

Oppgaver i R

Det anbefales at studentene på egenhånd utfører oppgave for å få økt trening i å bruke R. Oppgavene blir gjennomgått på forelesningene og danner basis for den kunnskapen som dere trenger for bl.a. å kunne skrive resultatene fra laboppgavene.

Innholdsfortegnelse:

Oppgave 1 Varianse og lineære modeller	1
Appendiks oppgave 1	11
Oppgave 2 Regresjon	23
Appendiks oppgave 2	34
Oppgave 3 Blokkdesign	44
Oppgave 4 Tellinger	47
Appendiks oppgave 4	49
Oppgave 5: Ikke-lineær regresjon	51
Oppgave 6: Tidsserieanalyse av klimadata fra Blindern	53

Oppgave 1 Varianse og lineære modeller

Eksperiment: Du arbeider i et firma som skal selge nye gjødselblandinger brukt i dyrking av korn. Det utføres et feltforsøk hvor effekten av tre gjødseltyper (FERTIL 1, 2 og 3) på avlingen (YIELD) undersøkes. Det er ti forsøksfelt (ekssperimentelle enheter) for hver gjødseltype og de 30 feltene høstes og avlingen i tonn for hvert felt bestemmes. Er det signifikante forskjeller mellom effekten av gjødseltypene FERTIL 1, 2 og 3 ? Hvilken gjødseltype vil du anbefale som gir størst avling ?

Vi skal undersøke i hvilken grad FERTIL kan forklare variasjonen i datasettet.

Datasett fra: Grafen, A. & Hails, R.: *Modern statistics for the life sciences*. Oxford University Press 2003

Modell: $YIELD \sim FERTIL + error$

Responsvariabel \sim Forklaringsvariabel, og \sim betyr modellert av.

```
#Last inn datafil oppg1.txt inn i R vi internett eller lokal
server:
oppg1<-
read.table("http://www.mn.uio.no/bio/tjenester/kunnskap/plantefys/matematik
k/oppg1.txt",header=T)
eller:
oppg1<-read.table("//platon/bio-kurs/bio2150/R/oppg1.txt",header=T)
attach(oppg1)
names(oppg1) #angir navnene til et objekt
[1] "FERTIL" "YIELD"
```

Først må man sjekke datasettet. Oppsummering datasettet angir minimums- og maksimumsverdi, gjennomsnitt og median, 1. og 3. kvartil.

```
summary(oppg1)
```

```
FERTIL      YIELD
F1:10  Min.    :3.070
F2:10  1st Qu.:3.828
F3:10  Median  :4.540
       Mean   :4.644
       3rd Qu.:5.048
       Max.   :7.140
```

Man kan se på strukturen til datasettet med

```
str(oppg1).
```

Vi kan se hvilke klasser et objekt har:

```
sapply(oppg1, class)
```

```
FERTIL      YIELD
"factor" "numeric"
```

FERTIL er en faktorvariabel (prediktor, forklaringsvariabel, uavhengig variabel) med tre nivåer med ialt 30 eksperimentelle enheter (forsøksenheter). YIELD er responsvariabel (avhengig variabel).

```
dim(oppg1)
```

```
[1] 30  2
```

De 11 første dataene i datasettet, men generelt gir **head(oppg1)** de 6 første verdiene i objekt.

```
oppg1[1:11,]
```

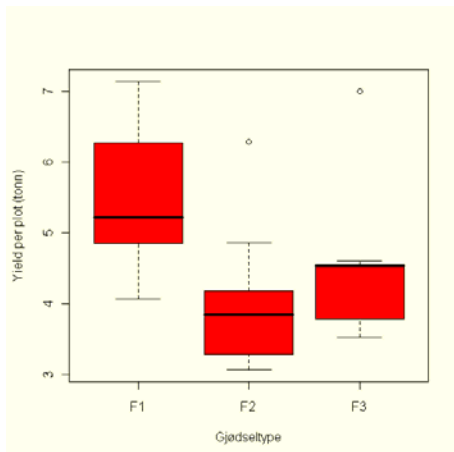
```
FERTIL YIELD
1      F1  6.27
2      F1  5.36
3      F1  6.39
4      F1  4.85
5      F1  5.99
6      F1  7.14
7      F1  5.08
8      F1  4.07
9      F1  4.35
10     F1  4.95
11     F2  3.07
```

Generelt kan vi se hva som står i en kolonne ved **oppg1\$YIELD** eller **oppg1[["YIELD"]]** eller **oppg1[, "YIELD"]** som gir samme resultat. \$ brukes for å angi et element i et objekt, og klammeparentes [] brukes for å angi hvilke elementer man har i en vektor eller matrise.

```
#Et boksplot av data
```

```
par(bg="lightyellow")
```

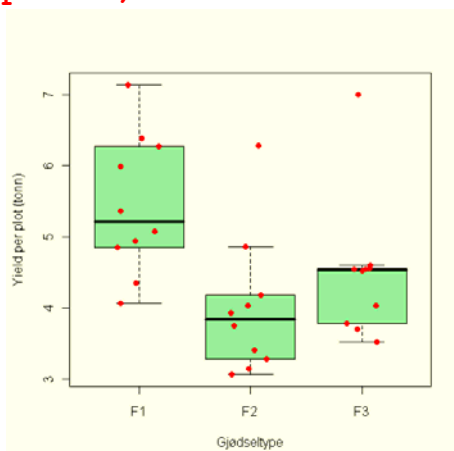
```
plot(FERTIL, YIELD, col="red", xlab="Gjødseltype", ylab="Yield per plot (tonn)", data=oppg1)
```



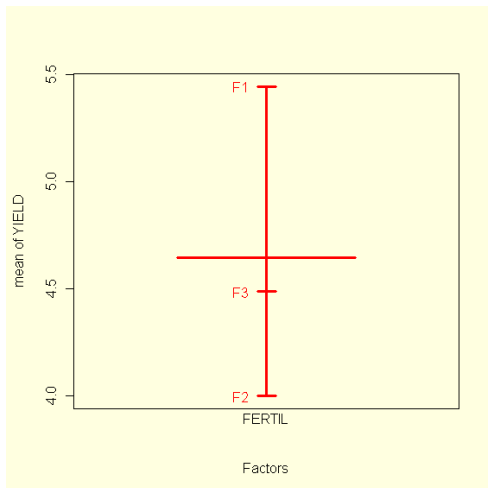
Boksplot viser median (svart strek midt i boksen=50% persentil). Boksen omfatter 1. og 3. kvartil (interkvartilområde=25%-75% persentil). Værhår (vertikal prikket strek) omfatter måleverdier som ligger 1.5 ganger interkvartilavstanden. Utliggere er observasjoner som ligger utenfor 1.5 ganger interkvartilavstanden. For en normalfordeling vil medianen ligge omtrent midt i boksen og værhårene går omtrent like langt ut på begge sider av boksen. Kommandoen **notch=T** (hakk, innskjæring, Notchplot) viser et hakk eller innsnitt for 95% konfidensintervallet rundt median, og hvis de to innsnittene ikke overlapper så er medianverdiene signifikant forskjellig. Hvis innhakkene ikke overlapper hverandre er medianene signifikant forskjellig på 5% nivå.

Her er en annen måte hvor datapunktene vises samtidig:

```
par(bg="lightyellow")
boxplot(YIELD~FERTIL,col="lightgreen",xlab="Gjødseltype",ylab="Yield per plot (tonn)",
outline=F,data=oppg1)
#bruker jitter for å spre punktene
points(jitter(as.numeric(oppg1$FERTIL)),oppg1$YIELD,col="red",
pch=16)
```



```
#Ser på design til eksperimentet
plot.design (oppg1,col=2,lwd=3)
```



```
#Sjekking av data, finn antall måledata
```

```
n<-length(YIELD)
```

```
n
```

```
[1] 30
```

```
#Finn antall nivåer av faktoren FERTIL
```

```
levels(FERTIL)
```

```
[1] "F1" "F2" "F3"
```

I vårt tilfelle hvor faktorene er angitt med bokstavkombinasjoner er det greit, men i noen tilfeller er faktorene kodet som tall. Da må man gi beskjed til R om at variabelen er en faktor ved å angi variabelnavnet på nytt som en faktor for eksempel

```
FERTIL<-factor(FERTIL).
```

```
#Er FERTIL en faktor ?
```

```
is.factor(FERTIL)
```

```
[1] TRUE
```

```
#Finner forventet verdi (stormiddeltallet (M)) til YIELD
```

```
mean(YIELD)
```

```
[1] 4.643667
```

```
#Finn variansen til YIELD
```

```
var(YIELD)
```

```
[1] 1.256721
```

```
#Gi et sammendrag av dataene
```

```
summary(oppg1)
```

```
FERTIL      YIELD
F1:10  Min.   :3.070
F2:10  1st Qu.:3.828
F3:10  Median :4.540
       Mean  :4.644
       3rd Qu.:5.048
       Max.  :7.140
```

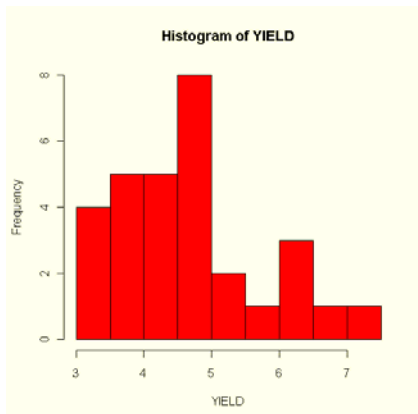
```
#Spredningen på dataene største og minste
```

```
range(YIELD)
```

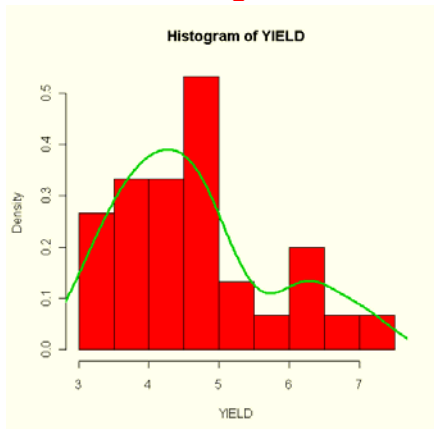
```
[1] 3.07 7.14
```

```
#Lag et histogram av YIELD
```

```
hist(YIELD,col="red")
```

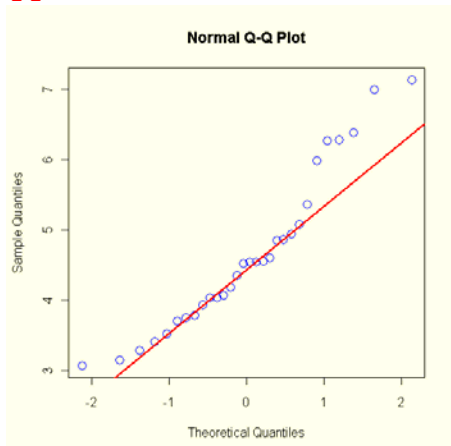


```
#histogram sannsynlighetstetthet
hist(YIELD,prob=T,col=2)
lines(density(YIELD),col=3,lwd=3)
```



Allerede nå kan vi se avvik skjevfordelte data

```
#Sjekk normalfordeling av YIELD
qqnorm(YIELD,col=4,cex=1.5)
qqline(YIELD,col=2,lwd=2)
```



Vi ser at datapunktene ligger på en bueform og det gjenspeiles også i histogrammet som skjevfordeling (skew).

ANOVA

```
#Anovamodell
```

```
summary(aov(YIELD~FERTIL))
```

```
      Df Sum Sq Mean Sq F value    Pr(>F)
FERTIL  2 10.8227   5.4114   5.7024 0.008594 **
Residuals 27 25.6221   0.9490
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi har for kvadratsummene: SST= SSF+SSE

36.4449= 10.8227 + 25.6221

Middelkvadratsummen (Mean Sq) er kvadratsummen (Sum Sq) dividert på antall frihetsgrader.

F-verdien er forholdet mellom middelkvadratene:

F=FMS/EMS= 5.4114/0.9490= 5.7024

Er denne F-verdien signifikant? Vi sammenligner med den kritiske verdien av **F** for p=0.05 dvs. 0.95, og henholdsvis 2 og 27 frihetsgrader. Nullhypotesen (H_0) om at det ikke er noen forskjell forkastes. Konklusjon: det er signifikante forskjeller i effekten av gjødseltype, gjødseltype forklarer variasjon i biomasse (YIELD). Vi kan også bruke:

```
anova(lm(YIELD~FERTIL))
```

```
#kritisk verdi for F
```

```
qf(0.95,2,27)
```

```
[1] 3.354131
```

```
#Sannsynligheten for å få F-verdien vi har funnet
```

```
1-pf(5.7024,2,27)
```

```
[1] 0.00859422
```

Sannsynligheten for å få en slik verdi vi fikk hvis forventede verdier (middeltallene, gjennomsnittene) var like er ca. 0.8%, konklusjon det er signifikant forskjell mellom gjødseltypene.

Lineær regresjonsmodell

Den lineære modellen er av type:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

hvor β_0 er intercept og β_1 er stigningskoeffisient i en regresjonsbasert tilnærming. Vi ønsker å finne konfidensintervall for β_0 og β_1 . Konfidensintervall til parameterestimatene er en god måte å bestemme størrelsen på en effekt. Konfidensintervall er nyttige selv om nullhypotesen ikke blir forkastet. Vi antar at $\varepsilon \sim N(0, \sigma^2)$ dvs. feilene (residualene) er uavhengige, har lik varians og er normalfordelt. Modellen er bare en approksimasjon av underliggende realiteter. Vi bruker en-veis ANOVA og fordeler totalvariansen i respons på faktor og error.

Generelt betyr `lm(YIELD~FERTIL)`, tilpass en lineære modell av YIELD som funksjon av FERTIL.

Den enkleste modellen er **nullmodellen**. Den sier noe om variasjonen i biomasse, men helt uavhengig av gjødseltype.

```
#Nullmodell
```

```
mod0<-lm(YIELD~1)
```

summary(mod0)

```
Call:
lm(formula = YIELD ~ 1)
Residuals:
    Min       1Q   Median       3Q      Max
-1.5737 -0.8162 -0.1037  0.4038  2.4963
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.6437      0.2047   22.69  <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.121 on 29 degrees of freedom
```

Vi ser at nullmodellen gir intercept (β_0) som er lik forventet verdi for alle YIELD, altså stormiddeltallet ("grand mean")

En ANOVA av nullmodellen:

anova(mod0)

Analysis of Variance Table

Response: YIELD

```
            Df Sum Sq Mean Sq F value Pr(>F)
Residuals  29 36.445   1.257
```

Vi ser at vi får bare en kvadratsum, og det er totalkvadratsummen SST (også kalt SSY nedenfor)

#Stor modell

```
mod1<-lm(YIELD~FERTIL)
```

summary(mod1)

```
Call:
aov(formula = YIELD ~ FERTIL)
Residuals:
    Min       1Q   Median       3Q      Max
-1.3750 -0.6715 -0.0720  0.1740  2.5130
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4450      0.3081  17.676 2.28e-16 ***
FERTILF2     -1.4460      0.4357  -3.319  0.00259 **
FERTILF3     -0.9580      0.4357  -2.199  0.03663 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.9742 on 27 degrees of freedom
Multiple R-Squared:  0.297,    Adjusted R-squared:  0.2449
F-statistic: 5.702 on 2 and 27 DF,  p-value: 0.008594
```

coef(mod1) #bare koeffisientverdiene i modellen

```
(Intercept)  FERTILF2  FERTILF3
    5.445    -1.446    -0.958
```

Vi ser at intercept (β_0) er lik forventet verdi (middeltallet, gjennomsnitt) for F1 og at estimatene for F2 og F3 er lik avvikene fra intercept. FERTILF1 blir brukt som referanse og settes lik 0. Det faktornivået som kommer først i alfabetet blir satt lik intercept, altså her F1. Fra p-verdi og F-statistic konkluderer vi med at det er en forskjell mellom gruppene FERTIL 1,2 og 3. Ved å dividere estimatet på

standardfeilen får man **t-verdien**. Vi kan se bort fra fortegnet på t-verdien siden det er de absolutte forskjellene vi ser på. t-observator (den kritiske verdien for t):

```
#kritisk verdi for t
```

```
qt(0.975,27)
```

```
[1] 2.051831
```

Siden vår tabellverdi er større enn den kritiske verdien for t forkastes nullhypotesen om at det ikke er noen forskjeller mellom F1, F2 og F3. Det er imidlertid bedre å se på sannsynligheten, vi bruker kommandoen **qt** og ganger med 2 siden vi bruker en **to-halet test**:

```
#Sannsynligheten for å få t-verdier iflg. tabell
```

```
2*pt(-3.319,27)
```

```
[1] 0.002592942
```

```
2*pt(-2.199,27)
```

```
[1] 0.03662561
```

Dette er sannsynlighetene du finner i tabellen.

R2 viser hvor mye av variasjonen som forklares av FERTIL, dvs. SSF/SST, dvs. 29.7% av variasjonen

