Karol Pichurski

321663

# Podstawy Analizy Danych – Lista 2

Residuals:

Min 1Q Median 3Q Max

-588.79 -263.96 63.52 200.68 498.66

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 311.84969 170.33684 1.831 0.0801 .

dlugosc\_trasy 1.25533 0.04972 25.248 <2e-16 \*\*\*

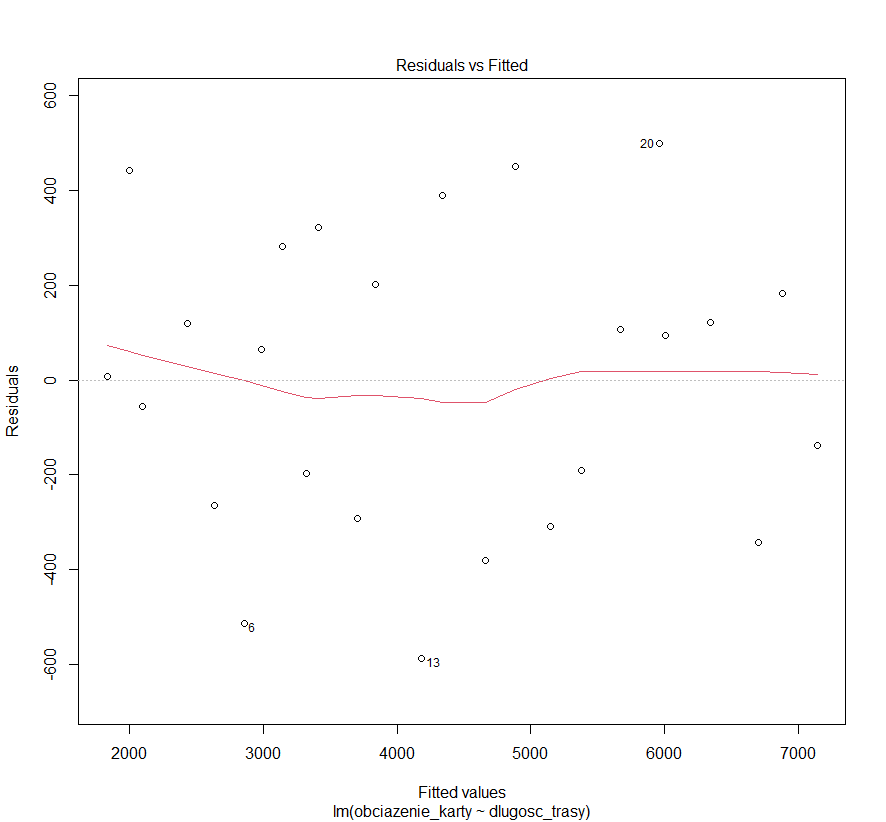
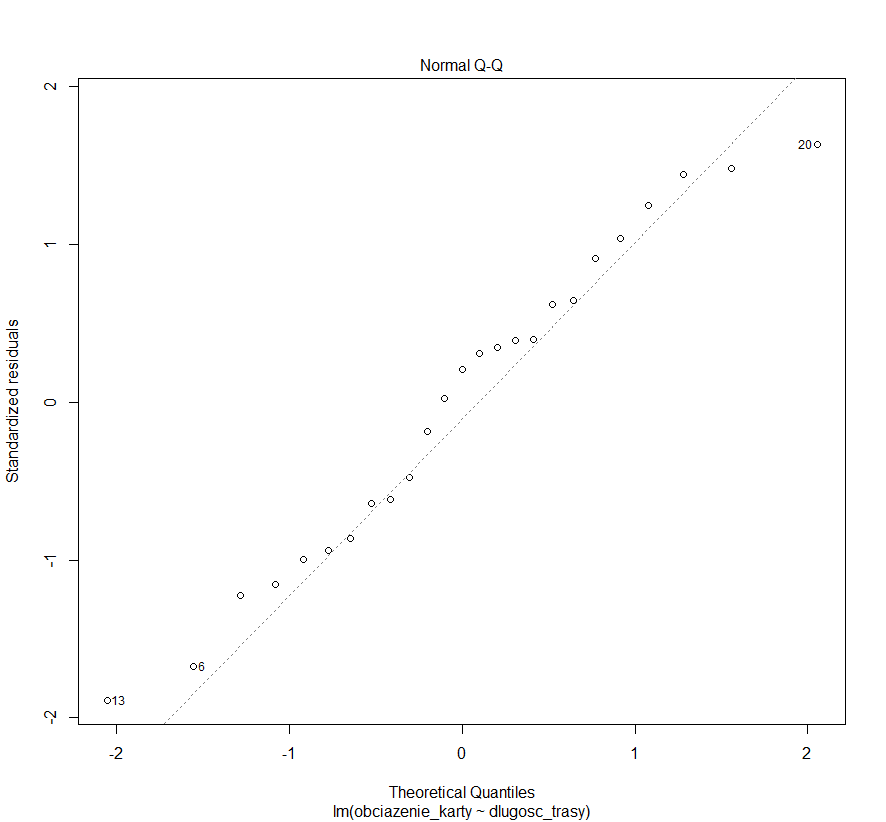
---

