STK1000: Løsningsforslag Uke 37

H2023

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).

Section 2.1: Oppgave 2.2

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.

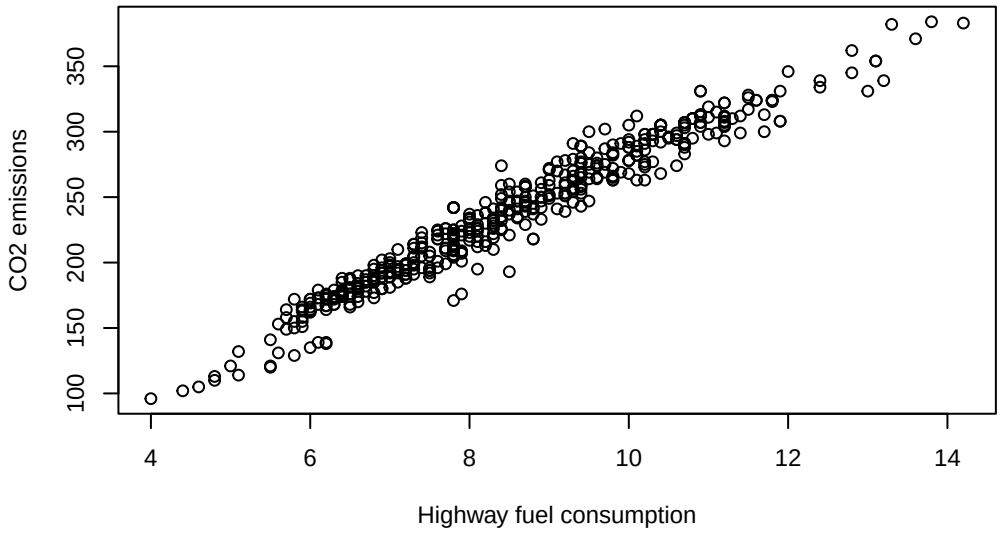
Section 2.2: Oppgave 2.11

a)

```
options(width=100)
data = read.csv('../../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-011CANFREG.csv')
head(data)
```

```
  Year  Make                Model          Class EngineSize Cylinders Transmission FuelType
1 2018  AUDI                   A3     SUBCOMPACT        2.0         4           AM7        X
2 2018  AUDI            A3 QUATTRO     SUBCOMPACT        2.0         4           AM6        X
3 2018  AUDI A3 CABRIOLET QUATTRO     SUBCOMPACT        2.0         4           AM6        X
4 2018  AUDI      TT COUPE QUATTRO     SUBCOMPACT        2.0         4           AM6        X
5 2018  AUDI   TT ROADSTER QUATTRO     TWO-SEATER        2.0         4           AM6        X
6 2018 BUICK               ENCLAVE SUV - STANDARD        3.6         6           AS9        X
  FuelConsCity FuelConsHwy FuelConsComb MPG CO2 MPGCity MPGHwy
1          9.1         6.8          8.0  35 188      31     41
2          9.7         7.5          8.7  32 205      29     38
3         10.8         8.0          9.5  30 223      26     35
4         10.1         7.8          9.1  31 209      28     36
5         10.1         7.8          9.1  31 209      28     36
6         12.9         9.0         11.2  25 263      22     31
```

```r
par(cex=0.75)
plot(data$FuelConsHwy, data$CO2, xlab = 'Highway fuel consumption', ylab = 'CO2 emissions')
```



