

# Raport z Laboratoriów 1 i 2

Adrian Siwak, numer albumu 242084

2023-03-10

## Spis treści

<b>1</b>	<b>1</b>
<b>2</b>	<b>6</b>
<b>3</b>	<b>7</b>
<b>4</b>	<b>8</b>
<b>5</b>	<b>8</b>
<b>6</b>	<b>8</b>
<b>7</b>	<b>8</b>
<b>8</b>	<b>9</b>
<b>9</b>	<b>12</b>
<b>10</b>	<b>13</b>
a . . . . .	13
b . . . . .	13
c . . . . .	14
d . . . . .	14
<b>Zadania teoretyczne</b>	<b>16</b>

<b>1</b>	<b>16</b>
<b>2</b>	<b>18</b>
<b>3</b>	<b>19</b>

## **1**

Dla każdej ze zmiennych y i x pojawiających się w pliku wyznacz podstawowe wskaźniki numeryczne charakteryzujące próbę: średnią, wariancję, odchylenie standardowe, medianę, pierwszy i trzeci kwartyl, minimum i maksimum. Skonstruuuj też histogramy i box-ploty.

```
dane<-read.table("lab1.txt",sep="",header=TRUE)

x_mean = mean(dane$x)
y_mean = mean(dane$y)
```

wartość średniej kolumny x to :

5.020899,

natomiat dla kolumny y:

11.135093.

```
x_var = var(dane$x)
y_var = var(dane$y)
```

wartość wariancji kolumny x:

6.70506474232222,

natomiat dla kolumny y:

28.8127662531828.

```
x_sd = sd(dane$x)
y_sd = sd(dane$y)
```

wartość odchylenia standardowego dla kolumny x:

2.58941397662139,

natomiat dla kolumny y:

2.58941397662139.

```
x_median = median(dane$x)
y_median = median(dane$y)
```

Wartość mediany dla kolumny x:

5.23575,

natomiat dla kolumny y:

11.40875.

```
x_quanrtiles = quantile(dane[,1],c(0.25,.75))
y_quanrtiles = quantile(dane[,2],c(0.25,.75))
```

Wartość pierwszego kwantylu dla kolumny x:

3.016625.

Wartość trzeciego kwantylu to:

6.866625.

Dla kolumny y - pierwszy kwantyl:

7.329675,

a trzeci kwantyl wynosi: 14.966675.

```
x_min = min(dane[,1])
y_min = min(dane[,2])
```

Wartość minimalna kolumny x:

0.1052,

natomiat kolumny y to:

0.2316.

```
x_max = max(dane[,1])
y_max = max(dane[,2])
```

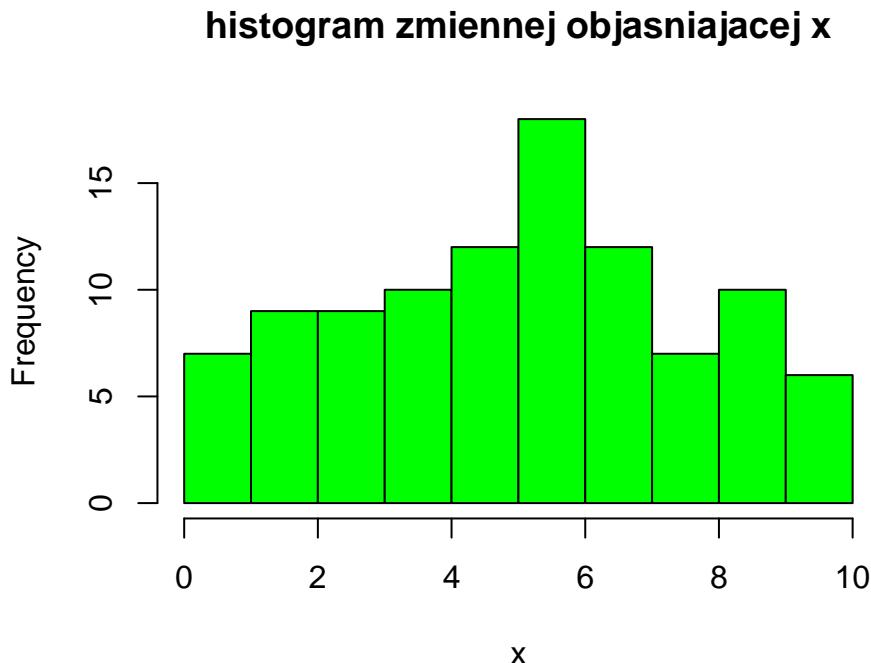
Wartość maksymalna kolumny x:

9.7658,

natomiat kolumny y to:

21.0094.

