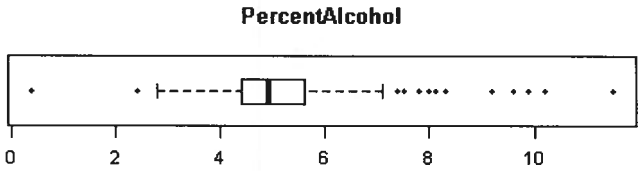
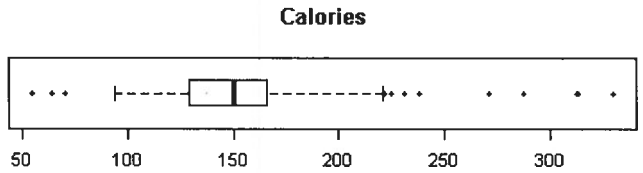
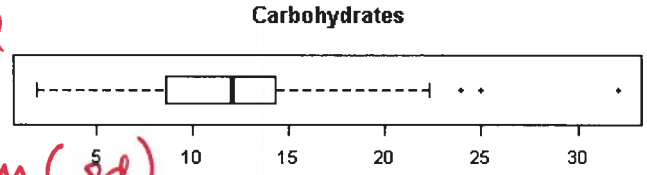
```
attach(ex02.26beer)

par(mfrow=c(3,1))
boxplot(PercentAlcohol,horizontal=TRUE,main="PercentAlcohol")
boxplot(Calories,horizontal=TRUE,main="Calories")
boxplot(Carbohydrates,horizontal=TRUE,main="Carbohydrates")
```







Ingen av disse er helt symmetriske, men ingen er heller helt skjeve. Ser at det er outliere i alle tre variabler, og regner derfor med at sd er noe større enn det vi vil forvente ved n-ford. Velger likevel å oppsummere vha mean(sd)

$\bar{x} = 5.2$, sd = 1.4

$\bar{x} = 154$, sd = 44

$\bar{x} = 12.0$, sd = 4.9

```
> c(summary(PercentAlcohol),sd(PercentAlcohol))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.400000  4.400000  4.900000  5.229000  5.600000 11.500000  1.428737
> c(summary(Calories),sd(Calories))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 55.00000 129.00000 150.00000 154.10000 166.00000 330.00000  44.49254
> c(summary(Carbohydrates),sd(Carbohydrates))
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 1.900000  8.600000 12.000000 11.960000 14.300000 32.100000  4.905587
```

sd

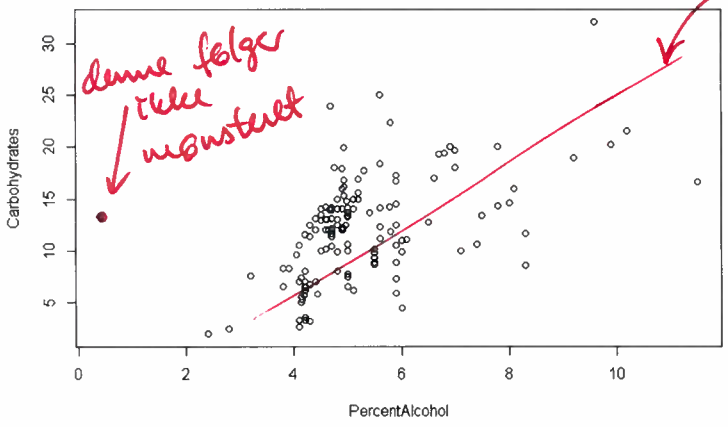
```
which.min(PercentAlcohol)
> which.min(PercentAlcohol)
[1] 104
> PercentAlcohol[104]
[1] 0.4
> PercentAlcohol[which.min(PercentAlcohol)]
[1] 0.4
```

identifiserer outlier

tall nr 104 i vektoren

fjerner den fra plottet

```
plot(PercentAlcohol,Carbohydrates)
```

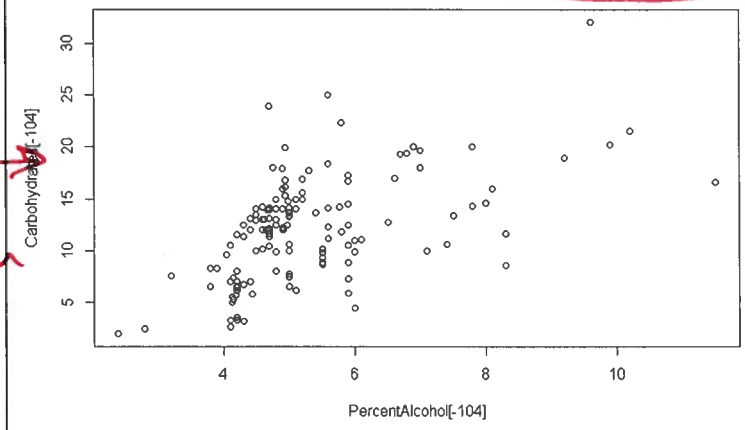


denne følger ikke mønsteret

håndtegnet

samme plott, men uten outlier.

```
plot(PercentAlcohol[-104],Carbohydrates[-104])
```