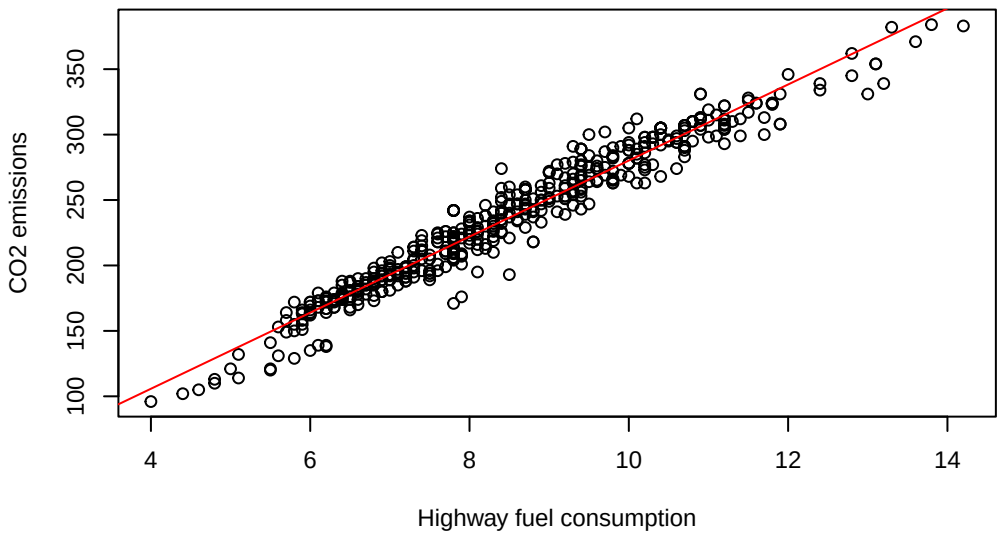
b) Vi ser det er en lineær trend (form) i positiv retning. Styrken (*strength*) er høy da det er et veldig tydelig mønster.

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.

Section 2.2: Oppgave 2.12

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.

```r
par(cex=0.75) # størrelsen på teksten i figuren blir mindre
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.

Section 2.2: Oppgave 2.13

a) Vi leser inn den større versjonen av datasettet og sjekker at de to versjonene er som beskrevet i oppgavene, før vi lager figuren.

```
options(width=100)
data_all = read.csv('../../ips10e_csv_data_sets/ips10e_ch2_csv_data_sets/EX02-013CANFUEL.csv')
head(data_all)
```

```
  Year  Make             Model       Class EngineSize Cylinders Transmission FuelType FuelConsCity
1 2018 ACURA               ILX     COMPACT        2.4         4          AM8        Z          9.4
2 2018 ACURA       MDX SH-AWD SUV - SMALL        3.5         6          AS9        Z         12.6
3 2018 ACURA MDX SH-AWD ELITE SUV - SMALL        3.5         6          AS9        Z         12.2
4 2018 ACURA           RDX AWD SUV - SMALL        3.5         6          AS6        Z         12.4
5 2018 ACURA       RLX HYBRID    MID-SIZE        3.5         6          AM7        Z          8.4
6 2018 ACURA               TLX     COMPACT        2.4         4          AM8        Z         10.0
  FuelConsHwy FuelConsComb MPG CO2 MPGCity MPGHwy
1         6.8          8.2  34 192      30     41
2         9.0         11.0  26 259      22     31
3         9.0         10.7  26 251      23     31
4         8.7         10.7  26 250      23     32
5         8.2          8.4  34 196      34     34
6         7.1          8.7  32 205      28     40
```

```
dim(data)
```

```
[1] 502  15
```

```
dim(data_all)
```

```
[1] 1045   15
```

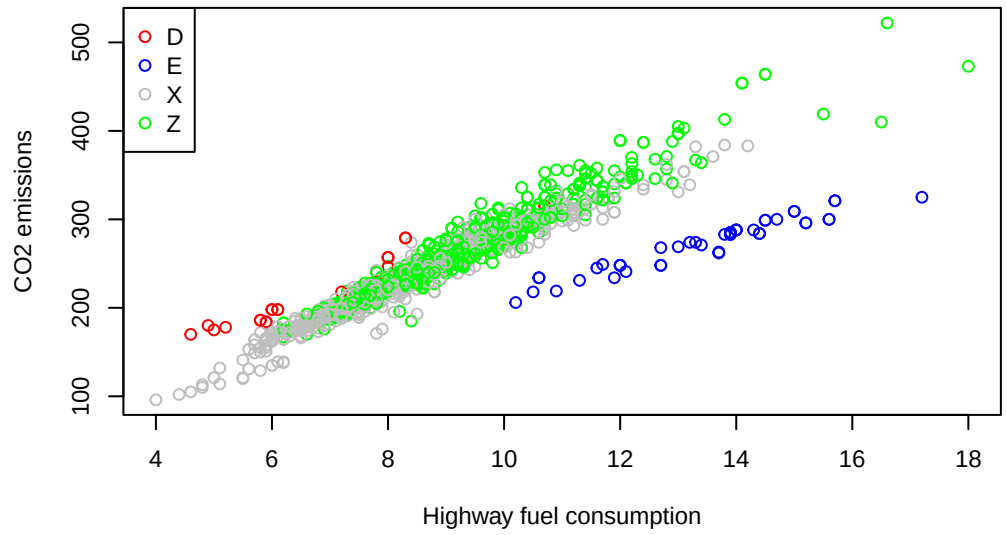
```
table(data$FuelType)
```

