STK1000: Løsningsforslag Uke 37

2022

Oppgave 2.2, Section 2.1 Exercises

a) Trolig er prisen på et vaskemiddel delvis avhengig av hvor gode test-score produktet har fått, og i så fall kan vi si at testscore er en forklaringsvariabel og pris er en respons. Du kunne like gjerne vurdert at det ikke er en sammenheng, og da er det riktignok likegyldig hvilken variabel du velger eller finner naturlig å tolke som forklaringsvariabel og respons.

b) I en studie der de ønsker å undersøke ukedagsrytmer og studievaner for statistikk, studenter, kan vi sette 'ukedag' som forklaringsvariabel og 'mengde tid brukt på å jobbe med faget' som responsvariabel.

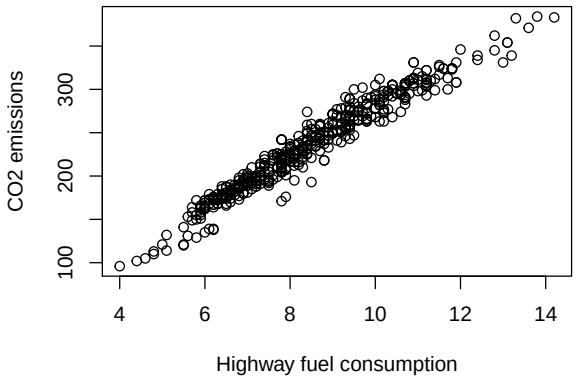
c) I denne studien ønsker de trolig å se hvordan det å være i en viss aldersgruppe påvirker om barnet får i seg nok kalsium. I så fall, er aldersgruppe forklaringsvariabelen og 'får i seg nok kalsium' respons.

d) Antall enheter med alkohol konsumert (i et tidsrom) kan være med på å forklare alkoholprosenten i blodet til en person. Vi har derfor at enheter med alkohol er en forklaringsvariabel og alkoholprosent i blodet er en respons.

Oppgave 2.11(R), Section 2.2 Exercises

a)

```
data = read.csv('../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-011CANFREG.csv')
plot(data$FuelConsHwy, data$CO2, xlab = 'Highway fuel consumption', ylab = 'CO2 emissions')
```



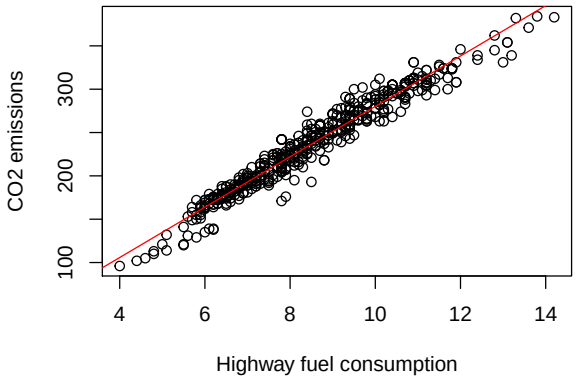
b) Vi ser det er en lineær trend (form) i positiv retning. Styrken (*strength*) er høy da det er et veldig tydelig mønster.

c) Her ser det ut som om alle datapunktene følger samme mønster, og det er ingen store avvik i x-verdier. Med andre ord, ingen uteliggere.

Oppgave 2.12(R)

a) En rett linje vil oppsummere forholde mellom drivstofforbruk og CO2-utslipp. Den vil også hjelpe på å evaluere styrken (*strength*). Vi ser at assosiasjonen er veldig lineær, og den rettlinja modellen oppsummererer sammenhengen godt.

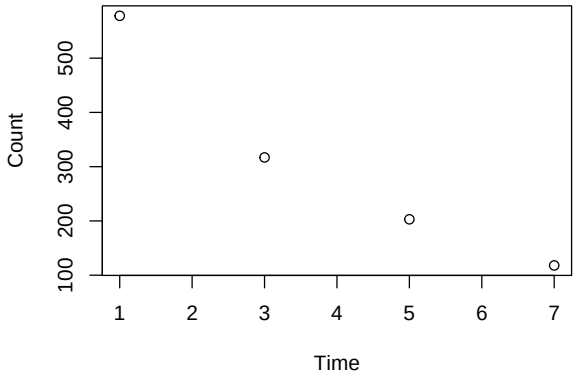
```
plot(data$FuelConsHwy, data$CO2, xlab = 'Highway fuel consumption', ylab = 'CO2 emissions')
abline(lm(data$CO2~data$FuelConsHwy), col = "red")
```



b) Ettersom assosiasjonen er veldig lineær vil ikke en *smoother* skille seg stor fra en rett linje.

