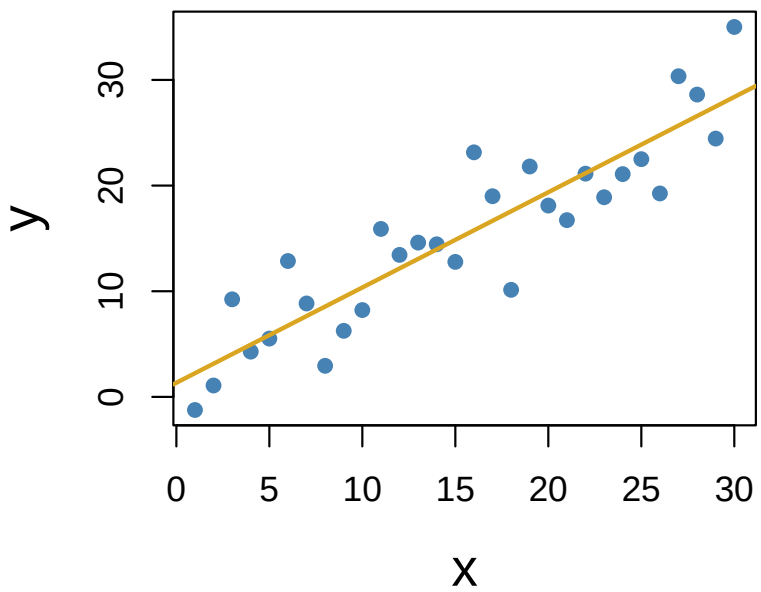
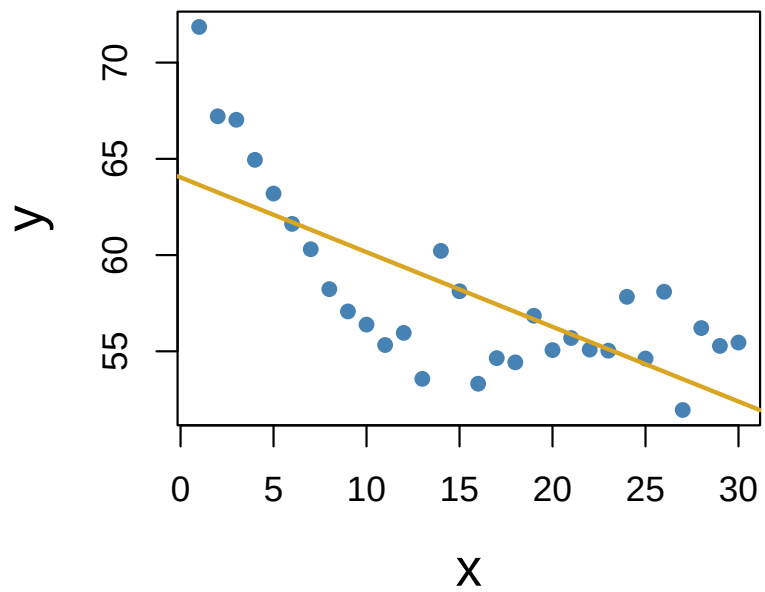
Lecture 8

karl.sigfrid@stat.su.se

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

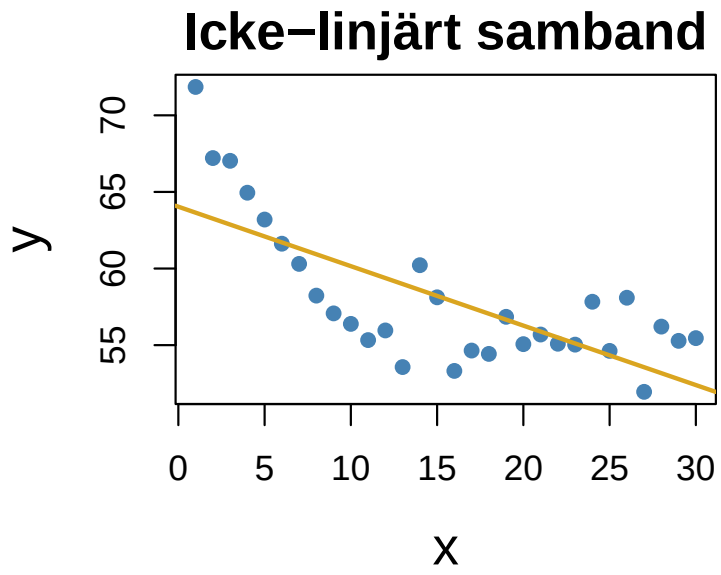
- Vi har redan introducerat enkel linjär regression. För att enkel linjär regression ska fungera väl bör sambandet mellan variablerna vara linjärt.
- Bilden till vänster visar ett samband som lämpar sig väl för linjär regression, och bilden till höger ett samband som inte kan fångas så väl med en rät linje.
- Om sambandet inte är linjärt kan vi ibland **göra** det linjärt genom en **transformation**.

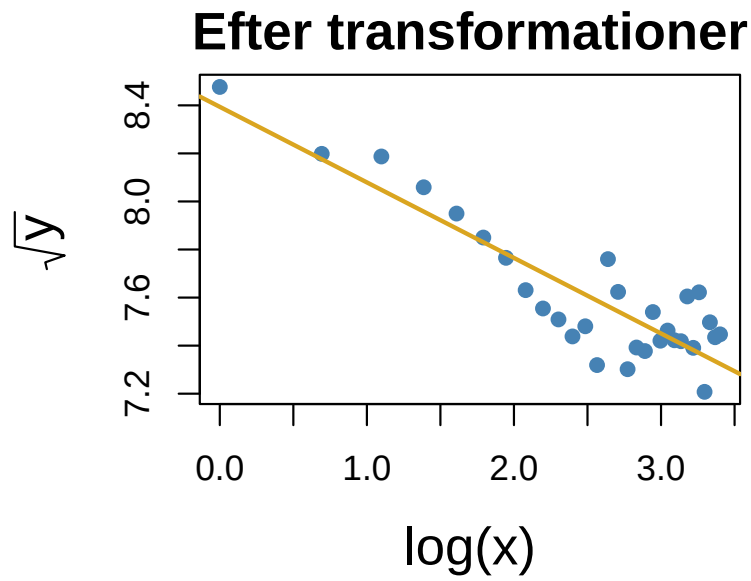




Transformationer

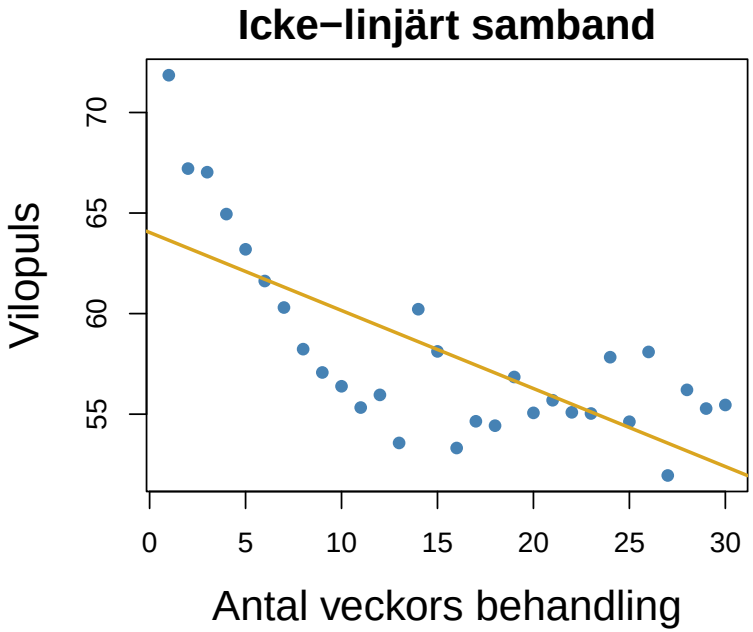
- I bilden till höger har vi transformerat våra variabler så att
  - $x \to \log(x)$
  - $y \to \sqrt{y}$
- Sambandet mellan de transformerade variablerna är någorlunda linjärt, så vi kan använda de transformerade variablerna i en simpel linjär regression!

