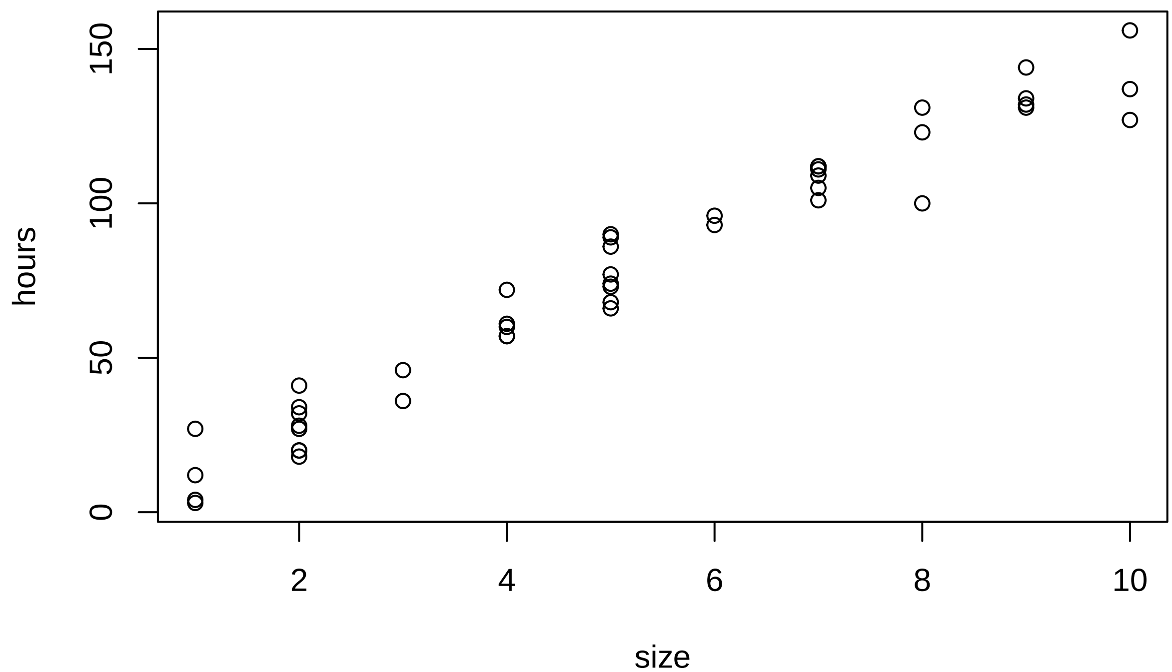
Sprawozdanie 2 Modele liniowe

Katarzyna Stasińska

2023

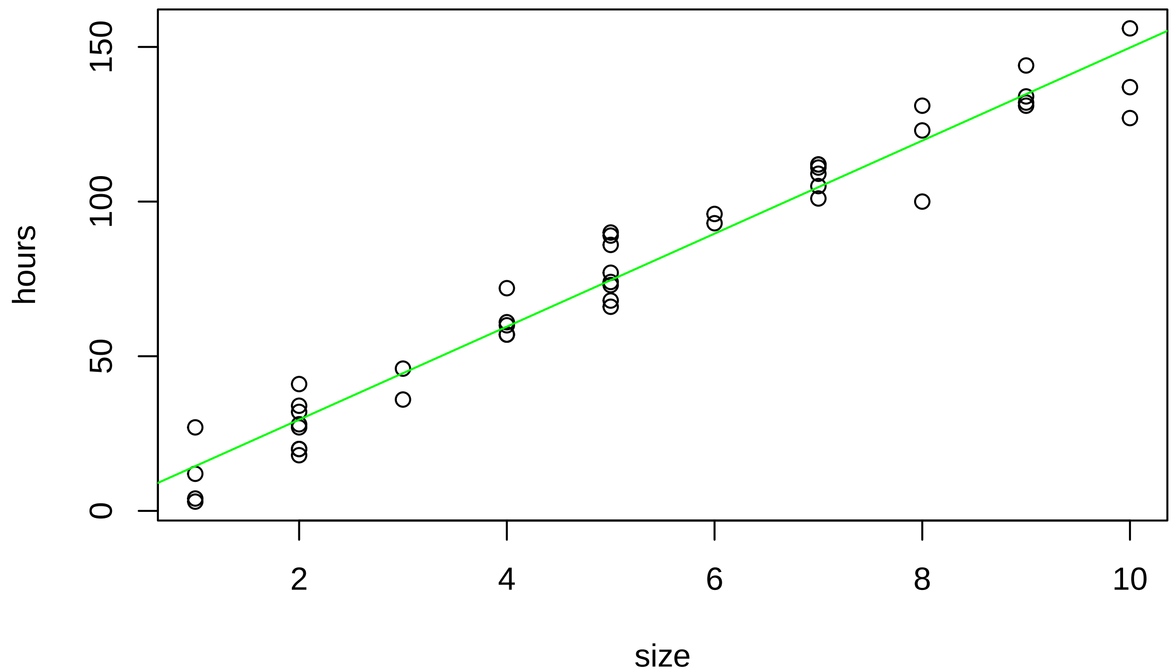
zadanie 1

Patrząc na wykres poniżej można stwierdzić, że zależność jest w przybliżeniu liniowa.



zadanie 2

Estimated regression equation jest opisane wzorem $Y_i = b_0 + b_1 X_i + \epsilon_i$, gdzie b_0 to wyraz wolny (intercept), b_1 to nachylenie (slope), oba są parametrami deterministycznymi, a ϵ_i to zmienna losowa z rozkładu $N(0, \sigma^2)$. Prosta regresji została oznaczona na wykresie kolorem zielonym.




```
## Korzystając z funkcji wbudowanych intercept =  -0.5801567 , a slope =  15.03525
```

Spróbujmy wyliczyć wartości tych parametrów korzystając z wiedzy teoretycznej.

```
slope=sum((dane$size-mean(dane$size))*(dane$hours-mean(dane$hours)))/sum((dane$size-mean(dane$size))^2)
intercept=mean(dane$hours-slope*mean(dane$size))
```

```
## Korzystając ze wzorów z wykładu intercept =  -0.5801567 , a slope =  15.03525
```

zadanie 3

Korzystając z funkcji wbudowanych możemy wyznaczyć przedziały ufności slope'a i intercepta.

```
## Przedziały ufności intercept = ( -6.234843 , 5.074529 ) slope = ( 14.06101 , 16.00949 )
```

Wyliczmy je wykorzystując teoretyczną wiedzę z wykładu.

```
