Modele liniowe - raport 2

Łukasz Rębisz

20.11.2022

Zadanie 1

Naszym zadaniem jest zbadanie prostej regresji liniowej danych zawartych w podanych pliku.

Teoretyczny model regresji liniowej

Zakładamy, że związek pomiędzy zmiennymi zależnymi i odpowiadającymi im wartościami zmiennych niezależnych jest postaci:

$$Y_i = f(X_i) + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, ..., n,$$

gdzie:

β_0 - wyraz wolny, deterministyczny parametr regresji,

β_1 - współczynnik kierunkowy, deterministyczny parametr regresji,

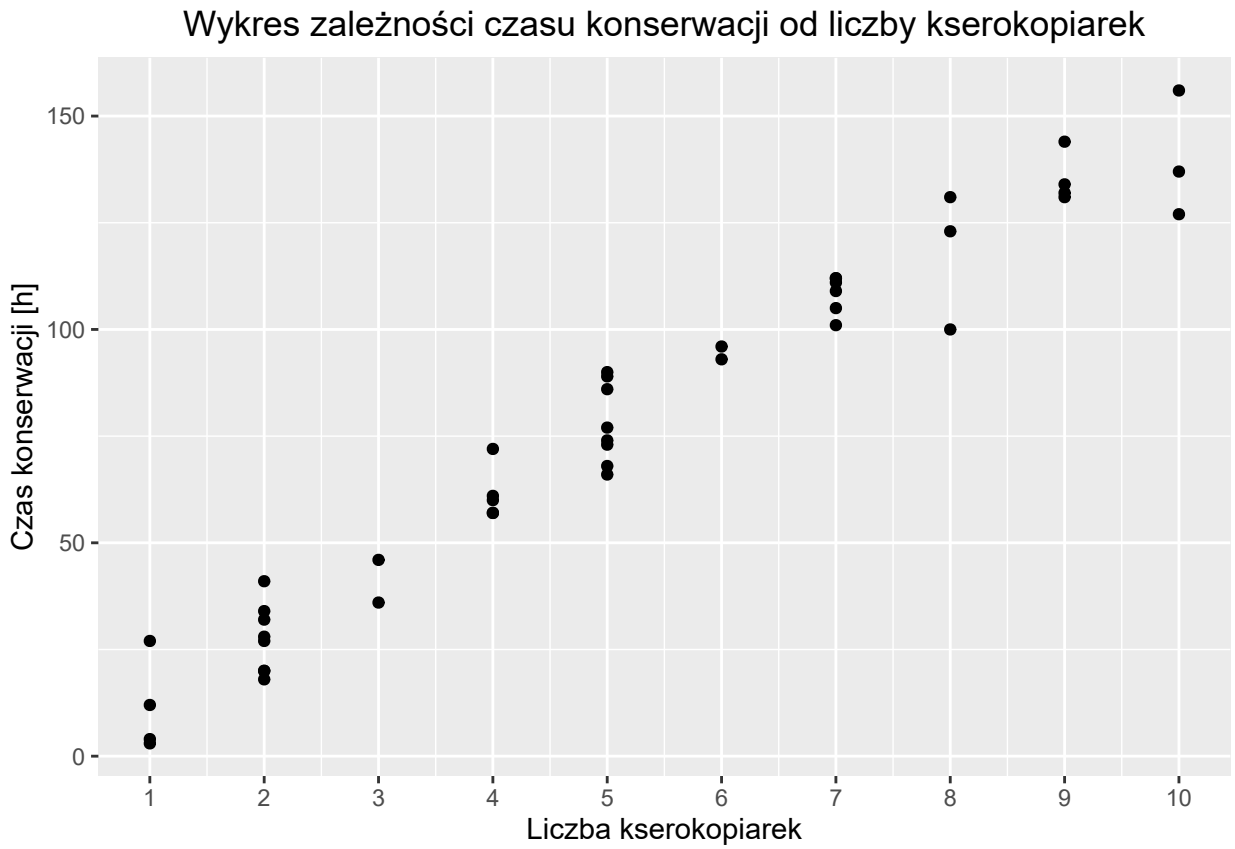
ϵ_i - błąd związany z i-tym pomiarem, zmienna losowa $N(0, \sigma^2)$. Zakładamy, że $\epsilon_1, ...\epsilon_n$ są niezależnymi zmiennymi losowymi.

W naszym przypadku badane dane opisują średni czas potrzebny do konserwacji danej liczby kserokopiarek.

Zapiszmy dane zawarte w pliku w postaci ramki danych.

```
data <- read.table(file="CH01PR20.txt",header = F, sep="")
colnames(data) <- c("Time", "Number")
```

Przedstawmy badane dane na wykresie.



Na podstawie wykresu możemy stwierdzić, że zależność pomiędzy zmienną X - liczbą kserokopiarek a zmienną Y - czasem konserwacji jest w przybliżeniu liniowa - punkty układają się w przybliżeniu wzdłuż pewnej prostej.

Zadanie 2

Zbadajmy regresję liniową pomiędzy zmiennymi X i Y.

a) równanie wyestymowanej regresji liniowej

Funkcje wbudowane

```
LinMod <- lm(Time~Number, data)
b_0 <- LinMod$coefficients[1]
b_1 <- LinMod$coefficients[2]
```

Równanie wyestymowanej regresji liniowej przy wykorzystaniu wbudowanej funkcji *lm*:

$Y = -0.58 + 15.04X$.

