Modele liniowe raport nr 2

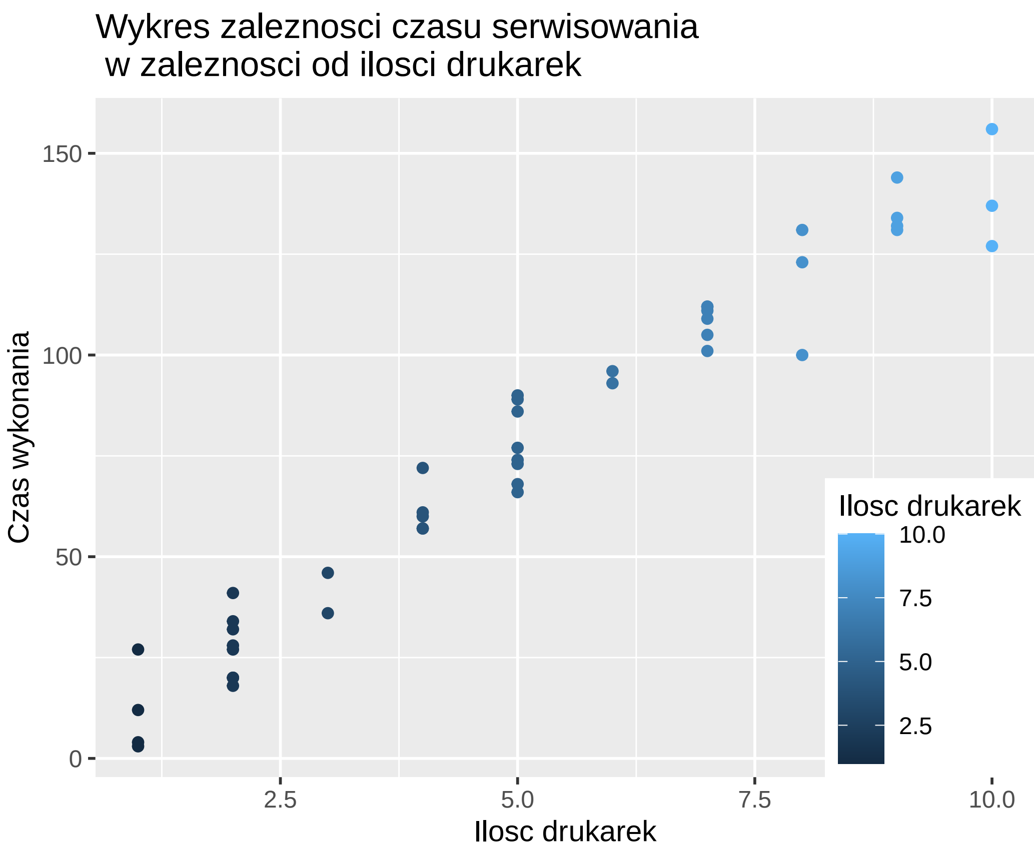
Dominik Mika

14 listopada 2020

W zadaniach 1-5 korzystamy z bazy danych **ch01pr20.txt**, która w pierwszej kolumnie zawiera informacje o czasie konserwowania, a w drugiej ilość drukarek serwisowanych w tym czasie .

1 Zadanie 1

W pierwszym zadaniu chcemy narysować nasze dane i zobaczyć, czy są one w przybliżeniu liniowe, czyli czy uzasadnione jest użycie modelu liniowego.



Widzimy, że relacja wygląda w przybliżeniu na liniową.

2 Zadanie 2

W tyma zadaniu chcemy stworzyć model liniowy, gdzie y - czas konserwacji, x - ilość drukarek. Następnie podamy wyestymowane równanie prostej regresji, podamy 95% przedział ufności dla współczynnika pochylenia oraz wykonamy test istotności współczynnika pochylenia i opiszemy odpowiednie wyniki.

