Statistisk översiktskurs - Föreläsning 7

Anders Fredriksson

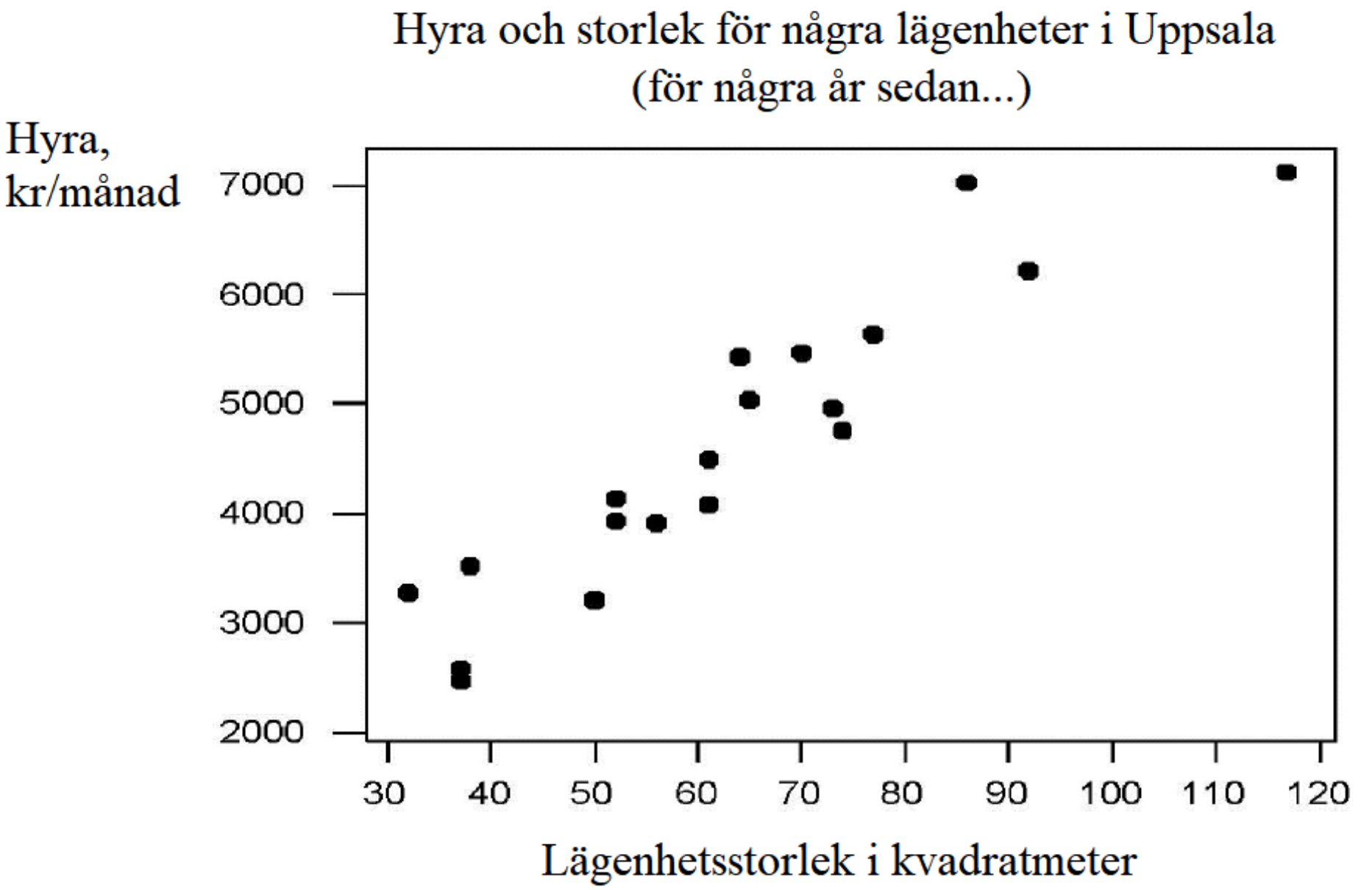
Statistiska Institutionen
Stockholms Universitet

