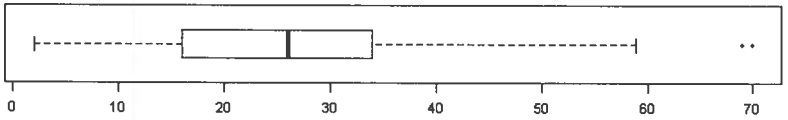
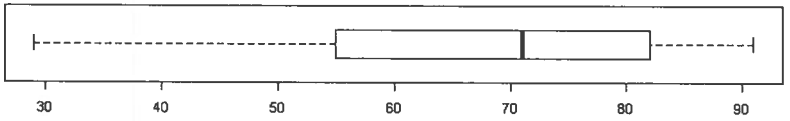
Fra boka: 10.32, 10.33, 10.34, 10.35, 10.3 og 10.37 (alle er basert på samme datasett).

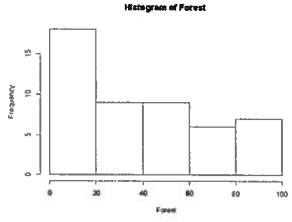
1

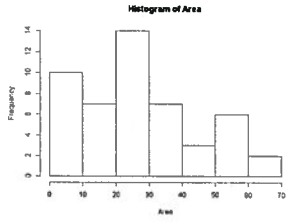
```
############ OPPGAVE 10.32
# Vannkvalitet. n=49 målinger i ulike områder.
# Forutsetter at datasettene til boka (i excel-format) er lastet ned fra hjemmesiden til boka,
# http://www.macmillanlearning.com/Catalog/studentresources/ips8e#t_922171
# Åpne riktig datasett (ex10-32.IBI.xls) i excel, og lagre som CSV-fil:
# File - Save as - File name:     ex10-32.IBI.csv
#                  Save as type:  CSV
# RStudio: Import dataset - From local file .... ex10-32.IBI.csv
#                                                Import

attach(ex10.32ibi)
# Nå heter variablene
# Forest , % skog i et område
# Area   , arealet av området*
# IBI    , Index of biotic integrity

# a)
hist(Forest) # Skjev med høyrehale
hist(Area)   # Ganske N-fordelt
hist(IBI)    # Skjev med venstrehale

par(mfrow=c(3,1))
boxplot(Forest,horizontal=TRUE)
boxplot(Area,horizontal=TRUE)
boxplot(IBI,horizontal=TRUE)

par(mfrow=c(1,1))
qqnorm(Forest)
qqnorm(Area)
qqnorm(IBI)

summary(ex10.32ibi)
c(mean(Forest),sd(Forest))
c(mean(Area),sd(Area))
c(mean(IBI),sd(IBI))
```

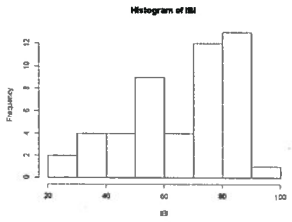
* Watershed area: Area of the entire region draining into a river, a river system, or other body of water.

