Statistik och dataanalys I

F7: Enkel linjär regression

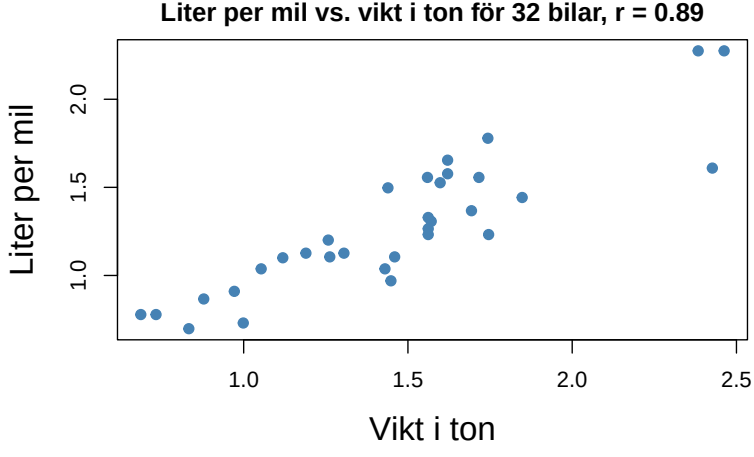
Valentin Zulj

Statistiska institutionen

Stockholm
University

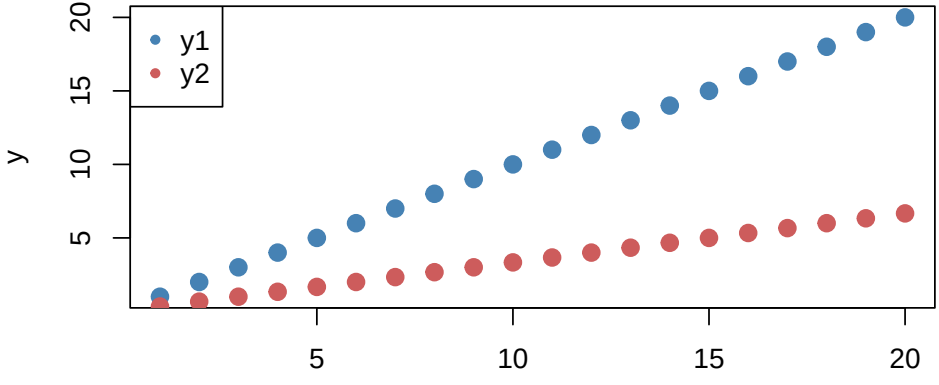
Vad har vi gjort hittills?

- Förra gången pratade vi om linjära samband och **korrelationkoefficienten**
- Vi tittade bl.a. på sambandet mellan vikt och bränsleförbrukning för bilar, och sa att det verkade vara relativt linjärt



Vad vi ska göra nu?

- I bilden nedan har vi $r_{xy_1} = 1$ och $r_{xy_2} = 1$ – båda sambanden är lika starka



- **Men:** vi ser samtidigt att sambanden **inte** är likadana!
- r_{xy} mäter styrkan i linjära samband, men det hjälper oss inte att beskriva hur sambanden ser ut mer konkret

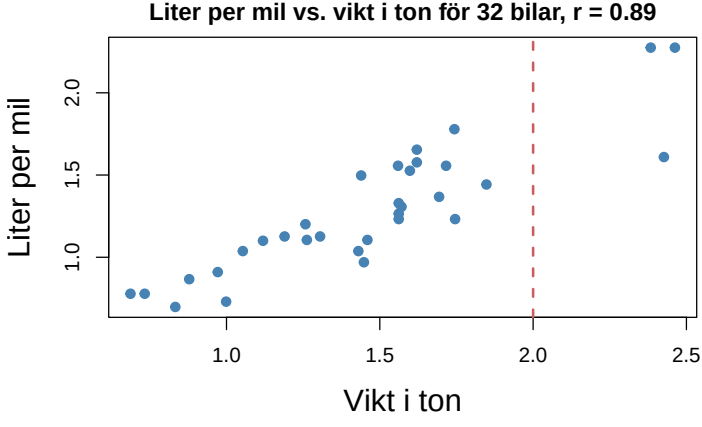
Vad vi ska göra nu?

- Antag att vi är intresserade av bränsleförbrukningen hos en bil som *inte finns med i vårt dataset*, men som vi vet väger *väger 2 ton*
- Även om korrelationen pekar på att vikt och förbrukting är linjärt relaterade, säger den inget om hur vi kan använda vikt för att förutspå förbrukning
- För att göra denna typ av **prediktion** kan använda **linjär regression**, som är vårt huvudfokus idag

Enkel linjär regression

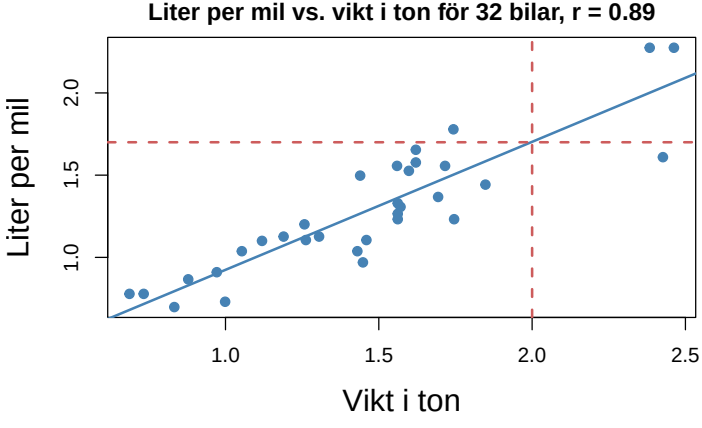
Uppskatta värdet av y när vi känner till värdet av x

- För att passa in i mönstret i vårt spridningsdiagram borde en bil som väger 2 ton förbruka ungefär 1.5–2 liter per mil
- En sådan "gissning" kan i vissa fall vara tillräcklig, men ofta vill vi ge en uppskattning i form av **en** siffra

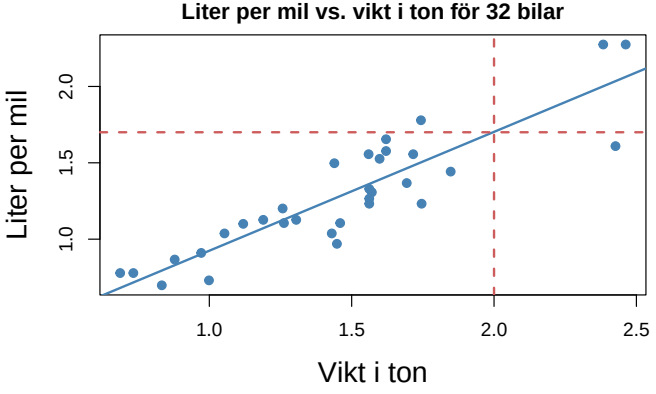


Uppskatta värdet av y när vi känner till värdet av x

- Om vi drar en rät linje genom punktsvärmen kan vi läsa av vilken bränsleförbrukning som, **enlig linjen**, motsvarar vår givna vikt
- När värdet på x-axeln (vikt i ton) är 2, så ser vi att värdet på y-axeln (liter/mil) är ≈ 1.7
- **Slutsats:** Vi uppskattar att en bil som väger 2 ton drar 1.7 liter per mil



