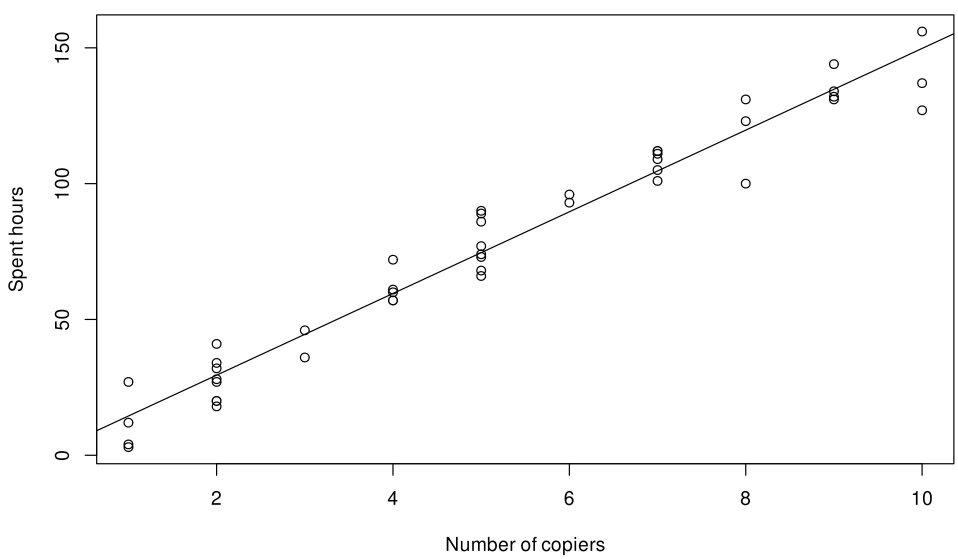
Modele Liniowe - Lista 2

Jakub Kuciński 309881

Listopad 2021

1 Zadanie 1



Rysunek 1: Dane wraz z prostą regresji.

Dane wraz z wyliczoną w R prostą regresji zostały przedstawione na rysunku 1. Widzimy, że dane układają się mniej więcej wzdłuż wyliczonej prostej regresji. Stąd możemy podejrzewać, że relacja między liczbą kserokopiarek a liczbą wymaganych godzin do ich konserwacji jest w przybliżeniu liniowa.

2 Zadanie 2

Szukane wartości wyznaczyłem przy pomocy funkcji *summary* oraz *confint*. Wyestymowane równanie regresji: $Y = 15.0352 \cdot X - 0.5802$. 95% przedział ufności dla nachylenia wynosi $[14.061010, 16.009486]$. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \frac{\bar{\beta}_1 - 0}{s(\bar{\beta}_1)}$, gdzie $s^2(\bar{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ i wynosi 31.123. Pochodzi z rozkładu t-studenta z $45 - 2 = 43$ stopniami swobody. Odpowiadająca p-wartość: $< 2.2e - 16$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $2.2e - 16$. Prawdopodobieństwo to jest bardzo bliskie zeru, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$.

3 Zadanie 3

Oczekiwany średni czas konserwacji dla 11 maszyn wynosi 164.8076h, a odpowiadający 95% przedział ufności to $[158.4754, 171.1397]$.

4 Zadanie 4

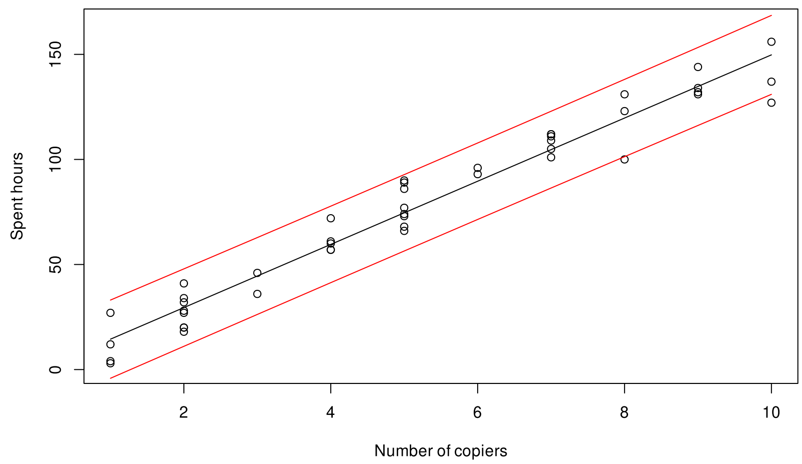
Oczekiwana wartość dla konserwacji 11 maszyn wynosi 164.8076h, a odpowiadający 95% przedział ufności to $[145.7491, 183.866]$. Widzimy, że wartość oczekiwana predykcji punktowej i wartość oczekiwana zmiennej zależnej są takie same. Jest to zgodne z faktem, że predykcja punktowa ma tę samą postać co estymator $E(Y_h)$, czyli $\bar{Y}_h = \bar{\mu}_h = \bar{\beta}_0 + \bar{\beta}_1 X_h$. Z kolei, również zgodnie z oczekiwaniami, przedział predykcyjny jest większy od przedziału ufności średniej.

