Matematisk Statistik: Modelbaseret Inferens

Lineær regression

Jens Ledet Jensen

Sammenhænge

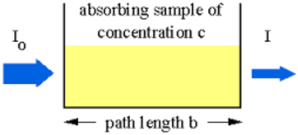
Specielt lineære sammenhænge

Dagens spørgsmål: hvor meget stiger bremselængden når hastigheden øges med 10 km/t ?

Beer-Lamberts lov: $\log \frac{I_0}{I} = \epsilon b c$

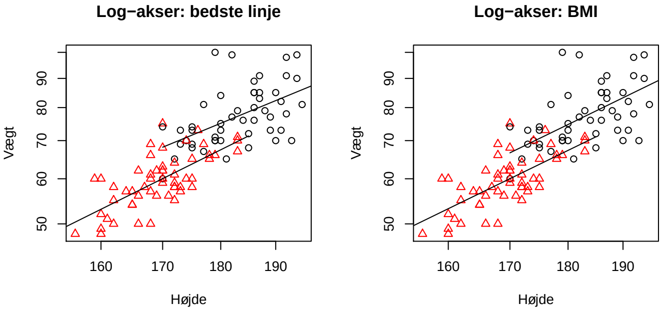
I, I_0: lysintensitet før og efter beholder

b: vejlængde, c: koncentration



Når vi kender koncentrationen kan vi sige hvor meget lys der absorberes

Sammenhængen er kausal: koncentration er årsag til absorption



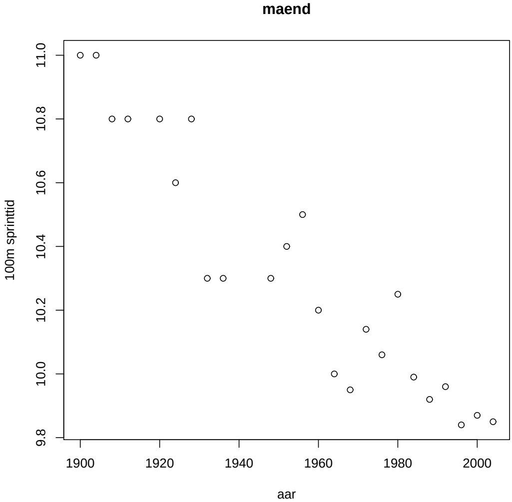
Ingen kausal sammenhæng: jeg kan ikke ud fra højden sige hvad vægten er

Men, der er ret stor sandsynlighed for at person A vejer mere end person B hvis A er 30 cm højere end B

Vægten kovarierer med højden (de er korrelerede)

Højden fortæller mig noget om fordelingen af vægten

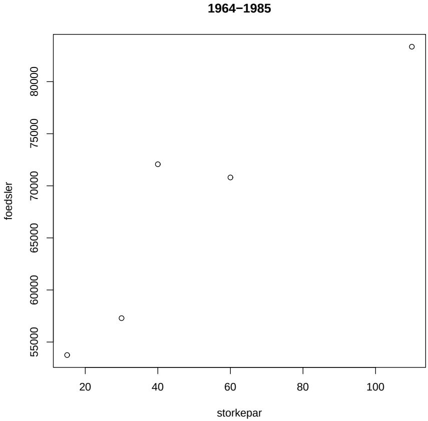
100 m sprinttid



OL år

Kausal sammenhæng / Biologisk samvariation ?

antal nyfødte



antal storke

https://www.tylervigen.com/spurious-correlations

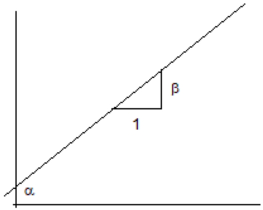
Forskelige former for sammehænge er omtalt

Næste: lineær sammenhæng og normalfordelte data

Ønsker at sige noget om respons x_i ud fra værdi af forklarende variabel t_i

$i = 1, \ldots, n$: observationsnummer

Lineære regressionsmodel:



$$E(X_i) = \alpha + \beta \cdot t_i$$

α: skæring med andenaksen

β: hældning, når t stiger med 1 stiger middelværdi med β

Forstå sammenhæng: estimere linjen, dvs α og β

Hvor meget mere absorberes når koncentration stiger med 1 g/cm^3 ?

Er der proportionalitet mellem antal kodelinjer i et program og cyclomatic complexity ?

Forstå linjens forklaringskraft: variationen omkring linjen

Hvor godt kender jeg vægten ud fra højden ?

