Gliwice, 10.06.2020 r.

Sprawozdanie z projektu

Metody Statystyczne

Temat: **Projekt 8**

Czembor Stanisław

Kałuziński Ryszard

Knura Tomasz

Mika Paweł

Pawlus Benedykt

Wylęgły Błażej

Problem

Młode małżeństwo zainteresowane kupnem mieszkania w Katowicach na podstawie dostępnych ofert zebrało dane dotyczące: ceny mieszkania, jego powierzchni i liczby pokoi. Dane z losowo wybranych ofert są następujące:

```
(345;57;2),  (253;33;2),  (183;45;3),  (412;97;4),  (398;94;4),  (375;96;4),
(399;85;4),  (224;49;3),  (362;99;3),  (229;44;2),  (324;70;3),  (378;66;2),
(222;44;2),  (331;87;4),  (303;73;4),  (209;31;1),  (331;63;3),  (267;51;2),
(182;49;2),  (301;53;2),  (310;62;2),  (241;52;2),  (190;44;1),  (183;43;2),
(261;72;3),  (341;79;4),  (340;68;3),  (318;79;3),  (405;78;3),  (245;49;2),
(210;54;3),  (236;57;2),  (211;52;2),  (186;50;1),  (186;54;2),  (262;59;3),
(187;50;3),  (124;52;2),  (306;56;2),  (292;57;3),  (314;55;2),  (167;51;2),
(256;58;3),   (260;51;2),   (130;58;3), (244;51;2), (238;58;3),  (296;55;3),
(248;58;2).
```

Polecenia do wykonania

1. Przedstawić graficznie dane dotyczące ceny mieszkania oraz powierzchni mieszkania:
   - bez uwzględnienia liczby pokoi,
   - z uwzględnieniem liczby pokoi.

2. Wyznaczyć modele regresji liniowej przedstawiające zależności:
   - ceny mieszkania od powierzchni mieszkania,
   - ceny mieszkania od powierzchni mieszkania i liczby pokoi.

3. Ocenić wyznaczone modele regresji.

4. Dla każdego z wyznaczonych modeli dokonać kompletnej diagnostyki.

5. Na wykresach prezentujących zbiór par (cena mieszkania powierzchnia mieszkania) dodać proste regresji oraz krzywe prezentujące korytarze ufności: dla prostej regresji oraz ceny mieszkania. Rozpatrzyć przypadki nieuwzględniania oraz uwzględniania liczby pokoi

6. Opracować histogramy rezyduów.

7. Sprawdzić, czy rezydua mają rozkład normalny.

Podstawy teoretyczne

- Miary położenia
  - średnia arytmetyczna

    $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - kwantyl

    kwantylem rzędu p (0 < p < 1) nazywamy liczbę x(p)

    $$P(X \leq x(p)) \geq p$$

    $$P(X \geq x(p)) \geq 1 - p$$

  - kwartyle - kwantyle rzędów p = 0.25, p = 0.5, p = 0.75
  - mediana - kwantyl rzędu p = 0.5

- Miary zróżnicowania
  - wariancja z próby

    $$S^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

  - odchylenie standardowe

    $$S$$

- Miary koncentracji
  - kurtoza

    $$Krt = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^4}{S^4}$$

- Statystyka Kołmogorowa jest miarą rozbieżności pomiędzy rozkładem empirycznym i hipotetycznym:

  $$D_n = \max_x |F_0(x) - F_n(x)|$$

  $F_0(x)$ – dystrybuanta rozkładu hipotetycznego.
  $F_n(x)$ – dystrybuanta empiryczna.
  Do wyznaczenia dystrybuanty rozkładu hipotetycznego, należy wykorzystać standaryzację rozkładu do postaci rozkładu normalnego.
  Następnie, należy wyznaczyć dystrybuantę empiryczną oraz dla każdej wartości wyznaczyć wektor modułu różnicy.
  Z tak przygotowanego wektora, należy wyznaczyć wartość maksymalną oraz odczytać wartość krytyczną $d_n(1-\alpha)$, gdzie obszarem krytycznym jest przedział:

  $$K_0 = < d_n(1-\alpha), 1 >$$

- Regresja liniowa

  Metoda zakłada, że pomiędzy zmiennymi wejściowymi (objaśniającymi) i wyjściowymi (objaśnianymi) istnieje mniej lub bardziej wyrazista zależność liniowa.