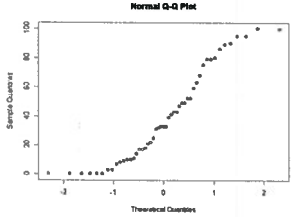


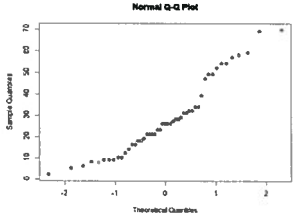


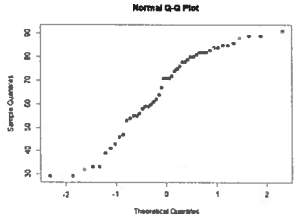












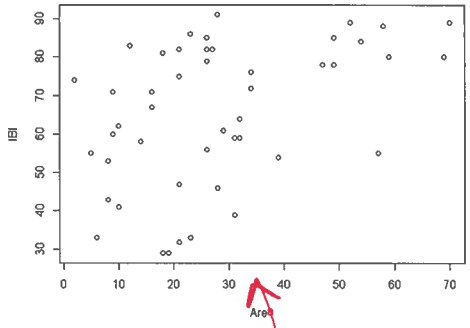
```
# Velger til slutt kanskje å oppsummere Area med mean og sd, IBI og Forest med median og kvartiler?
# Disse variablene er ikke helt symmetriske, og mean +/- 2*sd havner på utsiden av det som er mulig, både for Area og Forest:
# Area er et areal og må være positivt, og Forest er en andel og må være mellom 0 og 1.
# Men fordi skjevheten er slik at det ikke er noen superekstreme verdier som trekker opp eller ned, kunne jeg også,
# fra et mer pragmatisk ståsted, akseptere å bruke mean og sd som oppsummering av alle tre variabler. Jeg går ut i fra at
# de som leser oppsummeringen også ser at Area og Forest ikke kan være < 0, og at de dermed vil tolke et såpass stort sd
# (sammenlignet med mean) som et tegn på at data er noe skjeve, men ikke så skjeve at den som oppsummerer dem tar seg bryet
# med å bruke noe annet enn mean og sd. I så fall er det en god oppsummering.
# Men det krever en vurdering av både sender og mottaker.

> summary(ex10.32ibi)
     Forest             Area            IBI
 Min.   :  0.00   Min.   : 2.00   Min.   :29.00
 1st Qu.: 10.00   1st Qu.:16.00   1st Qu.:55.00
 Median : 33.00   Median :26.00   Median :71.00
 Mean   : 39.39   Mean   :28.29   Mean   :65.94
 3rd Qu.: 63.00   3rd Qu.:34.00   3rd Qu.:82.00
 Max.   :100.00   Max.   :70.00   Max.   :91.00

> c(mean(Forest),sd(Forest))
[1] 39.38776 32.20431              # Litt mange desimaler?

> c(mean(Area),sd(Area))
[1] 28.28571 17.71417

> c(mean(IBI),sd(IBI))
[1] 65.93878 18.27955
```

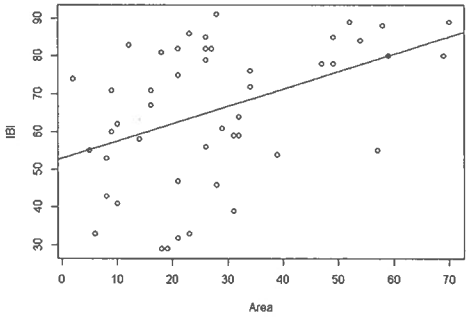
$\rightarrow \bar{x} = 39$, $sd = 32$

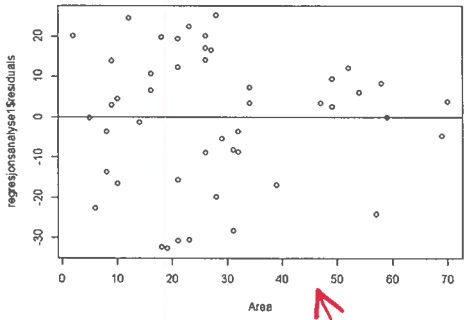
$\rightarrow \bar{x} = 28$, $sd = 18$

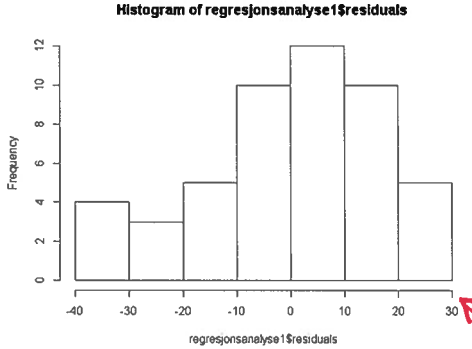
$\rightarrow \bar{x} = 66$, $sd = 18$

```
# b)
plot(Area, IBI) # Ser: Større variasjon i IBI-verdier i små områder (lave Area-verdier)
                # enn i store områder (høye Area-verdier) (Virker logisk)
```









$y = IBI$

$x = Area$

Modell: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma)$, uavh.

```
# e)
lm(IBI~Area)
summary(lm(IBI~Area))
```

Resultater s. (5)

```
plot(Area,IBI)
abline(lm(IBI~Area))
```

Problemstilling: Er det en sammenheng mellom Area og IBI?

H_0: Det er ingen (null) sammenheng: $\beta_1 = 0$

H_1: Det er en sammenheng: $\beta_1 \neq 0$

Antakelser: Lineær trend i sammenhengen, uavhengige & normalfordelte residualer.

sjekkes i scatterplot

sjekkes ved histogram av residualene

Residualene skal egentlig helle ikke avhenge av noen av variablene i analysen, og det sjekke vi ved å plotte dem mot x og mot y eller $\hat{y}$. Her har vi bare plottet mot x.