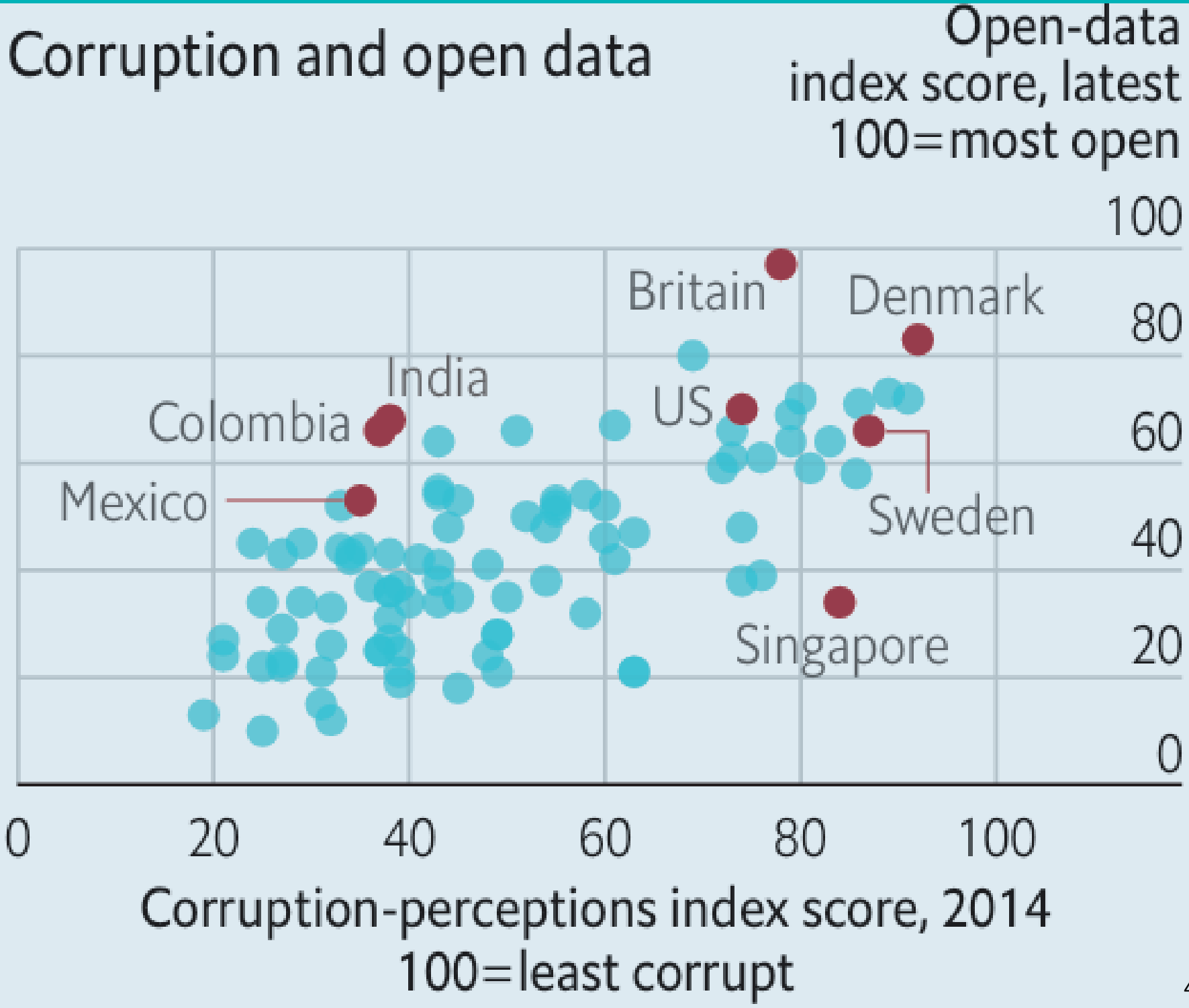
4 april 2025

Föreläsning 7 - innehåll (liknande innehåll på F8)

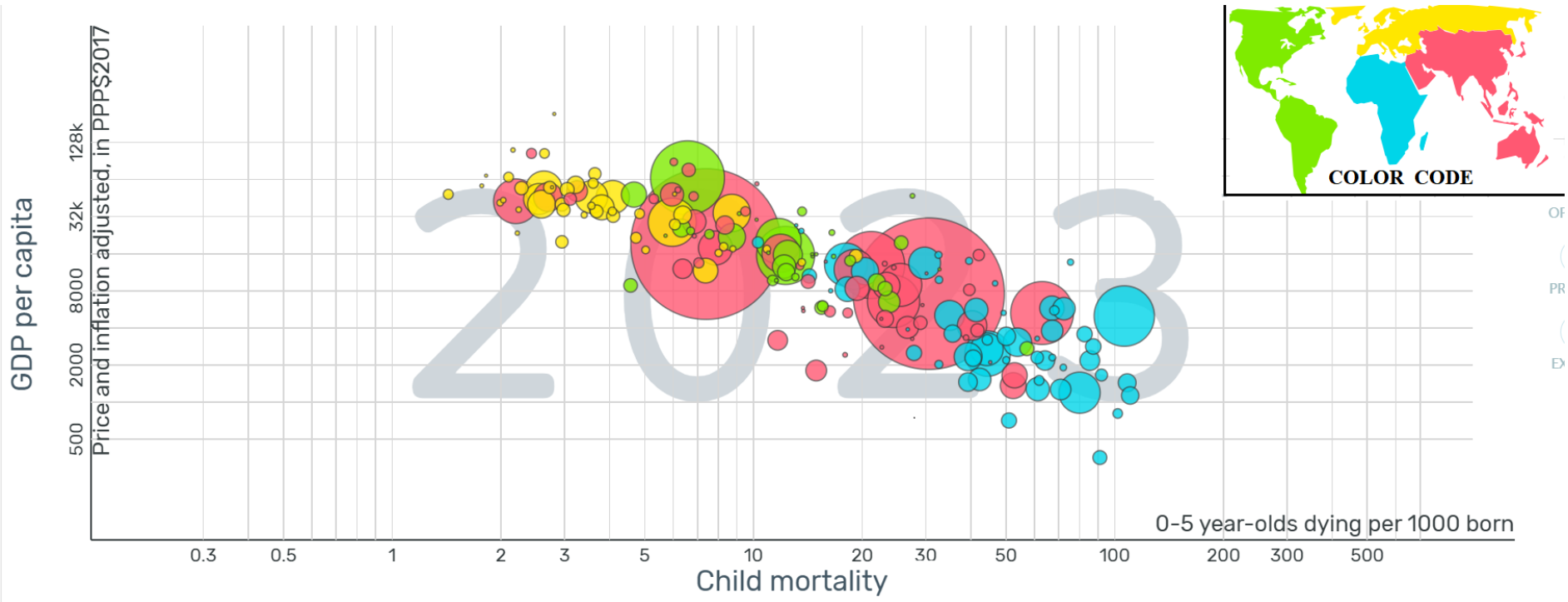
- Samband mellan variabler
  - beskriva samband
- Varians och kovarians
- Korrelation
- Linjär regression, introduktion
  - anpassa en rät linje till data (minsta kvadratmetoden)
  - ”förklara y mha x”
  - vi kan ha en eller flera x-variabler
- Vi har åtminstone två syften (relaterar till ”definitionen” av statistik)
  - beskriva data och samband
  - inferens
- Skilj på korrelation och kausalitet

Några spridningsdiagram - träna på att beskriva data




Corruption and open data
Open-data index score, latest
100=most open
100
80
60
40
20
0
Britain
Denmark
India
Colombia
US
Mexico
Sweden
Singapore
0
20
40
60
80
100
Corruption-perceptions index score, 2014
100=least corrupt

BNP per capita och barnadödlighet (storlek på cirkel - (relativ) folkmängd)



