Statistik och dataanalys I

F8: Transformationer och multipel linjär regression

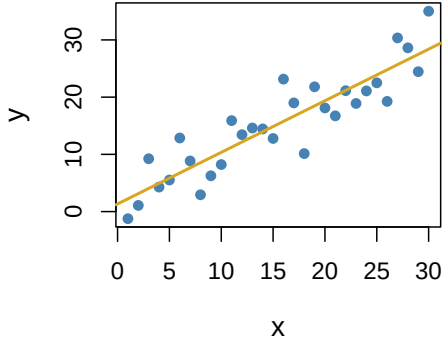
Valentin Zulj

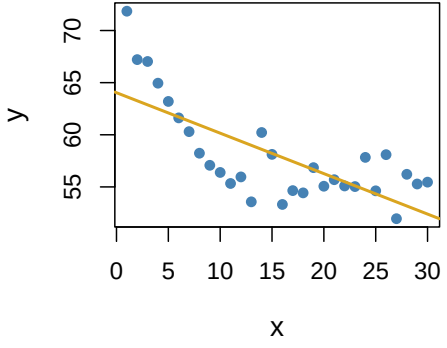
Statistiska institutionen

Stockholm
University

Vad har vi gjort hittills?

- Vi har introducerat enkel linjär regression, och beskrivit hur vi kan skatta, tolka, och använda en linjär regressionsmodell
- Vi har tittat på olika sätt att utvärdera en modell, och kontrollera modellantaganden (m.h.a. t.ex. R^2 och residualanalys)
- Vi har även sagt att linjär regression lämpar sig för linjära samband (bild till vänster), men inte för icke-linjära samband (bild till höger)





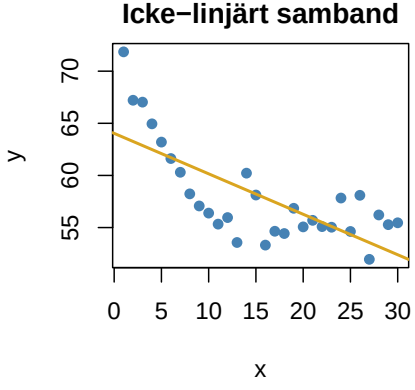
Vad vi ska göra nu?

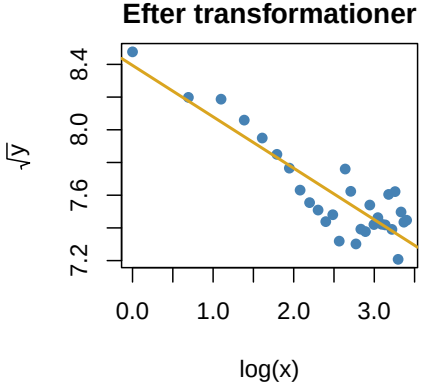
- Antag nu att vi vill analysera ett samband som **inte** är linjärt när variablerna mäts på ursprunglig skala
- Ett sätt angripa detta problem är att försöka **transformera** variabler för att **göra** sambandet linjärt
- Vi har tidigare pratat om t.ex. **log-transformationer** tidigare, och vi ska nu bredda diskussionen om transformationer en aning
- Vi ska även utvidga diskussion av linjär regression, och komplettera den **enkla** regressionsmodellen med en **multipel** regressionsmodell
- Vi använder multipel linjär regression för att studera samband mellan en responsvariabel, y, och **flera** förklarande variabler, $x_1, \ldots x_p$

Transformationer

Transformation till linjäritet

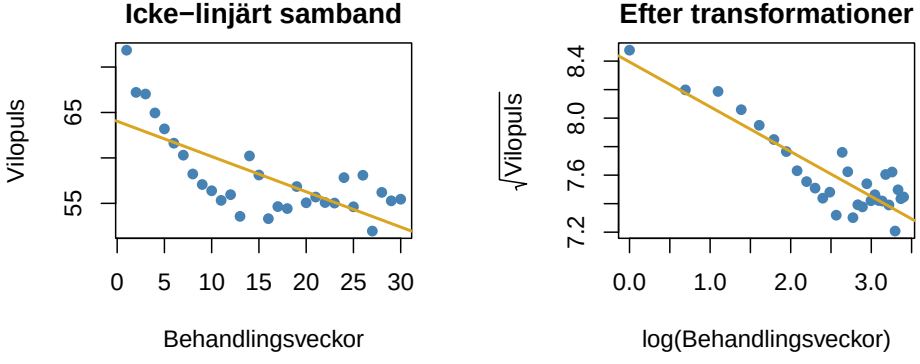
- Vi betraktar det icke-linjära sambandet och försöker transformera det till linjäritet
- Transformationerna $x \to \log(x)$ och $y \to \sqrt{y}$ ger oss sambandet som finns i den högra bilden





