# Lecture 6

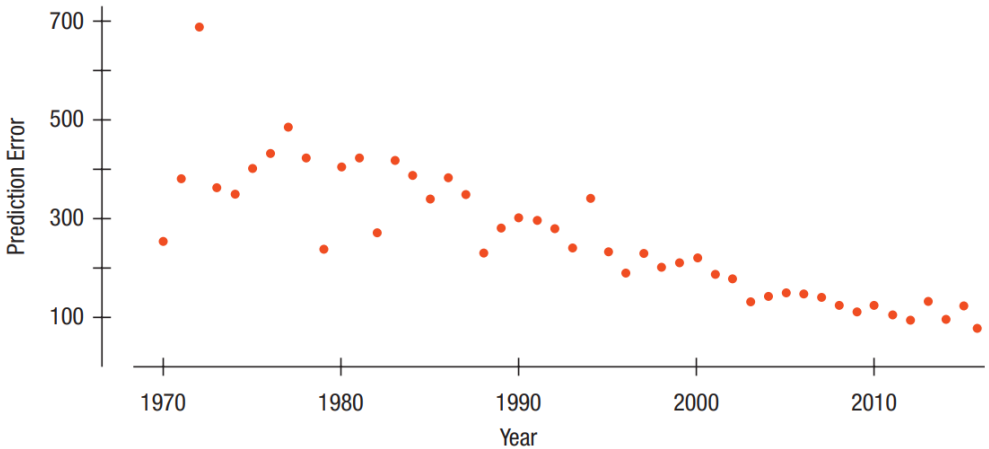
karl.sigfrid@stat.su.se

# Vad har vi gjort hittills, och vad vi ska göra nu

- I statistik studerar vi ofta samband mellan variabler.
- Vi har hittills studerat
  - Samband mellan två eller flera kategoriska variabler, med korstabeller och stapeldiagram.
  - Samband mellan en numerisk och en kategorisk variabel, med låddiagram och histogram.
- Nu ska vi studera **samband mellan två numeriska variabler**.

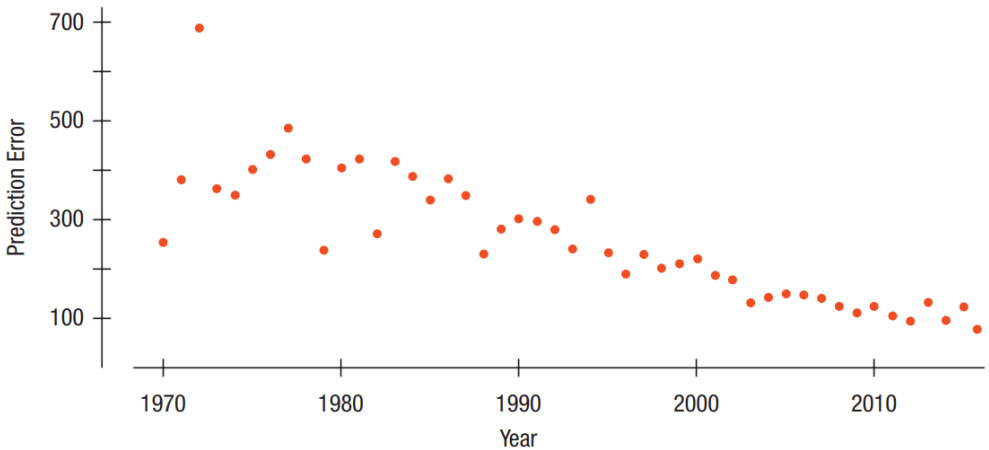
# Tidsserier

- Vi har redan tittat på diagram över tidsserier.
- Tidsserier, där vi studerar en numerisk variabel över en tidsperiod, är exempel på samband mellan två numeriska variabler. Detta eftersom även tiden är en numerisk variabel.
- Figuren visar en tidsserie av årsvisa genomsnittliga prediktionsfel i nautiska mil i lokalisering av orkaner i Atlanten:



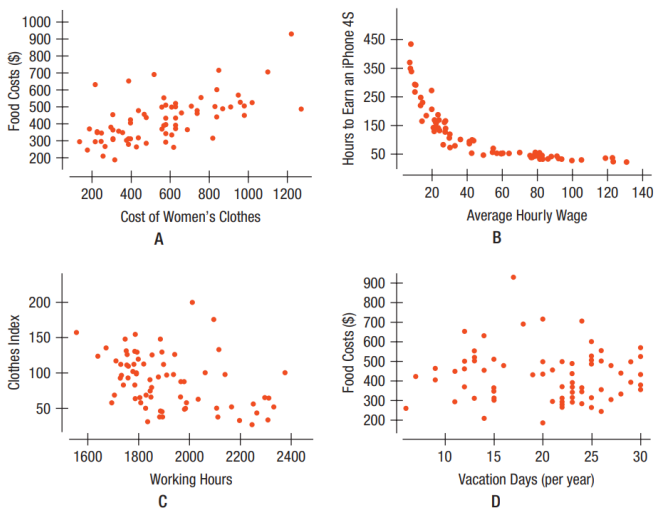
# Spridningsdiagram (scatter plots)

- Samband mellan två numeriska variabler kan illustreras i **spridningsdiagram (scatter plot)**.
- Det här diagrammet med en tidsserie är alltså ett exempel på ett spridningsdiagram.



# Spridningsdiagram (scatter plots)

- Vi har fyra spridningsdiagram med data om prisnivåer i 73 städer. Varje röd punkt representerar ett land. I diagrammen kan vi notera
  - Sambandets riktning. Är det positivt eller negativt?
  - Sambandets styrka. Är det starkt (tydligt) eller svagt (otydligt)?
  - Huruvida sambandet är linjärt. Följer mönstret en rak linje?