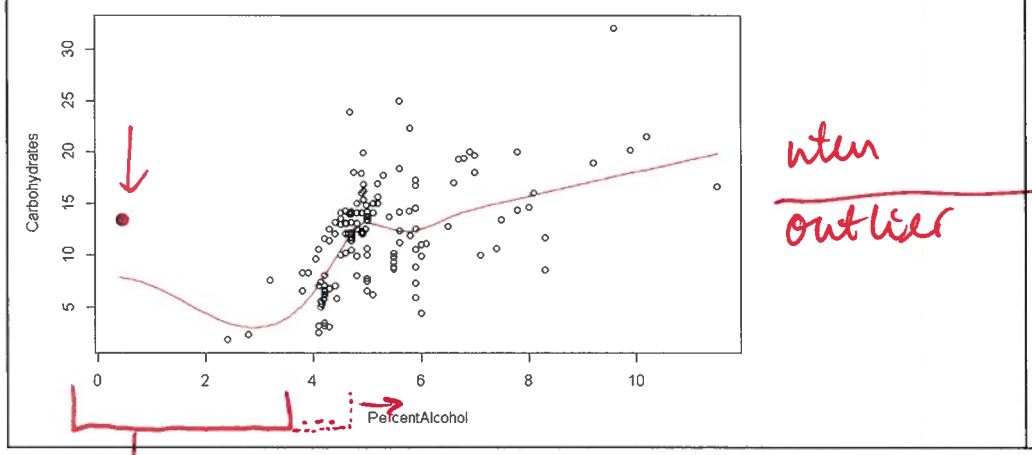
generelt: høye verdier av % alkohol → høye karbohydrat-tall, og } positiv sammenheng
lave ——— lave ———

Forventer en positiv korrelasjon.

Ingen klare ikke-lineære smh, som f.eks  etc.

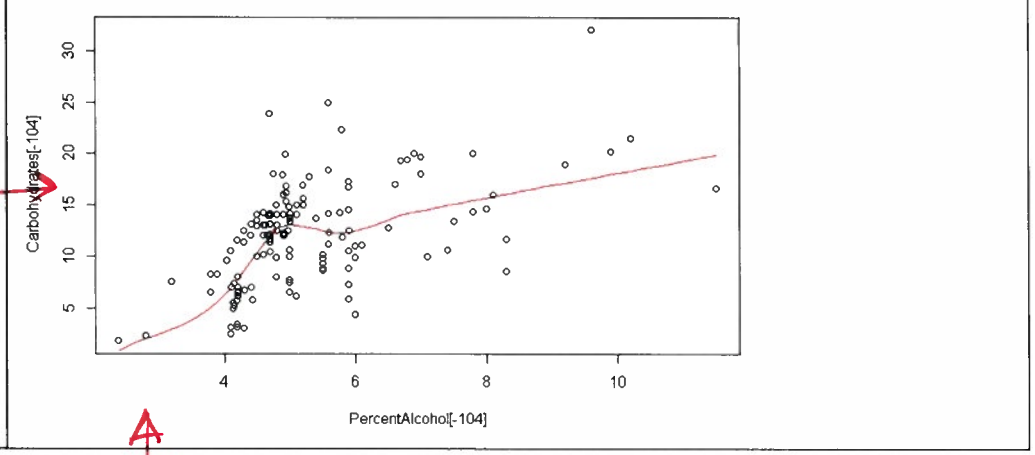
↳ en linear smh fanger opp det generelle i sammenhengen.

```
lines(loess.smooth(PercentAlcohol,Carbohydrates),col="red")
```



uten outlier

```
lines(loess.smooth(PercentAlcohol[-104],Carbohydrates[-104]))
```



Lokalt tilpasset regresjonslinje: tilpasse linja til et glidende vindu, for å se hva data fortelle om sammenhengen. Viktig å ikke legge stor vekt på endene. Der er det lite data, og den lokale tilpassingen blir usikker.