  Przewidywanie wartości zmiennych objaśnianych (y) na podstawie wartości zmiennych objaśniających (x) jest możliwe dzięki znalezieniu tzw. modelu regresji.

  W praktyce polega to na podaniu równania prostej, zwanej prostą regresji o postaci:

$$\hat{y} = b_0 + b_1 x$$

  gdzie:

  $\hat{y}$ - szacowana wartość zmiennej objaśnianej

  b_0 - punkt przecięcia linii regresji z osią y

  b_1 - nachylenie linii regresji

  b_0, b_1 - współczynniki regresji

  Często zmienna objaśniana zależna jest nie od jednej ale od kilku (wielu) zmiennych objaśniających. Będziemy zatem rozważać ogólne równanie regresji postaci:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m$$

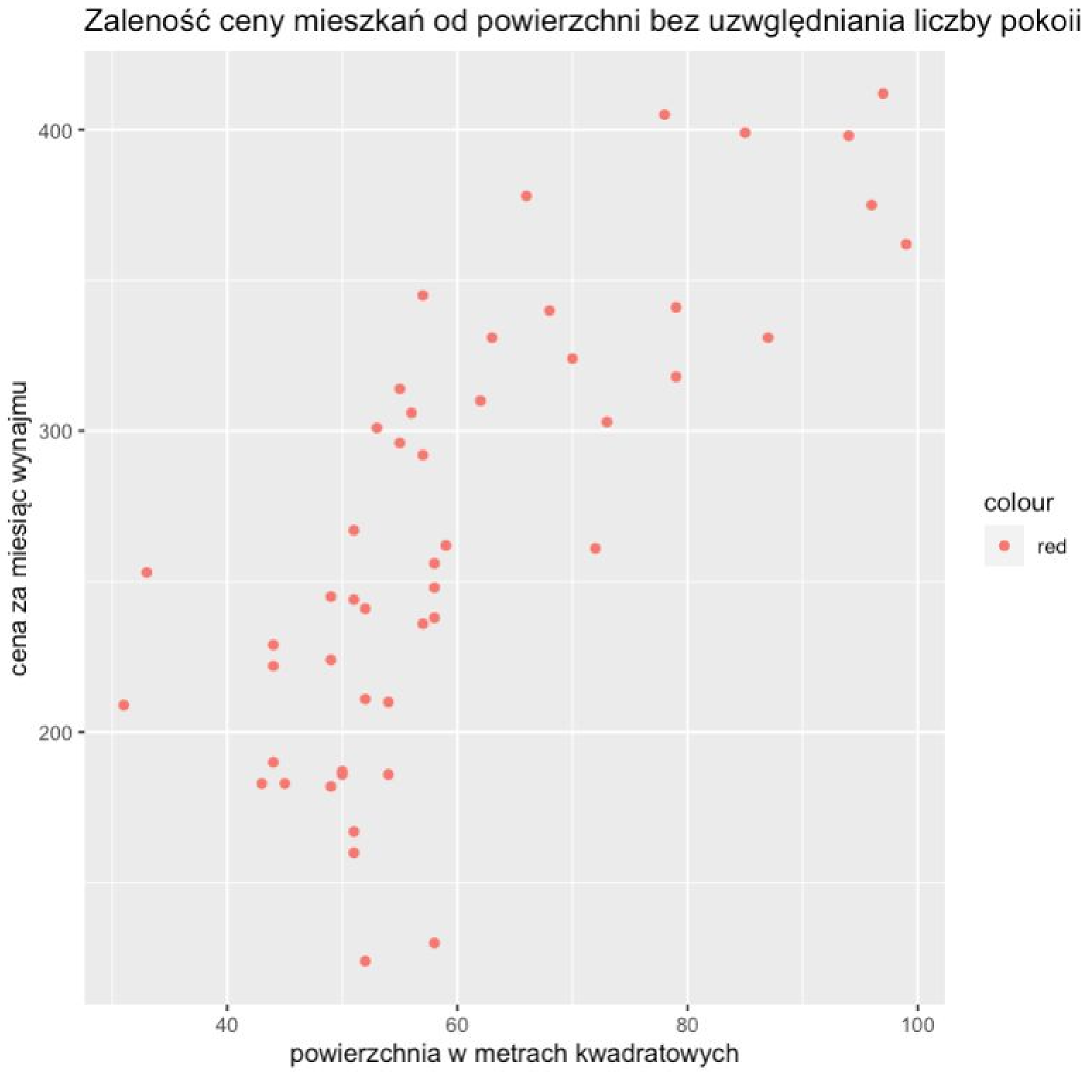
  gdzie m oznacza liczbę zmiennych objaśniających