2.1 a)

Tworzymy model liniowy:

$$Y_i = \hat{\beta}_1 X_i + \hat{\beta}_0$$

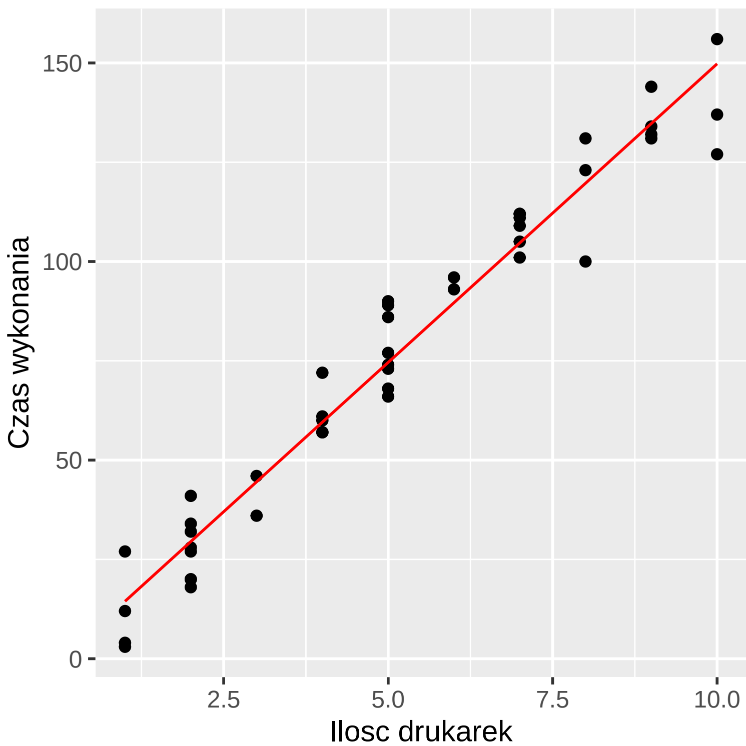
Wzory na estymatory współczynników prostej możemy otrzymać dwoma sposobami, czyli metodą najmniejszych kwadratów lub metodą największej wiarogodności, które są równoważne. Mamy wzory:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Po estymacji nasza prosta jest zadania następującym wzorem:

$$Y = 15.035X - 0.58$$

Poniżej znajduję się prosta regresji.



2.2 b)

Chcemy skonstruować 95% przedział ufności dla współczynnika kierunkowego. Końce przedziałów zadane są następującym wzorem:

$\hat{\beta}_1 \pm t_c s(\hat{\beta}_1)$, gdzie t_c to kwantyl z rozkładu studenta z parametrami $t_c(1-\frac{\alpha}{2}, n-2) = 2.017$,

$s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$, $s^2 = \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-2}$.

Po wykonaniu obliczeń otrzymujemy następujący przedział:

$$[14.061, 16.009]$$

2.3 c)

Chcąc zbadać istotność $\hat{\beta}_1$ wykonamy test statystyczn, czy $\hat{\beta}_1 = 0$ na poziomie istotności $\alpha = 0.05$.

$H_0 : \hat{\beta}_1 = 0$
$H_1 : \hat{\beta}_1 \neq 0$

Nasza statystyka testowa jest równa $T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = 31.123$

To oznacza, że odrzucamy naszą hipotezę zerową, bo $T = 31.123 \notin [14.061, 16.009]$. Dodatkowo P-wartość jest bliska 0 (2×10^{-17}), czyli widzimy że nasz Y mocno zależy od X.

3 Zadanie 3

W tym zadaniu chcemy oszacować średni czas serwisowania dla 11 maszyn i stworzyć dla niego 95% przedział ufności.

Wiemy, że $\mathbb{E}(y_h) = \mu_h = \beta_0 + \beta_1 X_h$. Wtedy estymatorem średniej jest: $\hat{\mu_h} = \hat{\beta}_0 + \hat{\beta}_1 X_h$

Wstawiamy za X_h 11, a estymatory naszych współczynników są znane, czyli:

$$\hat{\mu} = -0.58 + 15.035 \cdot 11 = 164.808$$

Teraz chcemy stworzyć przedział ufności dla średniej. Ma on następującą postać:

$$\hat{\mu_h} \pm t_c s(\hat{\mu_h})$$

,

gdzie $s^2(\hat{\mu_h}) = s^2(\frac{1}{n} + \frac{(X_h-\overline{X})^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2})$ oraz t_c to odpowiedni kwantyl z rozkładu studenta. Podstawiając dostajemy przedział:

$$[158.475, 171.14]$$

4 Zadanie 4

Tym razem musimy wyznaczyć predykcję czasu serwisowania 11 maszyn oraz skonstruować 95% przedział predykcyjny dla czasu.