```
10.8227/36.445
```

```
[1] 0.2969598
```

	FERTIL	
	0.0000	1
<i>YIELD</i> ~ 5.445 +	-1.446	2 + <i>error</i>
	-0.9580	3

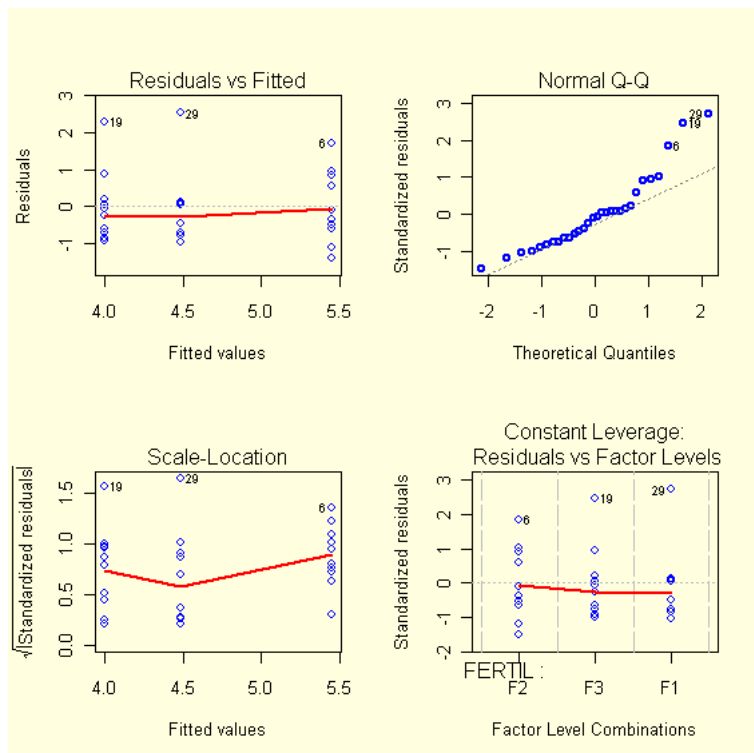
Regner vi ut dette får vi:

	FERTIL	
	5.4450	1
<i>YIELD</i> ~ 3.9990		2 + <i>error</i>
	4.4870	3

```
#Plotter modellen
```

```
par(mfrow=c(2,2),bg="lightyellow")
```

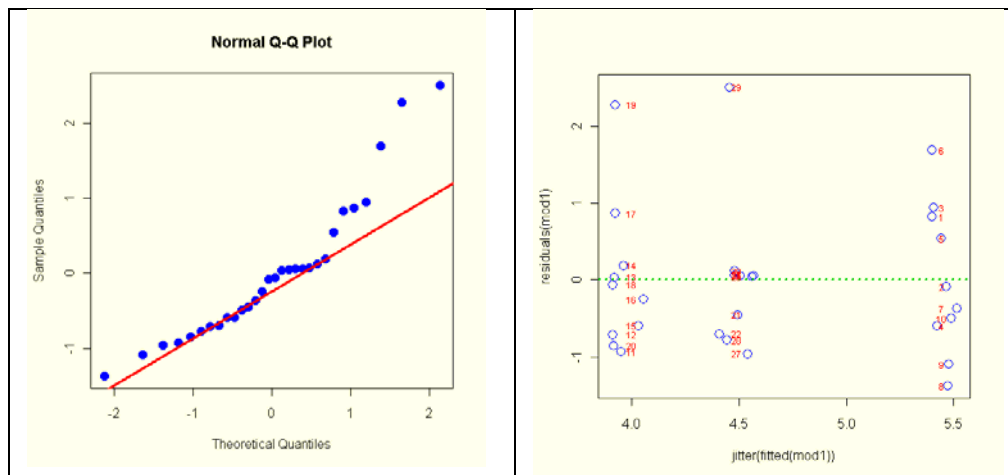
```
plot(mod1,col="blue",lwd=2)
```

Diagnostikk for ANOVA modellen

Linære modeller forutsetter residualene (feilene) er normalfordelte, uavhengige og har samme (konstant) varianse. Vi må sjekke for ikke-konstant varianse (**heteroskedastitet**). Må plote residualer og tilpassete verdier og lage et QQ-plot av residualene, og oppfylles forutsetningene vil punktene bli liggende på omtrent en rett linje. Når data prediktert av modellen plottes mot residualene skal dette gi punkter som ligger tilfeldig uten noe mønster på hver side av den horisontale nulllinjen.

```
qqnorm(residuals(mod1), col=4, cex=1.5, pch=16)
qqline(residuals(mod1), col=2, lwd=3)
plot(jitter(fitted(mod1)), residuals(mod1), col=4, cex=1.5)
abline(h=0, lty=3, col=3, lwd=2)
text(fitted(mod1), residuals(mod1), col=2,
      labels=rownames(oppg1), cex=0.7) #nummer på punktene
```



Vi ser at residualene ikke er helt normalfordelte, det blir en "bananform" noe som gjenspeiles i den tidligere kvantil-kvantil-testen.

Designmatrisen viser kodingen:

model.matrix(mod1)

```
(Intercept) FERTILF2 FERTILF3
1           1           0           0
2           1           0           0
3           1           0           0
4           1           0           0
5           1           0           0
6           1           0           0
7           1           0           0
8           1           0           0
9           1           0           0
10          1           0           0
11          1           1           0
12          1           1           0
13          1           1           0
14          1           1           0
15          1           1           0
16          1           1           0
17          1           1           0
18          1           1           0
19          1           1           0
20          1           1           0
21          1           0           1
22          1           0           1
23          1           0           1
24          1           0           1
25          1           0           1
26          1           0           1
27          1           0           1
28          1           0           1
29          1           0           1
30          1           0           1
```

```
attr("assign")
```

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

```
attr("contrasts")
```

```
attr("contrasts")$FERTIL
```

```
[1] "contr.treatment"
```

Vi kan lage en modell uten intercept og ser da at vi får forventede verdier direkte (gjennomsnittsverdier, middelværdier). R^2 blir ikke riktig siden vi har utelatt intercept. F-test tilsvare nullhypotesen om at forventet gjennomsnittsrespons er lik 0. Dette er uinteressant.

mod2<-lm(YIELD~FERTIL-1)

summary(mod2)

```

Call:
lm(formula = YIELD ~ FERTIL - 1)

Residuals:
    Min       1Q   Median       3Q      Max
-1.3750 -0.6715 -0.0720  0.1740  2.5130

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
FERTILF1    5.4450     0.3081   17.68 2.28e-16 ***
FERTILF2    3.9990     0.3081   12.98 4.02e-13 ***
FERTILF3    4.4870     0.3081   14.57 2.62e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9742 on 27 degrees of freedom
Multiple R-squared:  0.9625,    Adjusted R-squared:  0.9583
F-statistic:  231 on 3 and 27 DF,  p-value: < 2.2e-16

```

Vi kan se hvilke navn på objekter vi får i modellen:

```
names(mod1)
```

```

 [1] "coefficients" "residuals"      "effects"        "rank"
 [5] "fitted.values" "assign"         "qr"            "df.residual"
 [9] "contrasts"    "xlevels"       "call"          "terms"
[13] "model"

```

Appendiks oppgave 1

Appendiks1 gir mer utfyllende informasjon om oppgave 1.

En annen mulighet til å teste for normalitet er **Shapiro-Wilk test**:

```
shapiro.test(YIELD)
```

```
Shapiro-Wilk normality test
```

```
data: YIELD
```

```
W = 0.9272, p-value = 0.04148
```

Dette gjenspeiler det vi finner i QQ-plottet, om at avvikene fra normalfordelingen er i grenseland for det vi kan tillate. Hvis $p < 0.05$ er ikke dataene normalfordelte, og p må være > 0.05 for at vi skal konkludere med normalfordeling.

Homogen varianse er en forutsetning for anova og regresjon og dette kan testes i en **Bartlett test** eller **fligner.test**. Hvis vi bare har to prøver hvor variansene skal sammenlignes bruker vi en Fishers F-test (**var.test**).

```
bartlett.test(YIELD~FERTIL)
```

```
Bartlett test of homogeneity of variances
```

```
data: YIELD by FERTIL
```

```
Bartlett's K-squared = 2e-04, df = 2, p-value = 1
```

Vi kan konkludere med at variansen er homogen, hvis $p < 0.05$ hadde vi ikke hatt homogen og konstant varianse. Stor p -verdi og vi konkluderer at det er ingen evidens for ikke-konstant varianse

Variansanalyse tester om det er signifikante forskjeller mellom prøver, men angir ikke hvem som er forskjellig fra hvem. Det finnes imidlertid en rekke post-hoc tester som gjør

dette bl.a. Tukey's Honestly Significant Difference
(**TukeyHSD()**)

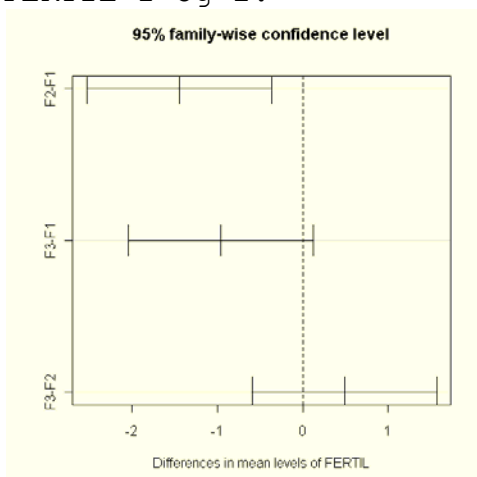
```
TukeyHSD(aov(YIELD~FERTIL))
```

```
Tukey multiple comparisons of means
 95% family-wise confidence level
Fit: aov(formula = YIELD ~ FERTIL)
$FERTIL
      diff      lwr      upr    p adj
F2-F1 -1.446 -2.5261662 -0.3658338 0.0070788
F3-F1 -0.958 -2.0381662  0.1221662 0.0894812
F3-F2  0.488 -0.5921662  1.5681662 0.5102335
```

```
par(bg="lightyellow")
```

```
plot(TukeyHSD(aov(YIELD~FERTIL)))
```

Hvis horisontale streker krysser striplet linje er det ingen signifikant forskjell. Det er signifikant forskjell mellom FERTIL 1 og 2.



Det går også an å lage seg en funksjon av x funksjon(x) som vi kaller varianse og deretter setter vi inn YIELD i stedet for x i funksjonen og vi ser at vi får samme verdi for variansen som ovenfor

```
varianse<-function(x) {sum((x-mean(x))^2/(length(x)-1))}  
varianse(YIELD)  
[1] 1.256721
```

Kommandoen **tapply** er nyttig for å finne forventet verdi (middeltall, gjennomsnitt) eller varianse for datasettet:

```
tapply(YIELD, FERTIL, mean)
```

```
      F1      F2      F3  
5.445 3.999 4.487
```

```
tapply(YIELD, FERTIL, var)
```

```
      F1      F2      F3  
0.9525389 0.9442989 0.9500678
```

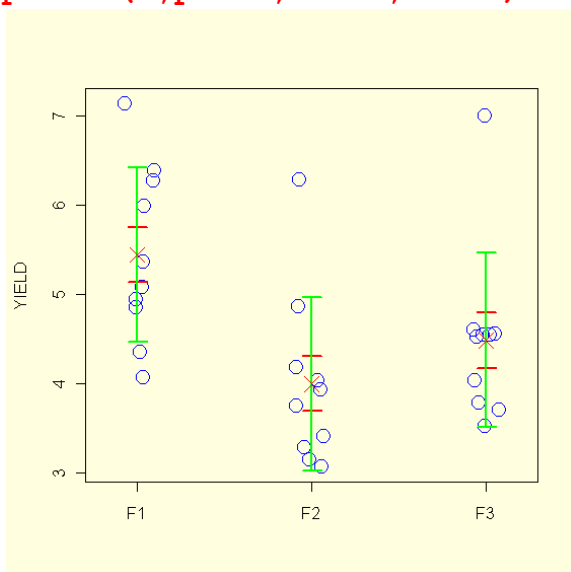
Vi kan sjekke at vi har like mange tall for hver FERTIL:

```
tapply(YIELD, FERTIL, length)
```

```
      F1  F2  F3  
10 10 10
```

Vi kan plote både standardavvik (grønn) og standardfeil (rød) (Dahlgaard, P.: *Introductory statistics with R*. 2e. Springer 2008 s. 135):

```
stripchart(YIELD~FERTIL,method="jitter",vert=T,pch=1,cex=2,col=4)
m<-tapply(YIELD,FERTIL,mean)
s<-tapply(YIELD,FERTIL,sd)
n<-tapply(YIELD,FERTIL,length)
se<-s/sqrt(n)
arrows(1:3,m+se,1:3,m-se,angle=90,
code=3,length=0.1,lwd=2,col=2)
arrows(1:3,m+s,1:3,m-s,angle=90,
code=3,length=0.1,lwd=2,col=3)
points(m,pch=4,cex=2,col=2)
```



Her vises en annen måte å finne middeltallene for F1, F2 og F3:

```
mean(YIELD[FERTIL=="F1"])
```

```
[1] 5.445
```

```
mean(YIELD[FERTIL=="F2"])
```

```
[1] 3.999
```

```
mean(YIELD[FERTIL=="F3"])
```

```
[1] 4.487
```

Vi kan for eksempel se på estimatene til de to første tallene i koeffisientmodellen:

```
mean(YIELD[FERTIL=="F2"])-mean(YIELD[FERTIL=="F1"])
```

```
[1] -1.446
```

Tilsvarende for det siste estimatet:

```
mean(YIELD[FERTIL=="F3"])-mean(YIELD[FERTIL=="F1"])
```

```
[1] -0.958
```

Tilsvarende kan man beregne variansen:

```
varF1<-var(YIELD[FERTIL=="F1"])
```

```

varF1
[1] 0.9525389
varF2<-var (YIELD[FERTIL=="F2"])
varF2
[1] 0.9442989
varF3<-var (YIELD[FERTIL=="F3"])
varF3
[1] 0.9500678
Hvordan finner man tallene i ANOVA-tabellen ?
Finn avvik fra stormiddeltallet (MY):
YIELD-mean (YIELD)
[1] 1.62633333 0.71633333 1.74633333 0.20633333 1.34633333 2.49633333
[7] 0.43633333 -0.57366667 -0.29366667 0.30633333 -1.57366667 -1.35366667
[13] -0.60366667 -0.45366667 -1.23366667 -0.89366667 0.22633333 -0.70366667
[19] 1.63633333 -1.49366667 -0.60366667 -0.85366667 -0.08366667 -0.09366667
[25] -0.09366667 -0.11366667 -1.11366667 -0.93366667 2.35633333 -0.03366667
Kvadrer avvikene fra stormiddeltallet:
(YIELD-mean (YIELD) ) ^2
[1] 2.644960111 0.513133444 3.049680111 0.042573444 1.812613444 6.231680111
[7] 0.190386778 0.329093444 0.086240111 0.093840111 2.476426778 1.832413444
[13] 0.364413444 0.205813444 1.521933444 0.798640111 0.051226778 0.495146778
[19] 2.677586778 2.231040111 0.364413444 0.728746778 0.007000111 0.008773444
[25] 0.008773444 0.012920111 1.240253444 0.871733444 5.552306778 0.001133444

```

Summer de kvadrerte avvikene fra stormiddeltallet og finn SSY:

```

SSY<-sum ( (YIELD-mean (YIELD) ) ^2)
SSY
[1] 36.4449
Finn avvikene (MF) fra stormiddeltallet til de enkelte
middeltallene for de 3 gjødseltypene:
MF<-tapply (YIELD, FERTIL, mean) -mean (YIELD)
MF
      F1      F2      F3
0.8013333 -0.6446667 -0.1566667

```

Vi deler opp YIELD fordelt på de 3 gjødseltypene:

```

F1<-YIELD[1:10];F1
[1] 6.27 5.36 6.39 4.85 5.99 7.14 5.08 4.07 4.35 4.95
F2<-YIELD[11:20];F2
[1] 3.07 3.29 4.04 4.19 3.41 3.75 4.87 3.94 6.28 3.15
F3<-YIELD[21:30];F3
[1] 4.04 3.79 4.56 4.55 4.55 4.53 3.53 3.71 7.00 4.61

```

En annen måte å plukke ut en del av et datasett er å bruke kommandoen **subset**.

```

F1<-subset (YIELD, FERTIL=="F1") #plukker ut F1
F2<-subset (YIELD, FERTIL=="F2") #plukker ut F2
F3<-subset (YIELD, FERTIL=="F3") #plukker ut F3

```

Beregn avvik av YIELD fra gjennomsnittsverdiene for de 3 gjødseltypene: FY, her fordelt på F1, F2 og F3

```

FY1<-F1-mean (F1) ;FY1
[1] 0.825 -0.085 0.945 -0.595 0.545 1.695 -0.365 -1.375 -1.095 -0.495
FY2<-F2-mean (F2) ;FY2
[1] -0.929 -0.709 0.041 0.191 -0.589 -0.249 0.871 -0.059 2.281 -0.849
FY3<-F3-mean (F3) ;FY3
[1] -0.447 -0.697 0.073 0.063 0.063 0.043 -0.957 -0.777 2.513 0.123

```

Beregn feilkvadratsummen SSE (kobler sammen vektorene FY1, FY2, FY3):

```
FY<-c(FY1, FY2, FY3) ;FY
```

```
[1] 0.825 -0.085 0.945 -0.595 0.545 1.695 -0.365 -1.375 -1.095 -0.495
[11] -0.929 -0.709 0.041 0.191 -0.589 -0.249 0.871 -0.059 2.281 -0.849
[21] -0.447 -0.697 0.073 0.063 0.063 0.043 -0.957 -0.777 2.513 0.123
```

```
SSE<-sum(FY^2) ;SSE
```

```
[1] 25.62215
```

Hvis vi nå samler alle dataene vi har fått i en tabell:

Plot	FERTIL	M	F	Y	MY	MF	FY
1	F1	4.6436	5.445	6.27	1.63	0.80	0.82
2	F1	4.6436	5.445	5.36	0.72	0.80	-0.09
3	F1	4.6436	5.445	6.39	1.75	0.80	0.94
4	F1	4.6436	5.445	4.85	0.21	0.80	-0.60
5	F1	4.6436	5.445	5.99	1.35	0.80	0.55
6	F1	4.6436	5.445	7.14	2.50	0.80	1.70
7	F1	4.6436	5.445	5.08	0.44	0.80	-0.37
8	F1	4.6436	5.445	4.07	-0.57	0.80	-1.38
9	F1	4.6436	5.445	4.35	-0.29	0.80	-1.10
10	F1	4.6436	5.445	4.95	0.31	0.80	-0.50
11	F2	4.6436	3.999	3.07	-1.57	-0.64	-0.93
12	F2	4.6436	3.999	3.29	-1.35	-0.64	-0.71
13	F2	4.6436	3.999	4.04	-0.60	-0.64	0.04
14	F2	4.6436	3.999	4.19	-0.45	-0.64	0.19
15	F2	4.6436	3.999	3.41	-1.23	-0.64	-0.59
16	F2	4.6436	3.999	3.75	-0.89	-0.64	-0.25
17	F2	4.6436	3.999	4.87	0.23	-0.64	0.87
18	F2	4.6436	3.999	3.94	-0.70	-0.64	-0.06
19	F2	4.6436	3.999	6.28	1.64	-0.64	2.28
20	F2	4.6436	3.999	3.15	-1.49	-0.64	-0.85
21	F3	4.6436	4.487	4.04	-0.60	-0.16	-0.45
22	F3	4.6436	4.487	3.79	-0.85	-0.16	-0.70
23	F3	4.6436	4.487	4.56	-0.08	-0.16	0.07
24	F3	4.6436	4.487	4.55	-0.09	-0.16	0.06
25	F3	4.6436	4.487	4.55	-0.09	-0.16	0.06
26	F3	4.6436	4.487	4.53	-0.11	-0.16	0.04
27	F3	4.6436	4.487	3.53	-1.11	-0.16	-0.96
28	F3	4.6436	4.487	3.71	-0.93	-0.16	-0.78
29	F3	4.6436	4.487	7.00	2.36	-0.16	2.51
30	F3	4.6436	4.487	4.61	-0.03	-0.16	0.12
d.f.		1	3	30	29	2	27
SS					36.44	10.82	25.62

$SSY (36.44) = SSF (10.82) + SSE (25.62)$

Her er et eksempel på hvordan datasettet kan splittes på de enkelte faktorene. Den lager en liste med vektorer med basis til nivåene til en faktor:

```
sy<-split(YIELD, FERTIL) ; sy
```

```
$F1
```

```
[1] 6.27 5.36 6.39 4.85 5.99 7.14 5.08 4.07 4.35 4.95
```

```
$F2
[1] 3.07 3.29 4.04 4.19 3.41 3.75 4.87 3.94 6.28 3.15
```

```
$F3
[1] 4.04 3.79 4.56 4.55 4.55 4.53 3.53 3.71 7.00 4.61
```

En annen måte å plukke ut en del av et datasett er å bruke kommandoen **subset**.

Eller bruk følgende hvor man lager objektene f1, f2, f3:

```
f1<-YIELD[FERTIL=="F1"]
f2<-YIELD[FERTIL=="F2"]
f3<-YIELD[FERTIL=="F3"]
```

Kommandoen **aggregate** sammen med **list** kan brukes til å samle data i grupper for eksempel beregne gjennomsnitt av YIELD.

