Tællende Aktivitet 2

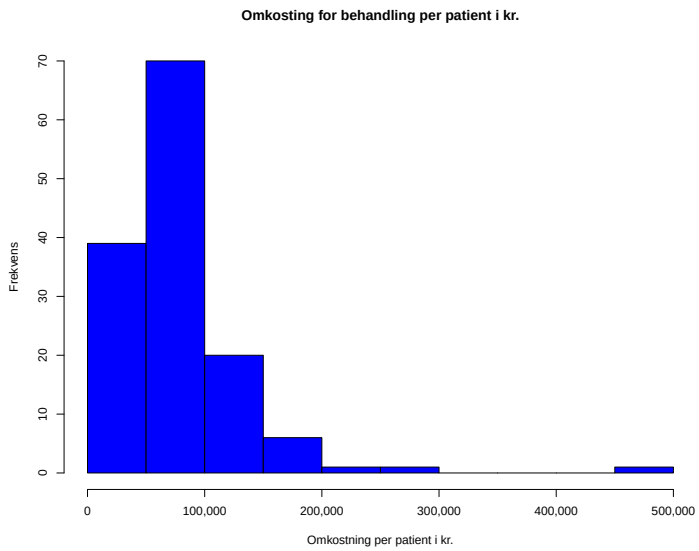
Peter Heilbo Ratgen
perat17

22. november 2020

1 Opgave 1

1.1 Deskriptiv analyse

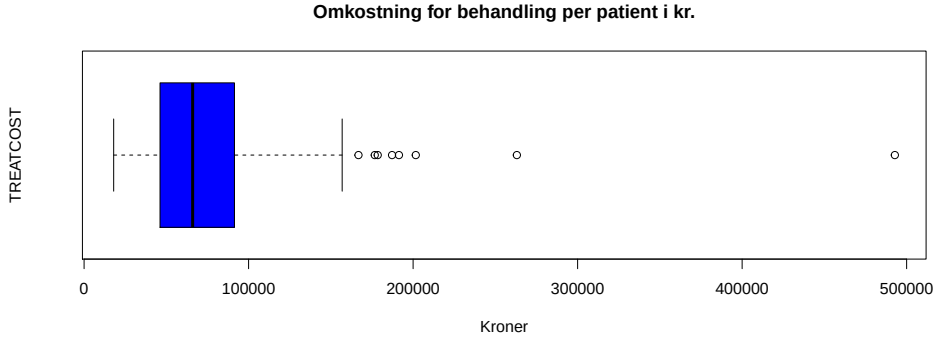
En deskriptiv analyse af udgifterne til behandling af patienter TREATCOST



Figur 1: Histogram af omkostningen for at behandle patienter

Gennemsnit	78754
Median	65992
Typeinterval (modus)	50000 - 10000
Standardafvigelse	55608.07
Standard error	4733.673
Minimum	17989
Maximum	492833
Q_1	46202
Q_3	91328

Ved at se på histogram og median samt gennemsnit ses at formen er højreskæv. Formen er på histogrammet er ikke symmetrisk. Hvis man betrager nedenstående boxplot ses det at der er en ekstrem observation. Der falder langt uden for andre observationer. Selvom dette er en outlier er dette datapunkt stadig vigtig at have med.



Figur 2: Boxplot af omkostning for at behandle patienter

1.2 Konfidensinterval for middelværdien

Vi opstiller et konfidensinterval på 95%. Dette betyder at 95% af data vil falde inden for dette interval. Vi udregner konfidensintervallet ved, at finde gennemsnittet og standardafvigelsen og vi anvender qnorm til at finde Z -scoren for den givne andel af observationsættet. Ud fra denne værdi kan konfidensintervallet ved at fratrække fejlmargenen fra gennemsnitsværdien.

Da ender vi med et konfidensinterval [69476.61 : 88032.26], her ligger 95% fra værdierne.

