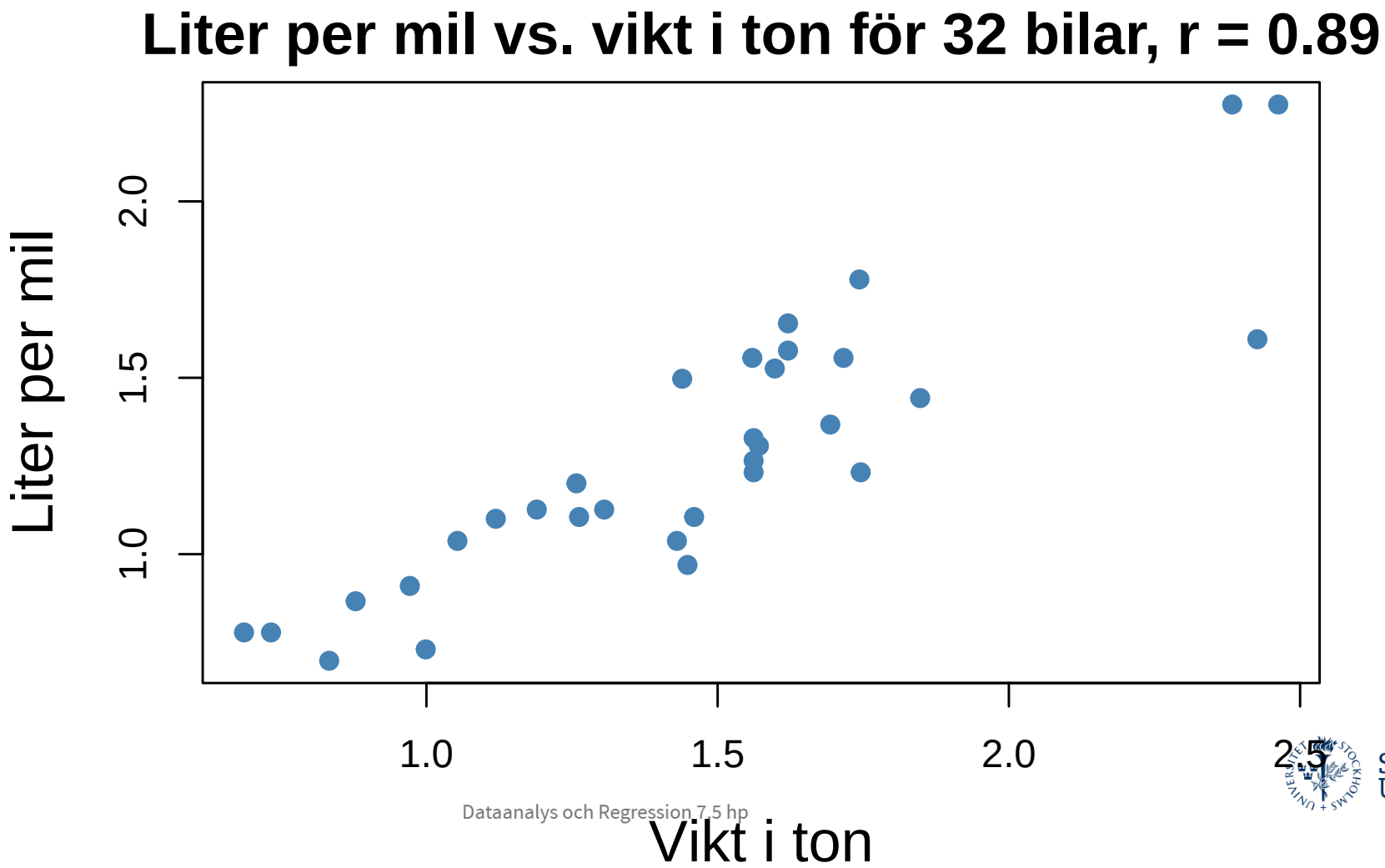
Lecture 7

karl.sigfrid@stat.su.se

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

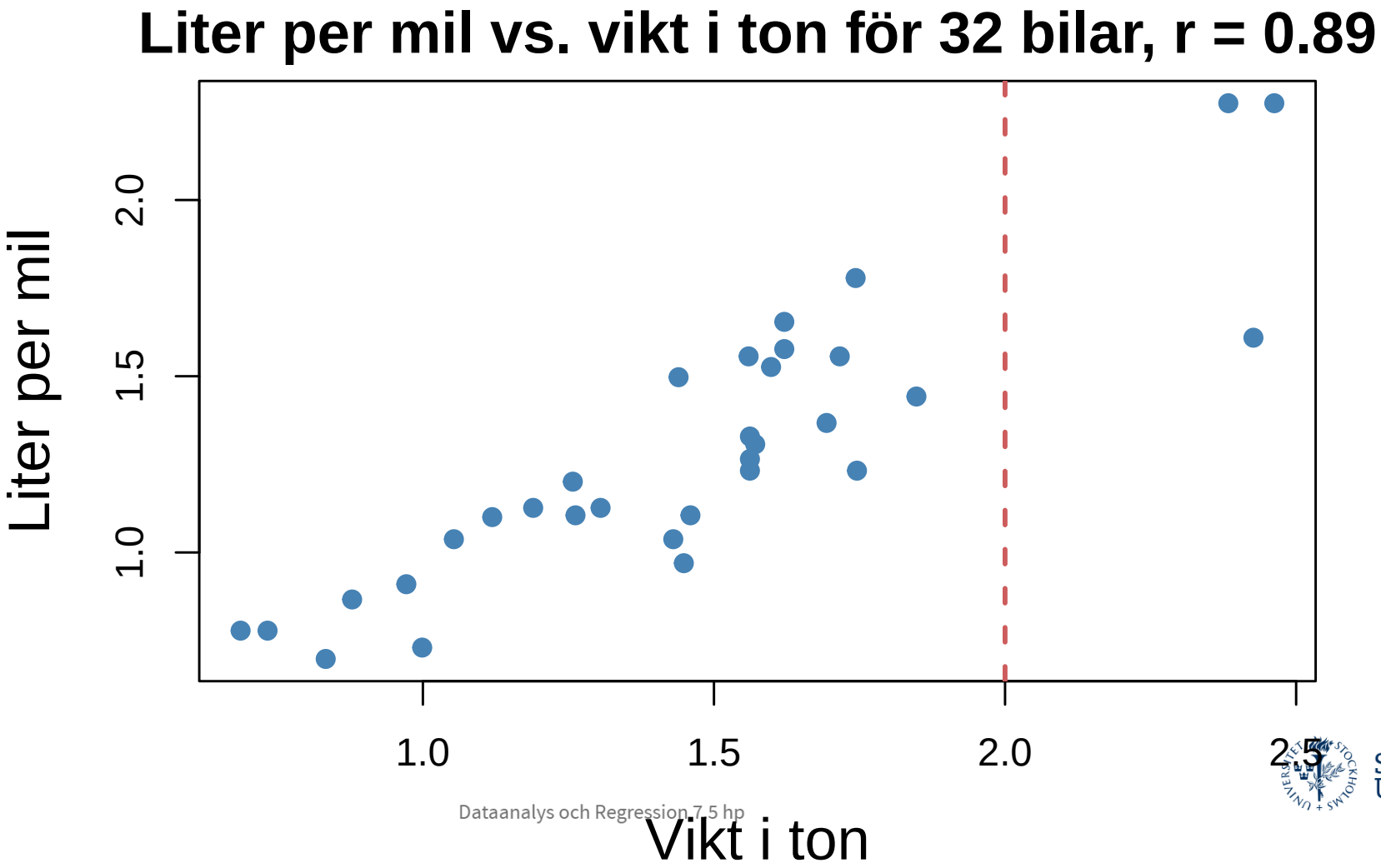
- Förra föreläsningen handlade om linjära samband och **korrelationkoefficienten**.
- Här ser vi sambandet mellan bränsleförbrukning (liter/mil) och vikt (ton) för 32 bilmodeller. Sambandet ser linjärt ut.



Stockholm University

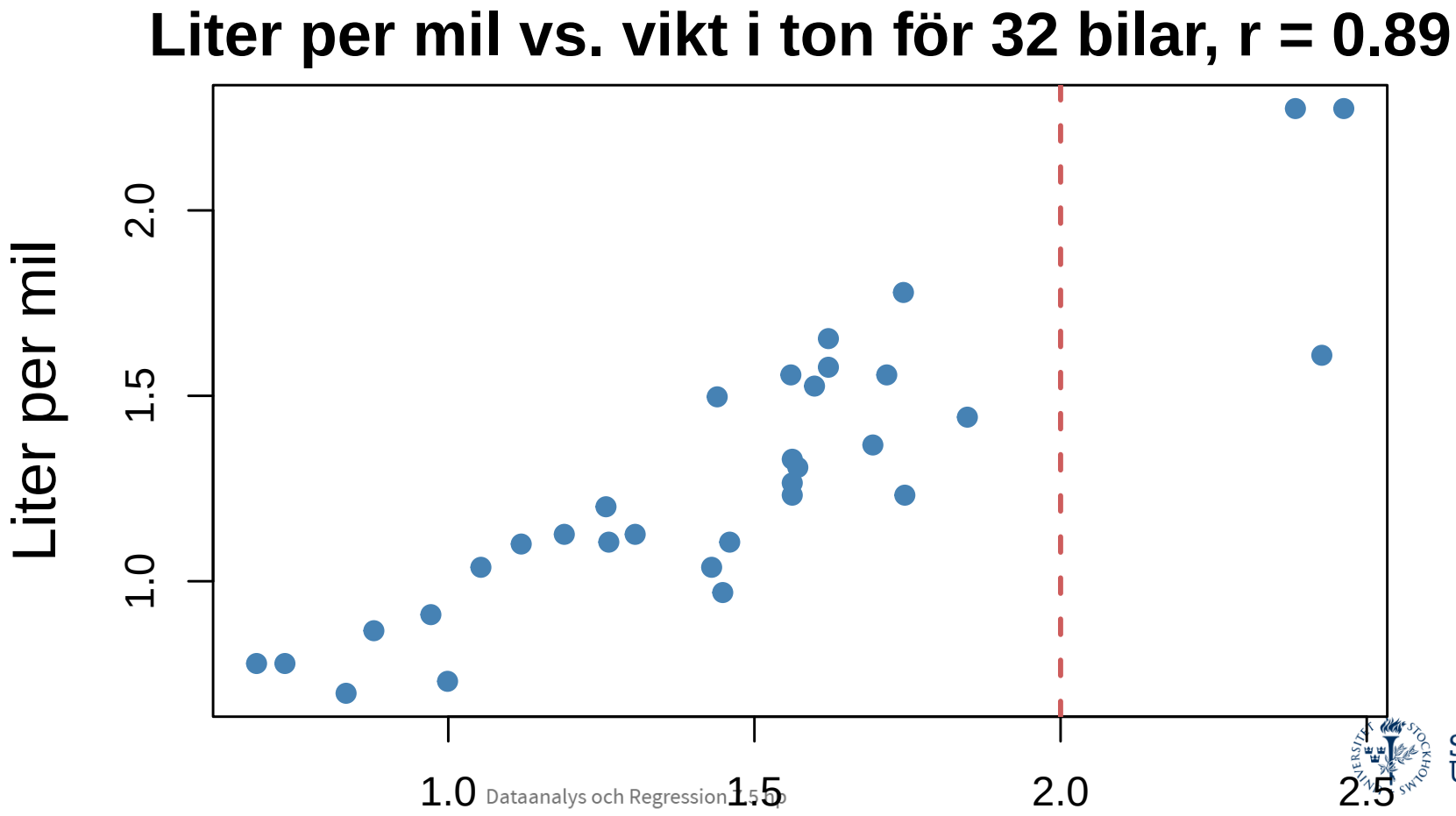
Uppskatta värdet av y när du känner till värdet av x

- Anta att vi är intresserade av bränsleförbrukningen hos en bil som *inte finns med i vårt dataset*. Vi vet dock att bilen *väger 2 ton*. Borde vi inte då kunna utnyttja det linjära sambandet för att göra en uppskattning av vad bränsleförbrukningen kan vara?