Własna implementacja funkcji

Wiemy, że estymatory $\hat{\beta}_0, \hat{\beta}_1$ wyznaczone za pomocą metody najmniejszych kwadratów (jak również metodą największej wiarogodności) wyrażają się następującymi wzorami:

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

Zaimplementujmy powyższe wzory:

```
beta_1 <- function(X, Y){
  return(sum((X-mean(X))*(Y-mean(Y)))/sum((X-mean(X))^2))
}

beta_0 <- function(X, Y){
  return(mean(Y) - beta_1(X, Y)*mean(X))
}

X <- data$Number
Y <- data$Time

b_1_own <- beta_1(X, Y)
b_0_own <- beta_0(X, Y)
```

Równanie wystemowanej regresji liniowej przy wykorzystaniu powyższych funkcji *beta_1* i *beta_0*:

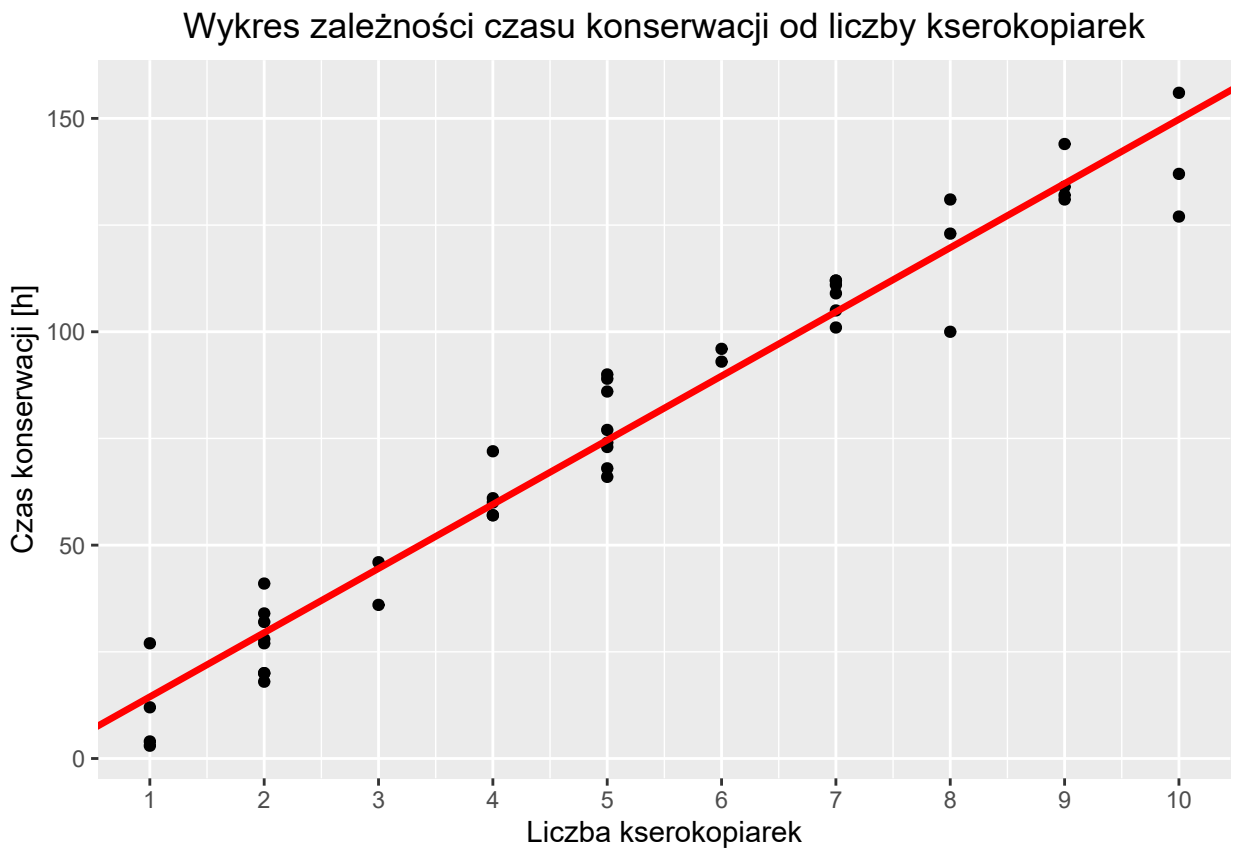
$Y = -0.58 + 15.04X$.

Różnica pomiędzy wartościami wyznaczonymi za pomocą funkcji wbudowanych a wartościami dla powyższych funkcji wynosi:

- $1.7763568 \times 10^{-15}$ dla estymatora $\hat{\beta}_1$,
- $3.1308289 \times 10^{-14}$ dla estymatora $\hat{\beta}_0$.

Otrzymujemy zatem praktycznie takie same wartości estymatorów. Przyjmijmy więc za wystemowane wartości $\hat{\beta}_0$ i $\hat{\beta}_1$ wartości obliczone przez funkcje wbudowane.

Zaznaczmy prostą $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ =-0.58 + 15.04X na wykresie.



b) 95% przedział ufności dla współczynnika kierunkowego prostej

Funkcje wbudowane

Przy pomocy funkcji wbudowanej *confint* otrzymujemy następujący 95% przedział ufności dla β_1:

```
conf_int_beta_1_R <- c(0,0)
conf_int_beta_1_R[1] <- confint(LinMod)[2,][["2.5 %"]]
conf_int_beta_1_R[2] <- confint(LinMod)[2,][["97.5 %"]]
```

Otrzymujemy następujący 95% przedziały ufności dla β_1: [14.061, 16.009].

Własna implementacja funkcji

Nieobciążony estymator parametru σ^2 ma postać:

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

gdzie n oznacza rozmiar próby.

Oznaczmy poprzez T następującą statystykę

$$T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)}, \quad \text{gdzie} \quad s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Wówczas statystyka T pochodzi z rozkładu studenta z n-2 stopniami swobody.

Na podstawie statystyki T przedział ufności na poziomie istotności α dla parametru β_1 ma postać

$$\hat{\beta}_1 \pm t_c s(\hat{\beta}_1),$$

gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n-2)$ jest kwantylem rzędu $1 - \frac{\alpha}{2}$ z rozkładu studenta z n-2 stopniami swobody.

Zaimplementujmy powyższe wzory w celu wyznaczenia 95% przedziału ufności dla β_1.

```
s2 <- function(X, Y){
  n <- length(X)
  b_0 <- beta_0(X,Y)
  b_1 <- beta_1(X,Y)

  return((1/(n-2))*sum((Y-b_0-b_1*X)^2))
}

conf_interval_b1 <- function(X, Y, alpha){
  n <- length(X)
  b_1 <- beta_1(X,Y)
  b_0 <- beta_0(X,Y)
  s2 <- s2(X,Y)

  t <- qt(1-alpha/2, df = n-2)
  SE <- sqrt(s2/sum((X-mean(X))^2))

  return(c(b_1-t*SE, b_1+t*SE))
}

conf_int_beta1 <- conf_interval_b1(X, Y, 0.05)
```

Otrzymujemy następujący 95% przedziały ufności dla β_1: [14.061, 16.009].