1.3 Sandsynlighed for samlede udgifter

Sandsynligheden for at en behandling kommer til at koste over 95.000 kr. er givet ved

$$\frac{\text{antal behandlinger over 95000}}{\text{antal behandlinger i alt}} = \frac{34}{138} = 0.2463768.$$

1.4 Sandsynlighed for samlede udgifter mellem 40000 og 65000

Sandsynlighed for at en behandling kommer til at koste mellem 40,000 og 65,000 er givet ved:

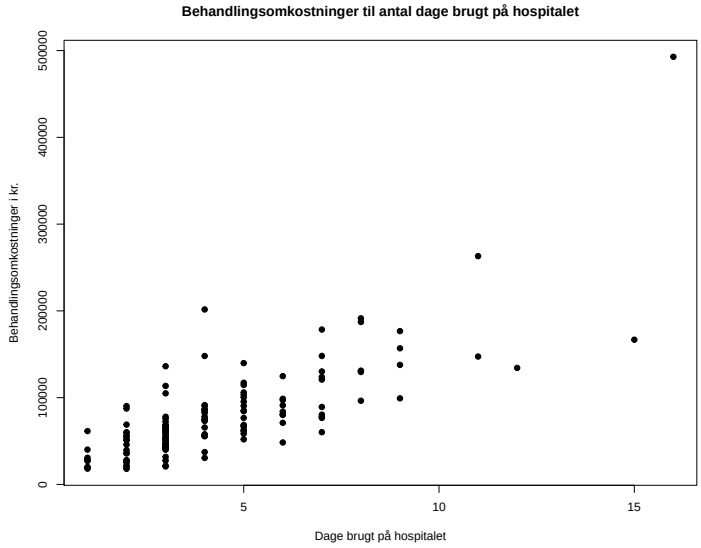
$$1 - \frac{\text{antal behandlinger under 40000} + \text{antal behandlinger over 65000}}{\text{antal behandlinger}}.$$

Da får vi:

$$1 - \frac{21 + 70}{138} = 0.3405797.$$

2 Opgave 2

Vi stater med at lave et scatterplot for at få et overblik over data:

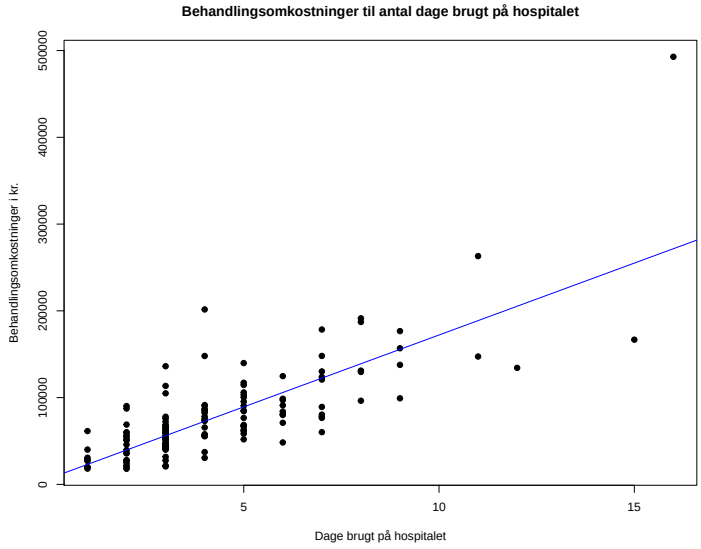


Figur 3: Scatterplot over dage på hospitalet og dennes sammenhæng med behandlindsomkostninger

Vi laver en regression på formen

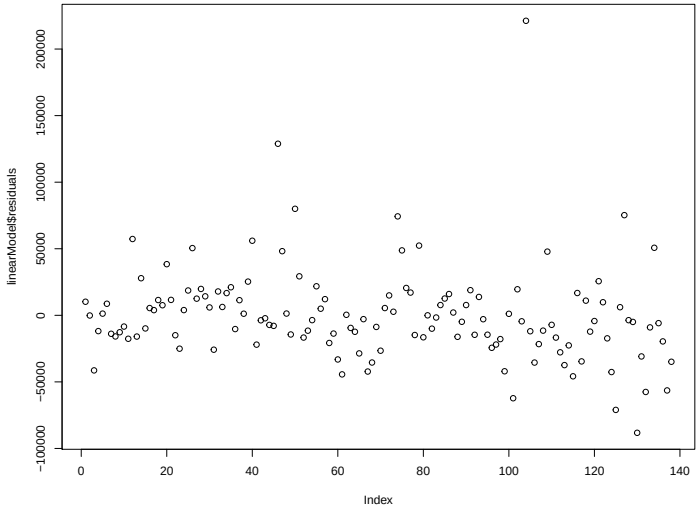
$$TREATCOST_i = \beta_0 + \beta_1 CAREDAYS_i + \epsilon_i \quad i = 1, \ldots, 138.$$

