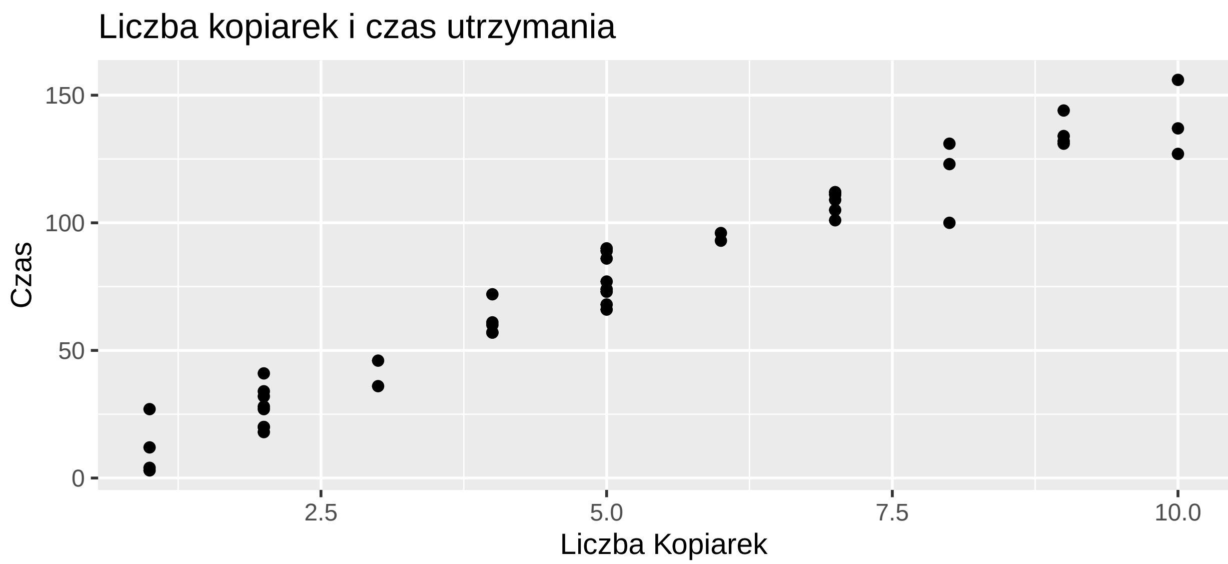
Raport 2

Magdalena Potok

2024-01-28

Zadanie 1

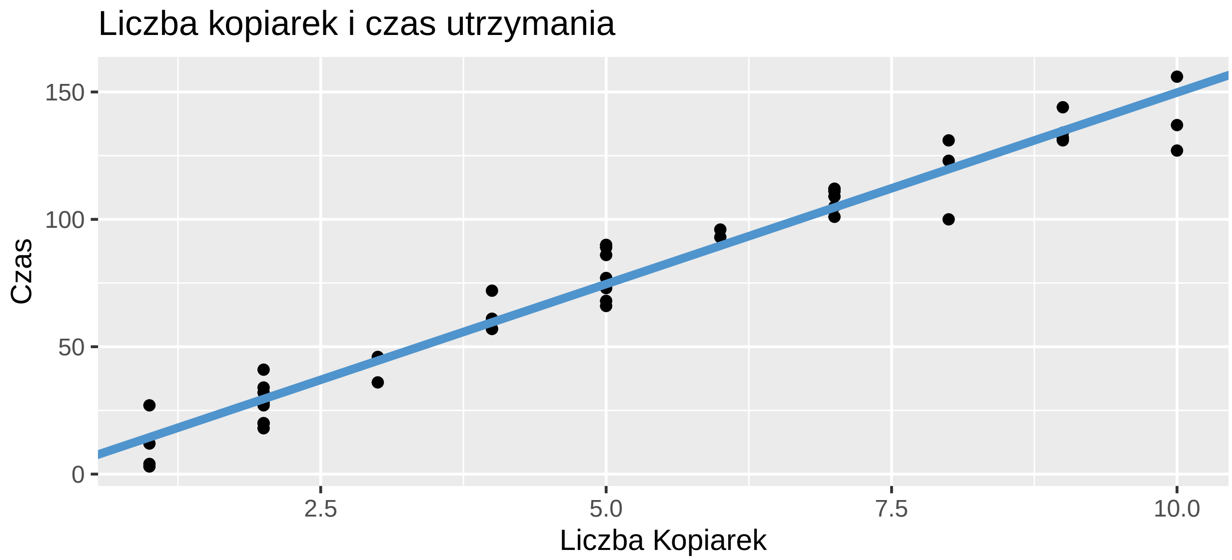
W zadaniu wczytałam dane z pliku **ch01pr20.txt**, która zawiera liczbę kopiarek oraz czas potrzebny na utrzymanie każdej z kopiarek.
Uzyskane dane przedstawię na poniższym wykresie.