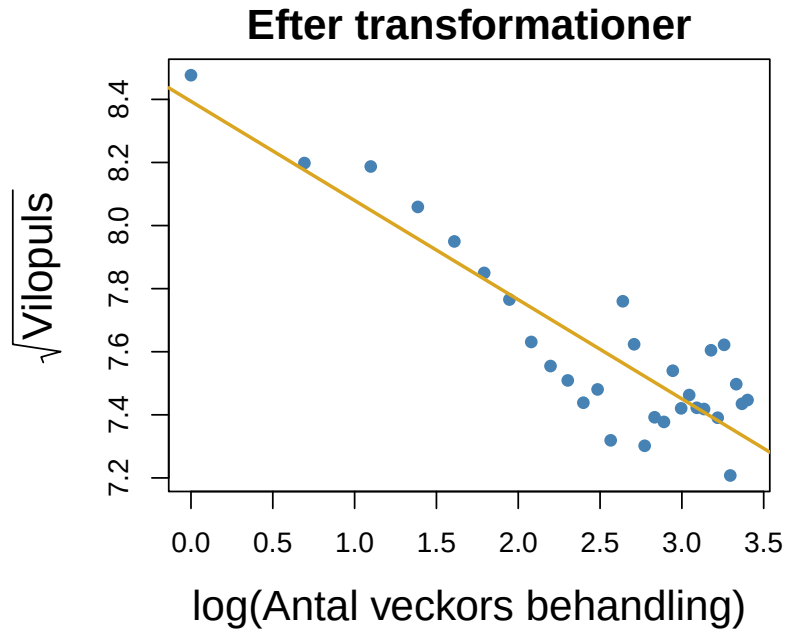




Transformationer - att transformera tillbaka

Låt oss säga att variabeln x i vårt exempel står för antalet veckor som ett antal sjukvårdspatienter har tagit ett visst läkemedel som sänker vilopulsen. Variabeln y är patienternas vilopuls mätt i hjärtslag per minut.





Stockholm University

Transformationer - att transformera tillbaka

Om vi använder våra transformerade variabler i en regressionsmodell blir modellen **inte**

$$\widehat{\text{vilopuls}} = b_0 + b_1 \cdot \text{behandlingstid}$$

Med de transformerade variablerna i vår regression blir modellen i stället

$$\widehat{\sqrt{\text{vilopuls}}} = b_0 + b_1 \cdot \log(\text{behandlingstid})$$

Genom att hitta värden för parametrarna b_0 och b_1, exempelvis med hjälp av *lm*-funktionen i R, ser vi att modellen blir

$$\widehat{\sqrt{\text{vilopuls}}} = 8.394 - 0.314 \cdot \log(\text{behandlingstid})$$

Transformationer - att transformera tillbaka

- När vi har funnit regressionslinen för våra transformerade variabler kan vi räkna ut estimatet $\widehat{\sqrt{\text{vilopuls}}}$ för ett visst värde av $\log(\text{behandlingstid})$.
- Men det är förmodligen **inte** vad vi är intresserade av!
- Vi vill kunna estimera variabeln vilopuls för ett visst värde av behandlingstid, dvs vill är intresserade av sambandet mellan våra **ursprungliga variabler**.
- För att estimera våra ursprungliga variabler med hjälp av modellen måste vi
  1. Transformera det värde av x som vi är intresserade av.
  2. Estimera det transformerade värdet av y-variabeln.
  3. **Transformera tillbaka** y-variabeln.

Transformationer - att transformera tillbaka

Följande visar hur vi kan estimera y i vårt exempel.

- Vår modell är

$$\widehat{\sqrt{\text{vilopuls}}} = 8.394 - 0.314 \cdot \log(\text{behandlingstid})$$

- Vi vill estimera vilopulsen för en patient som har tagit läkemedlet i fyra veckor, dvs behandlingstid $= 4$.
- Det betyder att $\log(\text{behandlingstid}) = \log(4) = 1.386$.
- Regressionsmodellen ger oss estimatet $\widehat{\sqrt{\text{vilopuls}}} = 8.394 - 0.314 \cdot 1.386 = 7.959$.

Transformationer - att transformera tillbaka

- Vi vet nu att $\widehat{\sqrt{\text{vilopuls}}} = 7.959$.
- Nu återstår att transformera $\widehat{\sqrt{\text{vilopuls}}}$ till $\widehat{\text{vilopuls}}$.
- Det gör vi med formeln $\widehat{\text{vilopuls}} = \left(\widehat{\sqrt{\text{vilopuls}}}\right)^2$.

$$\widehat{\text{vilopuls}} = \left(\widehat{\sqrt{\text{vilopuls}}}\right)^2 = 7.959^2 = 63.35$$

Slutsatsen blir att vi estimerar vilopulsen för en person som behandlats under fyra veckor till 63.35 slag per minut.

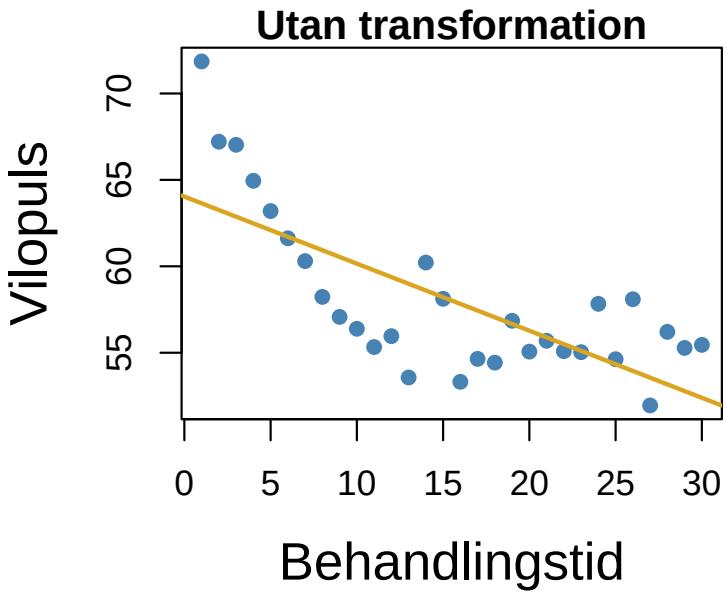
Transformationer - att transformera tillbaka