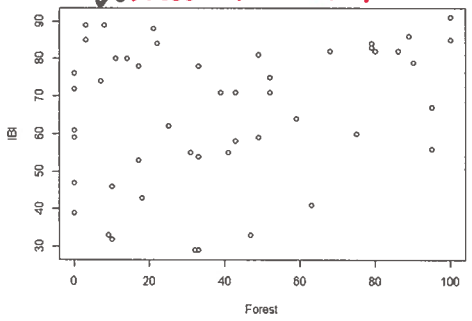
```
# f)
# Residualer (som vi ønsker skal være lik 0:
regresjonsanalyse1 <- lm(IBI~Area)
ls(regresjonsanalyse1)
regresjonsanalyse1$residuals

plot(Area, regresjonsanalyse1$residuals)
abline(h=0)
# ser at residualene er større for lave verdier av Area.
# Det betyr at modellen passer dårligst til de lave verdiene av Area.

# g)
hist(regresjonsanalyse1$residuals) # Tja. Normalfordelte?
```

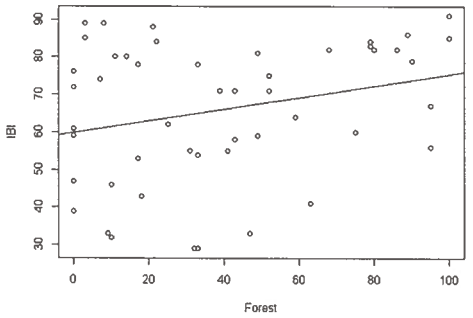
Plottene viser at Antakelsene er ikke helt etter læreboka, men ikke verre enn at vi godtar resultatene fra analysen.

```
############ OPPGAVE 10.33
# b)
plot(Forest, IBI) # Ser: Litt større variasjon i IBI-verdier i områder med lite skog
# enn i områder med mye skog (Virker logisk)
```

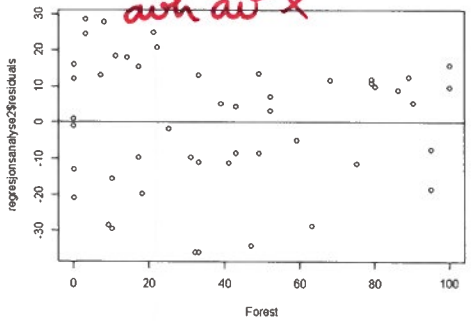
Antakelser:

Lineær trend? OK

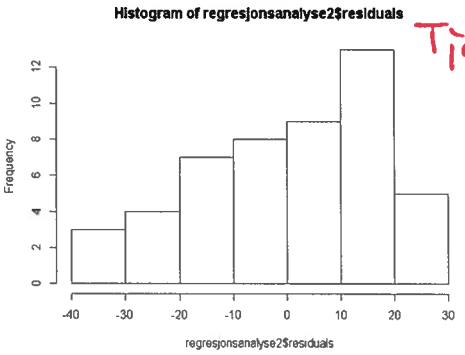




Residualer ikke OK avh. av x



N-ford. residualer?



Tja?

```
# e)
lm(IBI~Forest)
summary(lm(IBI~Forest))
```

$y = IBI$
$x = Area$

Modell (fortsatt) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma)$, uavh.

