STK-IN1050 - Statistikk for informatikere

17. Lineær regresjon

Matematisk institutt

2026-04-20

Prediksjon

Vi skal nå bevege oss fra statistisk inferens til **prediksjon**: bruke data til å si noe om framtidige eller ukjente verdier

Anta vi har n forskjellige kilder av informasjon $x_1, \ldots, x_n$ og vi ønsker å bruke denne informasjonen til å predikere verdien til en ny variabel y:

$$y = g(x_1, \ldots, x_n),$$

hvor g er en funksjon.

- $y \in \mathbb{R} \Rightarrow$ regresjon
- $y \in \{0, 1\} \Rightarrow$ klassifisering

Prediksjon

Vi kaller y en **responsvariabel**.

Vi kaller $x_1, \ldots, x_n$ **forklaringsvariabler**.

I tillegg har vi **confoundere**: andre variabler (som vi ikke har observert) som påvirker **både** forklaringsvariablene og responsvariabelen, slik at de skaper eller skjuler en sammenheng mellom dem.

Eksempel 1: Skjermtid og eksamensresultater

En foreleser samler inn data på hvor mange timer per dag studentene bruker foran skjerm (inkludert mobil, PC, nettbrett) i ukene før eksamen, og hvilken karakter de får på eksamen.

- Hva er en naturlig responsvariabel her?
- Hva kan være en forklaringsvariabel?
- Ville du brukt regresjon eller klassifisering her?
- Hva er minst to mulige confoundere som kan påvirke både skjermtid og eksamensresultater?

Eksempel 2: Kaffe og produktivitet

Et firma måler hvor mange kopper kaffe hver utvikler drikker per dag, og hvor mange linjer kode de i snitt skriver per dag.

- Hva er responsvariabelen?
- Hva er forklaringsvariabelen?
- Ville du brukt regresjon eller klassifisering her?
- Nevn mulige confoundere (både menneskelige og organisatoriske).
- Gir det mening å tolke en eventuell sammenheng som «kaffe øker/reduserer produktiviteten»?

Eksempel 3: Oppgaver og eksamenskarakter

I et innføringskurs i programmering registreres hvor mange oppgaver hver student leverer inn i løpet av semesteret, og eksamenskarakteren de får.

- Hva er responsvariabelen?
- Hva er forklaringsvariabelen?
- Hvilke andre variabler kunne det være lurt å måle i tillegg?

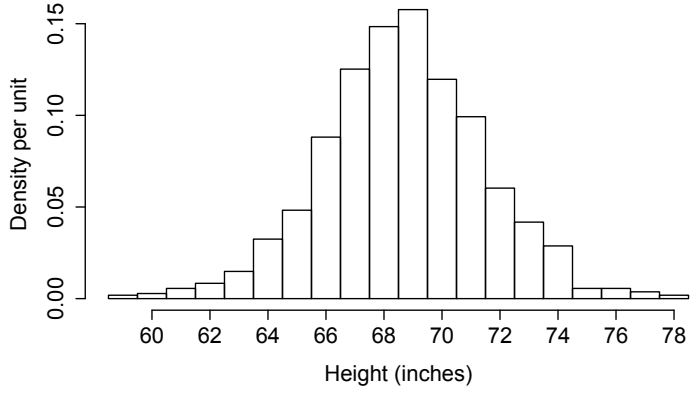
Eksempel 4: Tastatur og skrivehastighet

En informatikkstudent bytter fra et vanlig tastatur til et mekanisk tastatur. Man måler skrivehastighet (ord per minutt) før og etter byttet.

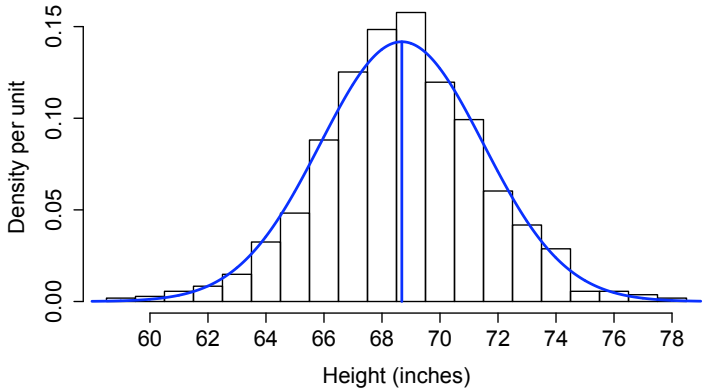
- Hvis vi skal modellere dette med regresjon, hva kan være forklaringsvariabel og responsvariabel?
- Hvilke andre variabler kan confoundere?
- Hvordan kunne man designet et eksperiment som reduserer effekten av confoundere?

