STK1000: Løsningsforslag Uke 37

2019

Oppgave 2.6

a) Det er naturlig å tro at det er lite sammenheng mellom hvor mye en person liker og syng og hvor mye personen liker å danse. (Men hvis du mener det er en sammenheng er det vilkårlig hvilken du velger som forklaringsvariabel og respons.)

b) Trolig er prisen på en bok delvis avhengig av antall sider, vi kan derfor si at antall sider er en forklaringsvariabel og pris er en respons.

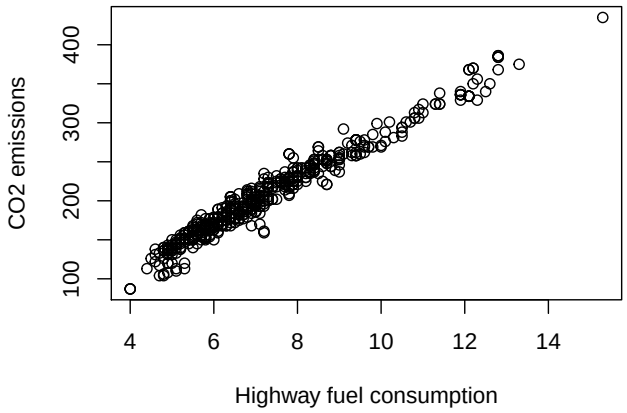
c) Antall enheter med alkohol konsumert (i et tidsrom) kan være med på å forklare alkoholprosenten i blodet til en person. Vi har derfor at enheter med alkohol er en forklaringsvariabel og alkoholprosent i blodet er en respons.

d) I denne studien ønsker de trolig å se hvordan D-vitamin påvirker "bone mineral content". Derfor er dose D-vitamin en forklaringsvariabel og "bone mineral content" respons.

Oppgave 2.21(R)

a)

```
data = read.csv('CSV/Chapter 2/EX02-021CANFREG.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.

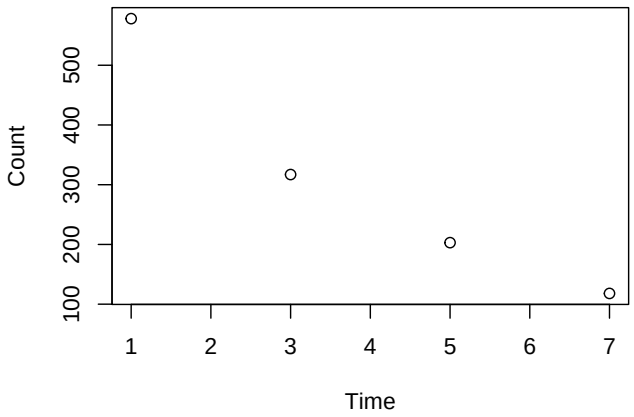
c) Vi har en observasjon som har en del høyere drivstofforbruk og CO2-utslipp en resten. Denne kan eventuelt regnes som en outlier.

d) En rett linje vil oppsummere forholde mellom drivstofforbruk og CO2-utslipp. Den vil også hjelpe på å evaluere styrken (*strength*).

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

Oppgave 2.32

a) Vi plotter tid på x-aksen og antall på y-aksen ettersom vi forventer at tiden forklarer antallet (tid er forklaringsvariabel og antall er respons).



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

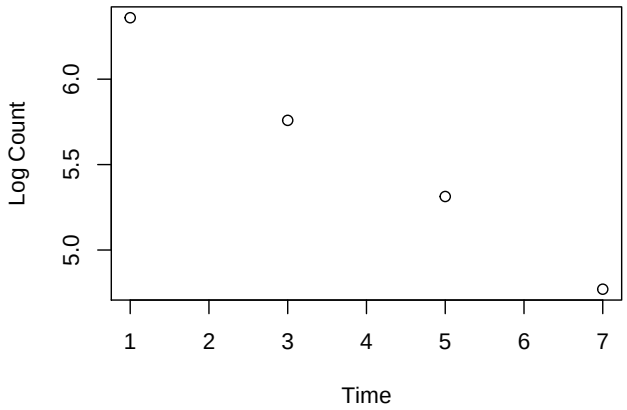
c) Formen har en svak positiv kurvatur og styrken ser ut til å være høy (tydelig mønster).

d) Det er rart å snakke om outliere i et slikt eksperiment.

e) Forholdet er ikke helt lineær, men har en positiv kurvatur. (**Ekstra:** Radioaktivitet synker typisk eksponentielt, noe som kan samsvare med kurven.)

Oppgave 2.33

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



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

c) Formen er lineær og styrken er høy.

d) Det er ingen outliers her.

e) Forholdet er veldig lineært. Dette tyder på at kurven i Oppgave 2.32 var eksponentielt synkende.

Oppgave 2.44(R)

Vi kan finne korrelasjonen med `cor` i R:

```r
data = read.csv('CSV/Chapter 2/EX02-044CANFREG.csv')
cor(data$FuelConsHwy, data$CO2)
```

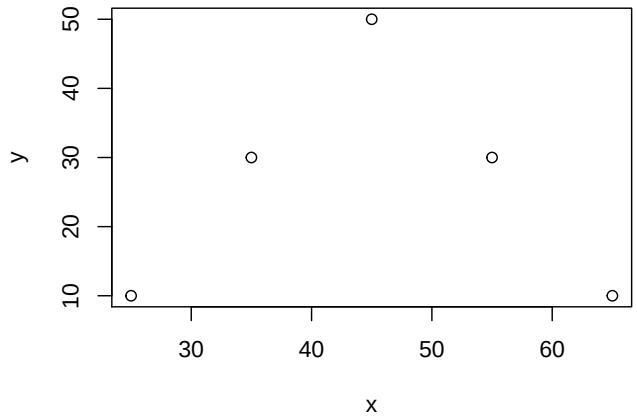
```
## [1] 0.9805158
```

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 2.46(R)

a)

```r
x = c(25, 35, 45, 55, 65)
y = c(10, 30, 50, 30, 10)
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.

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

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

d) Poenget her er å understreke at korrelasjon beskriver *lineære* sammenhenger mellom variabler og er ikke et mål på avhengighet eller styrke (*strength*) på mønster.

Oppgave 2.50

a) Vi må bruke formelen for korrelasjon på side 101 i boken (Chaper 2.3). Først estimerer vi standard-avvikene 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 2.51

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

a) Vi kan finn 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) Her er korrelasjonen et bra oppsummeringsmål ettersom dataene har et lineære forhold.

c) Hvis vi sammenligner med Oppgave 2.50 ser vi at begge har en korrelasjon som er nær -1. Dette tyder på en sterk sammenheng mellom dataene. Men ettersom dataene i Oppgave 2.50 ikke er lineære er ikke korrelasjonen et like godt oppsummeringsmål her.

Oppgave 2.66(R)

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('CSV/Chapter 2/EX02-066CANFREG.csv')
y = data$CO2
x = data$FuelConsHwy
model = lm(y ~ x)
model
```

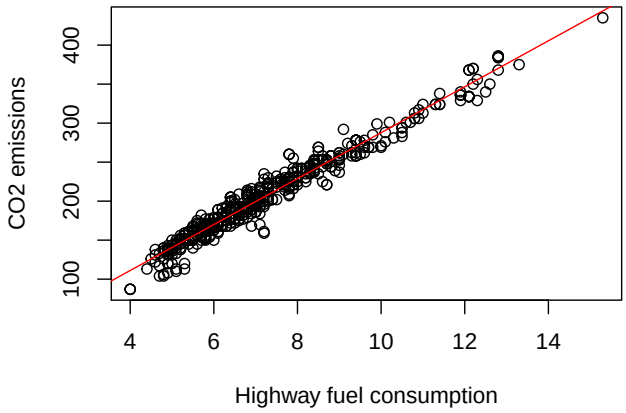
```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      -6.694       29.444
```

Altså får vi ligningen $\hat{y}$ = -6.694 + 29.444 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 $\hat{y} = 228.858$, 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)
##    228.8563
```

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

Oppgave 2.74(R eller kalkulator)

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 en negative differanse, mens resten er positiv. Mer også 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 vi ikke har den foreslått linje 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, $\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$.