```
  X
502
```

```
table(data_all$FuelType)
```

```
  D   E   X   Z
 20  47 502 476
```

```
ind = rep(NA, dim(data_all)[1]) # lager en vektor like lang som antall datapunkter i data_all
ind.tmp = which(data_all$FuelType=="D")
ind[ind.tmp] = "red"
ind.tmp = which(data_all$FuelType=="E")
ind[ind.tmp] = "blue"
ind.tmp = which(data_all$FuelType=="X")
ind[ind.tmp] = "gray"
ind.tmp = which(data_all$FuelType=="Z")
ind[ind.tmp] = "green"
par(cex=0.75)
plot(data_all$FuelConsHwy, data_all$CO2, col=ind,
     xlab = 'Highway fuel consumption', ylab = 'CO2 emissions')
legend("topleft", legend = c("D", "E", "X", "Z"),
       col=c("red", "blue", "gray", "green"), pch=1)
```



b) Ja, punkter i forskjellige farger faller i forskjellige ikke overlappende grupper.

Section 2.3: Oppgave 2.32

Vi kan finne korrelasjonen med `cor` i R:

```r
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 2.11.

Section 2.3: Oppgave 2.33

Vi regner korrelasjon for de tre andre typene:

```r
ind = which(data_all$FuelType == "D")
cor(data_all$FuelConsHwy[ind], data_all$CO2[ind])
```

```
[1] 0.9706754
```

```r
ind = which(data_all$FuelType == "E")
cor(data_all$FuelConsHwy[ind], data_all$CO2[ind])
```

```
[1] 0.9635226
```

```r
ind = which(data_all$FuelType == "Z")
cor(data_all$FuelConsHwy[ind], data_all$CO2[ind])
```

```
[1] 0.9657645
```

Det er svært høy korrelasjon som tilsier at CO2-utslipp kan i stor grad forklares av drivstoffbruk for hver type drivstoff. Det samsvarer bra med det vi ser på plottet i Oppgave 2.13.

Section 2.4: Oppgave 2.50

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`.

```r
y = data$CO2
x = data$FuelConsHwy
model = lm(y ~ x)
model
```

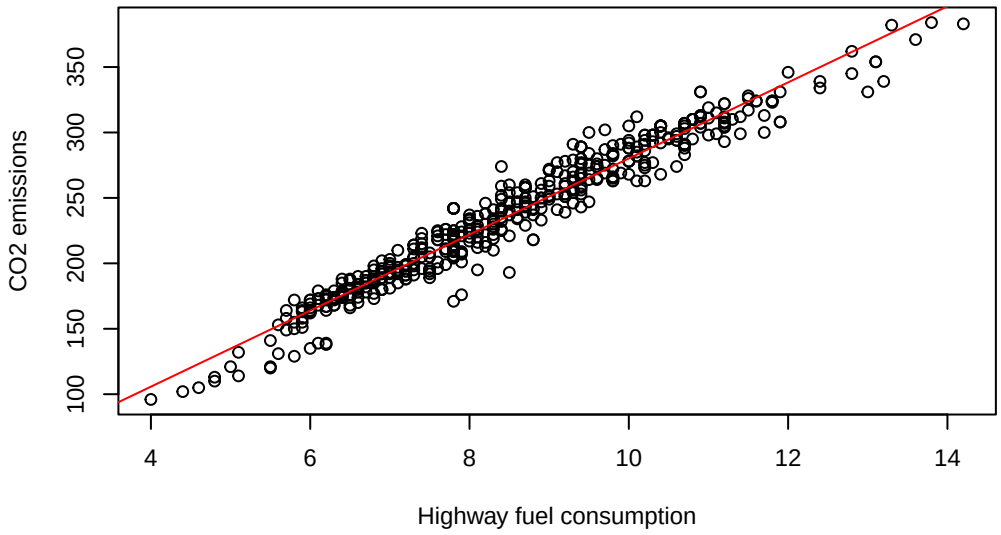
```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -10.49        29.07
```