```
fertilfaktor<-list(FERTIL=FERTIL)
YIELDmean<-aggregate(YIELD,by=fertilfaktor,mean)
YIELDmean
```

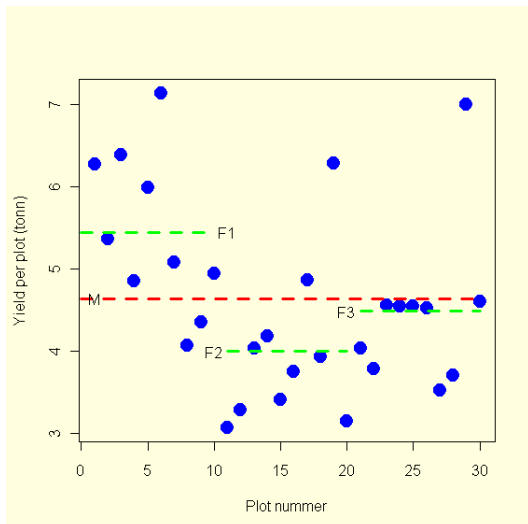
```
  FERTIL      x
1     F1 5.445
2     F2 3.999
3     F3 4.487
```

Kontrastene som er brukt i sammenligningene. Vi ser at F1 er brukt som referanse:

```
contrasts(FERTIL)
      F2 F3
F1    0  0
F2    1  0
F3    0  1
```

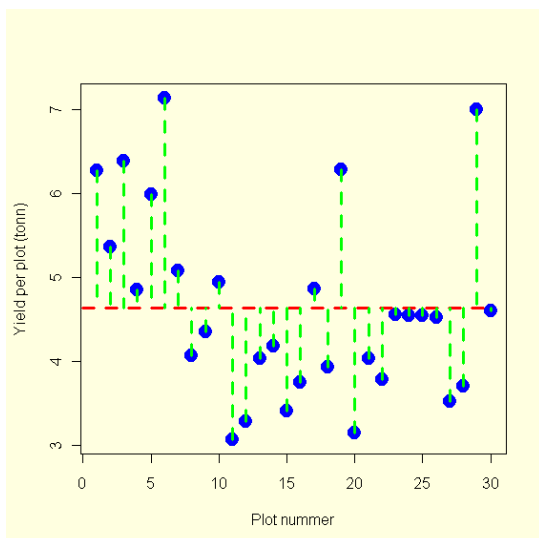
Nedenfor er det vist på figurer hva som skjer når vi utfører ANOVA på datasettet. Finner stormiddeltallet M og middeltallene for de 3 gjødseltypene F1, F2, F3

```
plot(YIELD,xlab="Plot nummer",ylab="Yield per plot
(tonn)",col="blue",pch=19,cex=2)
lines(c(0,30),c(mean(YIELD),mean(YIELD)),lty=2,col="red",lwd=3)
lines(c(0,10),c(mean(F1),mean(F1)),lty=2,col="green",lwd=3)
lines(c(11,20),c(mean(F2),mean(F2)),lty=2,col="green",lwd=3)
lines(c(21,30),c(mean(F3),mean(F3)),lty=2,col="green",lwd=3)
text(11,5.45,"F1")
text(10,4,"F2")
text(20,4.49,"F3")
text(1,4.64,"M")
```

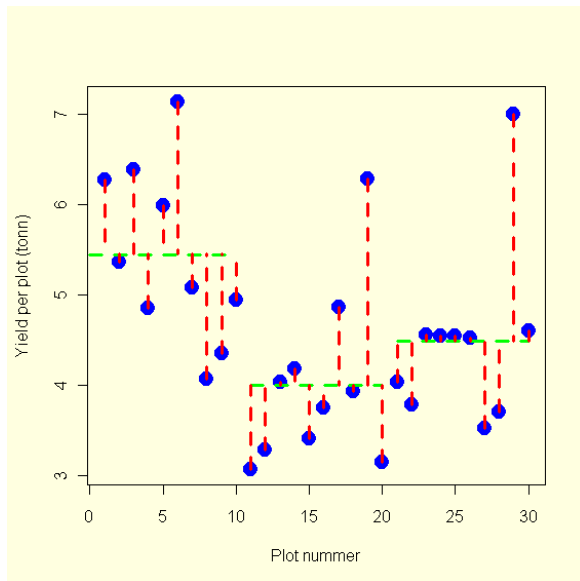
Deretter bestemmes avvik fra stormiddeltalet:

```
plot(YIELD, col="blue", pch=19, cex=2, xlab="Plot nummer", ylab="Yield per plot (tonn)")
lines(c(0,30), c(mean(YIELD), mean(YIELD)), lty=2, lwd=3, col="red")
for(i in 1:30) lines(c(i,i), c(YIELD[i], 4.64), lty=2, lwd=3, col="green")
```



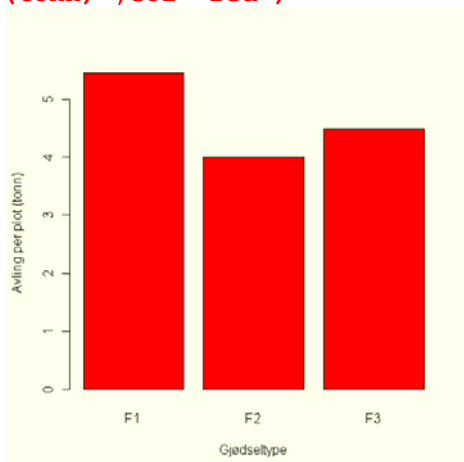
Så bestemmes avvik mellom stormiddeltalet og gjennomsnittene av F1, F2 og F3. Til slutt bestemmes avvik fra hver av gjennomsnittene for F1, F2 og F3:

```
plot(YIELD, xlab="Plot nummer", ylab="Yield per plot (tonn)", col="blue", pch=19, cex=2)
lines(c(0,10), c(mean(F1), mean(F1)), lty=2, lwd=3, col="green")
lines(c(11,20), c(mean(F2), mean(F2)), lty=2, lwd=3, col="green")
lines(c(21,30), c(mean(F3), mean(F3)), lty=2, lwd=3, col="green")
for(i in 1:10) lines(c(i,i), c(YIELD[i], 5.45), lty=2, lwd=3, col="red")
for(i in 11:20) lines(c(i,i), c(YIELD[i], 4.00), lty=2, lwd=3, col="red")
for(i in 21:30) lines(c(i,i), c(YIELD[i], 4.49), lty=2, lwd=3, col="red")
```



Hvis vi skal få R til å lage stolpediagram med error-bar må vi lage et program som gjør dette. Kommandoen `barplot` behøver en y-akse i form av en vektor, i vårt tilfelle bruker vi gjennomsnittsverdiene:

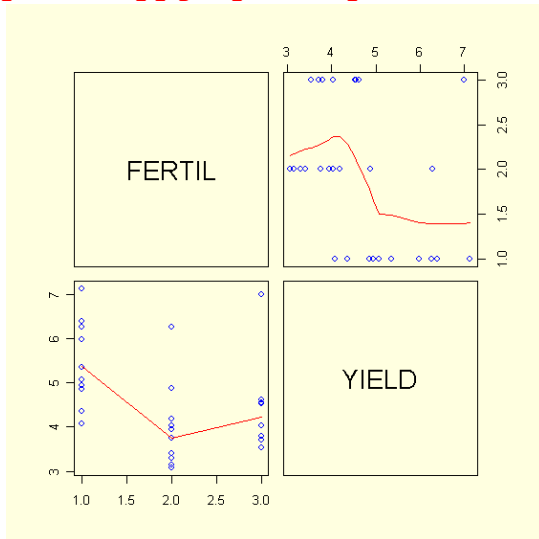
```
labels<-as.character(levels(FERTIL))
yakse<-as.vector(tapply(YIELD,FERTIL,mean))
barplot(yakse,names=labels,xlab="Gjødseltype",ylab="Avling per plot (tonn)",col="red")
```



Man kan også bruke kommandoen nedenfor for å lage et barplot (stolpediagram).

```
barplot(tapply(YIELD,list(FERTIL),mean),xlab="Gjødseltype",ylab="Yield per plot (tonn)",col="red")
```

```
pairs (opp1, panel=panel.smooth, col="blue")
```



Vi bruker ANOVA når vi skal sammenligne flere prøver. Skal vi sammenligne bare to prøver er det forskjellige kommandoer som kan brukes: Students t-test (**t.test**) som forutsetter normalfordeling eller Wilcoxon-rangeringssumtest (**wilcox.test**) som ikke forutsetter normalfordeling (ikke-parametrisk test). Andre tester er korrelasjonstest (**cor.test**), kji-kvadrattest på tellinger og kontingenstabeller (**chisq.test**, Fishers test på telldata (**fisher.test**) avhengig av prøvetype, forutsetninger og formål.

Vi kan gjøre en parvis sammenligning mellom F1 og F2. Det er en innebygget F-test (**var.test**) som sammenligner variansene. Er variansene forskjellige kan vi ikke sammenligne middeltallene med en t-test. Vi gjør først en F-test på F1 og F2:

```
F1<-YIELD[1:10];F1
F2<-YIELD[11:20];F2
var.test(F1,F2)
```

F test to compare two variances

```
data: F1 and F2
F = 1.0087, num df = 9, denom df = 9, p-value = 0.9899
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2505533 4.0611252
sample estimates:
ratio of variances
 1.008726
```

Vi beholder hypotesen, variansen er ikke signifikant forskjellig for F1 og F2.

Bruker Students t-test for to uavhengige prøver med konstant varians og normalfordelte errors. H_0 er at det ikke er noen forskjell mellom F1 og F2, med frihetsgrader $n_1 - n_2 - 2 = 18$ df. Er

det samme prøven som sammenlignes før og etter en behandling brukes parvis t-test og man legger inn kommandoen **paired=T**.

```
t.test(F1,F2,var.equal=T)  
Welch Two Sample t-test
```

```
data: F1 and F2  
t = 3.3201, df = 18, p-value = 0.003808  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 0.5309905 2.3610095  
sample estimates:  
mean of x mean of y  
 5.445      3.999
```

Konklusjon: H_0 forkastes, F1 og F2 er signifikant forskjellig. Vi kan se at vi får den samme p-verdi (halesannsynlighet) ved å lage en lineær modell med de to prøvene F1 og F2 som responsvariabel.

```
f1f2<-c(F1,F2);f1f2 #  
u<-factor(c(rep(1,10),rep(2,10)));u  
summary(lm(f1f2~u)) #u er prediktor
```

```
Call:  
lm(formula = f1f2 ~ u)  
Residuals:  
    Min       1Q   Median       3Q      Max  
-1.3750 -0.6235 -0.1670  0.6150  2.2810  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  5.4450     0.3080   17.68   8e-13 ***  
u2           -1.4460     0.4355   -3.32   0.00381 **  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.9739 on 18 degrees of freedom  
Multiple R-squared:  0.3798,    Adjusted R-squared:  0.3454  
F-statistic: 11.02 on 1 and 18 DF,  p-value: 0.003808
```

Ekvivalentene til kommandoene **pnorm** og **qnorm** for normalfordeling er **pt** og **qt** for t-fordelingen. kritisk verdi for t med df=18:

```
qt(0.975,18)  
[1] 2.100922
```

Beregnet t-verdi er større enn kritisk t-verdi og nullhypotesen om at F1 og F2 er like forkastes. Konfidensintervall= t-verdi x standardfeil

Wilcoxon er en ikke-parametrisk test som ikke forutsetter normalfordeling og er basert på rangeringer:

```
wilcox.test(F1,F2)
```

```
Wilcoxon rank sum test
```

```
data: F1 and F2  
W = 88, p-value = 0.002879  
alternative hypothesis: true mu is not equal to 0
```

Skew og kurtosis

(e. Crawly s. 285-289)

Skew (skjevfordeling), positiv skew til venstre og negativ skew til høyre:

```
skew<-function(x) {  
m3<-sum((x-mean(x))^3)/length(x)  
s3<-sqrt(var(x))^3  
m3/s3}
```

```
skew(YIELD)
```

```
[1] 0.6949096
```

Hvor sannsynlig er det å få en t-verdi 0.6949096 når det ikke er noen skew ?

```
1-pt(1.553865,29)
```

```
[1] 0.06553163
```

Vi ser at p-verdien er >0.05 som indikerer at det er sannsynlig at skewverdien ikke er forskjellig fra 0.

Kurtosis (spisshet):

```
kurtosis<-function(x) {  
m4<-sum((x-mean(x))^4)/length(x)  
s4<-var(x)^2  
m4/s4-3}
```

```
kurtosis(YIELD)
```

```
[1] -0.5169758
```

```
kurtosis(YIELD)/sqrt(24/length(YIELD))
```

```
[1] -0.5779965
```

```
1-pt(-0.5779965,29)
```

```
[1] 0.716136
```

Kruskal-Wallis rangerings sumtest `kruskal.test()` er et ikke-parametrisk alternativ til enveis anova.

```
kruskal.test(YIELD~FERTIL)
```

```
      Kruskal-Wallis rank sum test
```

```
data:  YIELD by FERTIL
```

```
Kruskal-Wallis chi-squared = 10.5847, df = 2, p-value = 0.00503
```

Konfidensintervall beregnet ved bootstraping

Ved bootstraping og resampling er det mulig å estimere gjennomsnittsverdi og konfidensintervall. Vi bruker kommandoen `sample(,replace=T)` for å lage et tilfeldig uttrekk fra prøven med tilbakelegging.

```
F1<-YIELD[1:10];F1
```

```
[1] 6.27 5.36 6.39 4.85 5.99 7.14 5.08 4.07 4.35 4.95
```

```
F2<-YIELD[11:20];F2
```

```
F3<-YIELD[21:30];F3
```

```
#resampler først for F1
```

```
Y<-sample(F1,replace=T);Y
```

```
[1] 4.95 4.85 5.08 4.85 6.27 4.95 5.36 5.36 4.95 4.07
```

```
#antall simuleringer
```

```
n<-10000
```

```
#Lager matriser for å lagre forventede verdier fra hver
```

```
#simulering
```

```
xmu<-matrix(NA,n,1) #matrise for gjennomsnitt
```

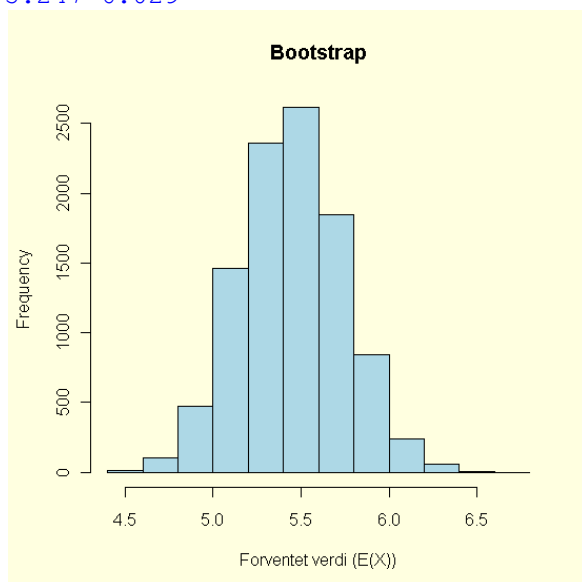
```
sdx<-matrix(NA,n,1) #matrise for standardavvik
```

```
for(i in 1:n)
```

```

    {
      xmu[i]<-mean(sample(F1,replace=T))
    }
par(bg="lightyellow")
#Lager et histogram over forventede verdier (gjennomsnitt)
hist(xmu,col="lightblue",xlab="Forventet verdi
(E(X))",main="Bootstrap")
#Bestemmer gjennomsnitt, standardfeil og kritisk t-verdi 9 df
ex<-mean(xmu);ex
[1] 5.447315
sta<-sd(xmu);sta
[1] 0.2917465
tobs<-qt(0.975,9);tobs
[1] 2.262157
ex+tobs*sta
[1] 6.107291
ex-tobs*sta
[1] 4.787338
quantile(xmu,c(0.25,.975))
 25% 97.5%
5.247 6.029

```



Bootstrap og resampling for F1. Gjennomsnittsverdi ved resampling estimert til 5.447, 95% konfidensintervall CI: 4.787-6.107. Kvantinintervallet for gjennomsnittet: 5.247-6.029. Gjør tilsvarende for F2 og F3.

Vi kan se hvilke objekter vi har laget:

```
objects()
```

Noen ganger kan det være lurt å renske opp i objektene du har samlet i minnet ved å fjerne dem:

```
rm(list=ls(all=TRUE))
```

Datsett i R

R inneholder flere datsett bl.a. i pakken datasets.