- Analysen vi landade i på förra bilden kallas **enkel linjär regressionsanalys**
- Syftet med enkel linjär regression är vanligtvis att
  - Undersöka hur sambandet mellan x och y ser ut
  - Göra **prediktioner** av värdet på y-variabeln, för givna värden på x-variabeln



Liten ordlista för enkel linjär regression (del 1)

- Att **göra en prediktion** (*to predict*) innebär att skatta ett värde (för y) i fall då vi inte kan mäta/observera y direkt
- Vi kallar vår uppskattning för en **prediktion**
  - "Syftet med vår modell är att göra en prediktion av bensinförbrukningen"
- Vi kan säga **estimera** eller **skatta** på ungefär samma sätt som att göra en prediktion
  - "Vi skattar bensinförbrukningen till 1.7 liter per mil"
- Substantivformerna är **estimat** och **skattning**
  - "Vår skattning är att bilen förbrukar 1.7 liter per mil"

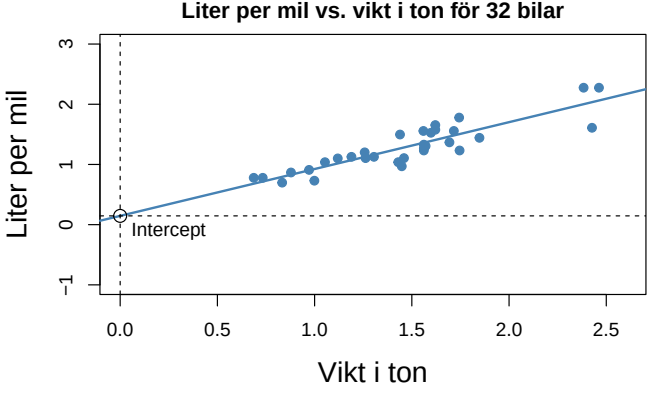
Liten ordlista för enkel linjär regression (del 2)

- Vår y-variabel kan kallas **responsvariabel**, eller **beroende variabeln*
- Vår x-variabel kan kallas **förklarande variabel**, eller **oberoende variabel**
- Vi kallar det **enkel linjär regression** (*simple linear regression*) då modellen för y bara bygger på **en enda** (en enkel) förklarande variabel
- I **multipel linjär regression** (*multiple linear regression*) använder vi **flera** förklarande variabler (mer om detta i F8)

Enkel linjär regression

- Rent intuitivt är linjär regression inte mer komplicerat än att dra en linje genom punktsvärmen, och läsa av vilket y-värde som passar ett givet x
- I praktiken krävs en del beräkningar för att bestämma exakt hur regressionslinjen bör dras för att ge en bra modell
- Vanligtvis vill vi också göra andra beräkningar som är kopplade till regressionsmodellen, t.ex. för att utvärdera hur bra modellen är
- Vi kommer titta närmare på dessa typer av beräkningar nu

- En regressionslinje kan definieras med två **regressionskoefficienter**:
  - **Intercept**: regressionslinjens y-värde när $x = 0$
  - **Lutning** (*slope*): anger hur många enheter y ökar/minskar när x ökar med en enhet
- Regressionslinjen på bilden har en lutning som är ungefär 0.8 och ett intercept strax över 0



Regressionslinjen (enkel linjär regression)

- Ren matematiskt skriver vi regressionslinjen som

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

- Här betecknar y vår responsvariabel (t.ex. förbrukning), och hatten över y indikerar att $\hat{y}$ **är en skattning av** y
- b_0 och b_1 betecknar linjens intercept respektive lutning (jämför med vad ni vet om **räta linjens ekvation**)
- Om vi har värden för b_0, b_1 och x (vikt) kan vi ta fram en skattning av y (förbrukning) med hjälp av ekvationen ovan
- Vid behov kan vi använda beskrivande variabelnamn istället för x och y, t.ex.

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

Skillnaden mellan y och $\hat{y}$

- Vi har redan nämnt "hatten" ovanför y i formeln

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

- Vi skriver som sagt $\hat{y}$ istället för y, då $\hat{y}$ är en **skattning** av y
- $\hat{y}$ står alltså **inte** för det verkliga värdet på y!
- Den som vill verka riktigt smart kan lära sig att hatten rent formellt heter cirkumflex
- I vårt första exempel tittade vi på en bil som väger 2 ton, och uppskatade att den förbrukar ≈ 1.7 liter/mil – vi hade alltså $\hat{y} = 1.7$
- Om vi hade skrivit $y = 1.7$ hade vi hävdat att bilens **verkliga** förbrukning är 1.7 liter/mil, med detta hade varit olämpligt då skattningar inte är perfekta

Tolkning av enkel linjär regressionsmodell

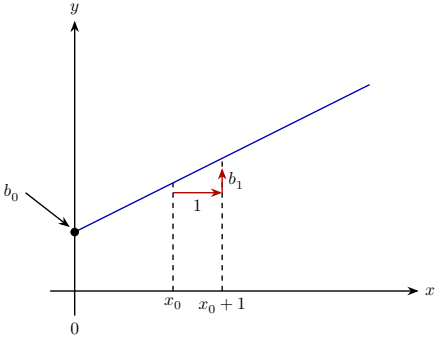
- Regressionsmodeller används ofta för att göra prediktioner, som vi redan visat
- Ett annat syfte kan vara att ge mer konkreta utsagor om sambandet mellan de två variablerna
- Vår modell skattar bensinförbrukning hos en bil, med hjälp av dess vikt
- I abstrakta termer kan vi skriva ut modellen som

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

- **Men:** b_0 och b_1 är generell notation för vanliga värden, som kan vara intressanta att tolka i sig
- Antag t.ex. att $b_0 = 0.146$ och $b_1 = 0.78$, vilket ger

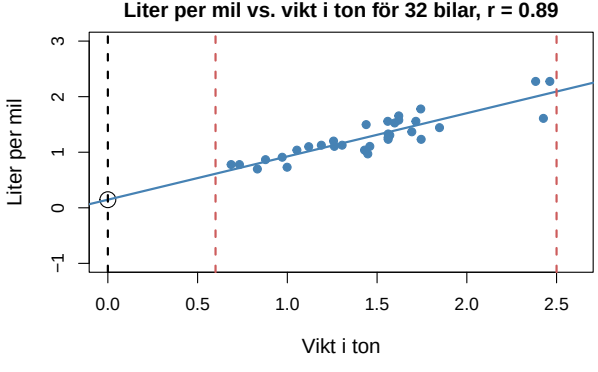
$$\widehat{\text{litermil}} = 0.146 + 0.78 \cdot \text{vikt}$$

- Vad betyder detta mer konkret?



- Vi har
$$\widehat{\text{litermil}} = 0.146 + 0.78 \cdot \text{vikt}$$
- $b_0 = 0.146$: en bil som väger 0 ton förbrukar, **enligt modellen**, 0.146 liter per mil
- I vårt fall är denna information inte meningsfull (varför?), så vi är försiktiga med att tolka b_0 bokstavligt
- $b_1 = 0.78$: en bil förbrukar, **enligt modellen**, ytterligare 0.78 liter/mil bensin för varje extra ton som bilen väger

- Modellen blir mer osäker ju längre bort från punktsvärmen vi kommer
- Vi bör därför vara försiktiga med att använda modellen för att prediktera värden **utanför** det intervall av x-axeln där våra datapunkter ligger
- I vårt exempel är det *risky business* att använda modellen för bilar som väger mindre än 0.5 ton eller mer är 2.5 ton

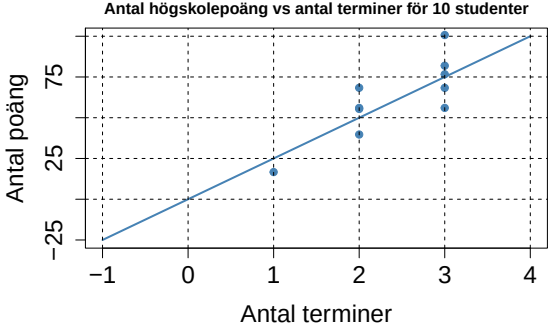


Stockholm University

OBS!