Dette giver denne regressionslinjre



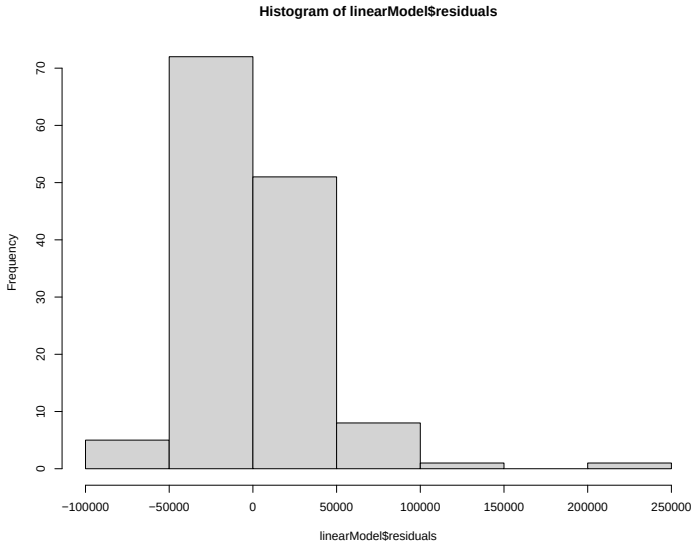
Figur 4: Scatterplot over dage på hospitalet og dennes sammenhæng med behandlindsomkostninger

Vi kigger på residulerne for regressionslinjen:



Figur 5: caption

Hvis vi kigger på residualerne de betegnes som tilfældige, der er dog en ekstrem observation som også set tidligere.

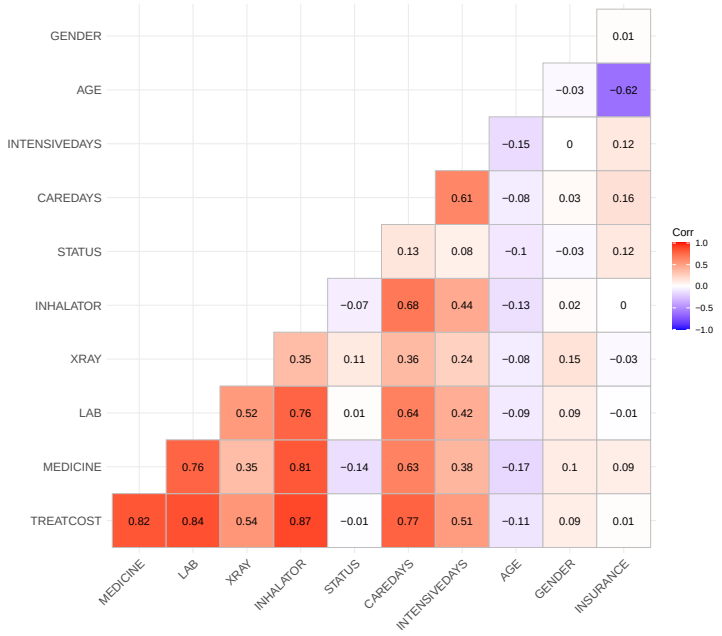


Figur 6: caption

Normalfordelingsplottet er ikke symmetrisk så det er ikke nødvendigvis normalfordelt, og vi ser også her den ekstreme observation.

3 Opgave 3

Vi opstiller en er en korrelationsmatrix. Denne viser at varibler som alder, køn og forsikring ikke har en stor indvirkning.