Realizacja Zadań

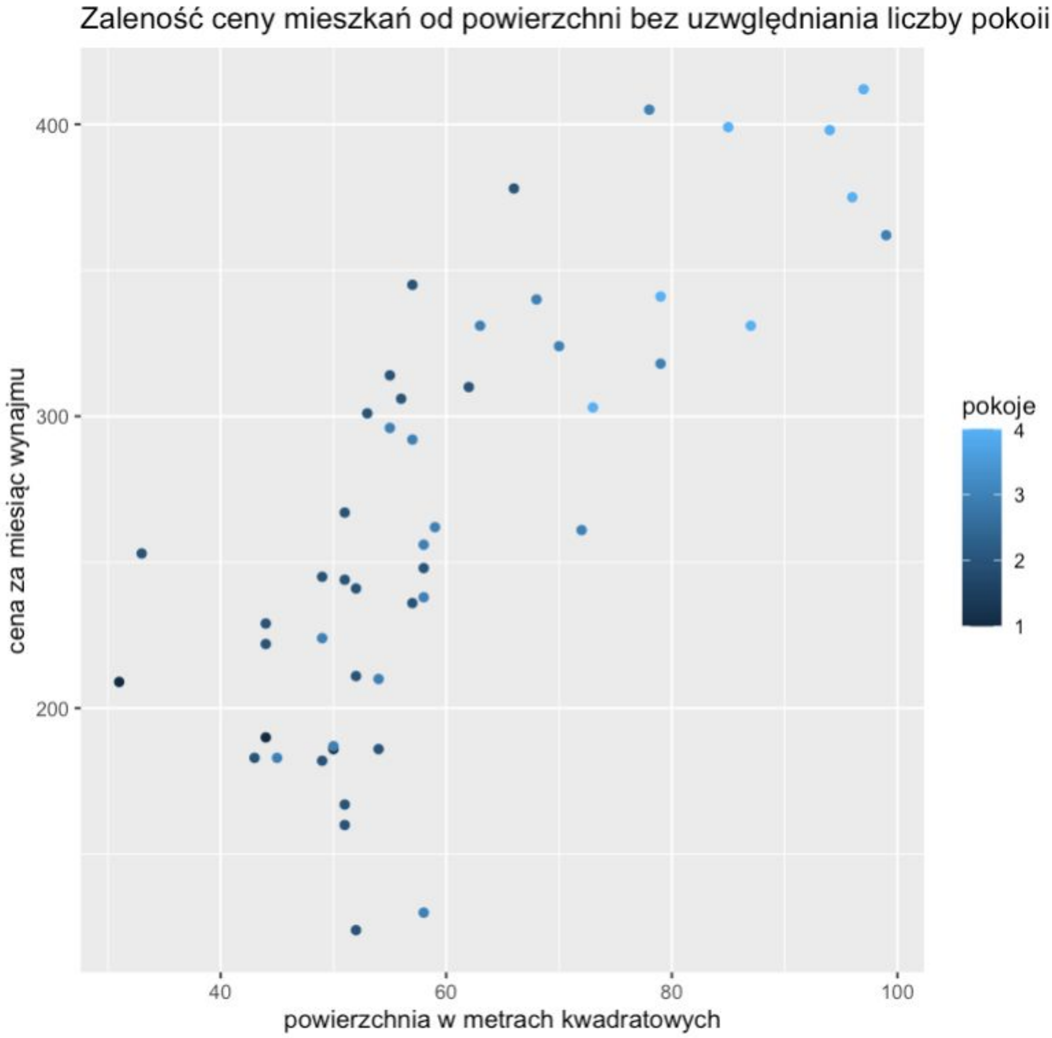
1. Na podstawie danych dostarczonych w treści zadania, sporządziliśmy graficzne reprezentacje wielkości reprezentujących cenę oraz powierzchnię mieszkań. W celu sporządzenia wykresów, została wykorzystana funkcja ggplot.