Invers regession: måler x, hvor meget ved jeg så om t ?

absorptionen måles til 0.3, hvad var koncentrationen ?

Prediktere x ud fra t, skøn over $\alpha + \beta t_0$

Hvad er middelvægten når højden er 175 cm ?

Fordeling af respons X_i:

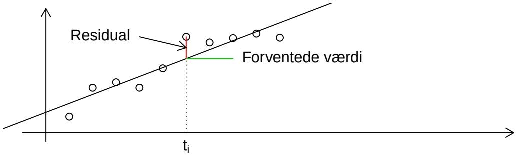
varians afhænger ikke af t_i: $\mathrm{Var}(X_i) = \sigma^2$

X_i er normalfordelt: $X_i \sim N(\alpha + \beta t_i, \sigma^2)$

Når skøn $\hat{\alpha}$ og $\hat{\beta}$ er fundet, kaldes

$r_i = x_i - (\hat{\alpha} + \hat{\beta} t_i)$ for residual, og

$\hat{\xi}_i = \hat{\alpha} + \hat{\beta} t_i$ kaldes for den forventede værdi (fitted value)



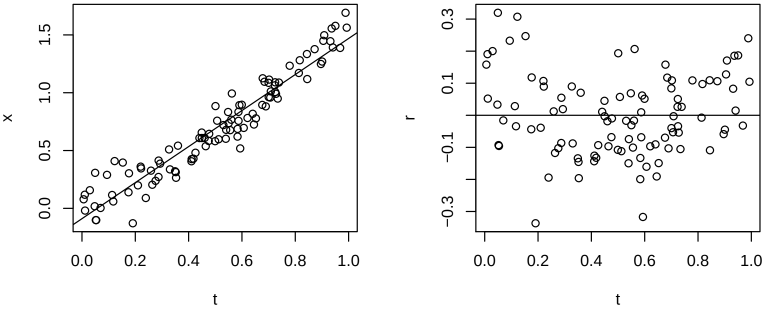
Start altid med en figur hvor x_i er afsat mod t_i

Lav dernæst skøn over α og β og lav:

Residualplot: r_i afsættes mod t_i (eller mod $\hat{\xi}_i$)

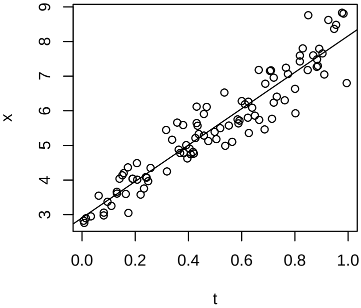
- se efter systematiske afvigelser
- se efter ikke-konstant varians

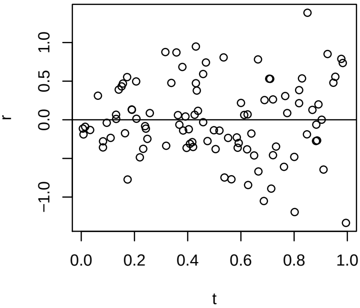
QQplot af residualer for at tjekke normalfordelingsantagelse



Data er simuleret fra en $N(\ t_i + 0.6t_i^2,\ 0.1^2\ )$-fordeling

Analyse lavet i model $X_i \sim N(\alpha + \beta t_i, \sigma^2)$





Data er simuleret fra en $N(\ 3 + 5t_i,\ 1 + 0.1 \cdot t_i^2\ )$-fordeling

Analyse lavet i model $X_i \sim N(\alpha + \beta t_i, \sigma^2)$

Opgave (a): overvej om hastighed og bremselængde i "cars"-datasæt kan beskrives med den lineære regressionsmodel

Lad h_i og b_i være hastighed og bremselægde for den i-te måling.
Vi vil undersøge modellen $B_i \sim N(\alpha + \beta h_i, \sigma^2)$, $i = 1, \ldots, 50$

```
h=cars[,1]
b=cars[,2]

# afsætte bremselængde mod hastighed
plot(h,b)
# indtegne skønnede linje
abline(lm(b~h))

# residualplot
plot(h,lm(b~h)$residuals)

# qqplot af residualer
qqnorm(lm(b~h)$residuals)

# Figur tyder både på systematisk afvigelse og stigende varians
```