Łatwo można zauważyć, że punkty na wykresie układają się w linie prostą, rosnącą. Oznacza to, że zależność między ilością kopiarek, a czasem ich utrzymania jest w przybliżeniu liniowa.

Zadanie 2

W tym zadaniu wyznaczę regresję liniową między czasem obsługi, a liczba obsługiwanych maszyn. Wynik został przedstawiony na wykresie poniżej.



Współczynniki funkcji liniowej, która wyznacza nam regresję liniową można wyliczyć za pomocą poleceń wbudowany w R przy pomocy funkcji **reg1$coefficients**, lub można wyliczyć je za pomocą teoretycznych wzorów podanych na wykładzie.

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

Tak wyliczyłam za pomocą funkcji dostępnych w R:

```
## Intercept:  -0.5801567
```

```
## Slope:  15.03525
```

Natomiast teoretyczne wartości wyliczyłam w następujący sposób:

```r
b1 <- sum((X- mean(X))*(Y-mean(Y)))/sum((X-mean(X))^2)
b0 <- mean(Y)-b1*mean(X)
```

```
## Intercept:  -0.5801567
```

```
## Slope:  15.03525
```

Zadanie 3

W tym zadaniu wyznaczę 95% przedział ufności dla wyliczonego powyżej slope'a oraz intercept'a. Zrobię to przy pomocy funkcji wbudowanej w R oraz wzorów teoretycznych.

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

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

Przedział ufności dla slope'a konstruujemy w sposób:

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

Przedział ufności dla intercept'a wygląda tak:

$$\hat{\beta}_0 \pm t_c s(\hat{\beta}_0),$$

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

Teraz pokażę wyliczenia teoretyczne dla $\hat{\beta}_0$ i $\hat{\beta}_0$:

```
n <- length(data[,1])
s <- 1/(n-2)* sum(((Y-b0-b1*X))^2)

sb0 <- s*(1/n+mean(X)^2/(sum((X-mean(X))^2)))
l_b0 <- b0 - qt(1-0.05/2, n - 2)*sqrt(sb0)
p_b0 <- b0 + qt(1-0.05/2, n - 2)*sqrt(sb0)

sb1 <- s/sum((X-mean(X))^2)
l_b1 <- b1 - qt(1-0.05/2, n - 2)*sqrt(sb1)
p_b1 <- b1 + qt(1-0.05/2, n - 2)*sqrt(sb1)
```

```
##    Parametr Lewy_koniec Prawy_koniec
## 1 Intercept   -6.234843     5.074529
## 2     Slope   14.061010    16.009486
```

Natomiast te same wyniki można otrzymać dużo krócej korzystając z funkcji wbudowanej w R, czyli *confint*.

```
ci <- confint(lm(V1~V2, data), level = 0.95)
```

```
##                2.5 %    97.5 %
## Intercept -6.234843  5.074529
## slope     14.061010 16.009486
```

Zadanie 4

Przeprowadzę testy istotności dla slope'a i intercepta. Wykonam to za pomocą wzorów teoretycznych oraz poleceń wbudowanych w R.

- Test istotności slope'a

W ocenie istotności współczynnika nachylenia (slope'a) kluczowym zagadnieniem jest to, czy zmienna wyjaśniana jest zależna od zmiennej wyjaśniającej. W równaniu $Y_i = \beta_1 X_i + \beta_0 + \epsilon_i$ współczynnik β_1 odpowiada za zależność między zmiennymi. Aby określić, czy są one ze sobą powiązane formułujemy następujące hipotezy:

$$H_0: \ \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq \ 0$$

Aby przetestować tę hipotezę musimy wyliczyć statystykę testową $T = \frac{\hat{\beta}_1 - 0}{s(\hat{\beta}_1)}$. Odrzucamy hipotezę zerową, 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. Inny sposób sprawdzenia hipotezy, to policzenie wartości $p = P(|z| > |T|)$, gdzie $z \sim t(n-2)$ oraz sprawdzenie czy wyliczona p-wartość jest mniejsza niż ustalony poziom istotności α - wtedy odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej.