Källa: Gapminder

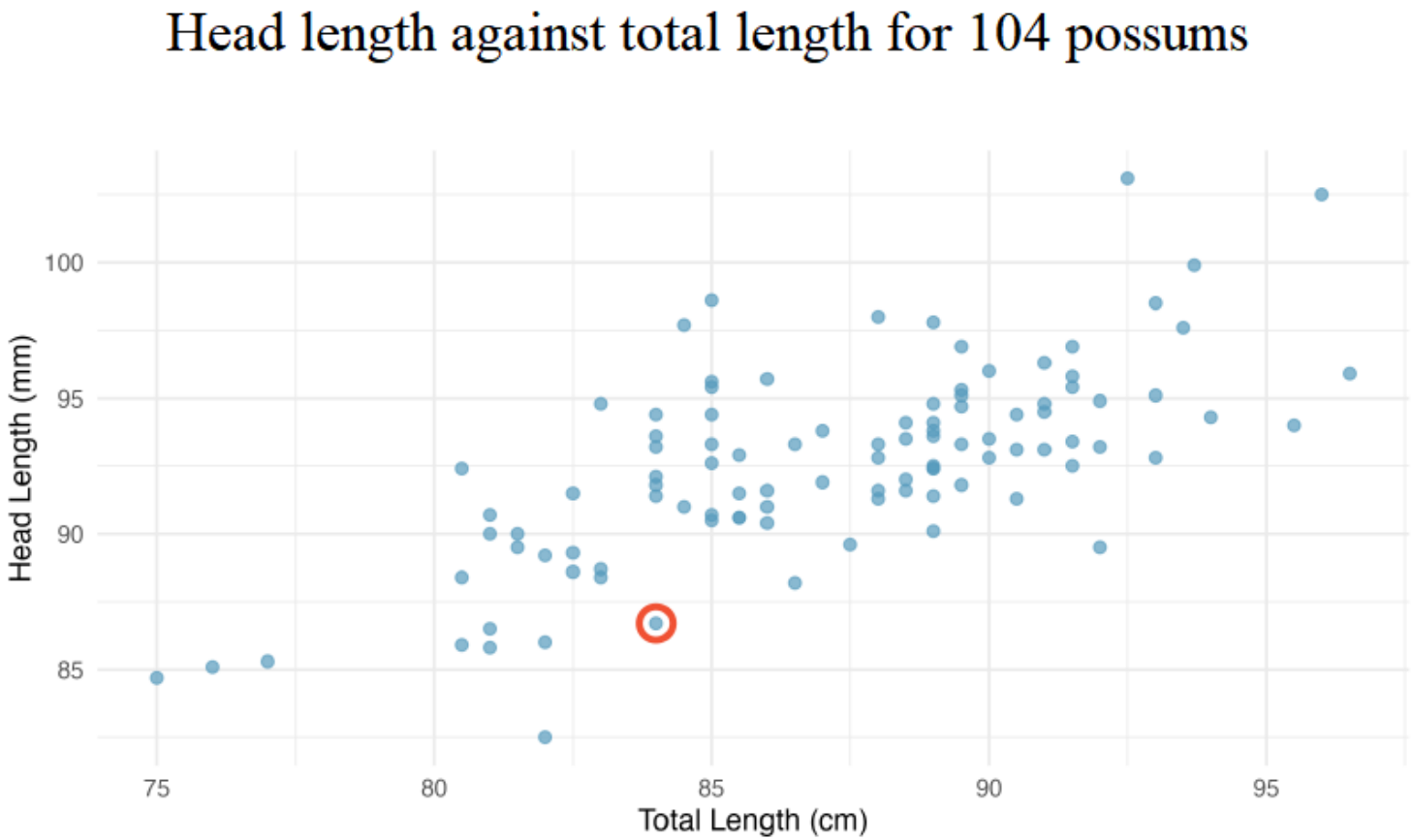


Figure 7.5: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 86.7 mm and total length 84 cm is highlighted.

Samband mellan två variabler

- Ofta är vi intresserade av om två variabler ”har med varandra att göra”, om de är **relaterade, associerade**
- Ex: Lön och utbildning, vikt och längd, tillväxt och skattesats
- Om vi är intresserade av hur en variabel (ex: lön) ”beror av” en annan variabel (ex: utbildning) kan vi kalla dessa
  - **y-variabel** och **x-variabel**
  - Mer formellt: **responsvariabel** och **förklaringsvariabel**
  - Förekommer också: **utfallssvariabel** och **prediktorvariabel**
  - Annan (inte så bra) terminologi: beroende variabel och oberoende variabel
  - På engelska (i olika varianter): y - x, explained - explanatory, response - predictor, dependent - independent
- I praktiken används terminologin även när vi inte har kausala samband (mer om detta följer)

Vi behöver begrepp och mått för att beskriva samvariation

- Titta först på formeln för varians (F2, en variabel):
- **Variansen** av n värden $x_1, x_2, \ldots, x_n$ skrivs som s_x^2 och ges av:

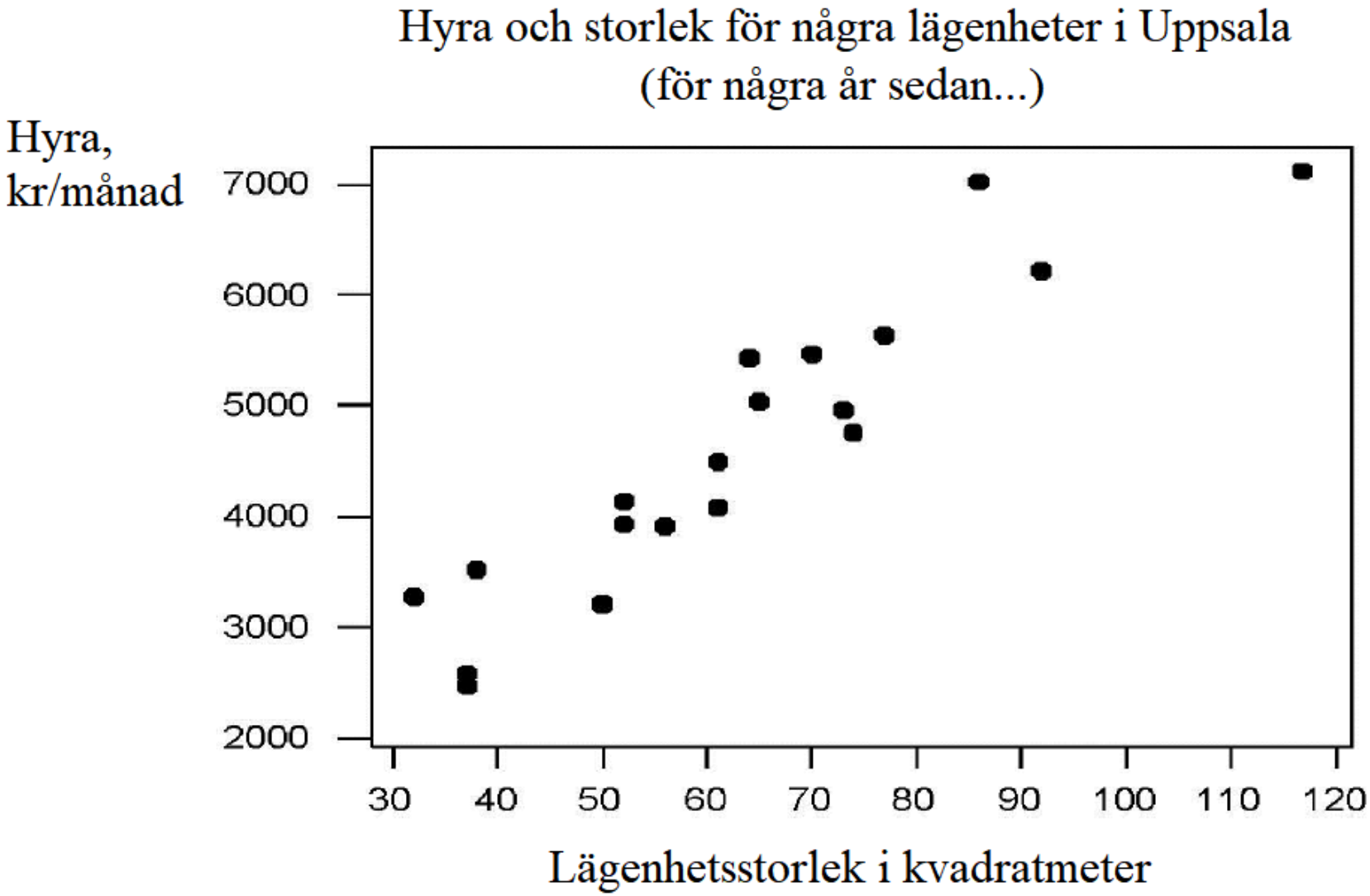
$$\text{Var(x)} = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