Oppgave 2.22 (R)

a) Tiden er den naturlige forklaringsvariabelen, og vi velger derfor denne til x-aksen. Det følger at vi velger 'antall' på y-aksen ettersom vi forventer at tiden forklarer antallet (tid er forklaringsvariabel og antall er respons).



b) Antallet synker med tiden og det er ingen observasjoner som er særlig overraskende.

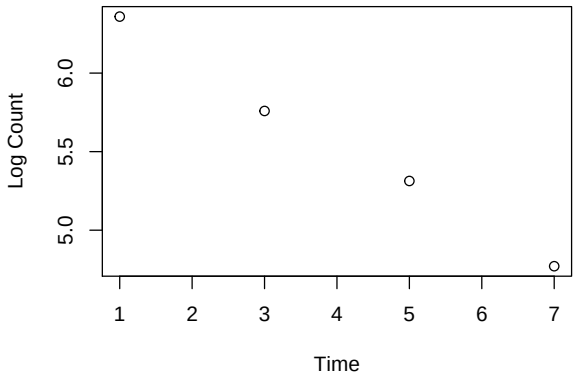
c) Formen har en svak positiv kurvatur og styrken ser ut til å være høy (tydelig mønster).

d) Det er tildels kunstig å snakke om uteliggere i et så lite eksperiment. uansett, det er ingen uteliggere her.

e) Forholdet er ikke helt lineært, men har en positiv kurvatur. (**Ekstra:** Radioaktivitet synker typisk eksponensielt, som kan samsvare med trenden som antydes i plottet.)

Oppgave 2.33 fra 9.de utgave av læreboka (R): 'Repeter stegene i ukesoppgaven over (2.22) med en log-transformasjon av antallene som responsvariabel'

a) Vi log-transformerer responsen (antall) og gjentar alt fra Oppgave 2.32:



b) Antallet synker med tiden og det er ingen observasjoner som er uventet.

c) Formen er lineær og styrken er høy.

d) Det er ingen uteliggere her.

e) Forholdet er veldig lineært. Dette antyder at kurven vi så i Oppgave 2.22 var eksponensielt synkende.

Oppgave 2.32(R), Section 2.3 Exercises

Vi kan finne korrelasjonen med `cor` i R:

```
data = read.csv('../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-011CANFREG.csv')
cor(data$FuelConsHwy, data$CO2)
```

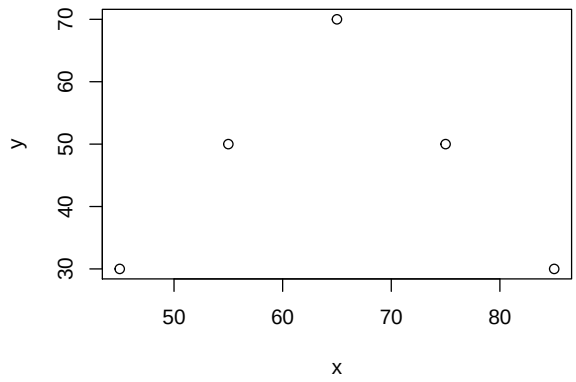
```
## [1] 0.9746135
```

Vi ser at det er høy korrelasjon (nært 1) som tilsier at CO2-utslipp kan i stor grad forklares av drivstofforbruk. Dette samsvarer bra med plottet i Oppgave 2.21.

Oppgave 2.34(R)

a)

```
x = c(45, 55, 65, 75, 85)
y = c(30, 50, 70, 50, 30)
plot(x, y)
```



b) Vi ser at det er et klart forhold mellom x og y, men det er ikke lineært.

c) Vi regner ut korrelasjonen og ser at den er 0.

```
cor(x, y)
```

```
## [1] 0
```

d) Poenget med denne oppgaven er å understreke at korrelasjon beskriver *lineære* sammenhenger mellom variabler, og at korrelasjon ikke er et mål på avhengighet eller styrke (*strength*) uavhengig av type mønster.