- Den räta linjen ser ut att passa den transformerade punksvärmen mycket bättre!

- Låt oss säga att
  - x är antalet veckor som en patient haft ett läkemedel för sänkt vilopuls
  - y står för samma patients uppmätta vilopuls (slag/minut)
- I figurerna från föregående slide får vi då följande axlar



Regression med transformerade variabler

- När vi använder de transformerade variablerna blir modellen **inte**

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

- När vi använder de transformerade variablerna får vi i stället

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

- Vi kan hitta värden för koefficienterna b_0 och b_1, t.ex. med `lm()`-funktionen i R, och få ut

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

Transformera tillbaka till originalskala

- Vi kan nu hitta skattningen $\widehat{\sqrt{\text{vilopuls}}}$ för ett givet värde på log(behandlingstid)
- **Men:** vi är troligtvis inte så intresserade av just $\sqrt{\text{vilopuls}}$ (vad betyder det ens?), utan vårt mål borde rimligtvis vara att skatta just vilopuls
- Vi är alltså intresserade av vår **ursprungliga variabel**
- För att skatta vilopulsen på ursprunglig skala behöver vi
  1. Transformera det värde av x som vi är intresserade av till $\log(x)$
  2. Skatta det transformerade värdet av y-variabeln, dvs $\sqrt{y}$ med vår modell
  3. **Transformera tillbaka** skattningen till den ursprungliga skalan

Transformera tillbaka till originalskala

- Vi har sedan tidigare att

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

- Antag nu att vi vill skatta vilopulsen för en patient som använt läkemedlet i fyra veckor
- Vi har alltså en patient med behandlingstid $= 4$, eller

$$\log(\text{behandlingstid}) = \log(4) = 1.386$$

- Regressionsmodellen ger oss nu skattningen

$$\widehat{\sqrt{\text{vilopuls}}} = 8.394 - 0.314 \cdot 1.386 = 7.959$$

Transformera tillbaka till originalskala

- Vi har så använt modellen för att få fram

$$\widehat{\sqrt{\text{vilopuls}}} = 7.959$$

- Det som återstår nu är att transformera skattningen tillbaka till ursprungsskalan, dvs att göra "avtransformationen"

$$\widehat{\sqrt{\text{vilopuls}}} \rightarrow \widehat{\text{vilopuls}}$$

- För att bli kvitt kvadratroten kan vi kvadrera skattningen, och få

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

- Vi får så dra slutsatsen att vilopulsen för en person som behandlats under fyra veckor skattas till ungefär 63 slag per minut

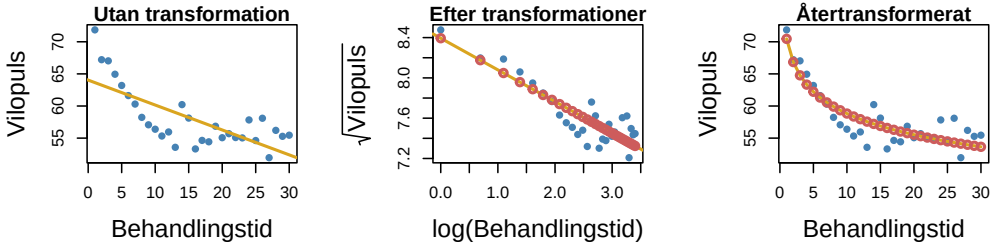
Transformera tillbaka till originalskala

- Tabellen nedan visar hur vi transformera tillbaka till originalskala efter några vanliga transformationer

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

Jämförelse med modell utan transformation

- Om vi skattar vilopulsen för många olika behandlingstider kan vi dra en linje mellan våra skattningar
- Linjerna som tagits fram med hjälp av transformation fångar mönstret mycket bättre än linjen som tagits fram utan transformation
- Med hjälp av transformationer har vi så lyckats fånga ett icke-linjärt samband med hjälp av enkel linjär regression!