- När vi tittade på korrelation betonade vi att korrelation inte är kausalitet
- Att x har en stark korrelation till y behöver inte betyda att x orsakar y
- Samma sak gäller när vi tolkar lutningsparametern b_1 i en regressionsmodell
- Bara för att $b_1 = 0.78$ kan vi inte utan vidare säga att en viktökning med ett ton *medför* att bränsleförbrukningen ökar med 0.78 liter/mil
- Förändringen i bränsleförbrukning kan också påverkas av hur vikten fördelas, hur ansträngd bilens motor redan är, etc
- **Regressionsmodeller visar samband, inte kausalitet!**
- För att kunna hävda att ett ton i ökad vikt orsakar en viss ökning i bränsleförbrukning behövs mer avancerade vetenskapliga metoder

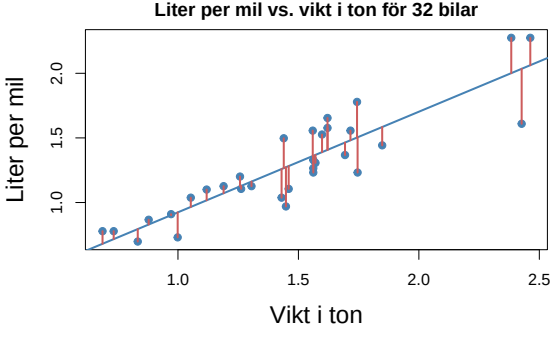
- Bilden visar en linjär regressionsmodell, där varje observation är en student
- x-axeln visar antalet avslutade terminer och y-axeln antalet avklarade högskolepoäng
- Fundera lite snabbt på frågorna
  1. Vilka värden har parametrarna b_0 och b_1?
  2. Hur ska parametrarna tolkas?
  3. Hur många poäng skattar modellen att en student har efter 2 terminer?



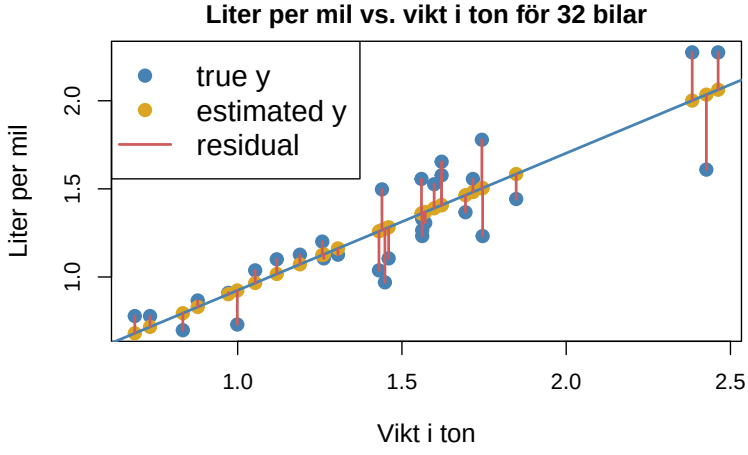
Övning i regressionsanalys

1. Vilka värden har parametrarna b_0 och b_1?
   - Interceptet är värdet på $\hat{y}$ då $x = 0$, alltså har vi $b_0 = 0$
   - $\hat{y}$ ökar med 25 poäng för varje termin som x ökar, alltså har vi $b_1 = 25$
2. Hur ska parametrarna tolkas?
   - b_1 kan tolkas som att det skattade antalet poäng ökar med 25 poäng per termin
   - b_0 kan tolkas som att en student som pluggat i 0 terminer har tagit 0 poäng, enligt modellen
3. Hur många poäng skattar modellen att en student har efter 2 terminer?
   - Vår modell är $\widehat{\text{poäng}} = 0 + 25 \cdot \text{terminer}$, och efter 2 terminer är uppskatningen att en student har $0 + 25 \cdot 2 = 50$ poäng

- För varje observation kan vi göra prediktion, $\hat{y} = b_0 + b_1 x$
- För varje observerat par (x, y) har vi både en prediktion från regressionslinjen $(\hat{y})$, och det *sanna* värdet y
- Skillnaden $e = y - \hat{y}$ kallas **residual** (*residual*), och mäter felet i prediktionen
- Vi har en residual för varje observation i datameterialet (se bild nedan)
- Residualerna kan vara positiva eller negativa – vad betyder det?



- Ju mindre residualerna är desto, bättre passar modellen ihop med observationerna

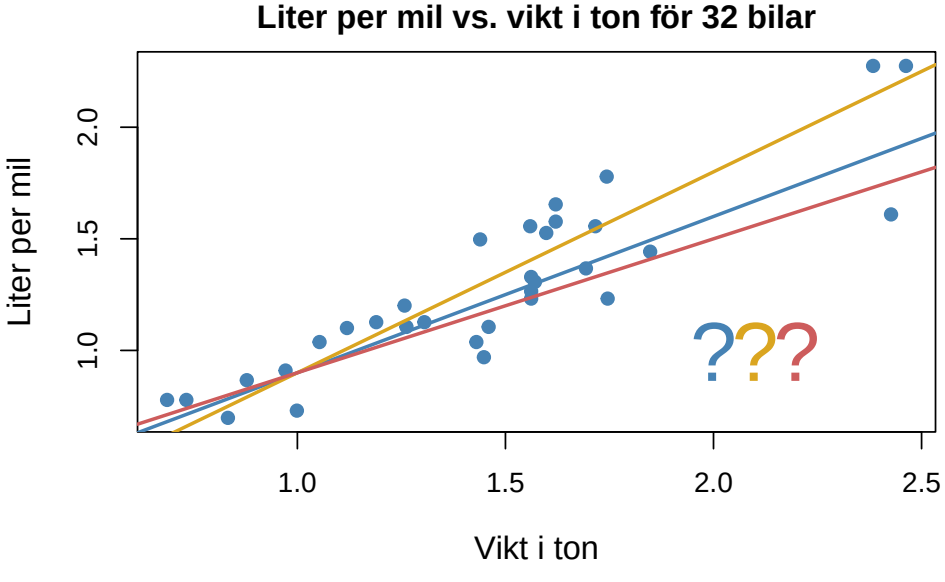


Residualer och prediktionsfel

- I formeln $e = y - \hat{y}$ står e alltså för residualen – e som i *error*
- Residualen för en observation är ett mått på hur stort fel modellen gör när den skattar just den observationen
- Ju mindre residualerna är desto bättre fångar modellen observationerna i vårt datamaterial
- Vårt mål är att hitta en regressionslinje med så små residualer (dvs fel) som möjligt
- Vi kan mäta den sammanlagda storleken på prediktionsfelen med summan

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$$

- Summan ovan används ofta för att bestämma **vilken linje** som bör användas i en regressionsmodell


Liter per mil vs. vikt i ton för 32 bilar
Liter per mil
1.0
1.5
2.0
1.0
1.5
2.0
2.5
Vikt i ton
???

