

## Föreläsning 8: Enkel linjär och icke-linjär regression

**Matias Quiroz**<sup>1</sup>

<sup>1</sup>Statistiska institutionen, Stockholms universitet

VT 2023

- ▶ Minsta kvadratmetoden för linjär regression.
- ▶ Prediktion i linjär regression.
- ▶ Regression mot medelvärdet.
- ▶ Residualanalys.
- ▶ R-kvadrat som mått på förklaringsgrad.
- ▶ Variansuppdelning i linjär regression.
- ▶ Icke-linjär regression via transformationer.

# Prediktion är mer användbart än korrelation!

- ▶ Förra föreläsningen gick vi igenom begreppet samband och **speciellt linjärt samband** i form av korrelation.
- ▶ Korrelationskoefficienten är ett mått på linjärt samband.
- ▶ Att veta styrkan och riktningen på det linjära sambandet mellan två variabler är viktigt.
- ▶ Men det är mycket viktigare (och roligare!) att, om  $y$  och  $x$  har ett samband, prediktera värdet på  $y$  givet ett värde på  $x$ .
- ▶ En modell som predikterar  $y$  givet ett värde på  $x$  kallas för **regression** i statistik. I machine learning kallas det AI (artificial intelligence) ☺.
- ▶ När vi antar att  $y$  beror linjärt på  $x$  så kallas det för enkel linjär regression (**linear regression** på engelska).
- ▶ När vi antar att  $y$  beror linjärt på  $k$  variabler  $x_1, x_2, \dots, x_k$ , så kallas modellen multipel linjär regression (**multiple linear regression** på engelska).

# Anpassa en rät linje: Minsta kvadratmetoden

- Fett mot protein för olika menyalternativ på Burger King:

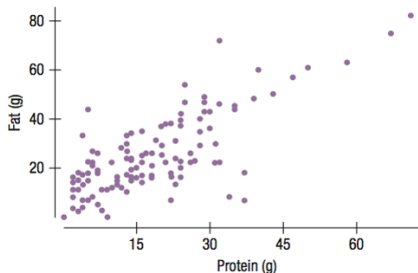


Figure 1: Figur 7.1 i De Veaux et al. (2021).

- Korrelationskoefficienten är  $r = 0.76$ .
- Det verkar finnas ett linjärt samband mellan  $y$  och  $x$ .
- Vilken rät linje ska vi välja för att beskriva sambandet?

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ Låt  $\hat{y}$  vara modellens predikterade genomsnittliga värde för ett givet  $x$ .
- ▶ Om vi antar en linjär modell, så predikterar vi enligt

$$\hat{y} = b_0 + b_1x,$$

där  $b_0$  är interceptet (prediktionen för  $\hat{y}$  när  $x = 0$ ) och  $b_1$  är linjens lutning.

- ▶ I praktiken finns det oändligt många värden på  $b_0$  och  $b_1$  vi kan välja.
- ▶ Vi vill välja  $b_0$  och  $b_1$  som anpassar data på ett optimalt sätt.
- ▶ Vad vi menar med “att anpassar data på ett optimalt sätt”? Vi vill minimera (i någon mening) modellens prediktionsfel.
- ▶ Prediktionsfelet mäts av **residualer**. En residual är skillnaden mellan en observation och modellens prediktion av observationen,

$$\text{Residual} = \text{Observation} - \text{Predikterat värde.}$$

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- En residual betecknas  $e$  och beräknas enligt

$$e = y - \hat{y} = y - (b_0 + b_1x),$$

eftersom  $\hat{y} = b_0 + b_1x$  för en linjär regression.

- Burger King datasetet med en anpassad linje (röd).

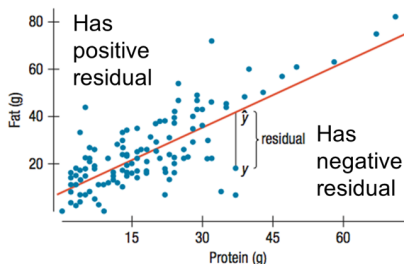


Figure 2: Från lärmaterialet skapat av utgivaren av De Veaux et al. (2021).

- Den röda linjen beskrivs av ekvationen  $\hat{y} = b_0 + b_1x$ .
- Observationer ovanför (under) linjen har positiva (negativa) residualer.

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ Vi vill att den rätta linjen i någon mening minimerar prediktionsfelen, dvs residualerna.
- ▶ Ett tänkbart mått som mäter det sammanlagda prediktionsfelet är residualkvadratsumman (**residual sum of squares** på engelska)

$$\sum e^2, \quad (1)$$

där  $e = y - \hat{y} = y - (b_0 + b_1x)$ .

- ▶ Varför kvadrerar vi residualerna? Vi vill inte att de positiva och negativa felen ska ta ut varandra.
- ▶ Man kan visa att

$$b_1 = r \frac{s_y}{s_x} \text{ och } b_0 = \bar{y} - b_1\bar{x},$$

ger en modell  $\hat{y} = b_0 + b_1x$  som minimerar (1).

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- $\hat{y} = b_0 + b_1x$  med

$$b_1 = r \frac{s_y}{s_x} \text{ och } b_0 = \bar{y} - b_1\bar{x},$$

kallas för **minsta kvadratanpassningen** av data.

- Metoden, dvs att hitta värden  $b_0$  och  $b_1$  som minimeras (1) kallas för minsta kvadratmetoden (**least squares method** på engelska).
- För Burger King exempel ger minsta kvadratmetoden

$$\widehat{Fat} = 8.4 + 0.91Protein,$$

dvs

$$b_0 = 8.4 \text{ och } b_1 = 0.91.$$

- Hur tolkas  $b_0$  och  $b_1$ ?



# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ I ekvationen  $\hat{y} = b_0 + b_1x$  är  $b_0$  interceptet och  $b_1$  linjens lutning.
- ▶ Interceptet är det predikterade värdet när  $x = 0$ :  $\hat{y} = b_0$ .
- ▶ För Burger King exemplet: Den predikterade genomsnittliga fettmängden för en produkt utan proteiner.
- ▶ Man ska vara försiktig med att tolka interceptet som något meningsfullt.
- ▶ Exempel: Modell som predikterar blodtryck som en linjär funktion av vikt. Meningsfull tolkning av interceptet?
- ▶ Viktigt att fråga sig om det är meningsfullt att tolka interceptet innan man gör det!

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ En lutning i en rät linje defineras som

$$\text{Lutning} = \frac{\text{Förändring i } y \text{ variabeln}}{\text{Förändring i } x \text{ variabeln}} = \frac{\Delta y}{\Delta x}. \quad (2)$$

- ▶ Lutningen har enheten: enhet  $y$ /enhet  $x$ , dvs “enhet  $y$  per enhet  $x$ ”.
- ▶ I Burger King exemplet: 0.91 gram fett per gram protein.
- ▶ Vi kan skriva om (2) som

$$\Delta x \cdot \text{Lutning} = \Delta y. \quad (3)$$

- ▶ Antag att  $x$  ökar en enhet, dvs  $x \rightarrow x + 1$ . Då är  $\Delta x = x + 1 - x = 1$ .
- ▶  $\Delta x = 1$  i (3) ger  $\Delta y = 1 \cdot \text{Lutning} = \text{Lutning}$ .
- ▶ Således kan vi tolka  $b_1$  som den förändringen i  $y$  som är associerad med en en-enhets ökning av  $x$ .

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ Det är frestande att säga: Att öka  $x$  med en enhet medför en ökning av  $y$  med  $b_1$  enheter.
- ▶ Detta är en kausal tolkning! Vi säger att  $x$  medför  $y$ .
- ▶ **Regression modellerar inte kausalitet.**
- ▶ **Regression modellerar samband.**
- ▶ Kausalitet kräver teoretiska resonemang kring variablerna i fråga.
- ▶ I Burger King exemplet:
  - ▶ Läger vi till ett gram protein så medför det i genomsnitt  $b_1 = 0.91$  extra gram fett (kausal tolkning).
  - ▶ Produkter med ett extra gram protein tenderar att (i genomsnitt) ha  $b_1 = 0.91$  extra fett (sambandstolkning).

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ Minsta kvadratanpassningens  $\hat{y} = b_0 + b_1x$  egenskaper:
  1. Minimerar residualkvadratsumman i (1).
  2. Residualerna summerar till 0, dvs  $\sum e = 0$ .
  3. Den anpassade linjen går genom punkten  $(\bar{x}, \bar{y})$ .
- ▶ Man kan visa 1. genom att minimera (1) med avseende på  $b_0$  och  $b_1$ . Se s.236 i De Veaux et al. (2021) (för den nyfikne studenten).
- ▶ Egenskap 2. kan visas genom att använda  $b_0 = \bar{y} - b_1\bar{x}$ ,

$$\begin{aligned}\sum e &= \sum y - (b_0 + b_1x) \\ &= \sum y - b_0 - b_1x \\ &= \sum y - (\bar{y} - b_1\bar{x}) - b_1x \\ &= \sum y - \bar{y} - b_1(x - \bar{x}) \\ &= \sum (y - \bar{y}) - b_1 \sum (x - \bar{x}),\end{aligned}$$

och utnyttja att både  $\sum (y - \bar{y})$  och  $\sum (x - \bar{x})$  är 0.

- För egenskap 3., det predikterade värdet för  $\bar{x}$

$$\begin{aligned}\hat{y} &= b_0 + b_1\bar{x} \\ &= \bar{y} - b_1\bar{x} + b_1\bar{x} \\ &= \bar{y},\end{aligned}$$

eftersom  $b_0 = \bar{y} - b_1\bar{x}$ . Alltså ligger punkten  $(\bar{x}, \bar{y})$  på linjen.

# Anpassa en rät linje: Minsta kvadratmetoden, forts

- ▶ Hur beräknar vi  $b_0$  och  $b_1$  i praktiken?
- ▶ Funktionen `lm` i R (linear models) räknar ut  $b_0$  och  $b_1$ .
- ▶ R får inte användas på tentan. Enkelt att räkna genom formlerna

$$b_1 = r \frac{s_y}{s_x} \text{ och } b_0 = \bar{y} - b_1 \bar{x}.$$

- ▶ Burger King exemplet:

| Protein                    | Fat                        |
|----------------------------|----------------------------|
| $\bar{x} = 18.0 \text{ g}$ | $\bar{y} = 24.8 \text{ g}$ |
| $s_x = 13.5 \text{ g}$     | $s_y = 16.2 \text{ g}$     |
| $r = 0.76$                 |                            |

Figure 3: Tabell från s.230 i De Veaux et al. (2021).

- ▶  $b_1 = 0.76 \cdot 16.2 / 13.5 = 0.912$  och  $b_0 = 24.8 - 0.912 \cdot 18 = 8.384$  (utan avrundning).

# Prediktion i linjär regression

- ▶ När vi anpassat modellen, dvs har ekvationen för linjen, börjar det roliga!
- ▶ Burger King gör reklam för en ny produkt med proteininnehåll 63 gram men det framgår inte från reklamen hur mycket fett den innehåller.
- ▶ Enligt vår modell kommer den i genomsnitt att den innehålla

$$\hat{y} = 8.4 + 0.91 \cdot 63 = 65.7$$

gram fett.

- ▶ Enkelt att prediktera i linjär regression:
  1. Anpassa modellen med minsta kvadratmetoden, dvs räkna  $b_0$  och  $b_1$ .
  2. Beräkna  $\hat{y} = b_0 + b_1x$  för det  $x$  värdet man vill prediktera responsvariabeln för.
- ▶ Var försiktig med att prediktera för  $x$ -värden utanför intervallet för de  $x$  värden användes för att anpassa modellen.
- ▶ Linjära modellen måste vara trovärdig. Mer om hur vi kollar detta snart.