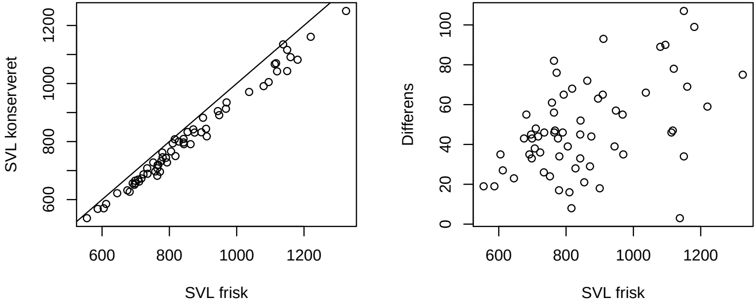
Opgave (b): Gentag undersøgelsen med *b* erstattet af sb=sqrt(b)

```
h=cars[,1]
sb=sqrt(cars[,2])

plot(h,sb)

abline(lm(sb~h))

plot(h,lm(sb~h)$residuals)

qqnorm(lm(sb~h)$residuals)

Konklusion:
```

Sammenhænge, specielt lineær sammenhæng, er omtalt

Overordnede formål med undersøgelse: ændrer slangerne sig på grund af temperaturstigning ?

Delproblem: Sammenhæng mellem længdemål, SVL (snout-vent-length) for frisk og for konserveret slange: prediktere friske ud fra konserverede



Forskel mellem konserveret og frisk bliver større, jo større frisk-længden er

Tyder på lineær regression med $\beta < 1$

Current Zoology Advance Access published January 5, 2017

Current Zoology, 112016, 1–7
doi: 10.1093/cz/zow112
Advance Access Publication Date: 25 December 2016
Article

OXFORD

Article

SVL_i, $Kons_i$: målte værdi af SVL for frisk henholdsvis konserveret slange

Model: $Kons_i \sim N(\alpha + \beta \cdot SVL_i, \sigma^2)$, $\quad i = 1, \ldots, 62$, uafhængige

Skøn over parametre (med generel notation: x, t, n):

maksimere likelihoodfunktion (produkt af tætheder):

tæthed for enkelt observation: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\left(x_i - (\alpha + \beta t_i)\right)^2 \right\}$

$L(\alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \left(x_i - (\alpha + \beta t_i)\right)^2 \right\}$

Mindste kvadraters metode:

Minimere $\sum_{i=1}^n \left(x_i - (\alpha + \beta t_i)\right)^2$ for at finde skøn $\hat{\alpha}$ og $\hat{\beta}$

Metode: differentiere og sætte lig med nul: Vis i webbog afsnit 3.2

$\hat{\beta} = \frac{SPD_{tx}}{SSD_t}$:

$SSD_t = \sum_{i=1}^{n}(t_i - \bar{t})^2$, sum of squared deviations

$SPD_{tx} = \sum_{i=1}^{n}(x_i - \bar{x})(t_i - \bar{t})$, sum of product of deviations

$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{t}$:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$

Slangedata: $\hat{\beta} = \frac{2121490}{2287244} = 0.9275$

$\hat{\alpha} = 810.7742 - 0.9275 \cdot 858.6452 = 14.3808$

Beregning i R: `lm(Kons~SVL)$coef`

Skøn over varians σ^2:

$$s_r^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(x_i - (\hat{\alpha} + \hat{\beta} t_i)\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(x_i - \hat{\xi}_i\right)^2$$

$\hat{\xi}_i = \hat{\alpha} + \hat{\beta} t_i$: forventede værdi = linjens værdi i t_i

Slangedata: $s_r^2 = 562.1823 = 23.7104^2$
(R: summary(lm(Kons~SVL))$sigma)

Generelt (lineær model M): $s^2(M) = \frac{1}{n-d}\sum_{i=1}^{n}\left(x_i - \hat{\xi}_i(M)\right)^2$

d = antal parametre i middelværdi

$\hat{\xi}_i(M)$: forventede værdi for i'te måling
= middelværdi med skønnede parametre indsat

$n-d$; sikrer at $E(s^2(M)) = \sigma^2$

s_r^2 er IKKE maksimum likelihood estimat fra $L(\alpha, \beta, \sigma^2)$

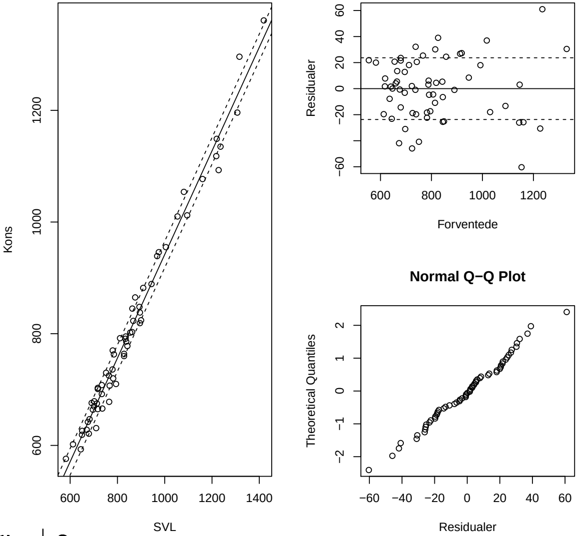
Men er maksimum likelihood estimat fra model:

$$\sum_i (X_i - \hat{\alpha} - \hat{\beta} t_i)^2 \sim \sigma^2 \chi^2(n-2)$$

Skift til eksempel 3.4 i webbog afsnit 3.2

Vis direkte beregning af $\hat{\alpha}$, $\hat{\beta}$ og s_r^2

vis alternativ: lm(x~t)$coef



Stiplede linjer: $\pm s_r$

Data med snout-vent-length af slanger er præsenteret og parameterestimater er fundet

Næste: Fordeling af parameterestimater $\rightarrow$ test+konfidensinterval
Fire tekniske slides

$X_1, \ldots, X_n$ er stokastiske $\rightarrow \hat{\beta}$, $\hat{\alpha}$ og s_r^2 er stokastiske
Hvad er fordelingen af $\hat{\beta}$, $\hat{\alpha}$ og s_r^2 ?

Ser på $\hat{\beta}$:

$$\hat{\beta} = \frac{SPD_{tx}}{SSD_t} = \frac{\sum_{i=1}^n (X_i - \bar{X})(t_i - \bar{t})}{SSD_t} = \sum_{i=1}^n X_i \frac{(t_i - \bar{t})}{SSD_t} \quad \sim N\big(\beta, \frac{\sigma^2}{SSD_t}\big)$$

X_i er stokastisk, $\quad \frac{(t_i - \bar{t})}{SSD_t}$ er fast

Regneregler:

"$a + b \cdot N(\mu, \sigma^2) = N(a + b\mu, b^2\sigma^2)$"

"$N(\mu_1, \sigma_1^2) + N(\mu_2, \sigma_2^2) = N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$"

s_r^2: husk at "$N(0,1)^2 + \cdots + N(0,1)^2 \sim \chi^2(k)$"

Model: $X_i \sim N(\alpha + \beta t_i,\ \sigma^2)$, $i = 1, \ldots, n$, uafhængige

$$\text{Skæring:} \quad \hat{\alpha}(X) \sim N\Big(\alpha, \sigma^2\Big[\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}\Big]\Big),$$

$$\text{Hældning:} \quad \hat{\beta}(X) \sim N\Big(\beta, \frac{\sigma^2}{SSD_t}\Big),$$

$$\text{Variansskøn:} \quad s_r^2(X) \sim \sigma^2\chi^2(n-2)/(n-2).$$

s_r^2 og $(\hat{\alpha}, \hat{\beta})$ er uafhængige

variansskøn er uafhængig af middelværdiskøn

Vise s_r^2 og $(\hat{\alpha}, \hat{\beta})$ er stokastisk uafhængige

$$s_r^2 = \sum R_i^2/(n-2),\ R_i = X_i - \hat{\alpha} - \hat{\beta} t_i$$

nok at vise at $(\hat{\alpha}, \hat{\beta})$ og $(R_1, \dots, R_n)$ er uafhængige

$(\hat{\alpha}, \hat{\beta})$ bestemt ved:

$R_1 + R_2 + R_3 + \cdots R_n = 0$ og
$t_1 R_1 + t_2 R_2 + t_3 R_3 + \cdots + t_n R_n = 0$

R_1, R_2 kan skrives som en funktion af $R_3, \dots, R_n$

nok at vise at $(\hat{\alpha}, \hat{\beta})$ og $(R_3, \dots, R_n)$ er uafhængige

Transformerer $(X_1, \ldots, X_n)$ over i $(\hat{\alpha}, \hat{\beta}, R_3, \ldots, R_n)$

Lineær transformation: Jacobiant er blot en konstant

Mangler blot at vise at tæthed for $(X_1, \ldots, X_n)$ kan skrives som

$g(\hat{\alpha}, \hat{\beta}) h(r_3, \ldots r_n)$

Tæthed: $\frac{1}{\sqrt{2\pi\sigma^2}^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (x_i - \alpha - \beta t_i)^2 \right\}$

se på $\sum_i (x_i - \alpha - \beta t_i)^2$

bruge: $x_i - \alpha - \beta t_i = x_i - \hat{\alpha} - \hat{\beta} t_i + \hat{\alpha} + \hat{\beta} t_i - \alpha - \beta t_i$