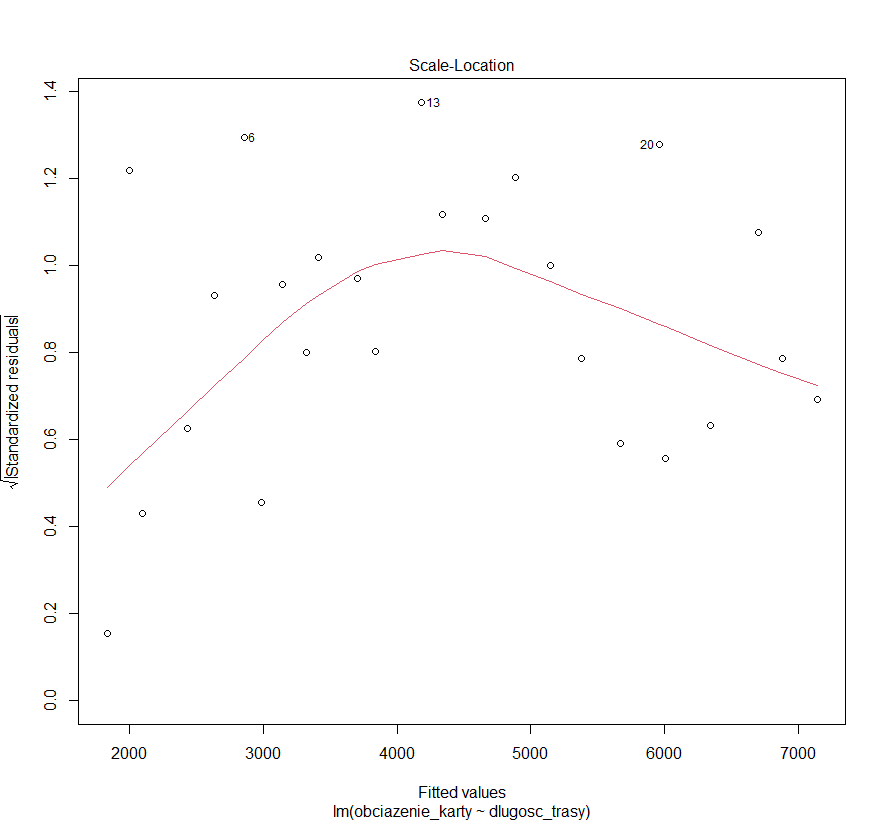
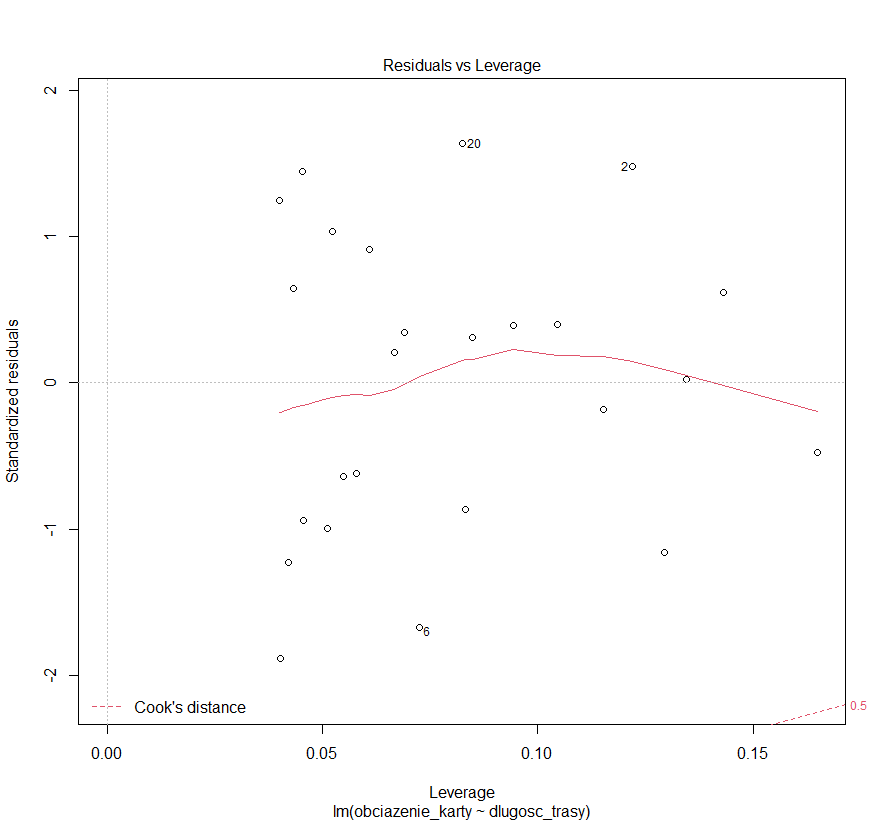
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 318.2 on 23 degrees of freedom