Predykcja tego czasu jest równa średniej z poprzedniego zadania, czyli 164.808. Sytuacja jest bardzo podobna, ponieważ przedział ma postać:

$$\hat{\mu_h} \pm t_c s(pred)$$

,

Widzimy, że tym razem wariancja jest inna. Zadaje się ona wzorem: $s^2(\hat{\mu_h}) = s^2(1 + \frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})$

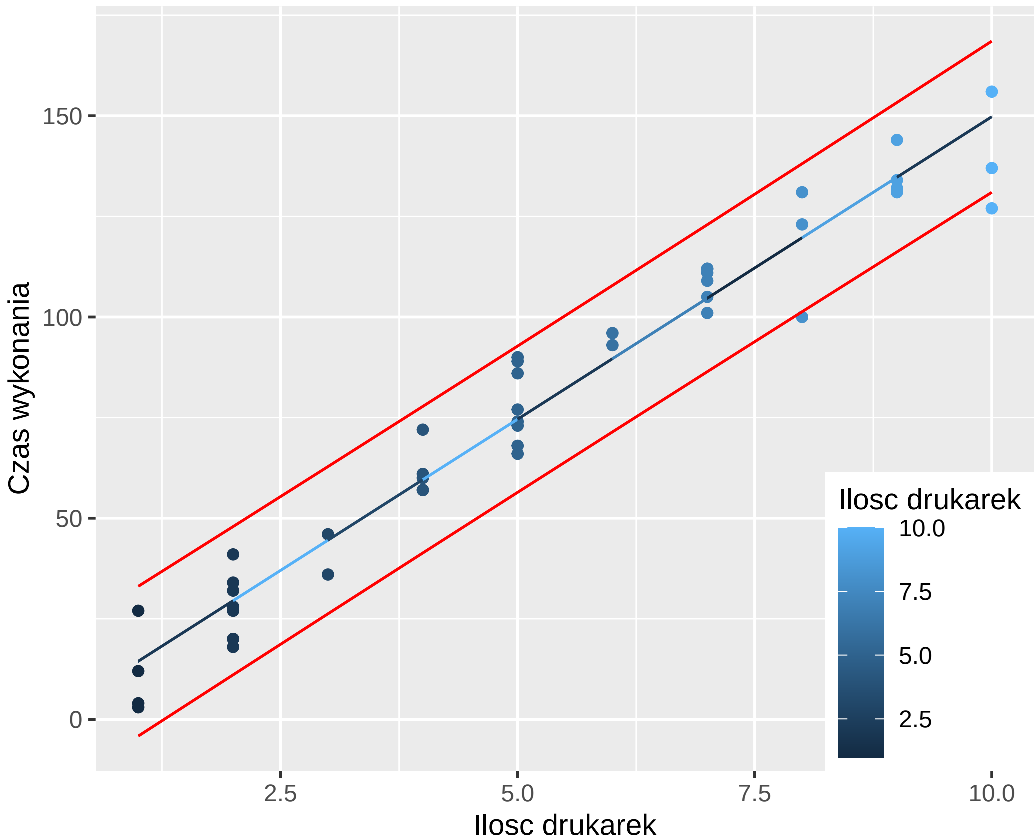
Dostajemy , więc następujący przedział

$$[145.749, 183.866]$$

Widzimy, że ma on większą długość. Wynika to właśnie z różnicy w naszych odchyleniach standartowych, ponieważ $s(pred) > s(\hat{\mu_h})$.

5 Zadanie 5

Mamy narysować nasze dane z 95% przedziałami predykcyjnymi dla każdej obserwacji.



Nasze czerwone proste to połączone końce przedziałów predykcyjnych dla każdej obserwacji. Przedziały te konstruowaliśmy tak samo jak w poprzednim zadaniu. Wiemy, że 95% naszych obserwacji powinno znajdować się w tym paśmie predykcyjnym i widzimy, że tak się dzieje.

6 Zadanie 6

Mamy dane: $n = 40$, $\sigma^2 = 120$, $SSX = \sum(X_i - \overline{X})^2 = 1000$.
Chcemy obliczyć moc testu, gdzie hipoteza zerowa to założenie, że $\beta_1 = 0$, dla poziomu istotności $\alpha = 0.05$, gdy $\beta_1 = 1$. Następnie chcemy narysować funkcję mocy testu dla β_1 od -2 do 2.