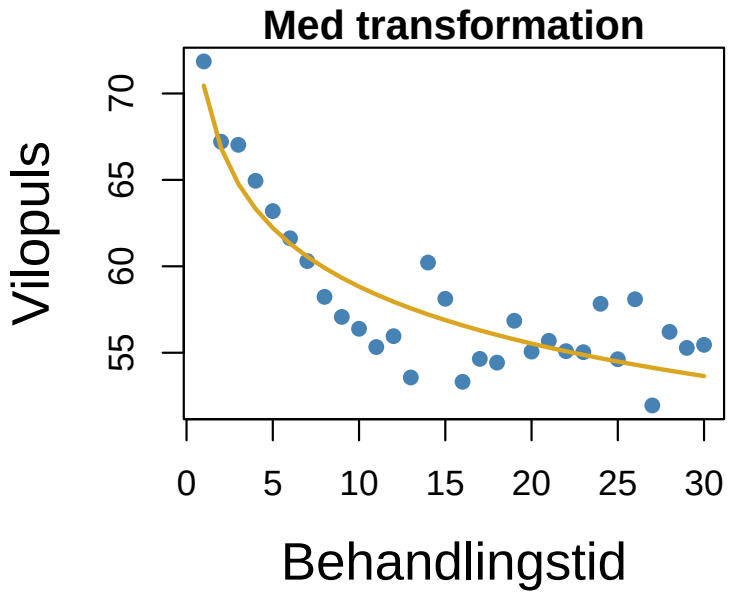
Den här tabellen visar hur vi transformerar tillbaka responsvariabeln y beroende på vilken transformation som vi använder för responsvariabeln i regressionsmodellen.

Transformation_av_responsvariabeln	Transformera_tillbaka
$y \to y^2$	$\hat{y} = \sqrt{\hat{y}^2}$
$y \to y$	$\hat{y} = \hat{y}$
$y \to \sqrt{y}$	$\hat{y} = \widehat{\left(\sqrt{y}\right)}^2$
$y \to \log(y)$	$\hat{y} = e^{\widehat{\log(y)}}$

Transformationer - att transformera tillbaka

- Om vi estimerar variabeln vilopuls för ett stort antal olika värden av behandlingstid kan vi dra en linje mellan våra estimat. Linjen illustreras i bilden till höger. Den fångar mönstret i observationerna mycket bättre än regressionslinjen till vänster, som vi får när vi inte transformerar.
- Genom transformationer har vi lyckats fånga ett icke-linjärt samband med hjälp av enkel linjär regression!



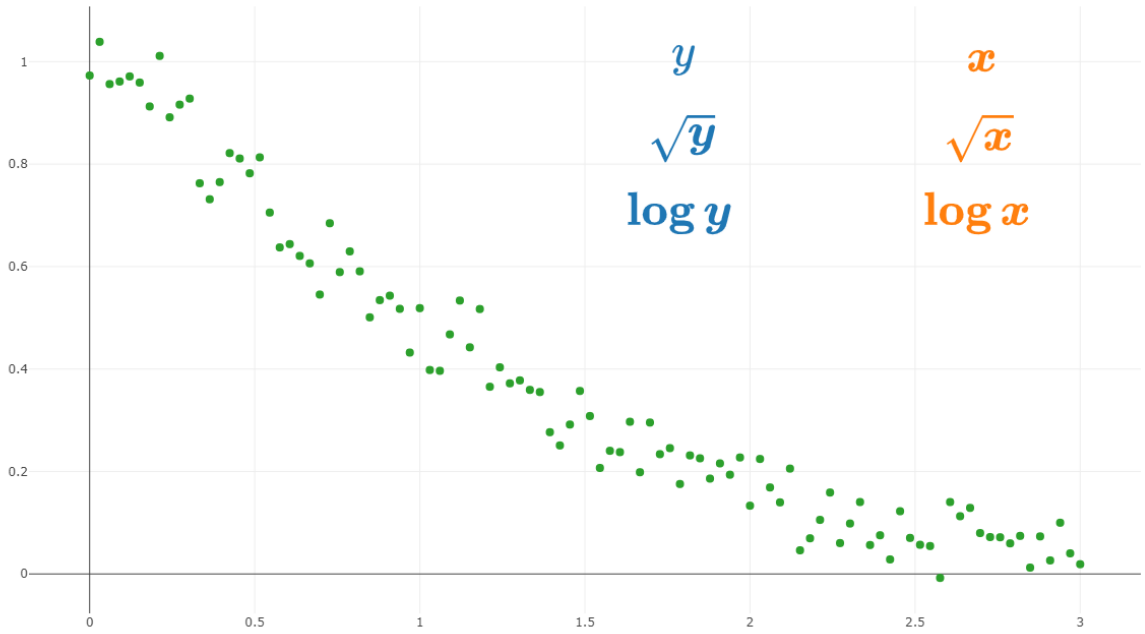


Transformationer - att välja transformationer

- I exemplet som vi just gick igenom transformerade vi y till $\sqrt{y}$ och x till $\log(x)$. Varför just dessa transformationer?
- Till viss del handlar det om att **pröva sig fram** tills vi hittar transformationer som gör linjen någorlunda rät.
- Vi behöver dock inte pröva i blindo. För en viss form på den urspungliga kurvan finns det några olika transformationer som vi rekommenderas att pröva.

Transformationer - att välja transformationer

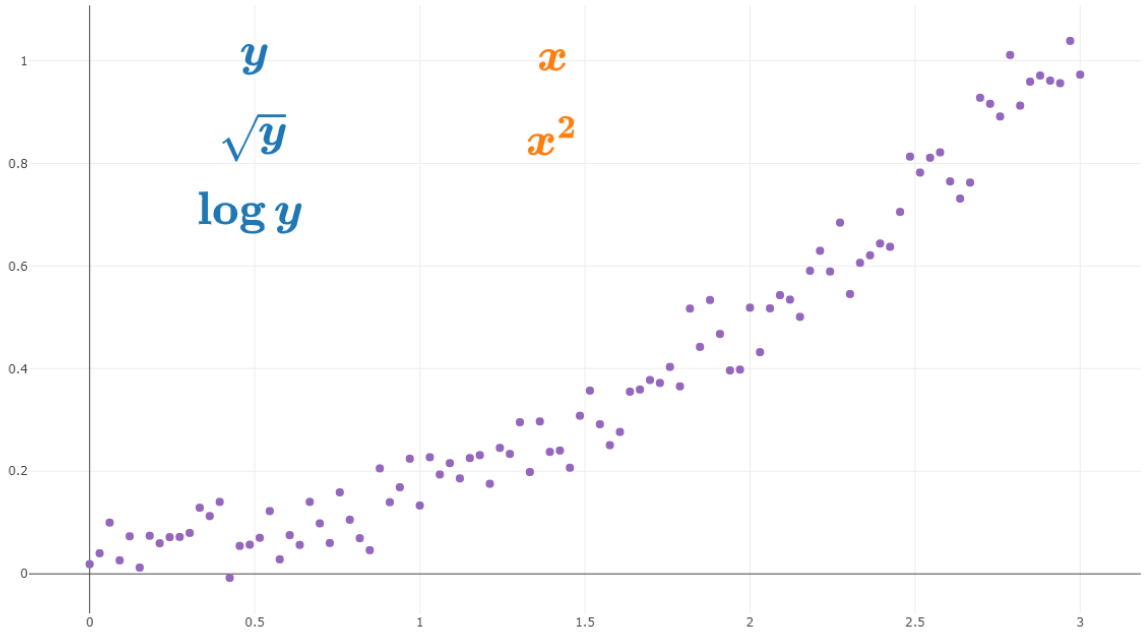
- Om vår graf ser ut som på den här bilden kan vi pröva
  - Att transformera y till $\sqrt{y}$ eller till $\log(y)$
  - Att transformera x till $\sqrt{x}$ eller till $\log(x)$