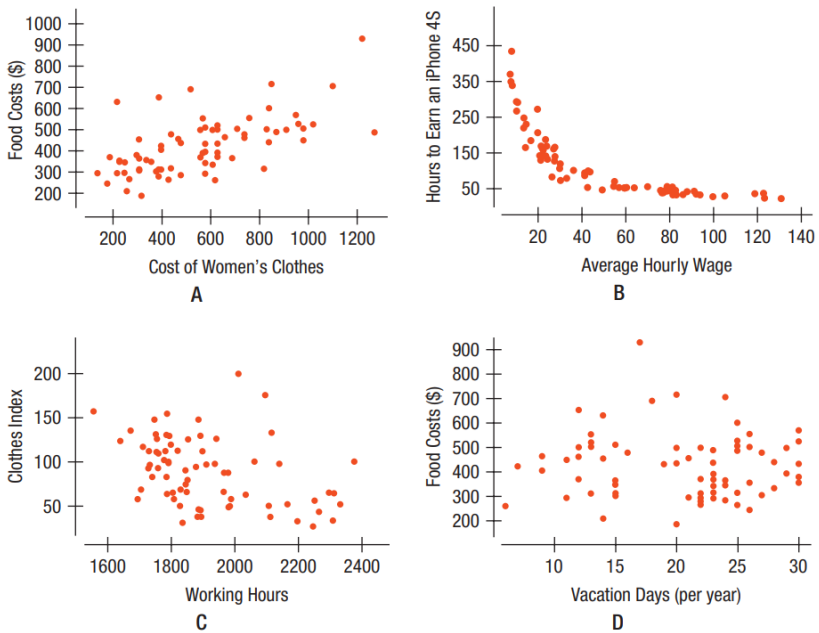





Stockholm University

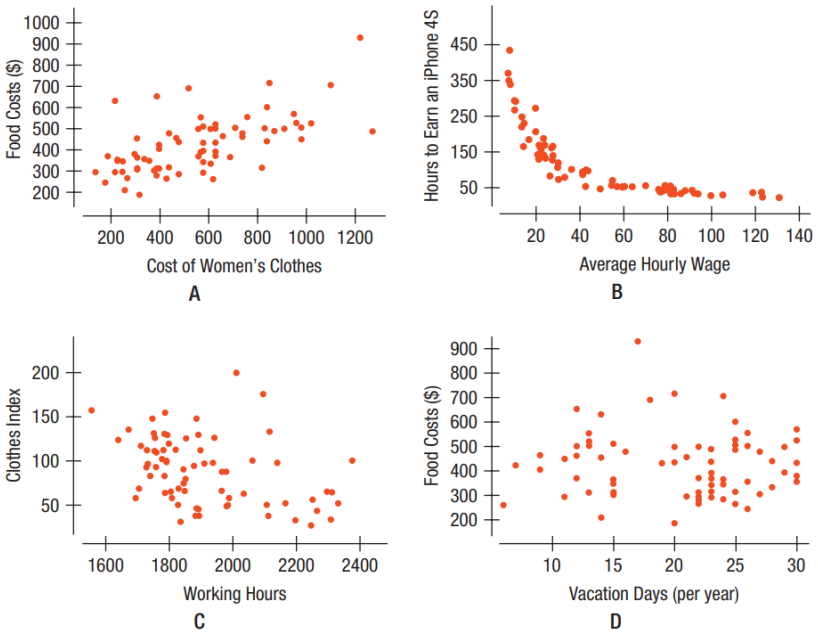
# Spridningsdiagram

- **Diagram A: Matpriser(y) vs. Klädpriser(x)**
  - Sambandet är positivt och ganska starkt (Vi ser det tydligt).
  - Sambandet är på ett ungefär linjärt.



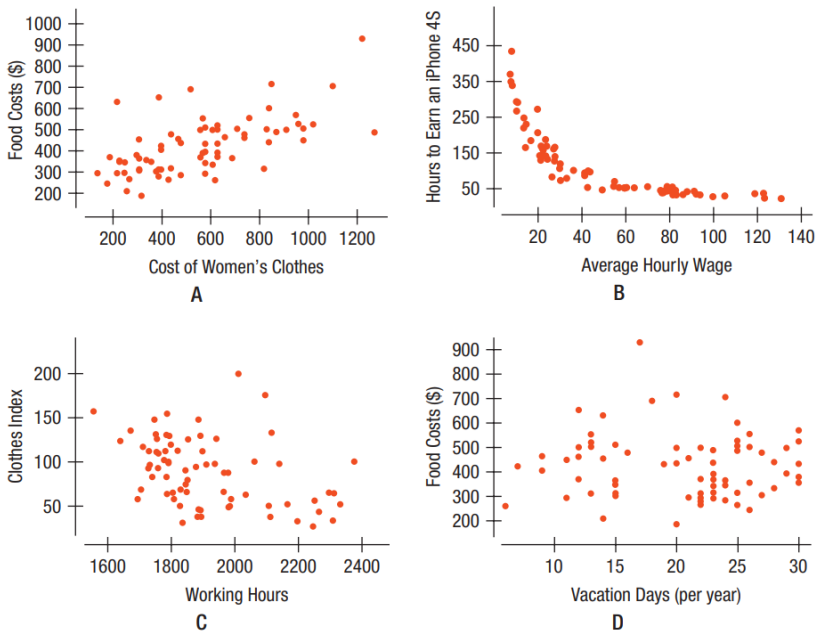
# Spridningsdiagram

- **Diagram B: Timmar att tjäna ihop till en Iphone 4S(y) vs. Medellön/tim(x)**
  - Sambandet är negativt.
  - Sambandet är *inte* linjärt.



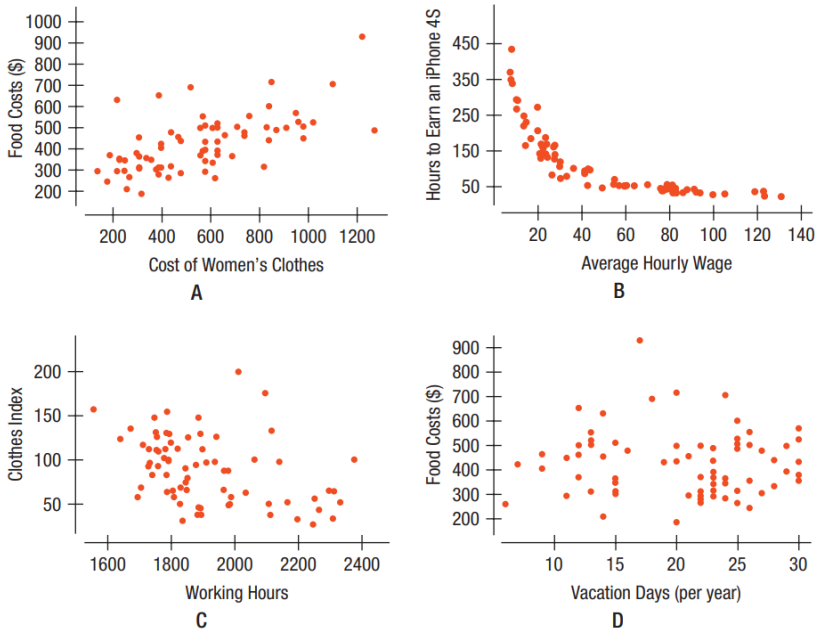
# Spridningsdiagram

- **Diagram C: Klädprisindex (y) vs. Arbetstid per år (x)**:
  - Sambandet är negativt och ganska svagt (Vi ser det inte så tydligt).
  - Sambandet är någorlunda linjärt.



# Spridningsdiagram

- **Diagram D: Matpriser (y) vs. Semesterdagar per år (x)**:
  - Inget samband, vare sig positivt eller negativt.

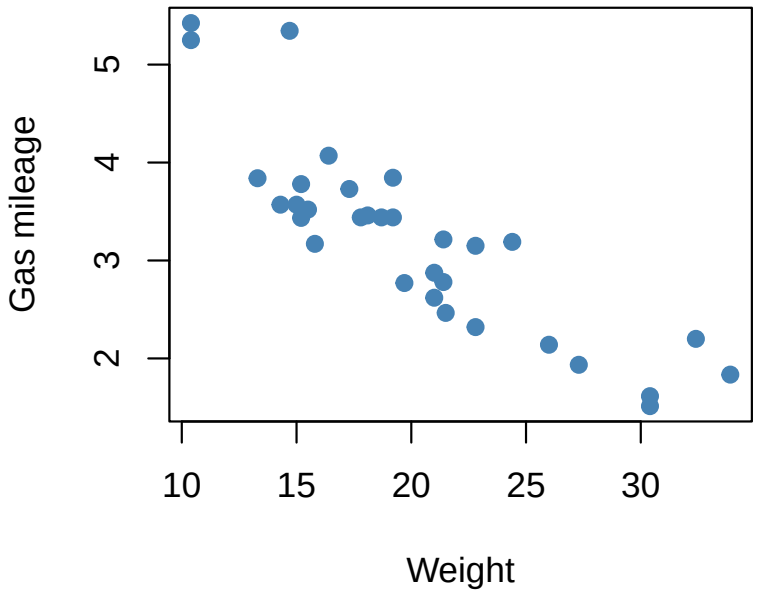


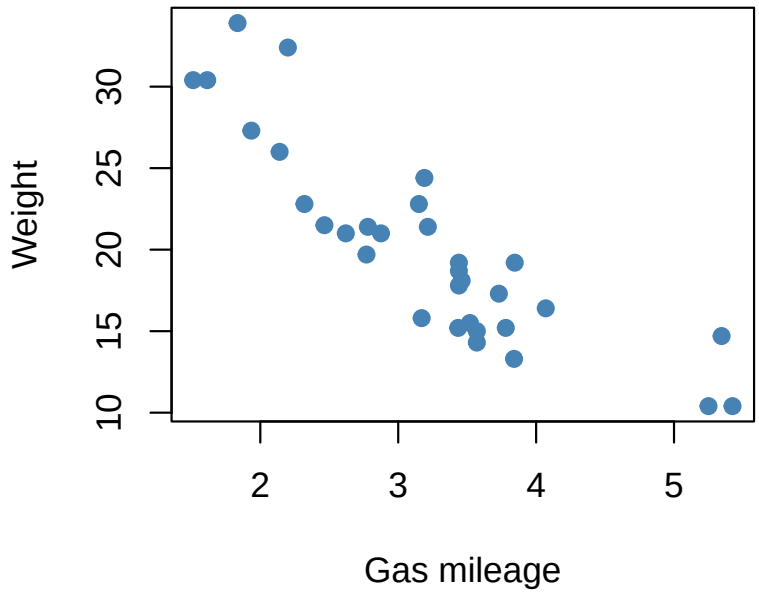
# Spridningsdiagram - X och Y

- Hur bestämmer vi vilken variabel som är y och vilken som är x?
- Ofta finns det inget korrekt svar, men vi kan använda tumregler för vad som är rimligast.
- Enklast att tänka i termer av prediktion.
  - Variabeln av intresse att prediktera väljs som y.
  - Variabeln som hjälper oss att prediktera väljs som x.
- Är det rimligt att fråga sig om variabeln på y-axeln påverkas av variabeln på x-axeln? Blir frågan mer rimlig om variablerna byter plats?