Multiple R-squared: 0.9652, Adjusted R-squared: 0.9637

F-statistic: 637.5 on 1 and 23 DF, p-value: < 2.2e-16





Całkowita suma kwadratów (SST): 66855898

Suma kwadratów błędów (SEE): 2328161.20112609

Regresyjna suma kwadratów (SSR): 64527736.7988739

Współczynnik determinacji R2: 0.965176427648521

Skorygowany współczynnik determinacji: 0.963662359285413

Minimalny znacząct współczynnik (MSR): 64527736.7988739

Nieobciążony estymator wariancji (MSE): 101224.400048961

Statystyka F: 637.472158567133

Obszar krytyczny wartości F: 4.27934430914465

Graniczny poziom istotności: 2.85083541898168e-18

Wartość t: 25.2482109973585

Wartość graniczna t: 2.06865761041905

P-value: 2.85083541898168e-18

Obserwacje odstające

1 2 3 4 5 6

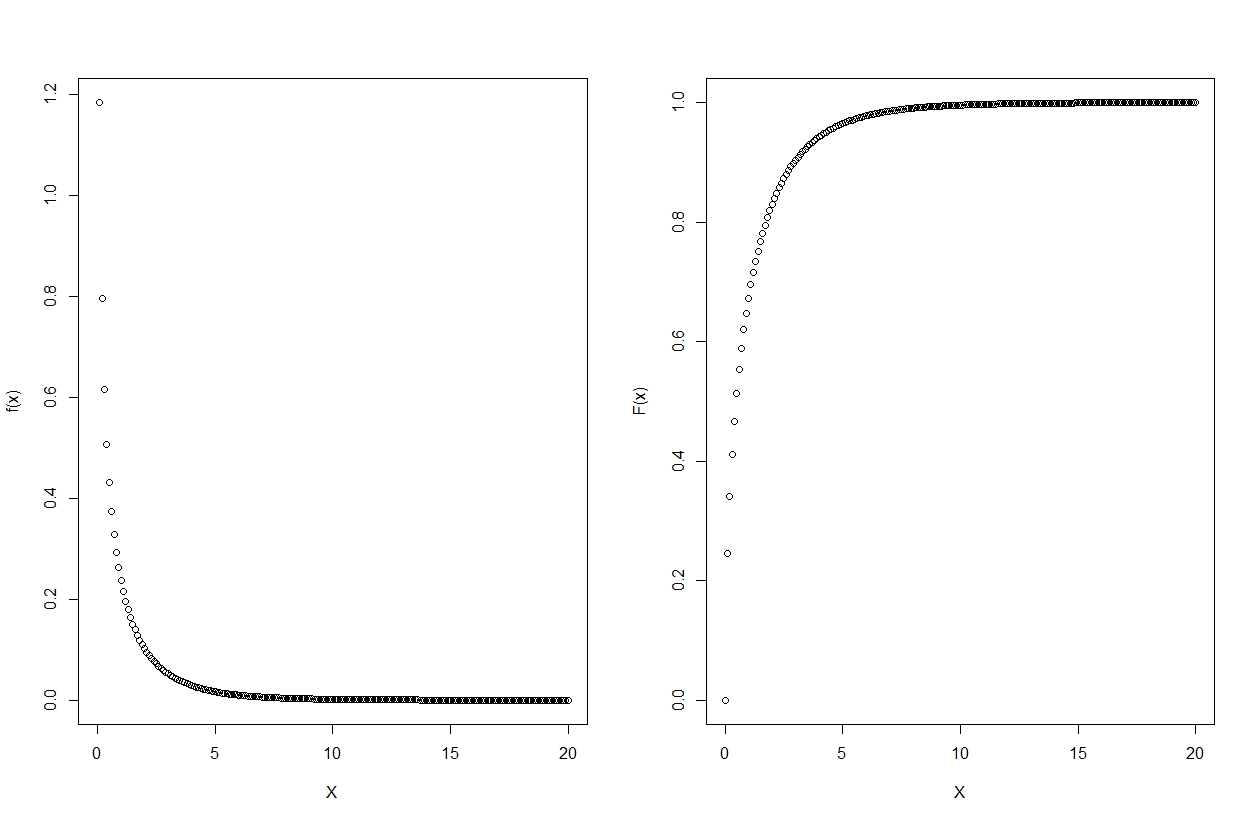
0.13448120 0.12204630 0.11529765 0.09428510 0.08312903 0.07240534

Obserwacje wpływowe

1 2 3 4 5 6

4.272197e-05 1.526064e-01 2.195828e-03 7.959113e-03 3.403324e-02 1.094555e-01

Rozkład gęstości i dystrybuanta:



Wartość t: 25.2482109973585

Wartość graniczna t: 2.06865761041905

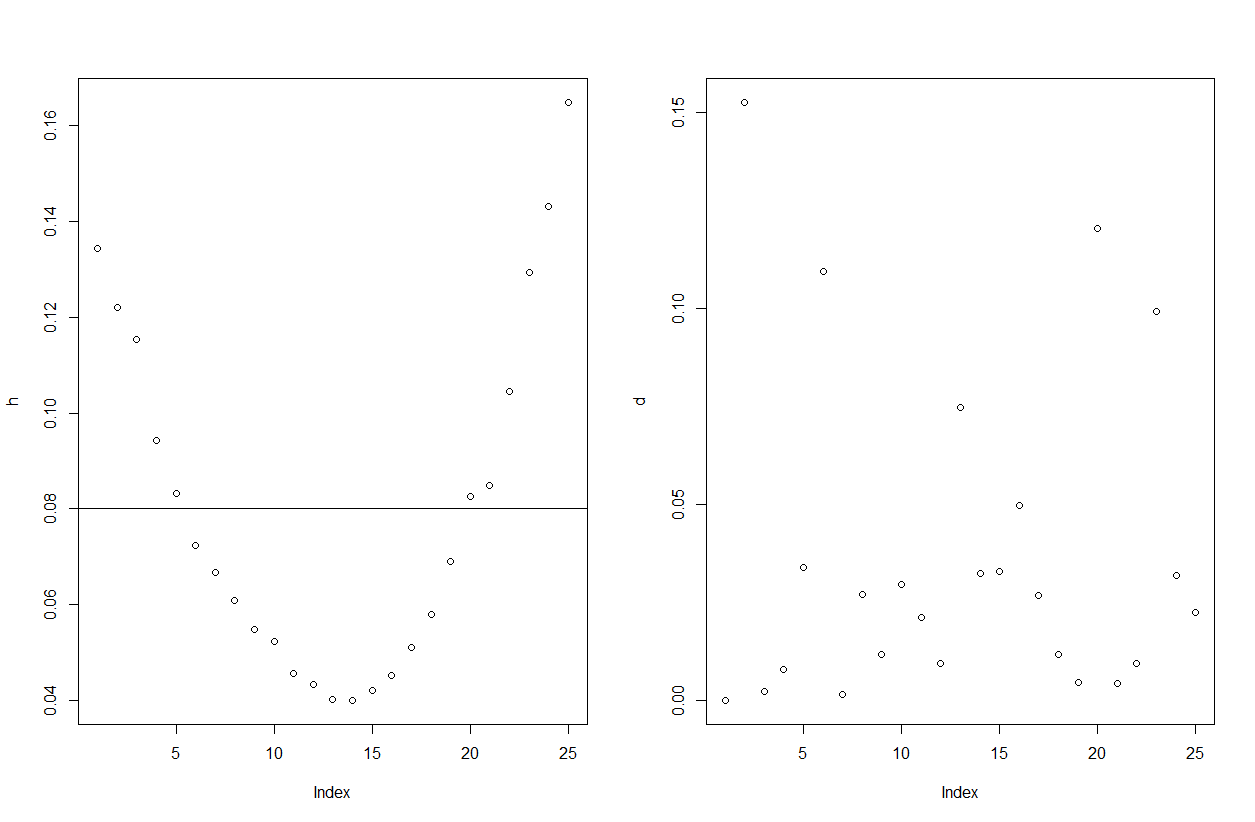
P-value: 2.85083541898162e-18

Obserwacje odstające:

0.13067904 0.09170261 0.11634307 0.08597515 0.09584245 0.09741400

Obserwacje wpływowe:

0.0076020471 0.1459189700 0.0004297142 0.0169607788 0.0232086186 0.1211448157



Ponieważ p-value (2e-16) jest mniejsze niż zakładany poziom istotności (d = 0.05), nie mamy podstaw do przyjęcia hipotezy zerowej.

Współczynnik determinacji R2 wynosi 0.9637 mamy dobre dopasowanie modelu.

Ponieważ wartość statystyki t (25.24) wpada do obszaru krytycznego (t > 2.07), hipotezę zerową należy odrzucić. Oznacza, to że nie ma podstaw do wnioskowania, że parametr jest nieistotny.

Ponieważ F (637.47) nie wpada do przedziału krytycznego (F > 4.28), zatem hipotezę zerową należy odrzucić.

Obserwacja wpływowa to taka, której usunięcie ze zbioru danych powoduje dużą zmianę wektora estymatorów regresji. Obserwacje odstające mogą być, ale nie muszą wpływowymi.