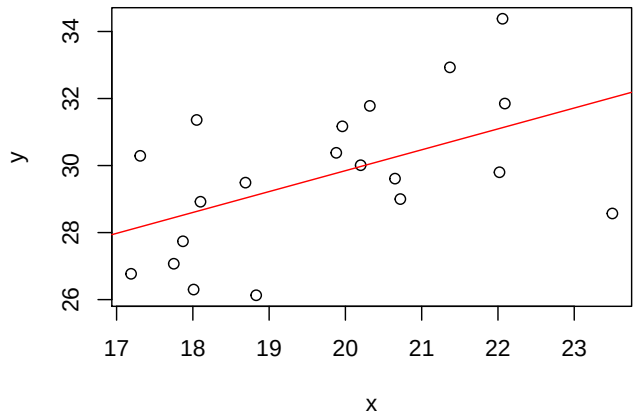
Vi ser at de predikerte verdiene er større for den nye ligningen, noe som gjør at alle avvikene er negative. Det tilsier at linjen ikke passer veldig bra til punktene våre (siden en litt mindre b_0 vil gi mindre avvik). Videre ser vi at kvadratavviket for den nye linjen er mye større for den nye linjen (ca 200 ganger større) som forteller oss at den første linjen er best.

Oppgave 2.80(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`.

```r
data = read.csv('CSV/Chapter 2/EX02-080GENDATA.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

```r
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4583 -1.3107  0.1481  1.4446  3.2492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.3804     4.7420   3.665  0.00177 **
## x             0.6233     0.2394   2.604  0.01794 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.939 on 18 degrees of freedom
## Multiple R-squared:  0.2737, Adjusted R-squared:  0.2333
## F-statistic: 6.782 on 1 and 18 DF,  p-value: 0.01794
```

Ligningen for regresjonslinjen er da $\hat{y} = 17.38 + 0.62\ x$.

c)

Fra utskriften over ser vi at vi har en $r^2 = 0.2737$, altså er 27.37 % av variansen i Y forklart av X.

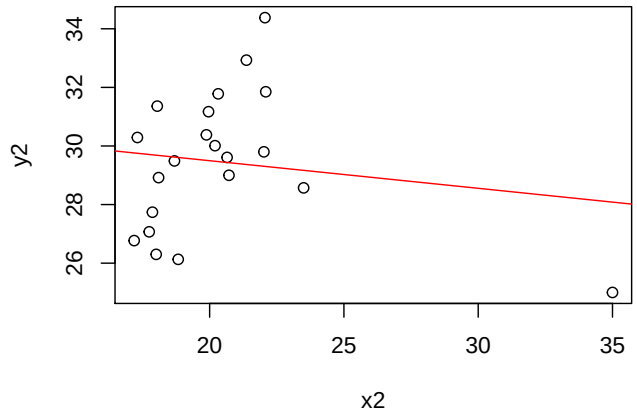
d)

Vi har altså funnet at bare 27.37 % 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 bra.

Oppgave 2.81(R)

Vi legger nå til en outlier 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)
```

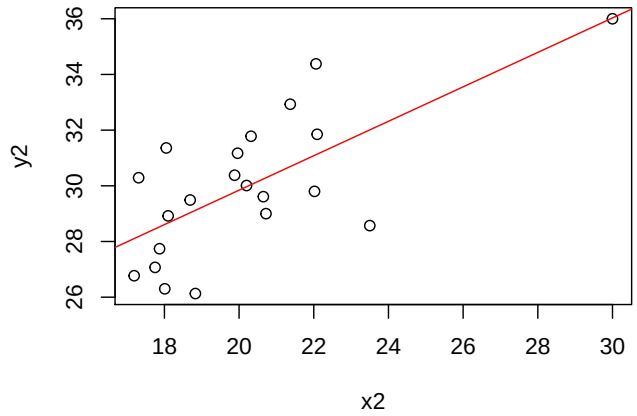
```
##
## Call:
## lm(formula = y2 ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4780 -1.9585  0.1735  1.6685  5.0764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.38274    2.96823   10.57 2.13e-09 ***
## x2          -0.09425    0.14279   -0.66    0.517
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.422 on 19 degrees of freedom
## Multiple R-squared:  0.02242,    Adjusted R-squared:  -0.02903
## F-statistic: 0.4357 on 1 and 19 DF,  p-value: 0.5171
```

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 2.24 % av variansen i Y for klart av X. Det er veldig tydelig at outlieren er ødeleggende for regresjonsanalysen.

Oppgave 2.82(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
## -3.4391 -1.2935  0.0423  1.3508  3.2617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.4717     2.9892   5.845 1.25e-05 ***
## x2            0.6186     0.1464   4.224 0.000459 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.887 on 19 degrees of freedom
## Multiple R-squared:  0.4843, Adjusted R-squared:  0.4572
## F-statistic: 17.84 on 1 and 19 DF,  p-value: 0.0004592
```

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 48.43 % av variansen til Y. Dette tilsvarer nesten 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 forsiktig.

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).