Oppgave 2.38 (Løs for hånd)

a) Vi bruker formelen for korrelasjon på side 92 i boka (Chaper 2.3). Først estimerer vi standardavvikene med formelen

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2},$$

og får da $s_x = 2.58$ og $s_y = 200.03$. Vi kan så regne ut korrelasjonen

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right),$$

som gir r = -0.964.

b) Ettersom vi konkluderte med at dataene ikke er lineære er det litt problematisk å bruke korrelasjonen som et oppsummeringsmål her. Likevel ser vi at kurvaturen er så svak at vi fremdels får en høy korrelasjon.

Oppgave 2.39 (for hånd eller i R)

```
x = c(1, 3, 5, 7)
y = c(578, 317, 203, 118)
cor(x, log(y))
```

```
## [1] -0.9985253
```

a) Vi kan beregne korrelasjonen på samme måte som i Oppgave 2.51, men vi må først log-transformerer antallet (*Counts*). Vi får da en korrelasjon på -0.999 ≈ -1.

b) I denne oppgaven er korrelasjonen et bra oppsummeringsmål, ettersom dataene har et lineært forhold.

c) Hvis vi sammenligner med Oppgave 2.38 ser vi at begge har en korrelasjon som er nær -1. Dette tyder på en sterk sammenheng mellom dataene. Men ettersom dataene i Oppgave 2.38 ikke er lineære, er ikke korrelasjonen et like godt oppsummeringsmål for de u-transformerte dataene

Oppgave 2.50(R), Section 2.4 Exercises

a)

Vi kan tilpasse en rett linje til datasettet med funksjonen `ln`. Her er `FuelConsHwy` forklaringsvariabelen x og `CO2` responsvariabelen y. Bruk `help(ln)` for å se dokumentasjonen til `ln`.

```
data = read.csv('../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-050CANFREG.csv')
y = data$CO2
x = data$FuelConsHwy
model = lm(y ~ x)
model
```

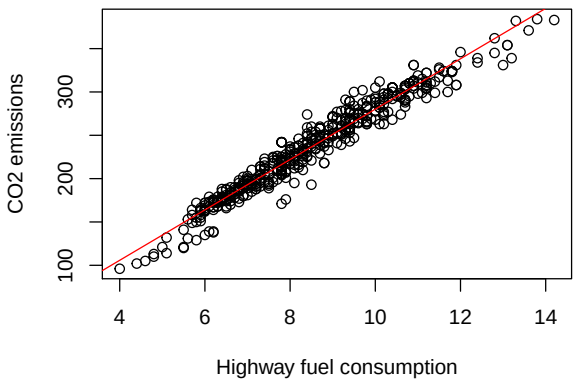
```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      -10.49        29.07
```

Altså får vi ligningen $\hat{y} = \text{-}10.494 + 29.066\ x$.

b)

Vi kan bruke `abline` for å plotte den tilpassede linjen.

```
plot(x, y, xlab='Highway fuel consumption', ylab='CO2 emissions')
abline(model, col = 'red')
```



c)

Vi ser at linjen passer veldig bra, noe vi forventet ettersom vi tidligere har konkludert med at dataene er veldig lineære.

d)

Her bruker vi svaret i a) hvor vi setter inn $x = 8$. Dette gir oss $\hat{y} = 222.034$, som vi ser passer bra med figuren i b). Vi kan også bruker `coef` i R for å hente ut de beregnede verdiene (koeffisientene)

```
x = 8
b0 = coef(model)[1]
b1 = coef(model)[2]
y = b0 + b1 * x
y
```

```
## (Intercept)
##    222.0333
```

Siden vi har med flere desimaler for b_0 og b_1 her, får vi et mer presist svar.