kwantyl=qt(0.975, length(dane$hours)-2)
s2=sum((dane$hours-intercept-slope*dane$size)^2)/(length(dane$hours)-2)
s2b1=s2/sum((dane$size-mean(dane$size))^2)
leftslope=slope-kwantyl*s2b1^(1/2)
rightslope=slope+kwantyl*s2b1^(1/2)
s2b0=s2*(1/length(dane$size)+(mean(dane$size))^2/sum((dane$size-mean(dane$size))^2))
leftintercept=intercept-kwantyl*s2b0^(1/2)
rightintercept=intercept+kwantyl*s2b0^(1/2)
```

```
## Przedziały ufności intercept = ( -6.234843 , 5.074529 ) slope = ( 14.06101 , 16.00949 )
```

zadanie 4

Ustalmy poziom istotności $\alpha = 0,05$

Test istotności slope'a korzystając z poleceń wbudowanych.

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

```
## statystyka testowa: 31.12326 , p-wartosc: 4.009032e-31
```

Możemy zauważyć, że $p < \alpha$, zatem odrzucamy hipotezę H_0, istnieje relacja między X a Y.

Test istotności slope'a korzystając ze wzorów teoretycznych.

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

```
tslope=slope/(s2b1)^(1/2)
pslope=2*(1-pt(abs(tslope),length(dane$size)-2))
```

```
## statystyka testowa: 31.12326 , p-wartosc: 0
```

Możemy zauważyć, że i w tym przypadku $p < \alpha$, choć zgubiliśmy gdzieś dokładność, bo $p = 0$, a nie jest bardzo bliskie zera. Ponownie odrzucamy hipotezę H_0, istnieje relacja między X a Y.