- Vi betraktar den räta linje som minimerar $\sum_i e_i^2$ som den **bästa** regressionslinjen
- Eftersom vi minimerar de **kvadrerade** prediktionsfelen, kallas denna skattningsmetod för **minsta kvadratmetoden** (*method of least squares*)

- Vi vet sedan tidigare att ekvationen för en regressionslinje ges av

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

- Den linje som minimerar kvadratsumman $\sum_i e_i^2$ har de specifika koefficientvärdena

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

- Vi känner igen korrelationskoefficienten (r), standardavvikelserna (s_x och s_y), och medelvärdena ($\bar{x}$ och $\bar{y}$) sedan tidigare!

- Vi kan beräkna de givna koefficienterna för hand (det kommer ni göra på räkneövningar!), men just nu tittar vi på beräkningar m.h.a. `R`
- Vi vill veta hur bränsleförbrukningen (`litermil`) samvarierar med vikt (`viktton`)

```
litermil <- mtcars$litermil
viktton <- mtcars$viktton
```

- Notera att jag har räknat om `mtcars` till det metriska systemet (originalet är i amerikanska enheter)
- Vi kan göra beräkningarna på två sätt, dels med formlerna från tidigare, dels med funktionen `lm()` – vi testar båda här!

- Vi vill beräkna

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

- Vi kan bryta upp allt i delar, och beräkna dem enskilt innan vi pusslar ihop

```
r <- cor(litermil, viktton) # Korrelation

sx <- sd(viktton)  # Standardavvikelse x
sy <- sd(litermil) # Standardavvikelse y

xbar <- mean(viktton)  # Medelvärde x
ybar <- mean(litermil) # Medelvärde y

print(c(r, sx, sy, xbar, ybar)) # Skriver ut värden
## [1] 0.8898927 0.4442197 0.3885382 1.4606315 1.2828131
```

- Vi kan nu gå rakt på formlerna

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

```
b1 <- r * (sy / sx)          # Börjar med b1 (behövs till b0)
b1                           # Skriver ut värde
## [1] 0.7783476

b0 <- ybar - b1 * xbar       # Beräknar b0
b0                           # Skriver ut värde
## [1] 0.1459341
```

- Vår modell är alltså

$$\widehat{\text{litermil}} = 0.145934 + 0.778348 \cdot \text{vikt}$$

- Notera: Vi kan beräkna r, s_x och s_y med egna formler istället för `cor()` och `sd()`

Enkel linjär regression - beräkning med lm()

- Vi kan också använda `lm()`, som är en inbygd funktion i `R`
- `lm` står för *linear model*, och det är säkrast att använda den i verkliga tillämpningar (för att minimera risk för fel etc)

```
my_regressionmodel <- lm(litermil ~ viktton, data=mtcars) # Skattar modell

summary_model <- summary(my_regressionmodel)   # Skapar sammanfattning

summary_model$coefficients                                  # Tar fram b0 och b1
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.1459341 0.11106493  1.313953 1.988214e-01
## viktton     0.7783476 0.07284538 10.684928 9.565824e-12
```

- Modellparametrarna återges i kolumnen `"Estimate"`, och vi har att
  - `(Intercept)` är vårt b_0,
  - `viktton` är vårt b_1 (b_1 hör till den variabeln `viktton`)