# Vad händer om vi får nya data?

- Nya data ger ny deskriptiv statistik. Villanis `widget1` och `widget2`.
- Om vi tar ett nytt stickprov ändras minsta kvadratanpassningen, dvs hela linjen.
- Nya  $b_0$  och  $b_1$ . Hur varierar lutningen  $b_1$  från stickprov till stickprov?
- Ett stickprov på 193 broar i New York state. Skick mot ålder vid inspektion.

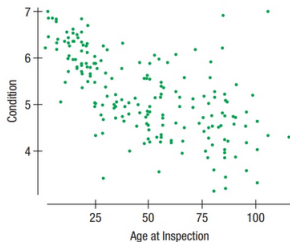


Figure 4: Figur från s.232 i De Veaux et al. (2021).

- $b_1 \approx -0.02$ . Skickmättet  $\downarrow$  med ca 0.02 per år en byggnad föråldras.



## Vad händer om vi får nya data?, forts.

- ▶ Beteckna lutningskoefficienten för populationen  $\beta_1$ .
- ▶ Stickprovets lutningskoefficient är  $b_1 \approx -0.02$ .  $b_1$  är en **skattning** av populationens lutningskoefficient  $\beta_1$ . Villani berättar mer.
- ▶ Finns det ett linjärt samband mellan skick och ålder **i populationen**?
- ▶ Antag att ett sådant samband inte finns. Då är  $\beta_1 = 0$ .
- ▶ Ett första resonemang:  $b_1 \approx -0.02$  vilket är ganska nära 0. Förmodligen finns inget samband.
- ▶ Resonemanget ovan tar inte hänsyn till att  $b_1$  **varierar från stickprov till stickprov**.

# Vad händer om vi får nya data?, forts.

- ▶ Det finns 17493 broar i New York state. Tag 1000 nya stickprov, varje stickprov bestående av 193 observationer.
- ▶ **Samplingfördelningen** för lutningskoefficienten  $b_1$ :

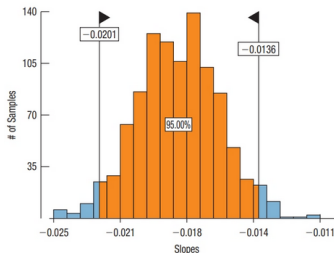


Figure 5: Figur från s.234 i De Veaux et al. (2021).

- ▶ Kommentarer:
  - ▶ 95% av alla stickprov ger lutningar mellan  $-0.0201$  och  $-0.0136$ .
  - ▶ Samplingfördelningen är (approximativt) symmetrisk.
  - ▶ Det negativa sambandet ( $b_1 < 0$ ) uppkommer vid varje stickprov.
  - ▶  $b_1 \approx 0.02$  är inte ett alltför ovanligt värde.

# Vad händer om vi får nya data?, forts.

- ▶ Simuleringarna motstrider att  $\beta_1 = 0$  (varenda simulering visade att skattningen var  $< 0$ ).
- ▶ Simuleringarna stödjer hypotesen att de finns ett negativt linjärt samband, dvs att  $\beta_1 < 0$ .
- ▶ Trots att  $b_1$  var ganska liten, så hade den tillräcklig lite spridning för att vi skulle kunna säga  $\beta_1 < 0$ .
- ▶ Vi kallar detta ett **statistiskt signifikant** resultat. **Resultatet beror inte på slumpen.**
- ▶ Ett resultat som beror på slumpen kallas för ett **statistiskt insignifikant** resultat.
- ▶ Del 2 kommer att gå igenom hur vi kan komma fram till samma slutsats utan att ta 1000 nya stickprov.

# Regression mot medelvärdet

- ▶ Regression mot medelvärdet (**regression to the mean** på engelska): En avvikande  $x$  observation resulterar i en prediktion  $\hat{y}$  som avviker mindre.
- ▶ Exempel: Antag att  $x$  avviker 2 SD ( $s_x$ ) från  $\bar{x}$ . Då avviker  $\hat{y} < 2$  SD ( $s_y$ ) från  $\bar{y}$ .
- ▶  $\hat{y}$  är närmare sitt medelvärde (mätt i SD  $s_y$ ) än vad  $x$  är närmare sitt medelvärde (mätt i SD  $s_x$ ).
- ▶ Regressionen har “dragit ner” prediktionen närmare medelvärdet för  $y$ , jämfört med hur långt ifrån  $x$  låg från medelvärdet för  $x$ .
- ▶ Låt oss göra en regression på standardiserade data för att förklara fenomenet.

- Standariserade data

$$z_x = \frac{x - \bar{x}}{s_x} \text{ och } z_y = \frac{y - \bar{y}}{s_y}.$$

- Medelvärdena efter standarisering är 0, dvs  $\bar{z}_x = 0$  och  $\bar{z}_y = 0$ .
- Standardavvikelserna efter standarisering är 1, dvs  $s_{z_x} = 1$  och  $s_{z_y} = 1$ .
- En minsta kvadratanpassning för regressionen  $z_y$  mot  $z_x$  ger

$$\hat{z}_y = b_0 + b_1 z_x,$$

där  $b_0 = \bar{z}_y - b_1 \bar{z}_x$  och  $b_1 = r \frac{s_{z_y}}{s_{z_x}}$ , och  $r$  är korrelationskoefficienten mellan  $z_y$  och  $z_x$ .

- Notera att  $b_0 = 0 - b_1 \cdot 0 = 0$  och  $b_1 = r \frac{1}{1} = r$ .