Test istotności intercepta korzystając z poleceń wbudowanych.

Hipotezy: $H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$

```
## statystyka testowa: -0.2069076 , p-wartosc: 0.8370587
```

Zauważmy, że $p > \alpha$, nie możemy odrzucić hipotezy H_0, nie wiemy czy X i Y są wzajemnie proporcjonalne.

Test istotności intercepta korzystając ze wzorów teoretycznych.

Hipotezy: $H_0 : \beta_0 = 0 \quad H_1 : \beta_0 \neq 0$

```
tintercept=intercept/(s2b0)^(1/2)
pintercept=2*(1-pt(abs(tintercept),length(dane$size)-2))
```

```
## statystyka testowa: -0.2069076 , p-wartosc: 0.8370587
```

Wyniki są takie same jak przy użyciu funkcji wbudowanych, stąd i wnioski są takie same, nie możemy odrzucić hipotezy H_0.

Wykorzystuję statystyki testowe z $n - 2$ stopniami swobody, gdzie n to rozmiar danych.

zadanie 5

Niech $E(Y_h) = \mu_h$, zatem estymator wartości oczekiwanej wyraża się wzorem $\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$.

Wyniki eksperymentu przeprowadzonego przy pomocy funkcji wbudowanych.

```
## Wartość estymatora wartości oczekiwanej czasu obsługi: 14.455091383812
## Przedział ufności dla k= 1 [ 9.63613979223049 , 19.2740429753935 ]
## Długość przedziału: 9.63790318316303
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 29.4903394255875
## Przedział ufności dla k= 2 [ 25.444675855645 , 33.5360029955299 ]
## Długość przedziału: 8.09132713988491
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 59.5608355091384
```

```
## Przedział ufności dla k= 4 [ 56.6707779183067 , 62.45089309997 ]
## Długość przedziału: 5.78011518166332
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 74.5960835509138
## Przedział ufności dla k= 5 [ 71.914223607454 , 77.2779434943736 ]
## Długość przedziału: 5.36371988691963
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 89.6313315926893
## Przedział ufności dla k= 6 [ 86.8152007658341 , 92.4474624195444 ]
## Długość przedziału: 5.63226165371032
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 104.666579634465
## Przedział ufności dla k= 7 [ 101.415873078774 , 107.917286190156 ]
## Długość przedziału: 6.50141311138202
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 119.70182767624
## Przedział ufności dla k= 8 [ 115.815714761247 , 123.587940591234 ]
## Długość przedziału: 7.77222582998724
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 164.807571801567
## Przedział ufności dla k= 11 [ 158.47543975806 , 171.139703845073 ]
## Długość przedziału: 12.6642640870123
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 375.301044386423
## Przedział ufności dla k= 25 [ 355.740113625425 , 394.861975147421 ]
## Długość przedziału: 39.1218615219951
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 1502.94464751958
## Przedział ufności dla k= 100 [ 1410.46143622461 , 1595.42785881455 ]
## Długość przedziału: 184.96642258994
```