   Zgodnie z poleceniem, zostały sporządzone dwa wykresy dotyczące stosunku ceny mieszkania od powierzchni:

   - bez uwzględniania liczby pokoi



- z uwzględnieniem liczby pokoi



2. Do wyznaczenia modeli regresji liniowej, wykorzystaliśmy funkcję lm. Jej wyniki prezentują się następująco
   - dla zależności bez uwzględnienia liczby pokoi

```
> reg1

Call:
lm(formula = cena ~ powierzchnia, data = mieszkania)

Coefficients:
 (Intercept)  powierzchnia
      52.433         3.577

> summary(reg1)

Call:
lm(formula = cena ~ powierzchnia, data = mieszkania)

Residuals:
     Min       1Q   Median       3Q      Max
-129.914  -30.410   -1.492   35.776   89.467

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   52.4329    27.7067   1.892   0.0646 .
powierzchnia   3.5773     0.4452   8.035 2.26e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.63 on 47 degrees of freedom
Multiple R-squared:  0.5787,    Adjusted R-squared:  0.5698
F-statistic: 64.57 on 1 and 47 DF,  p-value: 2.256e-10
```

- dla zależności z uwzględnieniem liczby pokoi

```
> reg2

Call:
lm(formula = cena ~ powierzchnia + pokoje, data = mieszkania)

Coefficients:
 (Intercept)  powierzchnia        pokoje
      53.731         3.801        -5.734

> summary(reg2)

Call:
lm(formula = cena ~ powierzchnia + pokoje, data = mieszkania)

Residuals:
     Min       1Q   Median       3Q      Max
-126.974  -28.903    1.225   38.827   86.093

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   53.7314    28.1288   1.910   0.0624 .
powierzchnia   3.8007     0.7015   5.418 2.14e-06 ***
pokoje        -5.7336    13.8253  -0.415   0.6803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50.07 on 46 degrees of freedom
Multiple R-squared:  0.5803,    Adjusted R-squared:  0.5621
F-statistic:  31.8 on 2 and 46 DF,  p-value: 2.126e-09
```

3. Funkcja summary wypisuje na ekranie wartości modelu regresji liniowej uzyskanego za pomocą funkcji lm umożliwiającego jego analizę:

   - Rozkład rezyduł, widoczny w 5 podpunktach sekcji residuals, nie jest symetryczny. Oznacza to, że niektóre wartości oszacowane modelu różnią się znacząco od wartości rzeczywistych

   - Z współczynników zawartych w sekcji coefficients możemy utworzyć równania regresji:

     - dla zależności bez uwzględnienia liczby pokoi
       $cena = 3.5773 \cdot powierzchnia + 52.4329$

       interpretując współczynnik nachylenia prostej regresji b = 3,5773 powiemy, że cena rośnie o 3,5773 punktu, jeśli powierzchnia rośnie o jedną jednostkę

     - dla zależności z uwzględnieniem liczby pokoi
       $cena = 3.8007 \cdot powierzchnia + -5.7336 \cdot pokoje + 53.7314$

       interpretując współczynnik nachylenia prostej regresji b_1 = 3,8007 powiemy, że cena rośnie o 3,8007 punktu, jeśli powierzchnia rośnie o jedną jednostkę. Zakładamy przy tym, że liczba pokoi jest stała.
       Z kolei interpretacja współczynnika b_2 = -5,7336 jest taka, że cena maleje o 5,7336 punktu, jeśli liczba pokoi rośnie o jedną jednostkę a powierzchnia jest stała.