- ▶ Minsta kvadratanpassningen är därför

$$\hat{z}_y = rz_x.$$

- ▶  $-1 < r < 1$ , dvs kan inte vara större eller mindre än 1.
- ▶ Exempel: Låt  $z_x = 2$  ( $x$  avviker 2 SD från  $\bar{x}$ ). För alla  $-1 < r < 1$ ,

$$\hat{z}_y = r \cdot 2 < 2, \text{ dvs } \hat{y} \text{ avviker } < 2 \text{ SD från } \bar{y}.$$

- ▶ En förutsättning för att använda linjär regression är att den linjära modellen måste vara trovärdig, dvs anpassa observerade data.
- ▶ “All models are wrong, but some are useful” – George E. P. Box.
- ▶ En **residualanalys** är en mycket viktig del av modellvalidering.
- ▶ Om modellen beskriver data på ett adekvat sätt, så kommer residualerna inte ha något tydligt mönster i sig. De beter sig slumpmässigt.
- ▶ Stora enskilda residualer ger information om outliers. Dessa bör examineras.
- ▶ Vad räknas som en stor residual? Om residualerna är normalfördelade, så vet vi att 99.7% ligger  $\pm 3$  SD från sitt medelvärde.
- ▶ Vad är residualernas medelvärde?
- ▶ Tips:  $\sum e = 0$  enligt egenskap 2. från slide 12.

## Residualanalys, forts.

►  $\bar{e} = \frac{\sum e}{n} = \frac{0}{n} = 0$ , dvs residualernas medelvärde är 0.

► Residualernas standardavvikelse kan räknas enligt

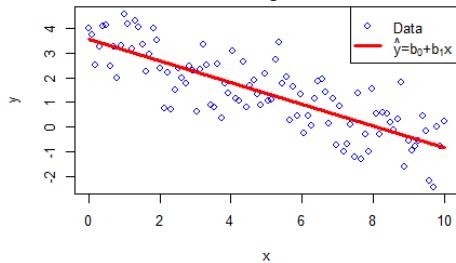
$$s_e = \sqrt{\frac{\sum e^2}{n-2}}.$$

- Varför delar vi med  $n-2$  istället för  $n$  eller  $n-1$ ? Vi behöver lära oss mer statistik innan vi kan förklara det.
- I Del 2 av kursen kommer vi att räkna **samplingfördelningen** för  $b_1$ . För att göra det behöver vi fler modellantaganden som medför:
1. **Residualerna är normalfördelade.**
  2. **Residualernas varians är konstant**, dvs beror inte på  $x$ .
- Vi kan undersöka 1. genom 68–95–99.7 regeln eller en normalfördelningsplot.
- Vi kan undersöka 2. genom att plotta  $e$  mot  $x$  och se om spridningen är ungefärlig densamma för alla  $x$ .