$= r_i - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) t_i$

Derfor: $\sum_i (x_i - \alpha - \beta t_i)^2 =$

$\sum_i r_i^2 - 2(\hat{\alpha} - \alpha) \sum_i r_i - 2(\hat{\beta} - \beta) \sum_i t_i r_i + \sum_i \left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) t_i\right)^2$

$= \sum_i r_i^2 + \sum_i \left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) t_i\right)^2$

$= \text{funktion}(r_3, \ldots, r_n) + \sum_i \left((\hat{\alpha} - \alpha) + (\hat{\beta} - \beta) t_i\right)^2$

idet $\sum_i r_i = 0$ og $\sum_i t_i r_i = 0$

Slut på bevis: retur til fordeling af $\hat{\alpha}$ og $\hat{\beta}$

$\text{sd}(\hat{\alpha}) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}}$, skøn over denne

$\text{sd}_s(\hat{\alpha}) = s_r\sqrt{\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}}$ kaldes standard error for $\hat{\alpha}$

$\text{sd}(\hat{\beta}) = \frac{\sigma}{\sqrt{SSD_t}}$, skøn over denne

$\text{sd}_s(\hat{\beta}) = \frac{s_r}{\sqrt{SSD_t}}$ kaldes standard error for $\hat{\beta}$

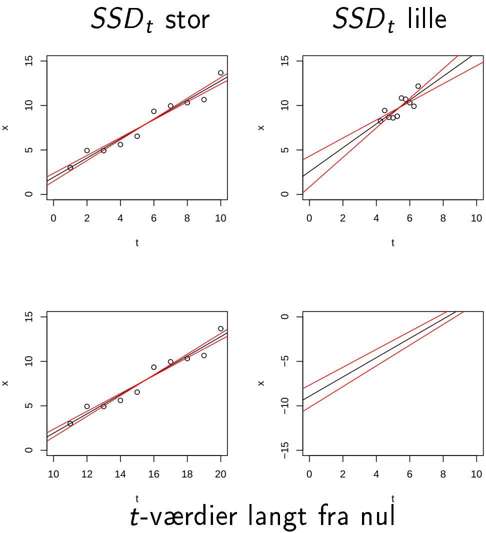
Generel t-teststørrelse: $\hat{\theta} \sim N(\theta, \sigma^2 C),\ s^2 \sim \sigma^2\chi^2(f)/f,\ \text{sd}_s(\hat{\theta}) = s\sqrt{C}$

$$T = \frac{\hat{\theta}-\theta_0}{\sqrt{s^2 C}} = \frac{\hat{\theta}-\theta_0}{\text{sd}_s(\hat{\theta})} \sim t(f)$$

Varians på estimater

$\hat{\beta}$: $\dfrac{\sigma^2}{SSD_t}$

$\hat{\alpha}$: $\sigma^2\left(\frac{1}{n} + \frac{\bar{t}^2}{SSD_t}\right)$



Teste $\beta = \beta_0$: $\quad T = \frac{\hat{\beta}-\beta_0}{\sqrt{s_r^2/SSD_t}} \sim t(n-2)$

$$\text{standard error for } \hat{\beta} = \sqrt{s_r^2/SSD_t}$$

Teste $\alpha = \alpha_0$: $\quad T = \frac{\hat{\alpha}-\alpha_0}{\sqrt{s_r^2\left(\frac{1}{n}+\frac{\bar{t}^2}{SSD_t}\right)}} \sim t(n-2)$

$$\text{standard error for } \hat{\alpha} = \sqrt{s_r^2\left(\frac{1}{n}+\frac{\bar{t}^2}{SSD_t}\right)}$$

95%-konfidensintervaller:

$$\hat{\beta} \pm t_0\sqrt{s_r^2/SSD_t}$$

$$\hat{\alpha} \pm t_0\sqrt{s_r^2\left(\frac{1}{n}+\frac{\bar{t}^2}{SSD_t}\right)}$$

$t_0 = t_{\text{inv}}(0.975, n-2)$, 97.5%-fraktil i $t(n-2)$-fordeling

De værdier af β der opfylder:

$$-t_0 \leq \frac{\hat{\beta}-\beta}{\sqrt{s_r^2/SSD_t}} \leq t_0$$

er de samme som værdierne i intervallet:

$$\hat{\beta} \pm t_0\sqrt{\frac{s_r^2}{SSD_t}} = [\hat{\beta} - t_0\sqrt{\frac{s_r^2}{SSD_t}},\ \hat{\beta} + t_0\sqrt{\frac{s_r^2}{SSD_t}}]$$