- test wzorami teoretycznymi

```
T <- b1/sqrt(sb1)
tc <- qt(1-0.05/2, n - 2)
T # statystyka t
```

```
## [1] 31.12326
```

```r
abs(T) > tc
```

```
## [1] TRUE
```

Zatem odrzucamy hipotezę zerową z 95% pewnością, ponieważ przyjęłam poziom istotności $\alpha = 0.05$. Sprawdźmy teraz test za pomocą wyliczenia p-wartości. Użyję do tego funkcji **pt(x, df)**, która oblicza dystrybuantę rozkładu t-studenta w punkcie x dla df stopni swobody. Mnożę otrzymaną wartość razy 2, ponieważ interesuje nasz obszar po lewej i po prawej stronie rozkładu t-Studenta.

```r
p_value <- 2 * (pt(-T,df =  n - 2))
p_value # p-wartość
```

```
## [1] 4.009032e-31
```

```r
p_value < 0.05
```

```
## [1] TRUE
```

Ponownie wyszło nam, że odrzucamy hipotezę zerową, czyli powinniśmy przyjąć, że nasze zmienne są skorelowane.

- test za pomocą poleceń wbudowanych w R

Aby sprawdzić hipotezę za pomocą poleceń wbudowanych w R posłużyłam się funkcją **summary**, w której możemy odczytać wartość statystyki t oraz p-wartość w następujący sposób:

```r
summary(reg1)$coefficients[2, "t value"] # statystyka t
```

```
## [1] 31.12326
```

```r
summary(reg1)$coefficients[2, "Pr(>|t|)"] # p-wartość
```

```
## [1] 4.009032e-31
```

Następnie testujemy naszą hipotezę na oba spodoby:

```r
abs(summary(reg1)$coefficients[2, "t value"]) > tc
```

```
## [1] TRUE
```

```r
summary(reg1)$coefficients[2, "Pr(>|t|)"] < 0.05
```

```
## [1] TRUE
```

A więc odrzucamy z 95% pewnością hipotezę zerową i ustalamy, że X i Y są skorelowane.

- Test istotności intercepta

W przypadku β_0 będziemy testować, czy intercept ma jakąś konkretną wartość β_{00}, która jest dowolną liczbą rzeczywistą. Aby przeprowadzić taki test konstruujemy następujące hipotezy:

$$H_0: \ \beta_0 = \beta_{00} \qquad vs \qquad H_1: \beta_0 \neq \ \ \beta_{00}$$

Aby uzyskać odpowiedź należy policzyć statystykę testową zadaną wzorem: $T = \frac{\hat{\beta}_0 - \beta_{00}}{s(\hat{\beta}_0)}$, gdzie $s(\hat{\beta}_0) = s^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$. Hipotezę odrzucamy, gdy $|T| > t_c$, gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n-2)$, lub gdy $p = P(|z| > T) < \alpha$, gdzie $z \sim t(n-2)$.
Przetestuję teraz istotność intercepta dla $\beta_{00} = 0$.

- test wzorami teoretycznymi

```r
b00 <- 0
T <- (b0 - b00)/sqrt(sb0)
```

```r
tc <- qt(1-0.05/2, n - 2)
T # statystyka t
```

```
## [1] -0.2069076
```

```r
abs(T) > tc
```

```
## [1] FALSE
```

Wnioskujemy więc, że z 95% pewnością nie jesteśmy w stanie odrzucić hipotezy zerowej, że $\beta_0 = 0$, nie oznacza to jednak, że tyle intercept wynosi.
Sprawdżmhy to samo przy pomocy p-wartości.

```r
p_value <- 2 * (1 - pt(abs(T), n - 2))
p_value # p-wartość
```

```
## [1] 0.8370587
```

```r
p_value < 0.05
```

```
## [1] FALSE
```

Mamy ten sam wniosek j.w.

- test za pomocą poleceń wbudowanych w R

Wykorzystamy polecenie summary do odczytania statystyki t oraz p-wartości.

```r
summary(reg1)$coefficients[1, "t value"]
```

```
## [1] -0.2069076
```

```r
summary(reg1)$coefficients[1, "Pr(>|t|)"]
```

```
## [1] 0.8370587
```

Wyniki wyszły te same, więc wyniki przeprowadzanego testu będą te same. Niestety wynik nie jest satysfakcjonujący, ponieważ nie odrzucenie hipotezy zerowej nie oznacza przyjęcie jej za prawdziwej.