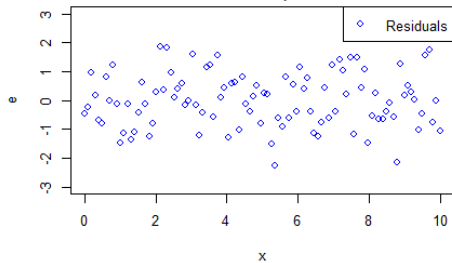


# Residualerna uppfyller modellantaganden

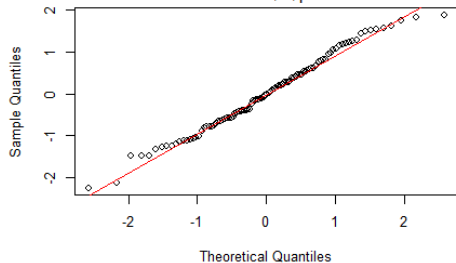
Data and regression fit



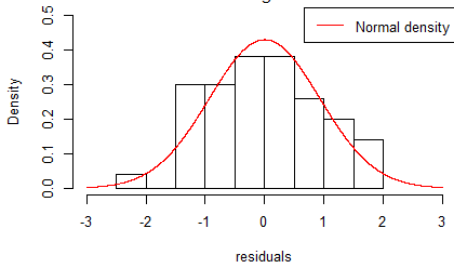
Residual plot



Normal Q-Q plot

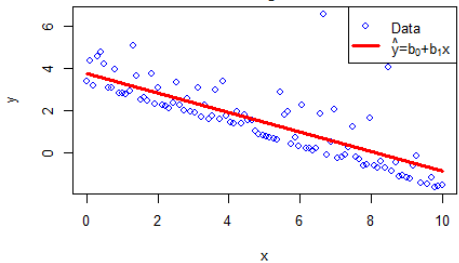


Normalised histogram of residuals

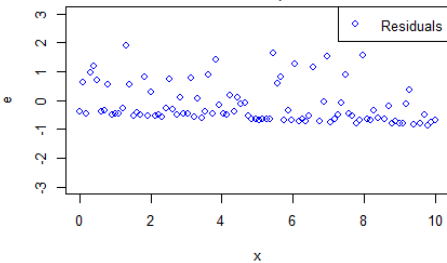


# Residualerna är inte normalfördelade (skeva åt höger)

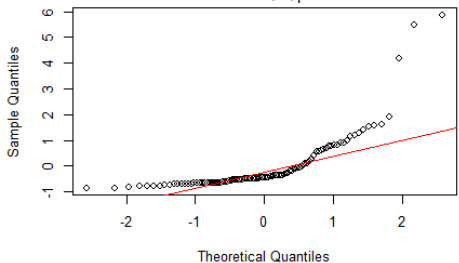
Data and regression fit



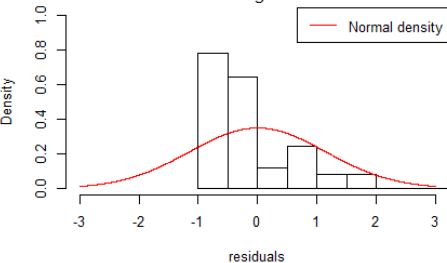
Residual plot



Normal Q-Q plot

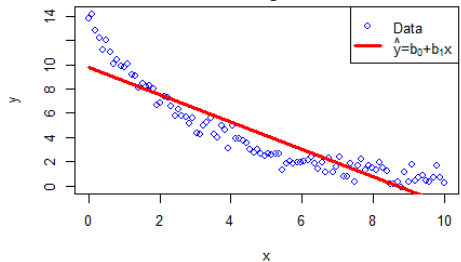


Normalised histogram of residuals

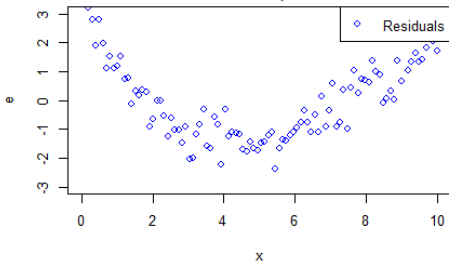


# Residualerna är inte slumpmässiga

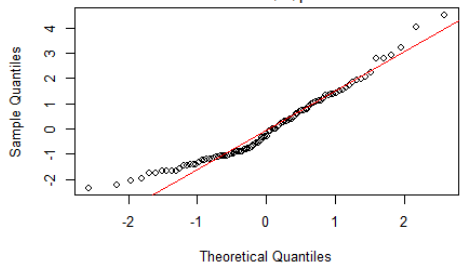
Data and regression fit



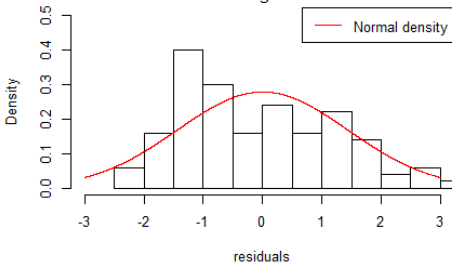
Residual plot



Normal Q-Q plot

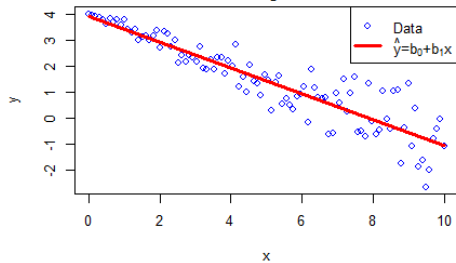


Normalised histogram of residuals

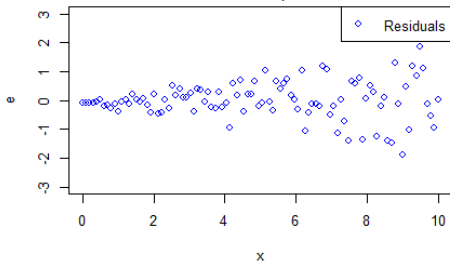


# Residualernas varians är inte konstant

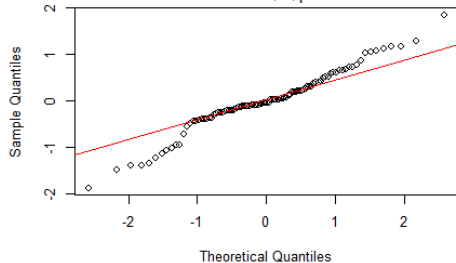
Data and regression fit



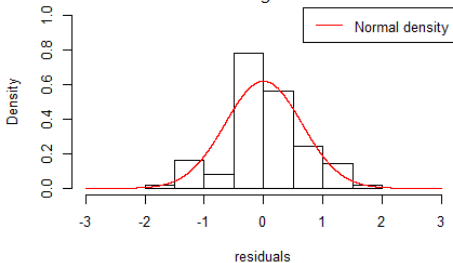
Residual plot



Normal Q-Q plot



Normalised histogram of residuals



# R-kvadrat: Mått på förklarad variation

- ▶ Varför anpassar vi regressionsmodeller?
  1. **Förstå sambandet** mellan  $y$  och  $x$  (positivt/negativt, inget linjärt samband).
  2. **Prediktera**  $y$  givet ett värde på  $x$ .
  3. **Förklara variationen** i  $y$  med hjälp av  $x$
- ▶ Vad menar vi med att “förklara variation i  $y$  med hjälp av  $x$ ” ?
- ▶ Antag först att vi bara har variabeln  $y$  (dvs inget  $x$ ). Ett tänkbart mått på variationen i  $y$  från Föreläsning 3 är den **totala kvadratsumman**,

$$SST = \sum (y - \bar{y})^2.$$

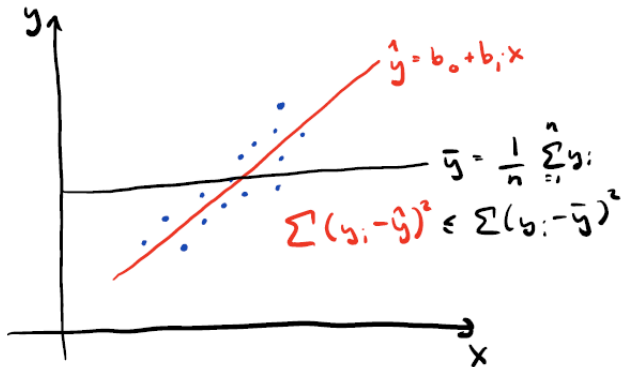
- ▶ Tag nu in  $x$  och skatta en regression  $\hat{y} = b_0 + b_1x$ . Ett tänkbart mått på variationen i  $y$  runt regressionslinjen är **residualkvadratsumman**

$$SSE = \sum (y - \hat{y})^2.$$

- ▶ Vilken av SST och SSE är störst?

## R-kvadrat: Mått på förklarad variation, forts.

- Man kan visa att  $SSE \leq SST$ . Ju mindre SSE är i förhållande till SST, desto mer av den totala variationen SST fångas av regressionen.



- Figuren visar hur en regression ( $\hat{y}$ ) lämnar mindre variation kvar jämfört med att inte ha en regression ( $\bar{y}$ ).