Prediksjoner

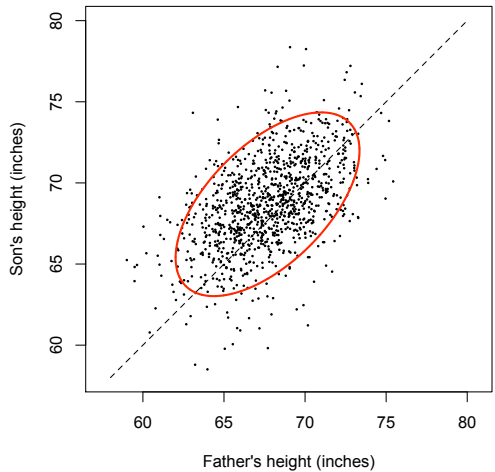
Histogrammet viser høyden til 1078 menn. Vi trekker en tilfeldig mann fra denne gruppen og skal predikere hvor høy han er.



Dataene er nærmest normalfordelt, så vår prediksjon er lik gjennomsnittet til dataene:

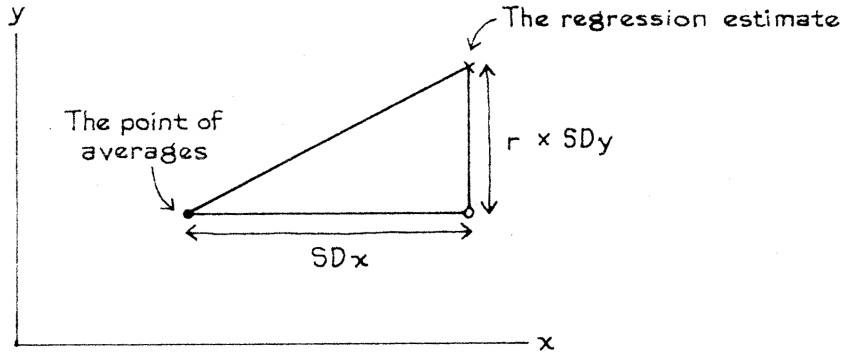


Vi har nå to korrelerte variabler: hvis vi kjenner verdien til den ene, da har vi noen informasjon om verdien til den andre og kan lage en bedre prediksjon.

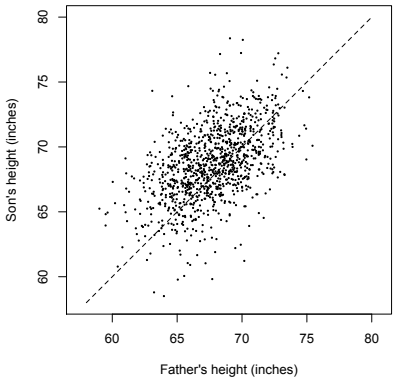


Regresjonslinjen

Regresjonslinjen for y på x estimerer gjennomsnittsverdien av y som tilsvarer hver verdi av x. Økning på 1 SD i x gir økning på bare r SD i y, hvor $r = \text{Cor}(x, y)$.



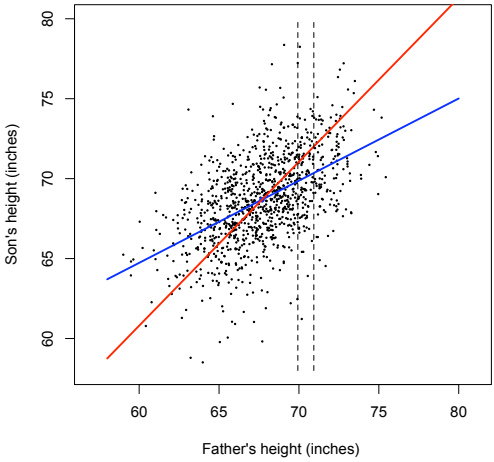
Eksempel: Høyden til far og sønn



Figuren viser høyden til 1078 far-sønn par:

- gjennomsnittlig høyde for far ≈ 68in ($\bar{x}$)
- SD for fars høyde ≈ 2.7in (s_x)
- gjennomsnittlig høyde for sønn ≈ 69in ($\bar{y}$)
- SD for sønns høyde ≈ 2.8in (s_y)
- r = 0.50