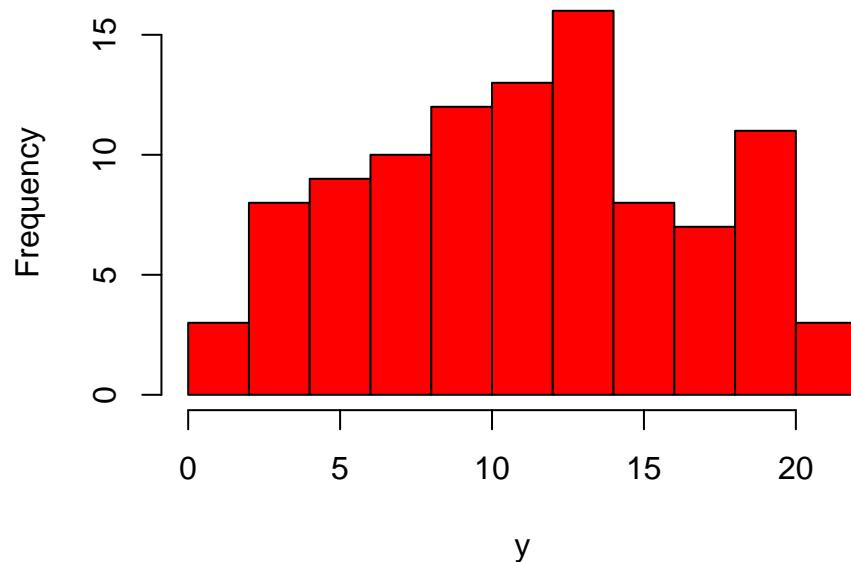
```
hist(dane$x,main="histogram zmiennej objaśniającej x",col = "green",xlab='x')
```



Rysunek 1: histogram x ZADANIE 1

```
hist(dane$y,main="histogram zmiennej objasnianej y",col = "red",xlab = 'y')
```

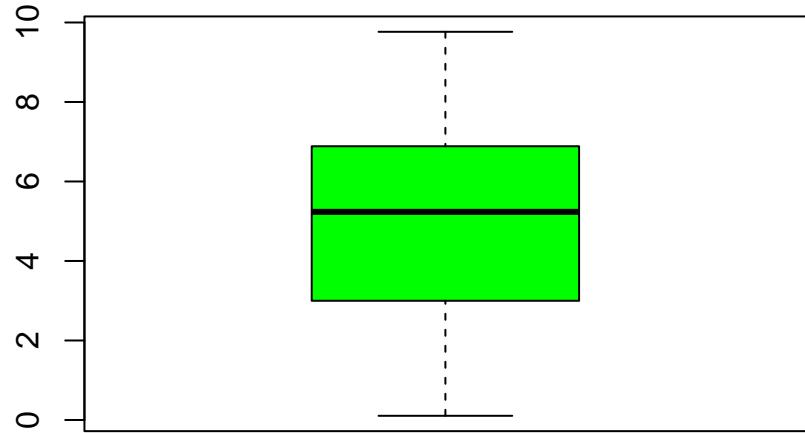
### histogram zmiennej objasianej y



Rysunek 2: histogram y ZADANIE 1

```
boxplot(dane$x, main="box-plot zmiennej objaśniającej x", col = "green")
```

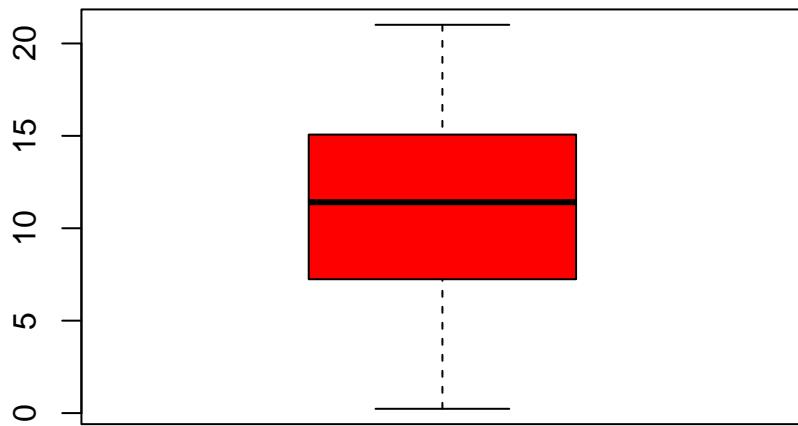
### box-plot zmiennej objasniajacej x



Rysunek 3: boxplot x ZADANIE 1

```
boxplot(dane$y, main="box-plot zmiennej objasnianej y", col = "red")
```