Różnica pomiędzy wartościami wyznaczonymi za pomocą funkcji wbudowanej i wyznaczonymi powyższymi funkcjami wynosi:

- dla lewego końca przedziału $1.7763568 \times 10^{-15}$,
- dla prawego końca przedziału $3.5527137 \times 10^{-15}$.

Otrzymujemy zatem praktycznie takie same przedziały ufności w przypadku funkcji wbudowanej i własnej implementacji funkcji.

c) Test istotności dla współczynnika kierunkowego

Zbadajmy, czy $\beta_1 \neq 0$. Wówczas bowiem pomiędzy zmiennymi X i Y NIE ma zależności liniowej.

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_1 : \beta_1 \neq 0$$

Statystyka testowa T jest postaci $T = \frac{\hat{\beta}_1 - 0}{s(\hat{\beta}_1)}$.

Odrzucamy hipotezę zerową na poziomie istotności α, gdy $|T| > t_c$, gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n-2)$ jest kwantylem rzędu $1 - \frac{\alpha}{2}$ z rozkładu studenta z n-2 stopniami swobody.

p-wartość dla tego problemu:

$p = P(|z| > |T|)$, gdzie $z \sim t(n\text{-}2)$.

Funkcje wbudowane

Wyznaczmy statystykę testową T oraz p-wartość, wykorzystując funkcje wbudowane.

```
T_R <- summary(LinMod)$coefficients[2,3]
df_R <- length(X)-2
p_val_R <- summary(LinMod)$coefficients[2,4]
```

- Statystyka testowa $T = 31.123$.
- Liczba stopni swobody wynosi n - 2 = 43.
- Natomiast p-wartość jest równa $p = 4.0090321 \times 10^{-31}$.

Na podstawie p-wartości wnioskujemy, że z prawdopodobieństwem równym 1-$p \approx 1$ odrzucamy hipotezę zerową, tzn. wnioskujemy, że $\beta_1 \neq 0$.

Własna implemetacja funkcji

```
T_statistic_beta1 <- function(X,Y){
  SE <- sqrt(s2(X,Y)/sum((X-mean(X))^2))
  return( c(beta_1(X,Y)/SE, length(X)-2))
}

T_own <- T_statistic_beta1(X, Y)
```

```
p_value_beta1 <- function(X,Y){
  T_stat <- T_statistic_beta1(X,Y)
  p <- 1-pt(q=abs(T_stat[1]), df=length(X)-2)
  return(2*p)
}

p_val_own <- p_value_beta1(X, Y)
```

Otrzymujemy następujące wartości:

- Statystyka testowa $T = 31.123$.
- Liczba stopni swobody wynosi n - 2 = 43.
- Natomiast p-wartość jest równa $p = 0$.

Na podstawie p-wartości wnioskujemy, że z prawdopodobieństwem równym 1-$p = 1$ odrzucamy hipotezę zerową, tzn. wnioskujemy, że $\beta_1 \neq 0$.

Różnica pomiędzy wyznaczonymi wartościami statystyki testowej T wynosi $7.1054274 \times 10^{-15}$, różnica pomiędzy wyznaczonymi p-wartościami jest równa $4.0090321 \times 10^{-31}$. Otrzymujemy zatem praktycznie takie same wartości statystyki testowej T oraz p-wartości przy wykorzystaniu własnych i wbudowanych funkcji.

Zadanie 3

Naszym zadaniem jest estymacja μ_h - wartości oczekiwanej Y_h (Y_h to wartość zmiennej objaśnianej dla punktu X_h, w zadaniu $X_h = 11$). Estymator μ_h dany jest następującą zależnością: $\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$.

Otrzymaliśmy analogiczne wyniki dla $\hat{\beta}_0, \hat{\beta}_1$, stosując funkcje wbudowane i własne implementacje. Przyjmijmy więc wartości wyznaczone przez funkcje wbudowane.

Zatem dla h=11 mamy $\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h \approx$ -0.58 + 15.04 ·11 = 164.81.

Przedział ufności

Przedział ufności dla μ_h wynosi (na poziomie ufności 95%):

Własna implementacja

Oznaczmy przez T następującą statystykę:

$$T = \frac{\hat{\mu}_h - \mathrm{E}(\hat{\mu}_h)}{s(\hat{\mu}_h)}, \quad \text{gdzie} \quad s^2(\hat{\mu}_h) = s^2 \left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right).$$

Wówczas statystyka T pochodzi z rozkładu studenta z n - 2 stopniami swobody. Na podstawie statystyki T możemy skonstruować przedział ufności o współczynniku ufności $1 - \alpha$ dla parametru $\mathrm{E}(Y_h)$:

$$\hat{\mu}_h \pm t_c s(\hat{\mu}_h).$$

Zaimplementujmy powyższe wzory:

```
conf_int_mean_Y_h <- function(X,Y, alpha, X_h){
  n <- length(X)
  b_0 <- beta_0(X,Y)
  b_1 <- beta_1(X,Y)
  mi_h <- b_0 + b_1*X_h

  s_mi_h <- sqrt(s2(X, Y)*(1/n + (X_h-mean(X))^2/sum((X-mean(X))^2)))
  t_c <- qt(1-alpha/2, n-2)
  return(c(mi_h-t_c*s_mi_h, mi_h+t_c*s_mi_h))
}
conf_int_mean_Y_h <- conf_int_mean_Y_h(X, Y, 0.05, 11)
```

Wykorzystując powyższą funkcję, otrzymujemy następujący 95% przedział ufnosci dla dla μ_h:

[158.48, 171.14].

Funkcje wbudowane

Wykorzystując funkcję *predict* otrzymujemy:

$\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h \approx 164.81$, czyli taką samą wartość, jak poprzednio (wykorzystując wzór teoretyczny na $\hat{\mu}_h$).

Wyznaczmy przedział ufności dla μ_h:

```
x_pred <- data.frame(Number=c(11))
conf_int_mean_Y_h_R <- predict(LinMod, x_pred, interval = "confidence")
```

Wykorzystując powyższą funkcję *predict*, otrzymujemy następujący 95% przedział ufnosci dla dla μ_h:

[158.48, 171.14]. Przedziały ufności otrzymane obiema metodami są identyczne (z dokładnością do co najmniej dwóch miejsc po przecinku).