Resultater s⑥

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

```
plot(Forest,IBI)
abline(lm(IBI~Forest))

# f)
# Residualer (som vi ønsker skal være lik 0:
regresjonsanalyse2 <- lm(IBI~Forest)
ls(regresjonsanalyse2)
regresjonsanalyse2$residuals

plot(Forest, regresjonsanalyse2$residuals)
abline(h=0)
# ser at residualene er litt større når det er lite skog
# Det betyr at modellen passer dårligst til områder med lite skog.

# g)
hist(regresjonsanalyse2$residuals) # Tja. Normalfordelte?
```

```
> # Oppgave 10.32 e)
> summary(lm(IBI~Area))

Call:
lm(formula = IBI ~ Area)

Residuals:
    Min      1Q  Median      3Q     Max
-32.666  -8.887   3.432  12.414  25.193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.9230     4.4835  11.804 1.17e-15 ***
Area          0.4602     0.1347   3.415  0.00132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 47 degrees of freedom
Multiple R-squared:  0.1988,    Adjusted R-squared:  0.1818
F-statistic: 11.67 on 1 and 47 DF,  p-value: 0.001322

>
> # f)
> # Residualer (som vi ønsker skal være ~~lik 0~~ ~N(0,σ))
> regresjonsanalyse1 <- lm(IBI~Area)
> ls(regresjonsanalyse1)
 [1] "assign"        "call"          "coefficients"  "df.residual"   "effects"       "fitted.values"
 [7] "model"         "qr"            "rank"          "residuals"     "terms"         "xlevels"
>
> regresjonsanalyse1$residuals
          1           2           3           4           5           6           7           8
-15.58621643  -5.26745777 -28.18776811  -8.64792328   3.43176639   7.43176639   9.52943886  12.14897336
          9          10          11          12          13          14          15          16
 20.15673177   3.86618033 -22.68388891 -19.80730261 -30.58621643  -0.07211282  -4.67366450   3.44974920
         17          18          19          20          21          22          23          24
 -3.60419924 -13.60419924   8.38804235   6.22866302   4.47549042 -24.15180248 -32.20575092 -32.66590609
         25          26          27          28          29          30          31          32
-16.86900945   2.52943886  13.93564559  -0.22373374  -1.36513025  13.93564559 -30.50652676  -8.18776811
         33          34          35          36          37          38          39          40
 19.79424908  10.71455941  12.41378357  -3.64792328 -16.52450958  17.11300773   2.93564559   6.22866302
         41          42          43          44          45          46          47          48
 24.55518009  19.41378357  16.65285256  22.49347324  14.11300773   6.71455941  -8.88699227  20.11300773
         49
 25.19269739
```

$IBI_i = \beta_0 + \beta_1 \cdot Area_i + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma)$

$\hat{\beta}_0 = 52.9230$, $SE(\hat{\beta}_0) = 4.4835$ p-verdi for $H_0: \beta_0 = 0$: 0.00000000000000117 Forkast H_0 på nivå 0.01

$\hat{\beta}_1 = 0.4602$, $SE(\hat{\beta}_1) = 0.1347$ p-verdi for $H_0: \beta_1 = 0$: 0.00132 Forkast H_0 på nivå 0.01

β_0: β_1:

estimat for σ

Altså: en enhets endring i Area gir en forventet endring på 0.46 enheter i IBI.

$\hat{\varepsilon}_i$, altså $(y_i - \hat{y}_i) = (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$, altså enkeltavvikene fra linja!

H_0 er de.

```
> # Oppgave 10.33 e)
> summary(lm(IBI~Forest))

Call:
lm(formula = IBI ~ Forest)

Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,  Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084

> regresjonsanalyse2 <- lm(IBI~Forest)
> ls(regresjonsanalyse2)
 [1] "assign"        "call"          "coefficients"  "df.residual"   "effects"       "fitted.values"
 [7] "model"         "qr"            "rank"          "residuals"     "terms"         "xlevels"
```

$IBI_i = \beta_0 + \beta_1 \cdot Forest_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma)$

$\hat{\beta}_0 = 59.90725$, $SE(\hat{\beta}_0) = 4.03957$ } p-verdi for $H_0: \beta_0 = 0$: 0.000... 2

Forkast H_0 på nivå 0.01: β_0 er signifikant forskjellig fra 0.

$\hat{\beta}_1 = 0.15313$, $SE(\hat{\beta}_1) = 0.07972$ } p-verdi for $H_0: \beta_1 = 0$: 0.06

Behold H_0 på nivå 0.01.

β_1 er ikke signifikant forskjellig fra 0.

β_0: β_1:

estimat for σ

regresjonsanalyse$residuals er residualene i analysen, altså avvikene fra linja: $(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$

Oppgave 10.34 Hvis du måtte velge: Hvilken ville du valgt?