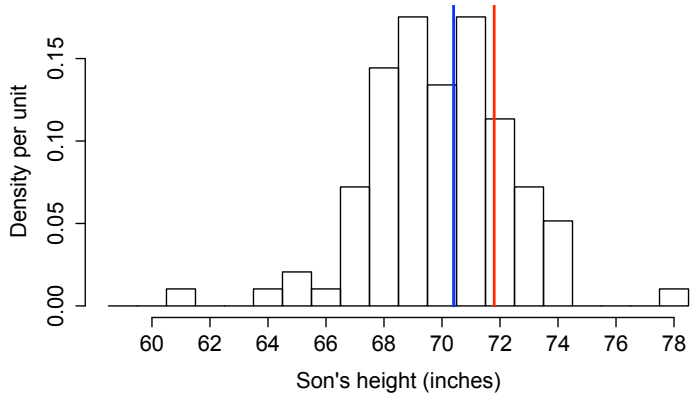
Eksempel: Høyden til far og sønn



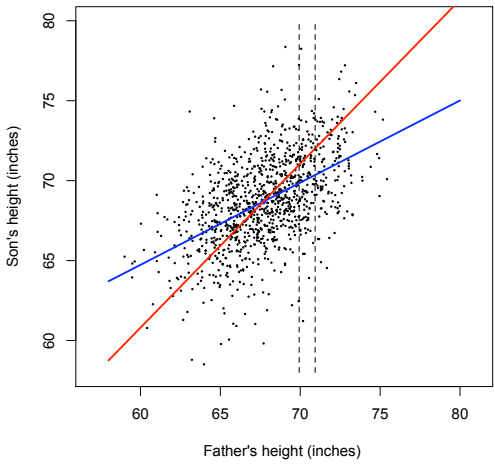
- De vertikale stripene viser fedre som er rundt 1 SD over gjennomsnittet $\bar{x}$ i høyde
- Den røde linje er SD-linjen, hvor økning på 1 SD i x gir økning på 1 SD i y
- Den blå linjen er regresjonslinjen

Eksempel: Høyden til far og sønn

Histogrammet viser høyden til sønner av fedre som er ca. 1 SD over gjennomsnittet $\bar{x}$ (SD-linje i rødt, regresjonslinje i blått).



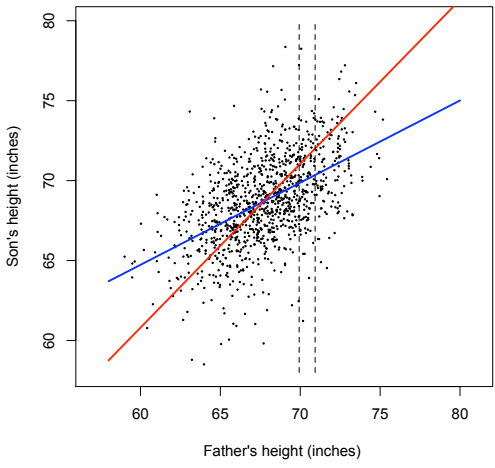
Eksempel: Høyden til far og sønn



Her kommer korrelasjonskoeffisienten $r = 0.5$ inn.

Når fedres høyde øker med 1 s_x, øker sønners høyde i snitt bare med 0.5 s_y.

Eksempel: Høyden til far og sønn



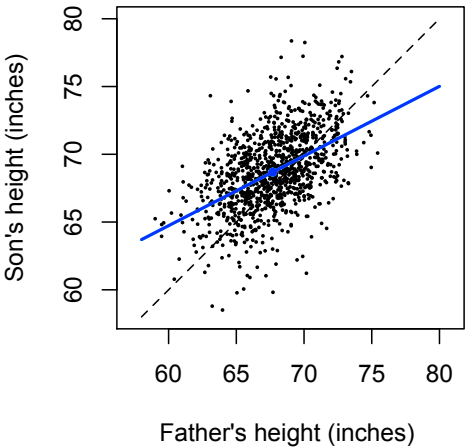
Se på fedre som er 1 s_x over gjennomsnittet,

$$\bar{x} + \text{SD}_x = 68\text{in} + 2.7\text{in} = 70.7\text{in}.$$

Den gjennomsnittlige høyden til sønnene deres vil være omtrent

$$\bar{y} + r \times s_y = 69\text{in} + 0.5 \times 2.8\text{in} = 70.4\text{in}.$$

Regresjonsmetoden for individer



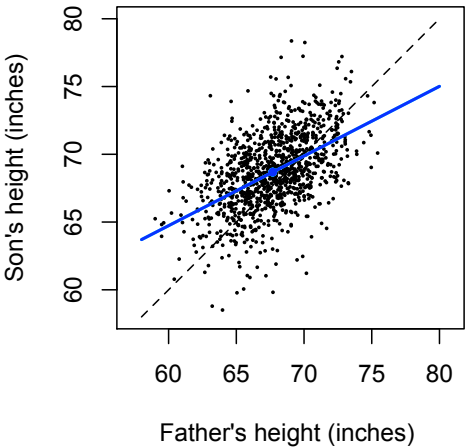
Alle punkter

(fars høyde, estimert høyde for sønn)

vil ligge på regresjonslinjen.

Regresjonslinjen går gjennom gjennomsnittspunktet: fedre med gjennomsnittlig høyde forventes også å ha sønner med gjennomsnittlig høyde.