Zadanie 4

Tym razem naszym celem jest wyznaczenie 95% **przedziału predykcyjnego** dla Y_h.

Predykcja punktowa Y_h jest postaci $\hat{Y}_h = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$. Ma ona dokładnie taką samą postać, jak wyznaczona w poprzednim zadaniu wartość estymatora $\mathrm{E}(Y_h)$. Z założenia nie potrafimy bowiem modelować zachowania błędu ϵ, który ma symetryczny rozkład. Błąd ϵ wpływa jednak znacząco na własności wariancji predykcji. Mamy:

$$T = \frac{Y_h - \hat{\mu}_h}{s(pred)} \sim t(n-2), \qquad \text{gdzie} \quad s^2(pred) = s^2 \left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).$$

Przedział predykcyjny dla Y_h jest postaci:

$$\hat{\mu}_h \pm t_c \cdot s(pred).$$

Zaimplementujmy powyższe wzory:

```
pred_int_Y_h <- function(X,Y, alpha, X_h){
  n <- length(X)
  b_0 <- beta_0(X,Y)
  b_1 <- beta_1(X,Y)
  mi_h <- b_0 + b_1*X_h

  s_mi_h <- sqrt(s2(X, Y)*(1 + 1/n + (X_h-mean(X))^2/sum((X-mean(X))^2)))
  t_c <- qt(1-alpha/2, n-2)
  return(c(mi_h-t_c*s_mi_h, mi_h+t_c*s_mi_h))
}
pred_int_Y_h <- pred_int_Y_h(X, Y, 0.05, 11)
```

Wykorzystując powyższą funkcję, otrzymujemy następujący 95% przedział predykcyjny dla Y_h:

[145.75, 183.87].

Funkcje wbudowane

Wyznaczmy 95% przedział predykcyjny dla Y_h, wykorzystując funkcje wbudowane:

```
x_pred <- data.frame(Number=c(11))
pred_int_Y_h_R <- predict(LinMod, x_pred, interval = "prediction")
```

Wykorzystując powyższą funkcję *predict*, otrzymujemy następujący 95% przedział predykcyjny dla Y_h:

[145.75, 183.87].

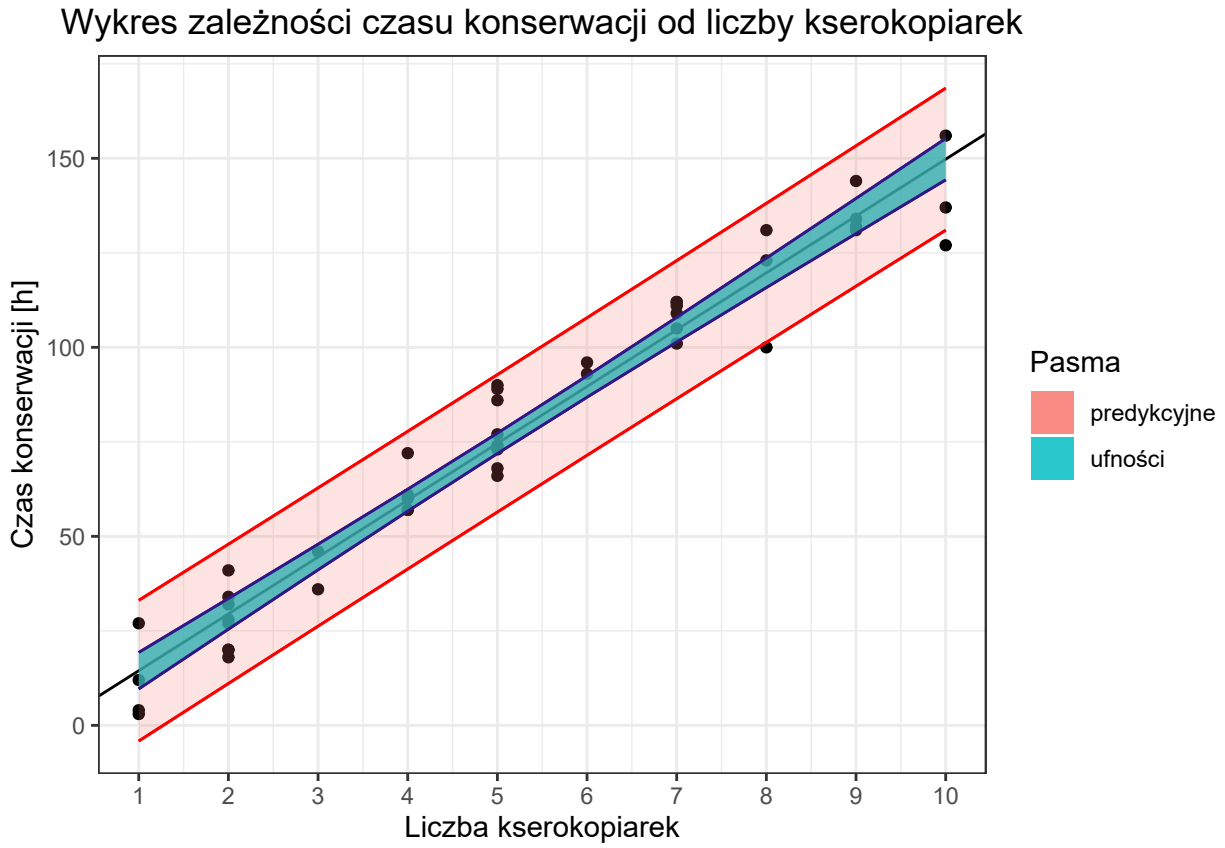
Przedziały predykcyjne otrzymane obiema metodami są identyczne (z dokładnością do co najmniej dwóch miejsc po przecinku).

Zadanie 5

Narysujmy ponownie wykres zależności czasu konserwacji od liczby kserokopiarek, zaznaczając przedziały ufności (**pasmo ufności** dla prostej regresji) i przedziały predykcyjne dla poszczególnych obserwacji.