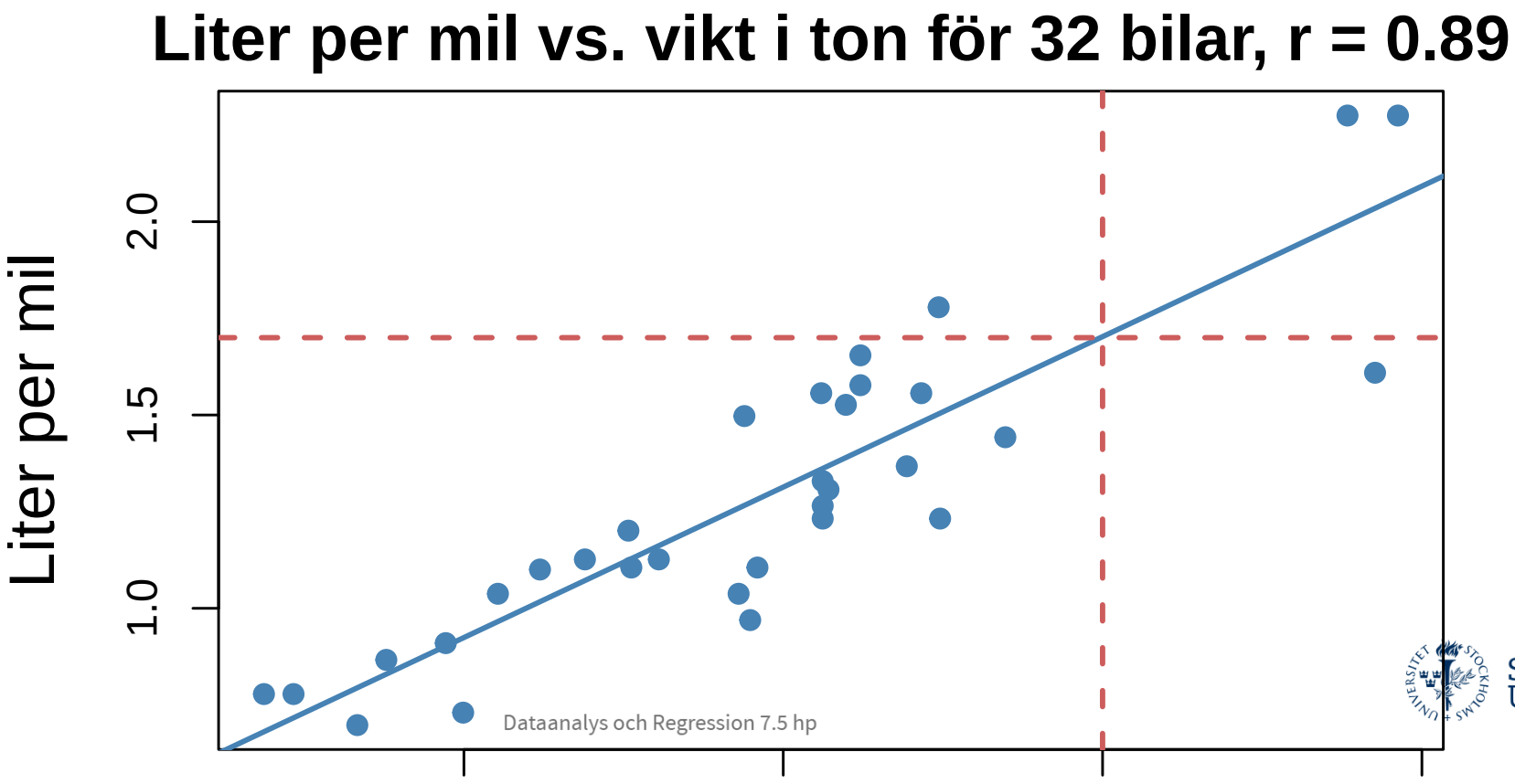
Uppskatta värdet av y när du känner till värdet av x

- För att passa in i mönstret i vårt spridningsdiagram borde en bil som väger 2 ton förbruka ungefär 1.5 - 2 ton.
- En sådan ungefärlig gissning kan i vissa fall vara tillräcklig, men hur gör vi om vi vill sätta **en** siffra på vår uppskattning av bensinförbrukningen.



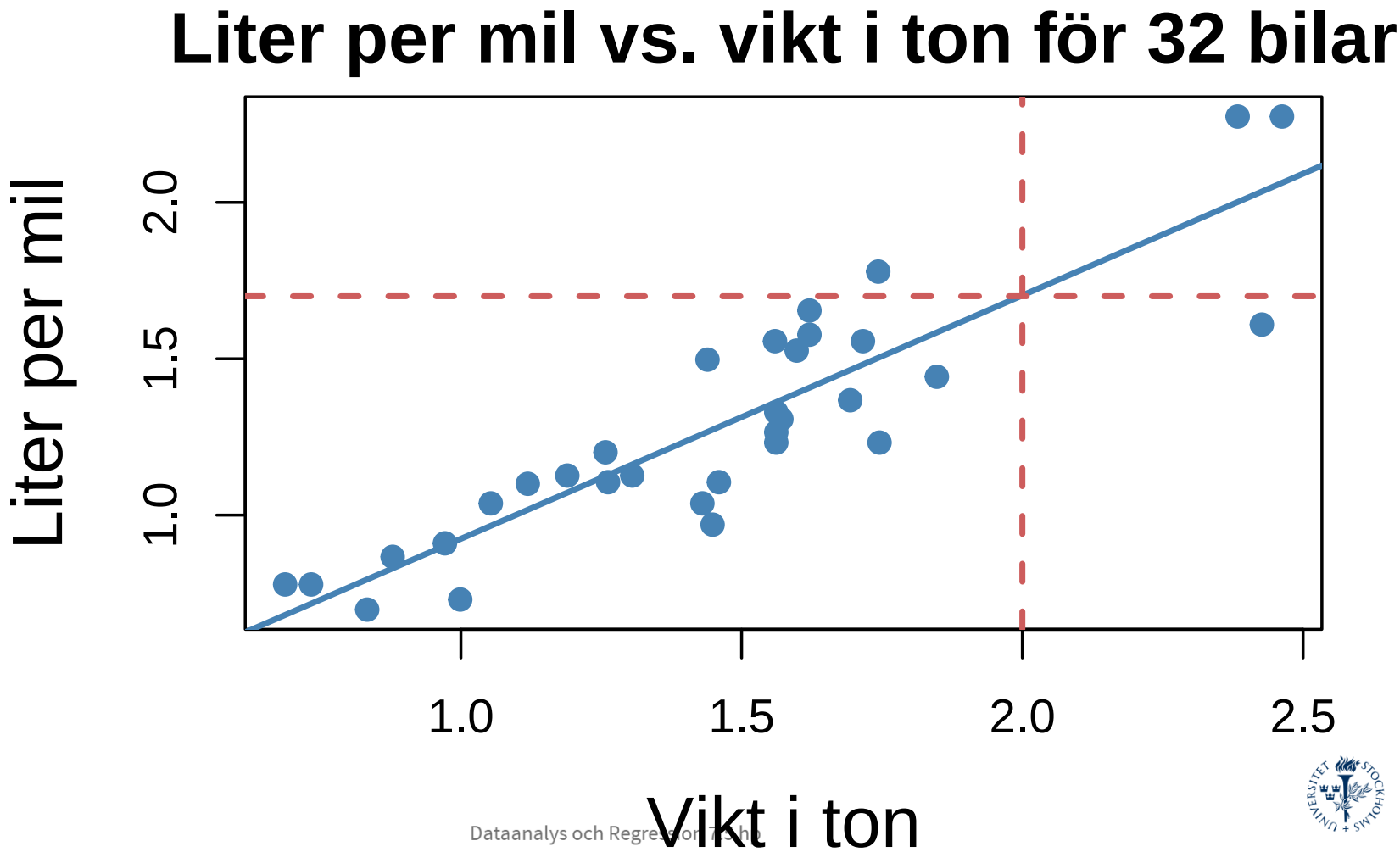
Uppskatta värdet av y när du känner till värdet av x

- En metod för att uppskatta bensinförbrukningen när vi känner till vikten på en bil kan vara att dra en linje rakt genom svärmen av punkter.
- Vi ser var vår linje är ungefär vid 1.7 på y-axeln när värdet på x-axeln är 2.
- Slutsats: Vi uppskattar att en bil som väger 2 ton drar 1.7 liter per mil.



Enkel linjär regression

- Det vi just gjorde var en **enkel linjär regression**.
- Syftet med enkel linjär regression är normalt att
  - **prediktera** värdet på y när vi känner till x.
  - undersöka hur sambandet mellan x och y ser ut.



Enkel linjär regression - liten ordlista, del 1

- Att **prediktera (predict)** är att göra en uppskattning av ett värde när vi inte kan göra en direkt observation (dvs när vi inte kan göra en mätning).
- Substantivformen av ordet är **prediktion**. Exempel: "Syftet med vår modell är att göra en prediktion av bensinförbrukningen."
- Vi kan också använda orden **estimera** eller **skatta** i ungefär samma mening som prediktera. Exempel: "Vi estimerar/skattar bensinförbrukningen till 1.7 liter per mil"
- Substantivformerna är **estimat** och **skattning**. Exempel: "Vårt estimat/Vår skattning är att bilen förbrukar 1.7 liter per mil."

Enkel linjär regression - liten ordlista, del 2

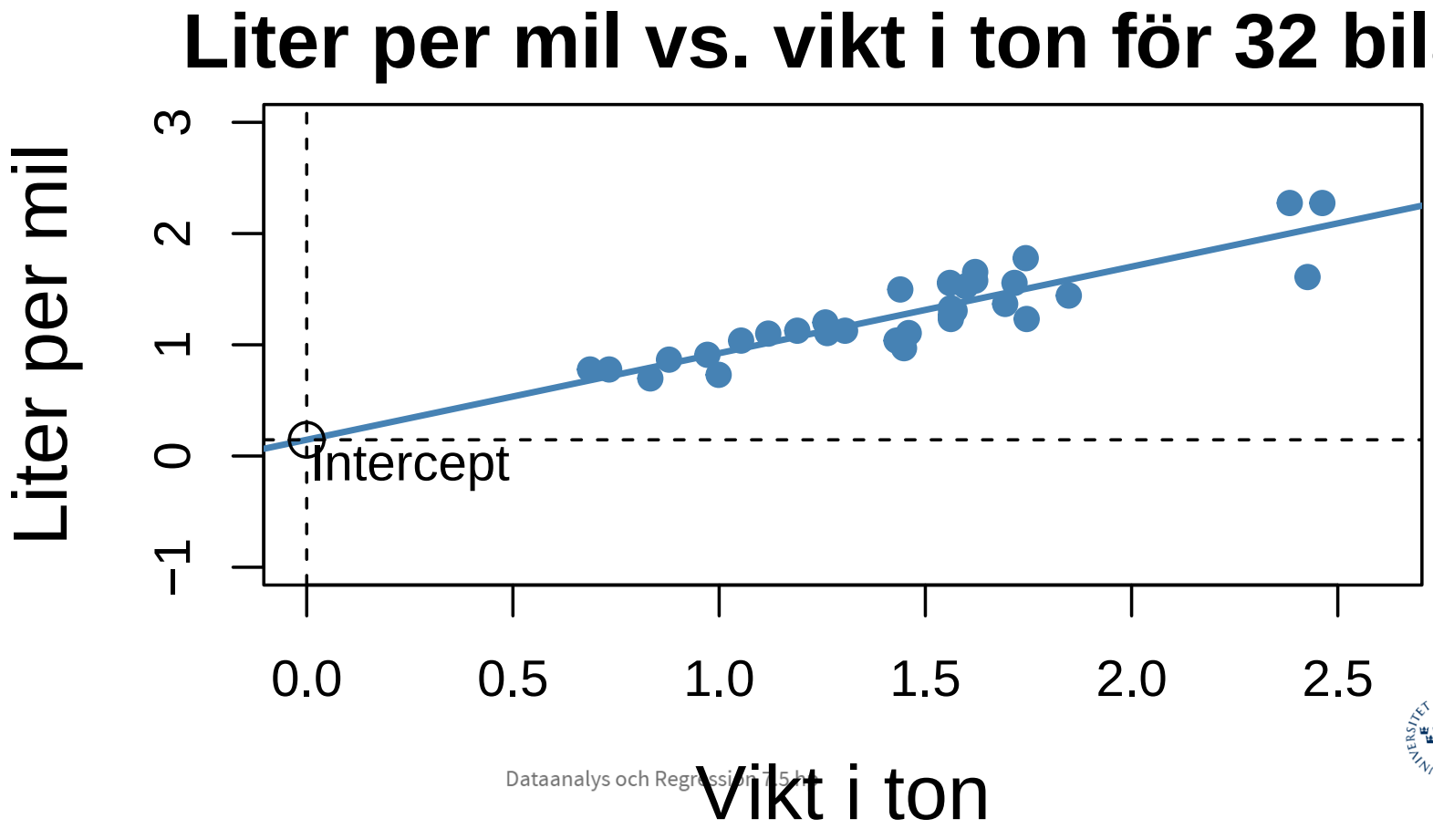
- Vår y-variabel kan kallas **responsvariabel**, eller den **beroende variabeln**.
- Vår x-variabel kan kallas **förklaringsvariabel**, eller den **oberoende variabeln**.
- **Enkel linjär regression (simple linear regression)** betyder att vi estimerar responsvariabeln med hjälp av *en enda* förklaringsvariabel.
- I **multipel linjär regression (multiple linear regression)** estimerar vi responsvariabeln med hjälp av flera förklaringsvariabler. (Det ska vi dock inte gå in på i dag.)

Enkel linjär regression

- Är enkel linjär regression inte mer komplicerat än att dra en linje rakt genom svärmen av observation och sedan läsa av vilket y-värde som mostsvarar ett visst x-värde längs linjen?
  - I princip fungerar det så.
  - I praktiken krävs matematiska beräkningar för att bestämma exakt hur regressionslinjen ska dras. Vi kommer att gå igenom hur de beräkningarna görs.
  - Vi kommer också att vilja göra andra beräkningar kopplade till regressionsmodellen, exempelvis för att utvärdera hur bra modellen är.

Enkel linjär regression - regressionslinjen

- En regressionslinje kan definieras med två **parametrar**:
  - **Interceptet** som är regressionslinjens y-värdet när $x = 0$.
  - **Lutningen (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.



Enkel linjär regression - regressionslinjen

En regressionslinje kan beskrivas matematiskt med formeln

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

där b_0 är interceptet och b_1 är lutningen.

- Känner vi till b_0, b_1 och x kan vi *estimera* y.
- Om vi vill kan vi välja att använda mer deskriptiva variabelnamn istället för x och y.

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

Enkel linjär regression - regressionslinjen

- Notera att vi har en hatt ovanför vårt y i formeln

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

- Att vi skriver $\hat{y}$ istället för y beror på att det handlar om ett **estimat** och inte om det verkliga värdet på y.
- I vårt första exempel estimerade vi att bränsleförbrukningen är 1.7 liter/mil för en bil som väger 2 ton. I det fallet kan vi skriva $\hat{y} = 1.7$.
- Om vi hade skrivit $y = 1.7$ hade vi hävdat att bilens **verkliga** bränsleförbrukning är 1.7 liter/mil. Det kan vi inte göra eftersom vårt estimat knappast är perfekt.