Stockholm University

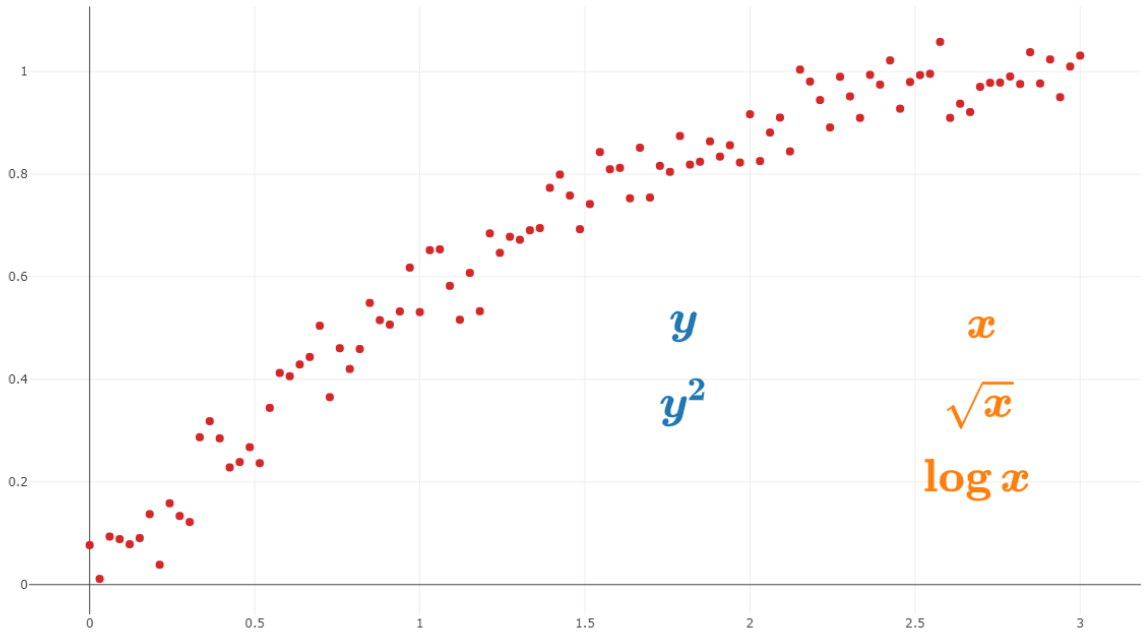
Transformationer - att välja transformationer

- Om vår graf ser ut som på den här bilden kan vi pröva
  - Att transformera y till $\sqrt{y}$ eller till $\log(y)$
  - Att transformera x till x^2



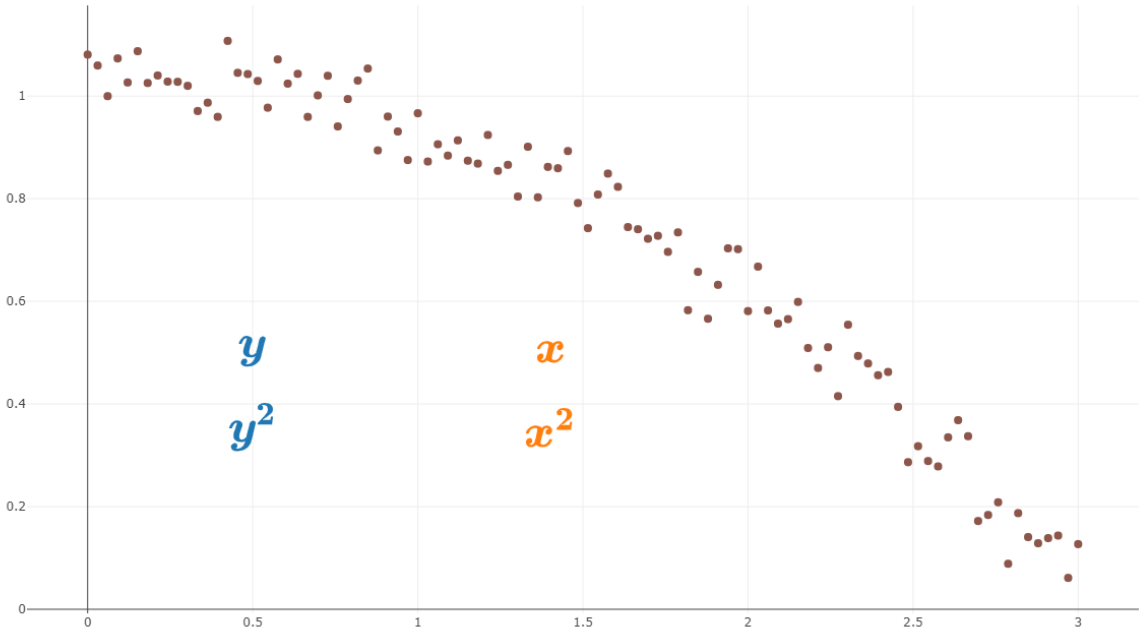
Transformationer - att välja transformationer

- Om vår graf ser ut som på den här bilden kan vi pröva
  - Att transformera y till y^2
  - Att transformera x till $\sqrt{x}$ eller till $\log(x)$