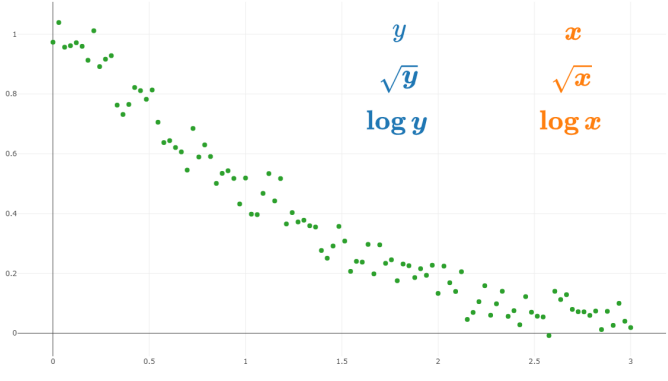
Tukeys cirkel – att välja transformation

Val av transformation

- I vårt exempel transformerade vi y till $\sqrt{y}$ och x till $\log(x)$ utan närmare eftertanke
- **Men:** Hur ska vi veta vilken typ av transformationer vi bör använda i andra fall? . . .
- Till viss del handlar det om att **pröva sig fram**, och testa olika transformationer tills någon ger ett relativt linjärt samband
- Vi behöver dock inte pröva i blindo, utan vissa specifika transformationer har visat sig passa specifika typer av samband
- Den matematiskt bevandrade kanske vet att logaritmen linjäriserar exponentialfunktionen, till exempel
- **Tukeys cirkel** hjälper oss att identifiera transformationer som kan vara vettiga, baserat på mönstret i vårt spridningsdiagram

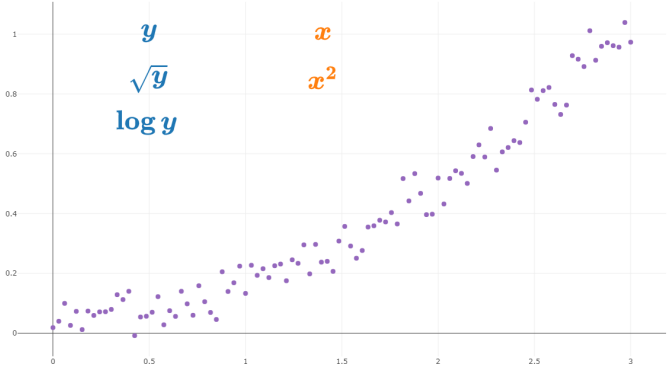
Val av transformation

- Om vårt spridningsdiagram ser ut som i bilden nedan kan vi testa att
  - Transformera y till $\sqrt{y}$ eller till $\log(y)$
  - Transformera x till $\sqrt{x}$ eller till $\log(x)$



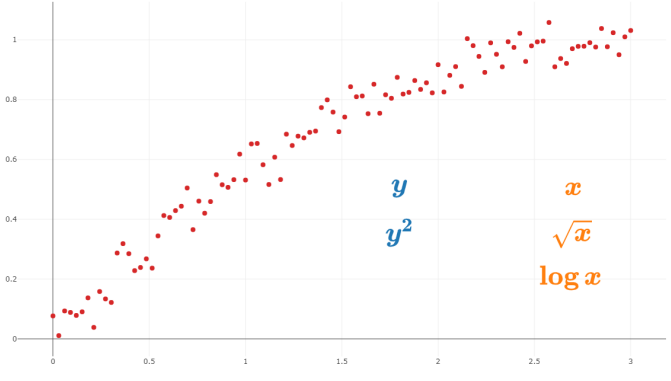
Val av transformation

- Om vårt spridningsdiagram ser ut som i bilden nedan kan vi testa att
  - Transformera y till $\sqrt{y}$ eller till $\log(y)$
  - Transformera x till x^2