Wyniki eksperymentu przy użyciu zaimplementowanych wzorów z wykładu.

```r
for (k in c(1, 2, 4, 5, 6, 7, 8, 11, 25, 100)){
  mu=intercept+slope*k
  smu2=s2*(1/length(dane$size)+(k-mean(dane$size))^2/(sum((dane$size-mean(dane$size))^2)))
  left=mu-kwantyl*smu2^(1/2)
  right=mu+kwantyl*smu2^(1/2)
  cat(paste("Wartość estymatora wartości oczekiwanej czasu obsługi:",mu,"\n"))
  cat(paste("Przedział ufności dla k=",k,"[",left,",",right,"]\n"))
  cat(paste("Długość przedziału=",right-left,"\n\n"))
}
```

```
## Wartość estymatora wartości oczekiwanej czasu obsługi: 14.455091383812
## Przedział ufności dla k= 1 [ 9.6361397922305 , 19.2740429753935 ]
## Długość przedziału= 9.63790318316303
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 29.4903394255875
## Przedział ufności dla k= 2 [ 25.444675855645 , 33.5360029955299 ]
## Długość przedziału= 8.09132713988491
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 59.5608355091384
## Przedział ufności dla k= 4 [ 56.6707779183067 , 62.45089309997 ]
## Długość przedziału= 5.78011518166332
##
```

```
## Wartość estymatora wartości oczekiwanej czasu obsługi: 74.5960835509138
## Przedział ufności dla k= 5 [ 71.914223607454 , 77.2779434943737 ]
## Długość przedziału= 5.36371988691963
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 89.6313315926893
## Przedział ufności dla k= 6 [ 86.8152007658341 , 92.4474624195445 ]
## Długość przedziału= 5.63226165371032
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 104.666579634465
## Przedział ufności dla k= 7 [ 101.415873078774 , 107.917286190156 ]
## Długość przedziału= 6.50141311138202
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 119.70182767624
## Przedział ufności dla k= 8 [ 115.815714761247 , 123.587940591234 ]
## Długość przedziału= 7.77222582998724
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 164.807571801567
## Przedział ufności dla k= 11 [ 158.47543975806 , 171.139703845073 ]
## Długość przedziału= 12.6642640870123
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 375.301044386423
## Przedział ufności dla k= 25 [ 355.740113625425 , 394.861975147421 ]
## Długość przedziału= 39.1218615219951
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 1502.94464751958
## Przedział ufności dla k= 100 [ 1410.46143622461 , 1595.42785881455 ]
## Długość przedziału= 184.96642258994
```

Im współrzędna x-owa punktu bliższa średniej wartości współrzędnej x-owej po wszystkich punktach, tym krótszy przedział ufności. Co zgadza się z teorią z wykładu, wpływ na długość przedziału ma $s(\hat{\mu}_h)$, które z kolei zależy od $(X_h - \bar{X})^2$. Średnia wartość współrzędnej x-owej po wszystkich punktach = 5.111111, stąd dla $k = 5$ możemy zaobserwować najkrótszy przedział ufności.

zadanie 6

Wartości przewidywanego czasu obsługi również wyrażają się wzorem $\hat{Y}_h = \hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$.

Wyniki eksperymentu przeprowadzonego przy pomocy funkcji wbudowanych.

```
## Wartość estymatora wartości oczekiwanej czasu obsługi: 14.455091383812
## Przedział ufności dla k= 1 [ -4.15543704559051 , 33.0656198132145 ]
## Długość przedziału: 37.221056858805
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 29.4903394255875
## Przedział ufności dla k= 2 [ 11.0648986520061 , 47.9157801991688 ]
## Długość przedziału: 36.8508815471626
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 59.5608355091384
## Przedział ufności dla k= 4 [ 41.3541909292627 , 77.7674800890141 ]
## Długość przedziału: 36.4132891597514
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 74.5960835509138
## Przedział ufności dla k= 5 [ 56.42132504534 , 92.7708420564876 ]
## Długość przedziału: 36.3495170111476
##
```

```
## Wartość estymatora wartości oczekiwanej czasu obsługi: 89.6313315926893
## Przedział ufności dla k= 6 [ 71.4362754835821 , 107.826387701796 ]
## Długość przedziału: 36.3901122182143
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 104.666579634465
## Przedział ufności dla k= 7 [ 86.3992161933218 , 122.933943075608 ]
## Długość przedziału: 36.5347268822859
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 119.70182767624
## Przedział ufności dla k= 8 [ 101.310760617266 , 138.092894735214 ]
## Długość przedziału: 36.7821341179477
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 164.807571801567
## Przedział ufności dla k= 11 [ 145.749099402237 , 183.866044200896 ]
## Długość przedziału: 38.1169447986583
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 375.301044386423
## Przedział ufności dla k= 25 [ 348.734910188611 , 401.867178584235 ]
## Długość przedziału: 53.1322683956231
##
## Wartość estymatora wartości oczekiwanej czasu obsługi: 1502.94464751958
## Przedział ufności dla k= 100 [ 1408.73066848563 , 1597.15862655353 ]
## Długość przedziału: 188.427958067901
```