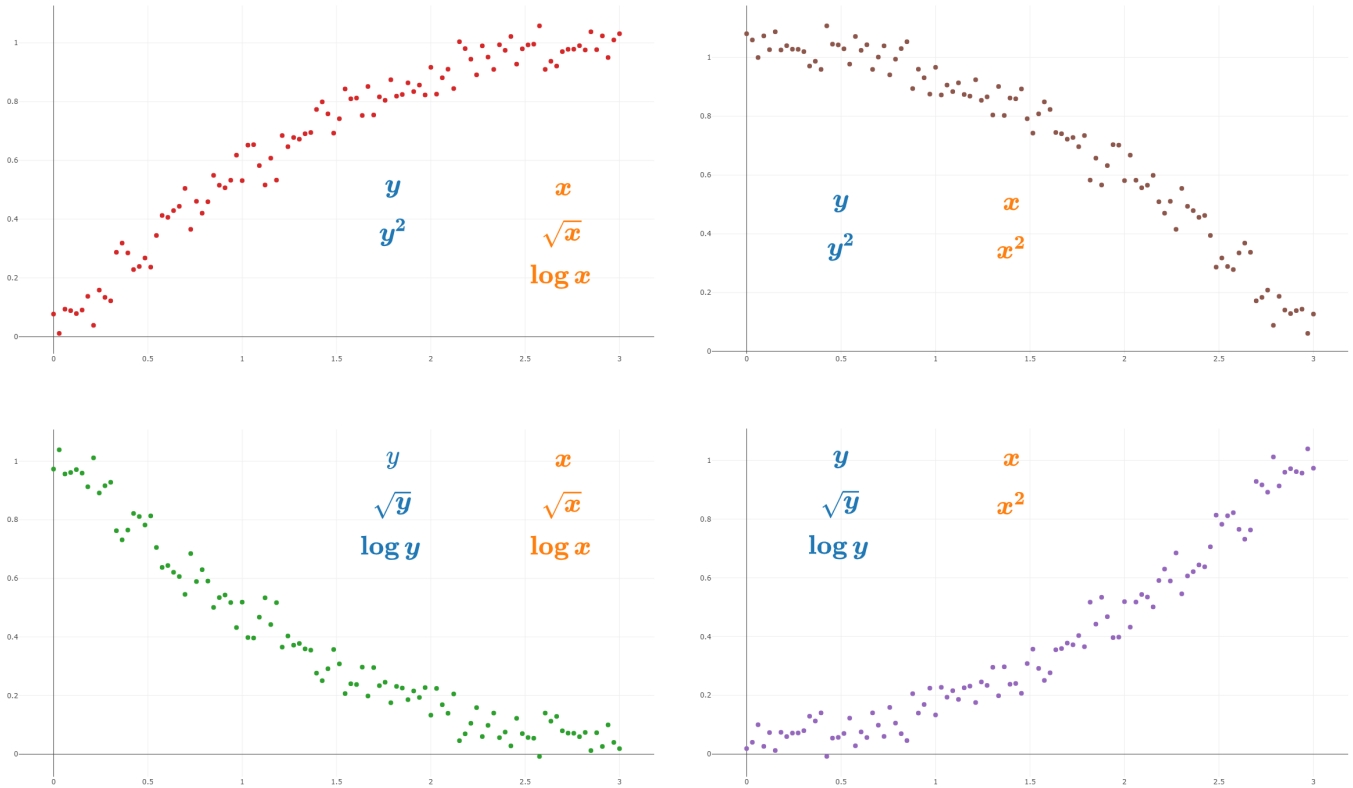
Transformationer - att välja transformationer

- Om vår graf ser ut som på den här bilden kan vi pröva
  - Att transformera y till y^2
  - Att transformera x till x^2



Transformationer - att välja transformationer

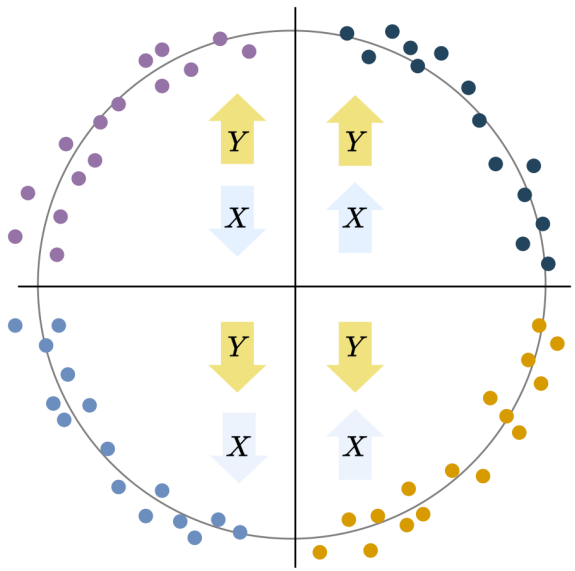
Om vi lägger graferna intill varandra bildar de en cirkel.



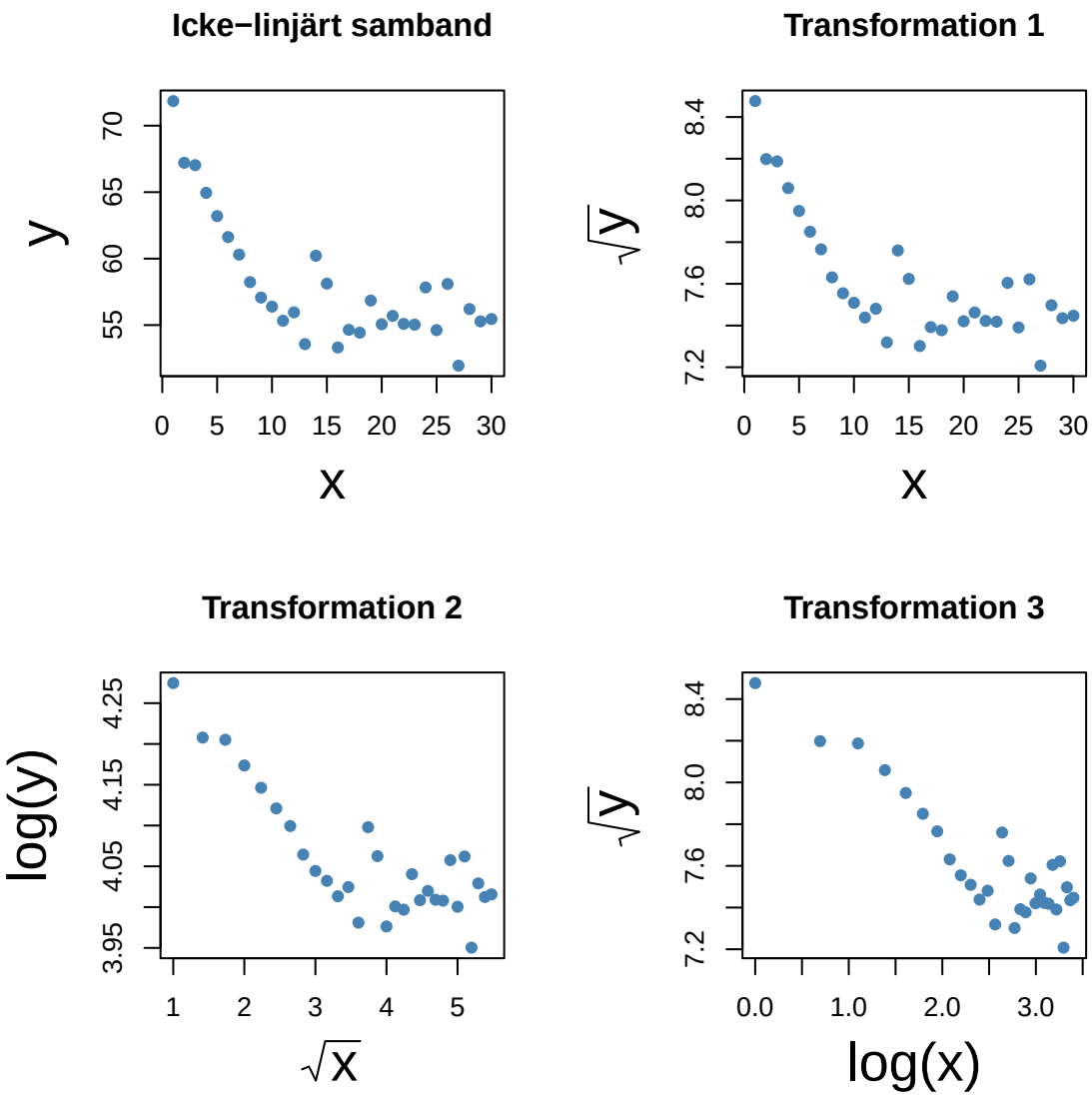
Transformationer - att välja transformationer

- Tukeys cirkel beskriver i kompakt form transformeringar som kan prövas.
- Varje fjärdedel av cirkeln representerar ett visst icke-linjärt mönster.
- Pilarna uppåt och nedåt indikerar vilka transformationer att pröva för ett visst mönster.