   - Z sekcji Multiple R Squared możemy odczytać współczynnik determinacji modelu R^2. W modelu nie uwzględniającym liczby pokoi wynosi on 57,87%, natomiast uwzględniając pokoje otrzymaliśmy 58,03% czyli po dodaniu zmiennej objaśniającej - liczby pokoi możemy wyjaśnić dodatkowo zaledwie 0,16% zmienności ceny.

- Błąd oszacowania obliczany jako standardowy błąd oszacowania (Residual standard error) wynosi 49,66 bez uwzględniania liczby pokoi oraz 50,07 z uwzględnieniem. Oznacza to, że estymacja ceny mieszkania na podstawie zawartości powierzchni zwykle różni się od właściwej wartości o 49,66 punktu, w przypadku estymacji na podstawie powierzchni oraz liczby pokoi wartość ta zwiększyła się do 50,07. Możemy więc przypuszczać, że zmienna liczby pokoi nie jest znacząco przydatna.

4. W celu przeprowadzenia kompletnej diagnostyki każdego wyznaczonego modelu, wykonaliśmy następujące testy:

   W opisie przeprowadzanych testów, posłużyliśmy się następującym (skrótowym) nazewnictwem poszczególnych modeli regresji:

   - model pierwszy - model regresji dla statystyki bez uwzględnienia liczby pokoi
   - model drugi - model regresji dla statystyki z uwzględnieniem liczby pokoi

   - Testy jednorodności wariancji
     - *Test Godfelda-Quandta* - polega na porównaniu wariancji modelu w dwóch grupach. Na podstawie wykresu kwadratów reszt oceniamy, czy reszty modelu da się podzielić na 2 części - początkową i końcową - o wyraźnie różnych wartościach kwadratów reszt. Jeżeli można wydzielić takie 2 części, to testujemy hipotezę o równości wariancji w obu częściach

       Dla modelu pierwszego:
       p = 0,1431, p > 0,05, zatem wariancje są homogeniczne
       Dla modelu drugiego:
       p = 0,1271, p > 0,05, zatem wariancje są homogeniczne

     - *Test Breuscha-Pagana* - Pozwala sprawdzić homoskedastyczność, czyli czy jakakolwiek zmienna różni się od innych wartością wariancji.

       Dla modelu pierwszego:
       p = 0,344, p > 0,05, zatem wariancje są homogeniczne
       Dla modelu drugiego:
       p = 0,4348, p > 0,05, zatem wariancje są homogeniczne

     - *Test Harrisona-McCabego* - pozwala na weryfikację hipotezy o posiadaniu tej samej, skończonej wartości składnika losowego

       Dla modelu pierwszego:
       p = 0,12, p > 0,05, zatem wariancje są homogeniczne
       Dla modelu drugiego:
       p = 0,124, p > 0,05, zatem wariancje są homogeniczne

   - Testy niezależności
     - *Test Durbina-Watsona* określający, czy skonstruowany model regresji jest dobrze dopasowany.

W celu przeprowadzenia testu, należy skorzystać z tablic rozkładu Durbina-Watsona. W przypadku rozpatrywanych przez nas modeli, przyjęliśmy liczbę predyktorów równą 2.

Dla tak określonej liczby predyktorów, przyjęliśmy wartości dl i dg, określające przedział wyników, dla których nie można stwierdzić, czy zachodzi autokorelacja reszt:
dl=1,46 dg=1,63

Dla modelu pierwszego:
  DW = 1,5975, dg > DW > dl, zatem nie ma autokorelacji między resztami
Dla modelu drugiego:
  DW =1,5975, dg > DW > dl, zatem nie ma autokorelacji między resztami

- *Test Breuscha-Godfreya* - W odróżnieniu od testu Durbina-Watsona, test ten potrafi wykryć autokorelację w dowolnej kolejności

  Dla modelu pierwszego:
    p = 0,2392 p > 0,05 i df = 1, zatem występuje autokorelacja w jednej kolejności
  Dla modelu drugiego:
    p = 0,2349, p > 0,05 i df = 1, zatem występuje autokorelacja w jednej kolejności

- Testy liniowości
  - *Test Harveya-Colliera* - pozwala określić, czy można przypuszczać, że dana zależność jest liniowa.