```
library(datasets)
```

```
data()
```

I pakken MASS (Venables & Ripley) er det et datasett fra et klassisk faktorielt eksperiment med undersøkelse av effekten av nitrogen (N), fosfat (P) og kalium (K) på vekst (yield) av erter, fordelt på 6 blokker, og som ligner på vårt labeksperiment.

```
library(MASS)
```

```
data(npk)
```

```
attach(npk)
```

```
names(npk)
```

```
npk
```

```
?npk
```

```
par(mar=c(12,4,4,2))
```

```
par(las=3) #vertikal aksetekst
```

```
boxplot(yield~N*P*K,xlab="NPK",ylab="yield")
```

```
summary(aov(yield ~ block + N*P*K)) #anova-tabell
```

Oppgave 2 Regresjon

Eksperiment: En skogeier ønsker å finne sammenhengen mellom volumet (VOLUME) av hoggbart skogvirke ut fra måling av høyden (HEIGHT) av trærne og diameter (DIAMET) målt 1.5 meter over bakken. Som statistisk utvalg hogges 31 trær, og fra disse måles volum, høyde og diameter 1.5 meter over bakken. Lag forslag til en modell som skogeieren kan benytte for å beregne hogstvolum ut fra diameter og høyde av trærne. Vær oppmerksom på korrelasjon mellom høyde og diameter.

Modell: $VOLUME \sim HEIGHT + DIAMET + error$

Vi laster inn datasettet:

```
oppg2<-
```

```
read.table("http://www.mn.uio.no/bio/tjenester/kunnskap/plantefys/matematikk/trees3.txt",header=T)
```

eller:

```
oppg2<-read.table("//platon/bio-kurs/bio2150/R/trees3.txt",header=T)
```

```
attach(oppg2)
```

```
names(oppg2)
```

```
[1] "DIAMET" "HEIGHT" "VOLUME"
```

```
summary(oppg2)
```

	DIAMET	HEIGHT	VOLUME
Min.	:0.2108	Min. :19.20	Min. :0.2888
1st Qu.:	:0.2807	1st Qu.:21.95	1st Qu.:0.5493
Median	:0.3277	Median :23.16	Median :0.6853
Mean	:0.3365	Mean :23.16	Mean :0.8543
3rd Qu.:	:0.3874	3rd Qu.:24.38	3rd Qu.:1.0562
Max.	:0.5232	Max. :26.52	Max. :2.1804

```
dim(oppg2)
```

```
[1] 31 3
```

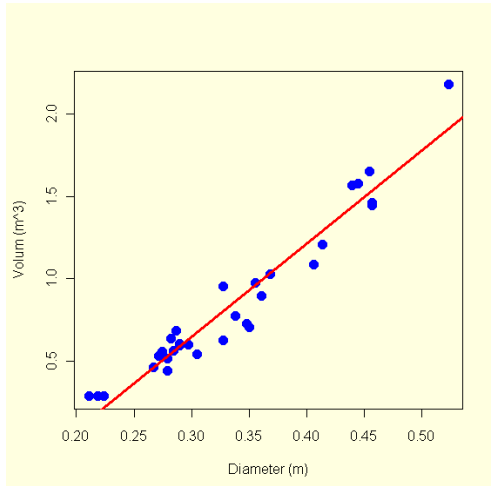
```
oppg2[1:5,]
```

	DIAMET	HEIGHT	VOLUME
1	0.21082	21.3360	0.2916635
2	0.21844	19.8120	0.2916635
3	0.22352	19.2024	0.2888318

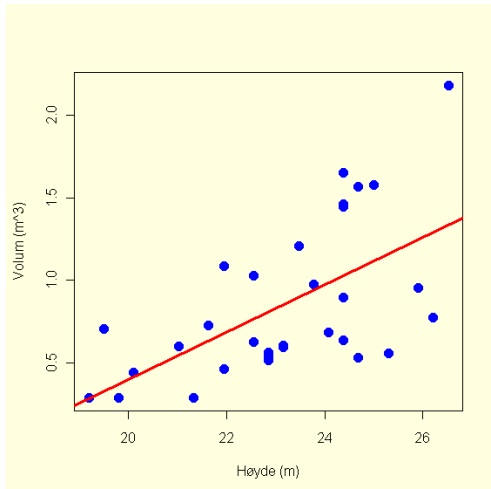
```
4 0.26670 21.9456 0.4643963
5 0.27178 24.6888 0.5323567
```

Vi plotter datasettet og trekker en linje ut fra en lineær modell:

```
par(bg="lightyellow")
plot(VOLUME~DIAMET, col="blue",pch=16,cex=1.5,xlab="Diameter
(m)",ylab="Volum (m^3)")
abline(lm(VOLUME~DIAMET), col="red", lwd=3)
```



```
plot(VOLUME~HEIGHT, col="blue",pch=16,cex=1.5,xlab="Høyde
(m)",ylab="Volum (m^3)")
abline(lm(VOLUME~HEIGHT), col="red", lwd=3)
```



En rett linje har formen:

$$y = \beta_0 + \beta_1 x$$

Vi kan finne forventet verdi av volum og høyde og lage et nytt aksesystem med origo i dette gjennomsnittet.

Vi har da fjernet skjæringspunktet med y-aksen β_0 og vi har en rett linje som går gjennom origo med stigningstall β_1 .

Vi svinger linjen rundt origo til de kvadrerte avvikene blir minst mulig:

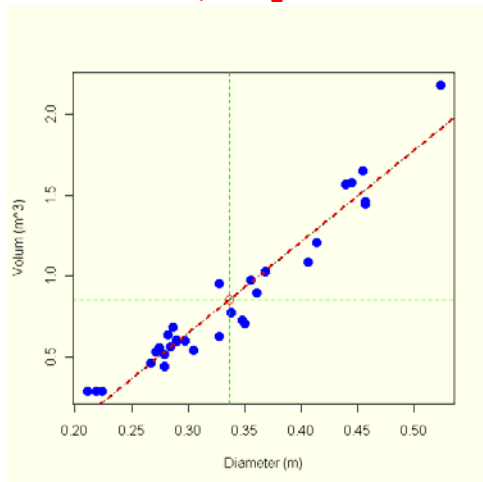
```
plot(VOLUME~DIAMET, col="blue",pch=16,cex=1.5,xlab="Diameter
(m)",ylab="Volum (m^3)")
mx<-mean(DIAMET)
my<-mean(VOLUME)
```



```

points(mx,my,cex=1.5,col=2)
abline(v=mx,h=my,lty=2,col=3)
#Nye x og y med origo i storgjennomsnitt
xny<-DIAMET-mx
yny<-VOLUME-mx
XY<-sum(xny*yny) #produktsummen
XX<-sum(xny^2) #kvadratsummen
b1<-XY/XX;b1 #finner stigningstallet
b0<-my-b1*mx;b0 #finner skjæringspunkt (intercept)
abline(b0,b1,lty=3,col=2,lwd=3) #trekker regresjonslinje
#Ser at dette stemmer med en lineær modell
mod<-lm(VOLUME~DIAMET)
abline(mod, lty=3)

```



$$y = -1.04612 + 5.64760x + \varepsilon$$

Deretter må de undersøkes om stigningstallet er signifikant forskjellig fra null, og hvor god regresjonslinjen er til å prediktere.

Vi kan også finne stigningstallet via følgende, hvor s_y og s_x er standardavviket for henholdsvis y og x og r er korrelasjonskoeffisienten

$$\beta_1 = r \cdot \frac{s_y}{s_x}$$

Korrelasjonskoeffisienten r er:

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

```

#Alternativ regnemåte for stigningstall
sdx<-sd(DIAMET) #standardavvik
sdy<-sd(VOLUME)
r<-cor(DIAMET,VOLUME) #korrelasjonskoeff
b12<-(sdy/sdx)*r;b12
5.647602
b02<-mean(VOLUME)-b12*mean(DIAMET);b02
[1] -1.046122

```

Konfidensintervallet for intercept β_0 , SE er standardfeilen og t er kritisk verdi for t ved $n-2$ frihetsgrader:

$$\beta_0 \pm t \cdot SE_{\beta_0}$$

Konfidensintervallet for stigningstallet β_1 blir:

$$\beta_1 \pm t \cdot SE_{\beta_1}$$

For å teste nullhypotesen $H_0: \beta_1=0$ så beregner vi t , finner hvor sannsynlig det er å få en slik t -verdi sammenlignet med dem kritiske t -verdi med $n-2$ frihetsgrader.

$$t = \frac{\beta_1}{SE_{\beta_1}}$$

Vi skal nå se på nærmere detaljer og starter først med en lineær regresjon og ser på sammenhengen mellom diameter og volum.

Modell: VOLUME~DIAMET

ANOVA-tabellen:

```
> modell<-lm(VOLUME~DIAMET)
```

```
> summary.aov(modell)
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
DIAMET     1  6.0794   6.0794  419.36 < 2.2e-16 ***
Residuals 29  0.4204   0.0145
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
modell<-lm(VOLUME~DIAMET)
```

```
summary(modell)
```

Call:

```
lm(formula = VOLUME ~ DIAMET)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-0.228386 -0.087972  0.004303  0.098961  0.271468
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.04612    0.09529  -10.98 7.62e-12 ***
DIAMET       5.64760    0.27578   20.48 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.1204 on 29 degrees of freedom

Multiple R-squared: 0.9353, Adjusted R-squared: 0.9331

F-statistic: 419.4 on 1 and 29 DF, p-value: < 2.2e-16

Vi ser at vi får de samme koeffisientene som tidligere:

En lineær regresjonslinje er av typen

$$\mu_y = \beta_0 + \beta_1 x + \varepsilon$$

hvor gjennomsnittsverdien μ endrer seg når x endres.

$$y = -1.04612 + 5.64760x + \varepsilon$$

Koeffisientene gir estimer av intercept β_0 og stigningskoeffisient β_1 , inkludert standardfeil, t-verdi og p-verdi og en eller flere * angir signifikansnivå. Residualene sier litt om residualane og fordeling av disse, middelverdiene av residualene skal være lik 0, og medianverdien bør ligge i nærheten av dette, og man kan se på kvartilene om dataene ser balansert ut. Verdien for skjæningspunkt med y-aksen, α , er interessant bare hvis man skal avgjøre om regresjonslinjen går gjennom origo eller ikke, men mest interesse knytter det seg til stigningskoeffisienten β . R^2 er kvadratet av Pearsons korrelasjonskoeffisient. F-test for hypotesen om at regresjonskoeffisienten er lik 0.

Formulert som en lineær modell:

$$VOLUME \sim -1.04612 + 5.64760 \cdot DIAMET + error$$

Vi kan gjøre det tilsvarende for sammenhengen mellom volum og høyde:

```
model2<-lm(VOLUME~HEIGHT)
```

```
summary.aov(model2)
```

```
      Df Sum Sq Mean Sq F value    Pr(>F)
HEIGHT  1  2.3263   2.3263  16.165 0.0003784 ***
Residuals 29  4.1735   0.1439
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2<-lm(VOLUME~HEIGHT)
```

```
summary(model2)
```

```
> Call:
```

```
lm(formula = VOLUME ~ HEIGHT)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.60242 -0.28018 -0.08195  0.34171  0.84532
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.46707     0.82892  -2.976 0.005835 **
HEIGHT       0.14338     0.03566   4.021 0.000378 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.3794 on 29 degrees of freedom
```

```
Multiple R-squared:  0.3579,    Adjusted R-squared:  0.3358
```

```
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

Som gir regresjonsmodellen:

$$VOLUME \sim -2.46707 + 0.14338 \cdot HEIGHT + error$$

Source	df	SS	MS	F	F tabell (5%)
Regression	1	SSR	SSR	$F=SSR/s^2$	$qf(0.95,1,n-2)$
Error	$n-2$	SSE	$s^2=SSE/(n-2)$		
Total	$n-1$	SST			

ANOVA-tabell for regresjon df - antall frihetsgrader, SS - kvadratsum, MS- middelkvadratsum, SST - totalkvadratsum, SSE - feilkvadratsum, SSR - regresjonskvadratsum

Varianse=SS/df

SST = SSR + SSE

$y = \beta_0 + \beta_1 x$

$R^2 = SSR/SST = 2.32631/6.499815 = 0.357904$

Adjusted $R^2 = 1 - [(SSE/(n-2))/(SST/(n-1))] = 1 - [(4.173505/29)/(6.499815/30)] = 0.3357628$

Fra $R^2 = 0.357904$ ser vi at 35.8% av variasjonen kan forklares med regresjonslinjen.

R^2 øker med hvor mye variasjon som blir forklart.

Adjusted R^2 er lik:

$$R_{adj}^2 = 1 - \left(\frac{SSE}{\frac{n-2}{SST}} \right) = 1 - \left(\frac{n-1}{n-p} \right) \cdot (1 - R^2)$$

SSE er feilkvadratsum, SST er totalkvadratsum, n er prøvestørrelsen, p er antall parametere i modellen og $p=2$ for vanlig lineære regresjon. Adjusted R^2 minsker med antall parametere som tillegges modellen.

t-verdien får vi ved å dividere estimatene på standardfeilen. F-verdien er:

$F=SSR/s^2 = 2.32631/0.1439140 = 16.16458$

#tabellverdi for kritisk verdi av F

qf(0.95,1,29)

[1] 4.182964

Sannsynligheten for å få en F-verdi vi har fått:

1-pf(16.16458,1,29)

[1] 0.0003783709

Det vil si vi forkaster nullhypotesen om et det ikke er noen lineær sammenheng mellom VOLUME og HEIGHT.

Responsvariabel VOLUME (y) og forklaringsvariabel HEIGHT (x)

Trenger følgende tall:

$\sum x = 718.1088$

$\sum x^2 = 16748.00$

$\sum y = 26.48475$

$\sum y^2 = 29.12697$

$\sum xy = 629.7384$

$(\sum y)^2 = 701.4418$

$(\sum x)^2 = 515680.2$

sum(HEIGHT)

[1] 718.1088

```

sum (VOLUME)
[1] 26.48475
sum (HEIGHT^2)
[1] 16748.00
sum (VOLUME^2)
[1] 29.12697
sum (HEIGHT*VOLUME)
[1] 629.7384
(sum (HEIGHT) ^2)
[1] 515680.2
(sum (VOLUME) ^2)
[1] 701.4418

```

Kovarianse mellom VOLUME og HEIGHT:

```

VOLUME*HEIGHT
[1] 6.222933 5.778438 5.546264 10.191455 13.143248 14.112507 8.886452
[8] 11.781281 15.604803 12.881730 16.500698 13.775036 14.037417 12.684944
[15] 12.363871 14.178965 24.796791 20.338029 15.748940 13.754322 23.225953
[22] 21.888152 23.184525 23.800776 28.311324 38.730637 39.421114 40.254867
[29] 35.559616 35.214377 57.818900

```

$$\mathbf{SST} = \sum y^2 - 1/n(\sum y)^2$$

$$\mathbf{SSX} = \sum x^2 - 1/n (\sum x)^2$$

$$\mathbf{SSXY} = \sum xy - 1/n(\sum x \sum y)$$

$$\mathbf{SST} = 29.12697 - 1/31*701.4418 = 6.499815$$

$$\mathbf{SST} = \mathbf{SSR} + \mathbf{SSE} = 6.0794 + 0.4202 = 6.4998$$

$$\mathbf{SSX} = 16748.00 - 1/31*515680.2 = 113.1548$$

$$\mathbf{SSXY} = 629.7384 - 1/31* (718.1088 *26.48475) = 16.22446$$

Stigningskoeffisienten β_1 :

$$\mathbf{\beta_1} = \mathbf{SSXY}/\mathbf{SSX} = 16.22446/113.1548 = 0.1433829$$

Skjæring med y-aksen, intercept (linje gjennom stormiddeltallet)

$$\mathbf{\beta_0} = y_m - bx_m = \sum y/31 - b*\sum x/31 = 26.48475/31 - 0.1433829 * 718.1088 /31 = -2.467089$$

$$\mathbf{VOLUME} = -2.467089 + 0.1433829*\mathbf{HEIGHT}$$

Regresjonskvadratsummen SSR:

$$\mathbf{SSR} = b*\mathbf{SSXY} = 0.1433829 *16.22446 = 2.32631$$

Feilkvadratsummen SSE:

$$\mathbf{SSE} = \mathbf{SST} - \mathbf{SSR} = 6.499815 - 2.32631 = 4.173505$$

Standardfeilen til regresjonsstigningskoeffisienten b:

$$s^2 = SSE/(n-2) = 4.173505/29 = 0.1439140$$

$$SE_b = \sqrt{s^2/SSX} = \sqrt{0.1439140/113.1548} = 0.03566277$$

Standardfeilen til skjæring med y-aksen (β_0): (x_m -gjennomsnitt av x , $\sqrt{\quad} = \text{sqrt}$)

$$SE_{\beta_0} = \sqrt{s^2[1/n + \sum x^2/SSX]} = \sqrt{((0.1439140 * 16748.00) / (31 * 113.1548))} = 0.8289258$$

t-verdier:

$$t_{\beta_0} = a/SE = -2.467089/0.8289258 = -2.976248$$

$$t_{\beta_1} = b/SE = 0.1433829 / 0.03566277 = 4.020521$$

Sannsynligheten for å få en slik t-verdi:

```
2*pt(-2.976248, 29)
[1] 0.005834442
```

95% konfidensintervall (CI_{β}) for β_1 :

$$CI_{\beta_1} = t(\alpha=0.025, df=(n-2)) \cdot SE_{\beta_1}$$

#kritisk verdi for t

```
qt(0.975, 29)
[1] 2.045230
```

Konfidensintervall for β_0 og β_1 :

$$CI_{\beta_1} = 0.1433829 \pm 2.045230 * 0.03566277 = 0.1433829 \pm 0.07293857$$

$$CI_{\beta_0} = -2.467089 \pm 2.045230 * 0.8289258 = -2.467089 \pm 1.695344$$

#95% konfidensintervall for parametere

```
confint(model2)
```

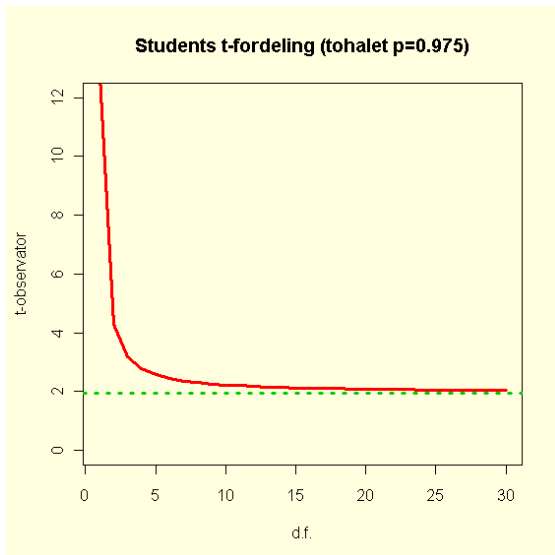
```
                2.5 %      97.5 %
(Intercept) -4.16240288 -0.7717291
HEIGHT       0.07044363  0.2163201
```

Som er det samme som vi har funnet over

Students t-fordeling:

Når n øker, dvs. antall d.f. øker så nærmer t-verdien seg 1.96 fra normalfordelingen:

```
plot(1:30, qt(0.975, 1:30), type="l", col="red", lwd=3, xlab = "d.f.",
     ylab="t-observator", ylim=c(0,12), main="Students t-fordeling (tohalet
     p=0.975)")
abline(h=1.96, lty=3, col=3, lwd=3)
```



Vi kan bruke modellen til å prediktere verdier. Hvor stort er volumet ved høyde 23 meter ?