Regresjonsmetoden for individer



Anta at vi trekker én far tilfeldig og skal gjette sønnens høyde.

Det beste gjettet er da det overordnede gjennomsnittet for sønners høyde, 69in.

Så får vi vite at faren er 55in høy. Da finnes et bedre gjett: gjennomsnittet til sønner av fedre som er 55in høye. Det estimeres med regresjonsmetoden til 62in.

Regresjon mot gjennomsnittet

Barn i barnehagealder deltar i et prosjekt for å øke deres IQ.

- De testes ved prosjekt-oppstart (pre-test)
- De testes igjen etter gjennomføring av prosjektet (post-test)

Resultater:

- Pre-test: gjennomsnittlig score = 100, SD = 15
- Post-test: gjennomsnittlig score = 100, SD = 15

Ser ut som prosjektet hadde liten effekt.

Men:

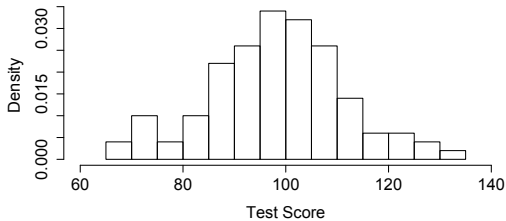
- IQ-en til barn under snittet på pre-test økte i snitt 5 poeng
- IQ-en til barn over snittet på pre-test falt i snitt ca. 5 poeng

Regresjon mot gjennomsnittet

Hva skjer her? Intent annet enn **tilfeldige feil**. Grunnlikningen er

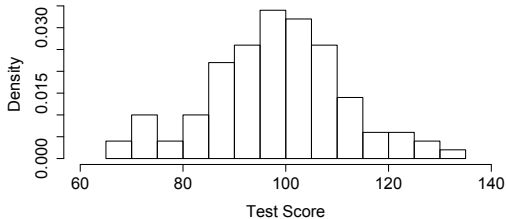
observert testskår = sann skår + tilfeldig feil

Anta at tilfeldig feil like ofte er negativ som positiv.



Anta også at skårene er normalfordelt, med snitt på 100 og SD på 15.

Regresjon mot gjennomsnittet



Se på et barn som får 130 på første test. To muligheter:

- sann skår under 130, med positiv tilfeldig feil
- sann skår over 130, med negativ tilfeldig feil

Første forklaring er mest sannsynlig, dersom det finnes flere barn med sann IQ litt under 130 enn litt over 130.

Regresjon mot gjennomsnittet

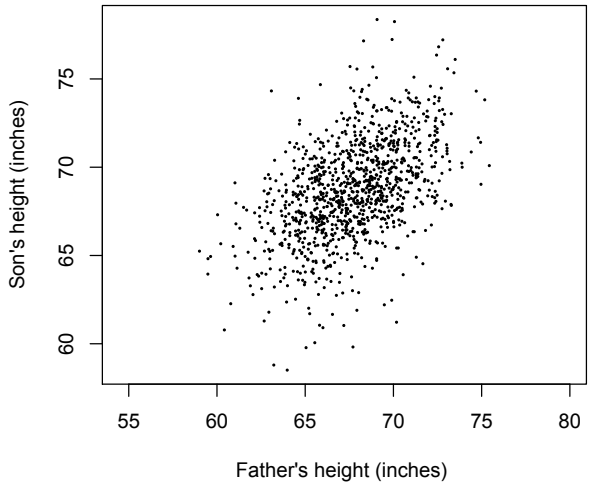
Skårer noen *over (under)* snittet første gang, er sann skår trolig litt *lavere (høyere)* enn observert. Ved ny test blir andre skår ofte litt *lavere (høyere)* enn første.

Ved test–retest vil den laveste gruppen på første test i snitt forbedre seg, og den høyeste gruppen falle tilbake. Dette er **regresjonseffekten**.

Å tolke regresjonseffekten som noe «virkelig» og viktig, og ikke bare som tilfeldige feil, er en feilslutning og kalles «**regression fallacy**».

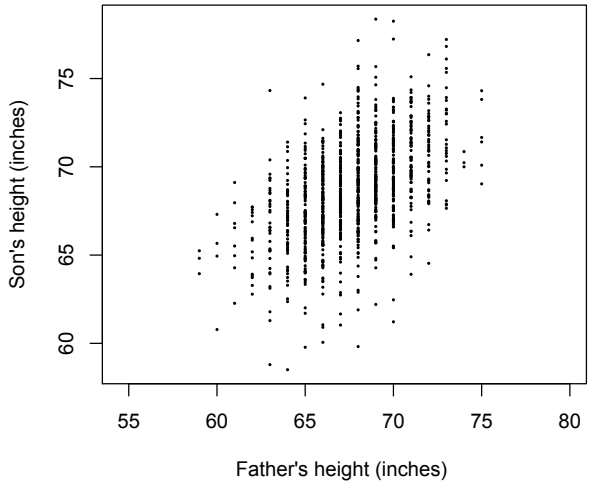
Intuisjson om estimering: Graf over gjennomsnitt