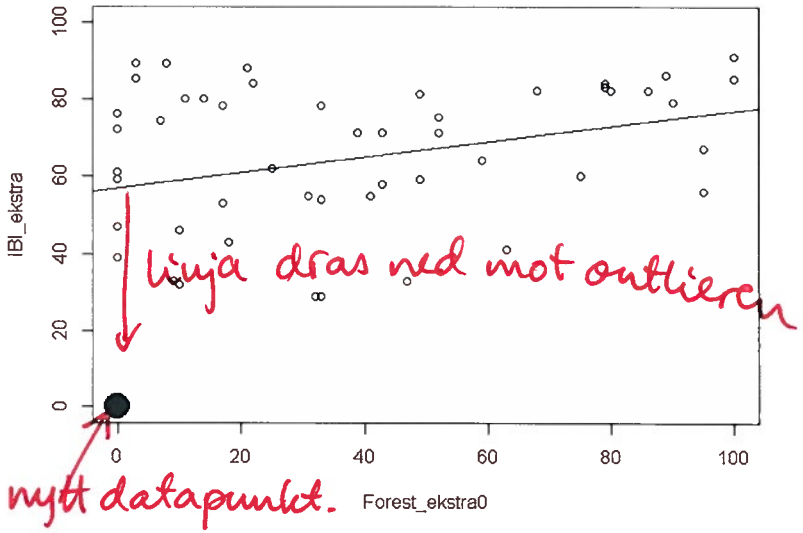
Svar: Velger i så fall den som predikerer best, altså i dette tilfellet den som har en kovariat som forklarer størst andel av variasjonen i y

```
> summary(lm(IBI~Area))

Call:
lm(formula = IBI ~ Area)

Residuals:
    Min      1Q  Median      3Q     Max
-32.666  -8.887   3.432  12.414  25.193

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.9230     4.4835  11.804 1.17e-15 ***
Area          0.4602     0.1347   3.415  0.00132 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.53 on 47 degrees of freedom
Multiple R-squared:  0.1988,   Adjusted R-squared:  0.1818                    # Jeg ville valgt denne!
F-statistic: 11.67 on 1 and 47 DF,  p-value: 0.001322

> summary(lm(IBI~Forest))

Call:
lm(formula = IBI ~ Forest)

Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,  Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084
```

Oppgave 10.35 Effekten av en outlier

Legger til en ekstra «måling» i stedet for å bytte ut en. Utgangspunktet er `IBI` og `Forest`. Lager en ny IBI-variabel og to nye Forest-variabler.

```
IBI_ekstra      <- c(IBI,0)
Forest_ekstra0  <- c(Forest,0)
plot(Forest_ekstra0,IBI_ekstra,ylim=c(0,100))
points(0,0,pch=16,cex=3)
abline(lm(IBI_ekstra~Forest_ekstra0))
```

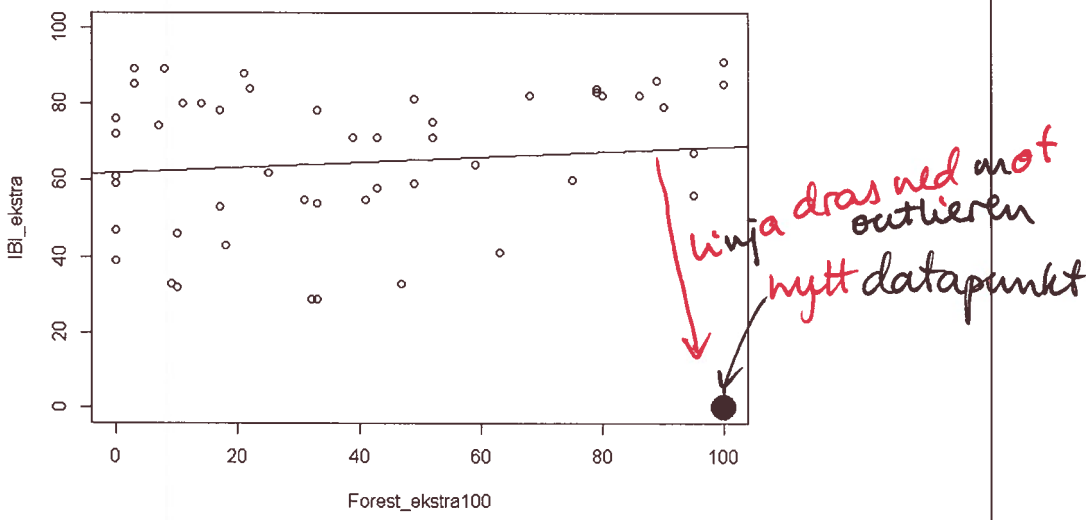



```
> summary(lm(IBI_ekstra~Forest_ekstra0))

Residuals:
    Min      1Q  Median      3Q     Max
-56.969 -10.064   4.111  14.292  31.436

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.96922    4.32222   13.18   <2e-16 ***
Forest_ekstra0  0.19821    0.08617    2.30   0.0258 *
---

Residual standard error: 19.52 on 48 degrees of freedom
Multiple R-squared:  0.09928,  Adjusted R-squared:  0.08052
F-statistic: 5.291 on 1 and 48 DF,  p-value: 0.02583
```

