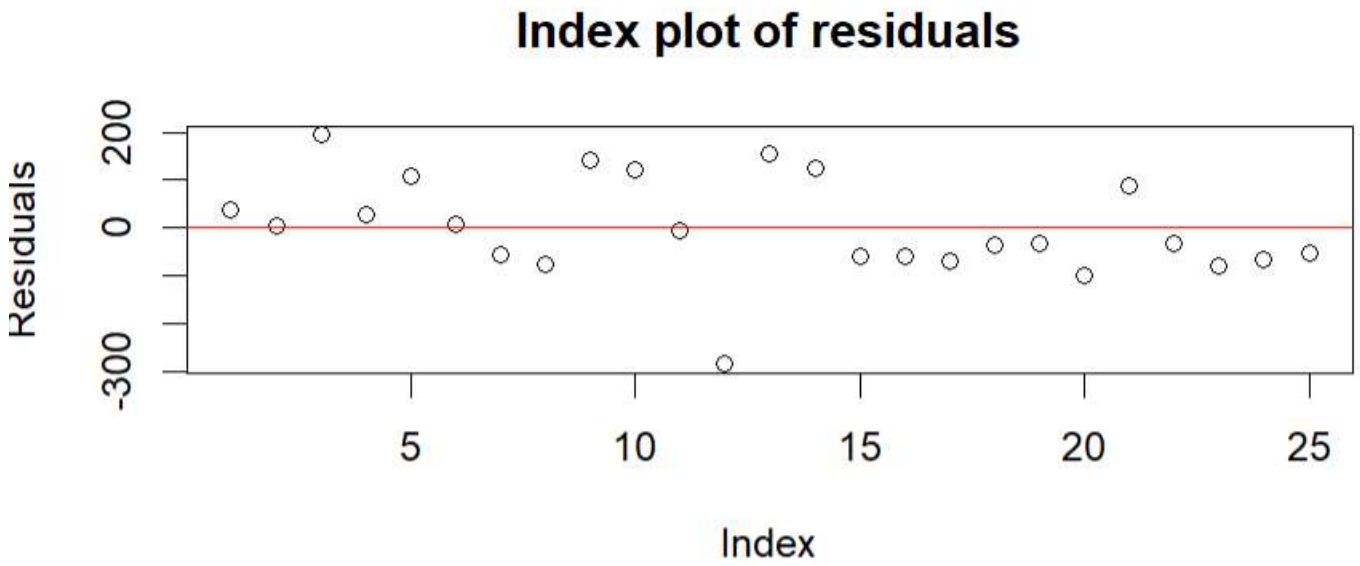
```
#Model ze wszystkimi zmiennymi
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                642.759    127.515   5.041 5.45e-05 ***
`Lata użytkowania budynku`  -4.225      1.706  -2.477   0.0219 *
`Odległość od centrum`     -19.194      6.790  -2.827   0.0101 *
Metraż                       3.643      1.665   2.188   0.0401 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.5 on 21 degrees of freedom
Multiple R-squared:  0.4749,    Adjusted R-squared:  0.3999
F-statistic: 6.332 on 3 and 21 DF,  p-value: 0.003149

#Model bez zmiennej "Metraż"

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                885.442     68.090  13.004 8.37e-12 ***
`Lata użytkowania budynku`  -5.416      1.750  -3.094   0.0053 **
`Odległość od centrum`     -19.837      7.344  -2.701   0.0130 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 120.7 on 22 degrees of freedom
Multiple R-squared:  0.3553,    Adjusted R-squared:  0.2966
F-statistic: 6.061 on 2 and 22 DF,  p-value: 0.008003
```