- ”Byt ut” en av faktorerna (ena parentesen) till y-variabeln

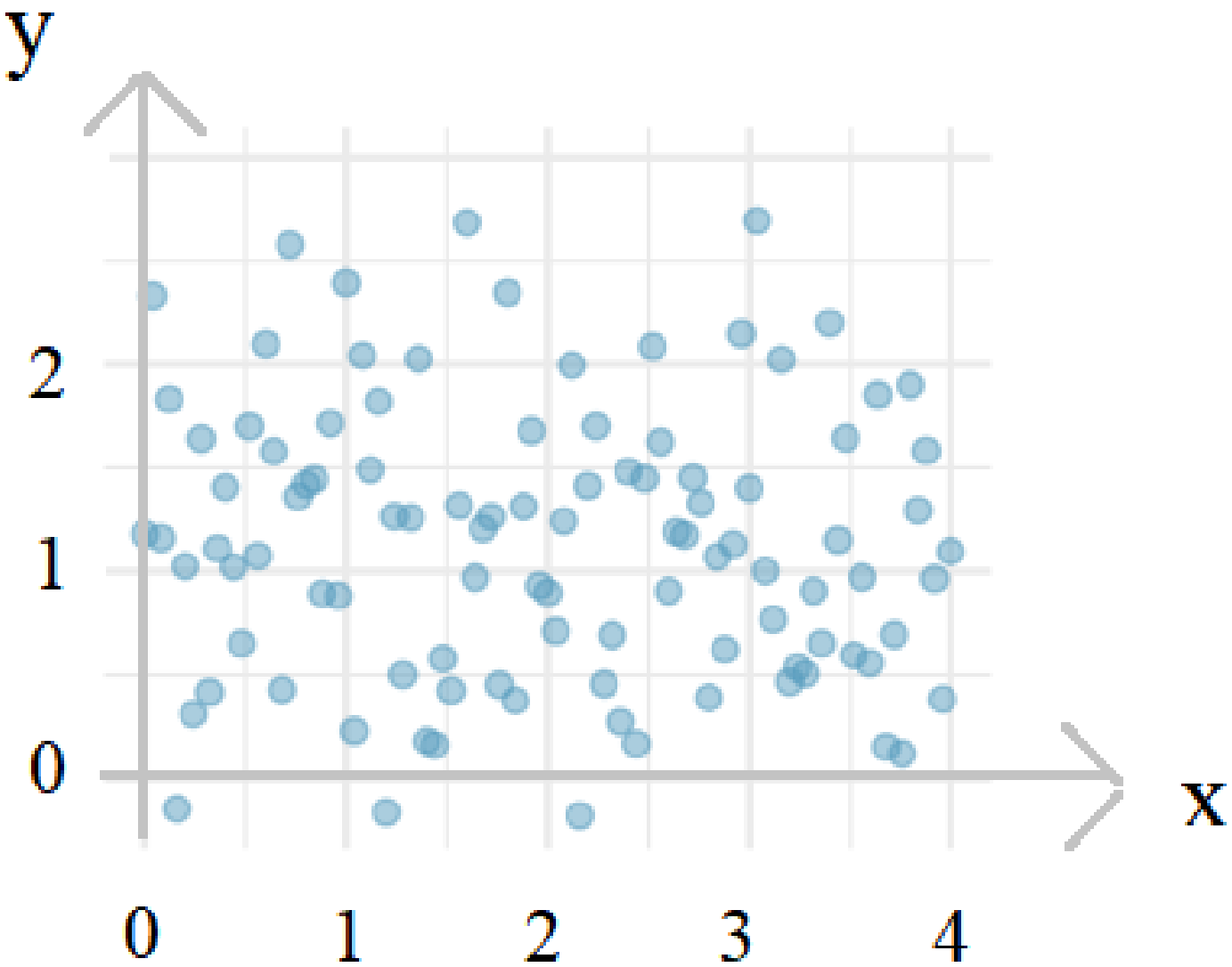
$$\text{Covar(x,y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- **Kovariansen** (covariance) är ett mått på hur mycket x och y samvarierar