Oppgave 2.58(R)

a)

```
b0 = 6.594
b1 = -0.2606
x = c(1, 3, 5, 7)
y = b0 + b1 * x
y
```

```
## [1] 6.3334 5.8122 5.2910 4.7698
```

b)

```
log_counts = c(6.35957, 5.75890, 5.31321, 4.77068)
log_counts - y
```

```
## [1]  0.02617 -0.05330  0.02221  0.00088
```

Vi ser at vi bare har en negativ differanse i dette tilfellet, mens resten er positive. Merk at det negative avviket er større enn noen av de positive avvikene.

c)

```
sum((log_counts - y)^2)
```

```
## [1] 0.004019817
```

d)

Vi repeterer a), b) og c) for den nye ligningen

```
b0 = 7
b1 = -0.2
y = b0 + b1 * x
# a)
print(y)
```

```
## [1] 6.8 6.4 6.0 5.6
```

```
# b)
print(log_counts - y)
```

```
## [1] -0.44043 -0.64110 -0.68679 -0.82932
```

```
# c)
sum((log_counts - y)^2)
```

```
## [1] 1.76444
```

I b) ser vi at alle avvikene nå er negative. Dette forteller oss at den foreslåtte linja ikke passer så bra til dataene.

e)

Minste kvadraters metode går ut på at man skal finne den linjen som gir det minste kvadratavvikene til data punktene, $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$.

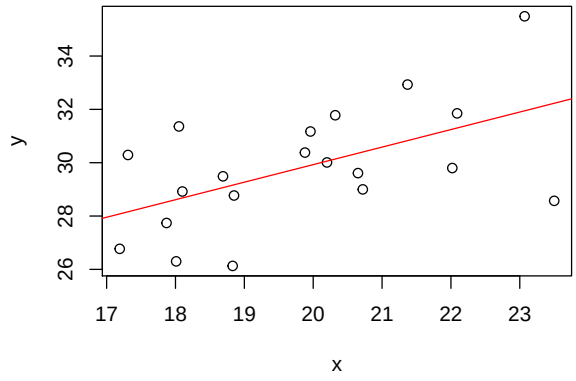
Vi ser at de predikerte verdiene er større for linja i deloppgave d, noe som gjør at alle avvikene er negative. Det tilsier at linja ikke passer veldig godt til punktene våre (siden en litt mindre verdi for b_0 ville gitt mindre avvik). Videre ser vi at kvadratavviket for den nye linjen er mye større for den nye linjen (ca 200 ganger større) som forteller oss at den første linja er best av de to.

Oppgave 2.64(R)

a) og b)

I koden under laster vi inn X og Y i R, for så å plotter dem. I tillegg bruker vi `lm` til å regne ut regresjonskoeffisientene og plotter linjen med `abline`.

```
data = read.csv('../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-064GENDATA.csv')
x = data$X
y = data$Y
plot(x, y)
model = lm(y ~ x)
abline(model, col = 'red')
```



Vi kan bruke `summary` for å få et sammendrag av modellen

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6585 -1.3323  0.0968  1.3625  3.5442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.7764     4.7325   3.545  0.00231 **
## x             0.6575     0.2376   2.768  0.01269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.959 on 18 degrees of freedom
## Multiple R-squared:  0.2985, Adjusted R-squared:  0.2595
## F-statistic:  7.66 on 1 and 18 DF,  p-value: 0.01269
```

Ligningen for regresjonslinjen er da $\hat{y} = 16.78 + 0.66\ x$.

c)

Fra utskriften over ser vi at vi har en $r^2 = 0.2985$, altså er 29.85 % av variansen i Y forklart av X.

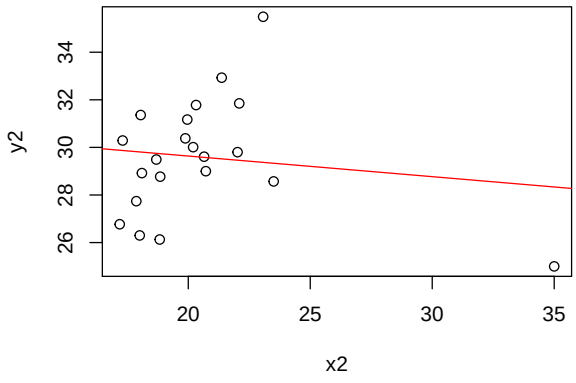
d)

Vi har altså funnet at bare 29.85 % av variansen i Y kan forklares av X. Altså er forholdet mellom X og Y ganske svakt. Dette ser vi også fra figuren i a) og b) ettersom linjen ikke beskriver dataene spesielt godt

Oppgave 2.65(R)

Vi legger nå til en uteligger og gjentar analysen fra 2.80.