5 Zadanie 5

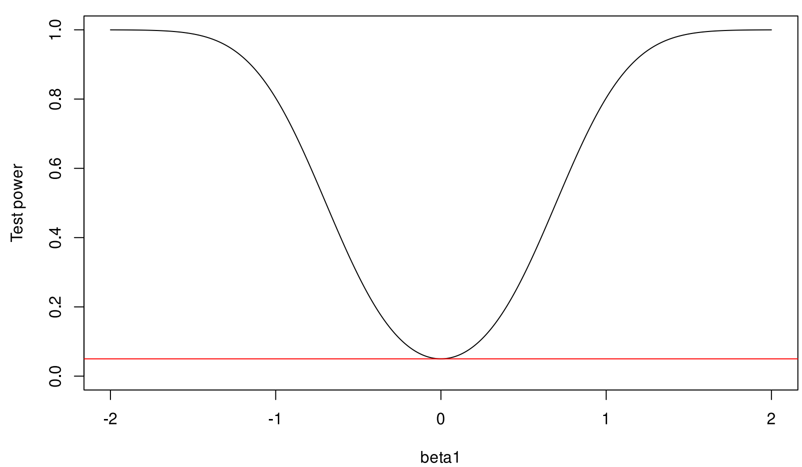
Rysunek 2 przedstawia 95% pasmo predykcyjne dla prostej regresji. Oczekujemy, że 5% obserwacji będzie leżało poza pasmem predykcyjnym. dla naszych 45 obserwacji oznacza to 2.25 punkty. W istocie obserwujemy, że dwie obserwacje (jedna dla $X = 8$ i druga dla $X = 10$) leżą poza pasmem predykcyjnym, a dokładniej poniżej dolnej krawędzi pasma.

6 Zadanie 6

Moc testu binarnego to prawdopodobieństwo odrzucenia H_0, gdy prawdziwa jest hipoteza H_1. Moc testu dla danych założeń oraz $\alpha = 0.05$, $\beta_1 = 1$ wyniosła 0.8032105. Oznacza to, że z prawdopodobieństwem 0.8032105 odrzucona zostanie hipoteza $H_0 : \beta_1 = 0$. Rysunek 3 przedstawia funkcję mocy testu dla β_1 z



Rysunek 2: 95% pasmo predykcyjne dla prostej regresji.



Rysunek 3: Funkcja mocy testu dla β_1 z przedziału $[-2, 2]$.

przedziału $[-2, 2]$. Pozioma czerwona linia oznacza poziom 0.05. Zauważmy, że jest on styczny z funkcją mocy testu w minimum 0. Jest to zgodne z faktem, że przyjęliśmy współczynnik ufności 0.95, więc w 5% przypadków odrzucilibyśmy hipotezę zerową mimo jej prawdziwości.

7 Zadanie 7

Statystyką T testujemy, czy $\bar{\beta_1}$ jest różna od 0: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Statystyka T ma postać: $T = \dfrac{\bar{\beta_1} - 0}{s(\bar{\beta_1})}$, gdzie $s^2(\bar{\beta_1}) = \dfrac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$. 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. Parametr α został ustalony na poziomie 0.05

W przypadkach a) i b), czyli gdy testowana hipoteza zerowa ($H_0 : \beta_1 = 0$) była zgodna ze sposobem generowania wektorów Y (ustaliliśmy $B_1 = 0$ w relacji $Y = 5 + \beta_1 X + \epsilon$), to hipoteza zerowa została odrzucona odpowiednio 47 i 59 razy z 1000 testów. Obserwowane częstotliwości są zgodne z parametrem *alpha*, który ustaliliśmy na poziomie 0.05, czyli oczekiwaliśmy około 50 odrzuceń.

W przypadkach c) i d), czyli gdy testowana hipoteza zerowa nie była zgodna ze sposobem generowania wektorów Y (ustaliliśmy $B_1 = 1.5$ w relacji $Y = 5 + \beta_1 X + \epsilon$), to hipoteza zerowa została odrzucona odpowiednio 342 i 350 razy z 1000 testów. Obserwowane częstotliwości są zgodne z wyliczoną teoretyczną wartością mocy równą 0.3476184, która oznacza, z jakim prawdopodobieństwem odrzucona zostanie hipoteza zerowa $H_0 : \beta_1 = 0$.