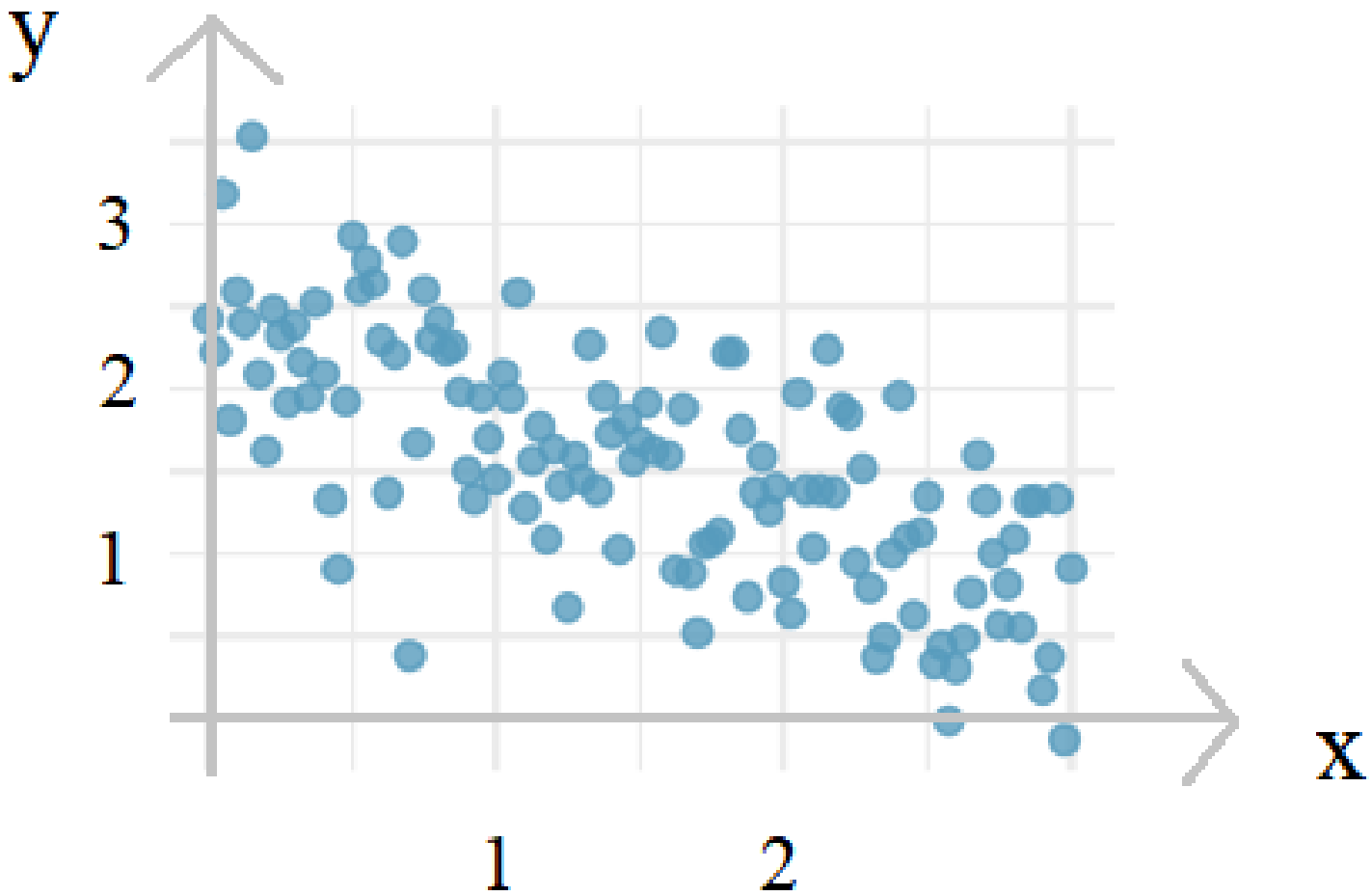
Om y är stor (i förhållande till sitt medelvärde) när x är stor blir kovariansen stor



Om y varierar oberoende av x blir blir kovariansen liten



Om y är liten när x är stor, och vice versa, blir kovariansen stor och negativ



Korrelation

- Ett potentiellt problem om vi vill beskriva samband är att kovariansen beror på urvalsstorleken och på vilken enhet x och y mäts i
- Om vi justerar kovariansformeln genom att dividera med standardavvikelserna, enligt nedan, får vi ett mått som alltid ligger mellan -1 och 1

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

- Vi har tagit fram **korrelationskoefficienten** (kallad r)
- Används för att beskriva **linjära samband**