Slump och signifikans

P-värde i regressionsoutput

- Vi tittar närmare på regressionsutskriften från `R`

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.1459341 0.11106493  1.313953 1.988214e-01
## viktton     0.7783476 0.07284538 10.684928 9.565824e-12
```

- Kolumnen längst till höger – med titel `Pr(>|t|)` – innehåller **p-värden**, som har med **statistiskt signifikans** att göra
- Det är vanligt att betrakta en koefficientskattning som statistiskt signifikant om p-värdet är mindre än 0.05, men det är upp till bedömaren/kontexten var gränsen ska gå
- Vi kan använda signifikans för att fördjupa de diskussioner vi tidigare har haft för att avgöra om samband är slumpmässiga eller ej

- Vi kommer inte räkna på p-värden och signifikans för hand i den här delkursen (det kommer på Del 2), men vi ska titta lite närmare på signifikans som koncept
- Mer specifikt ska vi titta närmare på signifikansen i lutningskoefficenten b_1 (vi struntar vanligtvis i b_0, då den sällan är av särskilt intresse)
- Notera först att vårt datamaterial `mtcars` bara innehåller data om 32 bilmodeller
- Om vi hade haft 32 andra bilmodeller i våra data hade regressionslinjen troligtvis sett lite annorlunda ut
- Mattias Villini har tagit fram en widget (länk) som låter er titta mer på hur sådana ändringar kan uttrycka sig

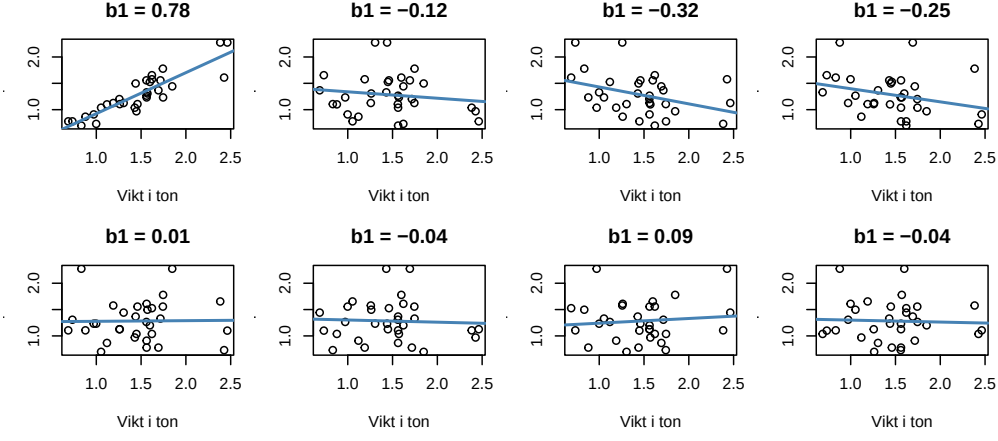
Slump och signifikans

- Vi säger att parametern b_1 är **signifikant** om vi drar slutsatsen att sambandet vi har hittat i våra data gäller generellt för **alla** bilmodeller
- Med andra ord drar vi slutsatsen att sambandet inte beror på **slumpen**, ungefär som vi har diskuterat tidigare
- När vi säger att ett samband **inte är signifikant** menar vi att sambandet i vår data mycket väl kan vara **orsakat av slumpen**
- När vi säger att ett samband **är signifikant** menar vi att sambandet i vår data med största sannolikhet ***inte* är ett resultat av slumpen**
- **Men:** Ett samband som är signifikant i statistisk mening behöver inte nödvändigtvis vara *betydelsefullt* i mer allmän bemärkelse!
- Det är vanligt att skilja på **statistisk signifikans** och **praktisk signifikans**, då det alltid är kontexten som "sätter värdet" på statistisk signifikans
- Inom medicin brukar man t.ex. prata mer konkret om **klinisk signifikans**
  - Vi kan t.ex. tänka oss att effekten av en behandling är statistiskt signifikant, men att effekten är såpass liten att läkare bedömer den som obetydlig

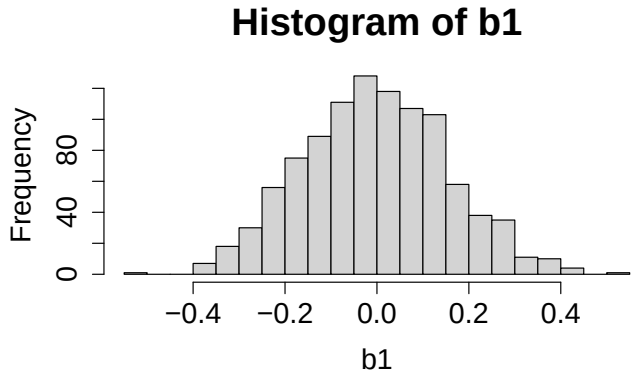
Slump och signifikans

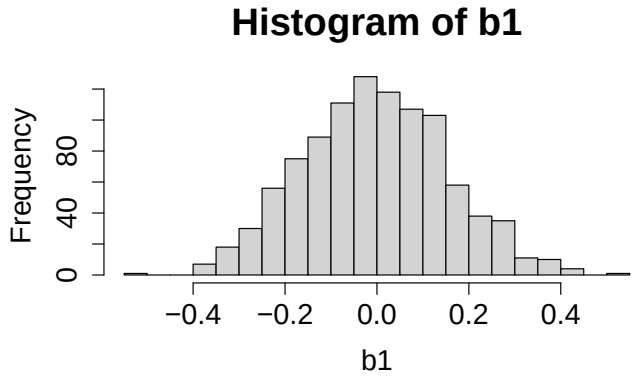
- Vi har redan sett att p-värdet för b_1 är väldigt litet i vårt exempel med bilars vikt och bensinförbrukning, och vi säger därför att koefficienten är signifikant (skild från 0)
- Vi bedömer det som troligt att det samband vi ser i våra data inte bara är orsakat av slumpen, och att de mycket väl kan finnas generellt
- Vi ska nu illustrera ett lite annorlunda sätt att tänka kring signifikans
- Vi utgår ifrån tanken att det **inte** finns någon samband mellan bensinförbrukning och vikt
- Detta skulle medföra att vilken bensinförbrukning som helst kan vara kopplad till vilken vikt som helst
- Vi ska visa några grafer för hur det slumpvisa utfallet kan bli om det saknas samband mellan vikt och bensinförbrukning (se nästa slide)