## R-kvadrat: Mått på förklarad variation, forts.

- ▶ SST är den **totala oförklarade variationen**.
- ▶ SSE är den variationen som är kvar efter regressionen, dvs **regressionens oförklarade variation**.
- ▶  $SSE/SST$  är **regressionens oförklarade variation som andel av den totala oförklarade variationen**.
- ▶ Eftersom  $0 \leq SSE \leq SST$ , så ligger  $0 \leq SSE/SST \leq 1$ .
- ▶ R-kvadrat, betecknad  $R^2$  **mäter andelen av den totala variationen som förklaras av regressionen**, dvs modellens **förklaringsgrad**.
- ▶ Den defineras enligt

$$R^2 = 1 - \frac{SSE}{SST},$$

eller i ord

Andel förklarad variation =  $1 - \text{Andel oförklarad variation}$ .

## R-kvadrat: Mått på förklarad variation, forts.

- ▶ Man kan visa att  $R^2$  också kan räknas genom  $r^2$ , därav namnet.
- ▶  $R^2$  ligger mellan 0 och 1. Ibland uttrycks den i procent och multipliceras då med 100.
- ▶  $R^2 \approx 1$  betyder att den linjära modellen har förklarat mestadels av variationen i  $y$  med hjälp av  $x$ .
- ▶  $R^2 \approx 0$  betyder att den linjära modellen förklarar nästan inget av variationen i  $y$  med hjälp av  $x$ .
- ▶ Det går inte att säga vad ett bra värde på  $R^2$  är, det beror på applikationen.
- ▶ I vetenskapliga experiment kan  $R^2$  vara runt 0.8-0.9 om modellen är bra.
- ▶ I de sociala vetenskaperna får man vara nöjd med 0.3-0.5. Svårare att prediktera variabler i t.ex ekonomi jämfört med fysik.



## R-kvadrat: Mått på förklarad variation, forts.

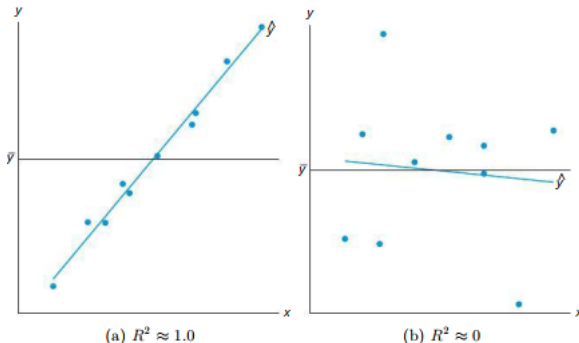


Figure 6: Figur 11.10 in Walpole et al. (2016).

- Figuren till vänster visar en linjär regression som fångar mestadels av den totala variationen.
- Figuren till höger visar en linjär regression som fångar väldigt lite av den totala variationen.

# Variansuppdelning i linjär regression

- ▶ Vi har definierat två olika variationer i linjär regression:
  - ▶ SST är den **totala oförklarade variationen** i  $y$ .
  - ▶ SSE är residualernas variation, dvs **regressionens oförklarade variation**.
- ▶ Det finns en tredje variation som mäter hur mycket prediktionen  $\hat{y}$  varierar kring  $\bar{y}$ .
- ▶ Denna kallas regressionens kvadratsumma (**sum of squares regression** på engelska),

$$SSR = \sum (\hat{y} - \bar{y})^2.$$

- ▶ Ett viktigt resultat är att SST kan dekomponeras enligt

$$SST = SSR + SSE,$$

dvs

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2.$$

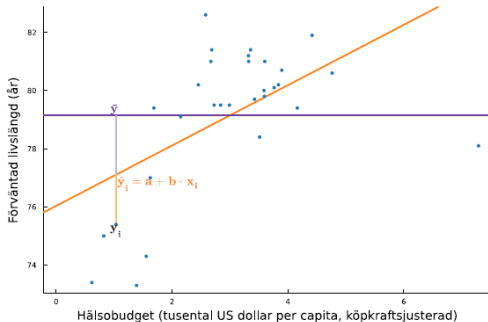
- ▶ Resultatet kallas för variansuppdelning. Analysis of variance (**ANOVA**) på engelska.

# Variansuppdelning i linjär regression, forts.

- Intuition för resultatet fås genom omskrivningen

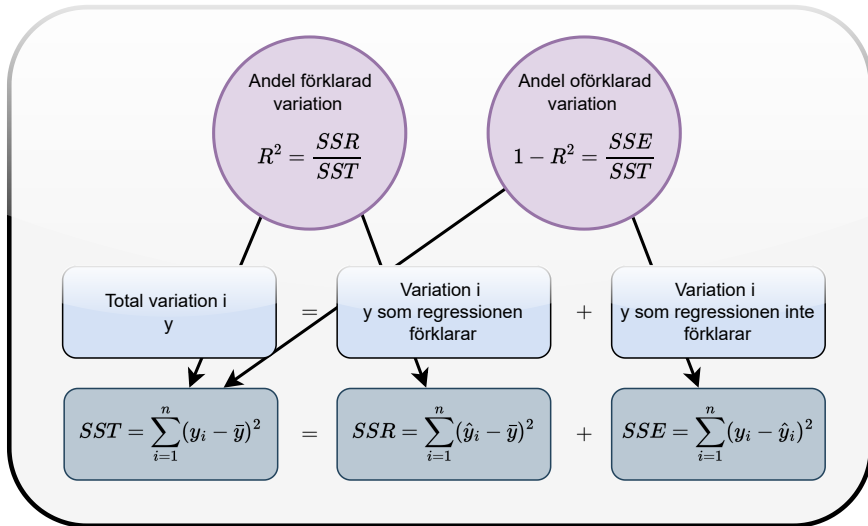
$$y - \bar{y} = y - \hat{y} + \hat{y} - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y}).$$

- Variansuppdelning i regressionen förväntad livslängd mot hälsobudget för olika länder:



- Notera att R-kvadrat också kan räknas som  $R^2 = \frac{SSR}{SST}$ .