Exempel på data och storlek på korrelationskoefficienter

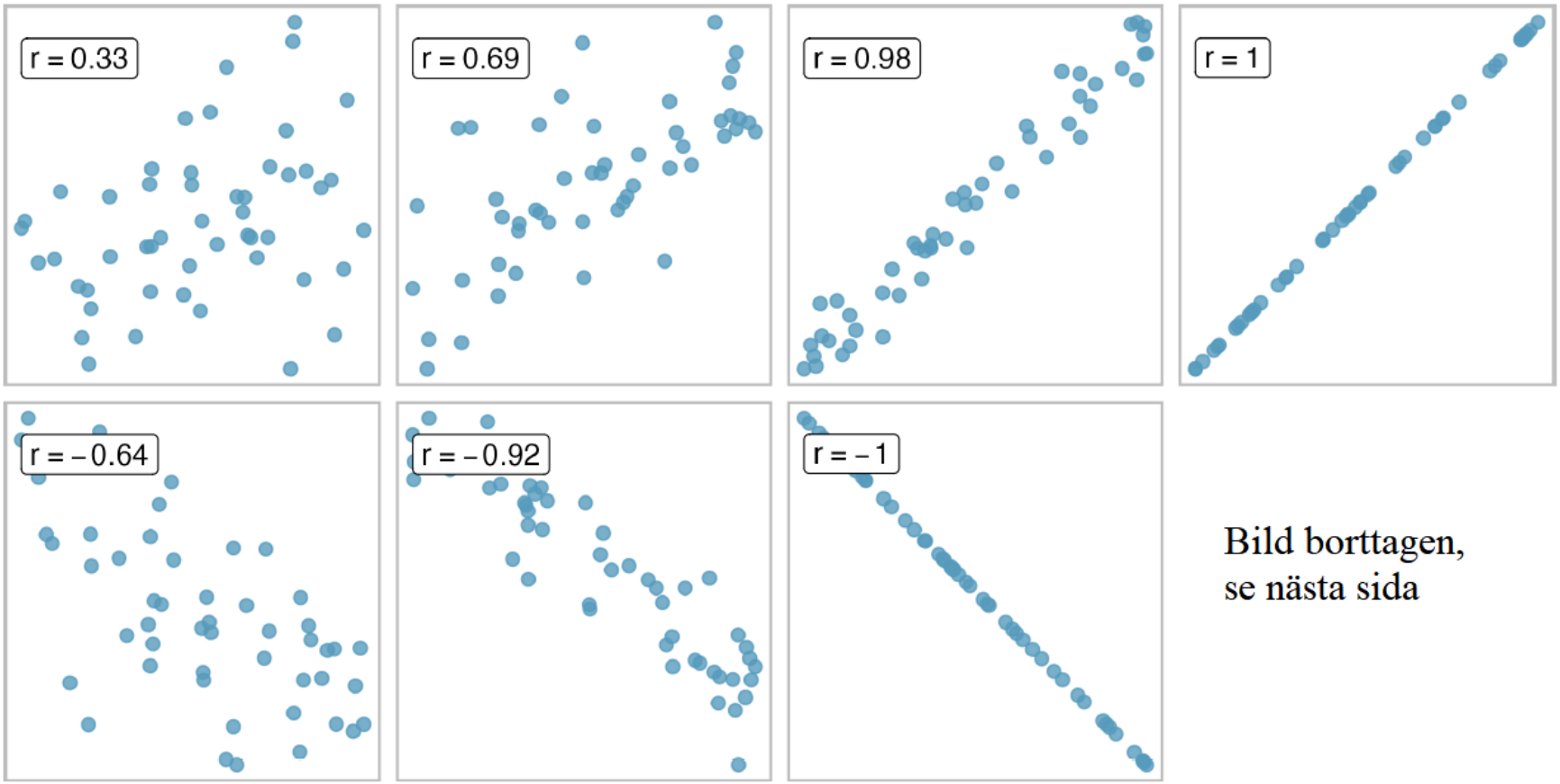


Bild borttagen,
se nästa sida

Figure 7.10: Sample scatterplots and their correlations. The first row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a lower value in the other.

Vad säger korrelationskoefficienten i dessa fall?

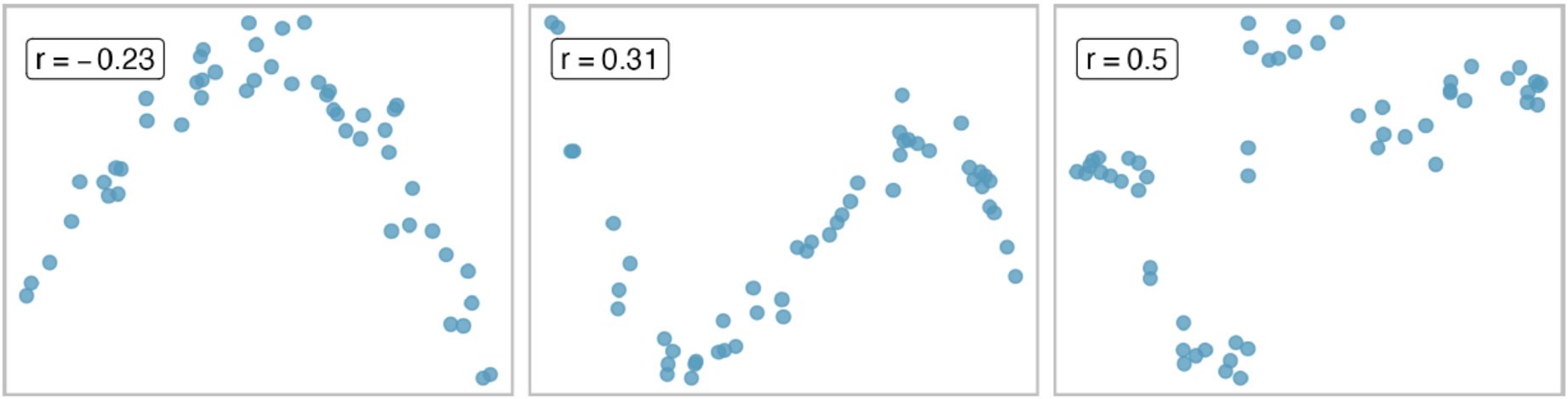
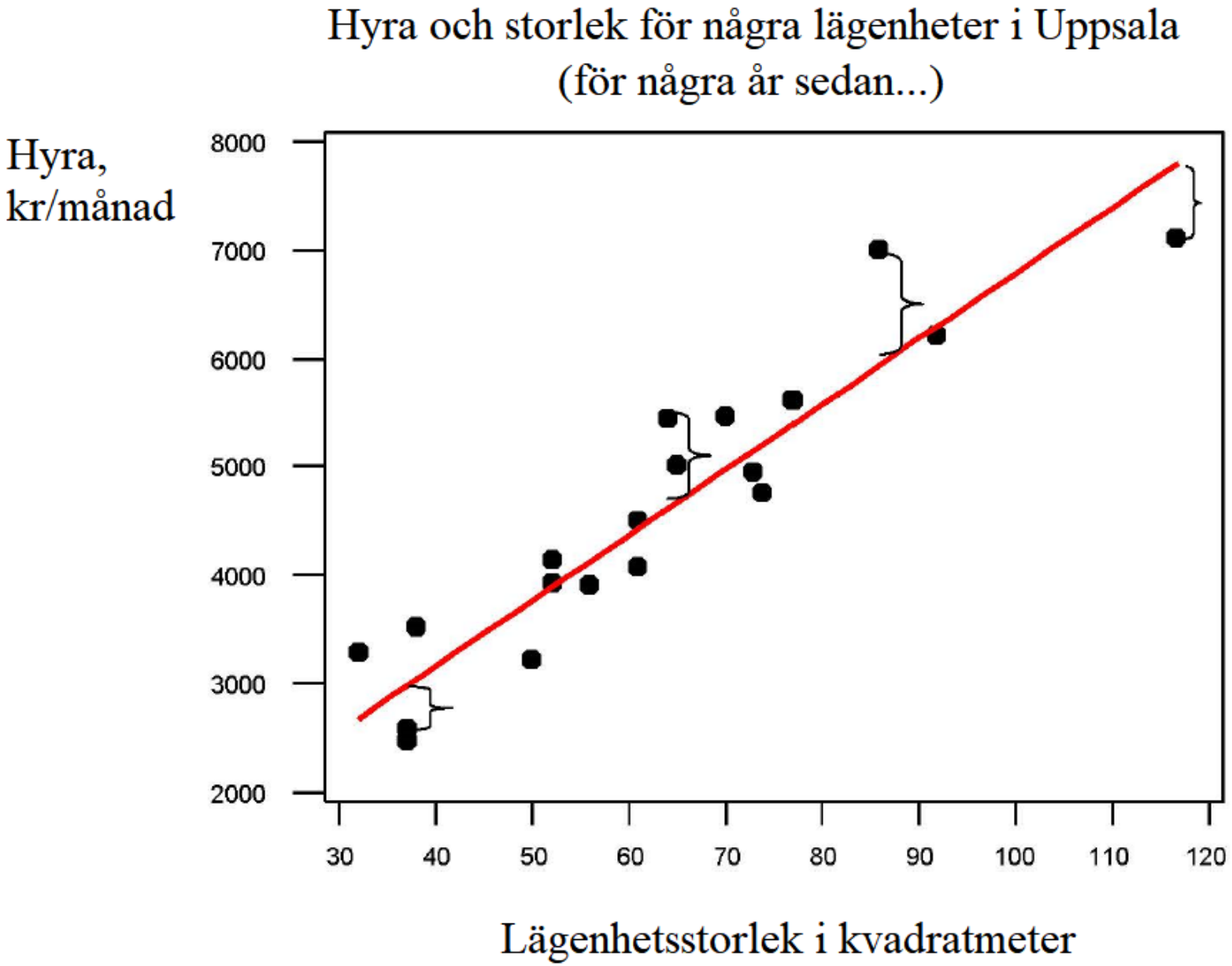