Figur 7: caption

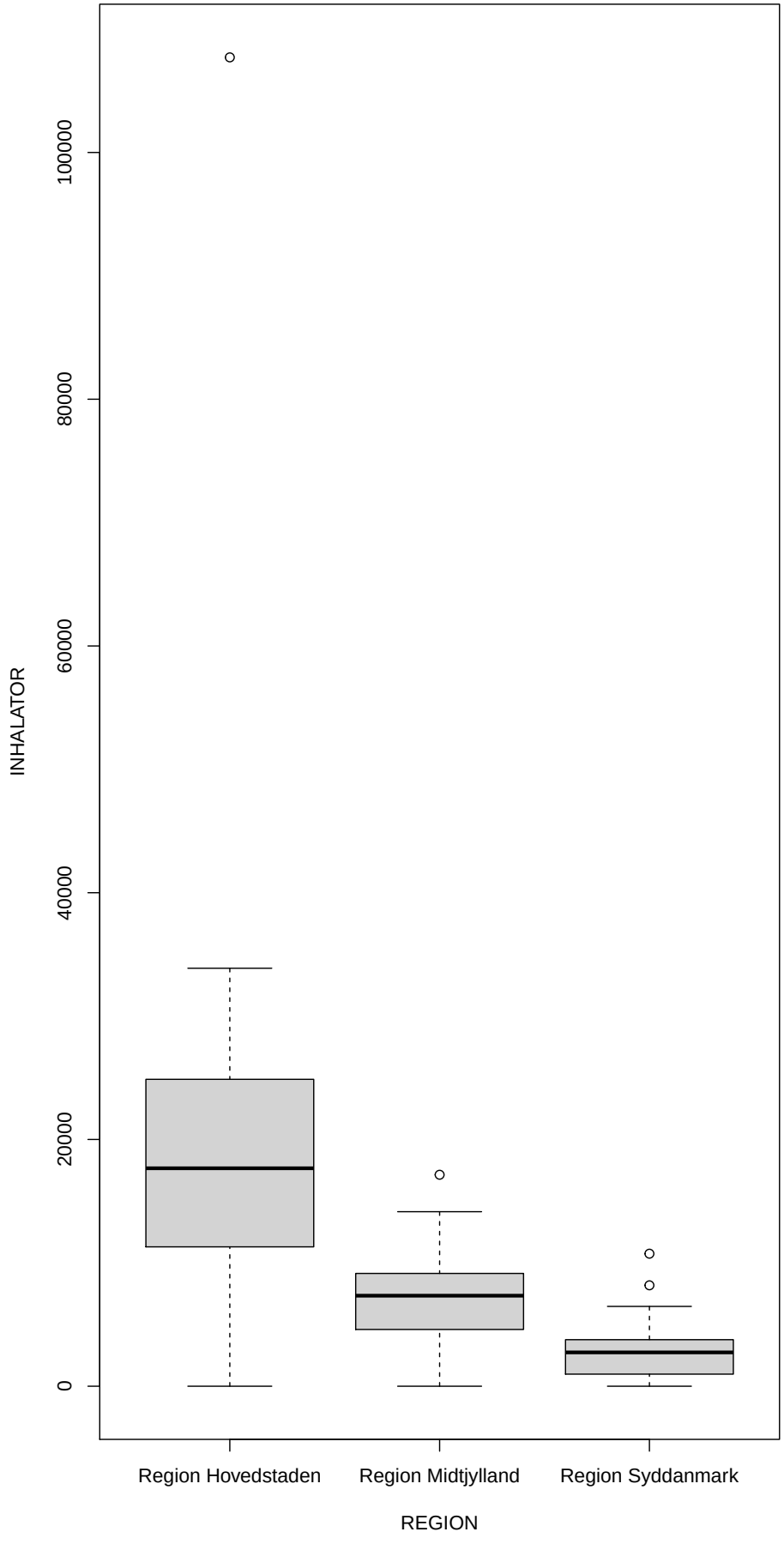
Vi skal elimiere de varibler der ikke er ikke er vigtige for modellen, da den mest komplekse model (den fulde model) ikke altid er den bedste. Til at eliminere variabler og maksimere R^2_{adj} elimineres de variabler der ved eliminering giver den højeste R^2_{adj}.

Efter at have elimineret status, age og gender ender vi med en model der ser således ud: MEDICINE+LAB+XRAY+INHALATOR+CAREDAYS+INTENSIVEDAYS+INSURANCE. Dette giver en forklaringsværdi på 0.888.

4 Opgave 4

Vi har en hypotese om at regionale forskellige kan have en indflydelse på brugen af respirator, og de udgifter der følger med. Vi har en null hypotese H_0 og en alternativ hypotese H_A. Hvor null hypotesen udgør et konservativt synspunkt og den alternative hypotese udgør noget nyere og mere interessant.
H_0: Der er ingen regionale forskelle i brugen af respirator.
H_A: Der er regionale forskelle i brugen af respirator.

Vi skal også bestemme et passende niveau for signifikans α. Hvis vælger et lavere α fx 0.01 er vi mindre tilbøjelige til at lave type 1 fejl (hvor H_0 fejlagtigt afvises til fordel for H_A), hvis vi vælger et højere α fx 0.10 er vi mindre tilbøjelige til at lave type 2 fejl (hvor H_0 fejlagtigt ikke afvises). I forhold til vores hypotese betyder en type 1-fejl at der fejlagtigt konstateres at der er regionale forskelle i brugen af respirator.



Figur 8: caption

Resultatet af ANOVA viser signifikans.

5 Opgave 5

Vi har en hypotese om at bestemte udgifter kan være relateret til køn, vi skal undersøge om middeludgifterne til røntgenundersøgelser er højere for mandlige end kvindelig patienter.