```
predict(model2, list(HEIGHT=23))
```

```
1  
0.8307173
```

Svar: 0.83 m³.

Vi kan finne den totale variansen SST også på en annen måte via nullmodellen `lm(VOLUME~1)`:

```
deviance(lm(VOLUME~1))
```

```
[1] 6.499813
```

Feilkvadratsummen (SSE) finner vi ved:

```
deviance(lm(VOLUME~HEIGHT))
```

```
[1] 4.173513
```

SST=SSR+SSE, dvs. SSR=6.499813-4.173513= 2.3263

Som vi også finner igjen i ANOVA-tabellen:

```
> summary(aov(lm(VOLUME~HEIGHT)))
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)        
HEIGHT     1  2.3263   2.3263  16.165 0.0003784 ***  
Residuals 29  4.1735   0.1439
```

R² som angir hvor stor prosent var variasjonen som blir forklart av regresjonsmodellen:

$$R^2 = \frac{SST - SSE}{SST} = \frac{6.499813 - 4.173513}{6.499813} = 0.3579026$$

Som er det samme som:

```
summary(lm(VOLUME~HEIGHT)) [[8]]
```

```
[1] 0.3579026
```

Korrelasjonskoeffisienten r :

$$r = \frac{SSXY}{\sqrt{SSX \cdot SST}} = \frac{16.22446}{\sqrt{113.1548 \cdot 6.499813}} = 0.5982509$$

Som er det samme som:

```
cor(VOLUME, HEIGHT)
```

```
[1] 0.5982497
```

Pearsons korrelasjonstest brukes som default, andre korrelasjonstester mellom parvise variable finner du ved **?cor.test**

```
cor.test(VOLUME,DIAMET)
```

```
Pearson's product-moment correlation
data:  VOLUME and DIAMET
t = 20.4783, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9322519 0.9841887
sample estimates:
      cor
0.9671194
```

Vi kan lage en modell med både høyde og diameter:

```
model3<-lm(VOLUME~HEIGHT+DIAMET)
```

```
summary(model3)
```

```
Call:
```

```
lm(formula = VOLUME ~ HEIGHT + DIAMET)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.181411 -0.075021 -0.008143  0.062306  0.240260
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.64203    0.24461  -6.713 2.75e-07 ***
HEIGHT       0.03152    0.01209   2.607  0.0145 *
DIAMET       5.24883    0.29461  17.816 < 2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1099 on 28 degrees of freedom
```

```
Multiple R-squared:  0.948,    Adjusted R-squared:  0.9442
```

```
F-statistic: 255 on 2 and 28 DF,  p-value: < 2.2e-16
```

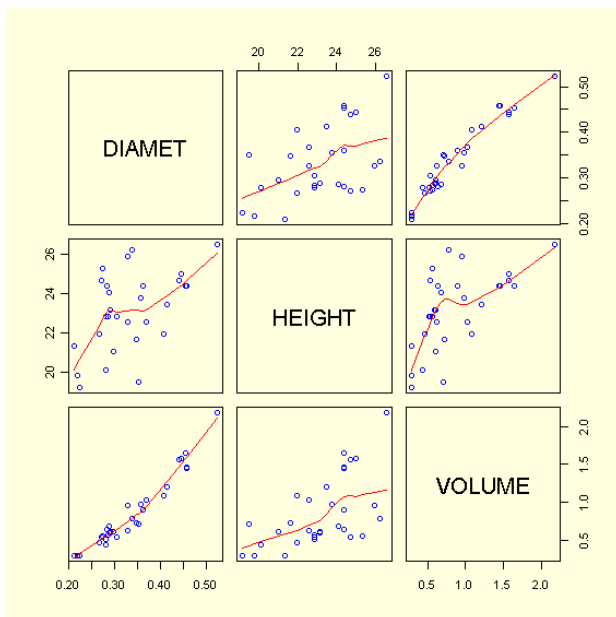
Som gir den lineære modellen

```
VOLUME ~ - 1.64203 + 0.03152 · HEIGHT + 5.24883 · DIAMET + error
```

La oss se på den utvidete modellen.

Vi kan plote i multipanel med kommandoen **pairs:**

```
pairs(oppg2,panel=panel.smooth,col="blue")
```

Matriser i statistiske modeller

Skrevet som matrise er en lineær regresjonsmodell lik

$$Y = X \cdot b + e$$

$$y_i = \beta_0 + \beta_1 \cdot x_i + e_i \quad \text{for } i = 1, 2, 3, \dots, n$$

hvor e er normalfordelte feil-ledd (error) og vi ønsker å bestemme b .

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 \cdot x_1 + e_1 \\ \beta_0 + \beta_1 \cdot x_2 + e_2 \\ \beta_0 + \beta_1 \cdot x_3 + e_3 \\ \vdots \\ \beta_0 + \beta_1 \cdot x_n + e_n \end{pmatrix}$$

Nå skal vi gjøre det samme regnestykket med datasett trees3, men nå i form av matriser:

```
VOLUME%*%HEIGHT
```

```
      [,1]
[1,] 629.7384
```

```
VOLUME%*%VOLUME
```

```
      [,1]
[1,] 29.12697
```

```
HEIGHT%*%HEIGHT
```

```
      [,1]
[1,] 16748.00
```

Vi lager matrisen Y med VOLUME.

```
Y<-VOLUME
```

I matrisen X må vi lage en kolonne med 1-tall og binde disse sammen med HEIGHT.

```
X<-cbind(1,HEIGHT)
```

```
Σ y2= 29.12697
```

```
t(Y)%*%Y
```

```
      [,1]  
[1,] 29.12697
```

Vi finner en matrise som inneholder n , \sqrt{x} og $\sqrt{x^2}$ ved å ta den transponerte matrisen tX og gange den med X :

```
tXX<-t(X)%*%X;tXX
```

```
      HEIGHT  
      31.0000  718.1088  
HEIGHT 718.1088 16748.0026
```

Vi finner en matrise som inneholder $\sum y$ og $\sum xy$:

```
tXY<-t(X)%*%Y;tXY
```

```
      [,1]  
      26.48475  
HEIGHT 629.73836
```

Vi løser matriseligningene og finner β_1

```
b<-solve(tXX,tXY);b
```

```
      [,1]  
      -2.4670660  
HEIGHT  0.1433819
```

Som vi ser er lik koeffisientene i modellen:

```
VOLUME ~ -2.46707 + 0.14338·HEIGHT + error
```

$$\begin{pmatrix} 31 & 718.1088 \\ 718.1088 & 16748.0026 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 26.48475 \\ 629.73836 \end{pmatrix} \rightarrow \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} -2.4670660 \\ 0.1433819 \end{pmatrix}$$

Som er det samme som de to ligningene:

$$\begin{aligned} \beta_0 \cdot n + \beta_1 \cdot \sum x &= \sum y \\ \beta_0 \cdot \sum x + \beta_1 \cdot \sum x^2 &= \sum xy \end{aligned}$$

De samme regneoperasjonene kan nå gjøres i matriseform hvis vi har $m > 1$ x-variable, noe som viser den store fleksibiliteten ved å regne med matriser.

Stigningskoeffisienten (slope) β_1

$$\beta_1 = \frac{\sum(x - \bar{x}) \cdot (y - \bar{y})}{\sum(x - \bar{x})^2}$$

Skjæringspunktet (intercept) :

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

Appendiks oppgave 2

Datasettet `trees` til oppgave 2 er hentet fra R og viser sammenhengen mellom volum (cubic ft), diameter (girth, inch) og høyde (height, ft) for 31 felte kirsebærtrær. Vi må regne om fra tommer (1 inch=2.54 cm) og fot (1 foot=30.38 cm) til meter

```
library(datasets)
```

```
data(trees)
```

```
trees2<-
```

```
transform(trees,DIAM=Girth*2.54/100,HEIGHT=Height*30.48/100,VOLUME=Volume*(30.48/100)^3)
```

```
trees2
```

```
trees3<-trees2[,c("DIAM", "HEIGHT", "VOLUME")]
```

Nedenfor vises det hvordan man ved regresjonen først finner stormiddeltallet og dreier linjen med senter i stormiddeltallet slik at de kvadrerte avvikene blir minst mulig. Vi har laget et nytt x-,y-koordinatsystem med origo i stormiddeltallet, og har fått en linje av typen $y=bx$

```
mean(VOLUME)
```

```
[1] 0.8543467
```

```
Antall målepunkter:
```

```
length(VOLUME)
```

```
[1] 31
```

```
Tilpasset prediktert linje:
```

```
fitted<-predict(lm(VOLUME~HEIGHT))
```

Trekker prikkete linjer fra hvert punkt ned til regresjonslinjen. Vi lager et nytt koordinatsystem med origo i stormiddeltallet. Regresjonslinjen går da gjennom origo og blir av typen $y=bx$. Regresjonslinjen vipres rundt det nye origo inntil de kvadrerte avvik blir så liten som mulig. De loddrette linjene fra hvert målepunkt ned/opp til regresjonslinjen blir mål på residualene.

```
par(bg="lightyellow")
```

```
plot(HEIGHT, VOLUME, xlab="Høyde", ylab="Volum", pch=19, cex=2, col="blue")
```

```
abline(lm(VOLUME~HEIGHT), col="red", lwd=3)
```

```
abline(h=mean(VOLUME), lty=2, lwd=3, col="orange")
```

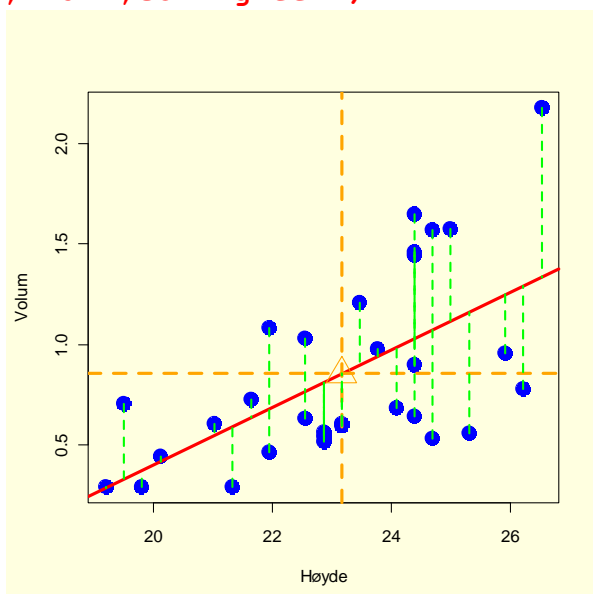
```
abline(v=mean(HEIGHT), lty=2, lwd=3, col="orange")
```

```
points(mean(HEIGHT), mean(VOLUME), pch=24, col="orange", cex=3)
```

```
fitted<-predict(lm(VOLUME~HEIGHT))
```

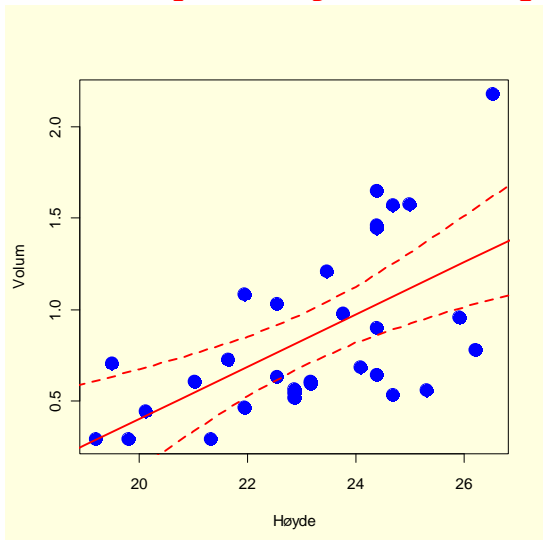
```
for(i in
```

```
1:31) lines(c(HEIGHT[i], HEIGHT[i]), c(VOLUME[i], fitted[i]), lty=2, lwd=2, col="green")
```



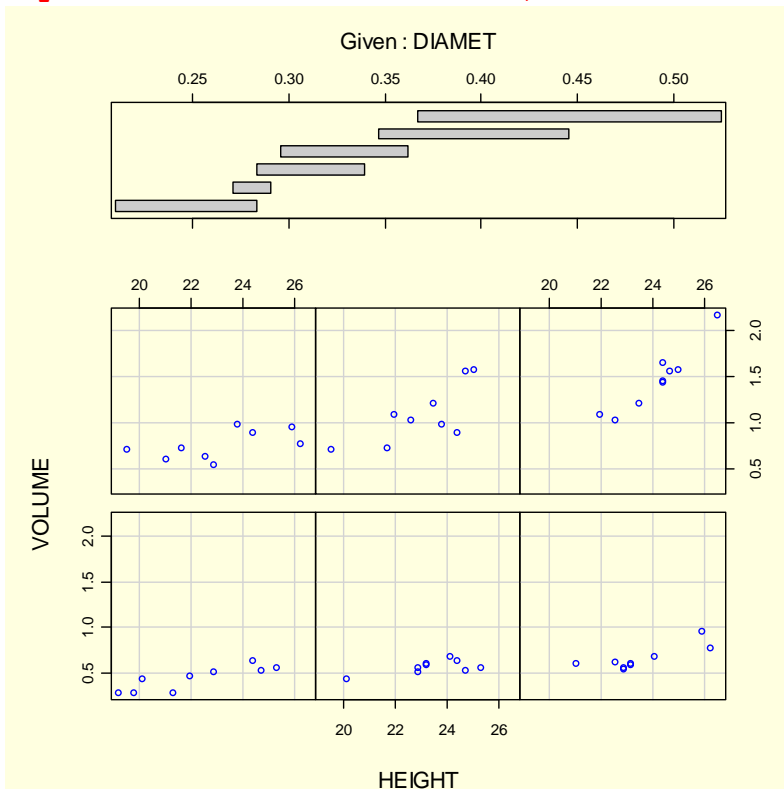
Konfidensintervall ved å lage en ny dataramme, bruke tilleggskommandoen `interval="conf"` og plotte med `matlines`.