### box-plot zmiennej objasianej y

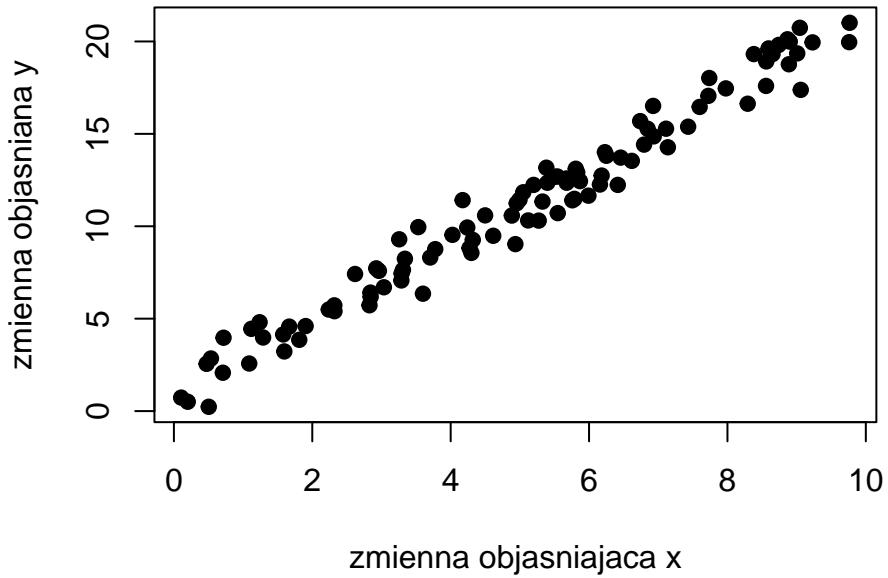


Rysunek 4: boxplot y ZADANIE 1

## 2

Wykonaj wykres rozproszenia zmiennych  $y$  i  $x$  i oblicz współczynnik korelacji próbkoowej tych zmiennych. Czy chmura punktów na tym wykresie ma (w przybliżeniu) charakter liniowy? Uzasadnij dlaczego można wykorzystać model regresji liniowej  $y = \beta_0 + \beta_1 \cdot x + \varepsilon$  do opisu zależności między zmiennymi  $y$  i  $x$ ?

```
plot(dane$x,dane$y,xlab="zmienna objaśniająca x", ylab="zmienna objaśniana y", pch=19)
```



Rysunek 5: wykres rozproszenia ZADANIE 2

```
corr=cor(dane$x,dane$y)
```

Jak wizadć z wykresu 5 chmura punktów ma charakter liniowy, natomiast współczynnik korelacji próbkoowej między  $x$  a  $y$  wynosi 0.985314282880089 więc możemy wnioskować o silnej liniowej korelacji pomiędzy tymi zmiennymi. Uzasadnione jest więc użycie modelu regresji liniowej.

### 3

Wyznacz wartości estymatorów najmniejszych kwadratów  $\hat{\beta}_0$  i  $\hat{\beta}_1$  parametrów  $\beta_0$  i  $\beta_1$ .

```
x<-dane$x
y<-dane$y
model<-lm(y~x)
b_0=model$coefficients[1]
b_1=model$coefficients[2]
```

Wartości estymatorów to:

$$\hat{\beta}_0 = 0.879819346096292,$$

$$\hat{\beta}_1 = 2.04251741648332.$$

## 4

Znajdź wartość estymatora  $\hat{\sigma}^2 := \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}$  parametru  $\sigma^2$ .

```
sigma_squared_hat=(summary(model)$sigma)**2
```

Wartość estymatora  $\hat{\sigma}^2$  wynosi 0.848630234264594.

## 5

Na poziomie istotności  $\alpha = 0,05$  zweryfikuj hipotezę  $H_0 : \beta_1 = 0$  przy hipotezie alternatywnej  $H_1 : \beta_1 \neq 0$ . Jaka jest p-wartość dla tego testu? Czy na podstawie tych wyników można stwierdzić, że rozpatrywany w tym przykładzie model regresji liniowej ma sens?

```
T_stat=summary(model)$coefficients[2,3]
t=qt(0.975,98)
p_value=summary(model)$coefficients[2,4]
```

Wartość statystyki  $T$  wynosi 57.1249198016699. Hipoteza zerowa może być odrzucona na poziomie istotności  $\alpha = 0,05$  gdy wartość bezwzględna statystyki  $T$  jest większa niż  $t_{\alpha/2,n-2}$  - kwantyl rzędu  $1 - \alpha/2$  rozkładu t-Studenta z  $n-2$  stopniami swobody. W tym przypadku:  $n = 100$ ,  $n - 2 = 98$ ,  $1 - \alpha/2 = 0,975$ ,  $t_{\alpha/2,n-2} = 1.98446745450848$ .  $57.1249198016699 > 1.98446745450848$  więc hipotezę zerową możemy odrzucić na poziomie istotności  $\alpha = 0,05$ . Ten test ma p-wartość równą 4.81959883049959e-77.