Vi starter med alle datene:



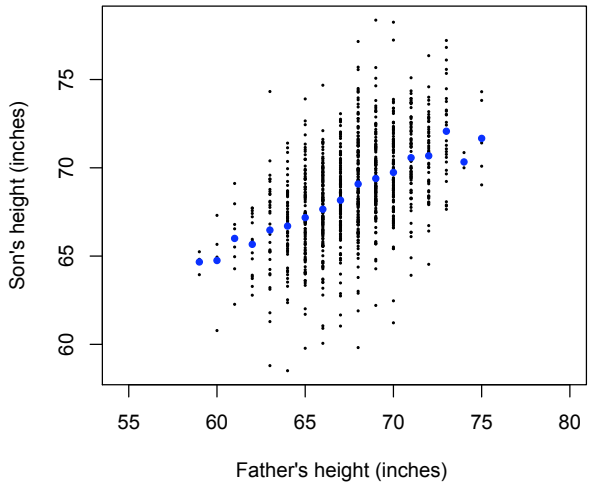
Intuisjson om estimering: Graf over gjennomsnitt

Vi runder høyden til faren til nærmeste tomme:



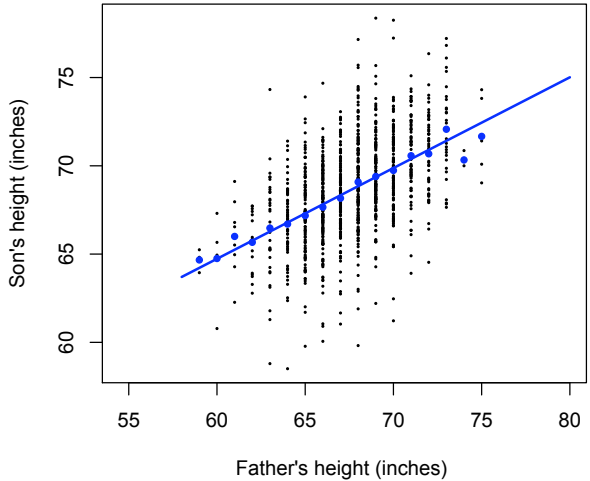
Intuisjson om estimering: Graf over gjennomsnitt

Vi finner gjennomsnittlig høyde til sønnen for hvert verdi av høyden til faren.



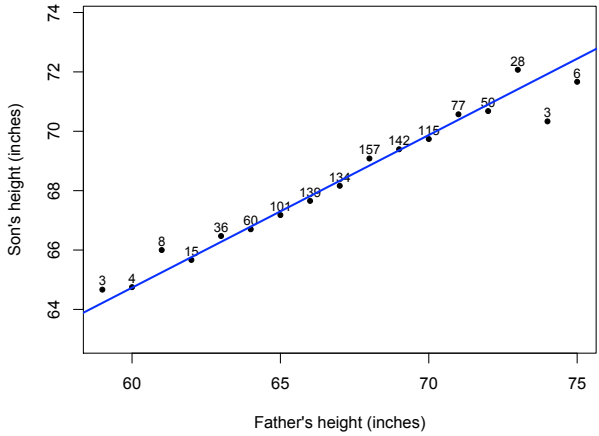
Intuisjson om estimering: Graf over gjennomsnitt

Vi finner den linjen som passer best til de blå punktene.



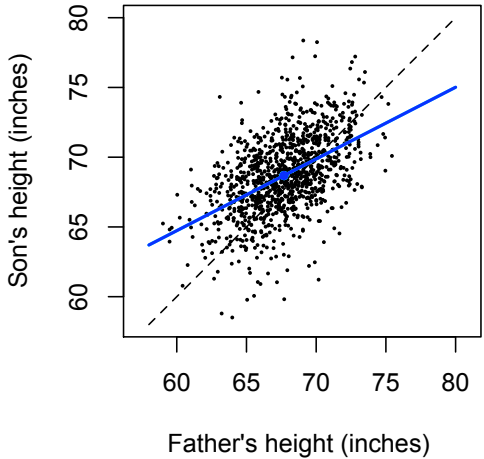
Intuisjson om estimering: Graf over gjennomsnitt

Her tar vi høyde for hvor mange datapunkter som ligger bak hver blå punkt.



Intuisjon om estimering: Graf over gjennomsnitt

Til slutt går vi tilbake til originaldataene, men beholder linjen.



En lineær regresjonsmodell

En **lineær regresjonsmodell** er gitt ved

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

for $i = 1, \ldots, n$ hvor y_i er **responsvariabelen**, x_i er **forklaringsvariabelen** og ε_i er **støyleddet** for observasjon i.