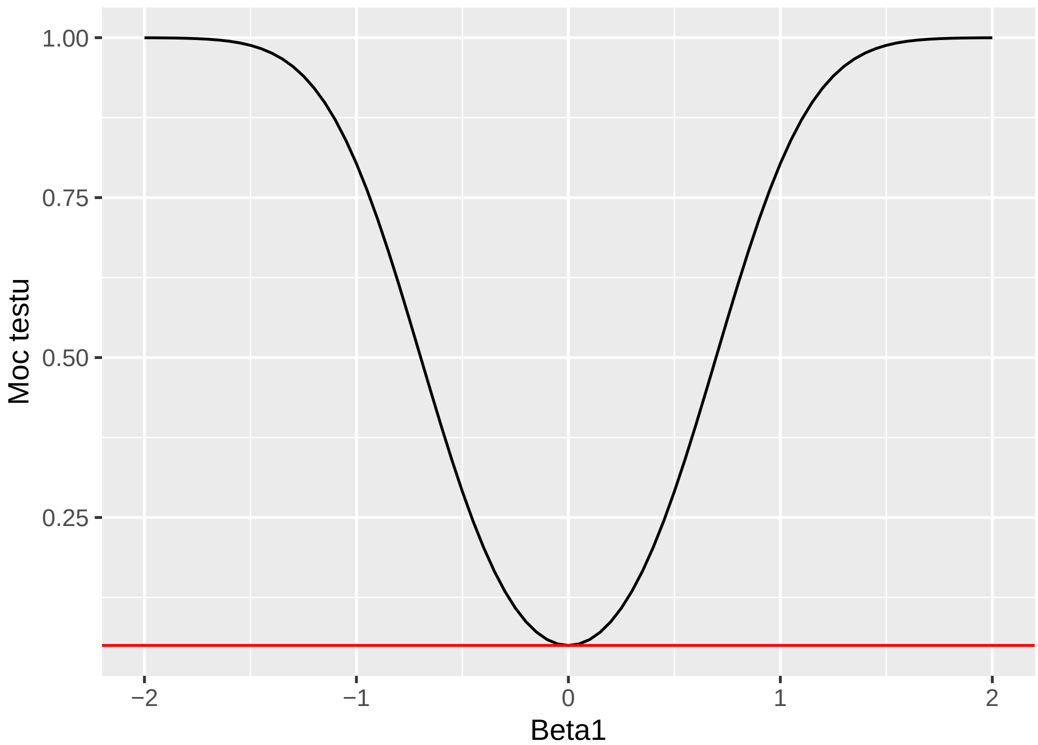
6.1 a)

Do wyznaczenia mocy testu potrzebujemy wyznaczyć parametr niecentralności δ, ponieważ nasza statystyka testowa $T \sim t(n-2, \delta)$.
Musimy, więc jeszcze policzyć $\sigma^2(\hat{\beta_1}) = \frac{\sigma^2}{\sum_{i-1}^{n}(X_i - \overline{X})^2} = 0.12$ Zakładaliśmy, że $\beta_1 = 1$, więc $\delta = \frac{\beta_1}{\sigma^2(\hat{\beta_1})} = 2.887$. Stąd statystyka $T \sim t(23, 2.887)$ oraz $t_c = t^*(0.975, 38)$
Możemy teraz wyznaczyć moc testu:

$$\pi(1) = P_{\beta_1=1}(|T| > t_c) = 1 + F_{\beta_1=1}(t_c) + F_{\beta_1=1}(-t_c) = 0.803$$

6.2 b)

Poniszy wykres otrzymaliśmy licząc moc testu dla $-2 < \beta_1 < 2$(co 0.05) i łącząc otrzymane punkty.



Widzimy że im β_1 jest bliższa 0 tym mniejsza jest moc testu. Funkcja ta przyjmuje minimum dla $\beta_1 = 0$ i wynosi 0.05, czyli jest to dokładnie prawdopodobieństwo odrzucenia hipotezy zerowej, gdy jest prawdziwa(błąd I-go rodzaju), czyli nasz poziom istotności $\alpha = 0.05$.