Wyniki eksperymentu przy użyciu zaimplementowanych wzorów z wykładu.

```r
for (k in c(1, 2, 4, 5, 6, 7, 8, 11, 25, 100)){
  mu=intercept+slope*k
  spred2=s2*(1+1/length(dane$size)+(k-mean(dane$size))^2/(sum((dane$size-mean(dane$size))^2)))
  left=mu-kwantyl*spred2^(1/2)
  right=mu+kwantyl*spred2^(1/2)
  cat(paste("Przewidywany czas obsługi:",mu,"\n"))
  cat(paste("Przedział ufności dla k=",k,"[",left,",",right,"]\n"))
  cat(paste("Długość przedziału=",right-left,"\n\n"))
}
```

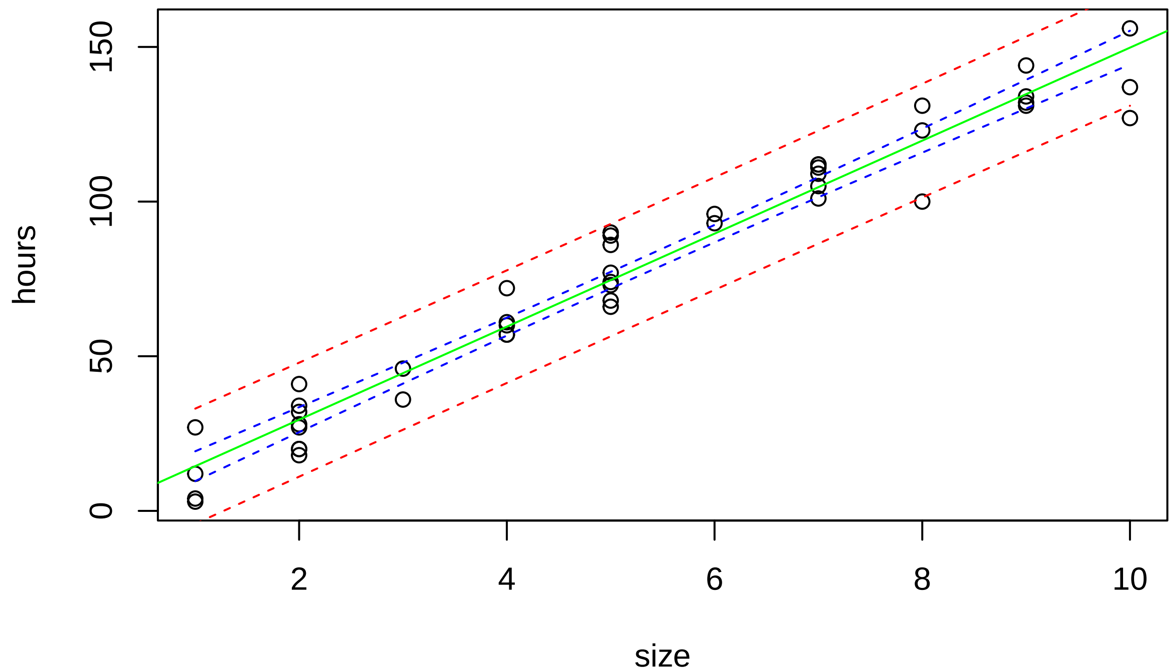
```
## Przewidywany czas obsługi: 14.455091383812
## Przedział ufności dla k= 1 [ -4.15543704559049 , 33.0656198132145 ]
## Długość przedziału= 37.221056858805
##
## Przewidywany czas obsługi: 29.4903394255875
## Przedział ufności dla k= 2 [ 11.0648986520062 , 47.9157801991688 ]
## Długość przedziału= 36.8508815471626
##
## Przewidywany czas obsługi: 59.5608355091384
## Przedział ufności dla k= 4 [ 41.3541909292627 , 77.7674800890141 ]
## Długość przedziału= 36.4132891597514
##
## Przewidywany czas obsługi: 74.5960835509138
## Przedział ufności dla k= 5 [ 56.42132504534 , 92.7708420564877 ]
## Długość przedziału= 36.3495170111476
##
## Przewidywany czas obsługi: 89.6313315926893
## Przedział ufności dla k= 6 [ 71.4362754835822 , 107.826387701796 ]
## Długość przedziału= 36.3901122182143
```

```
##
## Przewidywany czas obsługi: 104.666579634465
## Przedział ufności dla k= 7 [ 86.3992161933218 , 122.933943075608 ]
## Długość przedziału= 36.5347268822859
##
## Przewidywany czas obsługi: 119.70182767624
## Przedział ufności dla k= 8 [ 101.310760617266 , 138.092894735214 ]
## Długość przedziału= 36.7821341179477
##
## Przewidywany czas obsługi: 164.807571801567
## Przedział ufności dla k= 11 [ 145.749099402237 , 183.866044200896 ]
## Długość przedziału= 38.1169447986583
##
## Przewidywany czas obsługi: 375.301044386423
## Przedział ufności dla k= 25 [ 348.734910188611 , 401.867178584235 ]
## Długość przedziału= 53.1322683956231
##
## Przewidywany czas obsługi: 1502.94464751958
## Przedział ufności dla k= 100 [ 1408.73066848563 , 1597.15862655353 ]
## Długość przedziału= 188.427958067901
```