W tym celu wyznaczmy przedziały ufności i przedziały predykcyjne dla poszczególnych obserwacji, wykorzystując funkcję *predict*:

```
lwr <- predict(LinMod, interval = "prediction")[,2]
pwr <- predict(LinMod, interval = "prediction")[,3]

lwr_E <- predict(LinMod, interval = "confidence")[,2]
pwr_E <- predict(LinMod, interval = "confidence")[,3]
```



Analiza wykresu pozwala stwierdzić, że zgodnie z założeniami/wzorami teoretycznymi oba pasma są symetryczne względem prostej $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, bowiem wartość estymatora $\hat{\mu}_h$ jest równa $\hat{\beta}_0 + \hat{\beta}_1 X_h$ zarówno dla przedziałów ufności jak i przedziałów predykcyjnych. Dla dowolnego X_h przedział predykcyjny jest szerszy od przedziału ufności. Wynika to wprost ze wzorów - przedziały predykcyjne konstruuje się analogicznie do przedziałów ufności, mają jednak większą wariancję. Nie jest to zaskakujące. Przedziały ufności wyznaczamy bowiem dla wartości oczekiwanej $\mathrm{E}(Y_h)$, natomiast przedziały predykcyjne dla Y_h.

95% przedział predykcyjny dla Y_h oznacza, że tylko (co najwyżej) 5% obserwacji nie znajduje się w tym przedziale. Sprawdźmy, jaka część obserwacji znalazła się poza wyznczonym pasmem predykcyjnym.

Poza pasmem predykcyjnym znalazło się 4.444% obserwacji.

Zadanie 6

a)

Załóżmy, że rozmiar próby $n = 40$, wariancja błędu $\sigma^2 = 120$, SSX $= \sum_{i=1}^{n}(X_i - \bar{X})^2 = 1000$.

Naszym zadaniem jest wyznaczenie mocy testu. Zakładamy poziom istotności $\alpha = 0.05$. Przyjmujemy, że prawdziwa wartość β_1 wynosi 1.

Mamy:

$$\sigma^2(\hat{\beta}_1) = \sigma^2/SSX$$

$$\text{Parametr niecentralności} \quad \delta = \beta_1/\sigma(\hat{\beta}_1).$$

Statystyka testowa $T \sim t(n\text{-}2, \delta)$, zatem **moc testu** $\pi(\beta_1 = 1) = P_{\beta_1=1}(|T| > t_c) = F_{\beta_1=1}(-t_c) + 1 - F_{\beta_1=1}(t_c)$.

Zaimplemetujmy powyższe wzory:

```
power_of_rejection_H0 <- function(n, sigma2, SSX, alpha, beta1){
  sigma2_beta1 <- sigma2/SSX
  delta <- beta1/sqrt(sigma2_beta1)
  t_c <- qt(1-alpha/2, n-2)

  return(pt(-t_c, n-2, delta) + 1 - pt(t_c, n-2, delta))
}

n <- 25
sigma2 <- 120
SSX <- 1000
alpha = 0.05

(power_1 <- power_of_rejection_H0(n, sigma2, SSX, alpha, beta1=1))
```

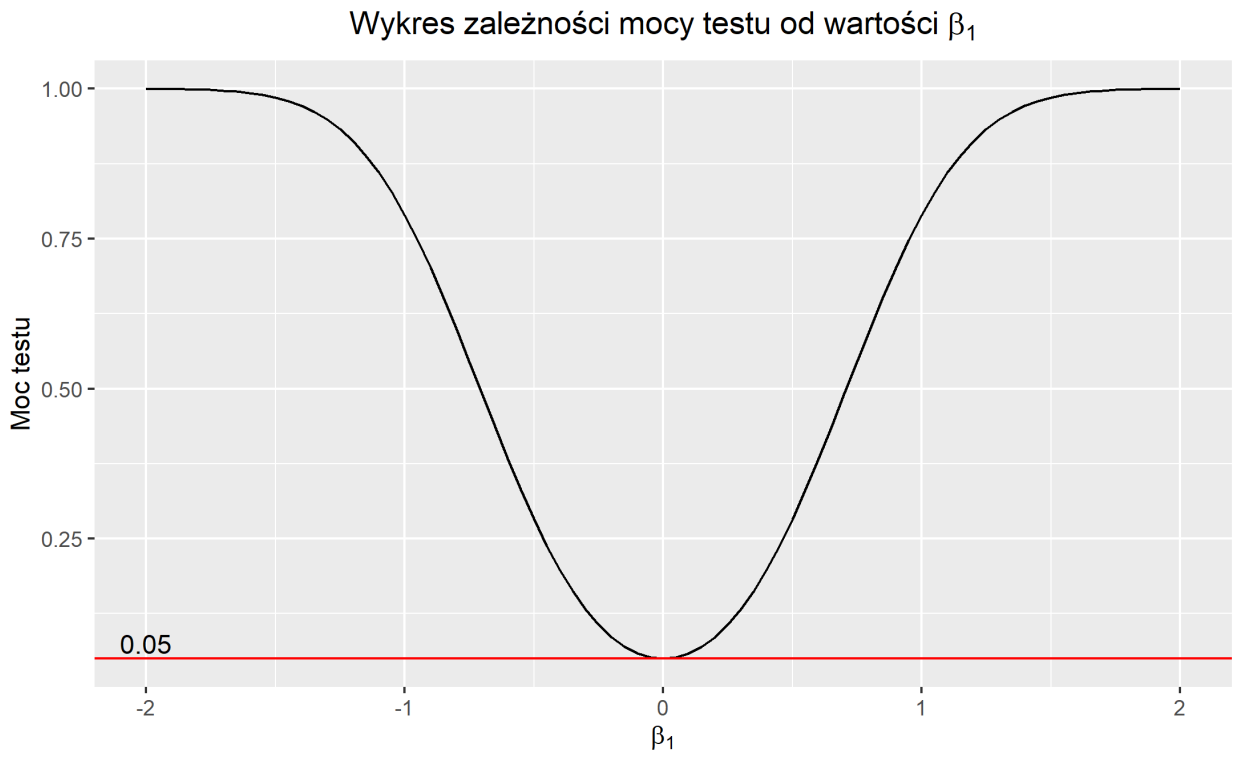
```
## [1] 0.7894431
```

Wyznaczona moc testu wynosi 0.7894431, to znaczy jeśli prawdziwa wartość $\beta_1 = 1$, to na poziomie istotności 0.05 odrzucamy hipotezę zerową mówiącą, że $\beta_1 = 0$ z prawdopodobieństwem $\pi(\beta_1 = 1) = 0.7894431$.