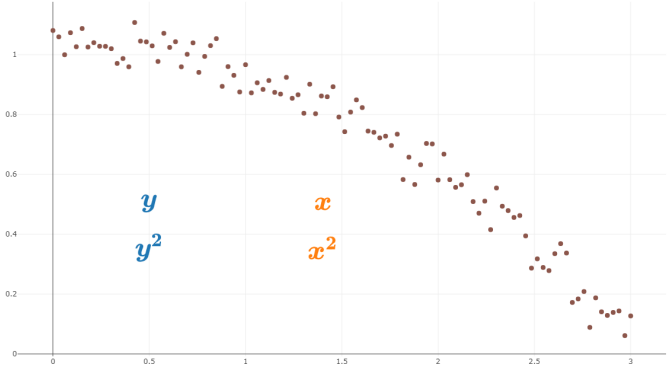


Val av transformation

- Om vårt spridningsdiagram ser ut som i bilden nedan kan vi testa att
  - Transformera y till y^2
  - Transformera x till $\sqrt{x}$ eller till $\log(x)$

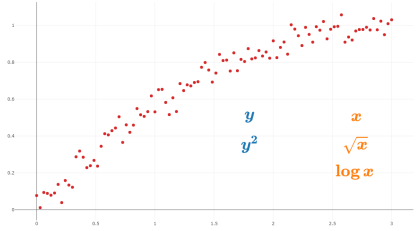


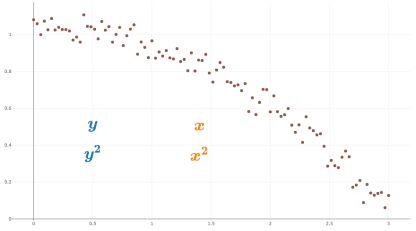
- Om vårt spridningsdiagram ser ut som i bilden nedan kan vi testa att
  - Transformera y till y^2
  - Transformera x till x^2

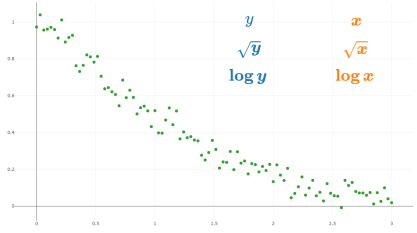


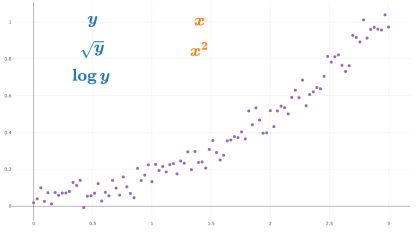
Val av transformation

- Sammantaget bildar de fyra bilderna en cirkel



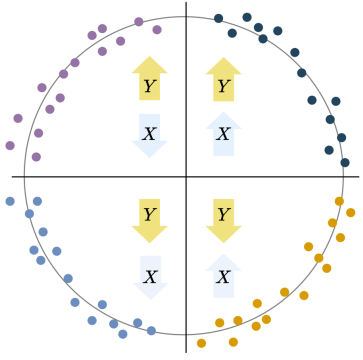




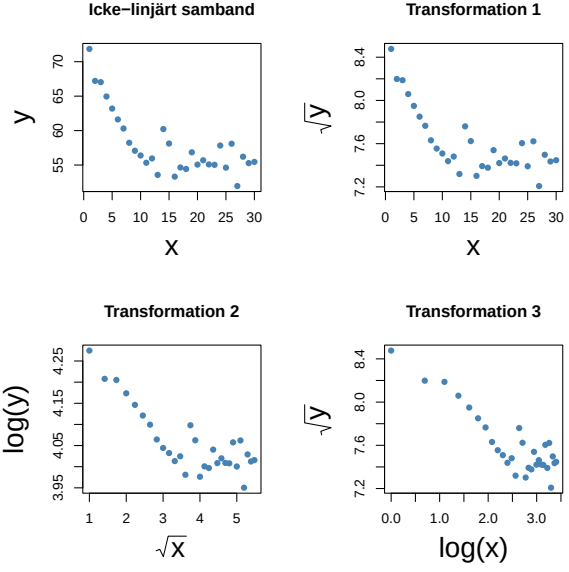


- Tukeys cirkel beskriver i kompakt form transformationer som kan vara lämpliga
- Varje fjärdedel av cirkeln representerar ett visst icke-linjärt mönster
- Pilarna uppåt och nedåt indikerar vilket “håll vi bör röra oss” på skalan till vänster när vi transformerar, om vi utgår från y och x

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	$-\frac{1}{y}$	$-\frac{1}{x}$



Exempel med Tukeys cirkel



- Figurerna 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", och därför valde vi den
- Ni kan testa att skatta och avtransformera de övriga linjerna, och se vilken som passar bäst!