# R-kvadrat och variansuppdelning



# Repetition av antaganden i linjär regression

- ▶ Både  $y$  och  $x$  måste vara numeriska variabler.
- ▶ Det finns andra regressionsmodeller man kan använda när  $y$  är en kategorisk variabel — fortsätt läsa statistik (SDA II)!
- ▶ Variablerna måste förhålla sig (approximativt) linjärt till varandra.
- ▶ Uppenbara outliers kan påverka minsta kvadratanpassningen. Anpassa linjen utan outliers för att kontrollera att resultaten blir ungefär desamma.
- ▶ Spridningen för  $y$  densamma för alla  $x$ . Residualernas varians måste vara konstant.
- ▶ Residualerna bör vara (approximativt) normalfördelade.
- ▶ Vad gör vi om variablerna inte förhåller sig linjärt?
- ▶ Förra föreläsningen nämnde vi att man kan transformera variablerna för att få ett mer linjärt förhållande.
- ▶ Icke-linjär regression via potenstransformationer!

# Icke-linjär regression via transformationer

## ► Ladder of powers. Stege av potenstransformationer:

Potenstransformationer för  $y$  och  $x$ .

| Stegnivå | $y$         | $x$         |
|----------|-------------|-------------|
| 1        | $y^2$       | $x^2$       |
| 2        | $y$         | $x$         |
| 3        | $y^{1/2}$   | $x^{1/2}$   |
| 4        | $\log(y)$   | $\log(x)$   |
| 5        | $-y^{-1/2}$ | $-x^{-1/2}$ |
| 6        | $-y^{-1}$   | $-x^{-1}$   |

Figure 7: Tabell från Lab 4. Ladder of powers.

- Om  $y$  och  $x$  inte förhåller sig linjärt så kan vi transformera variablerna för att få ett linjärt förhållande.

# Icke-linjär regression via transformationer, forts.

- ▶ Notera att  $y^{1/2} = \sqrt{y}$ ,  $-y^{-1/2} = -1/\sqrt{y}$  och  $-y^{-1} = -1/y$ .
- ▶ Varför har Stegnivå 5 och 6 negativa tecken?
- ▶ Negativa tecken på dessa stegnivåer ser till att **ordningen bevaras**.
- ▶ Utan det negativa tecknet på de två sista transformationer så bevaras inte ordningen. Exempel om sista transformationen inte har ett minustecken:

$$3 < 6, \text{ och för de transformerade värden } 1/3 > 1/6.$$

- ▶ Med det **negativa tecknet bevaras ordningen**, eftersom

$$3 < 6, \text{ och för de transformerade värden } -1/3 < -1/6.$$

- ▶ Plotta  $y$  mot  $x$  och gå upp och ner för stegen (dvs transformera) tills sambandet mellan de transformerade variablerna ter sig någorlunda linjärt.
- ▶ Ska vi röra oss upp eller ner längst stegen?
- ▶ Beror på formen på sambandet. Mer av en konst än vetenskap, men det finns tumregler. **John Tukeys cirkel**.
- ▶ Transformationer av  $x$  kallas **feature learning** i machine learning.



# Icke-linjär regression via transformationer, forts.

- John Tukeys cirkel för hur man rör sig upp och ner längst stegen:

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

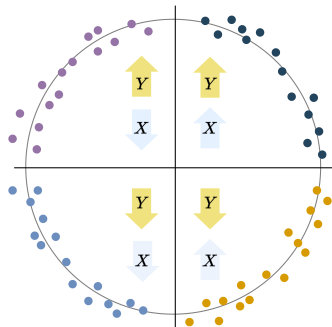
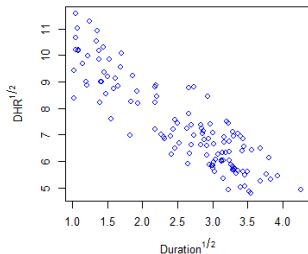
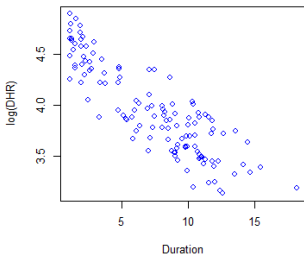
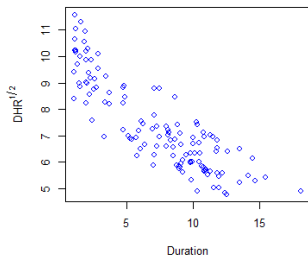
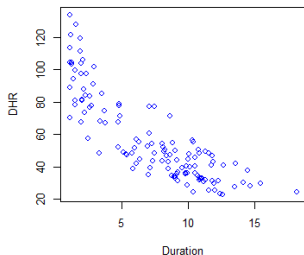


Figure 8: Stege av potenstransformationer (vänster) och Tukeys cirkel (höger).

- Om data är en avtagande funktion som är konvex (konkav uppåt) befinner vi oss på tredje kvadranten. Flytta  $x \downarrow$  och/eller  $y \downarrow$  i stegen.
- Om data är en avtagande funktion som är konkav befinner vi oss på första kvadranten. Flytta  $x \uparrow$  och/eller  $y \uparrow$  i stegen.

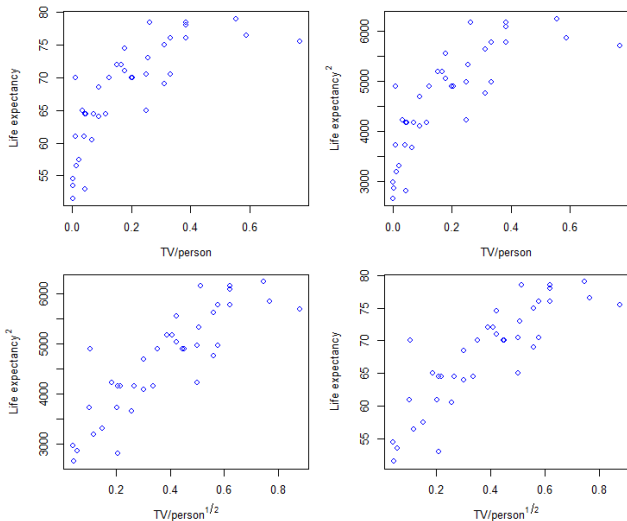
# Icke-linjär regression via transformationer, forts.