Zadanie 5

Policzę estymator wartości oczekiwanej czasu obsługi, której można oczekiwać, gdyby serwisowanych było $k \in \{1, 5, 8, 11, 25, 100\}$ maszyn oraz 95% przedział ufności dla tej wartości. Obliczę ten przedział za pomocą wzorów teoretycznych oraz poleceń wbudowanych w R.

- wzory teoretyczne

Dla tej wartości chcemy znaleźć przedział ufności:
$E(Y_h) = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$, gdzie $X_h =$ k, $\hat{\mu}_h \sim N(\mu_h, \sigma^2(\hat{\mu}_h))$
Aby wyznaczyć przedział, musimy policzyć:
$s^2(\hat{\mu}_h) = s^2(\frac{1}{n} + \frac{(X_h - \bar{X})}{\sum_{i=1}^{n}(X_i - \hat{X})^2})$
$T = \frac{\hat{\mu}_h - E(\hat{\mu}_h)}{s(\hat{\mu}_h)}$
Jak policzymy powyższe wartości, to przedział ufności na podstawie statystyki T o współczynniku ufności $1 - \alpha$ dla parametru $E(Y_h)$ kontruujemy:

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

```r
k <- c(1,5,8,11,25,100)
mu_h <- b0 + b1 * k
s2mu_h <- s * (1/n + (k - mean(X))^2/sum((X - mean(X))^2))
```

```r
LPU <- round(mu_h - tc * sqrt(s2mu_h),3)
PPU <- round(mu_h + tc * sqrt(s2mu_h),3)
```

k	Lewy_koniec	Prawy_koniec
1	9.636	19.274
5	71.914	77.278
8	115.816	123.588
11	158.475	171.140
25	355.740	394.862
100	1410.461	1595.428

- funkcja wbudowana w R

Funkcja którą użyłam to **predict**, liczy ona estymator wartości oczekiwanej oraz podaje przedział ufności.

```r
pred <- predict(reg1, newdata = data.frame(V2 = k), interval = 'confidence')
```

Wyniki powyższego polecenia przedstawię w tabeli:

k	estym_wart_oczekiwana	lewy_koniec	prawy_koniec	dlugosc_przedzialu
1	14.455	9.636	19.274	9.638
5	74.596	71.914	77.278	5.364
8	119.702	115.816	123.588	7.772
11	164.808	158.475	171.140	12.664
25	375.301	355.740	394.862	39.122
100	1502.945	1410.461	1595.428	184.966

Z tabelki można zauważyć, że najwęższy przedział jest dla k = 5, a co za tym idzie, mamy największą precyzję predykcji dla tego k. Dla k = 5 punkt przewidywany znajduje się najbliżej danych, czyli najbliżej średniej wartości X (która wynosi ok. 5.1). Dla pozostałch k im dalej znajdujemy się od 5, tym większa jest długość przedziału.
Jak spojrzymy na wzór teoretyczny wyliczania $s^2(\hat{u}_h)$, to widać tam, że im bliżej nasz punkt X_h jest średniej $\bar{X}$, tym odchylenie $s(\hat{u}_h)$ jest mniejsze, a co za tym idzie, przedział ufności $\hat{\mu}_h \pm t_c s(\hat{\mu}_h)$ jest węższy.

Zadanie 6

Podam przewidywany czas obsługi, który można oczekiwać, jeśli k maszyn było serwisowanych oraz 95% przedział predykcyjny dla tego czasu, gdzie $k \in \{1, 5, 8, 11, 25, 100\}$. Dla równania regresji liniowej $Y_h = \beta_0 + \beta_1 X_h + \epsilon$, która jesy **nową** zmienną zależną dla zmiennej niezależnej o wartości X_h obliczamy predykcję punktową w następujący sposób: $\hat{Y}_h = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$. Pozwala ona uzyskać konkretną wartość prognozowaną przez model dla określonego zestawu wartości zmiennych niezależnych. Policzę teraz prognozowany czas obsługi oraz 95% przedział predykcyjny dla wartości $X_h = k$.