Vi kaller **regresjonskoeffisientene** β_0 **skjæringspunktet** og β_1 **stigningstallet**.

En lineær regresjonsmodell

En **lineær regresjonsmodell** er gitt ved

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

for $i = 1, \ldots, n$ hvor y_i er **responsvariabelen**, x_i er **forklaringsvariabelen** og ε_i er **feilleddet** for observasjon i.

Vi gjør fire antakelser:

1. **Linearitet**: sammenhengen mellom responsen or forklaringsvariablene er (tilnærmet) lineær
2. **Uavhengige feil**: Feilleddene er uavhengig av hverandre
3. **Konstant varians**: Feilleddene har samme varians for alle verdier av x
4. **Normalfordeling**: Feilleddene er (tilnærmet) normalfordelte

Minste kvadraters kriterium

For å konstruere en estimator for β_0 og β_1 bruker vi **minste kvadraters kriteriet**

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1) &= \text{argmin}_{b_0, b_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \text{argmin}_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2
\end{aligned}
$$

hvor estimatene $(\hat{\beta}_0, \hat{\beta}_1)$ definerer **prediksjoner** $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ gitt x_i.

Eksakt formel

For å løse optimeringsproblemet kan vi bruke en generell optimeringsrutine for å finne $\hat{\beta}_0$ og $\hat{\beta}_1$. I dette tilfellet kan vi også utlede en eksakt formel for estimatene. La

$$f(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

hvor $\boldsymbol{b} = (b_0, b_1)$. Da gjelder at

$$\frac{\partial}{\partial b_0} f(\boldsymbol{b}) = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = -2(n\bar{y} - nb_0 - nb_1\bar{x});$$

$$\frac{\partial}{\partial b_1} f(\boldsymbol{b}) = -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) x_i = -2(\sum_{i=1}^{n} x_i y_i - b_0 n\bar{x} - b_1 \sum_{i=1}^{n} x_i^2).$$

Eksakt formel

Vi finner de optimale verdiene ved å sette de deriverte lik null, noe som gir ligningssystemet

$$nb_0 + n\bar{x}b_1 = n\bar{y}$$

$$b_0 n\bar{x} + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

som har løsningen

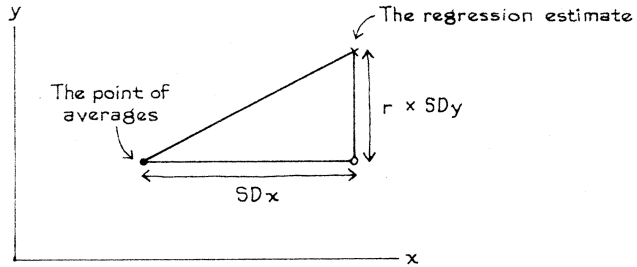
$$\hat{\beta}_1 = b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = r\frac{s_y}{s_x};$$

$$\hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Regresjonslinjen

Det følger at regresjonslinjen er en linje som skjærer punktet $(\overline{x}, \overline{y})$ og har stigningstall

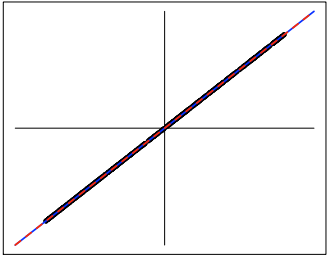
$$r\frac{s_y}{s_x}.$$

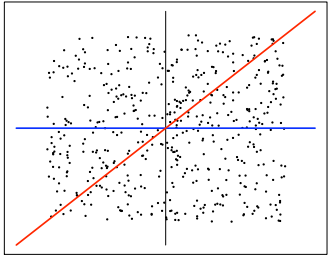


Regresjonslinjen

Hvis $r = 1$, da predikerer vi $\hat{y} = x$.

Hvis $r = 0$, da predikerer vi $\hat{y} = \overline{y}$.





Hvis r er mellom 0 og 1, vil vi predikere en verdi imellom, og regresjonsmetoden forteller oss nøyaktig hvilken.

Eksempel: Moores lov

Moores lov sier at antall transistorer som kan integreres på én integrert krets fordobles hver 24. måned.