## 6

Skonstruuj przedział ufności dla  $\beta_1$  na poziomie ufności 0.99.

```
c=confint(model,level=0.99)
```

Przedział ufności na poziomie ufności 0.99 wynosi [1.94859076356876, 2.13644406939789]

## 7

Dla  $x_0 = 1$  oblicz prognozowaną przez model wartość  $\hat{Y}(x_0)$ . Następnie wyznacz przedziały ufności na poziomie ufności 0.99 dla  $Y(x_0)$ .

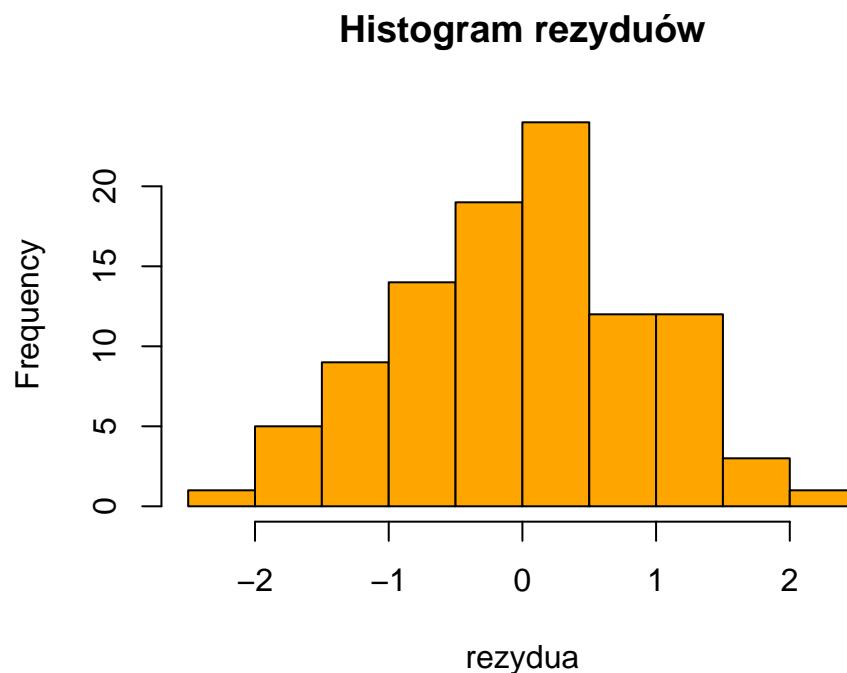
```
x_0<-data.frame(x=1)
pred=predict(model,newdata=x_0,interval=c("confidence"),level=0.99)
```

Wartość  $\hat{Y}(x_0)$  jest równa 2.92233676257962. Przedział ufności na poziomie ufności 0.99 dla  $Y(x_0)$  wynosi [2.47378764568139, 3.37088587947784].

## 8

Narysuj histogram i wykres kwantylowy dla rezyduów. Narysuj też wykresy rozproszenia dla prób  $(\hat{y}_1, e_1), \dots, (\hat{y}_n, e_n)$  i  $(x_1, e_1), \dots, (x_n, e_n)$ . Czy na podstawie tych rysunków można stwierdzić, że któreś założenie modelu regresji nie jest spełnione (patrz zadanie teoretyczne 3.)?