b)

Narysujmy wykres zależności mocy testu od prawdziwej wartości β_1.

```
beta1_vec <- seq(-2,2,0.05)
powers <- sapply(beta1_vec, power_of_rejection_H0, n=n, sigma2=sigma2, SSX=SSX, alpha=alpha)

data_power <- data.frame((beta1_vec), (powers))
```



Analiza wykresu pozwala stwierdzić, że im bardziej prawdziwa wartość β_1 różni się od 0 (czyli od hipotezy zerowej $H_0 : \beta_1 = 0$), tym większa jest moc testu. Wykres jest symetryczny. Badania wykonywaliśmy na poziomie istotności $\alpha = 0.05$. Zauważmy, że moc testu osiąga minimalną wartość równą α w momencie, gdy prawdziwa wartość $\beta_1 = 0$. Nie jest to zaskakujące. W tym przypadku prawdopodobieństwo odrzucenie hipotezy zerowej (prawdziwej!) wynosi właśnie założony poziom istotności α.

Dla jakiego β_1 otrzymujemy dużą moc?

Sprawdźmy, dla jakich wartości β_1 moc testu wynosi 1 z dokładnością do trzech miejsc po przecinku:

```
indx_min <- min(abs(data_power[,1][which(1-data_power[,2]<0.001)]))
```

Żądana dokładność zachodzi dla $|\beta_1| \geq 1.85$.

Zadanie 7

Naszym zadaniem jest wygenerownie wektora $X = (X_1, ..., X_{200})^T$ pochodzącego wielowymiarowego rozkładu normalnego $N(0, \frac{1}{200}I)$.

```
N <- 200
sigma_x <- 1/sqrt(200)
X_200 <- sigma_x*rnorm(N)
```

Następnie wygenerujmy $M = 1000$ wektorów Y pochodzących z następującego modelu liniowego:

$$Y = 5 + \beta_1 X + \epsilon.$$

Zbadajmy hipotezę zerową mówiącą, że $\beta_1 = 0$ w zależności od zadanych wartości/rozkładów β_1 i ϵ.

a)

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

Stwórzmy model regresji liniowej pomiędzy X a Y i zbadajmy hipotezę zerową $H_0 : \beta_1 = 0$. W tym celu dla każdego z $M = 1000$ powtórzeń eksperymentu zbadajmy hipotezę zerową na poziomie istotności $\alpha = 0.05$. Następnie obliczmy, w jakiej części eksperymentów odrzuciliśmy hipotezę zerową.

```
rejection_exercise_7_N <- function(X, alpha, beta1){
  n <- length(X)
  epsilon <- rnorm(n)
  Y <- 5 + beta1*X + epsilon

  data <- data.frame(X,Y)
  LinMod <- lm(Y~X, data)
  t <- summary(LinMod)$coefficients[2,3]
  t_c <- qt(1-alpha/2, n-2)

  if(abs(t)>t_c){return(1)} else{return(0)}

}

M <- 1000
alpha <- 0.05

(percent_of_rejections_a <- mean(replicate(M, rejection_exercise_7_N(X_200, alpha, beta1=0))))
```

```
## [1] 0.054
```

Odrzuciliśmy hipotezę zerową w 5.4% przypadków. Wynik zgadza się w przybliżeniu z teoretycznym prawdopodobieństwem popełnienia błędu I rodzaju. Skoro $\beta_1 = 0$, to hipoteza zerowa ($H_0 : \beta_1 = 0$) jest prawdziwa. Odrzucając prawdziwą hipotezę zerową, popełniamy błąd I rodzaju równy założonemu poziomowi istotności $\alpha = 0.05$. W założeniach modelu przyjmujemy, że ϵ mają rozkład normalny (tak jak w tym podpunkcie).

b)

$\beta_1 = 0, \quad \epsilon_1, ..., \epsilon_{200} \sim exp(\lambda = 1)$

W tym przypadku prawdziwa wartość β_1 wciąż wynosi 0. Zmienia się natomiast rozkład ϵ. Nie są spełnione założenia modelu regresji liniowej (ϵ nie ma rozkładu normalnego). Sprawdźmy, czy wpłynie to istotnie na procent odrzucenia hipotezy zerowej.

```
rejection_exercise_7_exp <- function(X, alpha, beta1){
  n <- length(X)
  epsilon <- rexp(n)
  Y <- 5 + beta1*X + epsilon

  data <- data.frame(X,Y)
  LinMod <- lm(Y~X, data)
  t <- summary(LinMod)$coefficients[2,3]
  t_c <- qt(1-alpha/2, n-2)

  if(abs(t)>t_c){return(1)} else{return(0)}

}

(percent_of_rejections_b <- mean(replicate(M, rejection_exercise_7_exp(X_200, alpha, beta1=0))))
```

```
## [1] 0.044
```

Odrzuciliśmy hipotezę zerową w 4.4% przypadków. Zatem otrzymany wynik również zgadza się w przybliżeniu z teoretyczną wartością popełnienia błędu I rodzaju równą $\alpha = 0.05$. Najprawdopodobniej spowodowane jest to przyjętym rozkładem wykładniczym dla ϵ. Rozkład wykładniczy nie jest symetryczny, ale tak jak w przypadku standardowego rozkładu normalnego największa masa prawdopodobieństwa skupia się wokół wartości 0 (wartości są jednak zawsze nieujemne). Z tego powodu błąd ϵ o takim rozkładzie nie zaburza w znaczący sposób modelu regresji liniowej.

c)

W tym przypadku $\beta_1 = 1.5, \quad \epsilon \sim N(0, I)$, zatem prawdziwa jest hipoteza alternatywna. Błąd ϵ pochodzi z rozkładu normalnego zgodnie teoretycznymi założeniami modelu regresji liniowej.