- Den första figuren visar vårt ursprungliga samband, och de övriga figurerna visar hur det slumpvisa sambandet kan se ut om det saknas det helt saknas samband mellan variablerna



- Under hypotesen om att variablerna inte har någon samband kan vi slumpa fram 1 000 datamaterial – alla med 32 bilar var
- För varje datamaterial kan vi beräkna lutningskoefficienten b_1, och jämföra med vår riktiga observation
- Histogrammet visar hur de 1000 lutningskoefficienterna b_1 fördelar sig



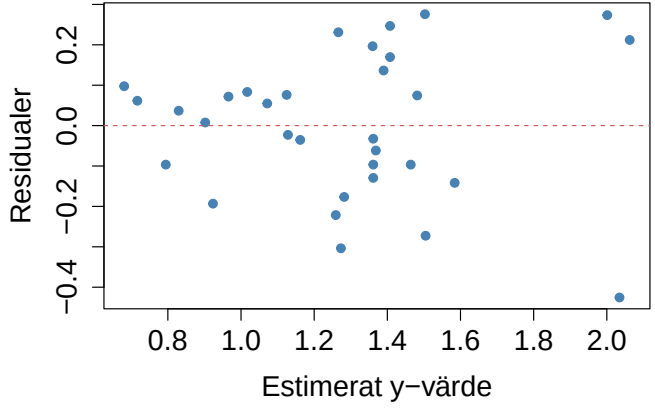


- Vi ser att **ingen** av de slumpvisa koefficienterna är så stor som 0.78
- Slutsatsen blir att slumpen troligtvis inte orsakade det samband som vi såg i våra data, och därför betraktar vi b_1 som signifikant

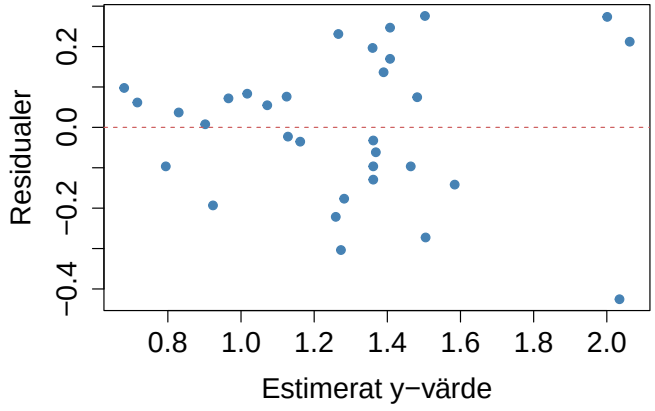
Modellantaganden och residualanalys

- Vi har redan introducerat residualerna, och beskrivit hur de används för att hitta den bästa regressionslinjen (linjen som minimerar $\sum_i e_i^2$)
- Nu återvänder vi till residualerna, och går igenom hur de kan användas för att utvärdera våra **modellantaganden**
- Modellantaganden är antaganden som måste vara uppfyllda för att våra resultat ska vara tillförlitliga

- Statistiska modeller bygger ofta på antaganden, regressionsanalys likaså
- När vi använder regressionsanalys gör vi följande två antaganden:
  1. Residualerna är **normalfördelade**
  2. Residualernas **varians är konstant**
- Dessa antaganden är främst relevanta när vi uttrycka osäkerheten i våra skattade värden
- Vi ska inte göra sådana beräkningar här (de kommer på Del 2), men det är bra att redan nu skapa vanan att kontrollera residualernas mönster
- Vi använder framför allt två typer av grafer för residualanalysen
  - Residualgrafer (residual plots)
  - Normalfördelningsgrafer



- Figuren visar en **residualgraf** för vår regressionsmodellmodell, där x-axeln visar våra $\widehat{y}$, och y-axeln visar våra residualer e
- Om residualplotten inte visar något tydligt mönster, utan ser slumpmässig ut, så är det ett gott tecken på att modellen fångar upp sambandet bra

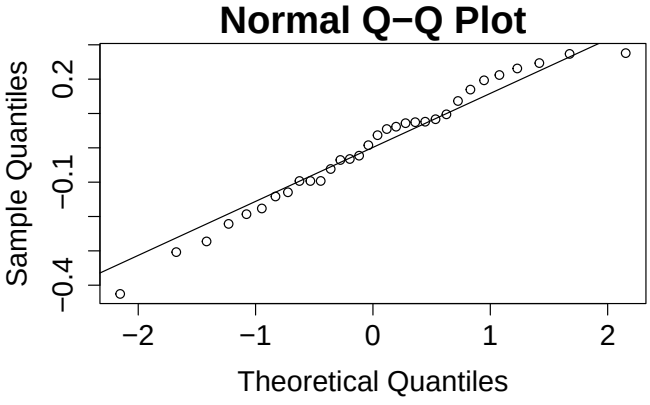


- Om vi ser tydliga **outliers** bland residualerna så representerar de observationer som inte passar väl in i modellen
- Det kan finnas anledning att studera dessa observationer för att förstå varför de avviker

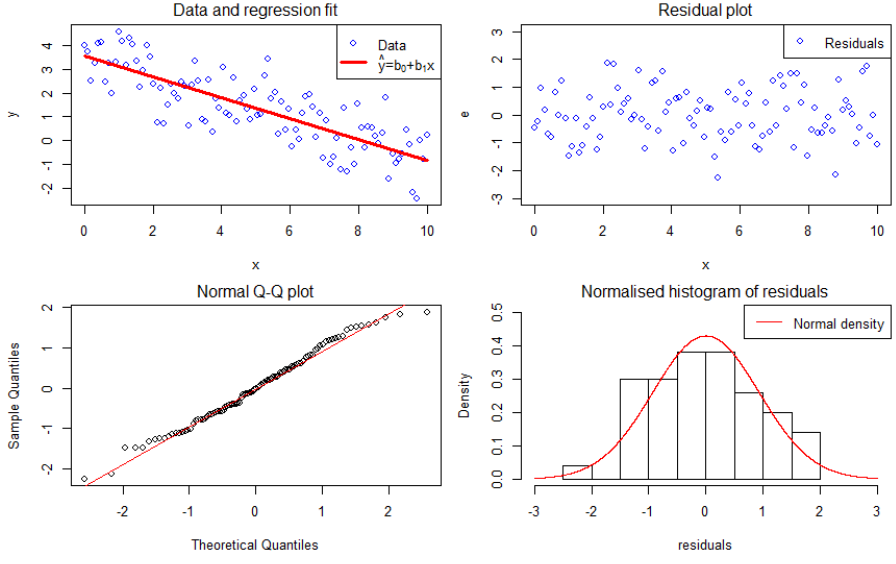
Sammanfattning av diskussion om residualgrafer

"A scatterplot of the residuals vs. the x-values should be the most boring scatterplot you've ever seen. It shouldn't have any interesting fatures, like direction or shape."

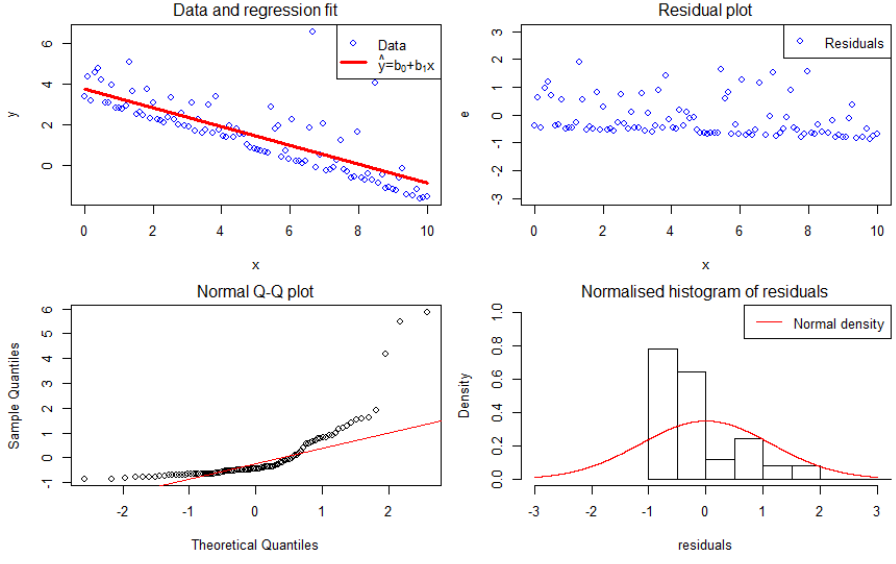
De Veaux et al (2021), s. 238



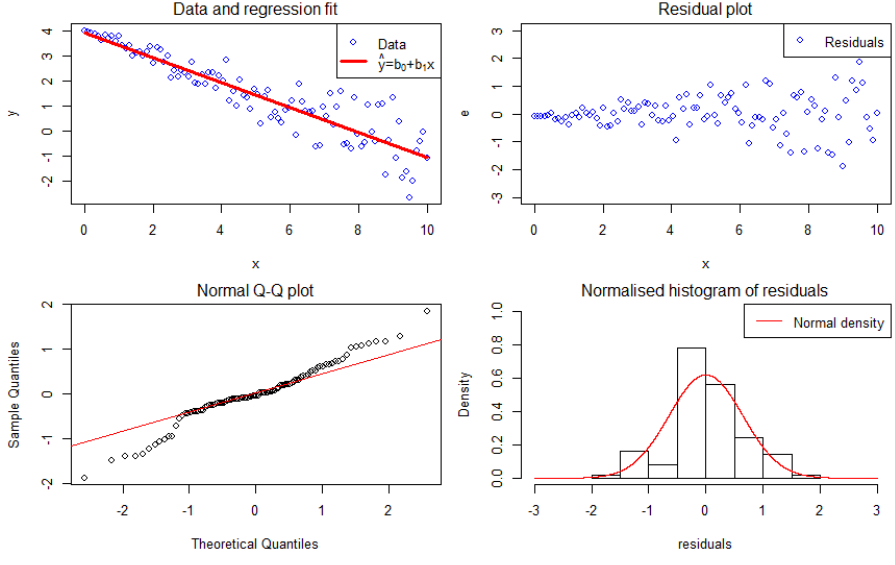
- På den här bilden ser vi en **normalfördelninggraf** för samma modell
- Om residualerna är normalfördelade ska punkterna följa linjen på ett ungefär, vilket de i det här fallet gör


Data and regression fit
Data
ŷ=b0+b1x
y
x
Residual plot
Residuals
e
x
Normal Q-Q plot
Sample Quantiles
Theoretical Quantiles
Normalised histogram of residuals
Normal density
Density
residuals

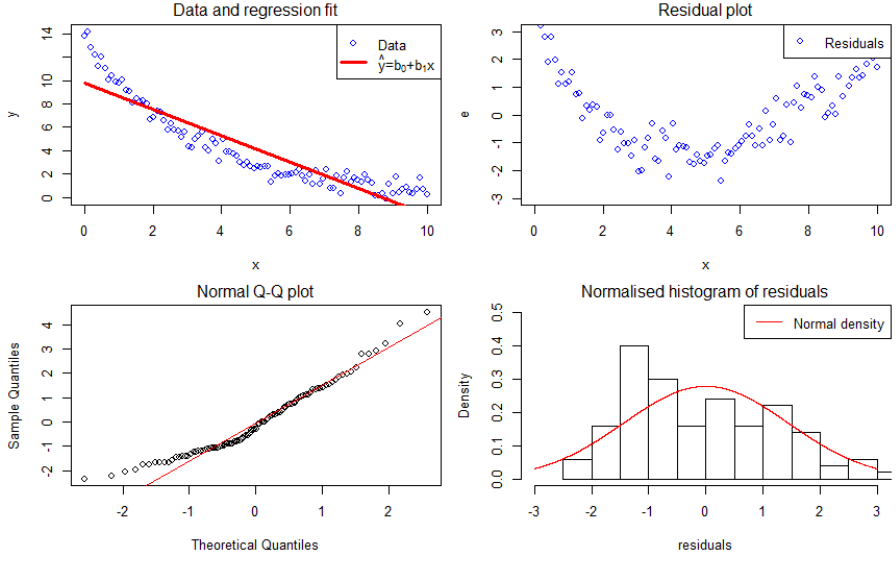
Residualer som inte är normalfördelade



Residualer som inte har konstant varians

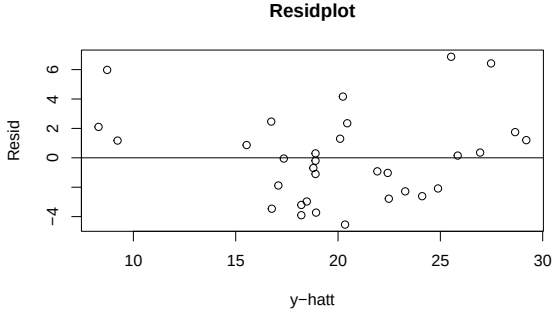


Residualer som inte har slumpmässigt mönster

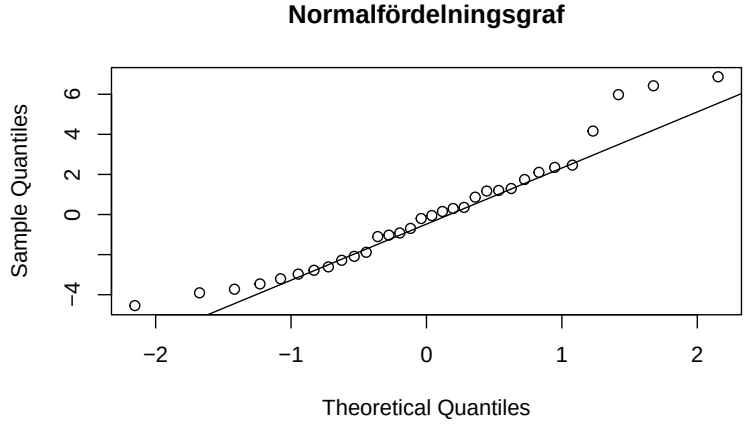


Kodexempel – skapa residualplot i R