Przedziały predykcyjne są znacząco dłuższe od przedziałów ufności. Zależność między długością przedziałów predykcyjnych jest taka sama, jak zależność między długością przedziałów ufności. Im współrzędna x-owa punktu bliższa średniej wartości współrzędnej x-owej po wszystkich punktach, tym krótszy przedział predykcyjny. Wynika to z tego samego argumentu co w zadaniu 5.

zadanie 7

Na wykresie poniżej, czerwony kolor wyznacza przedziały predykcyjne, a niebieski przedziały ufności.



Przedziały ufności są zawsze mniejsze od przedziałów predykcyjnych, ponieważ wpływ na długość przedziału ma wariancja (im większa tym dłuższy przedział). A wariancja błędu predykcji zmiennej Y jest większa od wariancji estymatora $\hat{\mu}_h$

$Var(Y_h - \hat{\mu}_h) = Var(Y_h) + Var(\hat{\mu}_h) > Var(\hat{\mu}_h)$, bo wariancja jest nieujemna.

zadanie 8

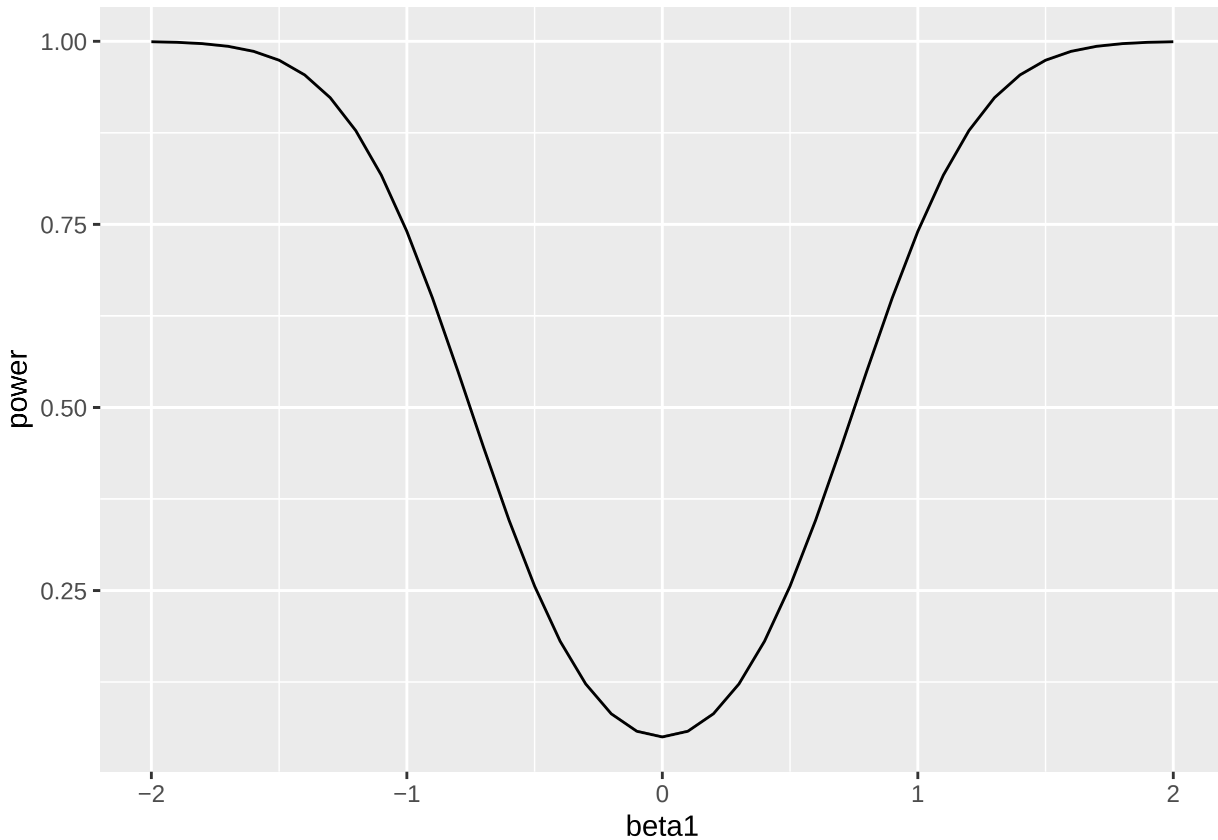
a)