Se relativt lineært ut, hvis vi ser litt stort på det. Vi har i mest fall ingen (kjemiske) teorier som kan forklare hvorfor det evt skulle være en "kul" i midten, og da vil vi heller ikke tilpasse parametere til det kurve-

```
cor(PercentAlcohol,Carbohydrates)
cor.test(PercentAlcohol,Carbohydrates)
cor(PercentAlcohol,Carbohydrates)^2
```

```
> cor(PercentAlcohol,Carbohydrates)
[1] 0.5210898
> cor.test(PercentAlcohol,Carbohydrates)

        Pearson's product-moment correlation

data:  PercentAlcohol and Carbohydrates
t = 7.5023, df = 151, p-value = 5.005e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3950790 0.6278535
sample estimates:
      cor
0.5210898

> cor(PercentAlcohol,Carbohydrates)^2
[1] 0.2715346
```

1A

```
cor(PercentAlcohol[-104],Carbohydrates[-104])
cor.test(PercentAlcohol[-104],Carbohydrates[-104])
cor(PercentAlcohol[-104],Carbohydrates[-104])^2
```

```
> cor(PercentAlcohol[-104],Carbohydrates[-104])
[1] 0.5484863
> cor.test(PercentAlcohol[-104],Carbohydrates[-104])

        Pearson's product-moment correlation

data:  PercentAlcohol[-104] and Carbohydrates[-104]
t = 8.0338, df = 150, p-value = 2.576e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4265300 0.6508548
sample estimates:
      cor
0.5484863

> cor(PercentAlcohol[-104],Carbohydrates[-104])^2
[1] 0.3008373
```

1B

Korrelasjon 0.52 med outlier, 0.55 uten. Positiv korrelasjon som forventet, og den øker når outlier er borte.

Hypotesetest: Parameter: Korrelasjonen i populasjonen, ρ $\left(= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}\right)$

H_0: Ingen (null) sammenheng; $\rho = 0$

H_1: Sammenheng mellom alkohol% og karbohydrater; $\rho \neq 0$

Velger tosidig hypotese delvis av konvensjon & fordi det er konservativt, delvis fordi jeg vet svært lite om dette og kan derfor ikke utelukke både negative og positive sammenhenger.

Kanskje trengs mer sukker for å få høyere % alkohol → positiv smh } Tosidig hyp.
Kanskje brukes mest sukker når det lages alkohol → negativ smh

Bruker r, altså korrelasjonen i utvalget, som testobservator.

Velger signifikansnivå 0.01 denne gang.

Hva betyr signifikansnivå?

Det maksimalt akseptable nivå for type I-feil. Vi ønsker altså at

$$P(\text{type I-feil}) = P(\text{Forkaste } H_0 \mid H_0 \text{ sann}) = P(\text{"}r \text{ er veldig stor (høy positiv korrelasjon) eller veldig liten (høy negativ korrelasjon)"}, \text{til tross for at } \rho = 0) \leq 0.01$$

For å finne ut hvor stor $P(\text{type I-feil})$ vi ville fått dersom vi hadde forkastet H_0 med de observasjonene vi har nå, beregner vi p-verdien:

p-verdi = "signifikanssannsynligheten" = den sannsynligheten for type I feil vi måtte ha godtatt dersom vi valgte å forkaste H_0 på bakgrunn av det vi har observert

$$= P(\text{minst like ekstreme observasjoner som det vi har observert} \mid H_0)$$

$$= P(|r| > 0.52 \mid \rho = 0) = 0.0000000000005005$$

Fra R-utskrift

Ok, dette var lite sannsynlig under H_0. Forkast H_0.

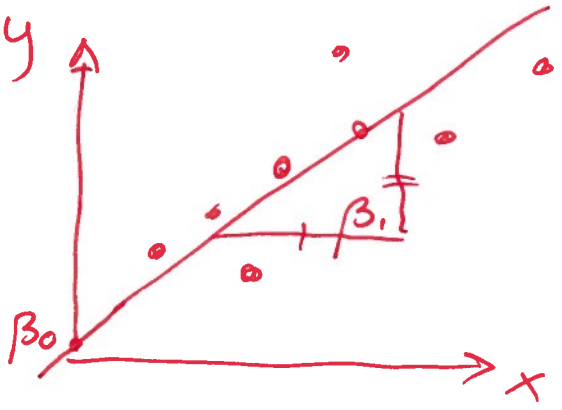
```
lm(Carbohydrates~PercentAlcohol)

> lm(Carbohydrates~PercentAlcohol)

Call:
lm(formula = Carbohydrates ~ PercentAlcohol)

Coefficients:
   (Intercept)  PercentAlcohol
         2.605           1.789
```