```
lmod2 <- lm(mpg ~ wt, data=mtcars) #Skapa en modell

mtcars$res <- resid(lmod2) # Skapa en vektor med residualerna

mtcars$y_hatt <- fitted(lmod2) # Estimerade y-värden

plot(mtcars$res ~ mtcars$y_hatt, ylab="Resid", xlab="y-hatt", main="Residplot")

abline(h=0) #Dra en linje genom residualgrafen vid 0
```



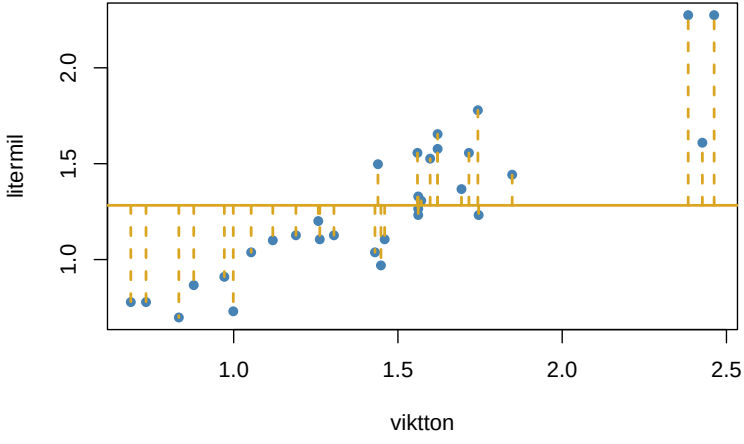
```r
qqnorm(resid(lmod2), main="Normalfördelningsgraf") #Rita graf