Rozważamy hipotezy: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 = 1$ Wyliczmy wartość $\pi(a) = P_{\beta_1=a}(|T| > t_c)$ dla $a = 1$

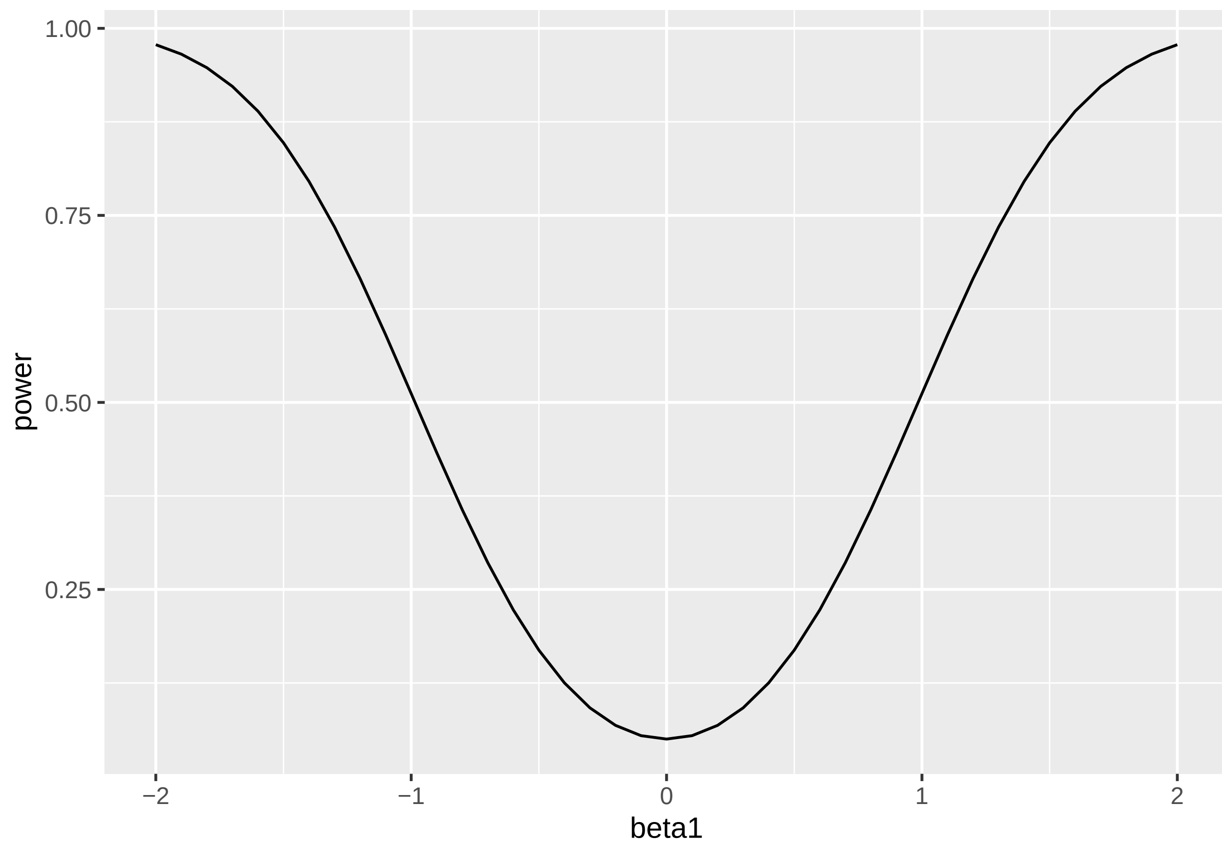
```
beta1 = 1
n = 40
s2 = 70
SSX = 500
s2b1 = s2/SSX
delta = beta1 / sqrt(s2/SSX)
kwantyl=qt(0.975,n-2)
power=pt(-kwantyl, n-2, ncp = delta) + 1 - pt(kwantyl, n-2, ncp = delta)
power
```