Multipel linjär regression

- Hittills har vi pratat om **enkel linjär regression** där modellen innehåller **en** förklarande variabel, dvs

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

- Vi kan dock tänka oss situationer där **flera** variabler variabler är relevanta
  - Om $y =$ månadslön, är det t.ex. rimligt att både utbildningsnivå och arbetslivserfarenhet förklarar y
  - Om $y =$ skörd från en jordgubbsplanta, kan både temperatur och näringsnivå i jorden vara rimliga x
- En regressionsmodell med **flera** förklarande variabler kallas för en **multipel** linjär regressionsmodell

- En multipel linjär regressionsmodell med två förklaringsvariabler skrivs som

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

- I detta fall är x_1 och x_2 två olika förklaringsvariabler
- Vi kan t.ex. ha

$$\underbrace{\widehat{\text{jordgubsskörd}}}_{\hat{y}} = b_0 + b_1 \cdot \underbrace{\text{lufttemperatur}}_{x_1} + b_2 \cdot \underbrace{\text{näringsnivå}}_{x_2}$$

- I allmänhet skriver vi en modell med k variabler som

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

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

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

- Vi kan använda `R` för att hitta modellens koefficienter, dvs b_0, b_1 och b_2
- Vi lägger till flera förklarande variabler m.h.a. plustecknet

```
lmod_cars <- lm(litermil ~ viktton + hp, data = mtcars)
lmod_cars$coefficients
  (Intercept)     viktton          hp
 0.149155845 0.598453723 0.001769319
```

- Med lite avrundning får vi

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

- Not: I `mtcars` har jag räknat om `mpg` och `wt` till skalorna liter/mil och kg

- För **enkel** linjär regression visade vi formler för beräkning av b_0 och b_1
- För multipel linjär regression är formlerna betydligt mer komplicerade, och ingår därför inte i denna kurs
- Vi följer dock samma princip som innan, och väljer de koefficienter som minimerar

$$\sum_{i=1}^{n} e_i^2, \text{ där } e_i^2 = (y_i - \hat{y})^2 \text{ står för en kvadrerad residual}$$

- Vi väljer alltså koefficienterna som ger den "bästa" modellen, i meningen att den minimerar summan av de kvadrerade residualerna

Generellt antal förklaringsvariabler

- Det finns få begränsningar i antalet förklaringsvariabler som vi kan ha med i en modell, men det krävs att
  - Vi behöver också ha fler datapunkter/observationer än variabler
  - Vi har en dator som är stark nog att klara av stora data (sällan ett problem för linjär regression)
- Vi såg tidigare att vi kan formulera den generella modellen

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

- Modellen ovan bygger på k stycken förklaringsvariabler, och enligt villkoret behöver vi alltid ha att $k < n$
- Stora modeller är relativt vanliga i praktiken, t.ex. när man vill analysera ett samband mellan två specifika variabler, men ta hänsyn till andra variabler som är relevanta

Val av förklaringsvariabler

- Vi hävdade tidigare att den enkla modellen från F7 blir **bättre** när vi lägger till `hästkrafter` som förklarande variabel
- Vad menar vi med detta, och kan vi verifiera att det stämmer?
- Ett enkelt sätt är att jämföra R^2 från modellerna – vi vet att modellen med störst R^2 förklarar störst andel av variationen i responsvariabeln
- På förra föreläsningen såg vi att R^2 blev 0.7919 när vi använde `vikt` som enda förklaringsvariabel
- Vi kan nu beräkna R^2 för den utökade modellen, och se vad som händer när vi lägger till `hästkrafter`