7 Zadanie 7

W zadaniu mamy wygenerować wektor $X = (X_1, \ldots, X_{200})^T$ z wielowymiarowego rozkładu normalnego $N(0, \frac{1}{200}I)$. Następnie wygeneruwać 1000 wektorów Y z modelu $Y = 5 + \beta_1 X + \epsilon$, gdzie

(a) $\beta_1 = 0$, $\epsilon \sim N(0, I)$,

(b) $\beta_1 = 0$, $\epsilon_1, \ldots, \epsilon_{200}$ są niezależne o rozkładzie eksponencjalnym z $\lambda = 1$,

(c) $\beta_1 = 1.5$, $\epsilon \sim N(0, I)$,

(d) $\beta_1 = 1.5$, $\epsilon_1, \ldots, \epsilon_{200}$ są niezależne o rozkładzie eksponencjalnym z $\lambda = 1$.

Następnie po każdym losowaniu wekotora Y będziemy testować hipotezę $\beta_1 = 0$ oraz policzymy średnią ilość odrzuceń tej hipotezy. Wyniki te porównamy w podpunkcie a), b) do teoretycznej wartości błędu pierwszego rodzaju ($\alpha = 0.05$), natomiast w podpunktach c), d) porównamy tą wartość do teoretycznej mocy testu, przy założeniu że ϵ ma rozkład normalny.

Otrzymaliśmy następujące wyniki: a)0.048, b)0.038, c)0.291, d)0.333

Następnie liczymy moc testu. Stosujemy do tego metodę z zadania 6. Tym razem nasze dane prezentują się następująco:

$$n = 200, \quad \sigma^2 = 1, \quad SSX = 0.9526$$

Z obliczeń mamy:

$$\sigma^2(\hat{\beta}_1) = 1.0498 \quad \text{oraz} \quad \delta = 1.464$$

.

Czyli nasza statystyka $T \sim t(198, 1.464)$, więc moc testu wynosi:

$$\pi(1.5) = 0.3078$$

Wracając do naszych prawdopodobieństw odrzucenia hipotezy, widzimy, że w podpunktach a), b) prawdopodobieństwo to jest bliskie 0.05, czyli naszemu poziomowi istotności, a w podpunktach c), b) jest bliskie mocy naszego testu.

8 Zadanie 8

Chcemy dopasować n=20 obserwacji do modelu $Y = \beta_0 + \beta_1 X + \epsilon$. Estymatory wynoszą $\hat{\beta_0} = 1$, $\hat{\beta_1} = 3$ oraz $s = 40$.

8.1 a)

W tym podpunkcie wiemy, że $s(\hat{\beta_1}) = 1$ i chcemy wyznaczyć 95% przedział ufności dla β_1.

Końce tego przedziału zadane są wzorem $\hat{\beta_1} \pm t_c s(\hat{\beta_1})$, gdzie $t_c = t^*(0.975, 18)$=2.101. Stąd nasz przedział ma postać $[0.899, 5.101]$.

8.2 b)

Chcemy zastanowić się, czy mamy dowód statystyczny na to, że Y zależy od X. Łatwo zauważyć, że przedział ufności dla β_1 nie zawiera 0, czyli możemy powiedzieć, że na 95% $\beta_1 \neq 0$, czyli Y zależy od X.

8.3 c)

W tym zadaniu mamy podany przedział ufności dla $\mathbb{E}(Y)$, gdy X=5([13,19]) i za pomocą niego chcemy wyznaczyć przedział predykcyjny. Znamy wzór na przedział ufności dla w. oczekiwanej:

$$\hat{\mu_5} \pm t_c s(\hat{\mu_5})$$

, gdzie t_c jest jak w podpunkcie a). Wiemy, że $\hat{\mu_5}$ jest środkiem naszego przedziału, czyli $\hat{\mu_5} = 16$ oraz $s(\hat{\mu_5}) = \frac{3}{t_c}$. Wiemy, że

$$s^2(pred) = s^2(1 + \frac{1}{n} + \frac{(X_5 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = s^2 + s^2(\frac{1}{n} + \frac{(X_5 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}) = s^2 + s^2(\hat{\mu_5})$$

Stąd końce naszego przedziału predykcyjnego prezentują się następująco:

$$\hat{\mu_5} \pm t_c \sqrt{s^2 + s^2(\hat{\mu_5})}$$

Podstawiając do wzoru mamy przedział predykcyjny:

$$[7.077, 24.923]$$