# Spridningsdiagram - X och Y

- Bilden visar räckvidden per gallon bensin mot vikten hos en bil.
- Vilken av följande frågor låter rimligast?
  - Brukar bensinförbrukningen vara högre för bilar som väger mer?
  - Brukar bilar vara tyngre om de förbrukar mer bensin?
- Om den första frågan låter mer rimlig väljer vi grafen till vänster.



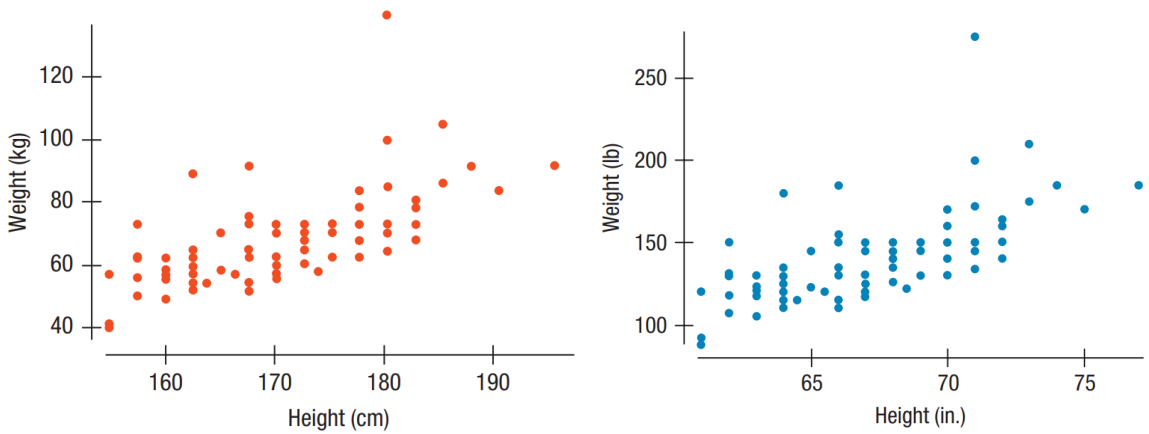



# Spridningsdiagram - X och Y

- Det finns flera olika namn som används för x- respektive y-variabeln.
- Variabeln y kallas i vissa sammanhang **responsvariabeln (response variable)**.
- Variabeln x kallas ibland för **förklaringsvariabeln (explanatory variable)**.
- Ett annat vanligt namn för y-variabeln är **beroende variabeln (dependent variable)**, och x-variabeln kallas då den **oberoende variabeln (independent variable)**.
- Andra vanligt förekommande namn för x variabeln är **prediktor (predictor)** och **kovariat (covariate)**.
- Inom maskininlärning kallas förklarande variabler **features**.

# Linjära samband mellan numeriska variabler

- Nu ska vi undersöka hur vi kan mäta hur *starkt* ett *linjärt* samband är mellan två numeriska variabler.
- Vi vill använda ett mått som är oberoende av vilka enheter vi använder.
- **Exempel:** Figur 6.2 och 6.3 i De Veaux et al. (2021) visar ett antal personers vikt och längd. Oavsett om viken anges i kg eller lbs ser vi samma samma samband. Den enda skillnaden är olika skalor på y-axeln i de två graferna.

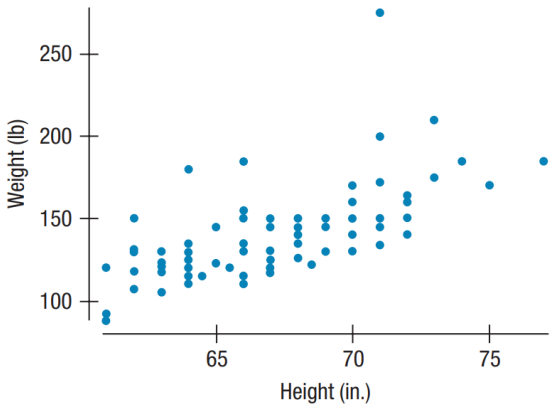





Stockholm University

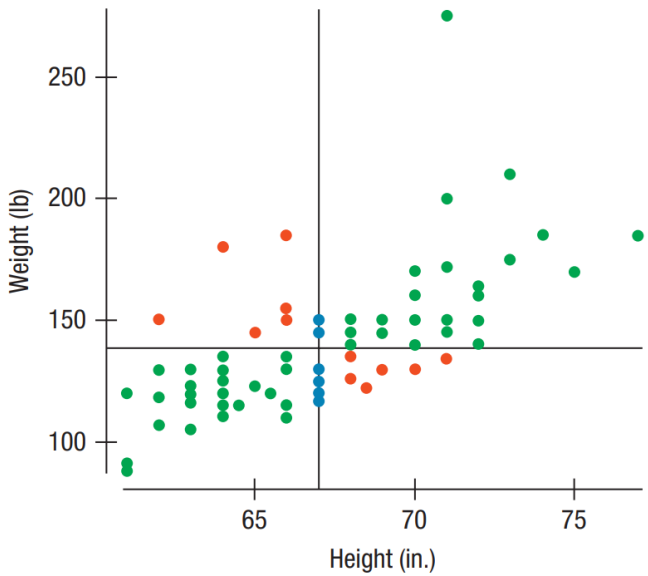
# Linjära samband mellan numeriska variabler

- Vi ser på bilden att sambandet mellan variablerna är positivt.
- De datapunkter som har ett stort värde på x-axeln tenderar att också ha ett stort värde på y-axeln.
- För att göra sambandet ännu tydligare kan vi markera variablernas medelvärden i grafen. Vi markerar $\bar{x}$ med ett vertikalt streck och $\bar{y}$ med ett horisontellt streck (se bilden på nästa sida).



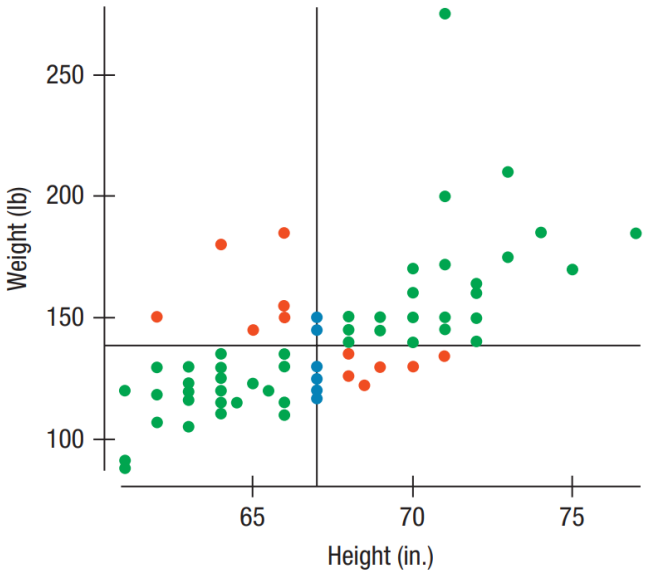
# Linjära samband mellan numeriska variabler

- Linjerna som representerar variablernas medelvärden delar in grafen i fyra delar, och färgen på punkterna skiljer sig åt mellan delarna.