Enkel linjär regression - regressionslinjen

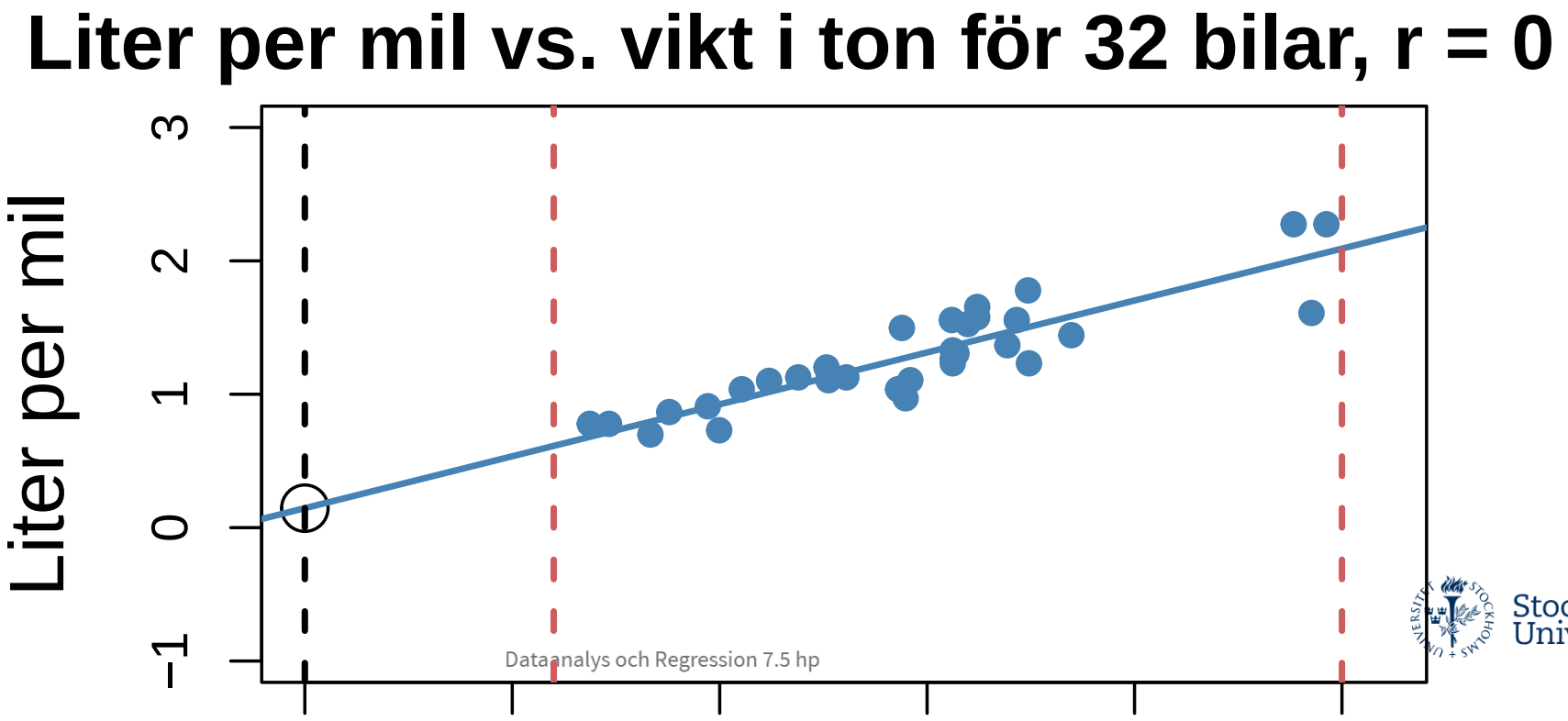
- Ett vanligt syfte med en regressionsmodell är att estimera värdet på responsvariabeln, så som vi precis gjorde.
- Vi kan också vara intresserade av att se sambandet mellan variablerna.
- Vi utgår från vår modell där vi estimerar bensinförbrukning hos en bil med hjälp av vikten. Den allmänna formeln, som vi redan har sett, kan skrivas som $\widehat{\text{litermil}} = b_0 + b_1 \text{vikt}$.
- Anta att $b_0 = 0.146$ och $b_1 = 0.78$. Det innebär att

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

- Parametervärdet $b_1 = 0.78$ betyder att en bil, enligt modellen, förbrukar ytterligare 0.78 liter/mil bensin för varje extra ton som bilen väger.

Enkel linjär regression - regressionslinjen

- Parametervärdet $b_0 = 0.146$ betyder att en bil som väger 0 ton enligt modellen förbrukar 0.146 liter per mil. Den informationen är förstås inte meningsfull, så var försiktig med att tolka interceptet bokstavligt.
- Var försiktig med att använda modellen för att prediktera värden **utanför** det intervall av x där våra datapunkter ligger. I vårt exempel är det vanskligt att använda modellen för bilar som väger mindre än 0.5 ton eller mer är 2.5 ton.



Enkel linjär regression - regressionslinjen

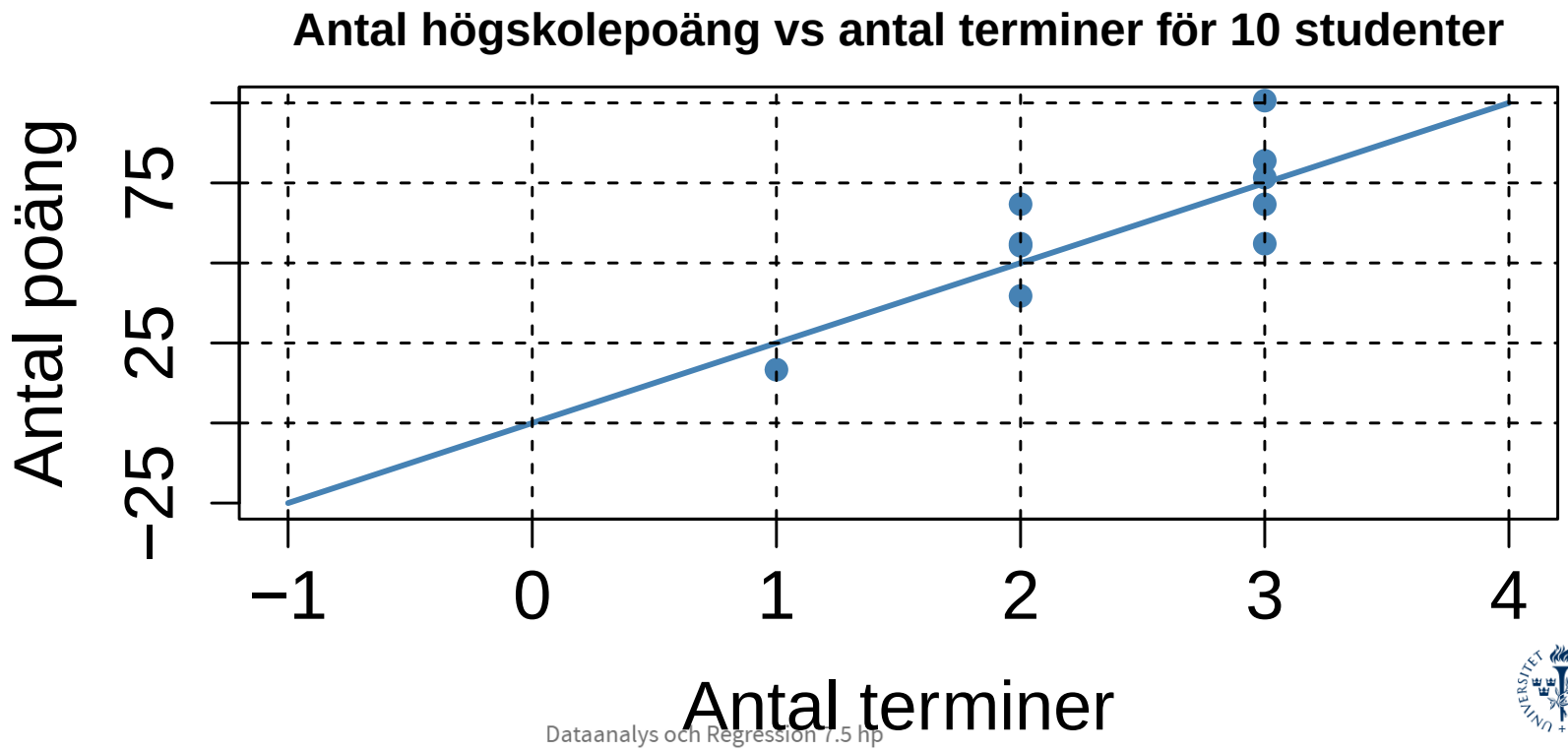
OBS!

- 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.
- Om $b_1 = 0.78$, som i vårt exempel, kan vi inte utan vidare säga att en viktökning med ett ton *medför* att bränsleförbrukningen ökar med 0.87 liter/mil.
- **Regressionsmodeller visar samband, inte kausalitet!**
- För att kunna hävda att ett ton i ökad vikt orsakar en viss ökning av bränsleförbrukningen måste andra vetenskapliga metoder användas.

Enkel linjär regression - regressionslinjen

En snabb övning: Bilden beskriver en enkel linjär regressionsmodell. Varje observation är en student. x-axeln visar antalet avslutade terminer och y-axeln antalet tagna poäng.

1. Vilka värden har parametrarna b_0 och b_1?
2. Hur ska parametrarna tolkas?
3. Hur många poäng estimerar modellen att en student har efter 2 terminer?

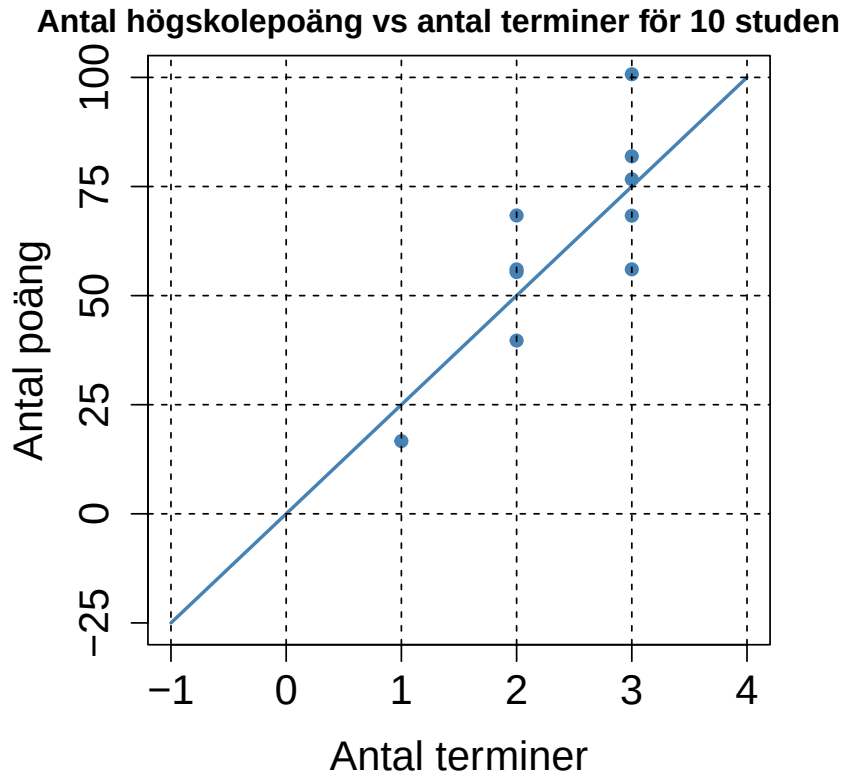


Stockholm University

Enkel linjär regression - regressionslinjen

Frågor

1. Vilka värden har parametrarna b_0 och b_1?
2. Hur ska parametrarna tolkas?
3. Hur många poäng estimerar modellen att en student har efter 2 terminer?