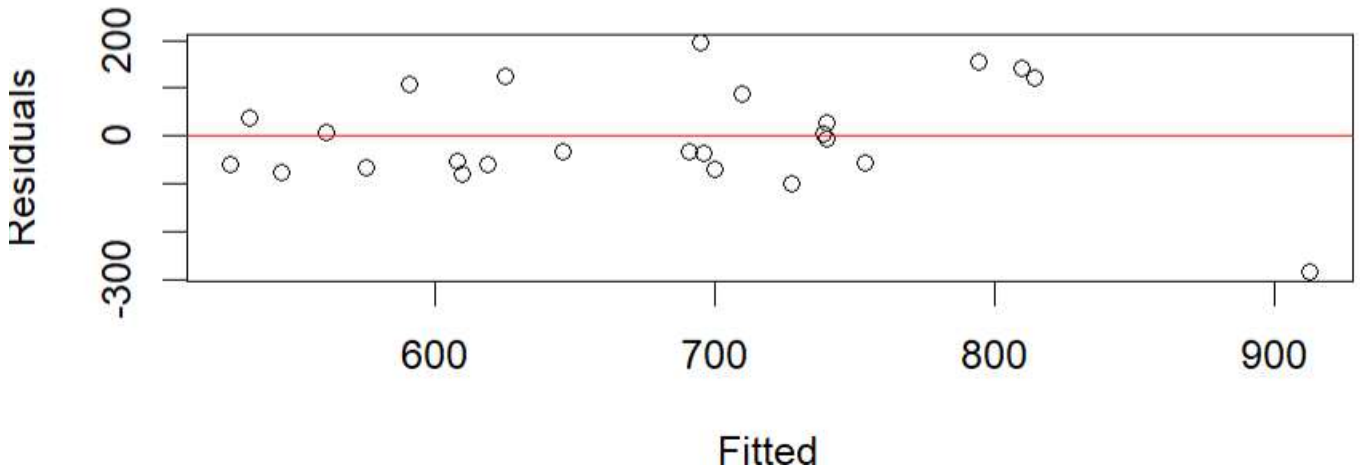
```
#Model bez zmiennej "Odległość od Centrum"
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                471.250    128.748   3.660  0.00138 **
`Lata użytkowania budynku`  -2.331      1.801  -1.294  0.20900
Metraż                       3.846      1.909   2.014  0.05635 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 128 on 22 degrees of freedom
Multiple R-squared:  0.2751,    Adjusted R-squared:  0.2092
F-statistic: 4.175 on 2 and 22 DF,  p-value: 0.02903

#Model bez zmiennej "Lata użytkowania budynku"
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              463.417    116.562   3.976  0.00064 ***
`Odległość od centrum`   -12.589      6.935  -1.815  0.08312 .
Metraż                     4.959      1.752   2.830  0.00975 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 123.8 on 22 degrees of freedom
Multiple R-squared:  0.3216,    Adjusted R-squared:  0.2599
F-statistic: 5.214 on 2 and 22 DF,  p-value: 0.01401
```

4. Analiza otrzymanych wyników empirycznych

- Interpretacja parametrów: (summary dla modelu z 3 zmiennymi na screenie u góry)
  - P-Value – im mniejszą wartość przyjmuje tym zmienna lepiej opisuje model, przyjmujemy, że te poniżej 0,05 są istotne statystycznie.
  - Błąd standardowy (Std. Error) – oznacza ile średnio myli się model dla każdej zmiennej na plus lub minus
  - Residual standard error (Błąd standardowy reszt): Wartość błędu standardowego reszt wynosi 111.5 na 21 stopniach swobody. Błąd standardowy reszt służy do oceny jakości prognoz modelu. Im niższa wartość, tym lepsza prognoza.
- Interpretacja testów:
  - Test T i Test F: summary również pokazuję 2 wartości które używane są w tym teście, w modelu p-value wynosi 0,003149 co oznacza to, że grupa zmiennych jest istotna dla modelu, ponadto F-Statistics wynosi 6,8 co również świadczy o dobrym dopasowaniu modelu.
- Analiza miar dopasowania – Im większa wartość R^2 tym lepsze dopasowanie modelu. W przypadku tego modelu wynosi 0.4749. Adjusted R^2 uwzględnia liczbę zmiennych niezależnych w modelu im mniejszy tym więcej niezależnych w modelu 0,3999.