β_0 β_1

```
lm(Carbohydrates[-104]~PercentAlcohol[-104])

> lm(Carbohydrates[-104]~PercentAlcohol[-104])

Call:
lm(formula = Carbohydrates[-104] ~ PercentAlcohol[-104])

Coefficients:
          (Intercept)  PercentAlcohol[-104]
                1.650                 1.958
```

En annen måte å tallfeste den lineære sammenhengen på, er å beregne den linja som passer best til data. La y = Carbohydrates, x = PercentAlcohol

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$



β_0 angir hvor linja skjærer y-aksen (intercept)

β_1 angir stigningstallet til linja, altså hvor mange enheter vi forventer at y skal øke når x øker med én enhet

Her er $\hat{\beta}_0 = 2.605$

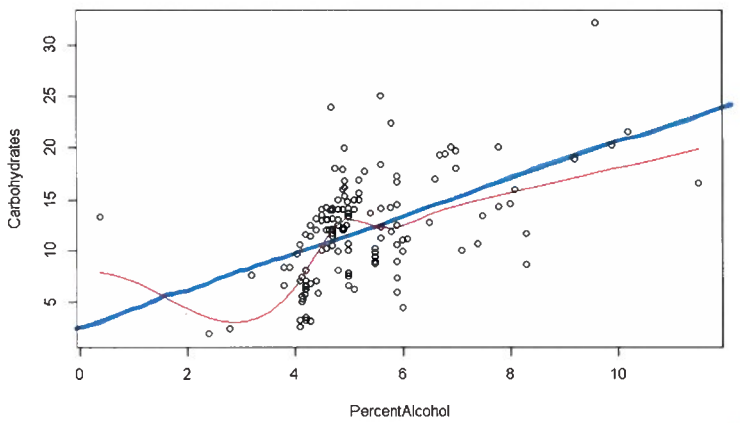
$\hat{\beta}_1 = 1.789$

→ det endres seg hvis vi fjerner outlieren. Nye estimat:

$\hat{\beta}_0 = 1.650$

$\hat{\beta}_1 = 1.958$

```
?abline
abline(2.605,1.789,col="blue")
```

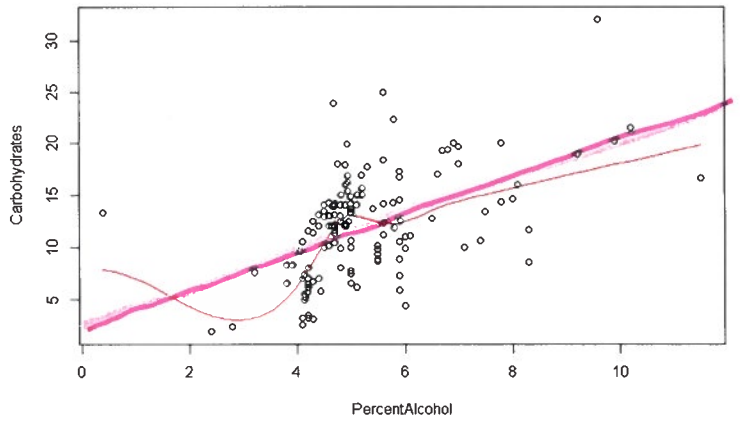



```
# abline(2.605,1.789,col="blue")
```

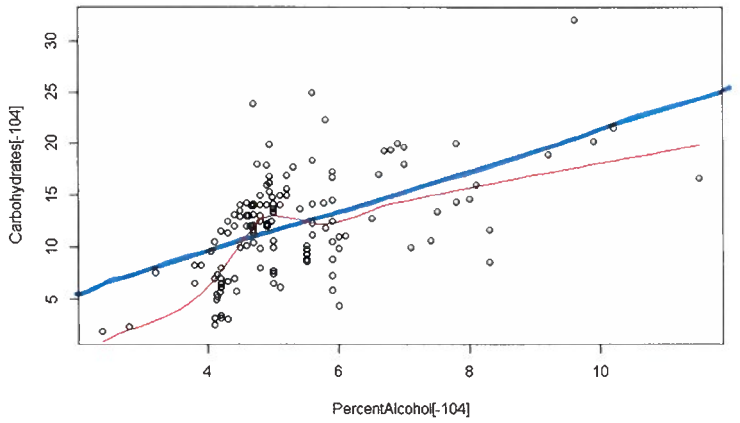
Disse estimatene måtte byttes ut med 1.650 , 1.958