Figure 7.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, because the relationship is not linear, the correlation is relatively weak.

Vi går vidare och vill skatta hur ett linjärt samband ser ut

- Vi vill ta fram ”den bästa” linje som karakteriserar hur sambandet mellan y och x ser ut - **Enkel linjär regression**
- Det finns olika sätt att göra en sådan skattning
- Vi väljer den linje för vilken summan av de kvadrerade avstånden, i y-led, mellan respektive datapunkt och linjen, är så liten som möjligt
- Metoden kallas **minsta kvadratmetoden** (Least Squares)
- Vi får fram en rät linje som vi bland annat kan använda för att prediktera - skatta y för ett nytt x-värde

Illustration, minsta kvadratmetoden



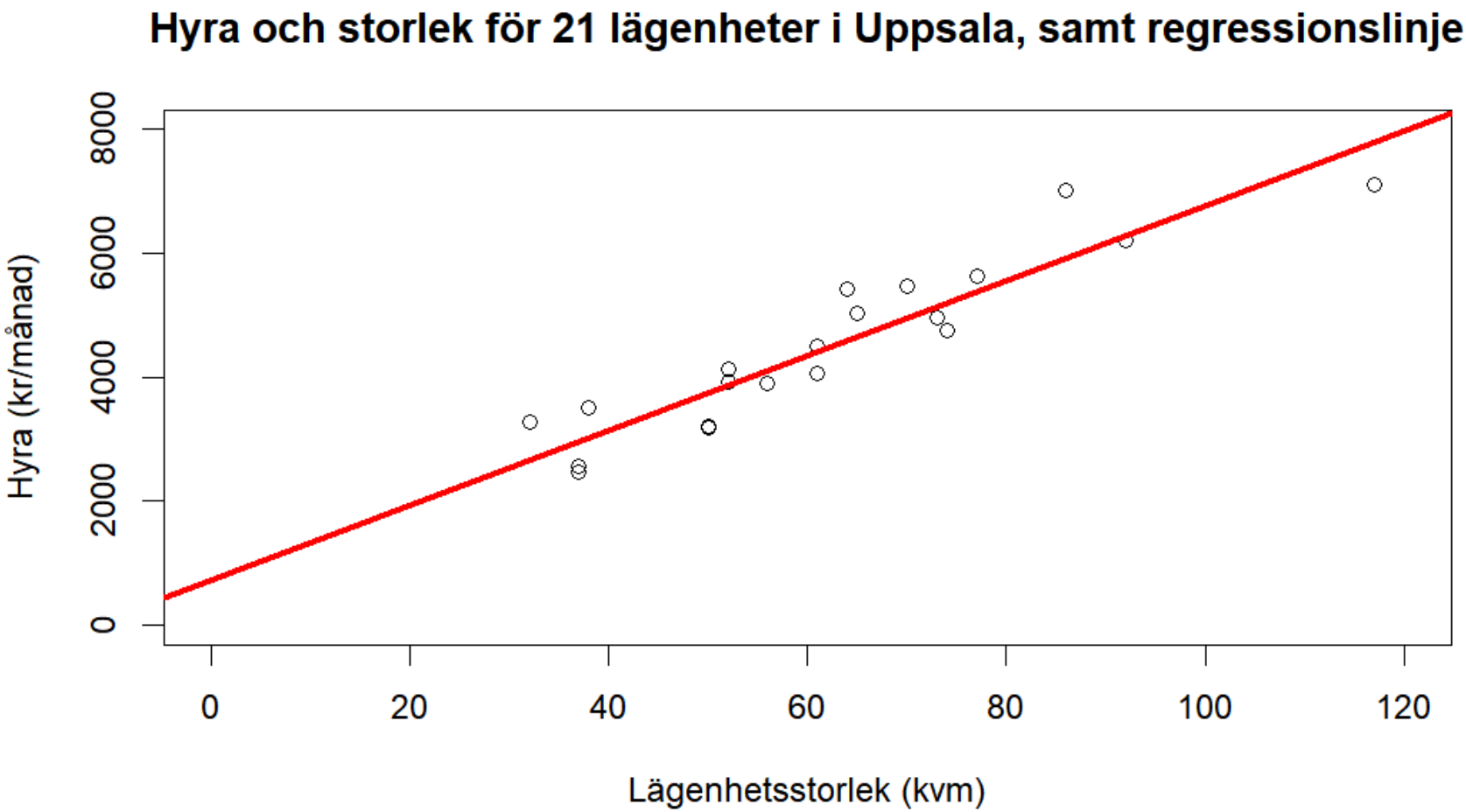
Vi analyserar i R, definiera data

```
Uppsaladata <- data.frame(
                yta=c(61, 50, 32, 74, 61, 70, 52, 64, 65, 38, 37,
                      37, 50, 117, 86, 50, 73, 77, 52, 56, 92),
                hyra=c(4490, 3211, 3265, 4750, 4063, 5471, 4120,
                       5432, 5020, 3512, 2456, 2560, 3179, 7110,
                       7019, 3199, 4953, 5623, 3919, 3898, 6219))
```