Vi ser på noen data:

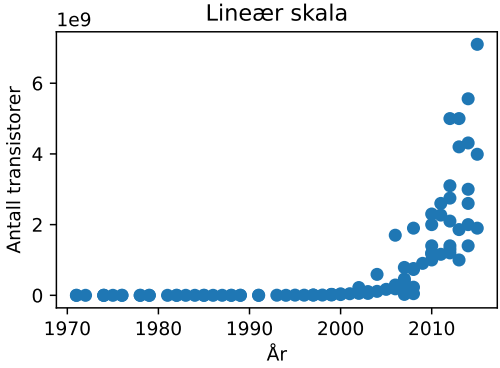
```
import pandas as pd

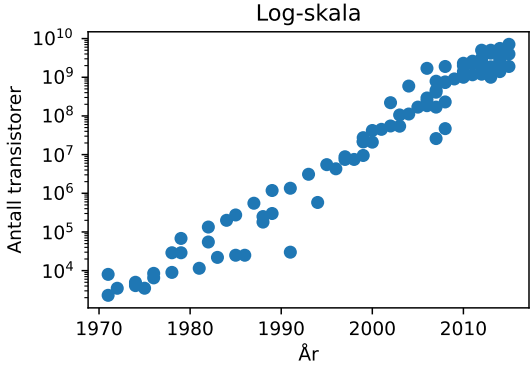
z = pd.read_excel("../data/Mooreslaw.xlsx")
print(z.iloc[:5, :3])
```

```
             Processor  Transistorcount  Date
0              TMS 1000             8000  1971
1           Intel 4004              2300  1971
2           Intel 8008              3500  1972
3  MOS Technology 6502              3510  1975
4        Motorola 6800              4100  1974
```

Eksempel: Moores lov

Vi plotter antall transistorer mot dato:





Eksempel: Moores lov

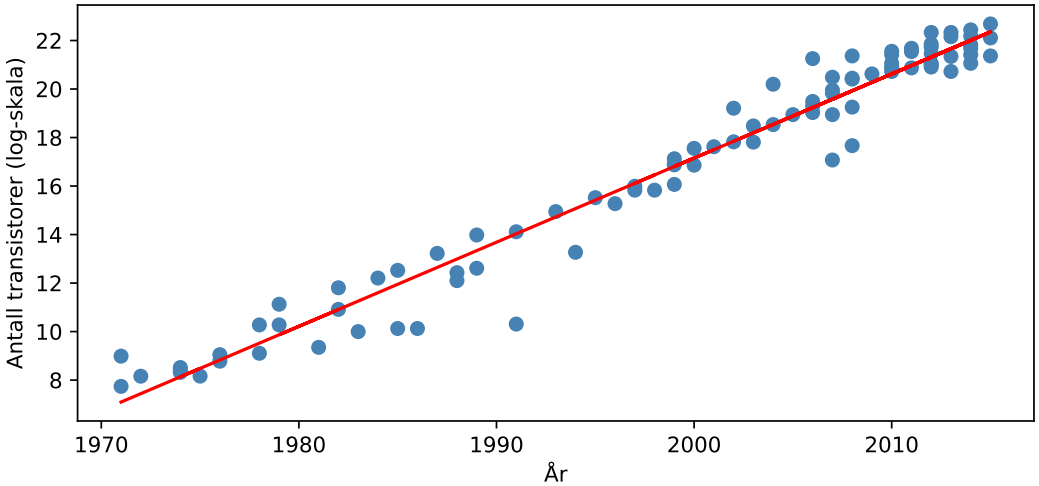
```
import numpy as np
import pandas as pd

def lin1reg(x,y):
    n = len(y)
    b1 = np.sum((x-np.mean(x))*(y-np.mean(y)))/np.sum((x-np.mean(x))**2)
    b0 = np.mean(y)-b1*np.mean(x)
    bhat = [b0,b1]
    return(bhat)

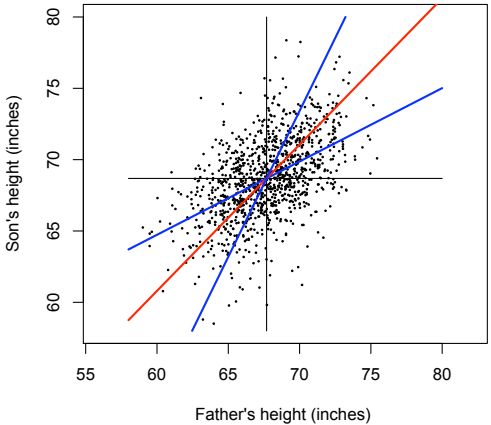
Tdata = pd.read_excel("../data/Mooreslaw.xlsx")
Tdata['logTranscount'] = np.log(Tdata['Transistorcount'])
y = Tdata['logTranscount']
x = Tdata['Date']
bhat = lin1reg(x,y)
print(f"{bhat[0]:.3f}, {bhat[1]:.3f}")
```

```
-676.569, 0.347
```