```
par(bg="lightyellow")
plot(HEIGHT, VOLUME, xlab="Høyde", ylab="Volum", pch=19, cex=2, col="blue")
hoyde<-data.frame(HEIGHT=18:27)
conf<-predict(lm(VOLUME~HEIGHT), int="c", newdata=hoyde)
predheight<-hoyde$HEIGHT
matlines(predheight, conf, lty=c(1,2,2), col="red", lwd=2)
```



Vi kan også plotte med kommandoen `coplot`:

```
coplot(VOLUME~HEIGHT | DIAMET, col="blue")
```



Vi ser på en mer kompleks modell om det er interaksjon mellom høyde og diameter:

```
model4<-lm(VOLUME~HEIGHT*DIAMET)
```

```
summary(model4)
```

```
Call:
```

```
lm(formula = VOLUME ~ HEIGHT * DIAMET)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.186383 -0.030222  0.008568  0.044291  0.132096
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.96508     0.67495   2.911  0.00713 **
HEIGHT        -0.12050     0.02879  -4.186  0.00027 ***
DIAMET        -6.52831     2.14198  -3.048  0.00511 **
HEIGHT:DIAMET  0.49251     0.08916   5.524  7.48e-06 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.0767 on 27 degrees of freedom
```

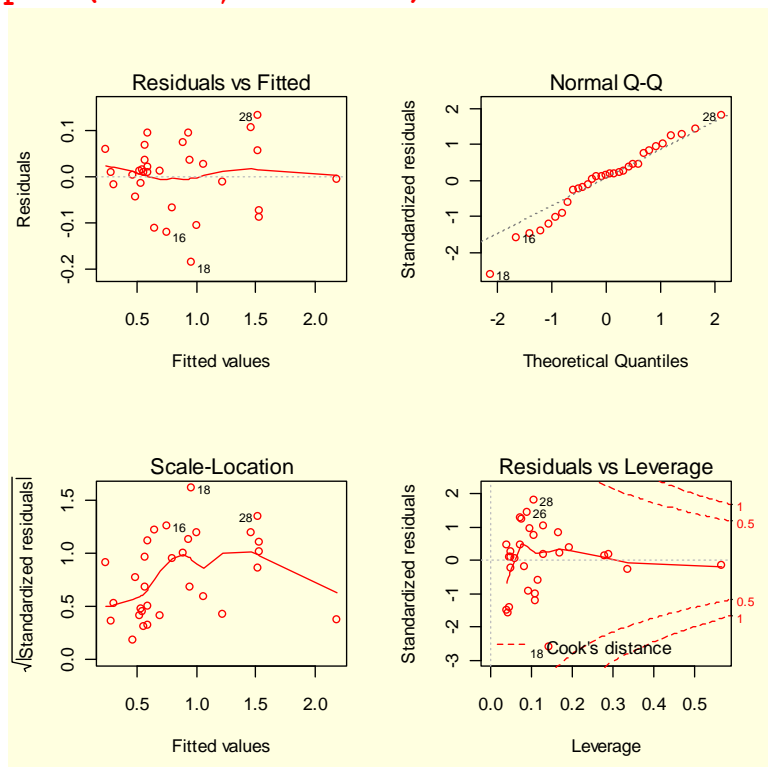
```
Multiple R-squared:  0.9756,    Adjusted R-squared:  0.9728
```

```
F-statistic: 359.3 on 3 and 27 DF,  p-value: < 2.2e-16
```

Vi plotter den komplekse modellen:

```
par(mfrow=c(2,2))
```

```
plot(model4,col="red")
```



```
step(model4)
```

```
Start:  AIC=-155.49
```

```
VOLUME ~ HEIGHT * DIAMET
```

	Df	Sum of Sq	RSS	AIC
<none>			0.159	-155.492
- HEIGHT:DIAMET	1	0.179	0.338	-134.051

```
Call:
lm(formula = VOLUME ~ HEIGHT * DIAMET)

Coefficients:
(Intercept)          HEIGHT          DIAMET  HEIGHT:DIAMET
      1.9651         -0.1205         -6.5283          0.4925
```

```
#Bestemmer AIC
```

```
AIC(model3)
```

```
[1] -44.07653
```

```
AIC(model4)
```

```
[1] -65.51734
```

Vi tester om modell1 og model2 er signifikant forskjellige ved å bruke ANOVA:

```
anova(model3,model4, test="F")
```

```
Analysis of Variance Table
```

```
Model 1: VOLUME ~ HEIGHT + DIAMET
```

```
Model 2: VOLUME ~ HEIGHT * DIAMET
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	0.33832				
2	27	0.15883	1	0.17949	30.512	7.484e-06 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Som viser at interaksjonsleddet HEIGHT:DIAMET er signifikant.

Hvis man ser på korrelasjonen mellom variablene ser man spesielt korrelasjonen mellom VOLUME og DIAMET som også går fram av figurene foran. Man kan også bruke `cor.test`

```
cor(HEIGHT,DIAMET)
```

```
[1] 0.5192801
```

```
cor(VOLUME,DIAMET)
```

```
[1] 0.9671194
```

```
cor(VOLUME,HEIGHT)
```

```
[1] 0.5982497
```

I tilfeller hvor det ikke er noen lineære interaksjon mellom avhengig og uavhengig variable kan man bruke Generaliserte additive modeller (`gam`) med glattingsfunksjon `s`. Dette er ikke påkrevet i vårt tilfelle, men her er vist hvordan dette eventuelt kan se ut.

```
library(mgcv)
```

```
This is mgcv 1.4-1
```

```
model5<-gam(VOLUME~s(HEIGHT)+s(DIAMET))
```

```
summary(model5)
```

```
Family: gaussian
```

```
Link function: identity
```

```
Formula:
```

```
VOLUME ~ s(HEIGHT) + s(DIAMET)
```

```
Parametric coefficients:
```

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85435     0.01364   62.65  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
              edf Ref.df      F p-value
s(HEIGHT)  1.000  1.500  10.67 0.00102 **
s(DIAMET)  2.693  3.193 216.45 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

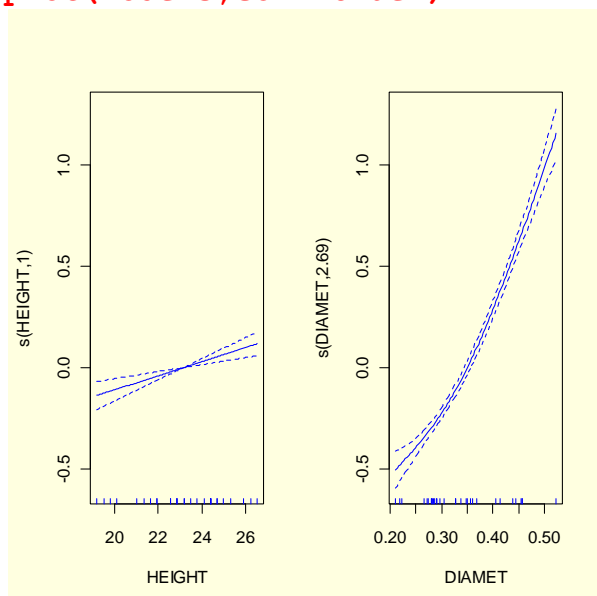
R-sq.(adj) =  0.973   Deviance explained = 97.7%
GCV score = 0.0067944   Scale est. = 0.0057657   n = 31

```

```

par(mfrow=c(1,2))
plot(model5,col="blue")

```



Vi kan i stedet lage en teoretisk betraktning med treet braktet som en konus hvor volumet blir:

$$\text{VOLUME} = \frac{1}{3} \cdot \pi \cdot \text{radius}^2 \cdot \text{høyde}$$

$$\text{Radius} = \frac{1}{2} \text{ diameter}$$

$$\text{VOLUME} = \frac{\pi \cdot \text{diameter}^2 \cdot \text{høyde}}{12}$$

Imidlertid foretar vi en empirisk justering:

$$\text{VOLUME} = \frac{\pi \cdot \text{diameter}^2 \cdot \text{høyde}}{10}$$

La D og H være henholdsvis diameter og høyde:

```
D<-seq(0.2,0.55,0.01)
```

```
H<-seq(20,27,0.2)
```

```
length(D)
```

```
[1] 36
```

```
length(H)
```

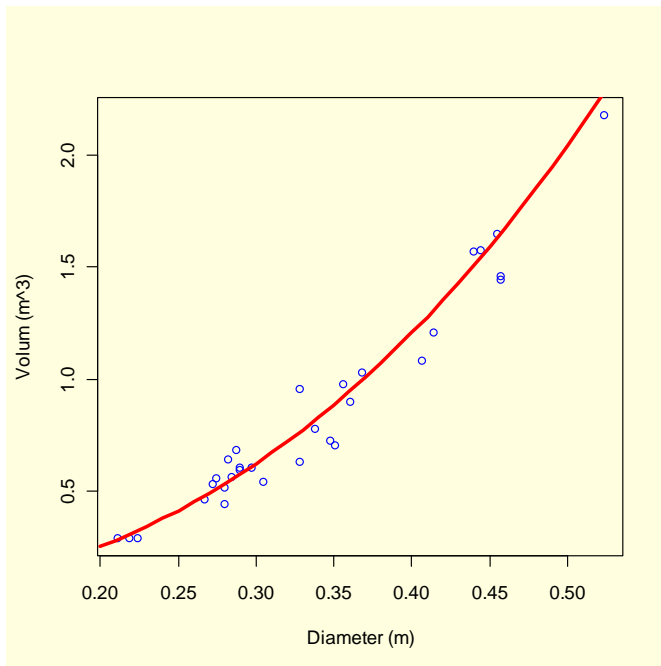
```
[1] 36
```

```
>
```

```
VOL<- (pi*D^2*H)/10
```

```
plot(VOLUME~DIAMET, col="blue",xlab="Diameter (m)",ylab="Volum (m^3)")
```

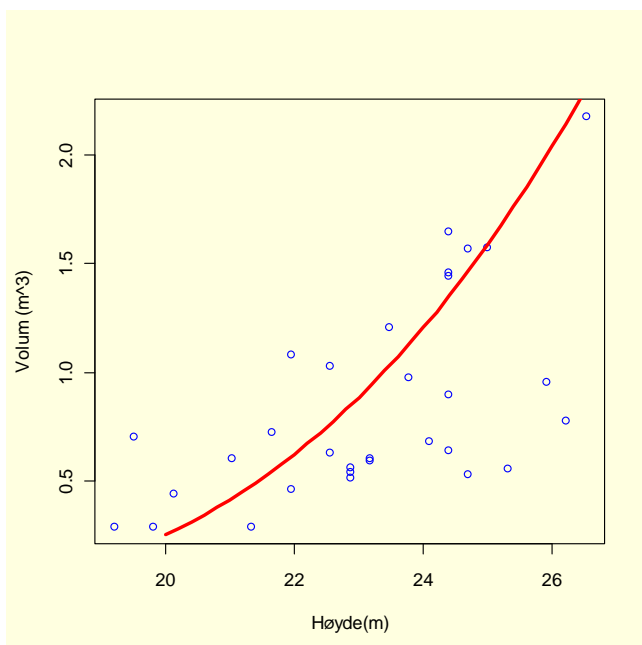
```
lines(VOL~D,col="red",lwd=3)
```



```

D<-seq(0.2,0.55,0.01)
H<-seq(20,27,0.2)
VOL<-(pi*D^2*H)/10
plot(VOLUME~HEIGHT, col="blue",xlab="Høyde (m)",ylab="Volum
(m^3)")
lines(VOL~H,col="red",lwd=3)

```



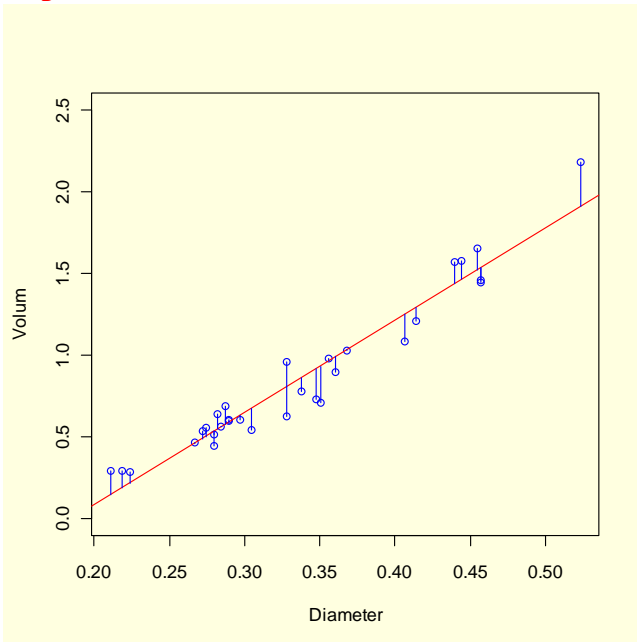
```

modell1<-lm(VOLUME~DIAMET)
fitted(modell1)
resid(modell1)
par(bg="lightyellow")
plot(DIAMET,VOLUME,ylim=c(0,2.5),xlab="Diameter",ylab="Volum",
col="blue")

```

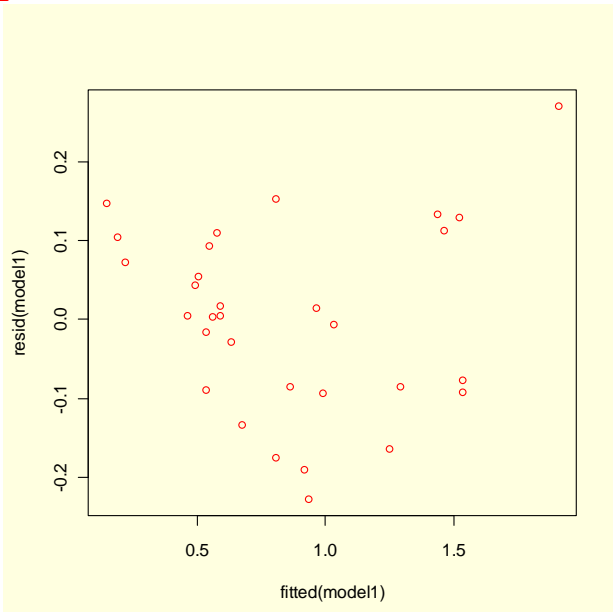


```
abline(lm(VOLUME~DIAMET), col="red")
segments(DIAMET, fitted(modell), DIAMET, VOLUME, col="blue")
```

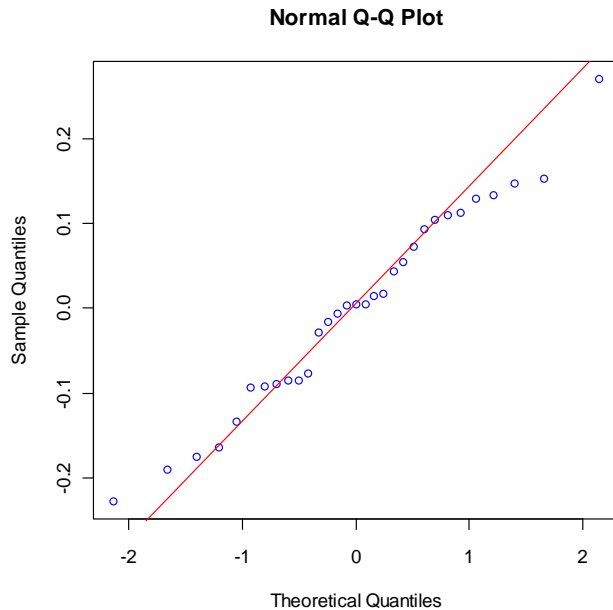


Se på tilpassete verdier og residualer:

```
par(bg="lightyellow")
plot(fitted(modell), resid(modell), col="red")
```



```
qqnorm(resid(modell), col="blue")
qqline(resid(modell), col="red")
```



```
shapiro.test(resid(modell))
Shapiro-Wilk normality test
```

```
data: resid(modell)
W = 0.9789, p-value = 0.7811
Som viser normalfordeling av residualene (p>0.05)
```

Vi kan også forsøke en selvstartende modell **SSlogis**:

```
par(bg="lightyellow")
logis<-nls (VOLUME~SSlogis (DIAMET,Asym,xmid,scal))
summary(logis)
predikt<-predict(logis)
plot(VOLUME~DIAMET, col="blue",xlab="Diameter
(m)",ylim=c(0,5),ylab="Volum (m^3)")
lines (DIAMET,predikt,col="red",lwd=3)
abline (h=coef(logis) ["Asym"],col="green",lty=2,lwd=3)
Formula: VOLUME ~ SSlogis(DIAMET, Asym, xmid, scal)
```

Parameters:

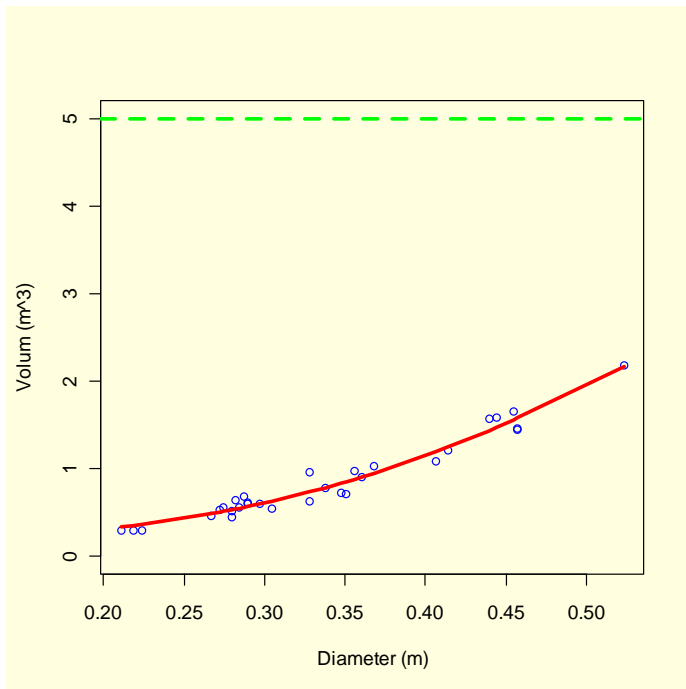
	Estimate	Std. Error	t value	Pr(> t)
Asym	4.99617	2.12330	2.353	0.0259 *
xmid	0.55847	0.09580	5.830	2.90e-06 ***
scal	0.13106	0.01759	7.452	4.08e-08 ***

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.0944 on 28 degrees of freedom
```

```
Number of iterations to convergence: 3
```

```
Achieved convergence tolerance: 5.625e-07
```

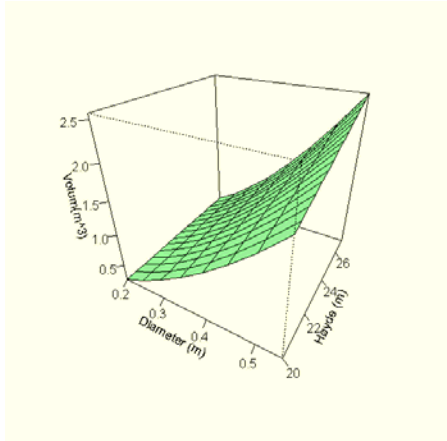


#3D av modellen

```

D<-seq(0.2,0.55,0.05) #diameter
H<-seq(20,27,0.5) #høyde
f<- function (D,H)(pi*D^2*H)/10 #volum
V <- outer(D,H,f)
V[is.na(V)]<- 1
op <- par(bg= "lightyellow")
persp(D,H,V,theta=30,phi=30,col="lightgreen",ticktype="detailed",
xlab="Diameter (m)",ylab="Høyde (m)",zlab="Volum(m^3)")