- za pomocą wzorów teoretycznych

Przedział predykcyjny jest postaci:

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

Widać, że szerokość tego przedziału zależy od s(pred), a liczy się go w następujący sposób:
$s^2(pred) = s^2(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \hat{X})^2})$
Po samym wzorze można dojść do wniosku, że przedział predykcyjny będzie szerszy niż przedział ufności estymacji. Wynika to z faktu, że zawsze $s^2(pred) > s^2(\hat{\mu}_h)$.

```
s2pred <-s * (1+ 1/n + (k - mean(X))^2/sum((X - mean(X))^2))
LPP <- round(mu_h - tc * sqrt(s2pred),3)
PPP <- round(mu_h + tc * sqrt(s2pred),3)
```

k	Lewy_koniec	Prawy_koniec
1	-4.155	33.066
5	56.421	92.771
8	101.311	138.093
11	145.749	183.866
25	348.735	401.867
100	1408.731	1597.159

- funkcja wbudowana w R

Funkcja wbudowana w R, która wyliczy nam przedział predykcyjny, to funkcja **predict**, jedyne co się różni od poprzedniego zadania, to fakt, że argument funkcji 'interval' przyjmuje wartość 'prediction'.

```
pred <- predict(reg1, newdata = data.frame(V2 = k), interval = 'prediction')
```

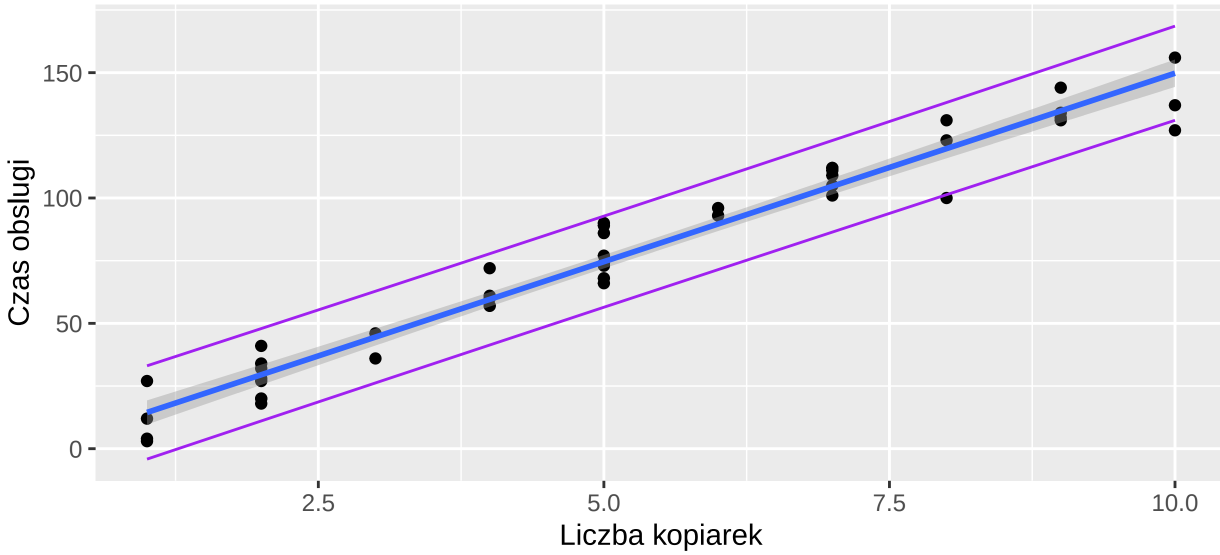
Wyniki powyższej funkcji przedstawię w postaci tabeli.

k	przewid_czas_obslugi	lewy_koniec	prawy_koniec	dlugosc_przedzialu
1	14.455	-4.155	33.066	37.221
5	74.596	56.421	92.771	36.350
8	119.702	101.311	138.093	36.782
11	164.808	145.749	183.866	38.117
25	375.301	348.735	401.867	53.132
100	1502.945	1408.731	1597.159	188.428

Ponownie najwęższy przedział jest dla k = 5, jest tak z tego samego powodu, co w zadaniu 5. Ten punkt znajduje się najbliżej średniej wartości X. Czyli dla wzoru $s^2(pred) = s^2(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \hat{X})^2})$ im bliżej $X_h = k$ jest $\bar{X}$, tym bardziej minimalizujemy różnicę $(X_h - \bar{X})^2$, tym $s^2(pred)$ jest mniejsze, a to jest liczba, która wpływa na długość przedziału.

Zadanie 7

Do wykresu danych z 1. zadania dołączę 95% przedziały ufności oraz przedziały predykcyjne.