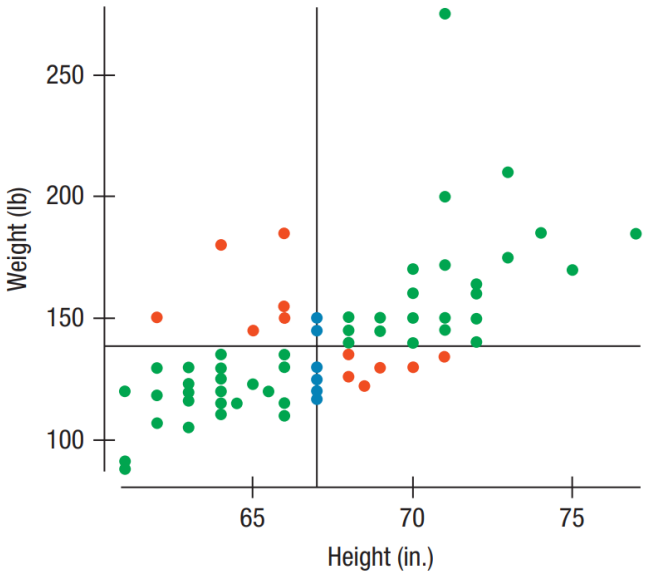
# Linjära samband mellan numeriska variabler

- Punkterna är **gröna** om *både x och y är större* än medelvärdet eller om *både x och y är mindre* än medelvärdet.
- Punkterna är **röda** om de är mindre än medelvärdet på den ena skalan och större än medelvärdet på den andra skalan.



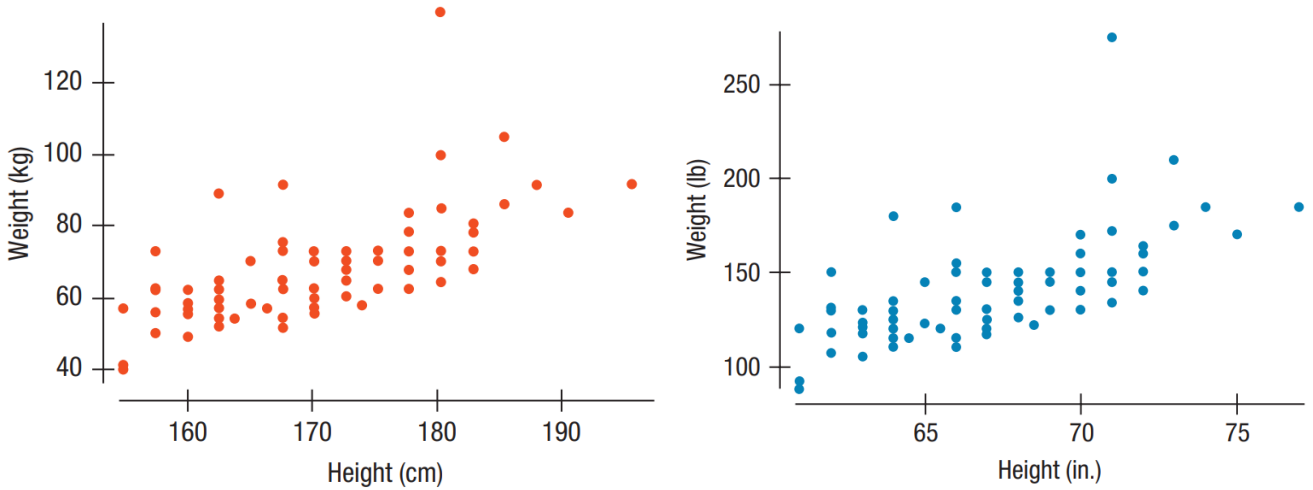
# Linjära samband mellan numeriska variabler

- De **gröna** punkterna påverkar sambandet i *positiv* riktning.
- De **röda** punkterna påverkar sambandet i *negativ* riktning.
- De gröna punkterna är fler. Det indikerar att sambandet är positivt.
- Fler röda punkter hade indikerat ett *negativt* samband.



# Linjära samband mellan numeriska variabler

- Att räkna punkter i olika sektioner av plotten är ett enkelt och intuitivt sätt att bedöma riktningen på sambandet.
- Men vi vill ofta kunna räkna ut *hur starkt* det linjära sambandet är.
- Vi vill ha ett mått som inte beror på vilka enheter vi använder för variabler. Oavsett om vi mäter längd i meter eller tum ska resultatet bli detsamma.

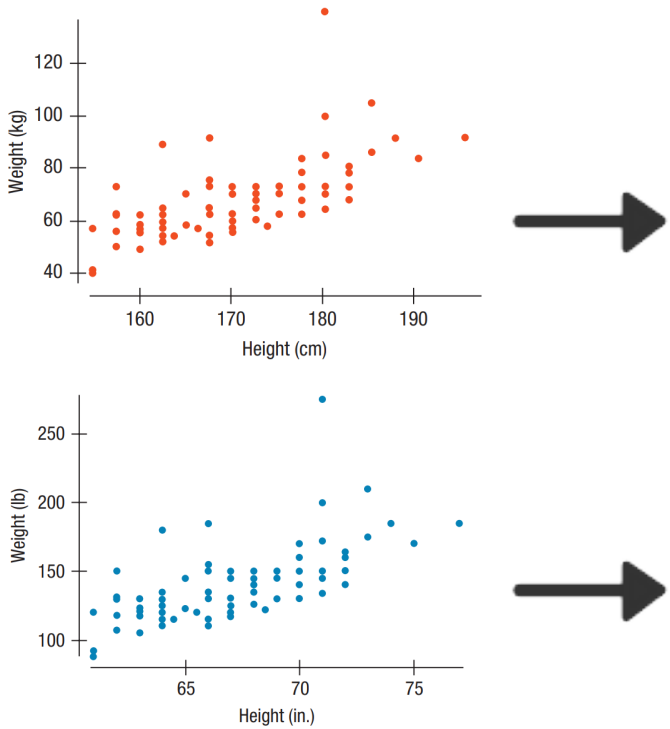


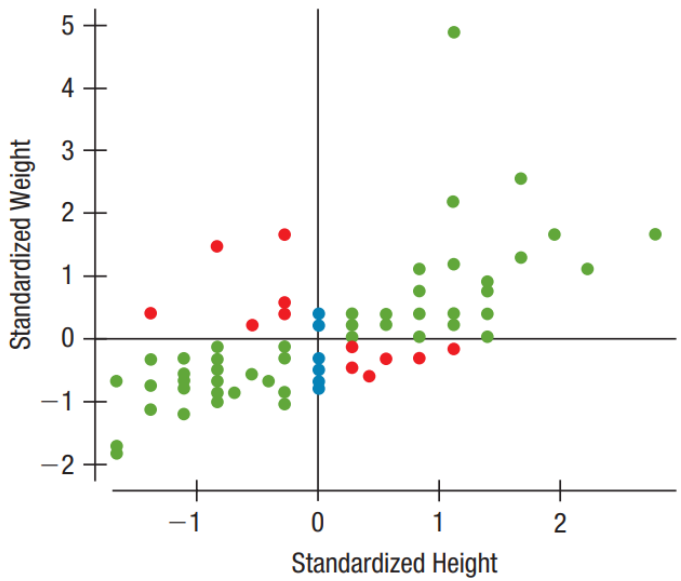



Stockholm University

# Standardiserade numeriska variabler

Vi har tidgare räknat ut **z-värdet** för numeriska variabler. z-värdet blir lika stort oavsett vilken enhet vi använder för den ursprungliga variabeln (exempelvis meter eller tum). Vi kallar därför z för en *standardiserad* variabel.