```

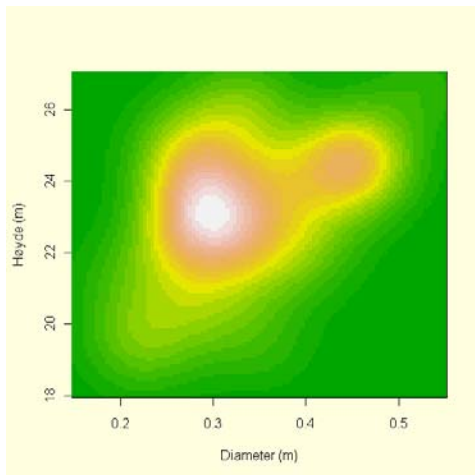


Med kommandoen `kde2d()` i pakken MASS kan man lage et todimensjonalt kjernetetthetsestimert lagt på et rutenett:

```

library(MASS)
?kde2d
tetthet<-kde2d(DIAMET,HEIGHT,n=100,lims=c(0.15,0.55,18,27))
image(tetthet, col = terrain.colors(20),xlab="Diameter
(m)",ylab="Høyde (m)")

```



Oppgave 3 Blokkdesign

Ekspesiment: En bonde ønsker å finne ut hvilken sort av bønner (BEAN 1-6) som gir størst avling (YIELD) på jorda han eier. For å unngå effekten av gradienter på forsøksfeltet deles forsøket inn i fire blokker (BLOCK 1-4). Er det noen signifikant forskjell mellom bønnesortene? Hvilken bønnesort vil du anbefale at bonden benytter?
 Datasett fra: Grafen, A. & Hails, R.: *Modern statistics for the life sciences*. Oxford University Press 2003

Modell: YIELD ~ BEAN + BLOCK + error

```

oppg3<-
read.table("http://www.mn.uio.no/bio/tjenester/kunnskap/plantefys/matematikk/oppg3.txt",header=T)

oppg3<-read.table("//platon/bio-kurs/bio2150/R/oppg3.txt",header=T)
attach(oppg3)
names(oppg3)
[1] "YIELD" "BLOCK" "BEAN"
summary(oppg3)
      YIELD      BLOCK BEAN
Min.   : 6.40    a:6   S1:4
1st Qu.:13.75   b:6   S2:4
Median :16.00   c:6   S3:4
Mean   :16.68   d:6   S4:4
3rd Qu.:19.98           S5:4
Max.   :25.60           S6:4

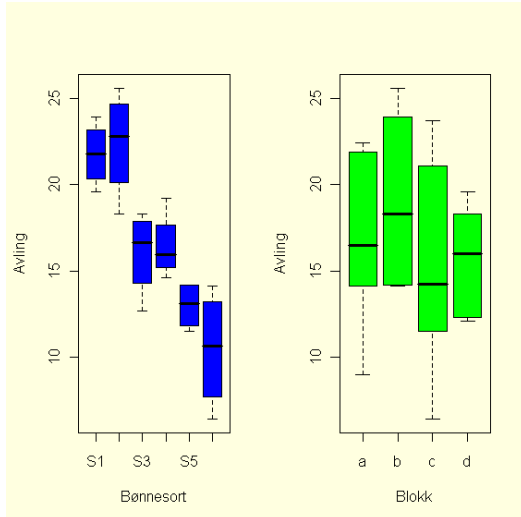
dim(oppg3)
[1] 24 3
oppg3[1:10,]
#Hele
oppg3
      YIELD BLOCK BEAN
1      9.0    a   S6
2     14.6    a   S4
3     18.3    a   S3
4     14.1    a   S5
5     21.9    a   S2
6     22.4    a   S1
7     14.2    b   S5

```

```
8 14.1 b S6
9 17.4 b S3
10 25.6 b S2
```

Plotting av datasettet:

```
par(mfrow=c(1,2),bg="lightyellow")
plot(BEAN,YIELD,col="blue",xlab="Bønnesort",ylab="Avling")
plot(BLOCK,YIELD,col="green",xlab="Blokk",ylab="Avling")
```



#ANOVA-tabell

```
summary(aov(YIELD~BEAN+BLOCK))
```

```
summary(aov(YIELD~BEAN+BLOCK))
      Df Sum Sq Mean Sq F value    Pr(>F)
BEAN    5  444.43   88.89  23.4757 1.341e-06 ***
BLOCK    3   52.89   17.63   4.6567  0.01713 *
Residuals 15   56.79    3.82
```

#Lineær modell

```
model<-lm(YIELD~BEAN+BLOCK)
```

```
summary(model)
```

Call:

```
lm(formula = YIELD ~ BEAN + BLOCK)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-3.1167 -1.1729  0.2333  0.8375  2.8083
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.792     1.192  18.288 1.15e-11 ***
BEANS2         0.625     1.376   0.454  0.65616
BEANS3        -5.675     1.376  -4.125  0.00090 ***
BEANS4        -5.325     1.376  -3.870  0.00151 **
BEANS5        -8.775     1.376  -6.378 1.24e-05 ***
BEANS6       -11.300     1.376  -8.213 6.22e-07 ***
BLOCKb         2.350     1.123   2.092  0.05388 .
BLOCKc        -1.517     1.123  -1.350  0.19703
BLOCKd        -1.000     1.123  -0.890  0.38745
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
Residual standard error: 1.946 on 15 degrees of freedom
Multiple R-Squared:  0.8975,    Adjusted R-squared:  0.8428
F-statistic: 16.42 on 8 and 15 DF,  p-value: 4.047e-06
```

```

model2<-lm(YIELD~BEAN)
summary(model2)

```

Call:

```
lm(formula = YIELD ~ BEAN)
```

Residuals:

```

  Min      1Q  Median      3Q      Max
-4.075 -1.456 -0.250  1.456  3.650

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.750      1.234   17.621 8.47e-13 ***
BEANS2         0.625      1.746    0.358 0.72447
BEANS3        -5.675      1.746   -3.251 0.00443 **
BEANS4        -5.325      1.746   -3.051 0.00688 **
BEANS5        -8.775      1.746   -5.027 8.76e-05 ***
BEANS6       -11.300      1.746   -6.474 4.34e-06 ***

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 2.469 on 18 degrees of freedom

Multiple R-Squared: 0.802, Adjusted R-squared: 0.7471

F-statistic: 14.59 on 5 and 18 DF, p-value: 8.579e-06

Hvis man sammenligner dette med middeltallene ser man hvordan estimatene i modellen kommer fram:

```
tapply(YIELD, BEAN, mean)
```

```

  S1      S2      S3      S4      S5      S6
21.750 22.375 16.075 16.425 12.975 10.450

```

```
anova(model, model2, test="F")
```

Analysis of Variance Table

Model 1: YIELD ~ BEAN + BLOCK

Model 2: YIELD ~ BEAN

```

  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1      15  56.795
2      18 109.690 -3   -52.895 4.6567 0.01713 *

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modell:

	BEANS		BLOCK	
	0.000	S1		
	0.625	S2	0.000	a
YIELD ~ 21.792 +	-5.675	S3	+ 2.350	b
	-5.325	S4	-1.517	c
	-8.775	S5	-1.000	d
	-11.300	S6		

+ error

BLOCKa og BEAN S1 er brukt som referanse (settes alfabetisk) og settes lik 0.

```

#Kontraster
contrasts (BEAN)
  S2 S3 S4 S5 S6
S1  0  0  0  0  0
S2  1  0  0  0  0
S3  0  1  0  0  0
S4  0  0  1  0  0
S5  0  0  0  1  0
S6  0  0  0  0  1
contrasts (BLOCK)
  b c d
a 0 0 0
b 1 0 0
c 0 1 0
d 0 0 1

```

En lineær modell betyr ikke nødvendigvis en rettlinjert sammenheng mellom responsvariabel og avhengig variabel. Noen modeller er ikke-lineære og kan bli linearisert ved transformering. "Fixed factors" er vanligvis krysset, dvs. alle kombinasjoner er representert, mens "random factors" er ofte nestet.

Se: Splitplotdesign og ANCOVA.

Oppgave 4 Telling

Eksperiment: Du skal utføre et feltforsøk for å teste hvordan forskjellige stammer av bygg (*Hordeum vulgare*) er resistent mot meldugg (*Erysiphe graminis*) under forskjellige vannforhold. Det er fem stammer (STRAIN) bygg som sammenlignes under fire forskjellige vanningsregimer (WATER). For hver behandling plukkes det tilfeldig ut 6 blader hvor antall flekker med soppinfeksjon (SPOTS) telles. Er det signifikant forskjell mellom stammene når det gjelder resistens (toleranse) for sopp sykdommen meldugg? WATER og STRAIN er kategoriske variable. Når det gjelder telling av hendelser eller objekter som skjer tilfeldig i tid eller rom bør man straks tenke på Poisson-fordeling.

Datsett fra: Grafen, A. & Hails, R.: *Modern statistics for the life sciences*. Oxford University Press 2003

Modell: SPOTS~WATER + STRAIN

```

oppg4<-
read.table("http://www.mn.uio.no/bio/tjenester/kunnskap/plantefys/matematikk/oppg4.txt",header=T)
eller:
oppg4<-read.table("//platon/bio-kurs/bio2150/R/oppg4.txt",header=T)
attach(oppg4)
names(oppg4)
[1] "WATER" "STRAIN" "SPOTS"
summary(oppg4)
WATER  STRAIN      SPOTS
W1:5   S1:4   Min.    : 5.0
W2:5   S2:4   1st Qu.: 5.0

```

```

W3:5   S3:4   Median : 6.5
W4:5   S4:4   Mean    : 9.3
        S5:4   3rd Qu.:14.0
        Max.   :19.0

```

oppg4 [1:6,]

```

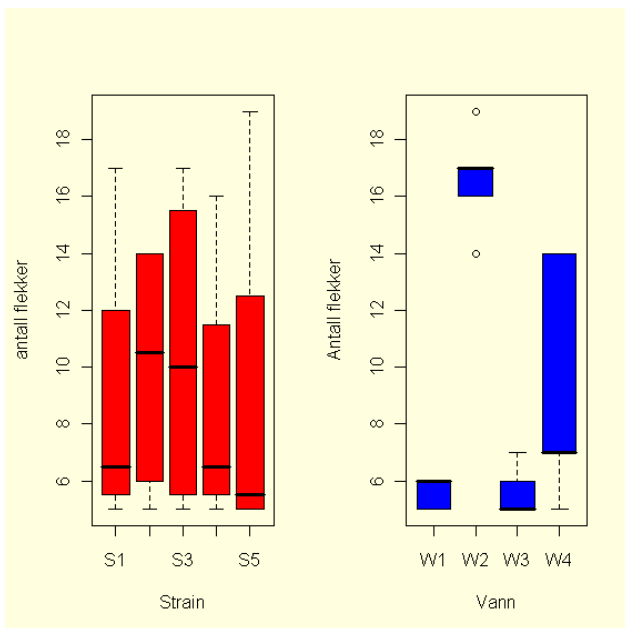
WATER STRAIN SPOTS
1     W1     S1     6
2     W1     S2     5
3     W1     S3     5
4     W1     S4     6
5     W1     S5     6
6     W2     S1    17

```

```

par(mfrow=c(1,2),bg="lightyellow")
plot(STRAIN,SPOTS,col="red",xlab="Strain",ylab="antall
flekker")
plot(WATER,SPOTS,col="blue",xlab="Vann",ylab="Antall flekker")

```



#ANOVA-tabell

```
summary(aov(SPOTS~WATER+STRAIN))
```

```

          Df Sum Sq Mean Sq F value    Pr(>F)
WATER      3  403.40  134.47  20.6607 4.955e-05 ***
STRAIN     4   12.70    3.18   0.4878  0.7448
Residuals 12   78.10    6.51
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Vi lager en matrise med navn flekker og setter inn verdiene for spots:

```

flekker<-
matrix(c(6,17,5,7,5,14,7,14,5,17,6,14,6,16,5,7,6,19,5,5),nrow=4)
flekker

```

```

      [,1] [,2] [,3] [,4] [,5]
[1,]    6    5    5    6    6
[2,]   17   14   17   16   19
[3,]    5    7    6    5    5
[4,]    7   14   14    7    5

```



```
#Kjikkvadrattest
chisq.test(flekker)
```

Pearson's Chi-squared test

```
data: flekker
X-squared = 7.8323, df = 12, p-value = 0.7981
```

```
#kritiske verdi
qchisq(0.95,1)
[1] 3.841459
```

Appendiks oppgave 4

Vi ser på de tilsvarende kontingenstabellene:

	S1	S2	S3	S4	S5	Total
W1	6	5	5	6	6	28
W2	17	14	17	16	19	83
W3	5	7	6	5	5	28
W4	7	14	14	7	5	47
	35	40	42	34	35	186

Kontingenstabellen for melduggresistens i bygg

S1	S2	S3	S4	S5	Tot
$28/186 \cdot 35/186$	$28/186 \cdot 40/186$	$28/186 \cdot 42/186$	$28/186 \cdot 34/186$	$28/186 \cdot 35/186$	28
$83/186 \cdot 35/186$	$83/186 \cdot 40/186$	$83/186 \cdot 42/186$	$83/186 \cdot 34/186$	$83/186 \cdot 35/186$	83
$28/186 \cdot 35/186$	$28/186 \cdot 40/186$	$28/186 \cdot 42/186$	$28/186 \cdot 34/186$	$28/186 \cdot 35/186$	28
$47/186 \cdot 35/186$	$47/186 \cdot 40/186$	$47/186 \cdot 42/186$	$47/186 \cdot 34/186$	$47/186 \cdot 35/186$	47
35	40	42	34	35	186

Sannsynligheten for kombinasjonene

	S1	S2	S3	S4	S5	Total
W1	5.2688	6.0215	6.3226	5.1182	5.2688	28
W2	15.6182	17.8494	18.7419	15.1720	15.6182	83
W3	5.2688	6.0215	6.3226	5.1183	5.2688	28
W4	8.8441	10.1075	10.6129	8.5914	8.8441	47
	35	40	42	34	35	186

Forventet frekvens. Tallene i tabellen over *186

I GLM kan man spesifisere forskjellige error-strukturer for eksempel **family=poisson** eller family = binomial. Linkfunksjone relaterer middelveiden til responsvariabel til dens lineære prediktor. Hvis man spesifiserer family settes følgende linkfunksjoner automatisk:

Error	Linkfunksjon
Normalfordelt	identitet
Poissonfordelt	log
Binomialfordelt	logit
Gammafordelt	resiprok

I generelle additive modeller (gam) bruker man glattingsfunksjon `s()`, som egentlig er "splines".

Modell med interaksjon:

```
model<-glm(SPOTS~WATER*STRAIN,poisson)
summary(model)
```

Modell uten interaksjon:

```
model2<-glm(SPOTS~WATER+STRAIN,poisson)
summary(model2)
model3<-glm(SPOTS~WATER,poisson)
summary(model3)
```

Call:

```
glm(formula = SPOTS ~ WATER, family = poisson)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-1.57711 -0.25829 -0.02519  0.16708  1.39776
```

Coefficients:

```
              Estimate Std. Error  z value Pr(>|z|)
(Intercept)  1.723e+00  1.890e-01   9.116 < 2e-16 ***
WATERW2      1.087e+00  2.185e-01   4.972 6.62e-07 ***
WATERW3     -1.233e-16  2.673e-01 -4.62e-16  1.0000
WATERW4      5.179e-01  2.387e-01   2.170  0.0300 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 49.6844 on 19 degrees of freedom
Residual deviance: 9.3124 on 16 degrees of freedom
AIC: 96.489
```

Number of Fisher Scoring iterations: 4

Test om model2 er signifikant forskjellig fra model3:

```
anova(model2,model3,test="Chi")
```

Analysis of Deviance Table

```
Model 1: SPOTS ~ WATER + WATER:STRAIN
Model 2: SPOTS ~ WATER
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1         0 -5.329e-15
2        16  9.3124 -16  -9.3124  0.9000
```

```
par(mfrow=c(2,2))
plot(model3)
```

```
#ANOVA-tabell
```

```

model<-aov (SPOTS~WATER+STRAIN)
summary (model)
      Df Sum Sq Mean Sq F value    Pr(>F)
WATER   3 403.40  134.47 20.6607 4.955e-05 ***
STRAIN   4  12.70    3.18  0.4878  0.7448
Residuals 12  78.10    6.51
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Tar kvadratroten av SPOTS:

```
SQRTSP<-c (sqrt (SPOTS))
```

```

model4<-glm (SQRTSP~WATER+STRAIN)
summary (model4)

```

```

par (mfrow=c (2,2))
plot (model4)

```

Vi kan prediktere forventede verdier. Funksjonen type="response" gjør at man ikke behøver å tilbaketransformere.

```
yv<-predict (model2, type="response")
```

```

yv
      1          2          3          4          5          6          7          8
5.268817 6.021505 6.322581 5.118280 5.268817 15.618280 17.849462 18.741935
      9         10         11         12         13         14         15         16
15.172043 15.618280 5.268817 6.021505 6.322581 5.118280 5.268817 8.844086
      17         18         19         20
10.107527 10.612903 8.591398 8.844086

```

Dette ser man er omtrent tallene i tabellen vist foran.