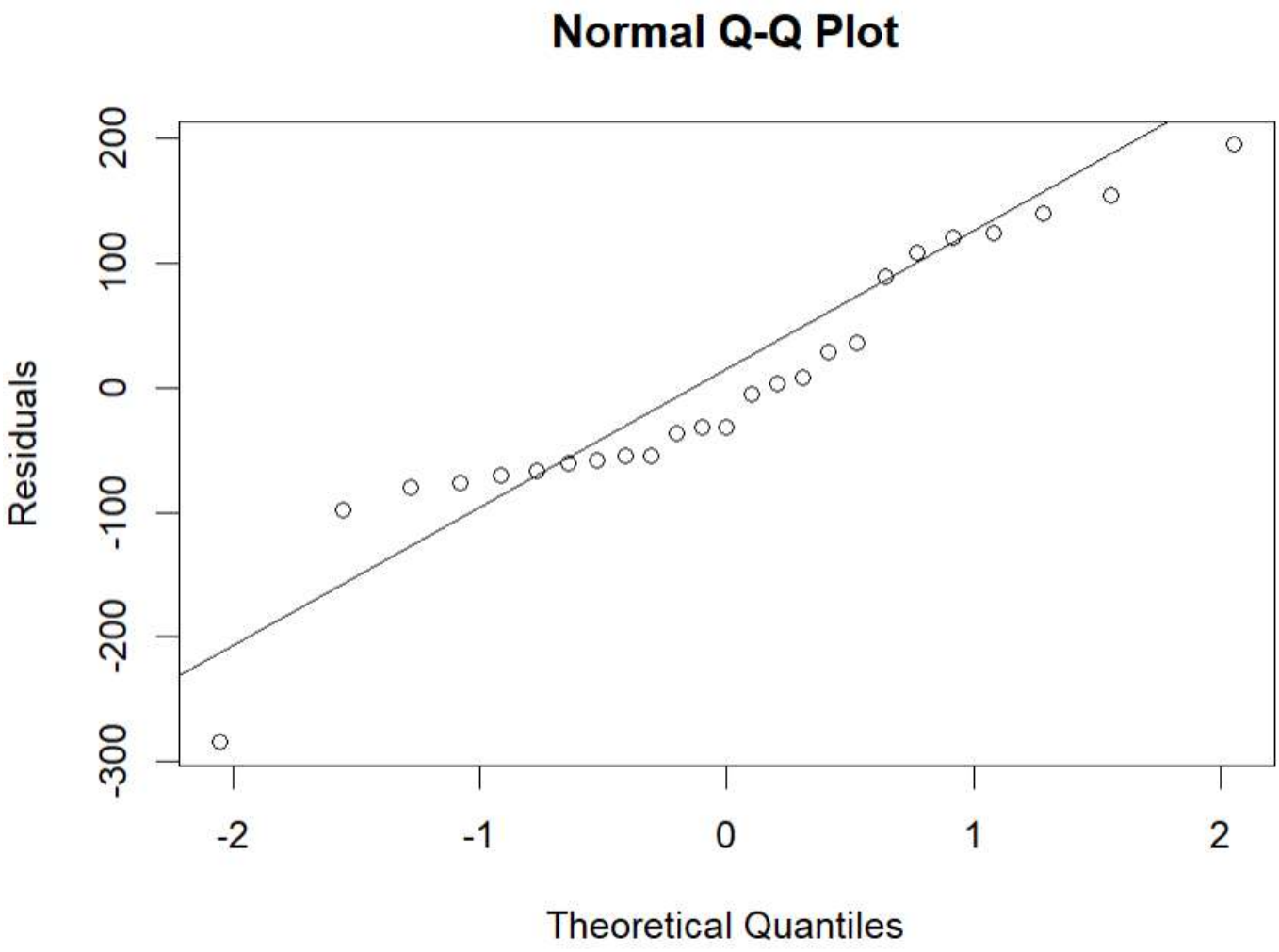
5. Diagnostyka Reszt
    - Wykres Reszt:





Na wykresie można zaobserwować wartości oscylujące wokół zera. W Okolicach indeksu 12 można zaobserwować skok co może oznaczać, że model jest zakłócony przez wyjątkowe wartości.

- Wykres Kwantyl-Kwantyl – na wykresie można zaobserwować punkty które są blisko prostej, oznacza to, że rozkład reszt jest normalny.