Generelt for konfidensinterval for middelværdiparameter i normalfordelingsmodel:

skøn $\pm$ $t_0\cdot$(standard error)

Hvis konservering forkorter længden af slanger med en procentdel forventer vi at linjen går gennem nul
Opgave: undersøg dette

Model: $Kons_i \sim N(\alpha + \beta \cdot SVL_i, \sigma^2)$, teste: $\alpha = 0$

Bruger Resultat 3.5. Beregnede værdier ("se vedhæftede R-kode"):

$$\hat{\alpha} = 14.3541, \quad s_r^2 = 562.1823$$

$$\bar{t} = 858.6452, \quad SSD_t = 2287244$$

T-teststørrelse: $t = \frac{\hat{\alpha} - \alpha_0}{\sqrt{s_r^2 \left(\frac{1}{n} + \frac{\bar{t}^2}{SSD_t} \right)}} = \frac{14.3541 - 0}{\sqrt{562.1823 \cdot \left(\frac{1}{62} + \frac{858.6452^2}{2287244} \right)}} = 1.0406$

P-værdi fra $t(60)$-fordeling: $2(1 - t_{\text{cdf}}(1.0406, 60)) = 0.3022$

R: `2*(1-pt(1.0406,62-2))`

Konklusion: da *p*værdi er langt over 0.05 strider data ikke mod skæring = 0

95%-konfidensinterval for skæring:

$t_0 = t_{\text{inv}}(0.975, 60) = 2.0003$ (R: `qt(0.975,60)`)

$\hat{\alpha} \pm t_0 \cdot \text{sd}_s(\hat{\alpha}) = 14.3541 \pm 2.0003 \cdot 13.7943 \approx [-13, 42]$

OBS: p-værdi større end 0.05 samme som at nul ligger i konfidensinterval

Teste $\alpha = 0$ i "cars"-data med model $Sb_i \sim N(\alpha + \beta h_i, \sigma^2)$

$n = 50$

$\hat{\alpha} = 1.277$

$\mathrm{sd}_s(\hat{\alpha}) = 0.4844$

Sceneskift

Model er analyseret

Næste: kørsel i R via lm

Benyt lm(respons∼forklarende variabel)

Model: $\text{Kons}_i \sim N(\alpha + \beta \cdot \text{SVL}_i, \sigma^2)$

lmUD=lm(Kons∼SVL)

Benyt summary på output:

summary(lmUD)

Finder $\hat{\beta}$, $\hat{\alpha}$, s_r^2 og laver automatisk t-test for $\beta = 0$ og for $\alpha = 0$

Konfidensintervaller: benyt confint på output fra lm

confint(lmUD)

dat=read.csv("SnakeFreshPres.csv",header=FALSE)

SVL=dat[,1]

Kons=dat[,2]

lmUD=lm(Kons∼SVL)

summary(lmUD)

Første del af output:

lm(formula = Kons ∼ SVL)

```
Residuals:
    Min      1Q  Median      3Q     Max
-60.362 -18.263   1.657  19.533  61.015
```

Parametertabel:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.35414   13.79426   1.041    0.302
SVL          0.92753    0.01568  59.162   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23.71 on 60 degrees of freedom
Multiple R-squared:  0.9831,Adjusted R-squared:  0.9829
F-statistic:  3500 on 1 and 60 DF,  p-value: < 2.2e-16
```

Std.Error: Standard Error, t value: $\frac{\text{estimate}-0}{\text{Std.Error}} \sim t(\text{degrees of freedom})$

Residual standard error: $= s_r$

Hver parameter har sin egen række: $\alpha =$ (Intercept), $\beta =$ SVL

Konfidenintervaller for α og β

confint(lmUD)

```
                  2.5 %     97.5 %
(Intercept) -13.238492 41.9467647
SVL           0.896171  0.9588911
```

Samme navne som i parametertabel

Plotte residualer mod forventede:
plot(lmUD$fitted.values, lmUD$residuals)

Adressere output: s_r: `summary(lmUD)$sigma`
$(\hat{\alpha}, \hat{\beta})$: `lmUD$coef`

Konfidensinterval for varians σ^2: webbog afsnit 2.6
med variansskøn s_r^2 og frihedsgradsantal $df = n - 2$

Prøv selv: "cars"-data