```
x2 = c(x, 35)
y2 = c(y, 25)
plot(x2, y2)
model = lm(y2 ~ x2)
abline(model, col = 'red')
```




```
summary(model)
```

```
##
## Call:
```

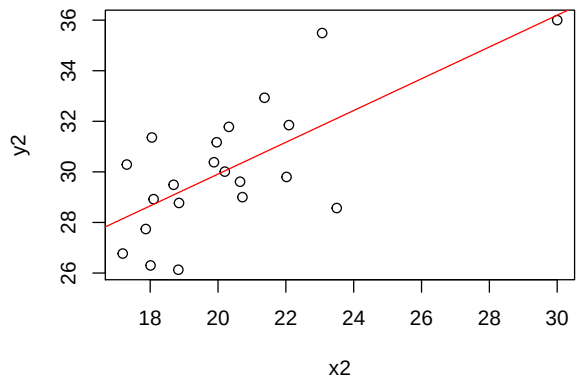
```
## lm(formula = y2 ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6082 -0.9665  0.0296  1.5298  6.1193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.37022    3.07780  10.192 3.87e-09 ***
## x2          -0.08667    0.14736  -0.588    0.563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 19 degrees of freedom
## Multiple R-squared:  0.01788,    Adjusted R-squared:  -0.03381
## F-statistic: 0.3459 on 1 and 19 DF,  p-value: 0.5634
```

men med en stor outlier. Plottet viser en klar trend i dataene, men med en stor outlier. Vi ser at denne ene oulieren gjør at vi nå får en synkende linje i stedet for en stigende (fortegnet på b_1 er nå negativt). Videre ser vi at linjen ikke beskriver dataene noe bra ettersom den er dominert av denne ene outlieren. Faktisk er bare 1.79 % av variansen i Y for klart av X. Det er veldig tydelig at outlieren er ødeleggende for regresjonsanalysen.

Oppgave 2.66(R)

a)

```
x2 = c(x, 30)
y2 = c(y, 36)
plot(x2, y2)
model = lm(y2 ~ x2)
abline(model, col = 'red')
```




```
summary(model)
```

```
##
## Call:
## lm(formula = y2 ~ x2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5416 -1.3649 -0.0282  1.2826  3.6486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.3465     3.0341   5.717 1.64e-05 ***
## x2            0.6283     0.1479   4.248 0.000435 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.908 on 19 degrees of freedom
## Multiple R-squared:  0.4871, Adjusted R-squared:  0.4601
## F-statistic: 18.04 on 1 and 19 DF,  p-value: 0.0004353
```

Vi har en stigende sammenheng mellom X og Y med en outlier for $X = 30$. Selv om vi har lagt til denne outlieren får vi en line som er veldig lik den i Oppgave 1.80 (se at koeffisientene er veldig like). Videre ser vi at X nå beskriver 48.71 % av variansen til Y. Dette tilsvarer noe nær en dobling fra Oppgave 1.80. Dette viser at selv om outlieren ikke forandrer regresjonslinjen noe særlig, har den et unaturlig høyt bidrag til analysen (en liten forandring i denne outlieren tilsvarer en stor forskjell i regresjonsanalysen).

b)

Når vi sammenligner resultatene i denne oppgave med Oppgave 2.81 ser vi at en outlier har en veldig dominerende effekt på regresjonsanalysen. Den påvirker både regresjonslinjen vi får gjennom minste kvadraters metode og hvor stor forklaringskraft forklaringsvariabelen X har. Dette kan gjøre av vi får en gal tolkning av forholdet mellom X og Y hvis man ikke er er bevisst effekten av uteliggere og tar hensyn til det i dataanalysen.

Midtveiseksamen Våren 2006

1) Interkvartilavstanden er $Q_3 - Q_1 = 5$, som gir a).
2) Variansen er `StDev` kvadrert = 11.86 som gir d).
3) b).
4) c).
5) c).
6) d).
7) Vi har at $r = \sqrt{r^2} = 0.5523$, som gir c). (Husk å tenke på fortegnet her).
8) d).