Enkel linjär regression - regressionslinjen

Svar

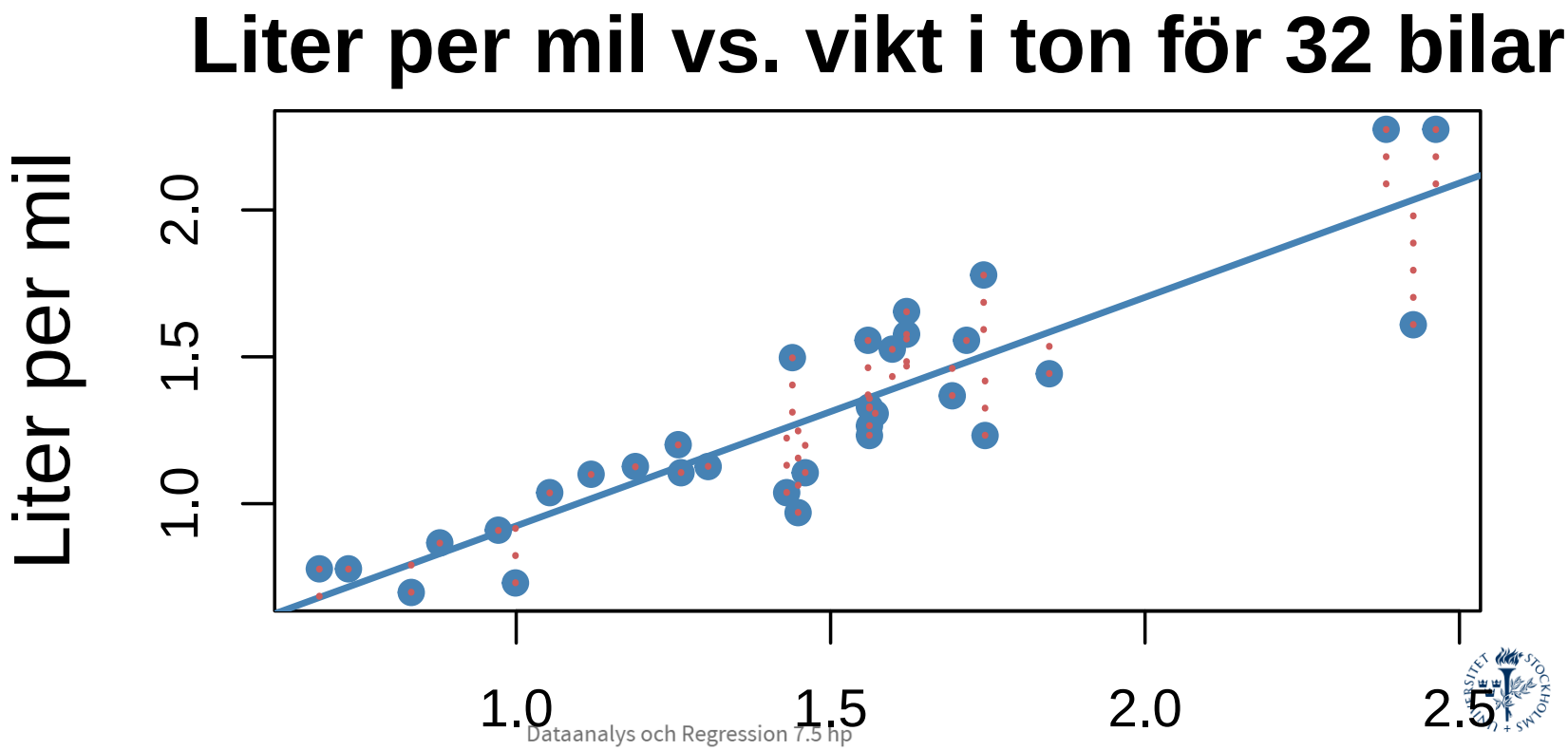
1. $\hat{y} = 0$ när $x = 0$, så $b_0 = 0$.$\hat{y}$ ökar med 25 poäng för varje termin som x ökar, så $b_1 = 25$.
2. b_1 kan tolkas som att antalet poäng ökar med 25 poäng per termin.b_0 kan tolkas som att en student som pluggat i 0 terminer har tagit 0 poäng.
3. Vår modell är $\widehat{\text{poäng}} = 0 + 25 \cdot \text{terminer}$.Om antalet terminer är 2 får vi alltså $\widehat{\text{poäng}} = 0 + 25 \cdot 2 = 50$, så modellen estimerar att en student tar 50 poäng på två terminer.

Regression mot medelvärdet

- **Regression mot medelvärdet (regression to the mean):** En observation med ett avvikande x-värde resulterar i en prediktion $\hat{y}$ som är mindre avvikande.
- Exempel: Anta att x befinner sig 2 standardavvikelser från $\bar{x}$. Då avviker $\hat{y}$ mindre än 2 standardavvikelser.
- Avståndet mellan $\hat{y}$ och $\bar{y}$ är alltså mindre än avståndet mellan x och $\bar{x}$, mätt i antalet standardavvikelser.
- Vi kommer inte att fördjupa oss i detta, men "regression to the mean" är ett begrepp som är bra att känna till betydelsen av.

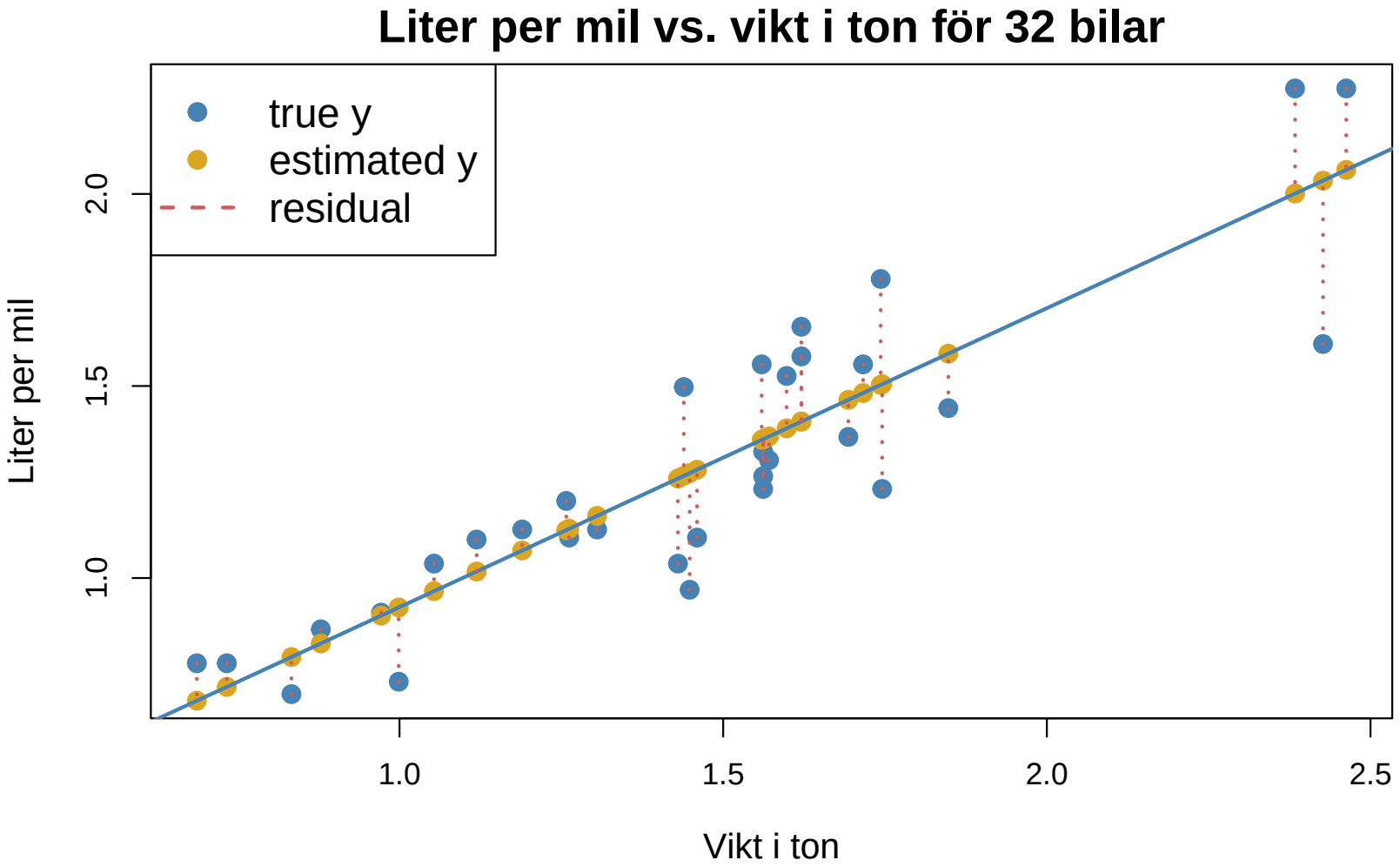
Enkel linjär regression - residualer

- För varje observation kan vi räkna ut ett estimat $\hat{y} = b_0 + b_1 x$.
- $\hat{y}$ är det värde som regressionslinjen har på y-axeln vid observationens x-värde. För varje observation har vi dessutom det *sanna* värdet y.
- Skillnaden $e = y - \hat{y}$ kallas **residualen (the residual)**. Varje observation har alltså en residual, som i bilden är markerad med en röd streckad linje. Residualerna kan vara positiva eller negativa.



Enkel linjär regression - residualer

Ju mindre residualerna är desto bättre passar modellen ihop med observationerna.



Stockholm University

Enkel linjär regression - residualer

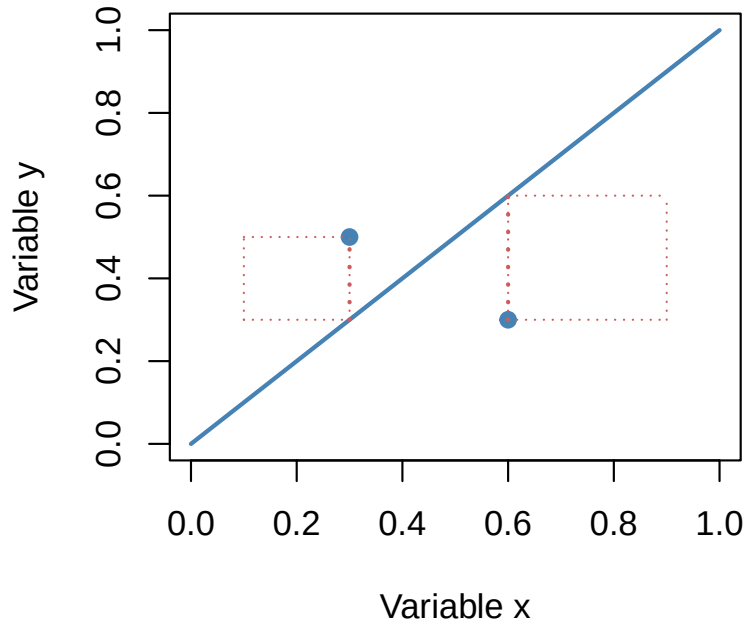
- I formeln $e = y - \hat{y}$ står e alltså för residualen. Det är e som i "error".
- Residualerna är ett mått på hur stort prediktionsfelet är för var och en av observationerna.
- Ju mindre residualerna är desto bättre fångar modellen observationerna i vårt datamaterial.
- Vårt mål är att hitta en regressionslinje med så små residualer som möjligt.
- Ett mått som vi kan använda för att mäta den sammanlagda storleken på prediktionsfelen är

$$\sum e^2$$

Enkel linjär regression - minsta kvadratmetoden

- Den räta linje som minimerar $\sum e^2$ betraktar vi som den bästa regressionslinjen.
- Metoden kallas**Minsta Kvadratmetoden(Least Squares Method)**
- Den linje som minimerar $\sum e^2$ kan beskrivas med formeln

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



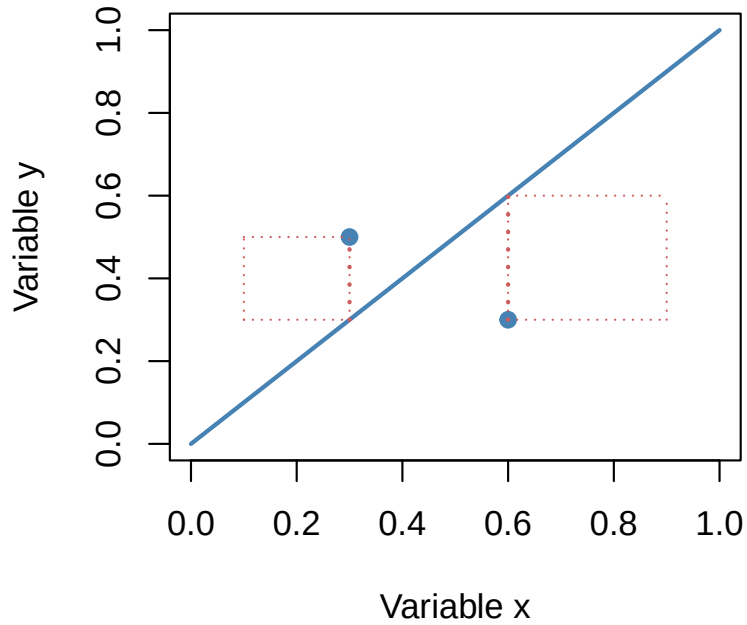
$$b_1 = r_{x,y}\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}$$