```
h=cars[,1]

sb=sqrt(cars[,2])

summary(lm(sb~h))
```

(a) Undersøg om data kan beskrives med den lineære regressionsmodel

plot af x mod t (plot(t,x))

indsæt linje (abline(lm(x∼t)))

residualplot (plot(t,lm(x∼t)$residuals))

qqplot (qqnorm(lm(x∼t)$residuals,datax=TRUE))

(b) Opstil en statistisk model for data

$X_i \sim N(\alpha + \beta t_i, \sigma^2)$, $i = 1, \ldots, n$

med relevante navne for X, t og n

(c) Angiv parameterskøn og konfidensintervaller

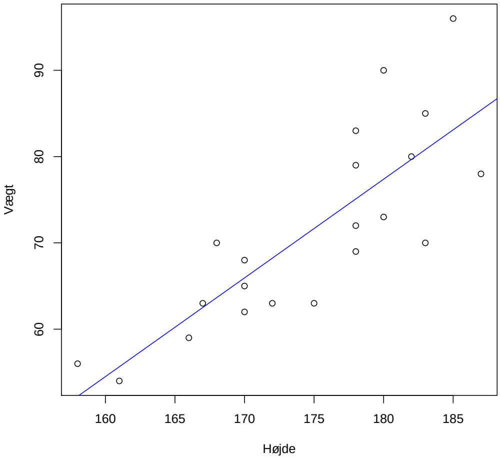
skøn: (summary(lm(x∼t))

konfidensintervaller: (confint(lm(x∼t))

Færdig med beskrivelse af lm og summary

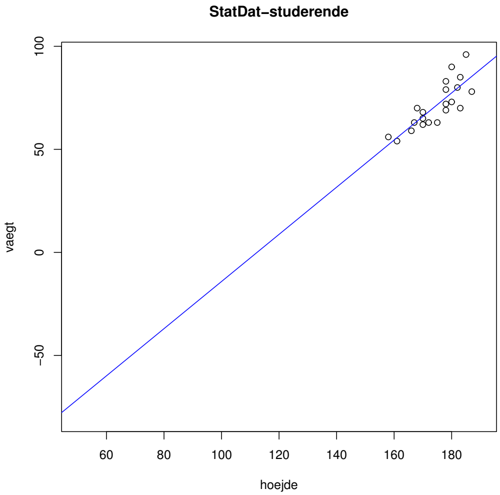
Næste: misbrug af sammenhæng

Fødselsvægt ud fra fødselshøjde



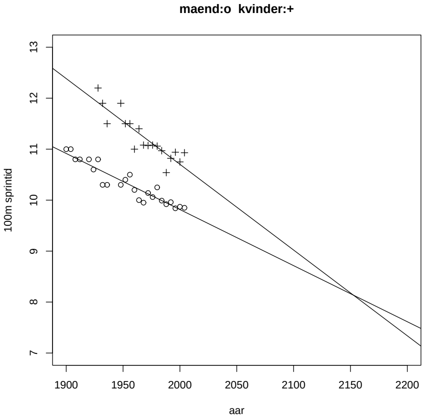
Fødselshøjde $\approx$ 50cm, gæt på vægt: $\hat{\alpha} + \hat{\beta} \cdot 50$

Misbrug af sammenhæng



Fødselshøjde $\approx$ 50cm, gæt på vægt: -70kg

Nature 2004:



I 2156 løber kvinderne hurtigere end mænd!

Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

The 2004 Olympic women's 100-metre sprint champion, Yuliya Nesterenko, is assured of fame and fortune. But we show here that — if current trends continue — it is the winner of the event in the 2156 Olympics whose name will be etched in sporting history forever, because this may be the first occasion on which the race is won in a faster time than the men's event.

The Athens Olympic Games could be viewed as another giant experiment in human athletic achievement. Are women narrowing the gap with men, or falling further behind? Some argue that the gains made by women in running events between the 1930s and the 1980s are decreasing as the women's achievements plateau[1]. Others contend that there is no evidence that athletes, male or female, are reaching the limits of their potential[1,2].

In a limited test, we plot the winning times of the men's and women's Olympic finals over the past 100 years (ref. 3; for data set, see supplementary information) against the competition date (Fig. 1). A range of curve-fitting procedures were tested (for methods, see supplementary information), but there was no evidence that the addition of extra parameters improved the model fit significantly from the simple linear relationships shown here. The remarkably strong linear trends that were first highlighted over ten years ago[2] persist for the Olympic 100-metre sprints. There is no indication that a plateau has been reached by either male or female athletes in the Olympic 100-metre sprint record.