Eksempel: Moores lov



To regresjonslinjer

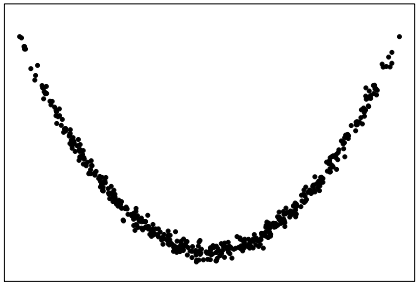


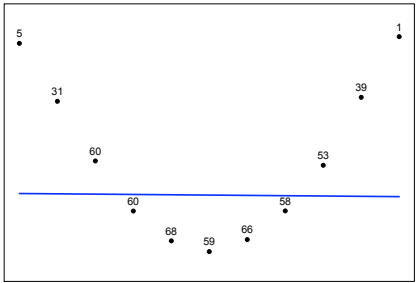
For hvert spredningsplott kan vi tegne to regresjonslinjer: én for å predikere y ut fra x, og en annen for å predikere x ut fra y.

Vi setter alltid responsen på y-aksen og forklaringsvariabelen på x-aksen.

Lineær regresjonsmodell er ikke alltid svaret

Hvis det er en ikke-lineær sammenheng mellom de to variablene, vil regresjonslinjen glatte ut signalet i dataene.





Ekstrapolering

Regresjonsmetoden kan brukes til å lage prediksjoner for enkelte datapunkter.

Men trekk varsomt hvis vi må ekstrapolere langt fra dataene, eller til en annen gruppe individer.

OL 2156



nature
International weekly journal of science
Login
Search This journal go Advanced search

Journal home > Archive > Brief Communications > Abstract

Journal home
Advance online publication
Current issue
Nature News
Archive

Supplements
Web focuses
Multimedia

Brief Communications

Nature **431**, 525 (30 September 2004) | doi:10.1038/431525a; Published online 29 September 2004

Athletics: Momentous sprint at the 2156 Olympics?

Andrew J. Tatem[1], Carlos A. Guerra[1], Peter M. Atkinson[2] & Simon I. Hay[1,3]

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.

Top

subscribe to *Nature*
Subscribe

ABSTRACT
Next
Table of contents
Full text
Download PDF
Send to a friend

OL 2156

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.

An Oxford University study found that women are running faster than they have ever done over 100m.

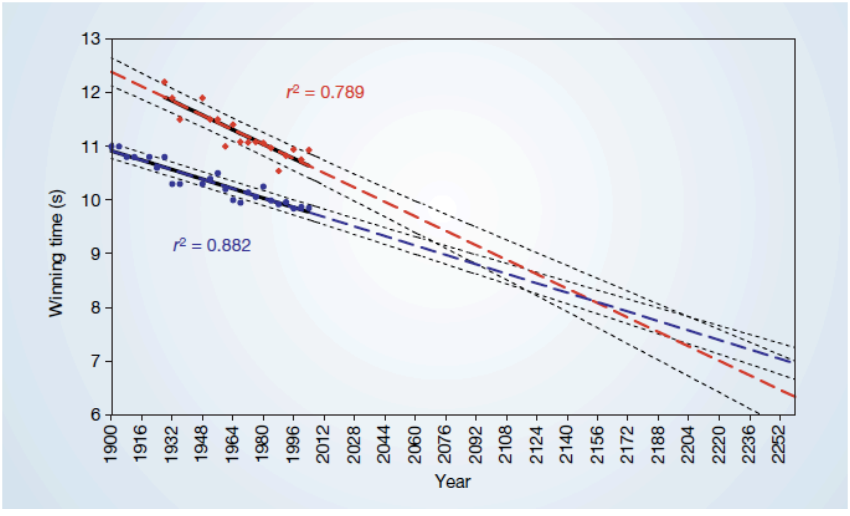


Women are set to become the dominant sprinters

Women of the future are on track to run faster than men

By Steve Connor Science Editor
Thursday, 30 September 2004

OL 2156


r^2 = 0.789
r^2 = 0.882
Winning time (s)
Year
13
12
11
10
9
8
7
6
1900
1916
1932
1948
1964
1980
1996
2012
2028
2044
2060
2076
2092
2108
2124
2140
2156
2172
2188
2204
2220
2236
2252

Prediksjon

Å predikere *y*-verdier basert på kunnskap om *x*:

- Uten informasjon er gjennomsnittet beste gjett.
- Hvis du kjenner *x*, er beste gjett for *y* gjennomsnittet av *y*-verdiene for dette *x*. Dette estimeres med regresjonsmetoden.

Åpne spørsmål

- Hva er egenskapene til $(\hat{\beta}_0, \hat{\beta}_1)$?
- Hvordan kan vi estimere variansen σ^2 til feilleddene ε_i?
- Hvor stor er usikkerheten knyttet til estimatorene?
- Er det avhengighet mellom estimatorene?
- Hvordan kan vi lage konfidensintervall og gjøre hypotesetester for β_0, β_1?
- Hvordan kan vi bruke ferdig implementerte funksjoner til slik analyse?