    Dla modelu pierwszego:
      p = 0,005935 p < 0,05, zatem nie mamy podstaw do odrzucenia hipotezy zerowej, że zależność nie jest liniowa
    Dla modelu drugiego:
      p = 0,01495, p < 0,05, zatem nie mamy podstaw do odrzucenia hipotezy zerowej, że zależność nie jest liniowa

  - *Test Rainbow* - pozwala określić, czy nawet jeśli badana zależność nie jest liniowa, to czy można stwierdzić że wiarygodnym dopasowaniem będzie dopasowanie wypracowane na podstawie wybranego fragmentu tego modelu.

    Dla modelu pierwszego:
      p = 0,3591 p > 0,05, zatem mamy podstawę do przypuszczenia, że można wypracować wiarygodne dopasowanie na podstawie fragmentu modelu
    Dla modelu drugiego:
      p = 0,4909, p < 0,05, zatem mamy podstawę do przypuszczenia, że można wypracować wiarygodne dopasowanie na podstawie fragmentu modelu

- *Test RESET Ramseya* - stosowany w celu sprawdzenia, czy to liniowa postać modelu (względem funkcji kwadratowej lub sześciennej) jest najlepszym możliwym do wybrania modelem

  Dla modelu pierwszego:

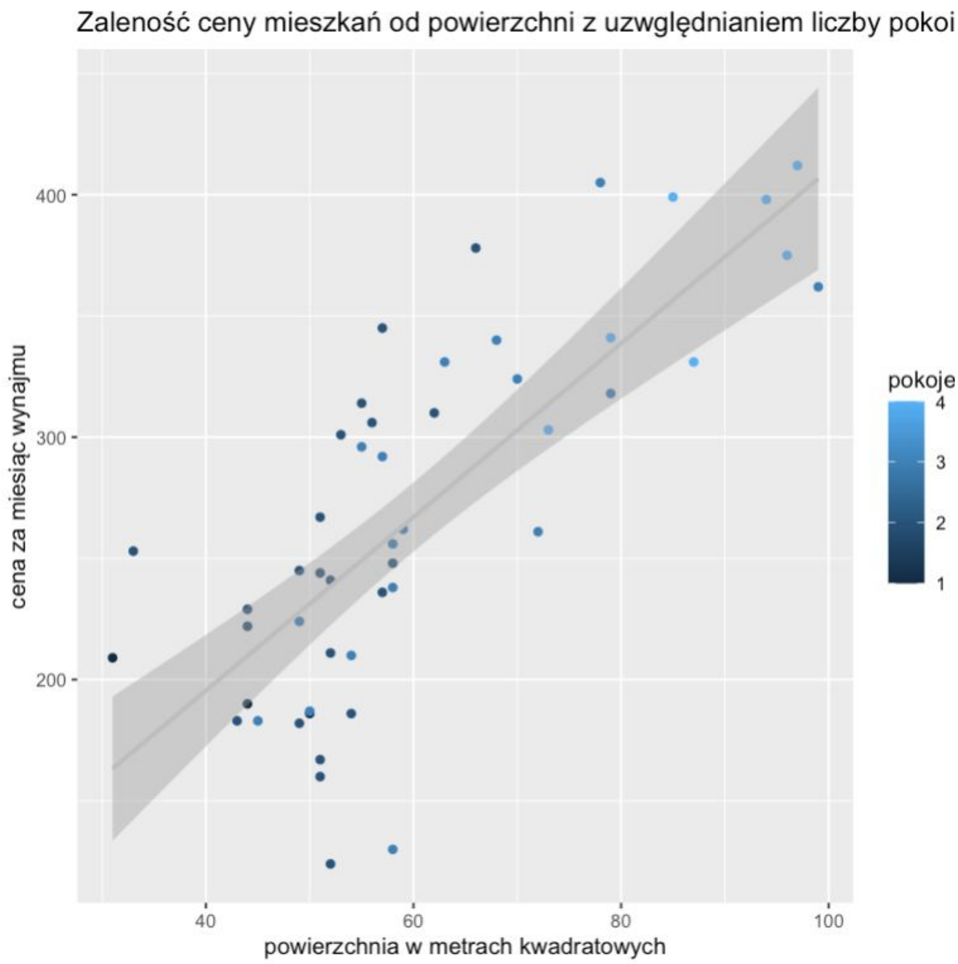
  p = 0,5829 p > 0,05, zatem mamy podstawę do przypuszczenia, że najlepszym modelem nie jest model liniowy