Potens	y	x
2	y^2	x^2
1	y	x
$\frac{1}{2}$	$\sqrt{y}$	$\sqrt{x}$
"0"	$\log y$	$\log x$
$-\frac{1}{2}$	$-\frac{1}{\sqrt{y}}$	$-\frac{1}{\sqrt{x}}$
-1	$-\frac{1}{y}$	$-\frac{1}{x}$



Transformationer - att välja transformationer



- Graferna till vänster visar några exempel på transformationer som vi hade kunnat pröva för vårt data om ett läkemedels effekt på vilopulsen.
- Den sista transformationen ger den mest räta linjen. Därför valde vi den.

Multipel linjär regression

- Hittills har våra modeller enbart innehållit **en** förklaringsvariabel x.

$$\hat{y} = b_0 + b_1 x$$

- Ofta har vi **flera** förklaringsvariabler som tillsammans förklarar responsvariabeln bättre än vad de kan göra var för sig.
- När vi använder mer än en förklaringsvariabel i en regression har vi en **multipel linjär regression**.
- En multipel linjär regressionsmodell med två förklaringsvariabler har formen

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2,$$

där x_1 och x_2 är våra förklaringsvariabler.

Multipel linjär regression

- I vårt exempel med bilar och bensinförbrukning har vi hittills förklarat bensinförbruningen med bilens vikt.
- Vi har dock fler variabler i vårt dataset, bland annat variabeln motorstyka, mätt i antalet hästkrafter.
- Det visar sig att modellen blir bättre om vi använder motorstyrkan som en ytterligare förklaringsvariabel i modellen.
- Vår nya modell blir

$$\widehat{\text{litermil}} = b_0 + b_1 \cdot \text{vikt} + b_2 \cdot \text{hästkrafter}$$

Multipel linjär regression

På samma sätt som tidigare kan vi använda R för att hitta modellens parameterar, dvs b_0, b_1 och b_2.

```
data(mtcars)
mtcars$litermil <-1 / (mtcars$mpg * 0.16 / 3.785)
mtcars$viktton <- mtcars$wt * 0.454
lmod_cars <- lm(litermil ~ viktton + hp, data=mtcars)
lmod_cars$coefficients
```

```
(Intercept)     viktton          hp
0.149155845 0.598453723 0.001769319
```

Vi ser att om vi avrundar blir modellen

$$\widehat{\text{litermil}} = 0.149 + 0.598 \cdot \text{vikt} + 0.00177 \cdot \text{hästkrafter}$$

Multipel linjär regression

- För enkel linjär regression, det vill säga linjär regression med bara en variabel, tittade vi på formler för att räkna ut b_0 och b_1.
- När vi gör multipel linjär regression är blir formlerna betydligt mer komplicerade. Vi kommer inte att gå igenom dem under den här kursen.
- Dock gäller fortfarande principen att vi väljer de värden på våra parametrar som minimerar

$$\sum_{i=1}^{n} e_i^2, \qquad e_i^2 = (y_i - \hat{y})^2$$

Vi minimerar alltså summan av de kvadrerade residualerna.

Multipel linjär regression

- Det finns ingen gräns för hur många förklaringsvariabler vi kan använda i en regressionsmodell.
- Ett generellt sätt att formulera modellen är

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k,$$

Vi säger i det här fallet att vi har k förklaringsvariabler, och k kan vara vilket antal som helst.

Multipel linjär regression - val av förklaringsvariabler

- Vi hävdade att vi kan få en bättre modell för en bils bensinförbrukning genom att använda båda variablerna *vikt* och *häftkrafter*.
- Hur kan vi se att det stämmer?
- Ett enkelt sätt är att jämföra R^2 för de båda modellerna. Den modell som har högre R^2 förklarar mer av variationen i responsvariabeln.
- På förra föreläsningen såg vi att R^2 blev 0.7919 när vi använde *vikt* som enda förklaringsvariabel.
- Nu ska vi se vad R^2 blir när vi lägger till förklaringsvariabeln *hästkrafter*.

Multipel linjär regression - val av förklaringsvariabler

Vi ser att $R^2 = 0.8471$, så modellen med två variabler ser ut att vara bättre.

```
summary(lmod_cars)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1491558  0.0968400   1.540  0.13435
viktton     0.5984537  0.0844168   7.089 8.45e-08 ***
hp          0.0017693  0.0005469   3.235  0.00303 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1571 on 29 degrees of freedom
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8365
F-statistic: 80.33 on 2 and 29 DF,  p-value: 1.494e-12
```

Multipel linjär regression - val av förklaringsvariabler

- Faktum är att R^2 **alltid** blir större när vi lägger till ytterligare variabler.
- Detta gäller även om det saknas samband mellan den tillagda förklaringsvariabeln och responsvariabeln.
- Att R^2 blir större när vi lägger till en till variabel betyder alltså inte med automatik att modellen blir bättre.
- En strategi för att avgöra vilka variabler som ska inkluderas kan vara att maximera det mått som kallas **adjusted R-squared**. Detta mått ökar med R^2, men minskar samtidigt med antalet förklaringsvariabler.

Multipel linjär regression - val av förklaringsvariabler

- Adjusted R-squared räknas ut

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(n-1)}{n-k-1},$$

där n är antalet observationer i vårt dataset och k antalet förklaringsvariabler i modellen.

Multipel linjär regression - val av förklaringsvariabler

- Låt oss jämföra **adjusted R-quared** för våra två modeller:
  - Den ursprungliga med bara förklaringsvariabel, *vikt*.
  - Den nya, som inkuderar även förklaringsvariabln *hästkrafter (hp)*.
- Vi börjar med att räkna ut adjusted R-squared med bara en förklaringsvariabel.

```
lmod_cars1 <- lm(litermil ~ viktton, data=mtcars)
summary_1 <- summary(lmod_cars1)
print(summary_1$adj.r.squared)
```

```
[1] 0.7849726
```

Multipel linjär regression - val av förklaringsvariabler

- Med enbart förklaringsvariabeln *vikt* blev adjusted R-squared ungefär 0.785.
- Nu undersöker vi vad vi får för resultat när vi lägger till variabeln *hästkrafter*.

```
lmod_cars2 <- lm(litermil ~ viktton + hp, data=mtcars)
summary_2 <- summary(lmod_cars2)
print(summary_2$adj.r.squared)
```

```
[1] 0.8365429
```

- Vi ser att adjusted R-squared tydligt ökade när vi inkluderade även förklaringsvariabeln *hästkrafter*, vilket är ett argument för att den större modellen är bättre.