Val av förklaringsvariabler

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

```
summary(lmod_cars)

  Call:
  lm(formula = litermil ~ viktton + hp, data = mtcars)

  Residuals:
       Min       1Q   Median       3Q      Max
  -0.39906 -0.11734  0.01813  0.09469  0.33636

  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
```

Kommentar om R^2

- Vi bör dock ta resultatet ovan med en nypa salt, då R^2 **alltid** blir större när vi lägger till ytterligare variabler
- Detta gäller **även om det helt saknas samband** mellan den nya förklaringsvariabeln och responsvariabeln (ökningen kan ibland vara väldigt liten)
- Eftersom R^2 blir större oavsett vilken variabel vi lägger till, betyder större R^2 inte nödvändigtvis att modellen blir bättre
- Som alternativ kan vi använda ett **justerat** R^2 (*adjusted R-squared*)
- R^2_{adj} är ett mått som ökar med R^2, men som samtidigt minskar med antalet förklaringsvariabler
- R^2_{adj} minskar så om den nya variabeln inte tillför tillräckligt mycket information

Definition av R^2_{adj}

- Justerat R^2 beräknas med formeln

$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(n-1)}{n-k-1} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

- Här är n antalet observationer i våra data, och k antalet förklaringsvariabler i vår modell
- Formeln är inte särskilt intuitiv, men i allmänhet gäller som tidigare att **större R^2_{adj} indikerar en bättre modell**

Val av förklaringsvariabler med R^2_{adj}

- Vi kan nu använda R^2_{adj} för att jämföra våra modeller
  - Den ursprungliga modellen använder bara **vikt** som förklaringsvariabel
  - Den nya modellen använder även **hästkrafter**
- Vi börjar med att räkna ut R^2_{adj} för den mindre modellen

```
lmod_cars1 <- lm(litermil ~ viktton, data=mtcars)
summary_1 <- summary(lmod_cars1)

summary_1$adj.r.squared
  [1] 0.7849726
```

Val av förklaringsvariabler med R^2_{adj}

- 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)

summary_2$adj.r.squared
  [1] 0.8365429
```

- Vi har alltså

	Bara vikt	Vikt och hp
R^2_{adj}	0.79	0.84

- Vi ser att R^2_{adj} stiger när vi lägger till `hästkrafter`, vilket indikerar att den nya variabeln tillför värdefull information som gör modellen bättre

Tolkning av koefficienter

- Vi har redan visat hur vi tolkar koefficienter i en enkel regressionsmodell
- Tolkningarna förändras inte jättemycket när vi lägger till fler variabler, men det finns en **viktig skillnad** att bära med sig
- **Enkel linjär regression:** $\hat{y}$ (det skattade värdet på y) ökar med b_1 enheter då x ökar med en enhet
- **Multipel linjär regression:** $\hat{y}$ ökar med b_j enheter då x_j ökar med en enhet, **givet att övriga förklaringsvariabler hålls konstanta**
- Den sista delen av tolkningen **krävs** för att tolkningen ska vara korrekt
- Om vi ändrar flera variabler samtidigt kan deras effekter på $\hat{y}$ förstärka eller ta ut varandra, och tolkningen av ett enskilt b_j blir då skev
- Om b_j är negativt får vi en "negativ ökning" i $\hat{y}$ när b_j ökar – med andra ord får vi en minskning i $\hat{y}$

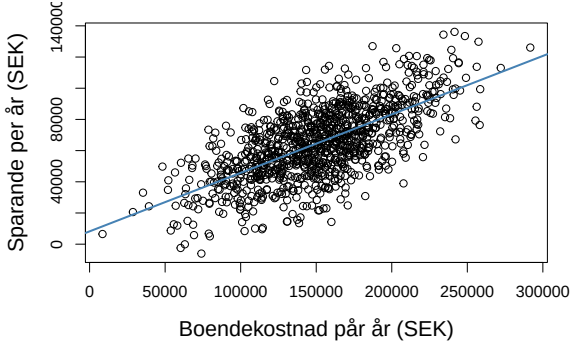
Tolkning av koefficienter

- Vi har följande koefficienter i vår multipla regressionsmodell