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

b)

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

```
par(cex=0.75)
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
a = coef(model)[1]
b = coef(model)[2]
y = b + b * x
y
```

```
       x
261.5931
```

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

Section 2.4: Oppgave 2.51

a)

Vi kan tilpasse en rett linje til det større datasettet med funksjonen `ln`. Her er igjen `FuelConsHwy` forklaringsvariabelen x og `CO2` responsvariabelen y.

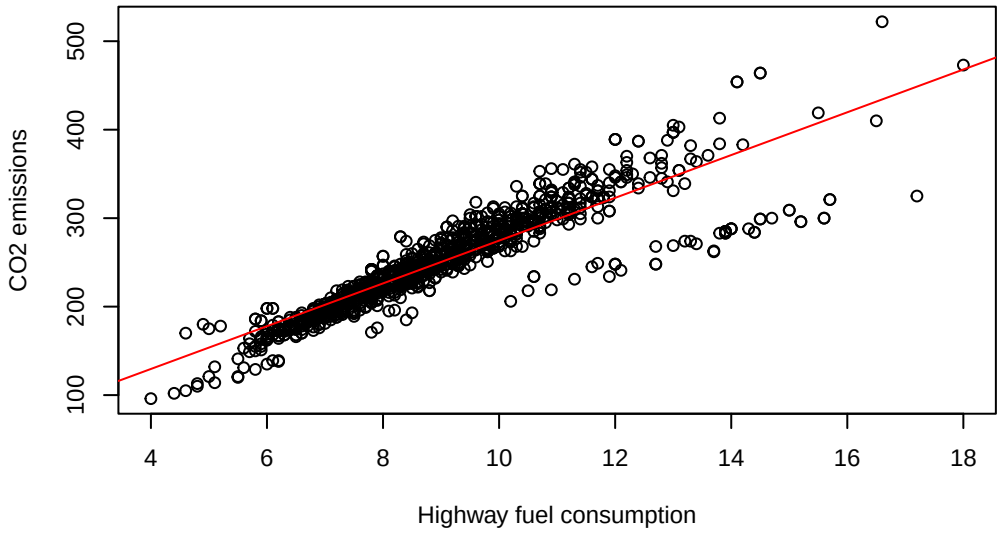
```
y = data_all$CO2
x = data_all$FuelConsHwy
model = lm(y ~ x)
model
```

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

Coefficients:
(Intercept)            x
      32.85        24.18
```

b)

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

```
par(cex=0.75)
plot(x, y, xlab='Highway fuel consumption', ylab='CO2 emissions')
abline(model, col = 'red')
```



c)

Vi ser at det er ikke mulig å tilpasse en rett linje gjennom hele datasettet på en god måte. Resultatene fra Oppgave 2.33 viser at CO2-utslipp kan i stor grad forklares av drivstoffbruk for hver type drivstoff, men CO2-utslippen er veldig forskjellig mellom typer drivstoff. Noen typer drivstoff fører til mye mer CO2-utslipp enn andre. For

eksempel, så ligger de blå punktene under de andre punktene i figuren i oppgave 2.13 som sier at etanol fører til mindre CO2-utslipp enn de andre typer drivstoff.