Multipel linjär regression - tolkning av parametrarna

- Tolkningen av parametrarna är nästan densamma som när vi hade en enda förklaringsvariabel. Då tolkade vi b_1 på följande sätt:
  - När värdet på x ökar med en enhet så ökar $\hat{y}$ med b_1 enheter.
- Med två eller fler förklaringsvariabler tolkar vi våra lutningsparametrar på följande sätt:
  - När värdet på x_k ökar med en enhet ökar $\hat{y}$ med b_k enheter, **givet att värdet på övriga förklaringsvariabler hålls konstant**.
  - Kom ihåg att om b_k är negativt minskar istället $\hat{y}$ när b_k ökar.

Multipel linjär regression - tolkning av parametrarna

- Vi har följande parametrar i vår multipla regressionsmodell.

```
(Intercept)     viktton          hp
0.149155845 0.598453723 0.001769319
```

- **Räkneexempel 1:** Anta att vi har två bilar som väger lika mycket. Den andra bilen har 1 hästkraft mer än den första bilen. Vi estimerar då att den andra bilen förbrukar 0.00177 liter mer per mil.
- **Räkneexempel 2:** Anta nu att vi har två bilar med samma antal hästkrafter. Den andra väger 0.2 ton *mindre* än den första. Modellen estimerar då att den andra bilen förbrukar $0.2 \cdot 0.598 = 0.1196$ liter per mil *mindre* än den första bilen.

Multipel linjär regression - tolkning av parametrarna

- Låt oss titta på ett nytt dataset, som vi kan använda för att undersöka sambandet mellan sparande, inkomst och boendekostnad.
- Vi är intresserade av att estimera hur mycket individerna i vårt dataset sparar årligen givet vilken inkomst och vilken boendekostnad de har.
- Vi börjar med att göra en linjär regressionsmodell med boendekostnad som enda förklaringsvariabel.

$$\widehat{\text{sparande}} = b_0 + b_1 \cdot \text{boendekostnad}$$

Multipel linjär regression - tolkning av parametrarna

- Vi räknar ut modellens parametrar i R.

$$\widehat{\text{sparande}} = b_0 + b_1 \cdot \text{boendekostnad}$$

```
lm(sparande ~ boendekostnad)$coefficients
```

```
  (Intercept) boendekostnad
 8206.3659478     0.3750134
```

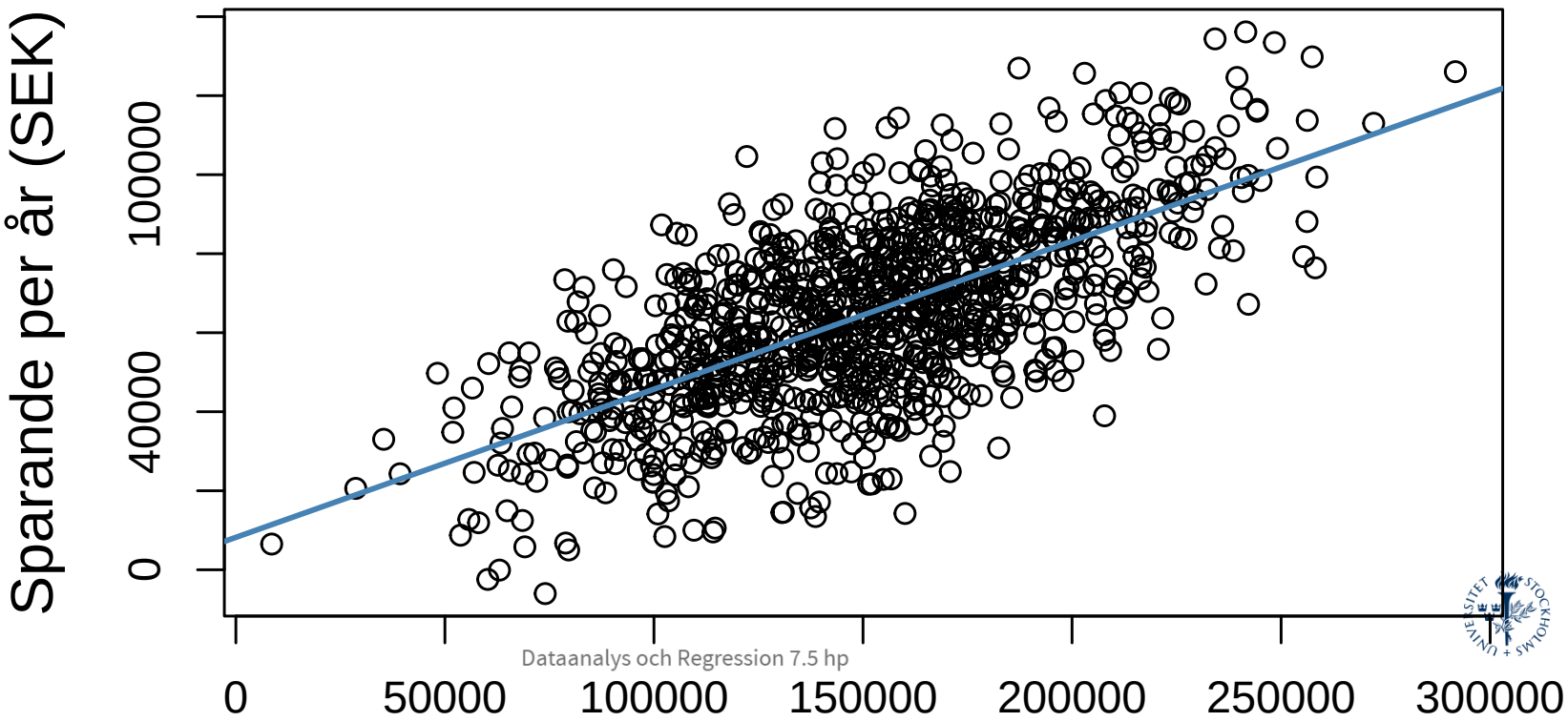
- Om vi avrundar blir vår regressionsmodell

$$\widehat{\text{sparande}} = 8,206 + 0.375 \cdot \text{boendekostnad}$$

Multipel linjär regression - tolkning av parametrarna

$$\widehat{\text{sparande}} = 8,206 + 0.375 \cdot \text{boendekostnad}$$

- För varje ytterligare krona en person lägger på boendet per år estimerar vår modell att personen sparar ytterligare 0.375 kronor per år.
- Om person A har en boendekostnad som är 10,000 kronor högre än person B, då estimerar vi alltså att person A sparar 3,750 kronor mer per år.



Multipel linjär regression - tolkning av parametrarna

Nu lägger vi till variabeln *inkomst* och får en ny modell..

```
lm(sparande ~ boendekostnad + inkomst)$coefficients
```

```
  (Intercept) boendekostnad       inkomst
-71.74833135   -0.08899729    0.19553001
```

$$\widehat{\text{sparande}} = -71.7 - 0.089 \cdot \text{boendekostnad} + 0.196 \cdot \text{inkomst}$$

- I den här modellen *minskar* sparandet med boendekostnaden. För varje ytterligare krona i boendekostnad estimerar vi att sparandet minskar med 0.089 kronor.
- När vi bara hade *boendekostnad* som förklaringsvariabel estimerade modellen att sparandet *ökade* med boendekostnaden.
- Betyder det att en av modellerna **har fel**? Vilken av dem i så fall?