Enkel linjär regression - minsta kvadratmetoden

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

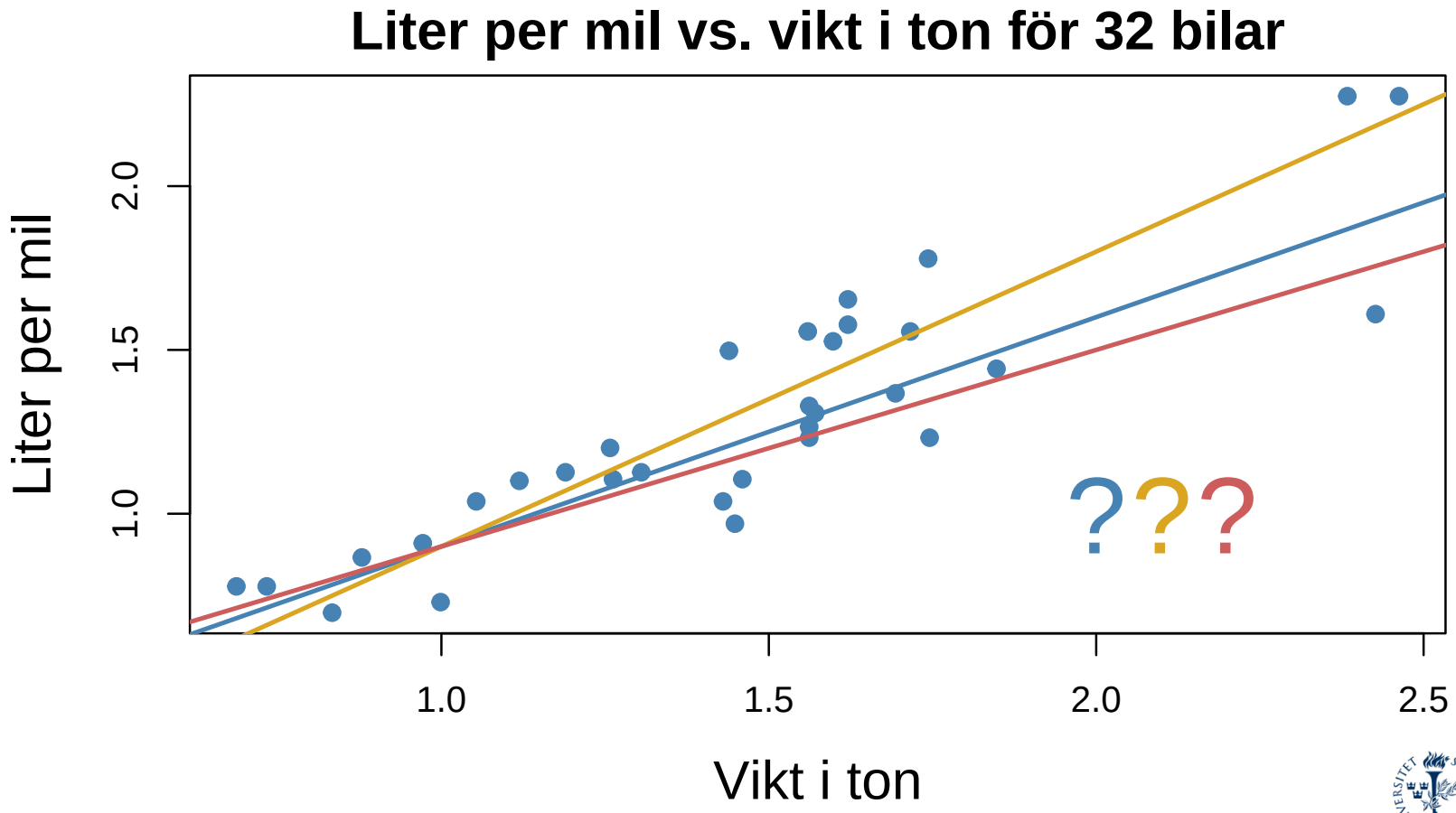
$$b_1 = r_{x,y}\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}$$

- Formlerna visar att vi kan räkna ut b_0 och b_1 med hjälp av
  - korrelationscoefficienten r
  - standardavvikelserna s_x och s_y
  - våra medelvärden $\bar{x}$ och $\bar{y}$



Enkel linjär regression - minsta kvadratmetoden i R

- Låt oss räkna ut regressionslinjens parametrar med R. Vi använder vårt dataset med bilar, med responsvariabeln *litermil* och förklaringsvariabeln *viktton*.
- Vi gör det på två sätt, först med formlerna och sedan med funktionen *lm*.



Stockholm University

Enkel linjär regression - minsta kvadratmetoden i R

- Vi använder formeln $b_1 = r_{x,y}\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}.$
- För att kunna göra det räknar vi först ut.
  - korrelationscoefficienten r
  - standardavvikelserna s_x och s_y
  - våra medelvärden $\bar{x}$ och $\bar{y}$

```
r <- cor(mtcars$litermil, mtcars$viktton)
sx <- sd(mtcars$viktton)
sy <- sd(mtcars$litermil)
xbar <- mean(mtcars$viktton)
ybar <- mean(mtcars$litermil)
sprintf("r=%.2f, sx=%.2f, sy=%.2f, xbar=%.2f, ybar=%.2f",
        r, sx, sy, xbar, ybar)
```

```
[1] "r=0.89, sx=0.44, sy=0.39, xbar=1.46, ybar=1.28"
```

Enkel linjär regression - minsta kvadratmetoden i R

- Nu är vi redo att räkna ut $b_1 = r_{x,y}\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b_1\bar{x}.$

```
b1 <- r * sy / sx
b0 <- ybar - b1 * xbar
sprintf("b0=%f, b1=%f", b0, b1)
```

```
[1] "b0=0.145934, b1=0.778348"
```

- Vi räknade ut att $b_0 = 0.14593$ och att $b_1 = 0.77835$.
- Vår modell är alltså $\widehat{\text{litermil}} = 0.145934 + 0.778348 \cdot \text{vikt}$
- Notera att vi använde funktionen *cor* för att räkna ut korrelationskoefficienten, och funktionen *sd* för att räkna ut standardavvikelserna. Vi har tidigare gått igenom hur även dessa beräkningar kan göras med de matematiska formlerna.

Enkel linjär regression - slump och signifikans

- Nu gör vi samma beräkning med funktionen *lm*.

```
my_regressionmodel <- lm(litermil ~ viktton, data=mtcars)
summary_model <- summary(my_regressionmodel)
summary_model$coefficients
```

```
             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 som vi är intresserade av hittar vi i kolumnen "Estimate".
- *(Intercept)* är vårt b_0, eftersom b_0 är interceptet.
- *viktton* är vårt b_1, eftersom b_1 hör till den förklaringsvariabeln.

Enkel linjär regression - slump och signifikans

```
             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 längs till höger har rubriken $Pr(>|t|)$. Värdet i den kolumnen kallas **p-värde** och har att göra med om en parameter är **statistiskt signifikant**. Ju mindre värdet är desto större statistisk signifikans har parametern.
- Det är vanligt i statistiska analyser att man betraktar en parameter som statistiskt signifikant om p-värdet är lägre än 0.05, men det är upp till bedömaren var gränsen ska går.
- Vi kommer inte att undersöka p-värden närmare i den här delkursen, men de återkommer i senare kurser.

Enkel linjär regression - slump och signifikans

- Vi ska inte beräkna signifikans matematiskt i den här delkursen, men vi ska gå igenom konceptet. Det vi mer specifikt är intresserade av är signifikansen av lutningskoefficienten b_1.
- Notera att vårt dataset *mtcars* bara inkluderar 32 bilmodeller.
- Om vi hade haft 32 andra bilmodeller i vår data hade regressionslinjen sett annorlunda ut.
- Mattias Villinis Widget: https://statisticssu.github.io/SDA1/observable/linreg_simple.html

Enkel linjär regression - slump och signifikans

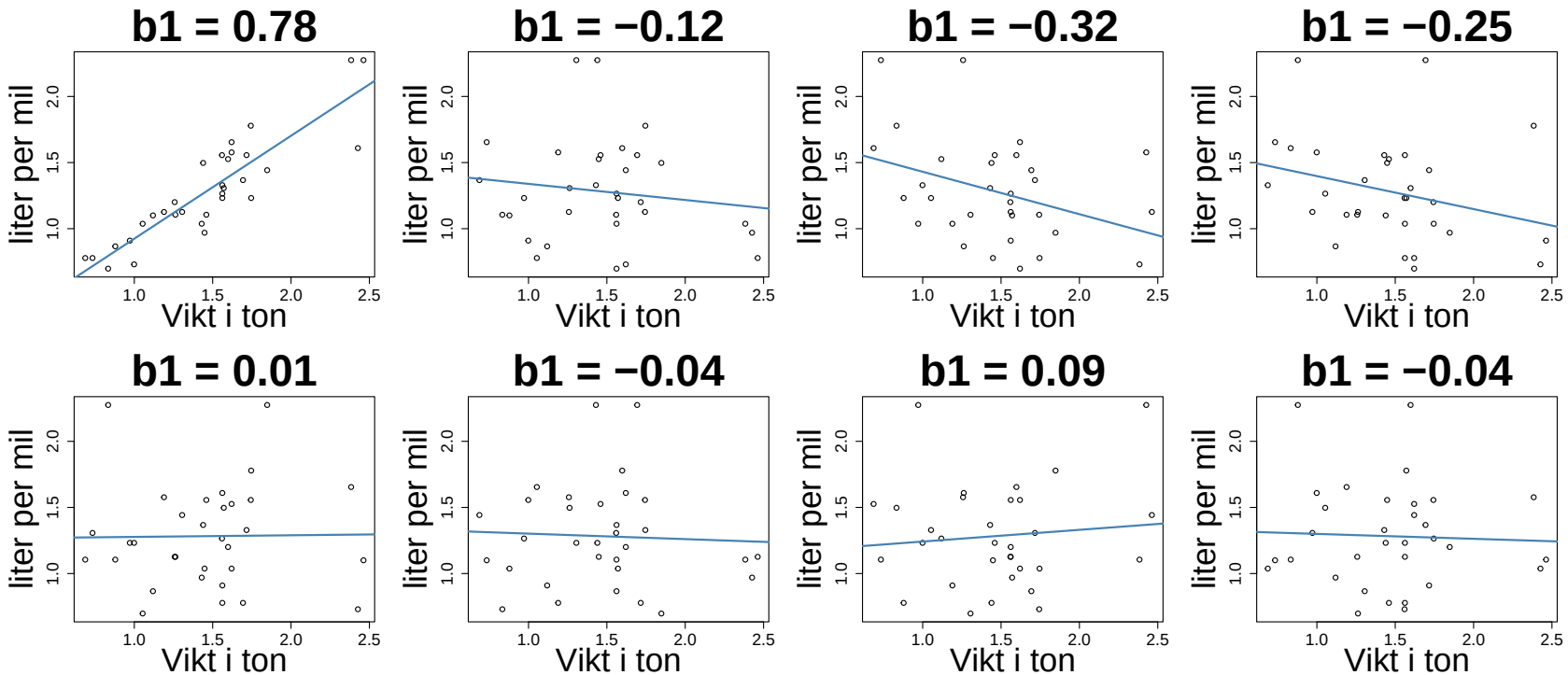
- Vi säger att parametern b_1 är **signifikant** om vi drar slutsatsen att sambandet vi har hittat i vår data gäller generellt för alla bilmodeller - inte bara för de 32 modellerna i vårt dataset.
- Sambandet som vi ser i vår data kan bero på att det finns ett generellt samband mellan variablerna, men det kan också bero på *slumpen*.
- 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.
- **Kom ihåg:** Att ett samband är signifikant i statistisk mening behöver inte betyda att sambandet är *betydelsefullt* i mer allmän bemärkelse.

Enkel linjär regression - slump och signifikans

- Vi har redan sett att p-värdet för b_1 är väldigt lågt i vårt exempel med bilars vikt och bensinförbrukning, vilket visar statistisk signifikans.
- Nu ska vi visa på ett annat sätt att b_1 är signifikant.
- Om det inte fanns ett samband mellan bensinförbrukning och vikt, då kunde vilken bensinförbrukning som helst ha varit kopplad till vilken vikt som helst i vår data.
- Vi ska visa några grafer för hur det slumpvisa utfallet kan bli om det saknas samband mellan vikt och bensinförbrukning. (Se nästa slide)

Enkel linjär regression - slump och signifikans

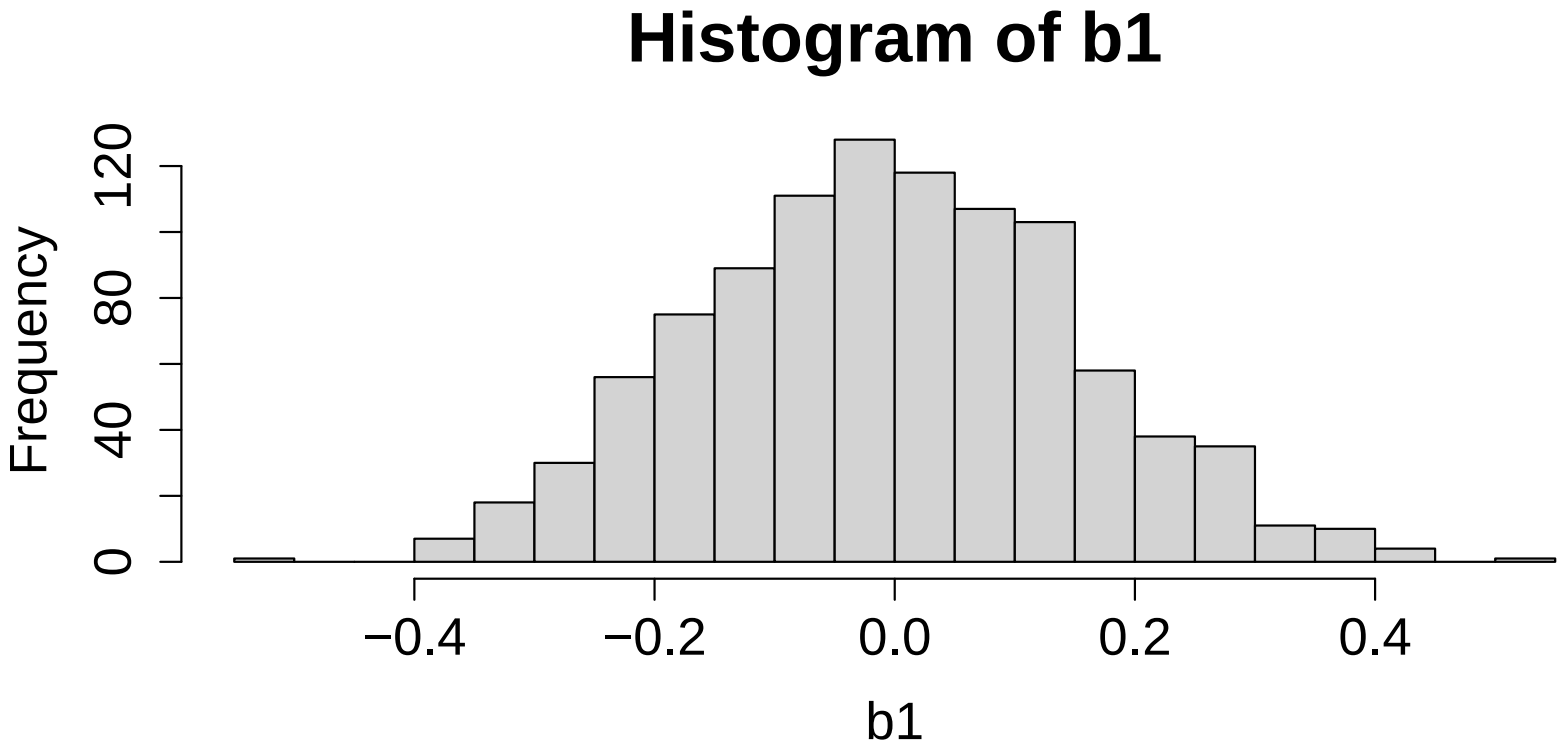
- Den första grafen är vår ursprungliga graf. De övriga graferna visar hur det slumpvisa utfallet kan bli om det saknas ett verkligt samband.