$\hat{\beta}_1$

Signifikant på 0.05 nivå

```
IBI_ekstra      <- c(IBI,0)
Forest_ekstra100  <- c(Forest,100)
plot(Forest_ekstra100,IBI_ekstra,ylim=c(0,100))
points(100,0,pch=16,cex=3)
abline(lm(IBI_ekstra~Forest_ekstra100))
```




```
> summary(lm(IBI_ekstra~Forest_ekstra100))

Residuals:
    Min      1Q  Median      3Q     Max
-68.744 -10.065   5.901  15.655  26.991

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.80108    4.60916  13.408   <2e-16 ***
Forest_ekstra100  0.06943    0.08844   0.785    0.436
---

Residual standard error: 20.43 on 48 degrees of freedom
Multiple R-squared:  0.01268,  Adjusted R-squared:  -0.007892
F-statistic: 0.6163 on 1 and 48 DF,  p-value: 0.4363
```

$\hat{\beta}_1$

ikke-signifikant på 0.05 nivå.

Oppsummering:

Outliere påvirker alle resultatene fra analysen: Estimater, S.E.-er, konfidensintervaller, p-verdier, residualer og andel forklart varians. Jo mer avvikende de er fra resten av strukturen I data, jo mer påvirker de resultatene.

Det oppgaven viser, men som også er sant, er at outliere får mindre betydning dersom antallet observasjoner er stort.

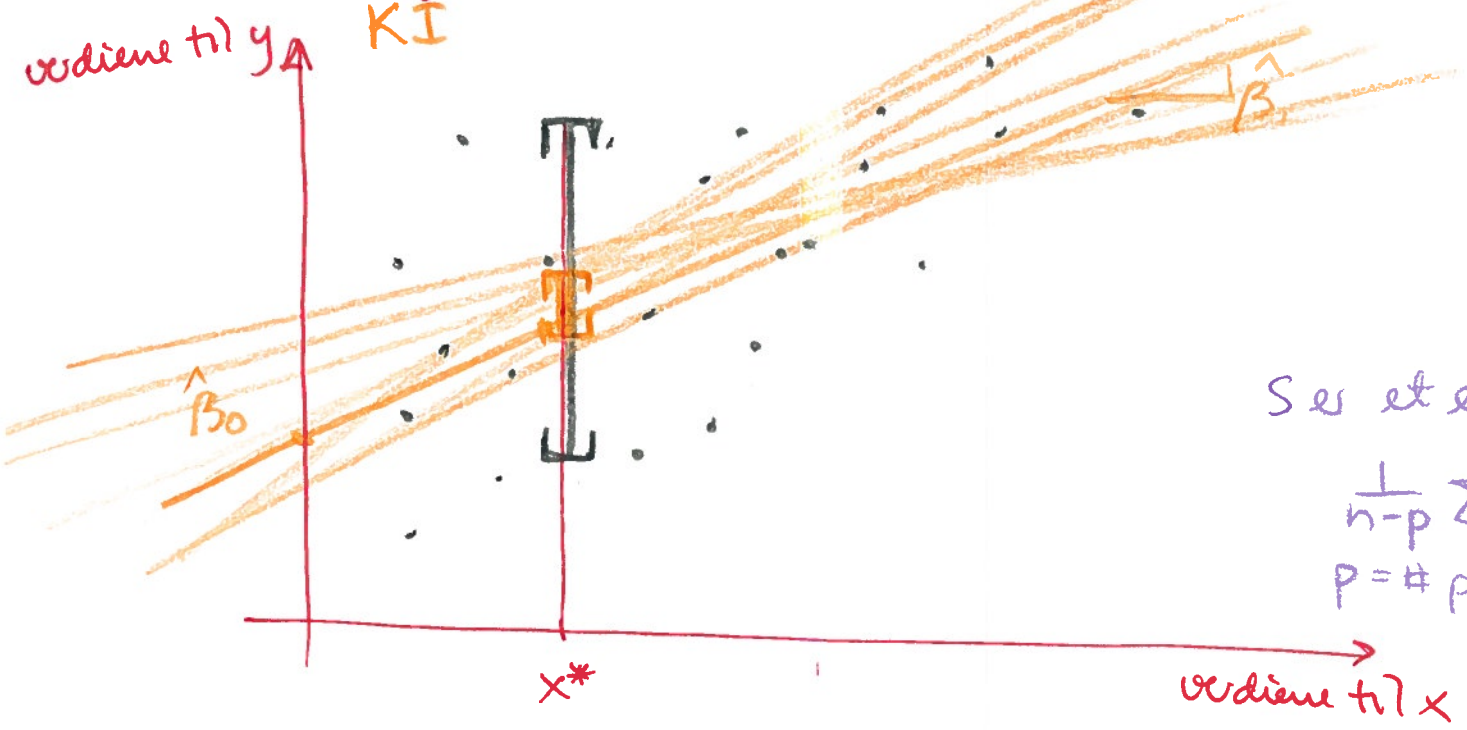
Konfidensintervall for mean response μ_Y når $x = x^* = 40$