Stockholm University

# Standardiserade numeriska variabler

- Följande exempel i R som illustrerar att två olika variabler som mäter samma vikt också får samma värde efter att ha standardiserats.
- Variabeln *weight_pounds* anger vikten på olika bilmodeller i 1000-tals pund.
- Genom att multiplicera med 0.454 får vi vikten i 1000-tals kg.

```
head(weight_pounds)
```

```
[1] 2.620 2.875 2.320 3.215 3.440 3.460
```

```
weight_kg <- weight_pounds * 0.454
head(weight_kg)
```

```
[1] 1.18948 1.30525 1.05328 1.45961 1.56176 1.57084
```

# Standardiserade numeriska variabler

Nu standardiserar vi var och en av variablerna till z-värden med formeln

$$z = \frac{x - \bar{x}}{s}, \quad \bar{x} = \frac{\sum x}{n}, \quad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}.$$

```
weight_pounds_z <- (weight_pounds - mean(weight_pounds)) /
  sd(weight_pounds)
head(weight_pounds_z) |> round(3)
```

```
[1] -0.610 -0.350 -0.917 -0.002  0.228  0.248
```

```
weight_kg_z <- (weight_kg - mean(weight_kg)) /
  sd(weight_kg)
head(weight_kg_z) |> round(3)
```

```
[1] -0.610 -0.350 -0.917 -0.002  0.228  0.248
```

Vi ser att den standardiserade variabeln får samma värden oavsett vilken viktenhet vi använde.

# Korrelationskoefficienten

Det mått som vi använder för att mäta det linjära sambandet kallas **korrelationskoefficienten (correlation coefficient)** och räknas ut med formeln

$$r = \frac{\sum z_x z_y}{n-1}$$

- Eftersom beräkningen använder de standardiserade variablerna $z_x$ och $z_y$ blir korrelationskoefficienten oberoende av enheterna för $x$ och $y$.
- Korrelationen är alltid ett tal mellan -1 och 1, dvs $-1 < r < 1$. Om $r$ ligger nära 1 eller -1 visar det att korrelationen är stark. Om $r$ ligger nära 0 visar det att korrelationen är svag.
- Vi kan ibland använda beteckningen $r_{,xy}$ för att betona att det är korrelationen mellan variablerna $x$ och $y$.

# Korrelationskoefficienten

- Om korrelationskoefficienten är positiv har vi ett positivt linjärt samband mellan variablerna.
- Omkorrelationskoefficienten är negativ har vi ett negativt linjärt samband.
- Korrelationsmåttet är *symmetriskt*. Korrelationen mellan $x$ och $y$ är samma sak som korrelationen mellan $y$ och $x$.

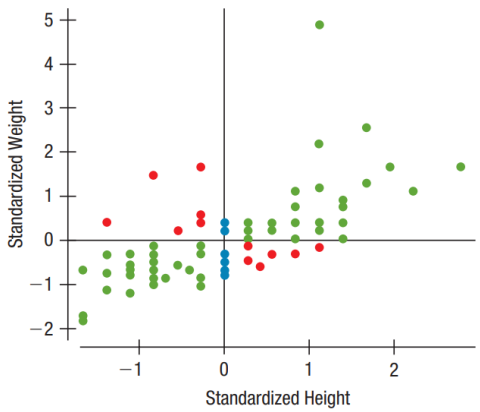
# Korrelationskoefficienten

$$r = \frac{\sum z_x z_y}{n-1}$$

- Vi ser i formeln för korrelationskoefficienten $r$ att korrelationen blir positiv om uttrycket $\sum z_x z_y$ är positivt, och att korrelationen är starkare när uttrycket är stort.
- När uttrycket $\sum z_x z_y$ är negativt blir korrelationen negativ, och om uttrycket är ett stort negativt tal är den negativa korrelationen starkare.

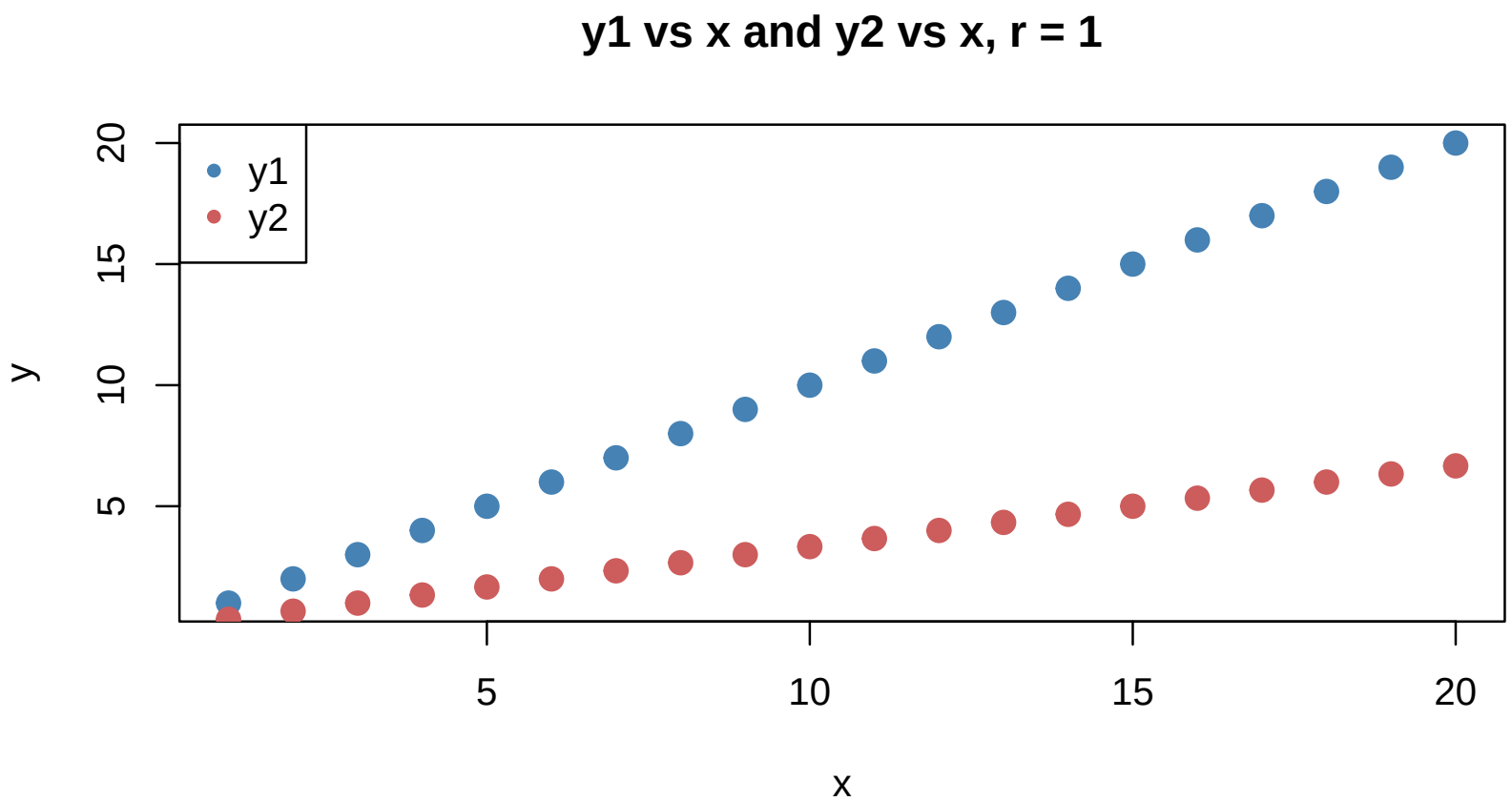
# Korrelationskoefficienten

- Notera att multiplikationen $z_x z_y$ är positiv när
  - både $z_x$ och $z_y$ är positiva.
  - både $z_x$ och $z_y$ är negativa
- På motsvarande sätt är multiplikationen $z_x z_y$ är negativ när
  - $z_x$ är positiv och $z_y$ är negativ, eller vice versa.