Szary obszar na wyrkesie przedstawia 95% przedział ufności dla podanych danych. Można zaobserować, że w punkcie X = 5 przedział ten jest najwęższy, a im dalej od 5 tym bardziej się rozszerza. Ten sam wniosek wysunęłam w zadaniu 5., jest tak, ponieważ mniejwięcej w tym miejscu znajduje się średnia danych X, więc s odpowiadająca za szerokość przedziału ufności jest minimalizowana w tym punkcie.
Fioletowymi liniami zaznaczyłam przedział predykcyjny. Widać że jest on wielokrotnie większy od przedziału ufności, powód tego napisałam już wyżej i jest to spowodowane wzorem s oraz s(pred), które odpowiadają za szerokość przedziałów. Przedział predykcyjny również minimalnie zwęża się w okolicach, gdzie X = 5, z tego samego względu co wyżej.

Zadanie 8

W tym zadaniu założę, że $n = 40, \sigma^2 = 70, SSX = \sum(X_i - \bar{X})^2 = 500$

(a) Obliczę moc odrzucenia $H_0 : \beta_1 = 0$, na poziomie istotności $\alpha = 0.05$, jeżeli prawdziwa jest wartość $\beta_1 = 1$.

Aby wyznaczyć funkcję mocy testu musimy obliczyć:

$$\pi(1) = P_{\beta_1=1}(|T| > t_c)$$

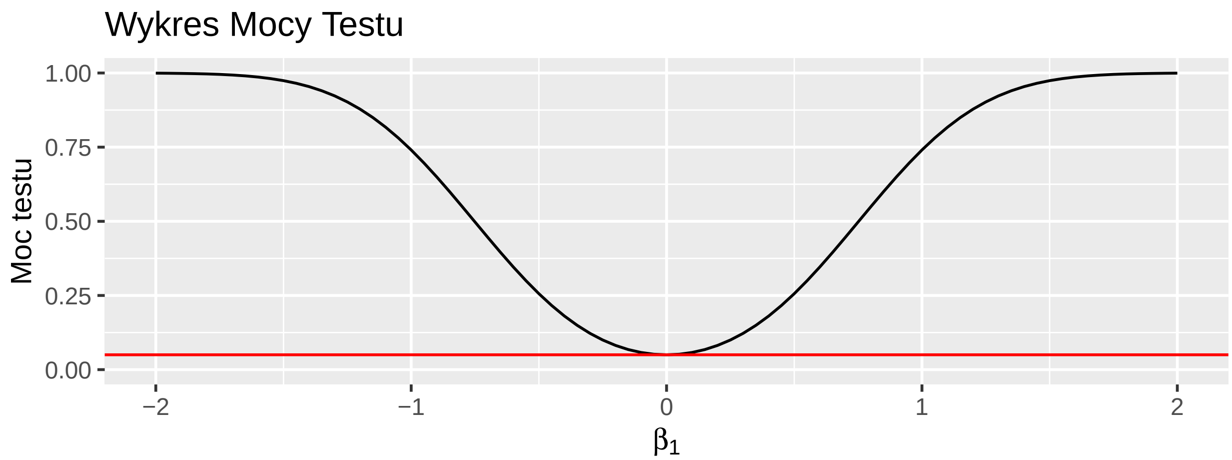
gdzie statystyka T ma niecentralny rozkład studenta z n - 2 stopniami swobody i paramtrem niecentralności $\delta = \frac{\beta_1}{\sigma(\hat{\beta}_1)}$. Wszystkie potrzebne wartości mamy podane w poleceniu, więc policzenie mocy funkcji tego testu w Rze wygląda następująco:

```
n <- 40
sig2 <- 70
ssx <- 500
alpha <- 0.05
sig2b1 <- sig2/ssx
df= n-2
tc <- qt(1-alpha/2,df)
beta1 = 1
delta <- beta1/sqrt(sig2b1)
(power <- 1 - pt(tc,df,delta) + pt(-tc,df,delta))
```

```
## [1] 0.740405
```

(b) Przedstawię na wykresie moc jako funkcję β_1 dla wartości $\beta_1 \in [-2, 2]$.