Oppgave 5: Ikke-lineær regresjon

Datasettet Puromycin inneholder reaksjonshastighet (rate, cpm/min) for en enzymreaksjon versus substratkonsentrasjon (conc, ppm) i celler med eller uten behandling med puromycin (state, treated/untreated). Vm er maksimal reaksjonshastighet, Km er substratkonsentrasjonen ved halvparten av maksimal reaksjonshastighet.

```

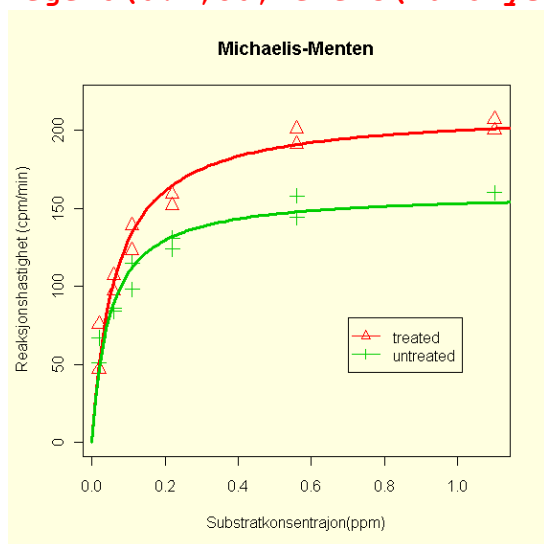
library (datasets)
data (Puromycin)
attach (Puromycin)
names (Puromycin)
Puromycin
?Puromycin
#Michaelis-Menten modell
par (bg="lightyellow")
#ikke-lineær modell nls for "treated"
nls1<-nls (rate~Vm*conc / (Km+conc), data=Puromycin,
subset=state=="treated", start=c (Vm=200, Km=0.05))
#ikke-lineær modell nls for "untreated"

```

```

mm2<-nls (rate~Vm*conc/ (Km +
conc) ,data=Puromycin ,subset=state=="untreated" ,
start=c (Vm=160 ,Km=0.05) )
plot (conc[state=="treated" ] ,rate[state=="treated" ] ,
ylim=c (0 ,220) ,pch=2 ,col=2 ,cex=1.5 ,
xlab="Substratkonsentrajon (ppm) " ,
ylab=" Reaksjonshastighet (cpm/min) " ,main="Michaelis-Menten")
points (conc[state=="untreated" ] ,rate[state=="untreated" ] ,col=3
,pch=3 ,cex=1.5)
#plotter predikterte linjer ifølge modellene
x2<-seq(0 ,1.2 ,0.01)
ym<-predict (mm1 ,list (conc=x2) )
lines (x2 ,ym ,col=2 ,lwd=3)
ym2<-predict (mm2 ,list (conc=x2) )
lines (x2 ,ym2 ,col=3 ,lwd=3)
legend (0.7 ,80 ,levels (Puromycin$state) ,col=2:3 ,lty=1 ,pch=2:3)

```



```

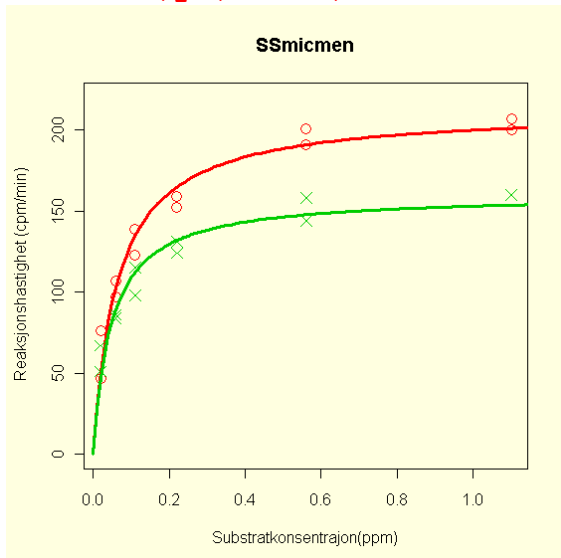
#eller via den selvstyrende funksjonen SSmicmen
#ikke-lineær regresjon selvstartende funksjon
plot (conc[state=="treated" ] ,rate[state=="treated" ] ,ylim=c (0 ,220) ,
col=2 ,cex=1.5 ,
xlab="Substratkonsentrajon (ppm) " , ylab=" Reaksjonshastighet
(cpm/min) " ,
main="SSmicmen")
#modell1
mod1<-
nls (rate~SSmicmen (conc ,Vm ,K) ,data=Puromycin ,subset=state=="tre
ated")
summary (mod1)
x2<-seq(0 ,1.2 ,0.01)
y3<-predict (mod1 ,list (conc=x2) )
lines (x2 ,y3 ,col=2 ,lwd=3)
points (conc[state=="untreated" ] ,rate[state=="untreated" ] ,col=3
,pch=4 ,cex=1.5)
#modell2

```

```

mod2<-
nls (rate~SSmicmen (conc ,Vm,K) ,data=Puromycin ,subset=state=="untreated")
summary (mod2)
y4<-predict (mod2 ,list (conc=x2) )
lines (x2 ,y4 ,col=3 ,lwd=3)

```



Formula: $\text{rate} \sim \text{SSmicmen}(\text{conc}, \text{Vm}, \text{K})$

Parameters:

	Estimate	Std. Error	t value	Pr(> t)	
Vm	1.603e+02	6.480e+00	24.734	1.38e-09	***
K	4.771e-02	7.782e-03	6.131	0.000173	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.773 on 9 degrees of freedom

Number of iterations to convergence: 5

Achieved convergence tolerance: 3.942e-06

Eksempler på andre selvstartende funksjoner:

SSasymp: asymptotisk regresjonsmodell

SSbiexp: biekspensiell modell

SSfpl: fire parameters logistisk modell

SSfol: første ordens kompartimentmodell

SSgompertz: Gompertz vekstmodell

SSlogis: logistisk regresjon

SSweibull: Weibull vekstmodell

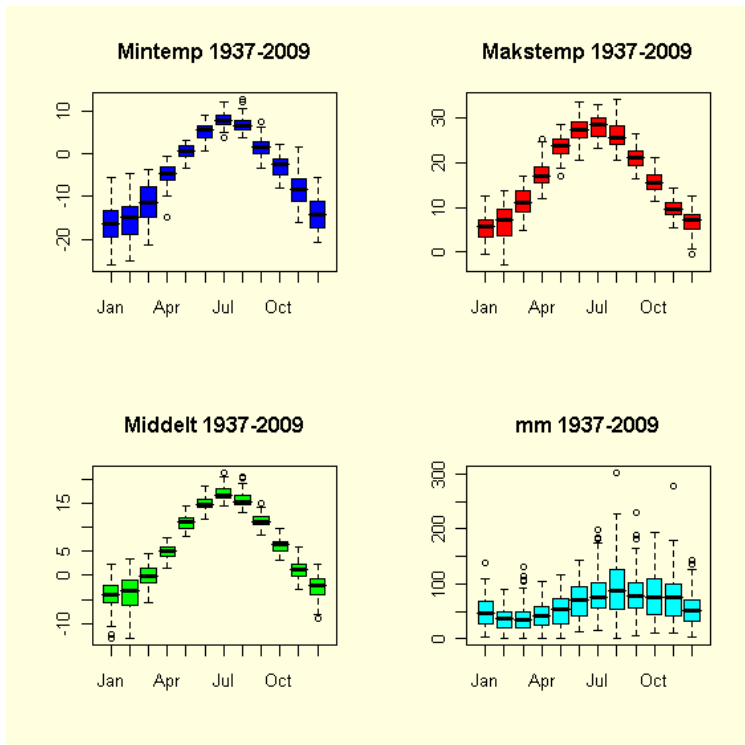
Oppgave 6: Tidsserieanalyse av klimadata fra Blindern

Observasjonsstudium: I dag mener man at observerte klimaendringer skyldes antropogene utslipp av karbondioksid (CO_2) fra forbrenning av fossilt brensel, metan (CH_4) fra landbruk og petroleumsindustri, lystgass (N_2O) fra

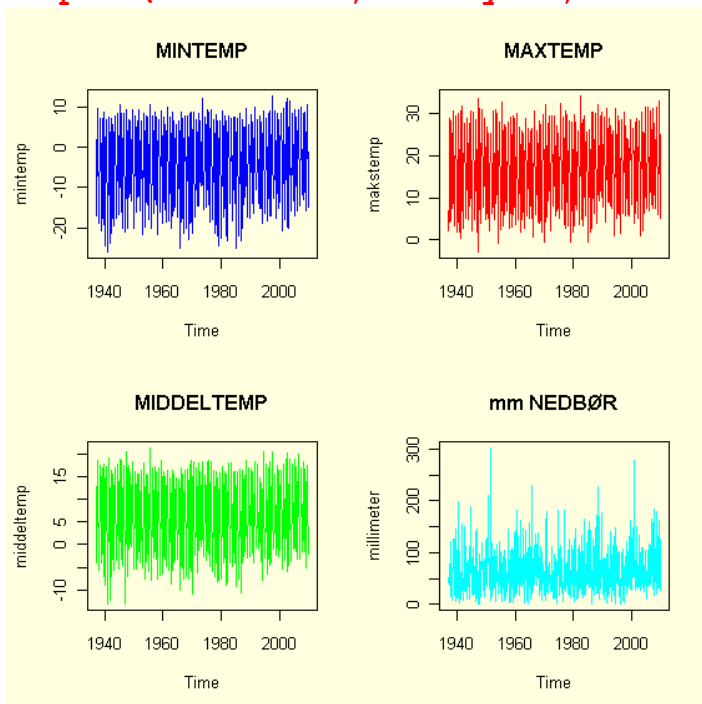
mineralgjødning og andre utslipp av drivhusgasser. Du har fått i oppgave å gå igjennom klimadata fra Blindern fra 1937-2009 for å undersøke om man kan observere en trend i klima. Datasettet inneholder månedlig minimums-, maksimums- og middeltemperatur samt millimeter nedbør. Datasettet er ikke helt komplett for desember 2009.

Laster inn datasettet

```
oppg6<-  
read.table("http://www.mn.uio.no/bio/tjenester/kunnskap/plantefys/matematikk/18700ny2.txt",header=T)  
eller:  
oppg6<-read.table("//platon/bio-kurs/bio2150/R/18700ny2.txt",header=T)  
attach(oppg6)  
names(oppg6)  
[1] "AAR" "MINT" "MDLT" "MAXT" "MM"  
#Boksplot for temperatur og nedbør  
mintemp<-ts(MINT,start=c(1937,1),frequency=12)  
makstemp<-ts(MAXT,start=c(1937,1),frequency=12)  
middeltemp<-ts(MDLT,start=c(1937,1),frequency=12)  
millimeter<-ts(MM,start=c(1937,1),frequency=12)  
par(mfrow=c(2,2),bg="lightyellow")  
boxplot(split(mintemp,cycle(mintemp)),col="blue",names=month.abb,main="Mintemp 1937-2009")  
boxplot(split(makstemp,cycle(makstemp)),col="red",names=month.abb,main="Makstemp 1937-2009")  
boxplot(split(middeltemp,cycle(middeltemp)),col="green",names=month.abb,main="Middelt 1937-2009")  
boxplot(split(millimeter,cycle(millimeter)),col="cyan",names=month.abb,main="mm 1937-2009")
```



```
ts.plot(mintemp, col="blue", main="MINTEMP")
ts.plot(makstemp, col="red", main="MAXTEMP")
ts.plot(middeltemp, col="green", main="MIDDELTEMP")
ts.plot(millimeter, col="cyan", main="mm NEDBØR")
```



summary (oppg6)

AAR	MINT	MDLT	MAXT
Min. : 1.194	Min. : -26.000	Min. : -13.100	Min. : -2.90
1st Qu.: 3.946	1st Qu.: -11.700	1st Qu.: -0.625	1st Qu.: 9.00
Median : 6.697	Median : -3.400	Median : 5.650	Median : 16.20
Mean : 6.697	Mean : -4.302	Mean : 5.994	Mean : 16.61
3rd Qu.: 9.449	3rd Qu.: 3.400	3rd Qu.: 13.000	3rd Qu.: 24.30
Max. : 12.201	Max. : 12.700	Max. : 21.300	Max. : 34.20

```

MM
Min.   : 0.00
1st Qu.: 34.00
Median : 57.95
Mean   : 63.72
3rd Qu.: 84.20
Max.   :303.00

```

#Minimumstemperatur

```
index<-1:876
```

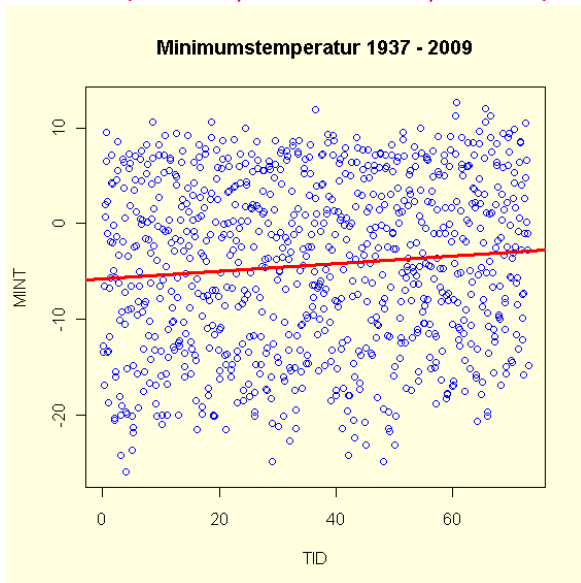
```
TID<-index/12
```

```
par(bg="lightyellow")
```

```
plot(TID,MINT,col="blue",main="Minimumstemperatur 1937 - 2009")
```

```
model<-lm(MINT~TID)
```

```
abline(model,col="red",lwd=3)
```



```
summary(model)
```

Call:

```
lm(formula = MINT ~ TID)
```

Residuals:

Min	1Q	Median	3Q	Max
-21.0586	-7.5691	0.7808	7.6272	16.2002

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-5.74919	0.60071	-9.571	< 2e-16 ***
TID	0.03961	0.01424	2.781	0.00553 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.882 on 874 degrees of freedom

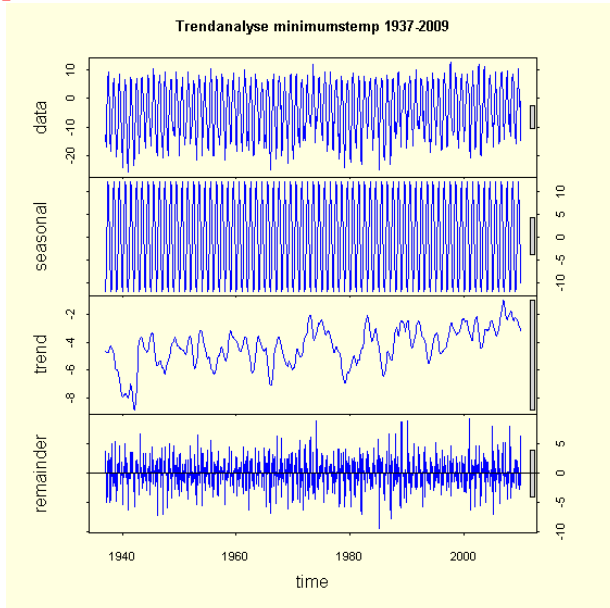
Multiple R-squared: 0.008774, Adjusted R-squared: 0.00764

F-statistic: 7.736 on 1 and 874 DF, p-value: 0.005529

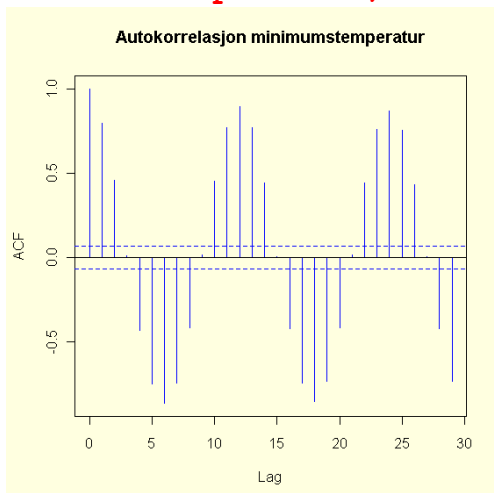
Nullhypotesen forkastes, det har skjedd en økning i minimumstemperaturen på Blindern I perioden 1937-2009:

#Trendanalyse

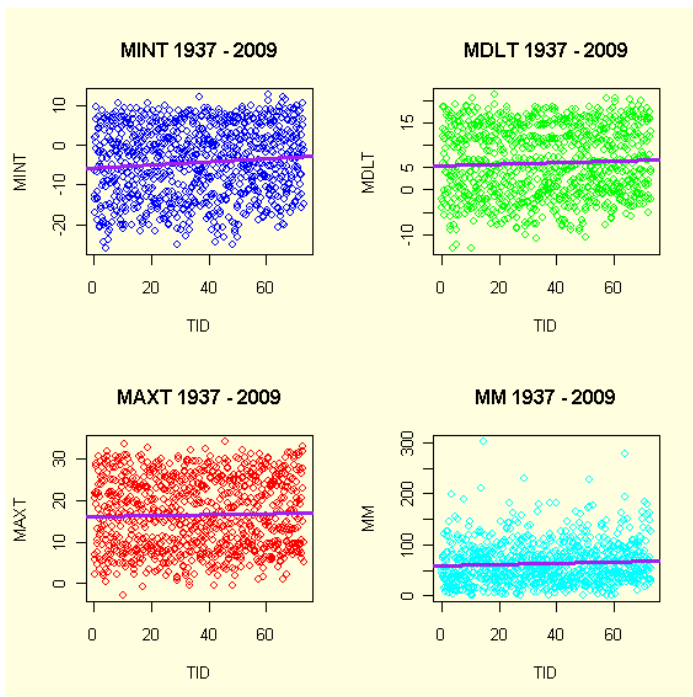

```
minimum<-stl(mintemp,"period")
plot(minimum,col="blue",main="Trendanalyse minimumstemp 1937-2009")
```



```
#Autokorrelasjon for studier av periodisitet
acf(MINT,col="blue", main="Autokorrelasjon
minimumstemperatur")
```



Tilsvarende for maksimumstemperatur, minimumstemperatur, og millimeter nedbør:



Det har ikke skjedd noen signifikant endring i middeltemperatur ($p=0.120$) og maksimumstempertur ($p=0.333$) i perioden. Nullhypotesen forkastes, det har skjedd en økning i minimumstemperaturen på Blindern i perioden 1937-2009 (0.00553). For nedbør en svak økning ($p=0.0376$)

Litteratur:

Crawley, M.J.: *The R book* John Wiley & Sons, Ltd. 2007.

Dahlggaard, P.: *Introductory statistics with R*. 2e. Springer 2008.

Grafen, A. & Hails, R.: *Modern statistics for the life sciences*. Oxford University Press 2003.

Whitlock, M.C. & Schluter, D.: *The analysis of biological data*. Roberts and Company publ. 2009.