$$\mu_Y = E(Y|x^*) = E(\beta_0 + \beta_1 x^*) = \beta_0 + \beta_1 x^*$$

$\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 x^*$ → dette er forventet verdi for linja når $x = x^*$

$$S.E.(\hat{\mu}_Y) = S \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\Sigma_i (x_i - \bar{x})^2}}$$ } dette er SE for linja

Hvor forventer vi at linja vil være?



Prediksjonsintervall for ny (enkelt-)verdi y når $x = x^* = 40$

Predikert verdi er den samme, altså

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ → dette er forventet verdi for y når $x = x^*$

, men S.E er ulik:

$$S.E(\hat{y}) = S \cdot \sqrt{\frac{1}{n} + 1 + \frac{(x^* - \bar{x})^2}{\Sigma (x_i - \bar{x})^2}}$$ } SE for enkelt-verdi av y.

Hvor forventer vi at en ny verdi av y vil være?

pred. intervall

S er et estimat for σ:

$$\frac{1}{n-p} \Sigma_i (y_i - \hat{y}_i)^2$$

p = # parameter. Her: p = 2

95% KI for μ_Y :

$\hat{\mu}_Y \pm t_{n-2} \cdot SE(\hat{\mu}_Y)$; $SE_{\hat{\mu}_Y} = s\cdot\sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

95% pred int. for y :

$\hat{y} \pm t_{n-2} \cdot SE(\hat{y})$; $SE(\hat{y}) = s\sqrt{\frac{1}{n} + 1 + \frac{(x^*-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$

Begge intervallene er basert på den foregående analysen R> lm(IBI ~ Area) (y ~ x)

$\hat{\mu}_Y$ og $\hat{y}$ er begge $\hat{\beta}_0 + \hat{\beta}_1 x^* = 52.9230 + 0.4602 \cdot 40 = 71.331$

t-fordelingen vi skal bruke er $t_{n-2} = t_{49-2} = t_{47}$, og det nærmeste vi kommer i Table D er t_{50}, som har $t_{0.025,50} = 2.009$

s er estimat for σ, og kalles "Residual standard error" i R-utskriften : $s = 17.79$

$\bar{x}$ og $\sum(x_i-\bar{x})^2$ henter vi fra den deskriptive statistikken for Area : $\bar{x} = 28.28571$

$sd(x) = 17.71417 = \sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2} \rightarrow \sum(x_i-\bar{x})^2 = (n-1)\cdot(sd(x))^2 = 48 \cdot 17.71417^2 = 15062.0$

$\rightarrow SE\hat{\mu}_Y = 17.79 \cdot \sqrt{\frac{1}{49} + \frac{(40-28.28571)^2}{15062}} = 3.0565$

$\rightarrow SE\,\hat{y} = 17.79 \cdot \sqrt{\frac{1}{49} + 1 + \frac{(40-28.28571)^2}{15062}} = 18.05066$

95% KI for μ_Y når $x = 40$:

$71.331 \pm 2.009 \cdot 3.0565 \rightarrow [65.2, 77.5]$

95% prediksjonsintervall for y når $x = 40$:

$71.331 \pm 2.009 \cdot 18.05066 \rightarrow [35.1, 107.6]$

(med forbehold om feilregning. Resonnementer er OK.)