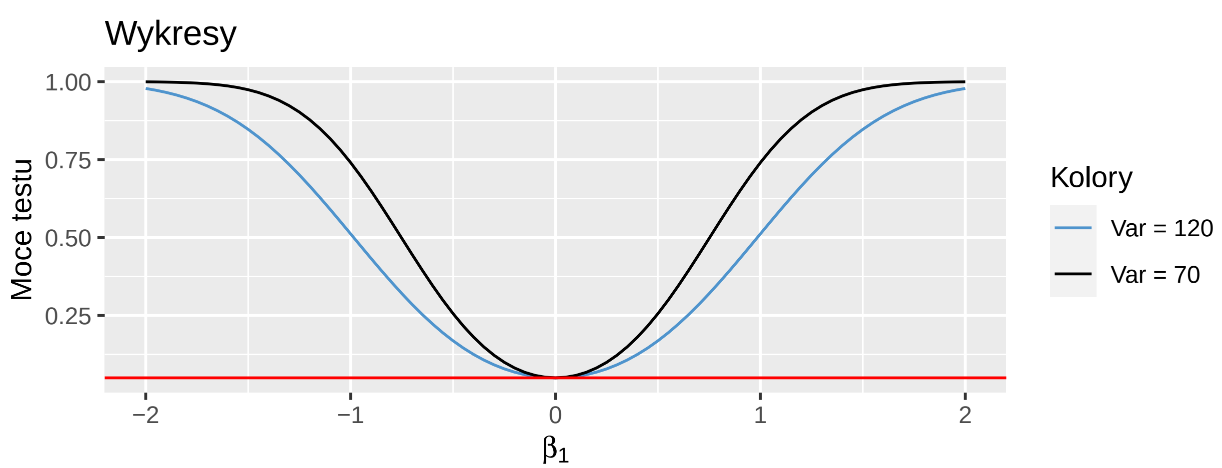
Wykres mocy testu prezentuje się tak:



Powyższy wykres pozwala zobaczyć, jak moc testu zależy od różnych wartości parametru β_1. Widać, że im bardziej oddalamy się od wartości $\beta_1 = 0$, tym moc testu odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej się zwieksza. Linia czerwona na wykresie jest linią $y = \alpha = 0.05$, mówi nam ona tyle, że punkty znajdujące się powyżej tej linii wskazują na sytuacje, w których jesteśmy w stanie odrzucić hipotezę zerową na poziomie istotności 0.05, gdy hipoteza alternatywna jest prawdziwa.

(c) Powtórzę zadanie (b) dla wartości $\sigma^2 = 120$ oraz dodam odpowiedni wykres do wykresu z zadania (b).

Do wykresu z podpuntktu (b) dodam wykres funkcji mocy dla $\sigma^2 = 120$, pozostałe parametry pozostają takie same.



Z wykresów można wyczytać, że większa wariancja zmniejsza moc testu. Jest to zgodne z naszą intuicją, bo przekonaliśmy się już we wcześniejszych zadaniach, że większa wariancja błędu wpływa na precyzję estymacji i na wartości statystyki testowej. Wyższa wariancja prowadzi do mniejszej mocy testu, czyli mniejszą zdolność testu do wykrywania istotnych różnic.

Zadanie 9

Wygeneruję wektor $X = (X_1, ..., X_{200})^T \sim N(0, \frac{1}{500}I)$ oraz wygeneruję 1000 wektorów Y z modelu $Y = 5 + \beta_1 X + \epsilon$, gdzie:

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

(b) $\beta_1 = 0$, $\epsilon_1, ..., \epsilon_{200}$ sa iid z rozkładu wykładniczego z parametrem $\lambda = 1$,

(c) $\beta_1 = 0$, $\epsilon_1, ..., \epsilon_{200}$ sa iid z rozkładu wykładniczego L(0,1),

(d) $\beta_1 = 2$, $\epsilon \sim N(0, I)$,

(e) $\beta_1 = 2$, $\epsilon_1, ..., \epsilon_{200}$ sa iid z rozkładu wykładniczego z parametrem $\lambda = 1$,

(f) $\beta_1 = 2$, $\epsilon_1, ..., \epsilon_{200}$ sa iid z rozkładu logistycznego L(0,1).

Dla każdego powtórzenia eksperymentu przetestuję hipotezę $H_0 : \beta_1 = 0$ i wyestymuję prawdopodobieństwo odrzucenia H_0 na podstawie częstości odrzuceń w próbie. Porównam te estymatory prawdopodobieństwa z teoretycznym prawdopodobieństwem I rodzaju (a, b, c) oraz teoretyczną mocą (d, e, f) obliczoną przy założeniu, że szum ma rozkład normalny.