Sprawdźmy, w jakiej części eksperymentów odrzucamy w tym przypadku hipotezę zerową.

```
(percent_of_rejections_c <- mean(replicate(M, rejection_exercise_7_N(X_200, alpha, beta1=1.5))))
```

```
## [1] 0.341
```

W tym przypadku odrzucamy hipotezę zerową (fałszywą, bo $\beta 1 = 1.5 \neq 0$) w 34.1% przypadków. Wartość ta stanowi wyestymowaną moc testu. Dla porównania teoretyczna moc testu (przy założeniach modelu regresji liniowej, w tym założeniu o rozkładzie normalnym ϵ) wynosi:

```
n <- length(X_200)
sigma2 <- 1 # wariancja epsilon
SSX <- sum((X_200-mean(X_200))^2)
(teoretic_power <- power_of_rejection_H0(n, sigma2, SSX, alpha, beta1=1.5))
```

```
## [1] 0.3325041
```

Zatem wyznaczona w eksperymencie moc testu jest w przybliżeniu równa mocy teoretycznej wynoszącej 33.25%.

d)

Sprawdźmy, jak wyestymowana moc testu zachowa się w sytuacji, gdy $\beta_1 = 1.5$, ale ϵ nie pochodzi z rozkładu normalnego, tylko z rozkładu wykładniczego $exp(\lambda = 1)$. W tym przypadku nie jest spełnione założenie modelu regresji liniowej.

Sprawdźmy, w jakiej części eksperymentów odrzucamy w tym przypadku hipotezę zerową.

```
(percent_of_rejections_d <- mean(replicate(M, rejection_exercise_7_exp(X_200, alpha, beta1=1.5))))
```

```
## [1] 0.317
```

W tym przypadku odrzucamy hipotezę zerową (fałszywą, bo $\beta_1 = 1.5 \neq 0$) w 31.7% przypadków. Wartość ta stanowi wyestymowaną moc testu. Dla porównania teoretyczna moc testu (przy założeniach modelu regresji liniowej, w tym założeniu o rozkładzie normalnym ϵ o wariancji równej wariancji zadanego rozkładu wykładniczego $var = \frac{1}{\lambda^2} = \frac{1}{1} = 1$) wynosi 33.25% (obliczyliśmy tę wartość w poprzednim podpunkcie). Zatem wyestymowana moc testu znów w przybliżeniu odpowiada teoretycznej wartości mocy testu, pomimo że zadany rozkład ϵ nie jest normalny.

Wnioski

Powyższe podpunkty pozwalają wysunąć następujące wnioski:

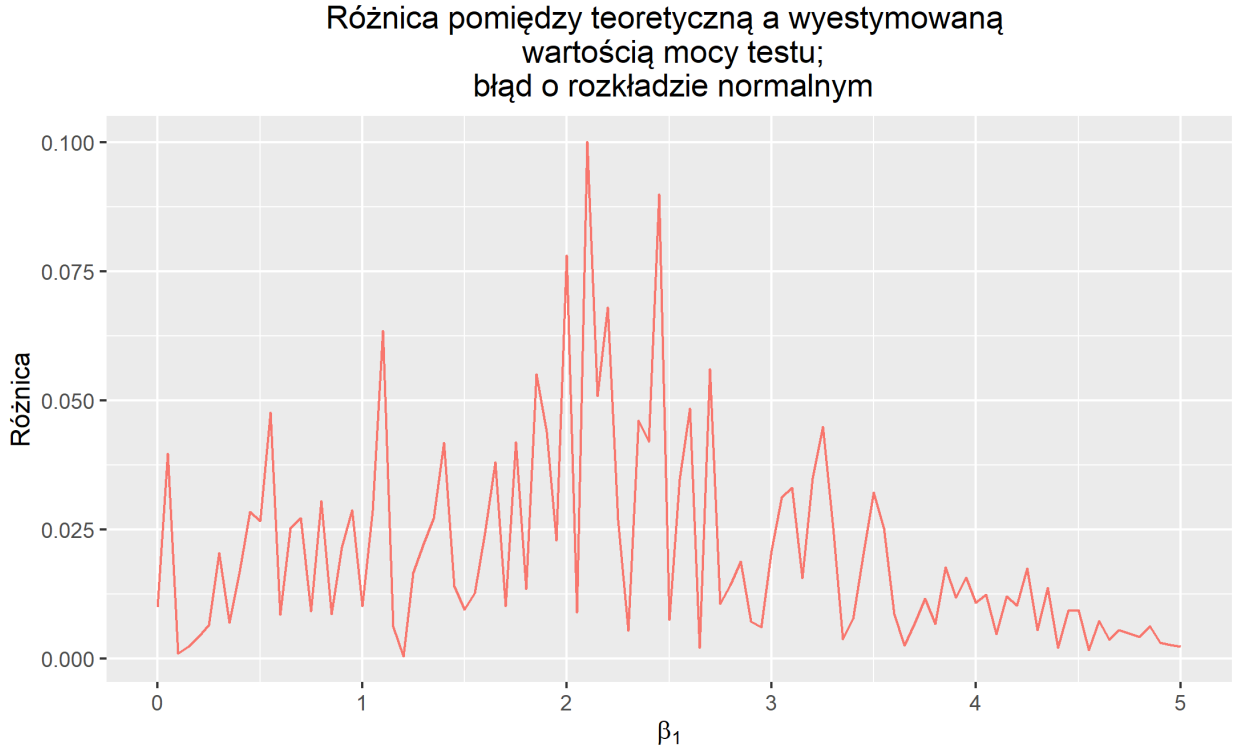
- W przypadku gdy ϵ ma rozkład normalny spełnione są założenia modelu liniowego. Liczba odrzuceń hipotezy zerowej zgadza się wówczas w przybliżeniu z teoretyczną wartością mocy testu (gdy prawdziwa jest hipoteza alternatywna) oraz z teoretyczną wartością popełnienia błędu I rodzaju (gdy prawdziwa jest hipoteza zerowa). Świadczy to o zgodności z założonym modelem regresji liniowej.
- Gdy ϵ nie pochodził z rozkładu normalnego, lecz wykładniczego $exp(\lambda = 1)$, to również otrzymaliśmy w przybliżeniu wyniki zgodne z teoretycznymi (dla mocy testu albo błędu I rodzaju). Zatem w tym przypadku mimo niespełnienia założenia modelu regresji liniowej o normalności rozkładu ϵ, otrzymaliśmy model w przybliżeniu zgodny z modelem regresji liniowej.