Velger heller den generelle måten:

```
abline(lm(Carbohydrates~PercentAlcohol),col="pink")
```




```
abline(lm(Carbohydrates[-104]~PercentAlcohol[-104]),col="blue")
```



Tegner inn denne linja på to ulike måter i R

```
summary(lm(Carbohydrates~PercentAlcohol))
> summary(lm(Carbohydrates~PercentAlcohol))

Call:
lm(formula = Carbohydrates ~ PercentAlcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-8.940 -3.145  0.586  2.628 12.904

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.6049     1.2924   2.016   0.0456 *
PercentAlcohol    1.7892     0.2385   7.502    5e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.201 on 151 degrees of freedom
Multiple R-squared:  0.2715,    Adjusted R-squared:  0.2667
F-statistic: 56.29 on 1 and 151 DF,  p-value: 5.005e-12
```

β0

β1

```
summary(lm(Carbohydrates[-104]~PercentAlcohol[-104]))
> summary(lm(Carbohydrates[-104]~PercentAlcohol[-104]))
```

uten outlier

```
Call:
lm(formula = Carbohydrates[-104] ~ PercentAlcohol[-104])

Residuals:
    Min      1Q  Median      3Q     Max
-9.3034 -3.0670  0.7311  2.6830 13.0661

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.6495     1.3253   1.245    0.215
PercentAlcohol[-104]  1.9583     0.2438   8.034 2.58e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.128 on 150 degrees of freedom
Multiple R-squared:  0.3008,    Adjusted R-squared:  0.2962
F-statistic: 64.54 on 1 and 150 DF,  p-value: 2.576e-13
```

Men $\hat{\beta}_0$ og $\hat{\beta}_1$ er som sagt estimater, og har usikkerhet knyttet til seg.
Vi kan både gjøre hypotesetester og beregne konfidensintervaller.
La oss ta hypotesetest først:

H_0 : Ingen (null) sammenheng mellom alkohol% og karbohydrater : $\beta_1 = 0$

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

Velger tosidige hypoteser av samme grunn som tidligere.

(OBS: R tester for om $\beta_0 = 0$ også, men det er ikke det som er interessant her.)

p-verdi = P(å observere minst like ekstreme verdier for $\hat{\beta}_1$ som det vi har observert $| \beta_1 = 0$)

$= P(|\hat{\beta}_1| > 1.7892 \; | \beta_1 = 0)$ = 0.000000000005005

Forkast H_0

i følge R-utskrift

Eksempel på at selv om effektmålet (r og $\hat{\beta}_1$) er forskjellig, så er metodene basert på samme antakelser og samme matematikk, og gir like konklusjoner.

Konfidensintervall for β_1:

Ta utgangspunkt i at $\frac{\hat{\beta}_1 - \beta_1}{\text{std.Error}(\hat{\beta}_1)} \sim T_{n-2}$ Her er $n = 153$, og $df = 151$

Da er

$$P\left(-t_{151,0.025} < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < t_{151,0.025}\right) = 0.95$$

$$P\left(-t_{151,0.005} < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < t_{151,0.005}\right) = 0.99$$

velger denne

vi finner i table D : $t_{100,0.025} = 1.984$ $t_{100,0.005} = 2.626$

eg $t_{1000,0.025} = 1.962$ $t_{1000,0.005} = 2.581$

$$P\left(-1.984 < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < 1.984\right) = 0.95$$

-2.626 2.686 0.99

$$P\left(\hat{\beta}_1 - 1.984 \cdot SE(\hat{\beta}_1) < \beta_1 < \hat{\beta}_1 + 1.984 \cdot SE(\hat{\beta}_1)\right) = 0.95$$

-2.686 $+2.686$ 0.99

99% KI for β_1 :

$1.7892 \pm 2.686 \cdot 0.2385 \rightarrow [1.15, 2.43]$

H_0-verdien for β_1

intervallet inneholder ikke 0

→ Forkast H_0 på nivå 0.01

95% KI for β_1 :

$1.7892 \pm 1.984 \cdot 0.2385 \rightarrow [1.32, 2.26]$

intervallet inneholder ikke 0

→ forkast H_0 på nivå 0.01