Vi kör en regression i R (mer detaljer och analys på F8)

```
> Uppsalamodell <- lm(hyra~yta, data=Uppsaladata)
> summary(Uppsalamodell)

Call:
lm(formula = hyra ~ yta, data = Uppsaladata)

Residuals:
    Min      1Q  Median      3Q     Max
-693.28 -450.36  -70.95  364.44 1092.24

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  720.923    370.244   1.947   0.0665 .
yta           60.533      5.713  10.595 2.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 525.5 on 19 degrees of freedom
Multiple R-squared:  0.8553,    Adjusted R-squared:  0.8476
F-statistic: 112.3 on 1 and 19 DF,  p-value: 2.057e-09
```

Skattad regressionslinje

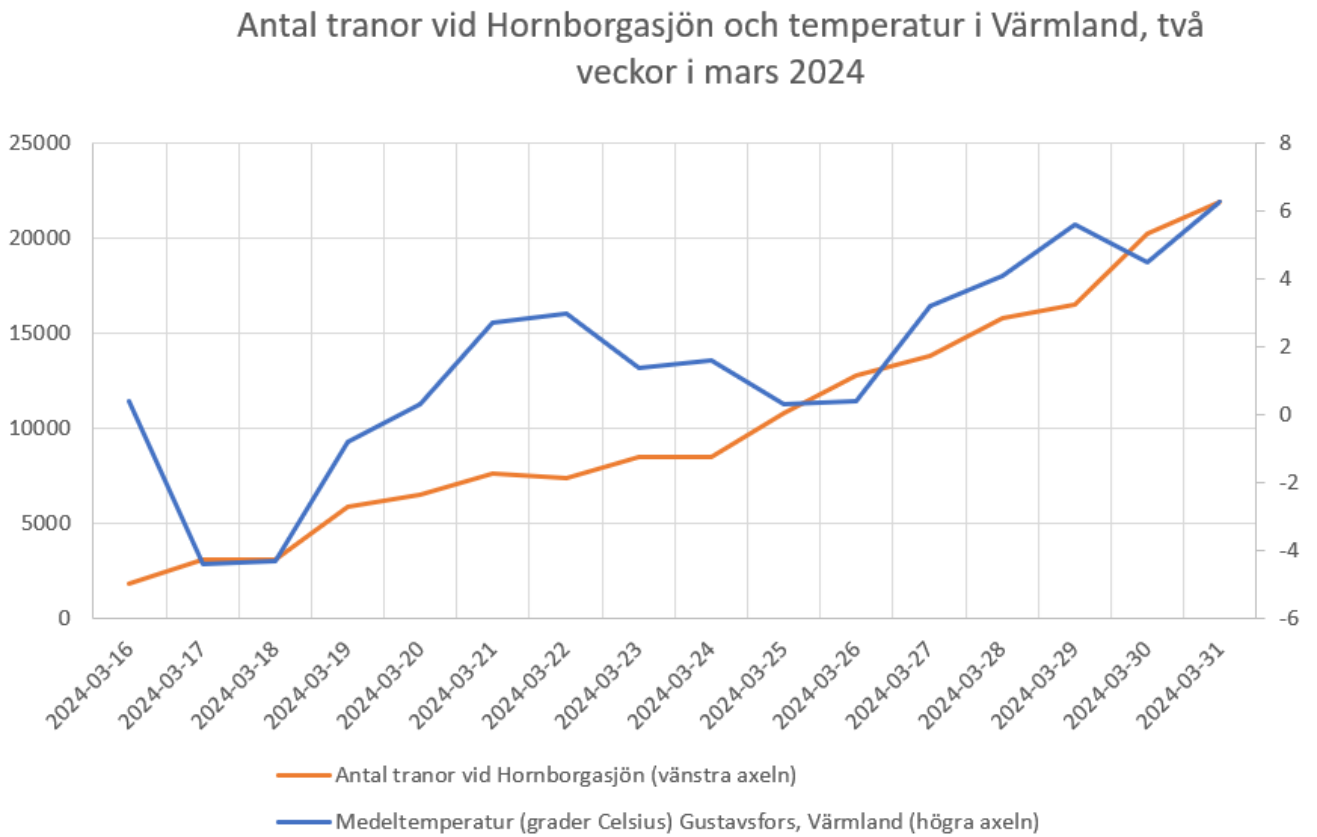


Korrelation är inte samma som kausalitet. Regressionssamband är inte samma som kausalitet.

- **Kausalitet** - orsakssamband, ”x orsakar y”
- Vi har sett spridningsdiagram på exv. lägenhetsstorlek och hyra, korruptionsmått och datatillgång, BNP/capita och barnadödlighet
- Kan vi säga något om kausalitet?
- På F3 diskuterade vi observationsstudier och experiment, och skillnader gällande vad vi eventuellt kan säga om kausalitet
- Om kausalitet, se kursboken, kap. 1-2, kapitel 7.2.3, 12.1, 16.1.1 (samma exempel som i F5)
- Vi skulle kunna ha kausala samband (medicin orsakar hälsoförbättring), men vi kan inte avgöra om vi har kausalitet med bas enbart i ett dataset, en korrelationskoefficient eller ett regressionsresultat.

Se upp för skensamband (spurious correlation)

- Tranor ”mellanlandar” i Västergötland på väg till bla. Värmland
- Korrelation mellan temperatur i Värmland och antal tranor vid Hornborgasjön, i grafen, är >0.8 (Källor: SMHI, Naturum)



- Googla ”spurious correlation” för kända exempel!

Denna version av dokumentet: 2025-04-04

Materialet i Statistisk översiktskurs har tagits fram av Ulf Högnäs och Anders Fredriksson, med inspiration och ibland direkt användande av material från andra kurser och personer, bland annat kurserna Statistik och dataanalys 1-3, med material av Michael Carlson, Ellinor Fackle Fornius, Jessica Franzén, Oskar Gustafsson, Oscar Oelrich, Mona Sfaxi, Karl Sigfrid, Mattias Villani, med flera.