Podpunkt	Wynik	Teoretyczna moc
a	0.054	0.05
b	0.051	0.05
c	0.052	0.05
d	0.053	0.43
e	0.057	0.43
f	0.051	0.43

Wyniki (kolumna "wynik") pokazują estymowaną częstość odrzuceń hipotezy zerowej, czyli estymator prawdopodobieństwa błędu I rodzaju - estymator α. Widać, że dla każdego z podpunktów wartość wyszła bardzo bliska prawdziwej wartości $\alpha = 0.05$. W drugiej kolumnie policzona została teoretyczna moc, czyli prowdopodobieństwo uniknięcia błędu II rodzaju.

$$\pi(a) = P_{\beta_1=a}(|T| > t_c) = P_{\beta_1=a}(T > t_c) + P_{\beta_1=a}(T < -t_c)$$

Dla podpunktów a, b i c $\beta_1 = 0$, zatem moc testu to tak naprawdę policzenie prawdopodobieństwa popełnienia błędu I rodzaju, czyli $P_{\beta_1=0}(|T| > t_c) = P_{\beta_1=0}(T > t_c) + P_{\beta_1=0}(T < -t_c)$ zakładamy że hipoteza zerowa jest prawdziwa, a liczymi prawdopodobieństwo przyjęcia hipotezy alternatywnej jako prawdziwą. Przyjęty poziom istotności w tym zadaniu, to $\alpha = 0.05$ i tyle właśnie wynosi teoretyczne prawdopodobieństwo błędu I rodzaju dla pierwszych trzech podpunktów. Moc ta jest bardzo niska, możemy więc wnioskować, że te modele mają niską zdolność do wykrywania efektów (Y i X są nieskorelowane).
Dla podopunktów d, e i f teoretyczna moc wynosi 0.43, co jest dużo większym wynikiem, wynika to z przyjętego parametru $\beta_1 = 2$. W tych przypadkach modele są bardziej skuteczne w wykrywaniu efektów i prowadzi to do wyższej teoretycznej mocy.

Zadania teoretyczne

Dla modelu liniowego $Y = \beta_0 + \beta_1 + \epsilon$ na podstawie n = 20 obserwacji uzyskano estymatory: $b_0 = 1, b_1 = 3$ i $s = 4$.

1. s(b_1) = 1, gdzie s(b_1) jest estymatorem odchylenia standardowego b_1. Skonstruuję 95% przedział ufności dla β_1

```
b1 <- 3
sb1 <- 1
n <- 20
tc <- qt(1-0.05/2, n - 2)
round((l_pu <- b1 - tc*sb1),3)
```

```
## [1] 0.899
```

```
round((p_pu <- b1 + tc*sb1),3)
```

```
## [1] 5.101
```

$$\hat{\beta}_1 \pm t_c s(\hat{\beta}_1) = [0.899; 5.101]$$

2. Czy masz statystyczne uzasadnienie dla twierdzenia, że Y zależy od X?

```
T <- b1/sb1
abs(T) > tc
```

```
## [1] TRUE
```

```
#odrzucam H0: beta1 = 0 na poziome istotnosci 0.05, czyli tak, zależy
```

$$T > t_c = t^*(1 - \frac{0.05}{2}, 18)$$

3. 95% przedział ufności dla E(Y), gdy X = 5 wynosi [13,19], znajdę odpowiedni przedział predykcyjny.

```
u_h <- 1 + 3 * 5
#s2u_h = 4^2*(1/20 + d)
#s2(pred) = 4^2 * (1 + 1/20 + d)
#u_h-tc*sqrt(s2u_h) = 13 (jedna niewiadoma -> chcemy wyznaczyc d)
# tc*sqrt(s2u_h) = u_h - 13
#sqrt(s2u_h) = (u_h - 13)/tc
# 4^2 * (1/20 + d) = [(u_h - 13)/tc]^2
# d = [(u_h - 13)/tc]^2/4^2 - 1/20
(u_h - 13)/tc
```

```
## [1] 1.427944
```

```
d <- ((u_h - 13)/tc)^2/4^2 - 1/20
s2pred <- 4^2 * (1 + 1/20 + d)

round(u_h - tc * sqrt(s2pred),3)
```

```
## [1] 7.077
```

```
round(u_h + tc * sqrt(s2pred),3)
```

```
## [1] 24.923
```

$$[\hat{\mu}_h - t_c s(pred), \hat{\mu}_h + t_c s(pred)] = [7.077; 24.923]$$