Extrapolation of these trends to the 2008

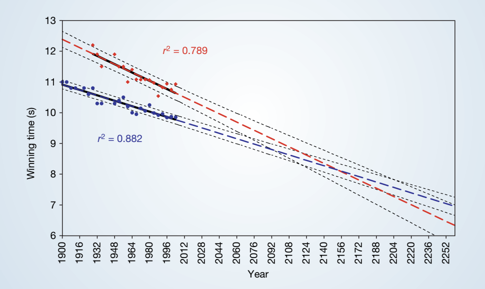


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

say that drug use explains why women's times were improving faster than men's, particularly as that improvement slowed after the introduction of drug testing[1]. However, no evidence for this is found here. By contrast, those who maintain that there could be a continuing decrease in gender gap point out that only a minority of the world's female population has been given the opportunity to compete (O. Anderson, www.pponline.co.uk/encyc/0151.htm).

Whether these trends will continue at the Beijing Olympics in 2008 remains to be seen.

Lung cancer

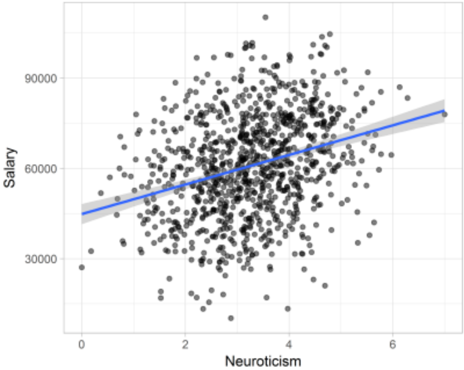
Intragenic ERBB2 kinase mutations in tumours

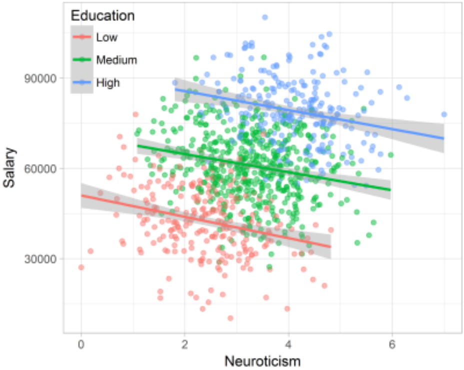
The protein-kinase family is the most frequently mutated gene family found in human cancer and faulty kinase enzymes are being investigated as promising targets for the design of antitumour therapies. We have sequenced the gene encoding the transmembrane protein tyrosine kinase ERBB2 (also known as HER2 or Neu) from

Ikke bruge sammenhæng udenfor data-område


THE TREND SEEN IN AGGREGATE
DATA MAY BE REVERSED WHEN

Simpsons paradoks





Fra generel linje:

$$X_i \sim N(\alpha + \beta t_i, \sigma^2)$$

til undermodel med proportionalitet mellem $E(X)$ og t:

$$X_i \sim N(\beta t_i, \sigma^2)$$

$$X_i \sim N(\beta t_i, \sigma^2), \quad \hat{\beta}_0 = \frac{\sum_{i=1}^n x_i t_i}{\sum_{i=1}^n t_i^2}, \quad s_{0r}^2 = \frac{1}{n-1} \sum_i (x_i - \hat{\beta} t_i)^2$$

R: lm(x∼-1+t)

```
summary(lm(Kons~-1+SVL))

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
SVL 0.944190   0.002935   321.6   <2e-16 ***

Residual standard error: 20.22 on 61 degrees of freedom
```

Konfidensinterval for β i model $X_i \sim N(\beta t_i, \sigma^2)$:

confint(lm(Kons $\sim$ -1+SVL))

```
        2.5 %    97.5 %
SVL 0.9383199 0.9500595
```

konfidensinterval: [0.938, 0.950], længde: 0.012

Konfidensinterval for β i model $X_i \sim N(\alpha + \beta t_i, \sigma^2)$:

[0.910, 0.969], længde: 0.59

Reduktion af model giver mere information om de resterende parametre (den store forskel her skyldes at SVL=0 ligger langt fra dataområdet)

Opsummering:

Model: $X_i \sim N(\alpha + \beta t_i, \sigma^2)$

lm(x∼t)

summary(lm(x∼t))

confint(lm(x∼t))

Slut for i dag efter en dag med mange slides