- Test niezależności

```
        Runs Test

data:  Model1X$res
statistic = -2.0871, runs = 8, n1 = 12, n2 = 12, n = 24, p-value = 0.03688
alternative hypothesis: nonrandomness
```

Z tego testu możemy zaobserwować, że p-value wynosi 0.03688, oznacza to, że istnieje mniej niż 3,7% szansy, że reszty są niezależne. Zważając na poziom istotności 0,05 mamy wystarczające dowody aby odrzucić hipotezę o niezależności reszt. Wartość statistic oznacza, że nasze dane nie są niezależne więc weźmiemy pod uwagę 3 inne modele (z 1 wyrzucona zmienna).

- Porównanie modeli pod kątem niezależności

```
        Runs Test

data:  ModelX$res
statistic = -2.0871, runs = 8, n1 = 12, n2 = 12, n = 24, p-value = 0.03688
alternative hypothesis: nonrandomness

> runs.test(Model1$res)

        Runs Test

data:  Model1$res
statistic = -0.43693, runs = 11, n1 = 11, n2 = 11, n = 22, p-value = 0.6622
alternative hypothesis: nonrandomness

> runs.test(Model2$res)

        Runs Test

data:  Model2$res
statistic = 0, runs = 13, n1 = 12, n2 = 12, n = 24, p-value = 1
alternative hypothesis: nonrandomness

> runs.test(Model3$res)

        Runs Test

data:  Model3$res
statistic = 0, runs = 13, n1 = 12, n2 = 12, n = 24, p-value = 1
alternative hypothesis: nonrandomness
```

Z powyższego screena można wywnioskować, że najbardziej niezależne reszty posiada model 2 i 3. W tym wypadku wybieram do dalszej kontynuacji projektu Model nr 3 czyli z wykluczoną Lata użytkowania budynku.

6. Zastosowanie Modelu:
   - Określenie współczynników dla danych zmiennych
   - Wyliczenie przewidywanej ceny (może się różnić o odchylenie standardowe)

Przewidywana cena dla mieszkania oddalonego o 10 km od centrum i powierzchni 43 metry kwadratowe