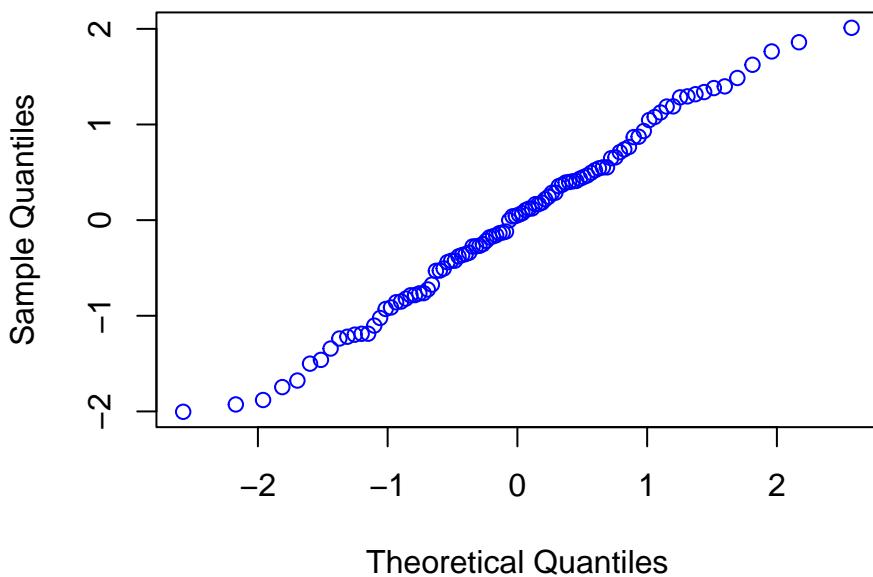
```
hist(model$residuals,main="Histogram rezyduów",xlab="rezydua",col = 'orange')
```



Rysunek 6: histogram rezyduów ZADANIE 8

```
qqnorm(model$residuals,main="Wykres kwantylowy rezyduów",col = 'blue')
```

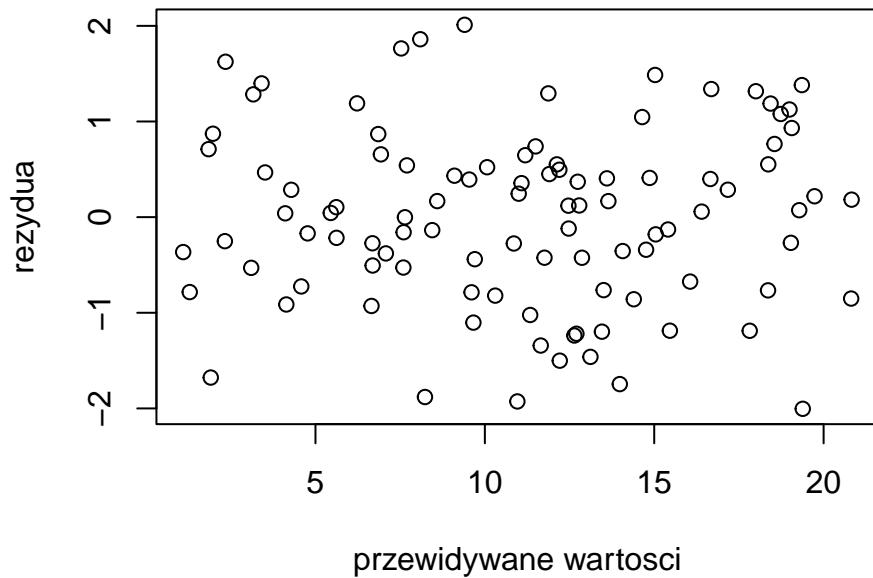
## Wykres kwantylowy rezyduów



Rysunek 7: Wykres kwantylowy rezyduów ZADANIE 8

```
y_hat=predict(model,newdata=list(dane$x))
plot(y_hat,model$residuals,xlab="przewidywane wartości",ylab = "rezydua",main="Wykres ro
```

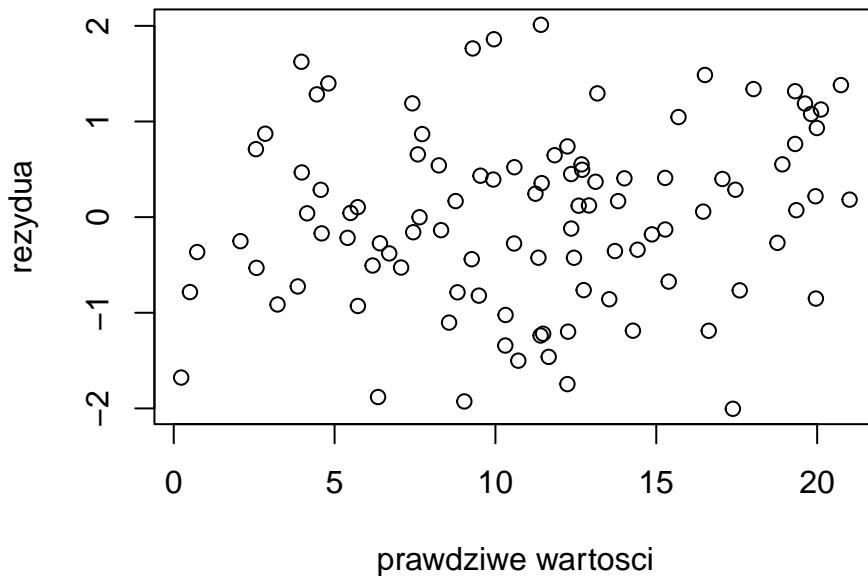
## Wykres rozproszenia 1



Rysunek 8: Wykres rozproszenia próba 1 ZADANIE 8

```
plot(y,model$residuals,xlab="prawdziwe wartości",ylab = "rezydua",main="Wykres rozprosze
```