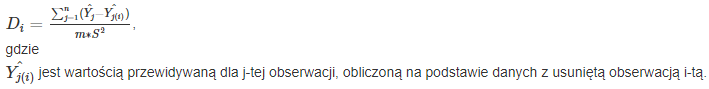
Wartości odstające przyjmuje za większe od dwukrotności ilości zmiennych podzielony przez ilość próbek:



Wartość odstająca:



Do sprawdzenia wartości wpływowych używam odległości Cooke’a, która sprawdza wpływ, jaki na prognozę znanych wartości zmiennej objaśnianej ma usunięcie ze zbioru danych i-tej obserwacji



Kod:

1. library(readxl)
3. dane = read\_excel("Zadanie\_domowe\_nr\_2\_2022\_2023\_KP.xlsx")
4. dlugosc\_trasy = dane$Długość\_trasy\_w\_milach
5. obciazenie\_karty = dane$`Obciążenie\_kart\_w USD`
6. model = lm(formula=obciazenie\_karty~dlugosc\_trasy, data=dane)
8. print(summary(model))
9. plot(model)
11. y\_mean = mean(obciazenie\_karty)
12. SST = sum((obciazenie\_karty - y\_mean) ^ 2)
13. my\_print("Całkowita suma kwadratów (SST)", SST)
15. SEE = sum(model$residuals ^ 2)
16. my\_print("Suma kwadratów błędów (SEE)", SEE)
18. SSR = sum((model$fitted.values - y\_mean) ^ 2)
19. my\_print("Regresyjna suma kwadratów (SSR)", SSR)
21. R2 = SSR / SST
22. my\_print("Współczynnik determinacji R2", R2)
24. #ilość próbek
25. n = nrow(dane)
26. p = 2
28. R2\_adj = 1 - (1 - R2) \* (n - 1) / (n - p)
29. my\_print("Skorygowany współczynnik determinacji", R2\_adj)
31. MSR = SSR / (1)
32. my\_print("Minimalny znacząct współczynnik (MSR)", MSR)
34. MSE = SEE / (n - p)
35. my\_print("Nieobciążony estymator wariancji (MSE)", MSE)
37. F = MSR / MSE
38. my\_print("Statystyka F", F)
40. par(mfrow=c(1, 2))
41. plot(seq(0, 20, 0.1), df(seq(0, 20, 0.1), p-1, n-p), xlab="X", ylab="f(x)")
42. plot(seq(0, 20, 0.1), pf(seq(0, 20, 0.1), p-1, n-p), xlab="X", ylab="F(x)")
44. obszar\_krytyczny = qf(0.95, p-1, n-p)
45. my\_print("Obszar krytyczny wartości F", obszar\_krytyczny)
47. graniczny\_poziom\_istotnosci = pf(F, p - 1, n - p, lower.tail=FALSE)
48. my\_print("Graniczny poziom istotności", graniczny\_poziom\_istotnosci)
50. alfa = model$coefficients
51. alfa\_se = **sqrt**(diag(vcov(model)))
52. t = alfa / alfa\_se
53. my\_print("Wartość t", t[2])
55. t\_gr = qt(1 - 0.05/2, df=n-p)
56. my\_print("Wartość graniczna t", t\_gr)
58. p\_value = 2\*pt(-**abs**(t), df=n-p, lower.tail=TRUE)
59. my\_print("P-value", p\_value[2])

62. h = hatvalues(model)
63. print("Obserwacje odstające")
64. print(head(h))
65. plot(h)
66. abline(2 / n, 0)
68. d = cooks.distance(model)
69. print("Obserwacje wpływowe")
70. print(head(d))
71. plot(d)

Źródła:

1. Strona: rpubs.com, autor: Marcin Mazurek, data dostępu: 16.01.2023, link: rpubs.com/mmazurek/320600