Multipel linjär regression - tolkning av parametrarna

- Nej, att våra modeller kan se ut att motsäga varandra betyder **inte** att någon av dem är fel.
- De båda modellerna ger oss olika information.
- Den första modellen säger att om vi slumpvis väljer ut två individer estimerar vi att den som har högst boendekostnad också är den som har högst sparande.

$$\widehat{\text{sparande}} = 8,206 + 0.375 \cdot \text{boendekostnad}$$

Multipel linjär regression - tolkning av parametrarna

- Den andra modellen säger att **givet värdet på de övriga variablerna** kan vi estimera att den som har högre boendekostnad har ett lägre sparande.

$$\widehat{\text{sparande}} = -71.7 - 0.089 \cdot \text{boendekostnad} + 0.196 \cdot \text{inkomst}$$

- Den övriga variabeln i modellen är **inkomst**. Modellen säger alltså att **givet en viss inkomst** estimerar vi att den som har lägre boendekostnad har ett större sparande.

Multipel linjär regression - tolkning av parametrarna

Om vi tänker på det i mer vardagliga termer är resultaten av båda våra regressionsanalyser logiska. De beskriver två olika scenarier.

Regression 1:

- Anta att vi samlar ihop en grupp helt slumpvis utvalda personer.
- Om vi jämför de personer som har dyrare bostäder med de personer som har billigare bostäder så har de med dyrare bostäder ofta högre inkomster. Det betyder att de också kan spara mer, trots att de har dyrare boende.

Regression 2:

- Vi samlar ihop en grupp personer där alla i gruppen har samma inkomst.
- **Inom den här gruppen** jämför vi dem som har dyrare bostäder med dem som har billigare bostäder. De som spenderar mindre på sitt boende har mer pengar över att spara, och därför är deras sparande högre.

Multipel linjär regression - tolkning av parametrarna

- I multipel linjär regression kan vi säga att sambandet mellan responsvariabel och en förklaringsvariabel **är betingat** på de övriga förklaringsvariablerna.
- Jämför med betingade proportioner från föreläsning 3, då vi exempelvis undersökte andelen som drabbades av prostatacancer *givet* att de sällan eller aldrig åt fisk.
- Det här understyker än en gång att de samband som vi hittar genom regressionsanalyser **inte** behöver betyda att vi har kausalitet.
- Vår första regression visar exempelvis inte att högre boendekostnader leder till större sparande, utan snarare att de som har högre boendekostnader kan spara mer **trots** de högre boendekostnaderna.

Multipel linjär regression - tolkning av parametrarna

- Låt oss titta på ett exempel till.
- Regressionsmodell 1 estimerar huspriser i dollar, med antalet sovrum som enda förklaringsvariabel.

$$\widehat{\text{price}} = 338,975 + 40,234 \cdot \text{bedroom}$$

- Regressionsmodell 2 inkluderar även boytan i kvadratfot (ft^2).

$$\widehat{\text{price}} = 308,100 + 135 \cdot \text{living area} - 43,347 \cdot \text{bedroom}$$

- Koefficienten för antalet sovrum har gått från positiv till negativ.
  - Vilka möjliga förklaringar kan man tänka sig?
  - Hur ska de två olika modellerna tolkas?

Multipel linjär regression - tolkning av parametrarna

- Enligt den första regressionsmodellen ökar priset på ett hus med antalet sovrum. Det verkar rimligt eftersom ett hus med fler sovrum ofta är större.
- Enligt den andra modellen *minskar* priset på ett hus med antalet sovrum, **givet husets storlek**.
- Om två hus är lika stora estimerar vi alltså att det hus som har färre sovrum är dyrare. Det skulle kunna bero på att huset med färre sovrum har större kök, vardagsrum, etc.

Multipel linjär regression - Fler än 2 förklaringsvariabler

- Med fler än två förklaringsvariabler anger modellen sambandet mellan responsvariabeln och förklaringsvariabeln **givet alla de övriga förklaringsvariablerna**.
- Låt oss definiera en modell som förklararen bils bränsleförbrukning med (1) bilens vikt i ton, (2) antalet hästkrafter (hp) och (3) antalet växlar (gear).

```
(Intercept)      viktton           hp         gear
0.352356646  0.540075281  0.001964705 -0.039753798
```

- Vi ser att coefficienten för antalet växlar är ungefär -0.04.
- Om vi har två bilar med samma vikt **och** samma antal hästkrafter, och bil 2 har en växel mer än bil 1, då estimerar vi att bil 2 drar 0.04 liter bensin *mindre* per mil.

Multipel linjär regression - förutsättningar

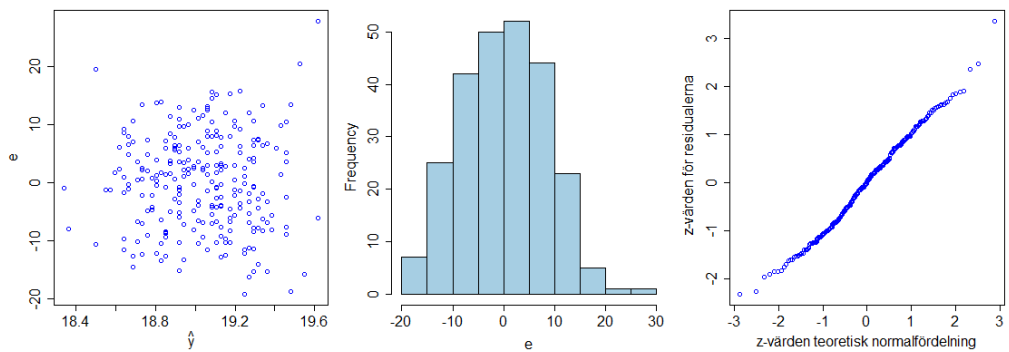
- y och alla $x_1, x_2, ..., x_k$ måste vara **numeriska variabler**.
- y måste förhålla sig någorlunda **linjärt** till vardera förklaringsvariabel.
- Det bör inte finnas uppenbara **outliers**, eftersom enstaka outliers kan ha stor påverkan på modellens parametrar.
- **Residualernas varians** bör vara konstant.
- Residualerna bör vara **normalfördelade**.
- Residualernas standardavvikelse räknas ut

$$s_e = \sqrt{\frac{\sum e^2}{n-k-1}},$$

där n är antalet observationer och k antalet förklaringsvariabler.

Multipel linjär regression - förutsättningar

- Dessa tre grafer tyder på att förutsättningarna är uppfyllda för en modell.
  - Spridningsdiagrammet till vänster visar att residualerna på y-axeln inte tycks förändras med våra predikterade värden på x-axeln.
  - Histogrammet i mitten visar att residualerna är approximativt normalfördelade.
  - Normalfördelningsgraften visar även den att residualerna är approximativt normalfördelade.



Multipel linjär regression - förutsättningar

- Ställ inte orimligt höga krav när du bedömer om regressionsmodellens antaganden är uppfyllda.
- Så länge en modell ger prediktioner som är bättre än $\bar{y}$ är modellen användbar i rätt sammanhang.
- En statistisk modell behöver inte vara perfekt. Det viktiga är att modellens brister **kommuniceras öppet!** Då vet den du kommunicerar med att modellens resultat ska tolkas med försiktighet.