qqline(resid(lmod2)) #Lägg till en linje i grafen
```



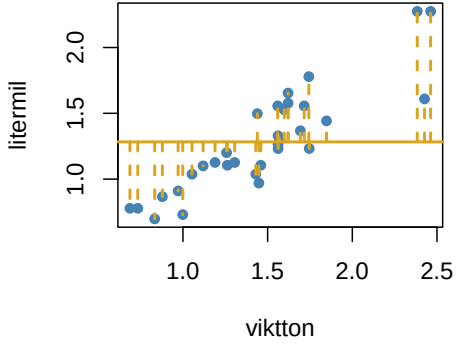
Förklaringsgraden R^2

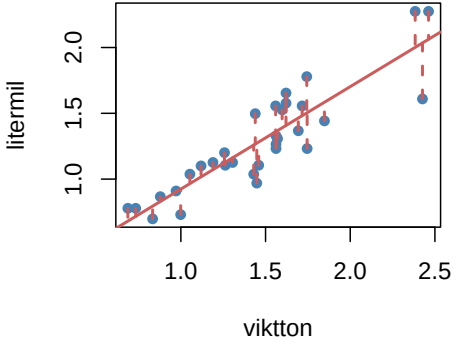
- R^2, som uttalas **R-kvadrat/R-två** (*en: R-squared*), kan användas som ett mått på hur bra en regressionsmodell är
- R^2 berättar hur väl modellen **förklarar variationen** i responsvariabeln
- För att förstå vad som menas med detta kan vi utgå ifrån den **enklaste tänkbara regressionsmodellen** som inte innehåller någon förklaringsvariabel alls
- Den enklaste sådana modellen innehåller bara interceptet, alltså $\hat{y} = b_0$
- I en regressionsmodell med bara interceptet blir $b_0 = \bar{y}$, dvs skattningen blir alltid detsamma som variabelns medelvärde

- Modellen $\hat{y} = b_0$ representeras av den gula regressionslinjen i bilden
- Modellen gör samma prediktion för alla bilar, och ger därmed **ingen** förklaring till varför förbrukningen varierar mellan modeller



- Vi lägger nu till `viktton` som förklarande variabel, och får modellen $\hat{y} = b_0 + b_1 x$ (höger bild)
- Modellen till höger **förklarar en stor del av variationen** med att högre vikt är associerat med högre bensinförbrukning
- Den variation som fortfarande är oförklarad illustreras med streckade linjer

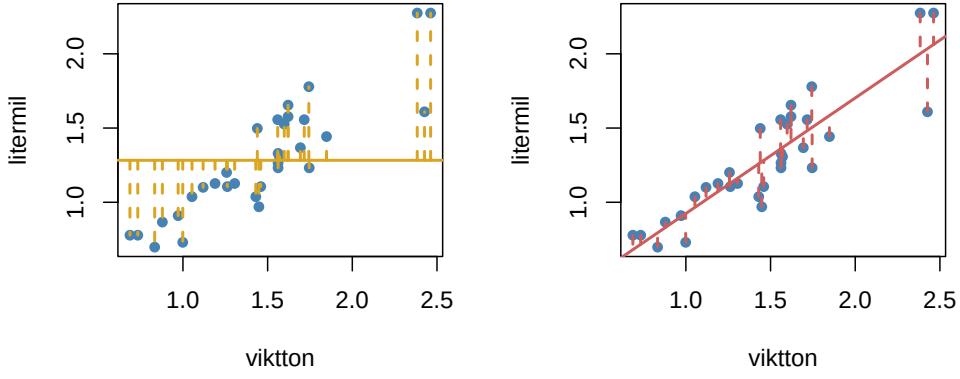




- R^2 anger hur stor **andel** av variationen i y som en modell förklarar
- Som exempel betyder alltså $R^2 = 0.6$ att modellen förklarar 60% av den totala variationen i y
- För att räkna ut R^2 för en modell behöver vi två nya begrepp
  - **SST (Sum of Squares Total):** Den totala variationen i responsvariabeln
  - **SSE (Sum of Squares Error):** Den variation som inte förklaras av modellen

- Vi kan räkna ut SST och SSE med formlerna

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad \text{och} \quad \text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y})^2$$



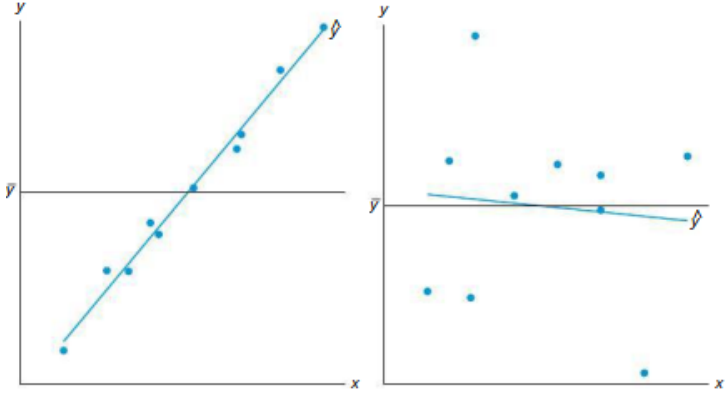
- När vi har räknat ut SST och SSE kan vi räkna ut R^2 som

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

- Kvoten SSE/SST är andelen av den totala variationen som modellen lämnar *oförklarad*
- $R^2 = 1 - SSE/SST$ är därför den andel som *är* förklarad
- $SSE \leq SST$, vilket innebär att $0 \leq R^2 \leq 1$
- **Notera:** I den enkla modellen är $\hat{y} = \bar{y}$, och SSE = SST, vilket ger $R^2 = 0$
- Om en modell har $R^2 = 1$ betyder det att alla observationer ligger exakt på regressionslinjen

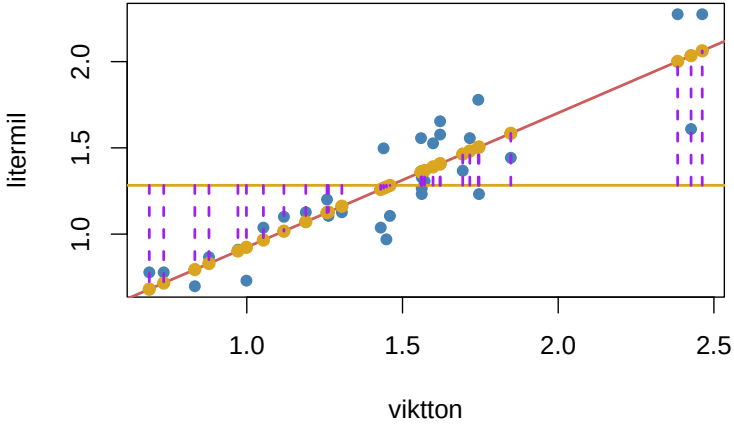
- I enkel linjär regression gäller att $R^2 = r^2$, där r är korrelationskoefficienten mellan y-variebeln och x-variabeln
- I multipel regression kan vi istället sätta r till korrelationen mellan y och $\hat{y}$
- Det går inte att säga vad som är ett bra värde på R^2 i allmänhet
- I vissa naturvetenskaper kan det förekomma att R^2 är nära 1
- I samhällsvetenskaper är det vanligt med modeller där R^2 är en bra bit under 0.5
- Det kan bero på att samhällsfenomen har många bidragande orsaker (inklusive slumpen), vilket gör det svårt att mäta/inkludera mer än ett fåtal av dem i en modell

- Nedan ser vi två figurer ur Walpole et al. (2016)
  - I figuren till vänster förklarar modellen nästan all variation
  - I figuren till höger förklara modellen nästan ingen del av variationen



(a) $R^2 \approx 1.0$ (b) $R^2 \approx 0$

- Vi ha talat om SST och SSE
- Ett tredje begrepp är **SSR (Sum of Squares Regression)**, som mäter variationen av $\hat{y}$ runt variabelns medelvärde $\bar{y}$
- Detta blir alltså ett mått på variationen som förklaras av modellens variabler



- Vi räknar ut SSR med formeln

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

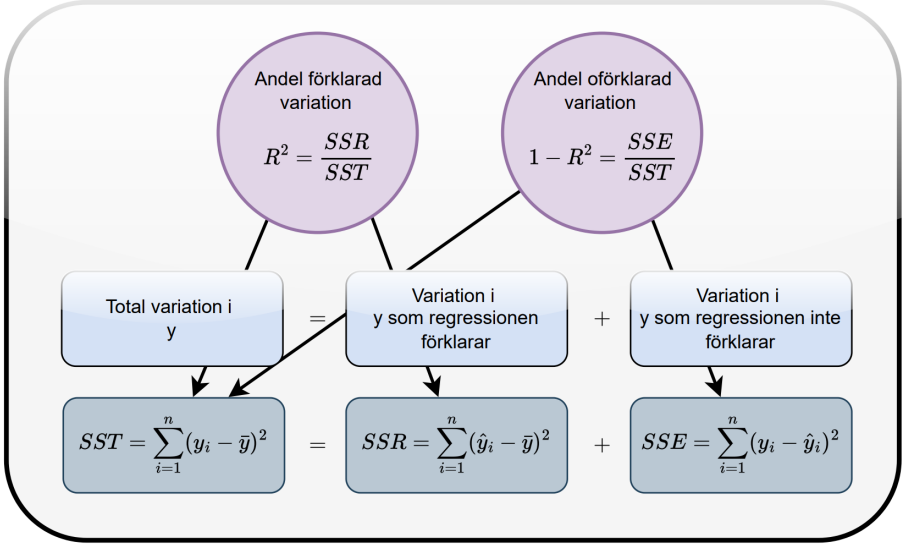
- SST, SSR och SSE förhåller sig till varandra på så sätt att

$$SST = SSR + SSE$$

- Vi kan därför även beräkna R^2 som

$$R^2 = \frac{SSR}{SST}$$

- Att dela upp variationen på det här sättet kallas **analysis of variance (ANOVA)**


Andel förklarad variation
$R^2 = \frac{SSR}{SST}$
Andel oförklarad variation
$1 - R^2 = \frac{SSE}{SST}$
Total variation i y
=
Variation i y som regressionen förklarar
+
Variation i y som regressionen inte förklarar
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
=
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
+
$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- Nu ska vi räkna ut R^2 för vår modell som estimerar responsvariabeln bensinförbrukning med förklaringsavariabeln vikt
- Först räknar vi ut SST, SSR och SSE

```
y <- litermil     # Utfall
y_bar <- mean(y)  # Medelvärde


lmod <- lm(litermil~viktton, data=mtcars) # Modell
y_hatt <- fitted(lmod)                    # Prediktioner


SST <- sum((y - y_bar)^2)
SSR <- sum((y_hatt - y_bar)^2)
SSE <- sum((y - y_hatt)^2)

print(c(SST, SSR, SSE))
## [1] 4.6798209 3.7059923 0.9738286
```

- Vi såg att $SST = 4.680$, och jämför med $SSR + SSE$ för att bekräfta $SSR + SSE = SST$

```
SSR + SSE
## [1] 4.679821
```

- Sedan räknar vi ut R^2 med formeln

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

```
R2 <- 1 - SSE/SST
R2
## [1] 0.791909
```

- Om vi använder `summary()` på vår modell så ser vi att `lm()` räknar ut R^2 per automatik!

```
summary(lmod)
##
## Call:
## lm(formula = litermil ~ viktton, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42543 -0.10480  0.02227  0.10729  0.27579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14593    0.11106   1.314    0.199
## viktton      0.77835    0.07285  10.685 9.57e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1802 on 30 degrees of freedom
## Multiple R-squared:  0.7919, Adjusted R-squared:  0.785
## F-statistic: 114.2 on 1 and 30 DF,  p-value: 9.566e-12
```

Tack för idag!!