```
lmod_cars$coefficients
  (Intercept)     viktton          hp
  0.149155845 0.598453723 0.001769319
```

- **Tolkning av b_1 (`viktton`):**
  - Vi har två bilar med *samma antal hästkrafter*, men den ena väger ett ton *mer* än den andra
  - Modellen uppskattar då att den tyngre bilen förbrukar ≈ 0.598 **liter per mil mer** än den tyngre bilen
- **Tolkning av b_2 (`hp`):**
  - Vi har två bilar som *väger exakt lika mycket*, men den ena bilen har 1 hästkraft mer än den andra
  - Modellen uppskattar att bilen med fler hästkrafter förbrukar **0.00177** liter mer per mil

- Låt oss titta på ett nytt datamaterial, som vi kan använda för att undersöka sambandet mellan sparande, inkomst och boendekostnad
- Vi är intresserade av att skatta hur mycket individerna i våra data sparar årligen, givet deras inkomst och deras boendekostnader
- Vi börjar med att skatta en enkel linjär regressionsmodell där boendekostnad är ensam förklaringsvariabel



Tolkning av koefficienter

- Vi skriver modellen som nedan, och beräknar dess koefficienter i R

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

```
lm(sparande ~ boendekostnad)$coefficients
    (Intercept) boendekostnad
   8206.3659478     0.3750134
```

- Efter avrundning kan vi även skriva ut modellen som

$$\widehat{\text{sparande}} = 8.206 + 0.375 \cdot \text{boendekostnad}$$

- **Tolkning av b_1:** För varje extra krona i bostadskostnad (per år), ökar vår uppskattning av sparandet med 0.375 kronor (per år)
- **Ytterligare analys:** Om person A har en boendekostnad som är 10 000 kronor högre än person B, så uppskattar vi att person A sparar $\approx$ 3 750 kronor mer per år

Tolkning av koefficienter

- Vi lägger nu till variabeln `inkomst`, och får en ny modell