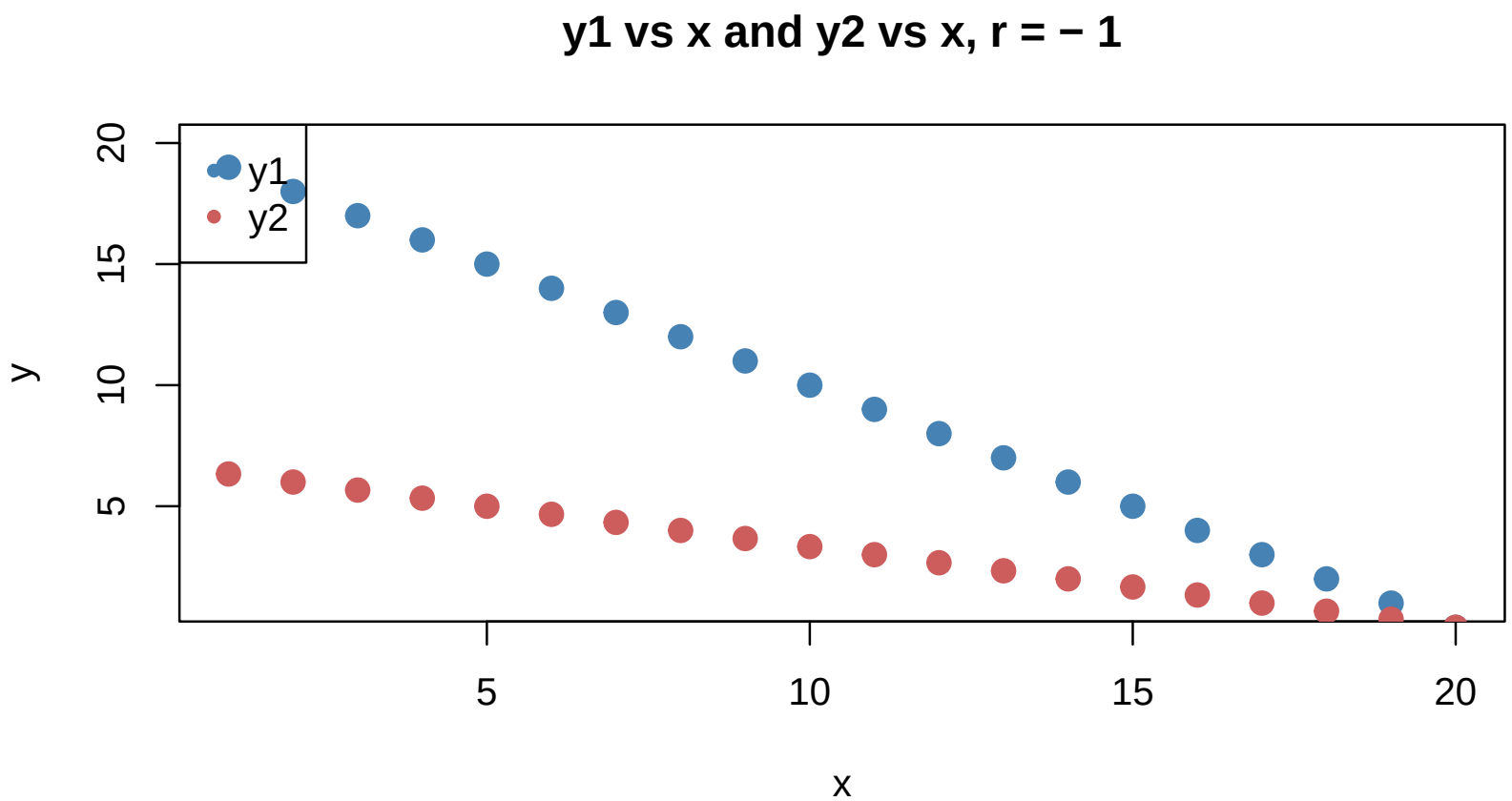
# Korrelationskoefficienten

Den här grafen visar två exempel på hur ett perfekt positivt linjärt samband kan se ut. Både $y1$ och $y2$ är perfekt korrelerade med $x$ trots att linjerna har olika lutning, dvs för båda sambanden gäller att $r = 1$.



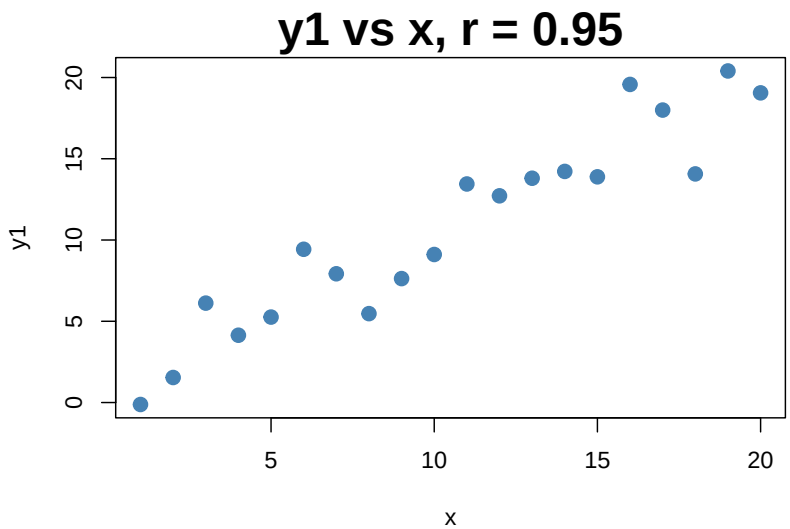
# Korrelationskoefficienten

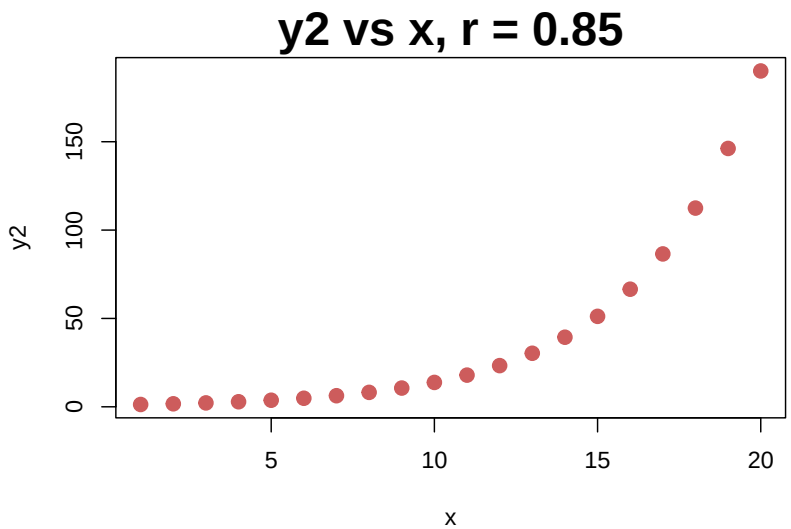
Den här grafen visar två exempel på hur ett perfekt negativt linjärt samband kan se ut. Både $y1$ och $y2$ är perfekt korrelerade med $x$ trots att linjerna har olika lutning, dvs för båda sambanden gäller att $r = -1$.



# Korrelationskoefficienten

- De här graferna visar två exempel på linjära samband som inte är perfekta.
- Den blå grafen visar ett en **starkare korrelation** än den röda grafen, trots att punkterna är utspridda medan de röda punkter följer en tydlig linje.
- Även om de blå punkterna är mer spridda följer de en **rak linje** medan de röda punkterna följer en böjd linje. Kom ihåg att korrelation mäter styrkan i det **linjära** sambandet.