Dokładność modelu w zależności od współczynnika kierunkowego

Rozkład normalny

Sprawdźmy, jak zachowa się model o $\epsilon \sim N(0,1)$ dla różnych wartości β_1:

```
percent_rejection <- function(beta1){
  return(mean(replicate(100, rejection_exercise_7_N(X_200, 0.05, beta1))))
}

vec_beta <- seq(0,5,0.05)
rejections <- sapply(vec_beta, percent_rejection)


n <- length(X_200)
sigma2 <- 1 # wariancja epsilon
SSX <- sum((X_200-mean(X_200))^2)
alpha = 0.05

teoretic_power_X_200 <- function(beta1){
  return(power_of_rejection_H0(n, sigma2, SSX, alpha, beta1))
}

teoretic_rejections <- sapply(vec_beta, teoretic_power_X_200)

data_for_plot <- data.frame(vec_beta, abs(rejections-teoretic_rejections))
```



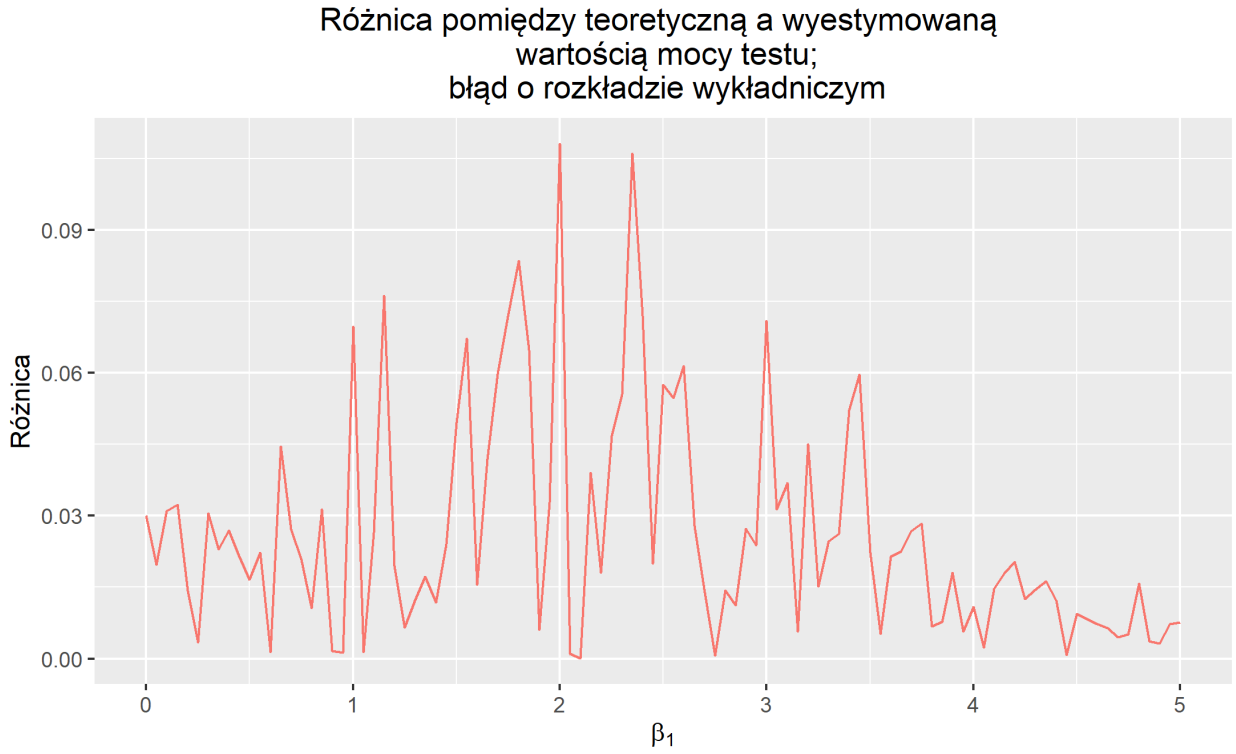
Analiza powyższego wykresu pozwala stwierdzić, że w przypadku wartości β_1 bliskich 0 różnica pomiędzy wyestymowaną a teoretyczną wartością mocy testu jest niewielka. Wówczas istotną rolę w modelu regresji liniowej pełni ϵ. W przypadku gdy ma on rozkład normalny (zgodny z modelem) nie zaburza on istotnie modelu - nie sprawia, że β_1 zdaje się mieć wartość istotnie większą od zera.

Również w przypadku dużych wartości β_1 (powyżej 3) różnica jest bardzo mała. W tej sytuacji z kolei czynnik $\beta_1 X$ znaczącą dominuje model. Wartości błędów ϵ nie wpływają praktycznie wcale na wartości Y.

Natomiast dla $\beta_1 \in (1, 3)$ obserwujemy znaczące odchylenia wartości wyestymowanych od teoretycznych. Dla takich β_1 trudno na podstawie badanej próby rozstrzygnąć prawdziwość hipotezy zerowej (prawdziwa wartość β_1 nie jest wówczas istotnie większe od zera, nie jest również w przybliżeniu równa zero).

Rozkład wykładniczy

Dokonajmy analogicznej analizy, zakładając tym razem, że $\epsilon \sim exp(\lambda = 1)$:



Otrzymujemy wyniki podobne do poprzedniej sytuacji. Dla wartości β_1 bliskich 0 różnica pomiędzy wyestymowaną a teoretyczną wartością mocy testu jest niewielka. Również dla dużych wartości β_1 (powyżej 3.5) różnice są niewielkie. Zatem błędy ϵ o rozkładzie wykładniczym zachowują się w modelu regresji liniowej podobnie jak błędy o rozkładzie normalnym - nie zaburzają istotnie modelu.

Podobnie jak w przypadku rozkładu normalnego błędu największe różnice występują dla $\beta_1 \in (1, 3)$. Wówczas trudno stwierdzić jednoznacznie prawdziwość hipotezy zerowej.

Niemniej jednak rozkład wykładniczy błędów ϵ sprawił, że model zachowuje się w podobny sposób jak założony teoretyczny model regresji liniowej.