8 Kod w R

```
    data = read.table("http://www.math.uni.wroc.pl/~mkos/Modele_liniowe/CH01PR20.txt", col.r

##### 1
plot(hours~copiers, data, ylab="Spent hours", xlab="Number of copiers")
abline(lm(hours~copiers, data))

##### 2
reg1 = lm(hours~copiers, data)
summary(reg1)
confint(reg1, level=0.95)
# plot(reg1)

##### 3
newdata = data.frame(copiers=c(11))
predict.lm(reg1, newdata, interval = "confidence", level=0.95)

##### 4
```

```
predict.lm(reg1, newdata, interval = "prediction", level=0.95)

##### 5
v1 = predict(reg1, interval = "prediction")
v1down = sort(v1[,2])
v1up = sort(v1[,3])
# v2 = predict(reg1, interval = "confidence")
# v2down = sort(v2[,2])
# v2up = sort(v2[,3])
m1 = sort(v1[,1])
copiers1 = sort(data$copiers)
hours1 = sort(data$hours)
plot(copiers1, m1, type='l', ylim=c(0,165),  ylab="Spent hours", xlab="Number of copiers")
points(data$copiers, data$hours)
points(copiers1, v1up, type='l', col='red')
points(copiers1, v1down, type='l', col='red')
dim(data)

##### 6
power = function(n, sig2, ssx, alpha, beta1){
  sig2b1 = sig2 / ssx
  dOfF = n - 2
  tc = qt(1-alpha/2, dOfF)
  delta = beta1 / sqrt(sig2b1)
  prob1 = function(delta){pt(tc, dOfF, delta)}
  prob2 = function(delta){pt(-tc, dOfF, delta)}
  powerOfBeta = 1 - prob1(delta) + prob2(delta)
  powerOfBeta
}
# a)
n = 40
sig2 = 120
ssx = 1000
alpha = 0.05
beta1 = 1
power(n, sig2, ssx, alpha, beta1)
# b)
beta1 = seq(from=-2.0, to=2.0, by=0.01)
plotData = power(n, sig2, ssx, alpha, beta1)
plot(beta1, plotData, type='l', ylim=c(0,1), ylab="Test power", xlab="beta1")
abline(h=0.05, col="red")

##### 7
n = 200
X = rnorm(n, 0, sqrt(1/n))
alpha = 0.05
```

```
sig2 = 1
ssx = sum((X - mean(X))^2)

# a)
beta1 = 0
rejected = 0
y_num = 1000 #1000
for (i in 1:y_num){
  Y = 5 + beta1* X + rnorm(n, 0, 1)
  reg1 = lm(Y~X)
  b0 = coef(reg1)[1]
  b1 = coef(reg1)[2]
  s2 = 1/(n-2) * sum((Y - b0 - b1 * X)^2)
  s2b = s2 / sum((X - mean(X))^2)
  t = (b1 - 0) / sqrt(s2b)
  tc = qt(1 - alpha/2, n - 2)
  if (abs(t) > tc) {
    rejected = rejected + 1
  }
}
rejected
# powinno wyjść około 5% z 1000 czylo 50 - ok

# b)
beta1 = 0
rejected = 0
y_num = 1000 #1000
for (i in 1:y_num){
  Y = 5 + beta1* X + rexp(n, 1)
  reg1 = lm(Y~X)
  b0 = coef(reg1)[1]
  b1 = coef(reg1)[2]
  s2 = 1/(n-2) * sum((Y - b0 - b1 * X)^2)
  s2b = s2 / sum((X - mean(X))^2)
  t = (b1 - 0) / sqrt(s2b)
  tc = qt(1 - alpha/2, n - 2)
  if (abs(t) > tc) {
    rejected = rejected + 1
  }
}
rejected
# powinno wyjść około 5% z 1000 czylo 50 - ok

# c)
beta1 = 1.5
rejected = 0
```

```
rejected2 = 0
rejected3 = 0
y_num = 1000 #1000
for (i in 1:y_num){
  Y = 5 + beta1* X + rnorm(n, 0, 1)
  reg1 = lm(Y~X)
  b0 = coef(reg1)[1]
  b1 = coef(reg1)[2]
  s2 = 1/(n-2) * sum((Y - b0 - b1 * X)^2)
  s2b = s2 / sum((X - mean(X))^2)
  t = (b1 - 0) / sqrt(s2b)
  tc = qt(1 - alpha/2, n - 2)
  if (abs(t) > tc) {
    rejected = rejected + 1
  }
}
rejected
power(n, sig2, ssx, alpha, beta1)

# d)
beta1 = 1.5
rejected = 0
y_num = 1000 #1000
for (i in 1:y_num){
  Y = 5 + beta1* X + rexp(n, 1)
  reg1 = lm(Y~X)
  b0 = coef(reg1)[1]
  b1 = coef(reg1)[2]
  s2 = 1/(n-2) * sum((Y - b0 - b1 * X)^2)
  s2b = s2 / sum((X - mean(X))^2)
  t = (b1 - 0) / sqrt(s2b)
  tc = qt(1 - alpha/2, n - 2)
  if (abs(t) > tc) {
    rejected = rejected + 1
  }
}
rejected
power(n, sig2, ssx, alpha, beta1)
```