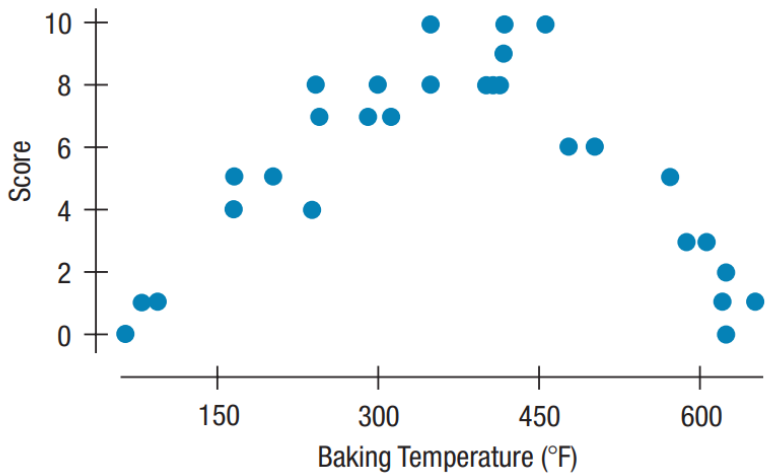






Stockholm University

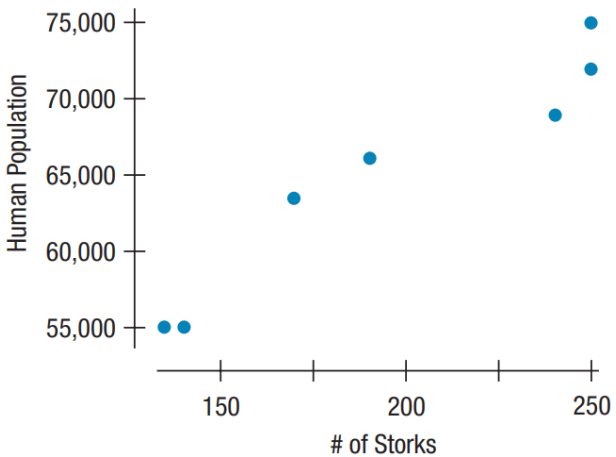
# Korrelationskoefficienten

- Att det finns ett samband betyder **inte** att det finns korrelation.
- Kom ihåg att korrelation är ett **linjärt** samband mellan två numeriska variabler.
- På den här grafen finns ett uppenbart samband mellan god smak (y) och baktemperatur (x) för kladdkakor. Ändå är korrelationen nära noll!



# Korrelation är inte kausalitet

- Figur 6.10 i De Veaux et al. (2021) visar ett tydligt linjärt samband mellan antalet storkar och befolkningsmängden under sju år i en stad i Tyskland.
- Det sägs ibland att storkar kommer med barn, och det skulle ju kunna förklara varför antalet människor blir fler när det kommer många storkar.
- Korrelationskoefficienten är $r = 0.97$, så det linjära sambandet är starkt.

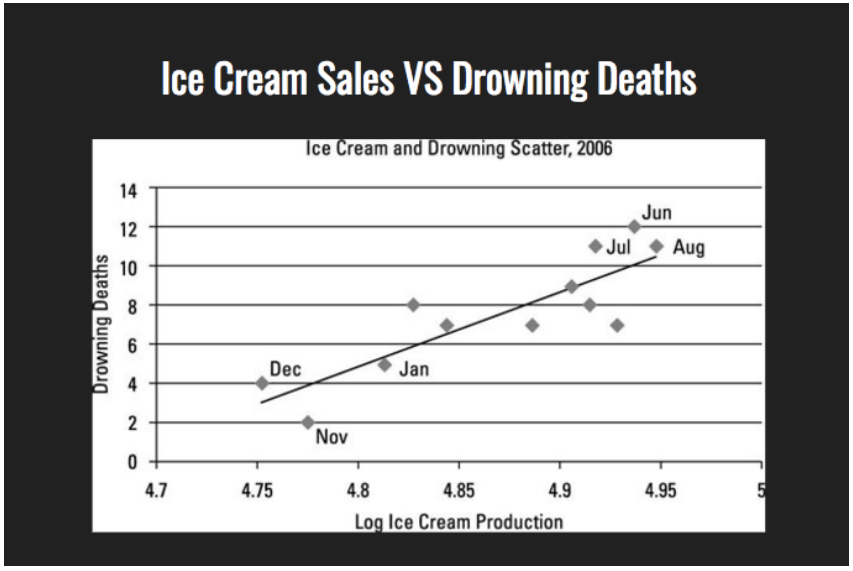


# Korrelation är inte kausalitet

- Slutsatsen om att fler storkar leder till fler människor är en kausal tolkning av korrelationen.
- Att vi ska undvika att dra snabba slutsatser av det här slaget betyder inte att det saknas kausalitet.
- Kanske är sambandet det omvända: Storkar bygger bon på skorstenar. Fler människor betyder fler hus med skorstenar där storkarna kan bygga bon.
- Med vissa samband kan det också vara så att det inte finns kausalitet åt något håll. Det kan istället vara en tredje variabel som förklarar både $x$ och $y$. En sådan osynlig variabel kallas för en **dold variabel (lurking variable)**.

# Korrelation är inte kausalitet

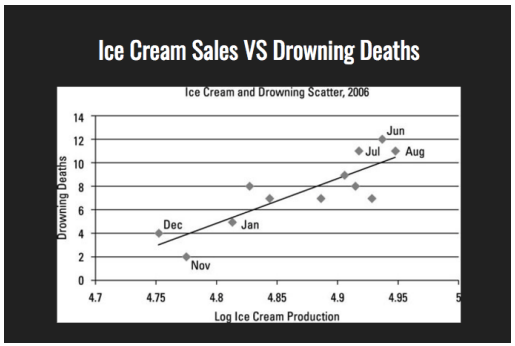
Figuren från Harvard Onlines Facebooksida visar antalet drunkningsolyckor (y) och mängden glass som produceras (x) under en månad.




**Kausal tolkning:** När det produceras mer glass måste folk äta upp glassen, och för mycket glass gör att människor inte orkar simma lika långt.

# Korrelation är inte kausalitet

**Omvänd kausal tolkning:** Folk blir deprimerade när de läser om drunkningsolyckor och tröstäter glass. Därför måste det produceras mer glass under månader då många drunknar.