```
lm(sparande ~ boendekostnad + inkomst)$coefficients
    (Intercept) boendekostnad       inkomst
   -71.74833135   -0.08899729    0.19553001
```

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

- **Tolkning av b_1:** För varje extra krona i (årlig) boendekostnad uppskattar modellen att en person sparar 0.089 kronor **mindre** per år, *givet att inkomsten är konstant*
- I denna modell **minskar** sparandet när boendekostnaden ökar!
- När vi bara hade `boendekostnad` som förklaringsvariabel uppskattade modellen att sparandet *ökade* med boendekostnaden
- Betyder det att en av modellerna **har fel**? Vilken av dem i så fall?

Tolkning av koefficienterna

- Bara för att modellerna ser ut att motsäga varandra betyder det **inte** att de har fel
- När vi skattar en modell kan vi se det som att vi "ställer en fråga" till vårt datamaterial
- Frågan blir i någon mening unik för varje modell, och **olika** modeller svarar därför på **olika** frågor
- Detta är precis vad som händer i exemplet med sparande – modellerna svarar på olika frågor, och ger oss olika information
- Båda svaren är dock rimliga i sin egen kontext!

Tolkning av koefficienter

- När vi skattar det första modellen ställer vi frågan *"hur förhåller sig variablerna* `sparande` *och* `boendekostnad` *till varandra?"*, och modellen ger svar därefter
- Modellen tar **inte** hänsyn till att andra variabler kan påverka `sparande`
- Kan det vara så att den som bor dyrare också har större inkomst, och därför kan spara mer? Det låter som en trolig förklaring!
- När vi skattar den andra modellen ställer vi frågan *"hur förhåller sig* `sparande` *och* `boendekostnad` *till varandra,* ***när vi även tar hänsyn till*** `inkomst`*?"*
- För att isolera sambandet mellan `sparande` och `boendekostnad` behöver vi välja en fast inkomstnivå, och se vad som händer med sparandet när boendekostnaden ökar
- Alltså: *givet att personens inkomst inte förändras*, så uppskattar modellen att sparandet kommer minska när boendekostnaden stiger
- När vi väljer modell så skapar vi en kontext inom vilken modellen måste tolkas – därför är bisatsen *givet att övriga variabler hålls konstanta* **extremt** viktig

Tolkning av koefficienter

- Båda modellerna ger logiska resultat, eftersom de beskriver två olika scenarier
- **Modell 1:** 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 lär personerna med dyra bostäder också ha högre inkomster
  - Det betyder att de också kan spara mer, trots att de har dyrare boende
- **Modell 2:** Vi samlar ihop en grupp personer där **alla har samma inkomst**
  - Inom denna grupp jämför vi dem som bor dyrt med dem som bor billigt
  - De som spenderar mindre på sitt boende har rimligtvis mer pengar över att spara, och därför är deras sparande troligtvis högre
- Även om *"olika frågor ger olika svar"* är en helt rimlig devis kan abstraktionen i våra modeller lätt förvilla, och göra modellerna svåra att tolka
- För att kunna dra korrekta slutsatser är det dock viktigt att vara noggrann med tolkningar – det är därför bra att diskutera med kursare/kollegor vid osäkerhet!

Tolkning av koefficienterna

- I multipel linjär regression säger vi att sambandet mellan y och den enskilda variabeln x_j **är betingat på övriga förklaringsvariabler**
- Jämför med betingade proportioner från F3, där vi 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 med hjälp av regressionsanalys **inte** är kausala per automatik
- Modell 1 visar t.ex. inte att högre boendekostnader leder till större sparande, utan snarare att de som har högre boendekostnader kan spara mer **trots** de högre boendekostnaderna

Tolkning av koefficienter

- Vi tittar på ett exempel till, där vi skattar skattar huspriser (i dollar)
- Regressionsmodell 1 har antalet sovrum som enda förklaringsvariabel

$$\widehat{\text{price}} = 338.975 + 40.234 \cdot \text{bedroom}$$

- Regressionsmodell 2 inkluderar även boytan i kvadratfot

$$\widehat{\text{price}} = 308.100 + 135 \cdot \text{living area} - 43.347 \cdot \text{bedroom}$$

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

Tolkning av koefficienter

- Enligt den första regressionsmodellen kostar hus med många sovrum mer än hus med få sovrum – rimligt då hus med många sovrum bör vara större
- Enligt den andra modellen kostar hus med många sovrum mindre än hus med få sovrum, givet husets storlek
- Om två hus är lika stora uppskattar vi alltså att det hus som har färre sovrum är dyrare
- Det skulle kunna bero på att huset med färre sovrum har större kök, större vardagsrum, mer plats för förvaring, osv

Fler än två förklaringsvariabler

- När vi har fler än två förklaringsvariabler måste vi göra tolkningar **givet alla de övriga förklaringsvariablerna**
- Vi tittar på en modell som förklarar 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 koefficienten för antalet växlar är ungefär -0.04, vad betyder detta?
- Om vi har två bilar med **samma** vikt **och** samma antal hästkrafter, där den **enda skillnaden** är att bil 1 har en växel mer än bil 2, så skattar modellen att bil 2 drar 0.04 liter bensin *mindre* per mil

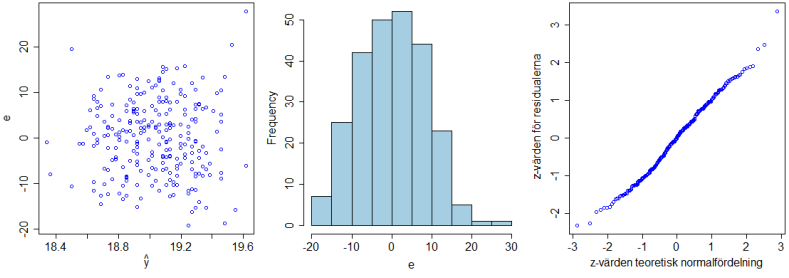
Förutsättningar för multipel linjär regression

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

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

- Här är n antalet observationer, och k antalet förklaringsvariabler

- Vi kan använda nedan figurer för att se om förutsättningarna är uppfyllda
  - 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
  - Normalfördelningsgrafen visar även den att residualerna är approximativt normalfördelade



Förutsättningar för multipel linjär regression

- Vi behöver inte ställa orimligt höga krav när vi bedömer om en regressionsmodell uppfyller antagandena
- 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, utan det viktiga är att modellens brister **kommuniceras öppet!**
- Då vet den som mottar informationen att modellens resultat bör tolkas med försiktighet

Tack för idag!!