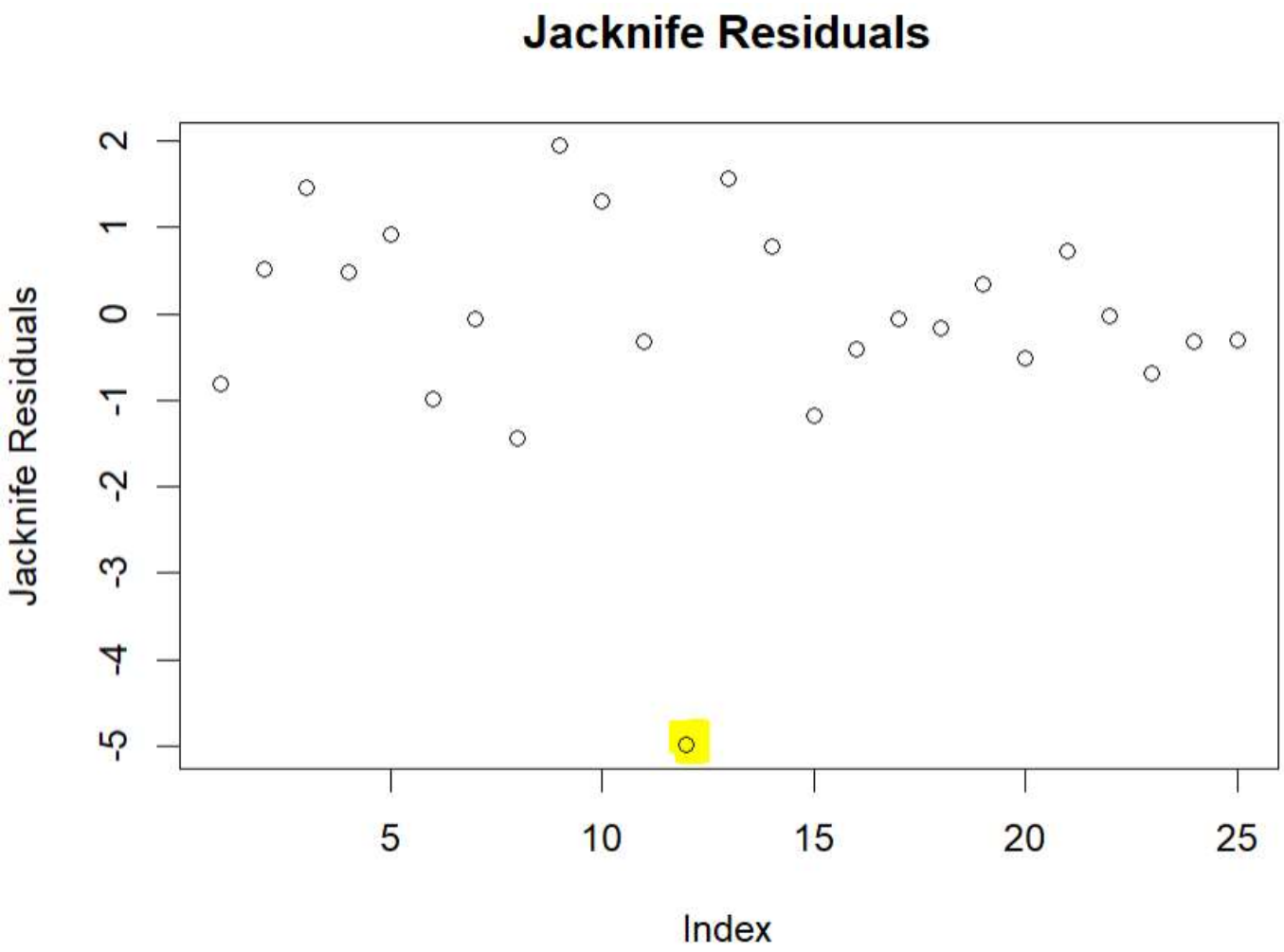
```
> print(our_forecast1)
(Intercept)
   673.4996
```

Przewidywana cena dla mieszkania oddalonego o 10 km od centrum i powierzchni 67 metrów kwadratowych

```
> print(our_forecast)
(Intercept)
    760.921
```

7. Analiza i wykluczenie wartości odstających

```
> #Analiza Outliers
> jack <- rstudent(Model3)
> plot(jack,ylab="Jacknife Residuals",main="Jacknife Residuals")
> jack[abs(jack)==max(abs(jack))]
       12
-4.977329
```



Z wykresu i tekstu w konsoli wynika, ze index 12 jest wartością odstającą.

Porównanie wyników po wyrzuceniu zmiennej 12.

```
> summary(Model3_Sub)

Call:
lm(formula = Model3, data = DaneIG, subset = -12)

Residuals:
     Min       1Q   Median       3Q      Max
-143.666  -41.209   -6.789   52.253  178.720

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            141.235    103.538   1.364  0.18698
`Odległość od centrum`  -15.732      4.849  -3.244  0.00388 **
Metraż                   10.973      1.713   6.404 2.39e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 85.86 on 21 degrees of freedom
Multiple R-squared:  0.6871,    Adjusted R-squared:  0.6573
F-statistic: 23.06 on 2 and 21 DF,  p-value: 5.026e-06

> summary(Model3)

Call:
lm(formula = Cena ~ `Odległość od centrum` + Metraż, data = DaneIG)

Residuals:
     Min       1Q   Median       3Q      Max
-296.769  -62.151   -7.504   88.889  211.731

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            463.417    116.562   3.976  0.00064 ***
`Odległość od centrum`  -12.589      6.935  -1.815  0.08312 .
Metraż                    4.959      1.752   2.830  0.00975 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 123.8 on 22 degrees of freedom
Multiple R-squared:  0.3216,    Adjusted R-squared:  0.2599
F-statistic: 5.214 on 2 and 22 DF,  p-value: 0.01401
```

Powyższy screen pokazuję, że usunięcie danej 12 znacząco poprawiło dopasowanie modelu i p-value.