Stockholm University

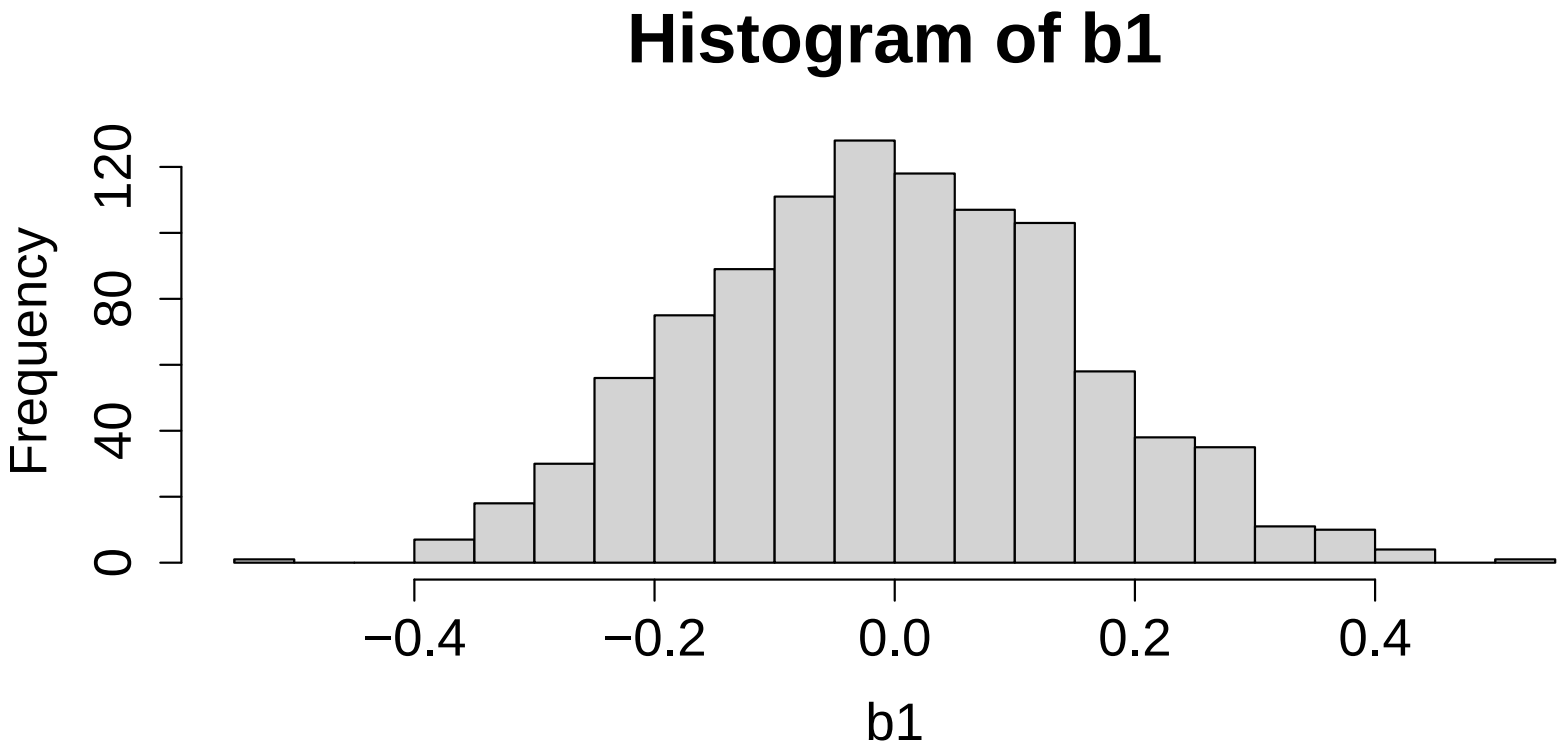
Enkel linjär regression - slump och signifikans

- Nu slumpar vi fram 1000 dataset utifrån hypotesen att det saknas koppling mellan vikt och bensinförbrukning.
- För varje dataset räknar vi ut lutningskoefficienten b_1.
- Histogrammet visar hur de 1000 lutningskoefficienterna b_1 fördelar sig.



Enkel linjär regression - slump och signifikans

- Vi ser att **ingen** av de slumpvisa koefficienterna är så stor som 0.78. Slutsatsen blir att slumpen knappast orsakade det samband som vi såg i vår data. Därför betraktar vi sambandet som signifikant.



Enkel linjär regression - residualanalys

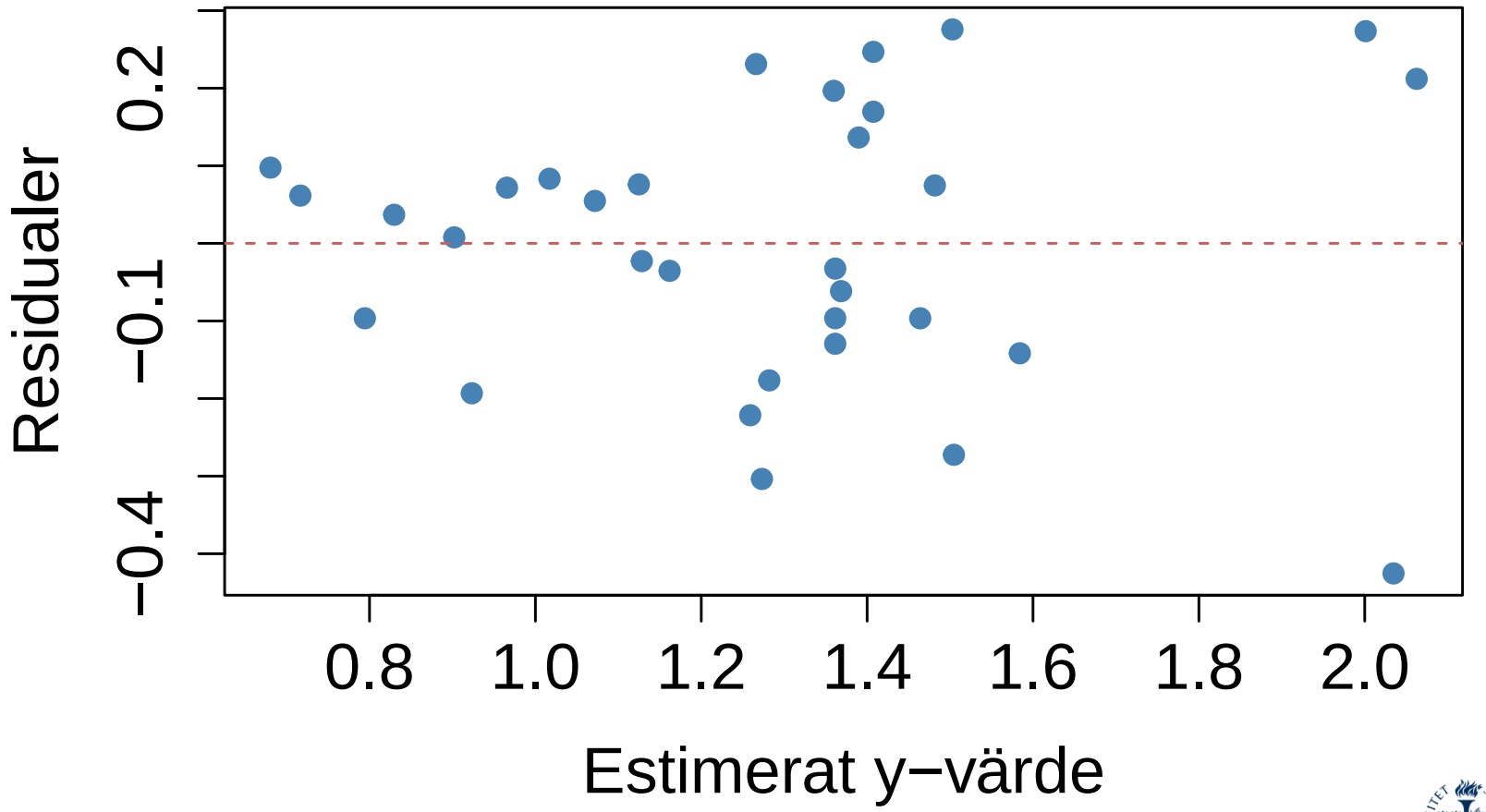
- Vi har redan introducerat residualerna.
- Vi har nämnt att de kan användas för att avgöra hur den bästa regressionslinjen ser ut (den linje som minimerar $\sum e^2$).
- Nu återvänder vi till residualerna, och går igenom hur de kan användas för att utvärdera våra **modellantaganden**.
- Modellantaganden är antaganden som måste vara uppfyllda för att våra resultat ska vara tillförlitliga.

Enkel linjär regression - residualanalys

- Statistiska modeller bygger ofta på antaganden.
- När vi gör regressionsanalys gör vi följande två antaganden:
  1. Residualerna är **normalfördelade**.
  2. Residualernas **varians är konstant**.
- Dessa antagande är i första hand relevanta när vi vill räkna ut felmarginalerna för våra estimerade värden. Det ska vi inte göra här, 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

Enkel linjär regression - residualanalys

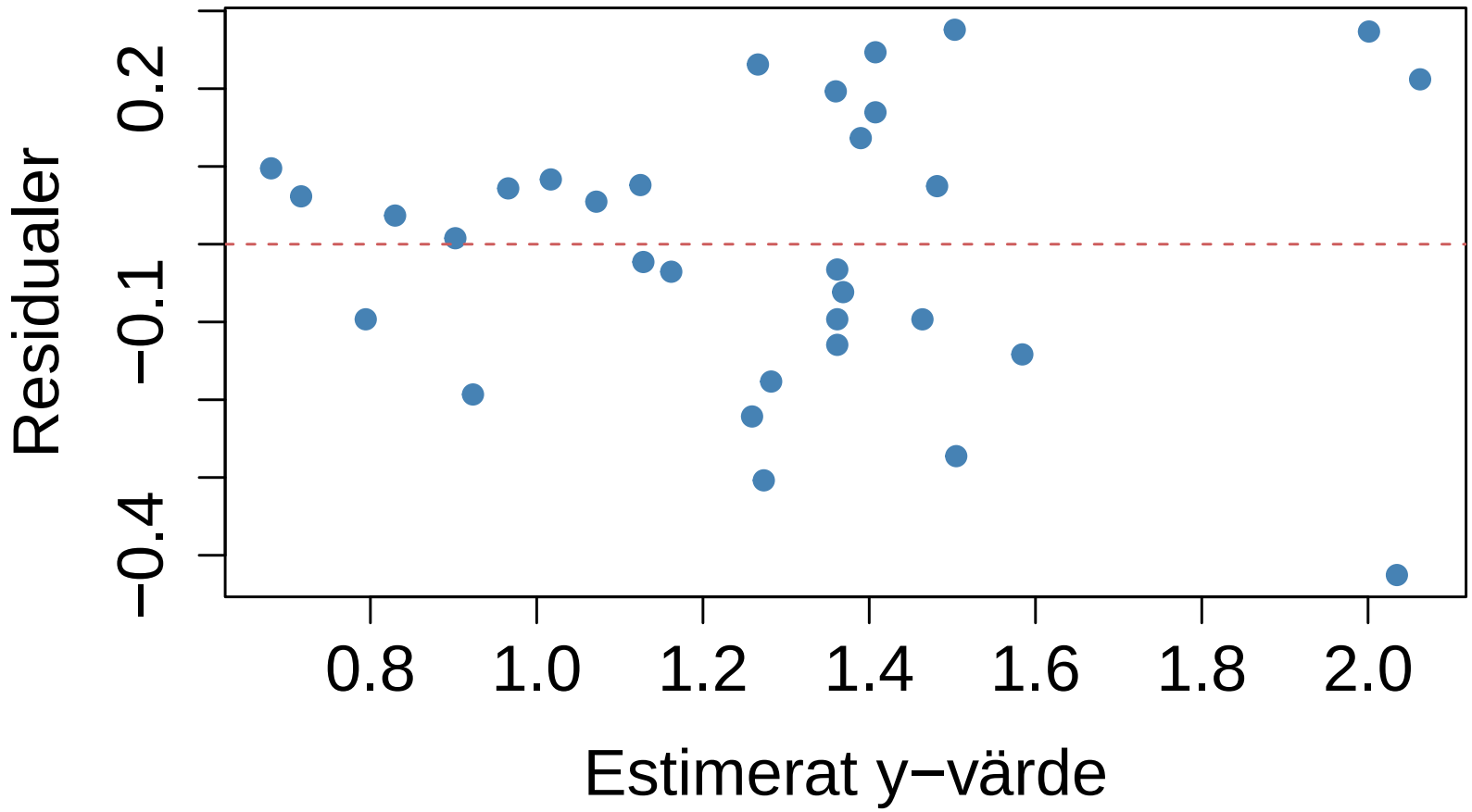
- Bilden visar en **residualgraf** för vår modell för bilmodellers vikt och bränsleförbrukning. På x-axeln har vi vår skattning av y, och på x-axeln har vi residualen e.
- Om residualplotten inte visar något tydligt mönster utan ser slumpmässig ut så är det ett gott tecken.