  Dla modelu drugiego:

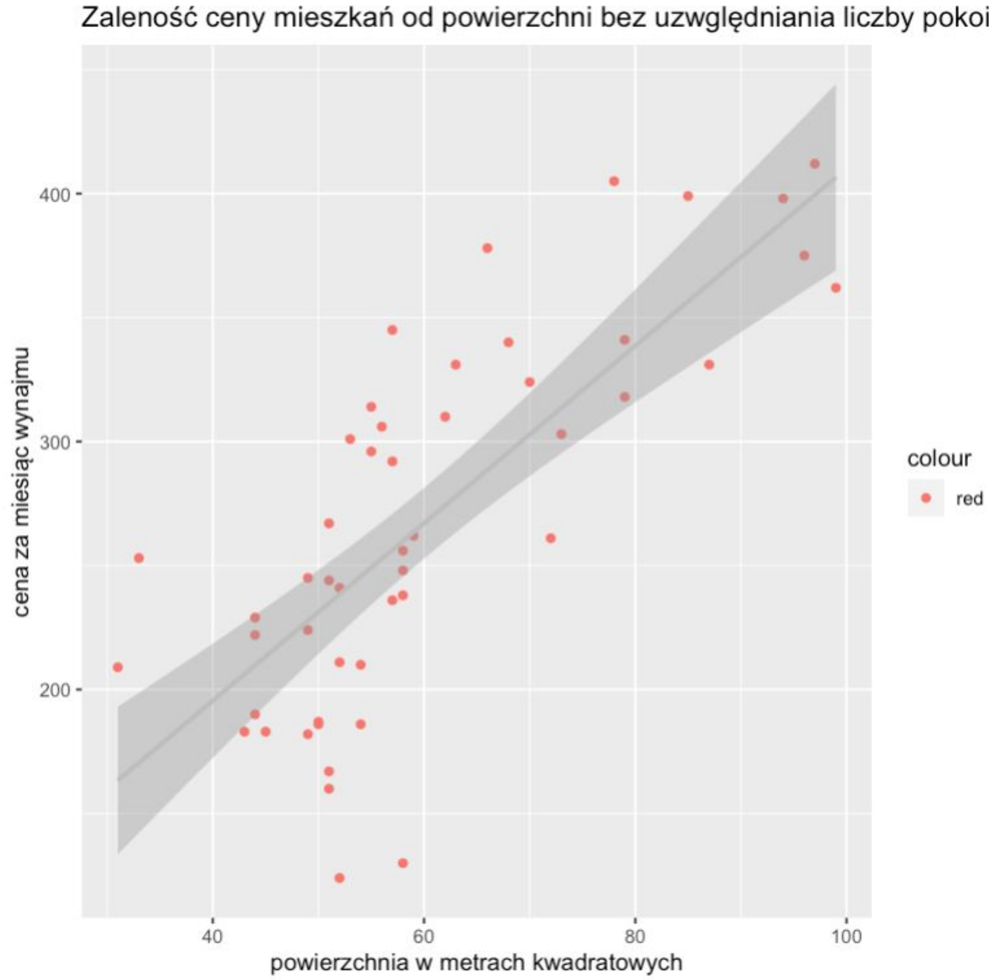
  p = 0,03014, p < 0,05, zatem nie mamy podstaw do odrzucenia hipotezy zerowej, że zależność liniowa jest najlepszym możliwym do wybrania modelem