6 R-kode

```
ttfamily
record <- read.csv("./data.csv", sep = ";", header = TRUE)

options(scipen = 10)

# Histogram for TREATCOST
h <- hist(record$TREATCOST)
pdf("plots/treatcost.pdf", width = 10, height = 8);
plot(h, main = "Omkosting for behandling per patient i kr.",
     xlab = "Omkostning per patient i kr.",
      ylab = "Frekvens", col = "blue", xaxt = "n")
axis(side = 1, at = axTicks(1),
    labels = formatC(axTicks(1), format = "d", big.mark = ","))
dev.off()

# Opsummering af data
summary(record[2])
sd(unlist(record[2]))
var(unlist(record[2]))
se <- sd(unlist(record[2]))/sqrt(length(unlist(record[2])))
print(se)

boxplot()

pdf("plots/costboxplot.pdf", width = 10, height = 4)
boxplot(record[2], main = "Omkostning for behandling per patient i kr."
        , col = "blue", horizontal = TRUE
        , xlab = "Kroner"
        , ylab = "TREATCOST", boxwex = 1.5)
dev.off()
treatMean = mean(unlist(record[2]))
treatSd = sd(unlist(record[2]))

error = qnorm(0.975)*treatSd/sqrt(length(unlist(record[2])))

#venstre
treatMean - error
#højre
treatMean+error

underOver <- (sum(record$TREATCOST < 40000) + sum(record$TREATCOST > 65000))/length(recor

1 - underOver

#overblik over data
pdf("plots/caredaysScatterPlot.pdf", width = 10, height = 8)
plot(record$CAREDAYS, record$TREATCOST, pch = 19, xlab = "Dage brugt på hospitalet", yla
dev.off()
```

```
pdf("plots/caredaysScatterPlotAbline.pdf", width = 10, height = 8)
plot(record$CAREDAYS, record$TREATCOST, pch = 19, xlab = "Dage brugt på hospitalet", yla
abline(lm(record$TREATCOST~record$CAREDAYS), col = "blue")
dev.off()

plot(record$CAREDAYS, record$TREATCOST)

linearModel <-lm(record$TREATCOST~record$CAREDAYS)
summary(linearModel)

pdf("plots/caredaysResiduals.pdf", width = 10, height = 8)
plot(linearModel$residuals)
dev.off()


pdf("plots/caredaysResidualsHist.pdf", width = 10, height = 8)
hist(linearModel$residuals)
dev.off()


library(ggcorrplot)

corretlationMatrix <- cor(record[,c(2:12)])
corretlationMatrix = round(corretlationMatrix, 2)

pdf("plots/corretlationMatrix.pdf", width = 10, height = 10)
ggcorrplot(corretlationMatrix, type = "lower",  lab = TRUE)
dev.off()

pdf("./plots/elim1.pdf", width = 10, height = 10)
# fulde model R2_adj 0.8859
elim1 <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+STATUS+CAREDAYS+INTENSIVEDAYS+AGE+GENI
plot(elim1$residuals)
dev.off()
summary(elim1)

elim2 <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+CAREDAYS+INTENSIVEDAYS+AGE+GENDER+INSU
plot(elim2$residuals)
summary(elim2)
# status 0.8867

elim3 <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+CAREDAYS+INTENSIVEDAYS+GENDER+INSURANC
plot(elim3$residuals)
summary(elim3)
# AGE 0.8874

pdf("./plots/elim4.pdf", width = 10, height = 10)
elim4 <- lm(TREATCOST~MEDICINE+LAB+XRAY+INHALATOR+CAREDAYS+INTENSIVEDAYS+INSURANCE, data
plot(elim4$residuals)
dev.off()
summary(elim4)
# GENDER 0.888


inregion <- record[,c(6,13)]
print(inregion)
```

```
region1 <- subset(inregion, REGION == 1)
region2 <- subset(inregion, REGION == 2)
region3 <- subset(inregion, REGION == 3)


pdf("plots/anovaboxplot.pdf", width = 7, height = 14)
boxplot(INHALATOR~REGION, data=record,
  height = 12,
  names = c("Region Hovedstaden", "Region Midtjylland", "Region Syddanmark"))
dev.off()

library(dplyr)
group_by(record["INHALATOR"], record["REGION"])%>%
  summarize(
    count = n(),
    mean = mean(INHALATOR),
    sd = sd(INHALATOR),
    var = var(INHALATOR)
  )

Anova <- aov(INHALATOR~REGION, data = record)
summary(Anova)
plot(Anova$residuals)
```