Stockholm University

Enkel linjär regression - residualanalys

- 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 förstår varför de skiljer ut sig.



Enkel linjär regression - residualanalys

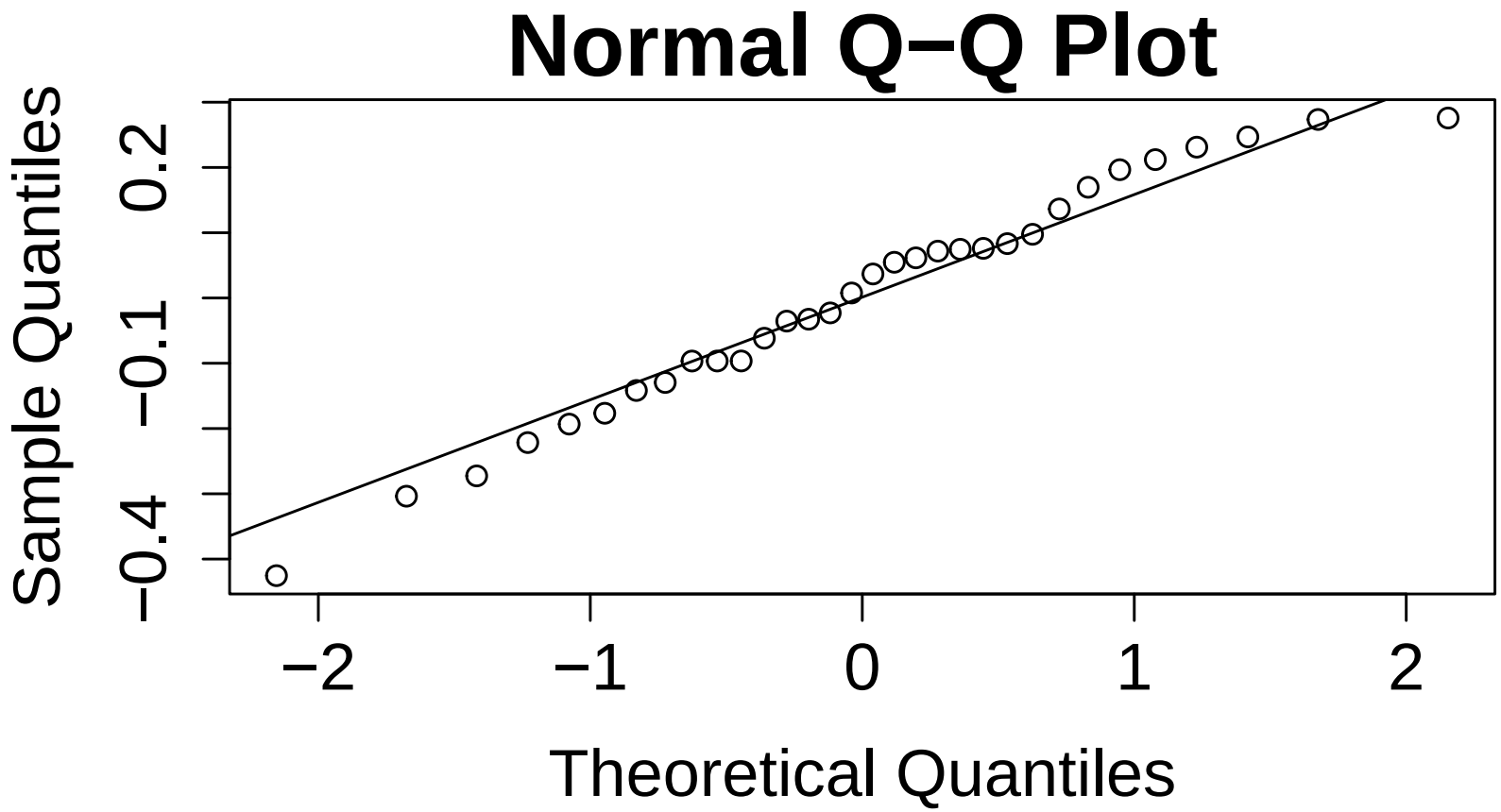
Sammanfattningsvis om residualgrafer

> *"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."*

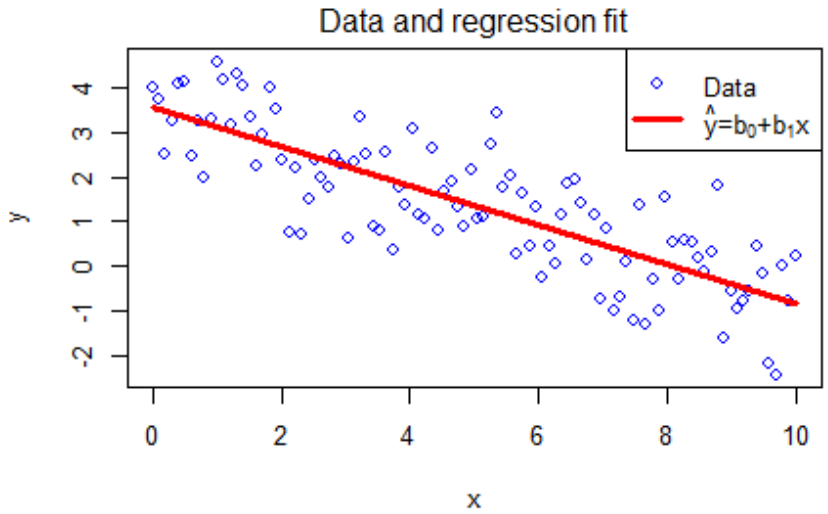
DeVeaux et al (2021), page 238

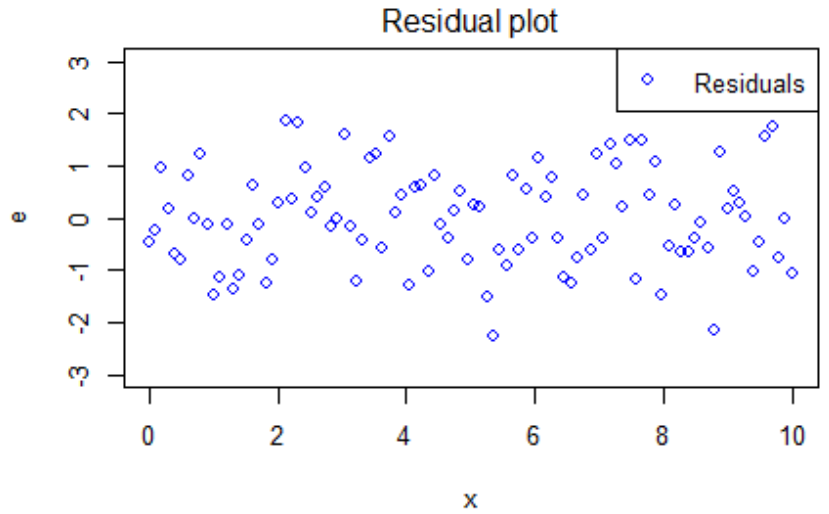
Enkel linjär regression - residualanalys

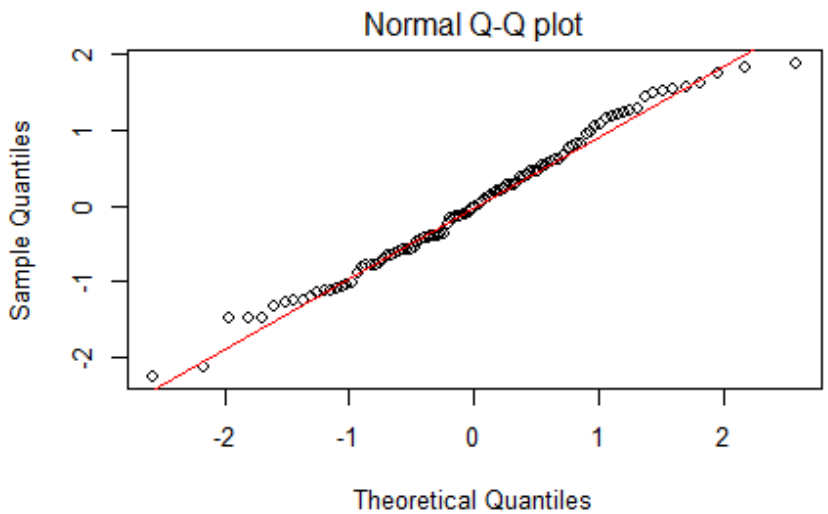
- På den här bilden ser vi en **normalfördelninggraf** för samma modell.
- Om residualerna är normalfördelade ska punkterna följa linjen på ett ungefär, vilket de i det här fallet gör.