## Wykres rozproszenia 2



Rysunek 9: Wykres rozproszenia próba 2 ZADANIE 8

Na podstawie wykresów można stwierdzić że założenia modelu są spełnione.

## 9

Zmodyfikuj dane z pliku lab1.txt, przyjmując, że ostatnią obserwacją jest 6, 9347 i 148, 6400, a nie 6, 9347 i 14, 8640. Wyznacz estymatory najmniejszych kwadratów parametrów  $\beta_0$  i  $\beta_1$  i porównaj je z estymatorami otrzymanymi poprzednio. Co zauważasz? Czy obserwacja (6, 9347; 148, 640) jest a) odstająca, b) wpływowa?

```
dane2=dane
dane2[100,1]=6.9347
dane2[100,2]=148.6400

x2<-dane2$x
y2<-dane2$y
model<-lm(y2~x2)
b_0_2=model$coefficients[1]
b_1_2=model$coefficients[2]
```

Poprzednio otrzymane estymatory miały wartości:

$$\hat{\beta}_0 = 0.879819346096292,$$

$$\hat{\beta}_0 = 2.04251741648332.$$

Po zmianie ostatniej obserwacji na (6, 9347; 148, 640):

$$\hat{\beta}_0^{nowe} = 0.281075792605789,$$

$$\hat{\beta}_0^{nowe} = 2.42820602593165.$$

Więc obserwacja (6, 9347; 148, 640) jest odstająca i wpływową.

## 10

Niech  $n = 100$ ,  $x_1, x_2, \dots, x_n$  iid  $U(0, 1)$ ,  $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_n$  iid  $N(0, \sigma^2)$ ,  $\sigma = 0.1$ ,  $\beta_0 = 1$ ,  $\beta_1 = 2$ .

### a

Wygeneruj obserwacje  $y_1, y_2, \dots, y_n$ , takie że  $y_i = \beta_0 + \beta_1 \cdot x_i + \mathcal{E}_i$ ,  $i = 1, \dots, n$ .

```
set.seed(1) #w celu reprodukowalności wyników ustawiam ziarno generatora
b0=rep(1,100)
set.seed(1)
x_new<-runif(100)
set.seed(1)
errors<-rnorm(100,0,0.1)

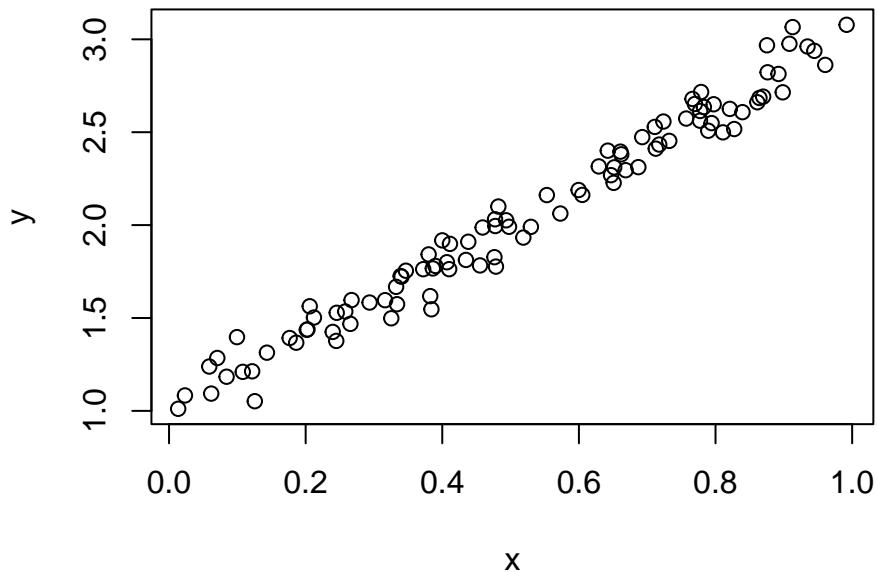
y_new<-b0+2*x_new+errors
```

### b

Wykonaj wykres rozproszenia dla próby  $(x_1, y_1), \dots, (x_n, y_n)$ . Czy chmura punktów ma (w przybliżeniu) charakter liniowy?

```
plot(x_new,y_new,main="wykres rozproszenia wygenerowanej próby",xlab="x",ylab="y")
```

### wykres rozproszenia wygenerowanej próby



Rysunek 10: Wykres rozproszenia ZADANIE 10b

Chmura punktów na wykresie 10 ma w przybliżeniu charakter liniowy.