5. Na wykresach prezentujących zbiór par, dodaliśmy proste regresji oraz korytarze ufności dla prostej regresji oraz ceny mieszkania. Wyniki, z podziałem na przypadki bez i z uwzględnieniem liczby pokoi, prezentują się następująco:

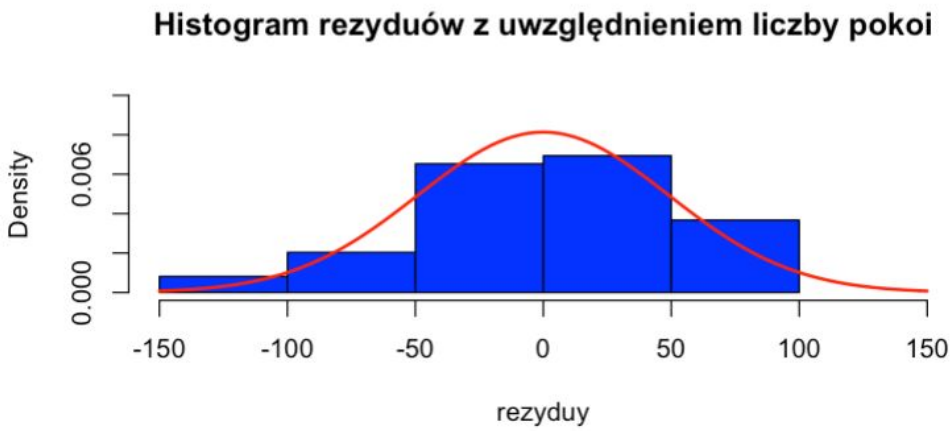
- z uwzględnieniem liczby pokoi



- bez uwzględniania liczby pokoi



6. Na podstawie otrzymanych zależności, opracowaliśmy histogramy rezyduów



7. W celu sprawdzenia, czy rezydua mają rozkład normalny, przeprowadziliśmy test zgodności Kołmogorowa–Lillieforsa, na poziomie istotności α = 0.05.

   Do przeprowadzenia testu skorzystaliśmy z funkcji lillie.test z pakietu nortest.
   - Dla rezyduów bez uwzględnienia liczby pokoi:

```
> lillie.test(residuals(reg1))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(reg1)
D = 0.057185, p-value = 0.955
```

   - Dla rezyduów z uwzględnieniem liczby pokoi:

```
> lillie.test(residuals(reg2))

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  residuals(reg2)
D = 0.057414, p-value = 0.9532
```

   Ponieważ wartości D w obu testach nie należą do obszaru krytycznego, stwierdzamy brak podstaw do odrzucenia hipotezy zerowej o normalności rozkładu.

   W celu weryfikacji poprawności funkcji lillie.test, napisaliśmy własny ciąg operacji weryfikujący obliczone wartości D:

   - Dla rezyduów bez uwzględnienia liczby pokoi:

```
> residuals1Sorted <- sort.default(residuals(reg1))
> fnx <- pnorm(residuals1Sorted, mean(residuals1Sorted), sd(residuals1Sorted))
>
> dn_plus <- max(seq(1:n)/n-fnx)
> dn_minus <- max(fnx-(seq(1:n)-1)/n)
>
> dn_prim = max(max(dn_plus), max(dn_minus))
> print(dn_prim)
[1] 0.05718495
```

   - Dla rezyduów z uwzględnieniem liczby pokoi:

```
> residuals2Sorted <- sort.default(residuals(reg2))
> fnx <- pnorm(residuals2Sorted, mean(residuals2Sorted), sd(residuals2Sorted))
>
> dn_plus <- max(seq(1:n)/n-fnx)
> dn_minus <- max(fnx-(seq(1:n)-1)/n)
>
> dn_prim = max(max(dn_plus), max(dn_minus));
> print(dn_prim)
[1] 0.05741406
```