OBS:

Dere må repetere
Begrepet conformdes.

Hvorfor jeg bruker store bokstaver?
Boka gjør det ikke så bra,
så derfor får dere en "tutorial"
på de neste sidene.

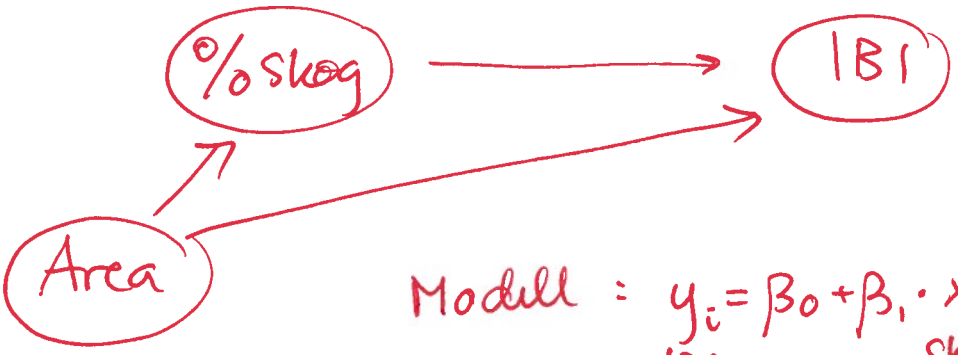
Sees på fredag.

Anta nå at vi visste mer om dette med vannkvalitet, nedfallsområdet til elver, og skogvekst. Da kunne vi brukt vår ekspertkunnskap til å formulere en modell for vannkvaliteten, der tres/skog ble brukt som en forklaringsvariabel, jfr analysen i 10.33. Altså,

Et estimat for denne sammenhengen er $\hat{\beta}_1 = 0.153$, p-verdi 0.06

Vi kunne også, dersom kunnskapen om feltet gjorde oss i stand til det, vurdere om denne sammenhengen ble feilaktig estimert pga konfunderende variables, eller det boka kaller "lurking variables".

Area er en slik. Jo større område med vannfylsig, jo større prosent tres (kan ekspertene si), og jo større område vann, jo høyere IBI (som er noe av det vi lurer på). Altså. Ekspertkunnskap må til for å vurdere om en (eller flere) variabler er konfunderende til den effekten vi ønsker å estimere. Her er det grunn til å tro det, og da ser det slik ut:



Modell: $y_{i\,(IBI)} = \beta_0 + \beta_1 \cdot x_{1i\,(Skog)} + \beta_2 \cdot x_{2i\,(Area)} + \varepsilon_i$ $\quad \varepsilon_i \sim N(0,\sigma)$

I så fall er Area en konfounder for sammenhengen mellom % skog og IBI, og må inn i analysen.

Justert estimat for β_1 er nå

$\hat{\beta}_1 = 0.23$, som er klart signifikant på nivå 0.05.

Konfunderende variable kan både svekke og forsterke effektestimater, og er viktige å ta hensyn til i observasjonelle studier.

Konfidensintervall for den justerte β_1:

$$\hat{\beta}_1 \pm t_{n-3} \cdot SE(\hat{\beta}_1)$$

t_{n-3}: OBS

```
> summary(lm(IBI~Forest))

Call:
lm(formula = IBI ~ Forest)

Residuals:
    Min      1Q  Median      3Q     Max
-35.961 -11.186   4.508  13.021  28.633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.90725    4.03957  14.830   <2e-16 ***
Forest       0.15313    0.07972   1.921   0.0608 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.79 on 47 degrees of freedom
Multiple R-squared:  0.07278,  Adjusted R-squared:  0.05305
F-statistic: 3.689 on 1 and 47 DF,  p-value: 0.06084

> summary(lm(IBI~Forest+Area))

Call:
lm(formula = IBI ~ Forest + Area)

Residuals:
    Min      1Q  Median      3Q     Max
-31.708  -8.636   3.206  10.677  30.596

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.62924    5.46144   7.439 2.01e-09 ***
Forest       0.23367    0.06944   3.365  0.00155 **
Area         0.56940    0.12624   4.511 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.97 on 46 degrees of freedom
Multiple R-squared:  0.3571,   Adjusted R-squared:  0.3292
F-statistic: 12.78 on 2 and 46 DF,  p-value: 3.864e-05
```