**c**

W oparciu o próbę  $(x_1, y_1), \dots, (x_n, y_n)$  wyznacz estymatory najmniejszych kwadratów  $(\hat{\beta}_0, \hat{\beta}_1)$  parametrów  $(\beta_0, \beta_1)$ . Porównaj wartości tych estymatorów z wartościami  $(\beta_0, \beta_1)$ .

```
model_new<-lm(y_new~x_new)
b_0_new=model_new$coefficients[1]
b_1_new=model_new$coefficients[2]
```

Otrzymane estymatory mają wartości:

$$\hat{\beta}_0 = 0.994990704506478,$$

$$\hat{\beta}_1 = 2.03070024582235.$$

Są one bardzo zbliżone do wartości parametrów  $(\beta_0, \beta_1) = (1, 2)$

**d**

Powtórz obliczenia dla  $\sigma = 0.5$  i  $\sigma = 1$ . Jak zmieniają się precyzja estymatorów i współczynnik  $R^2$ , gdy rośnie  $\sigma^2$

```

set.seed(1)
errors1<-rnorm(100,0,0.5)
set.seed(1)
errors2<-rnorm(100,0,1)

y_new1<-b0+2*x_new+errors1
y_new2<-b0+2*x_new+errors2

model_new1<-lm(y_new1~x_new)
b_0_new1=model_new1$coefficients[1]
b_1_new1=model_new1$coefficients[2]

model_new2<-lm(y_new2~x_new)
b_0_new2=model_new2$coefficients[1]
b_1_new2=model_new2$coefficients[2]

summary1=summary(model_new1)
summary2=summary(model_new2)

R_2_1=summary1$r.squared
R_2_2=summary2$r.squared

```

Dla  $\sigma = 0.5$ :

Otrzymane estymatory mają wartości:

$$\hat{\beta}_0 = 0.974953522532386,$$

$$\hat{\beta}_1 = 2.15350122911177.$$

Współczynnik  $R^2$  ma wartość:

$$R^2 = 0.624096638126024.$$

Dla  $\sigma = 1$ :

Otrzymane estymatory mają wartości:

$$\hat{\beta}_0 = 0.949907045064774,$$

$$\hat{\beta}_1 = 2.30700245822354.$$

Współczynnik  $R^2$  ma wartość:

$$R^2 = 0.322651496740211.$$

Precyzja estymatorów spada wraz ze wzrostem wartości  $\sigma^2$ , a wartość współczynnika  $R^2$  maleje.

# Zadania teoretyczne

## 1

Rozważmy model regresji liniowej z jedną zmienną objaśniającą  $y = \beta_0 + \beta_1 \cdot x + \varepsilon$ , tworzony na podstawie próby  $(x_1, y_1), \dots, (x_n, y_n)$ . ## 1 Wyprowadź wzory na estymatory najmniejszych kwadratów  $\hat{\beta}_0$  i  $\hat{\beta}_1$  parametrów  $\beta_0$  i  $\beta_1$ , wiedząc że  $\hat{\beta}_0$  i  $\hat{\beta}_1$  są rozwiązaniami następującego problemu minimalizacyjnego:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1)} [\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2]$$

The image shows handwritten mathematical steps on grid paper. At the top left, it says "ZAD 1". Below it, the function  $S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - \beta_0 - \beta_1 x_i]^2$  is written. Then, the estimators  $(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} S(\beta_0, \beta_1)$  are shown. The next step shows the partial derivative of  $S$  with respect to  $\beta_0$ :  $\frac{\partial S}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^n [y_i - \beta_0 - \beta_1 x_i]^2 = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i) \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_i) = \sum_{i=1}^n -2(y_i - \beta_0 - \beta_1 x_i)$ . This is followed by the equation  $\frac{\partial S}{\partial \beta_0} = 0$ . Below that, the equation  $-2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$  is written. Then, the equations  $\sum_{i=1}^n y_i - \sum_{i=1}^n \beta_0 - \beta_1 \sum_{i=1}^n x_i = 0$  and  $\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i = n\beta_0$  are shown. Finally, the estimator  $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$  is derived.