- Pingviners dykpuls mot tid för dykningen. Tredje kvadranten på Tukeys cirkel.



# Icke-linjär regression via transformationer, forts.

- Förväntad livslängd mot antal TV per capita. Andra kvadranten på Tukeys cirkel.



# Icke-linjär regression via transformationer, forts.

- ▶ När vi bestämt transformation(er) så anpassar vi en regression på transformerade variabler.
- ▶ I pingvinexemplet väljer vi transformationerna  $DHR^{1/2}$  och  $Duration^{1/2}$ .
- ▶ Vi anpassar modellen

$$\widehat{DHR^{1/2}} = 11.847 - 1.746Duration^{1/2}.$$

- ▶ Notera att anpassningen ger en prediktion i den transformerade skalan, dvs  $DHR^{1/2}$ .
- ▶ Hur går vi från  $\widehat{DHR^{1/2}} \Rightarrow \widehat{DHR}$ ?
- ▶ Vi behöver "reversera transformationen". En funktion som reverserar en transformation kallas för en inverstransformation.

# Icke-linjär regression via transformationer, forts.

Prediktion i originalskala (kolumn till höger) för olika transformerade responser (kolumn till vänster).

| Transformation av responsen | Prediktion i $y$ -skala ( $\hat{y}$ ) |
|-----------------------------|---------------------------------------|
| $y^2$                       | $(\widehat{y^2})^{1/2}$               |
| $y$                         | $\hat{y}$                             |
| $y^{1/2}$                   | $(\widehat{y^{1/2}})^2$               |
| $\log(y)$                   | $\exp(\widehat{\log(y)})$             |
| $-y^{-1/2}$                 | $(\widehat{-y^{-1/2}})^2$             |
| $-y^{-1}$                   | $-(\widehat{-y^{-1}})^{-1}$           |

Figure 9: Tabell från Lab 4. Prediktion i originalskala med olika transformerade responser.

- ▶ Antag att vi vill prediktera genomsnittliga dykpulsen för en pingvin som dyker 16 minuter, dvs  $Duration^{1/2} = 16^{1/2} = 4$ .
- ▶ I  $DHR^{1/2}$  skala är prediktionen

$$\widehat{DHR^{1/2}} = 11.847 - 1.746 \cdot 16^{1/2} = 4.863.$$

- ▶ I  $DHR$  skala blir prediktionen  $\widehat{DHR} = 4.863^2 \approx 23.65$  slag per minut.

# Icke-linjär regression via transformationer, forts.

- Notera att anpassningen är **linjär i de transformerade skalorna**.

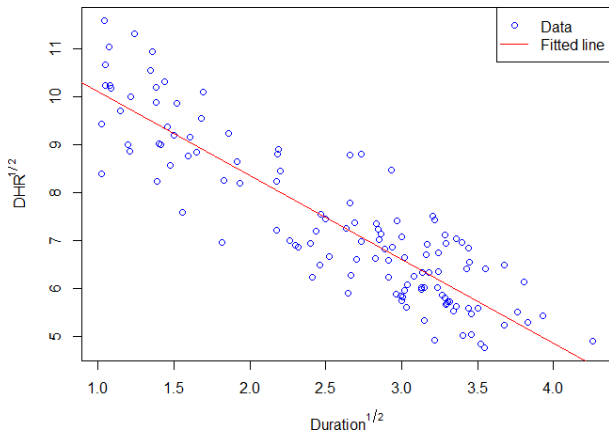


Figure 10: Data och anpassad regression i transformerad skala.

- Varför kallar vi det här för icke-linjär regression?

# Icke-linjär regression via transformationer, forts.

- Regressionen är **icke-linjär i originalskalan!**

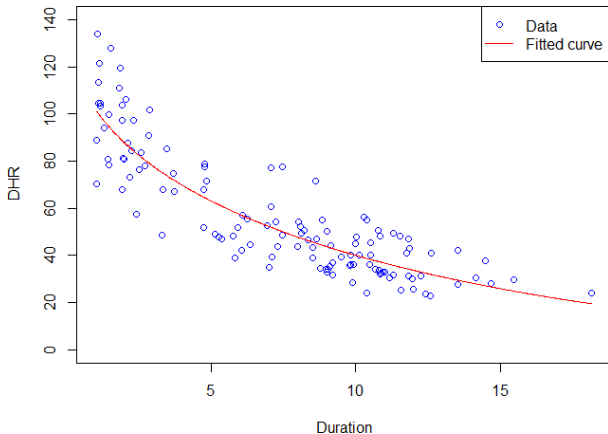


Figure 11: Data och anpassad regression i original skala.



# Transformationer i linjär regression, forts.

- ▶ Det finns tolkningar av  $b_1$  för vissa transformationer men vi går inte igenom dem i SDA I<sup>1</sup>.
- ▶ Är en modell användbar om vi inte kan tolka  $b_1$ ?
- ▶ Om målet är prediktion så bryr vi oss inte om att tolka koefficienter.
- ▶ **Deep learning** modeller saknar tolkningar av koefficienterna. Extremt bra på prediktion för många typer av data, speciellt bilder.
- ▶ Deep learning kan ses som en regression där man gör väldigt komplexa transformationer av  $x$  variabeln<sup>2</sup>.

---

<sup>1</sup>Några modeller i SDA II. Ännu fler i Generaliserade Linjära Modeller (GLM).

<sup>2</sup>Vår kurs Maskininlärning på masternivå lär ut deep learning.

# References I

- De Veaux, R. D., Velleman, P., and Bock, D. (2021). *Stats: Data and Models*. Pearson, Harlow, United Kingdom, fifth edition.
- Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2016). *Probability & Statistics for Engineers and Scientists*. Macmillan New York, 9 edition.