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

b)



c)



Zmiana wartości parametru σ^2 na większy wypłaszczyła wykres. Wzrosty i spadki nie są już aż tak gwałtowne.

zadanie 9

```
##             Beta = 0 Beta = 2
## normal         0.051    0.278
## exponential    0.047    0.232
## logistic       0.040    0.110
```

Hipotezę zerową odrzucam, gdy p-value jest mniejsze od $\alpha = 0.05$. Teoretycznie błąd pierwszego rodzaju jest równy parametrowi istotności α. Uzyskane wyniki są bardzo bliskiej tej wartości. Wyliczmy teoretyczne wartości funkcji mocy dla przykładów d), e) i f).

```
## Moc wynosi: 0.2542758
## Moc wynosi: 0.2542758
## Moc wynosi: 0.1103333
```

Możemy zwrócić uwagę, że teoretyczna moc jest bliska wynikom eksperymentu, szczególnie w przypadku f).

Zadania teoretyczne

zadanie 1

```r
cat(kwantyl=qt(0.975, 18))
```

2.100922

Przedziały ufności β_1 wyrażają się wzorem $b_1 \pm t_c s(b_1)$. Podstawiając dane z zadania do wzoru otrzymujemy lewy koniec przedziału $= 3 - t_c = 0.899078$ i prawy koniec przedziału $= 3 + t_c = 5.100922$.

zadanie 2

Rozważmy test istotości $\hat{\beta}_1$.

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

$T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = 3$ Jeśli $|T| > t_c$, gdzie t_c to kwantyl rzędu $1 - \alpha/2$ z 18 stopniami swobody, to odrzucamy hipotezę zerową. Zakładając, że rząd $= 0.975$ otrzymujemy $t_c = 2.100922$. Zatem zachodzi $|T| > t_c$, odrzucamy hipotezę zerową, X i Y są zależne.

zadanie 3

Z treści zadania wiemy, że $\hat{\mu} = 16$, bo jest to środek przedziału ufności, zatem $t_c s(\hat{\mu}) = 3$. Aby wyznaczyć predykcyjny przedział ufności potrzebujemy informacji o $s(pred)$. Zwróćmy uwagę, że

$$s^2(pred) - s^2(\hat{\mu}) = s^2 \tag{1}$$
$$t_c^2 \cdot s^2(pred) - t_c^2 \cdot s^2(\hat{\mu}) = t_c^2 \cdot s^2 \tag{2}$$
$$t_c^2 \cdot s^2(pred) - 9 = t_c^2 \cdot s^2 \tag{3}$$
$$t_c \cdot s(pred) = \sqrt{t_c^2 \cdot s^2 + 9} \tag{4}$$

Wiemy, że $t_c = 2.100922$ oraz $s = 4$, zatem $t_c s(pred) = 8.923115$. Szukany przedział predykcyjny wynosi $[7.076885, 24.92311]$.