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

—

$$\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$\frac{\partial S}{\partial \beta_1} = 0$$

$$0 = \sum_{i=1}^n -2(y_i x_i - \beta_0 x_i - \beta_1 x_i^2) / 2$$

$$0 = -\sum_{i=1}^n y_i x_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n x_i y_i = -\beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n x_i y_i = -(\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2$$

$$\sum_{i=1}^n x_i y_i = -\bar{y} \sum_{i=1}^n x_i - \beta_1 \bar{x} \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2$$

$$\begin{aligned}
 \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i &= \beta_1 \left( \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) \\
 \left\{ \begin{aligned} \sum_{i=1}^n x_i &= n \bar{x} \\ \sum_{i=1}^n x_i &= n \bar{x} \end{aligned} \right. & \\
 \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y} &= \beta_1 \left( \sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) \\
 * & \quad ** \\
 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 \\
 \beta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
 * * \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - 2 \bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \stackrel{*}{=} \sum_{i=1}^n x_i^2 - 2 \bar{x}^2 n + n \bar{x}^2 = \sum_{i=1}^n x_i^2 - \bar{x}^2 n \\
 * \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) = \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \bar{y} - \sum_{i=1}^n \bar{x} y_i + \sum_{i=1}^n \bar{x} \bar{y} \stackrel{*}{=} \\
 &= \sum_{i=1}^n x_i y_i - n \bar{y} \bar{x} - n \bar{y} \bar{x} + n \bar{x} \bar{y} = \sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}
 \end{aligned}$$

## 2

Uzasadnij, przeprowadzając odpowiednie rachunki, że  $\hat{\beta}_1$  można zapisać w równoważnej postaci  $\hat{\beta}_1 = r \frac{S_y}{S_x}$  gdzie  $r$  jest **współczynnikiem korelacji próbowej Pearsona**, a  $S_y^2$  i  $S_x^2$  oznaczają wariancje próbowe w próbach y-ów i x-ów.

2AD 2

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

$$S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$$

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$L = \hat{\beta}_1 = r \frac{S_y}{S_x} = r$$

$$P = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2 (n-1)}{\sum_{i=1}^n (y_i - \bar{y})^2 (n-1) (\sum_{i=1}^n (x_i - \bar{x})^2)^2}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = L$$

### 3

Niech  $e_1, \dots, e_n$  oznaczają kolejne rezydua, tzn. niech  $e_i := y_i - \hat{y}_i$ ,  $i = 1, \dots, n$ . Uzasadnij, przeprowadzając odpowiednie rachunki (!), następujący fakt: Jeśli model regresji liniowej poprawnie opisuje zależność między zmiennymi  $y$  i  $x$ , to

- współczynnik korelacji próbowej dla próby  $(\hat{y}_1, e_1), \dots, (\hat{y}_n, e_n)$  jest równy zero;
- współczynnik korelacji próbowej dla próby  $(x_1, e_1), \dots, (x_n, e_n)$  jest równy zero;

Jak powinny wyglądać wykresy rozproszenia dla tych prób?

ZAD 3

a)

$$r = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(e_i - \bar{e})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^n (e_i - \bar{e})^2}}$$

$$\begin{aligned} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})(e_i - \bar{e}) &\stackrel{\bar{e}=0}{=} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})e_i = \sum_{i=1}^n \hat{y}_i e_i - \bar{\hat{y}} \sum_{i=1}^n e_i \stackrel{\sum e_i = 0}{=} \sum_{i=1}^n \hat{y}_i e_i = \\ &= \hat{Y}^T e = (H Y)^T e = Y^T H e = Y^T H (I - H) Y = Y^T (H - H^2) Y \stackrel{H^2 = H}{=} Y^T (H - H) Y = 0. \end{aligned}$$

b)

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(e_i - \bar{e})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (e_i - \bar{e})^2}}$$

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})(e_i - \bar{e}) &\stackrel{\bar{e}=0}{=} \sum_{i=1}^n (x_i - \bar{x})e_i = \sum_{i=1}^n x_i e_i - \bar{x} \sum_{i=1}^n e_i \stackrel{\sum e_i = 0}{=} \sum_{i=1}^n x_i e_i = X^T e = \\ &= X^T (Y - \hat{Y}) = X^T Y - X^T \hat{Y} = X^T Y - X^T H Y = X^T Y - \underbrace{X^T X (X^T X)^{-1}}_I X^T Y = X^T Y - X^T Y = 0 \end{aligned}$$

Wykresy tych prób powinny oscylować wokół zera, bez widocznej zależności.