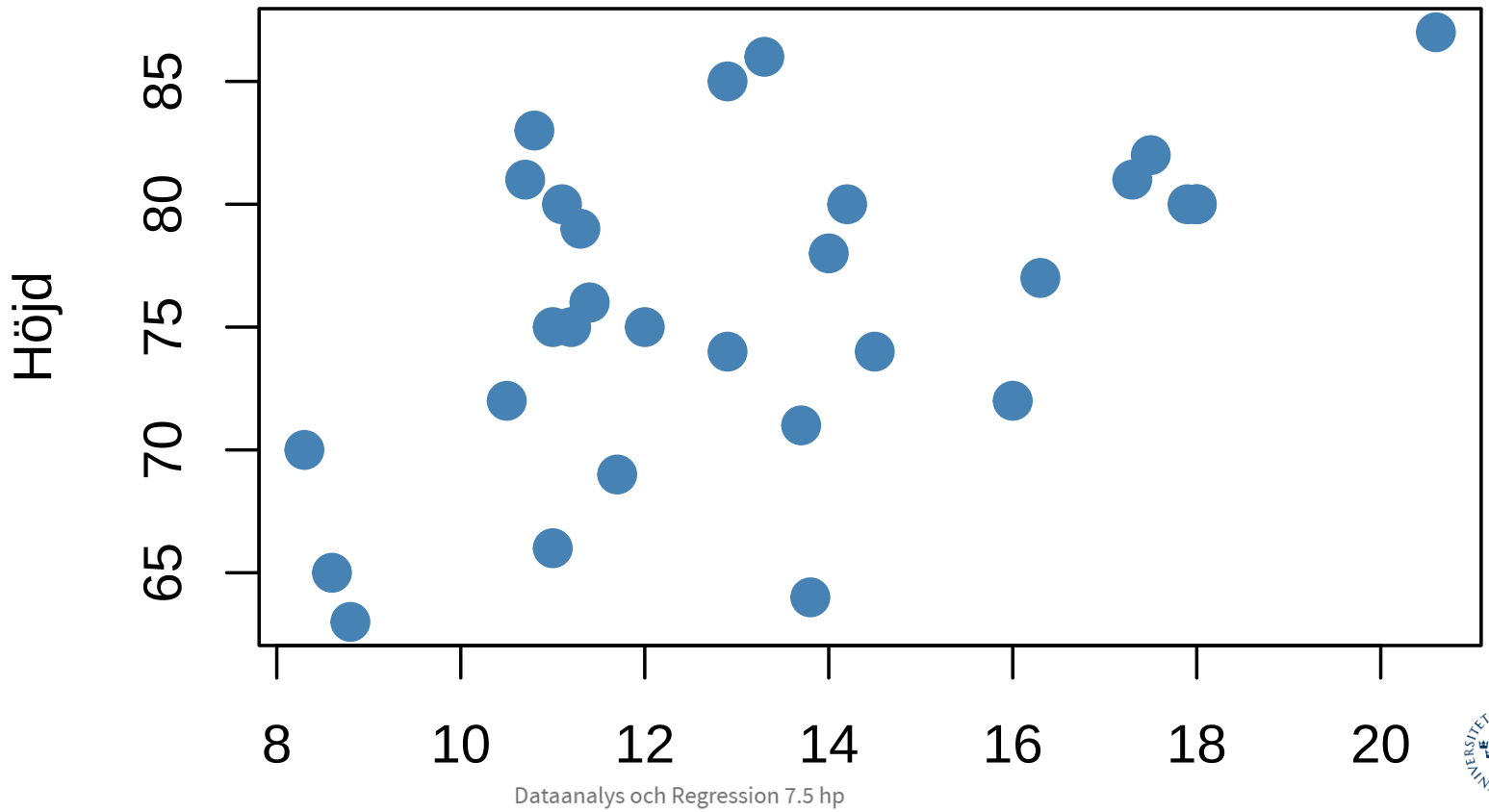



Ingen av de kausala tolkningarna är särskilt bra, men vi har en uppenbar *lurking variable.* **Årstiden** påverkar både antalet drunkningsolyckor och glassproduktionen. Folk badar mer och äter mer glass under sommarmånader.

# Korrelationer i R

- I R kan vi enkelt rita ett spridningsdiagram för 2 numeriska variabler.
- Vårt dataset *trees* anger omkrets (girth), höjd (height) och volym (volume) för 31 träd.

```
data(trees)
plot(x=trees$Girth, y=trees$Height, pch=19, cex=1.5,
     col="steelblue", xlab="Omkrets", ylab="Höjd")
```



# Korrelationer i R

Vi kan räkna ut korrelationskoefficienten mellan två variabler med functionen *cor*.

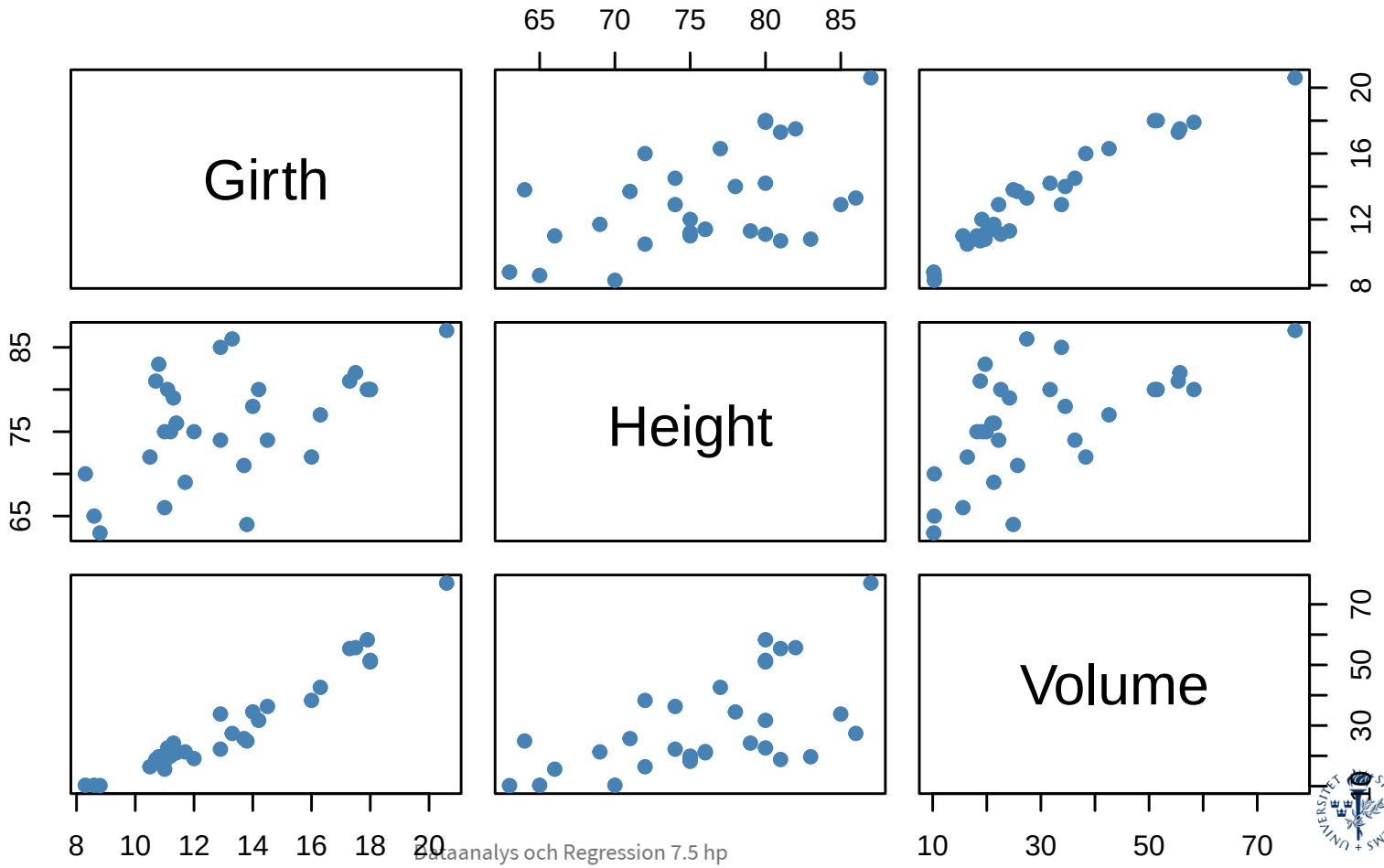
```
cor(trees$Girth, trees$Height) # Korrelation omkrets, höjd
```

```
[1] 0.5192801
```

# Samband mellan fler än två numeriska variabler

- Om vi har tre eller fler numeriska variabler kan vi göra parvisa spridningsdiagram i R med funktionen *pairs*.

```
pairs(trees, col="steelblue", pch=19, cex=1)
```



# Samband mellan fler än två numeriska variabler

För att studera parvis korrelation mellan flera variabler kan vi göra en korrelationstabell. Här ser vi tabell 6.1 i De Veaux et al. (2021). Det är en korrelationstabell för olika finansiella mått från Forbes:

| | Assets | Sales | Market Value | Profits | Cash Flow | Employees |
|---|---|---|---|---|---|---|
| **Assets** | 1.000 | | | | | |
| **Sales** | 0.746 | 1.000 | | | | |
| **Market Value** | 0.682 | 0.879 | 1.000 | | | |
| **Profits** | 0.602 | 0.814 | 0.968 | 1.000 | | |
| **Cash Flow** | 0.641 | 0.855 | 0.970 | 0.989 | 1.000 | |
| **Employees** | 0.594 | 0.924 | 0.818 | 0.762 | 0.787 | 1.000 |

- Varför är är alla värden längs diagonalen 1?
- Vad kan vi säga om den tomma delen av tabellen ovanför diagonalen?

# Samband mellan fler än två numeriska variabler

- I R kan vi använda funktionen *cor* även för att skapa en korrelationstabell. I stället för att ange två variabler som argument anger vi en hel dataframe som argument.
- **Viktigt!** För att det ska fungera måste alla variabler i vår dataframe vara numeriska.

```
cor(trees)
```

```
            Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

- Vad ser du om du jämför matrisens nedre vänstra del med dess övre högra del?