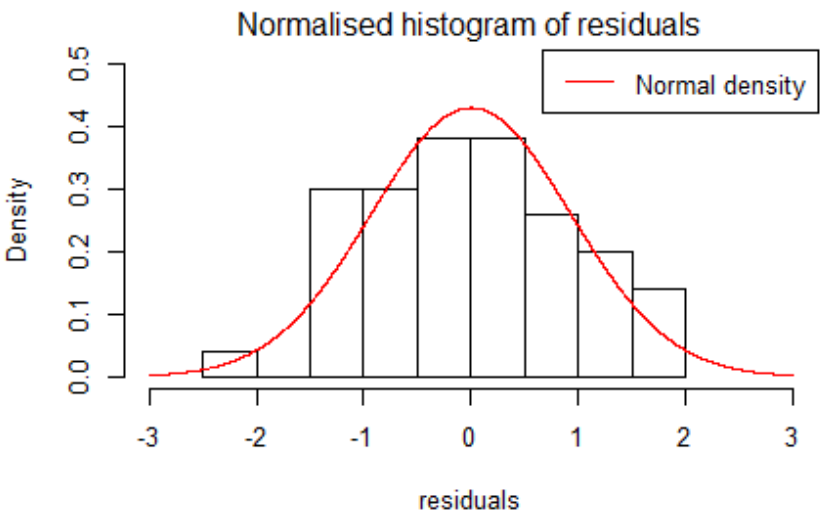


Enkel linjär regression - antaganden uppfyllda



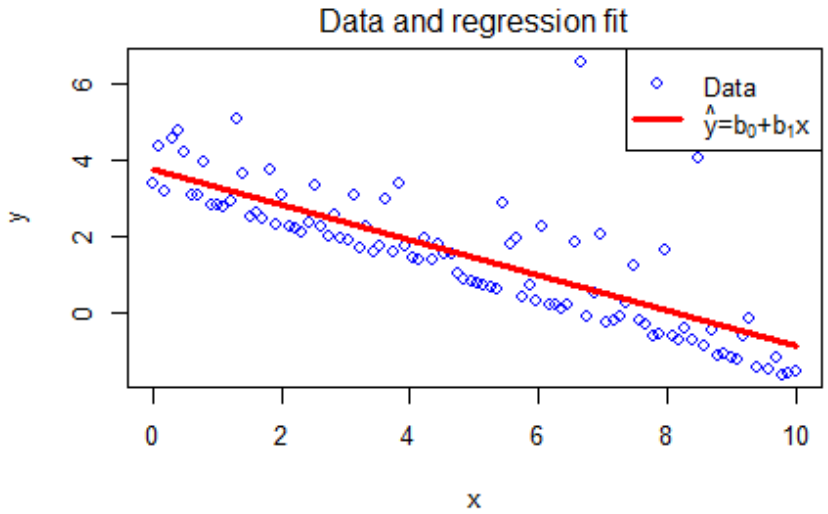


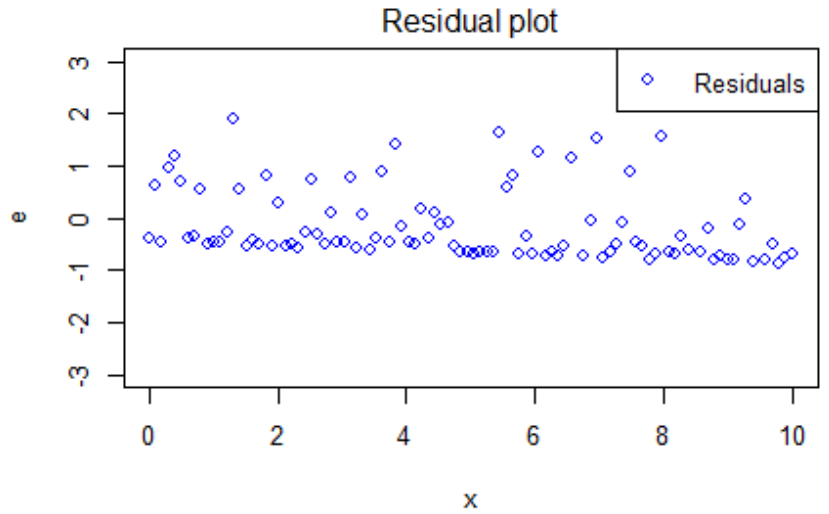


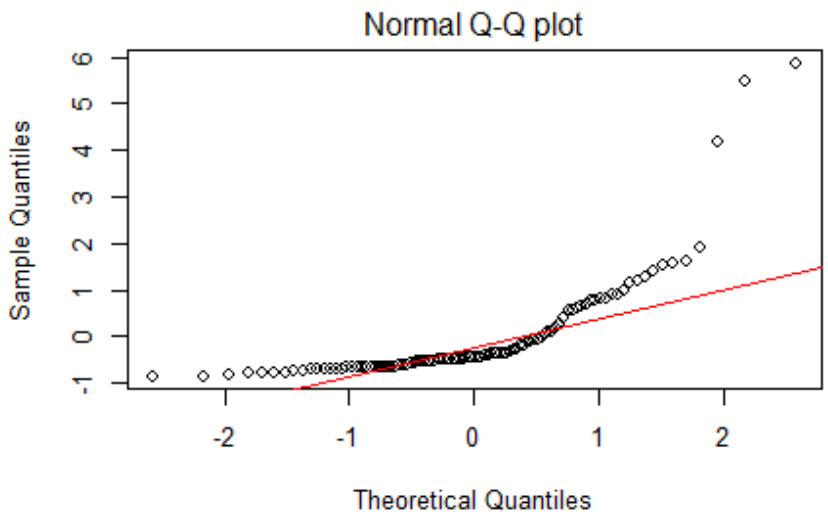


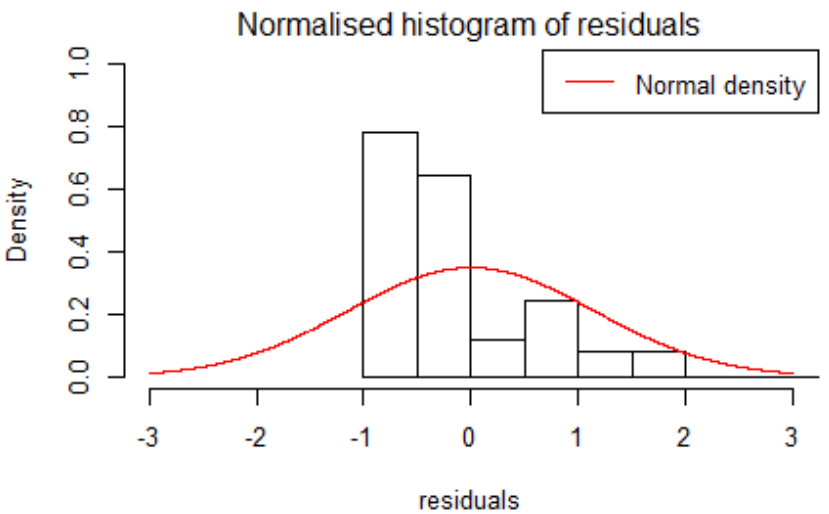
Stockholm University

Enkel linjär regression - inte normalfördelade



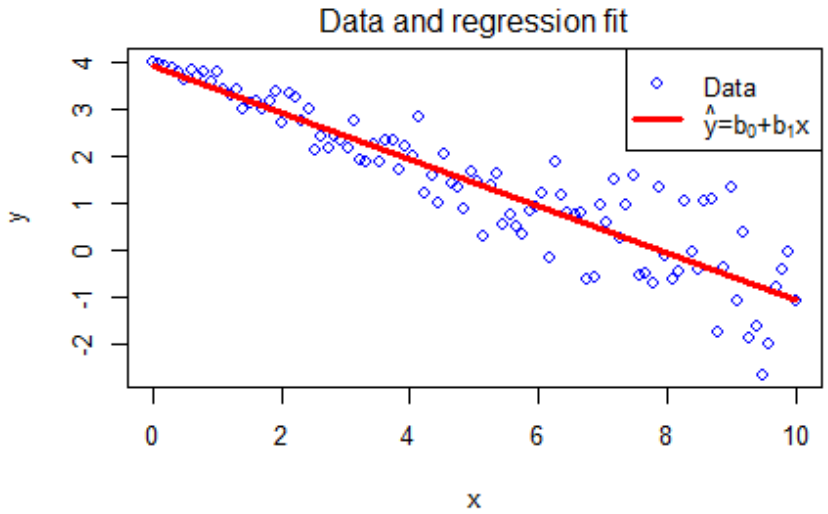


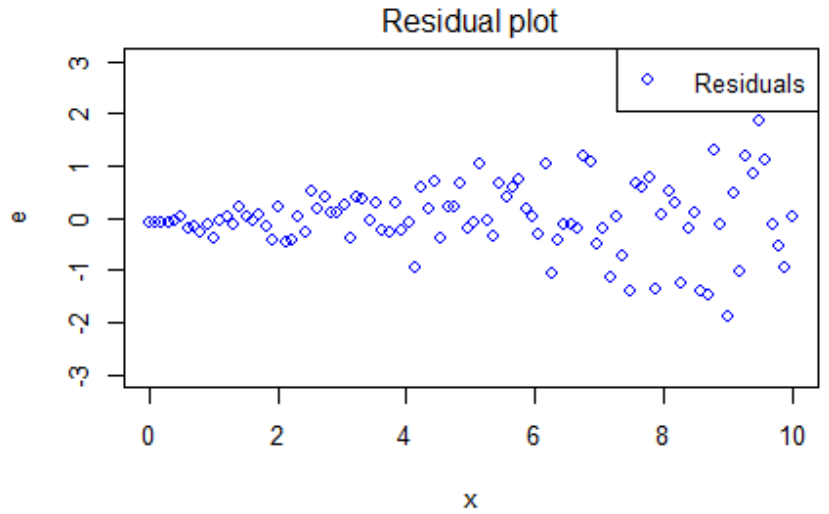


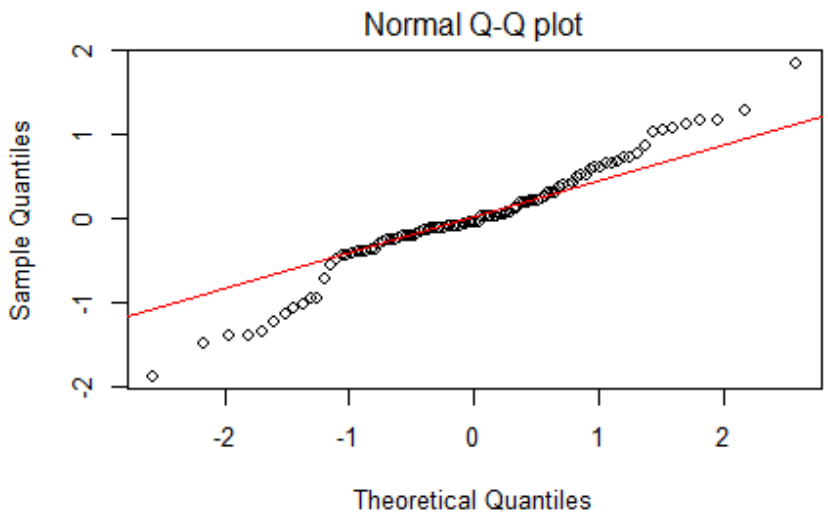


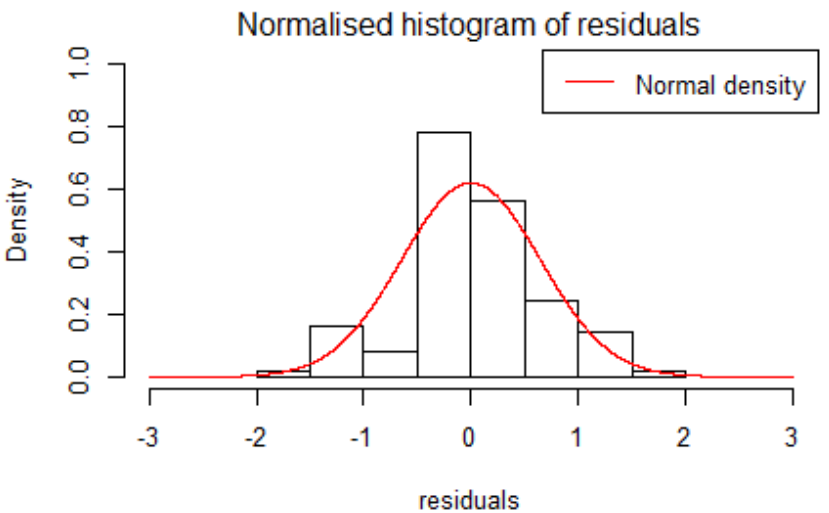
Stockholm University

Enkel linjär regression - inte konstant varians



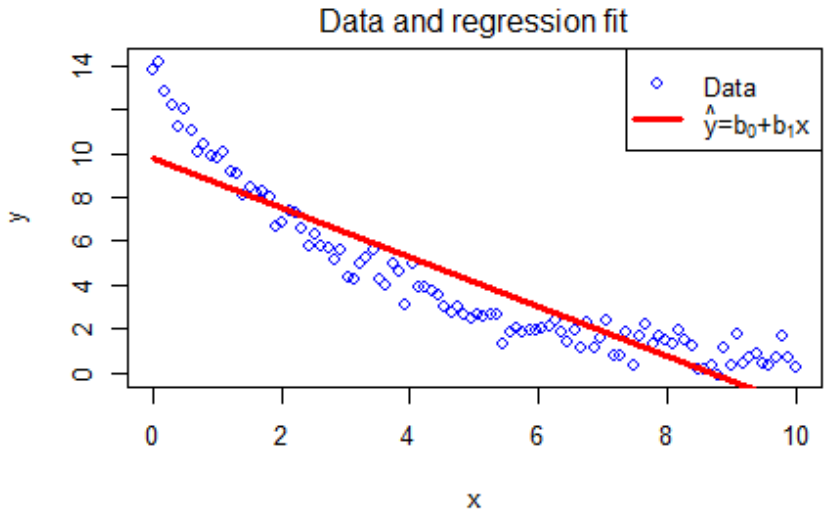


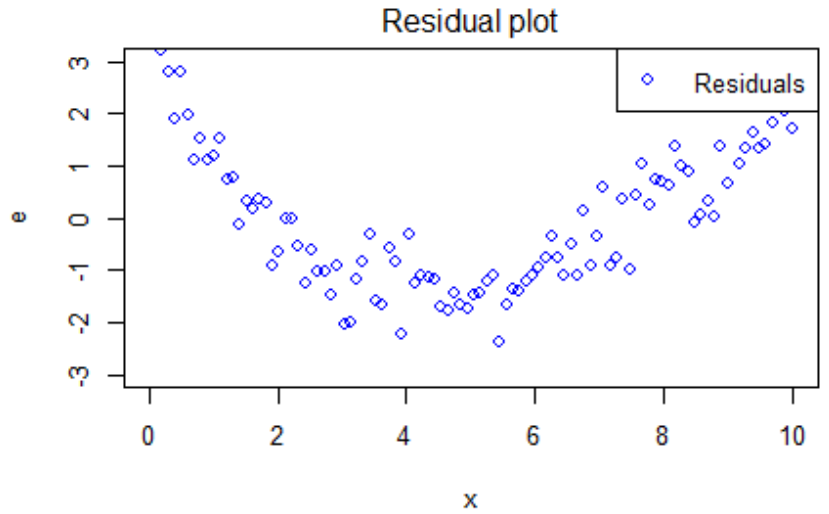


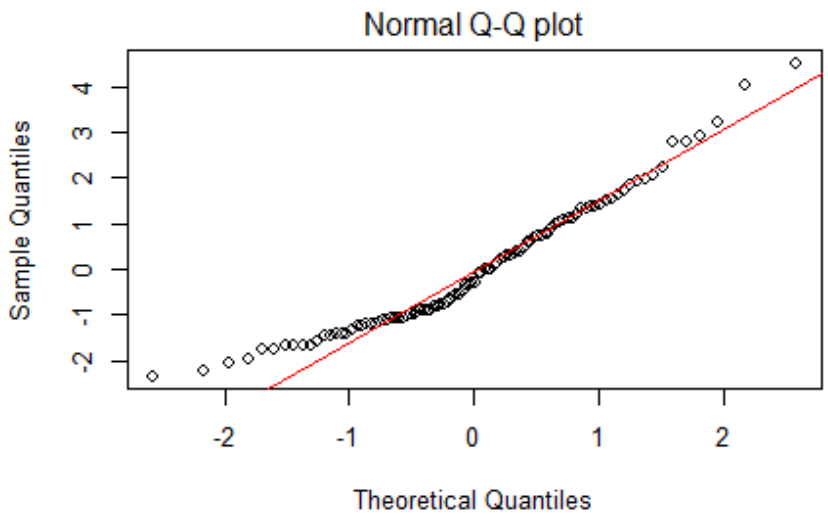


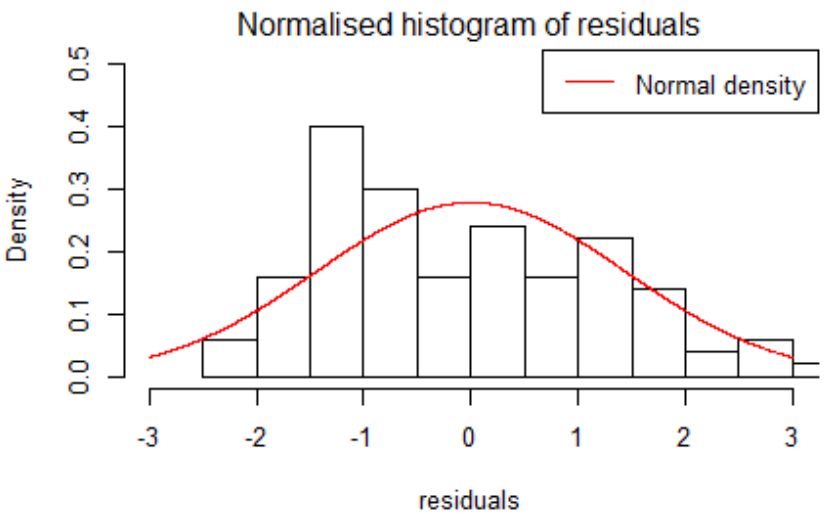
Stockholm University

Enkel linjär regression - inte slumpmässigt mönster







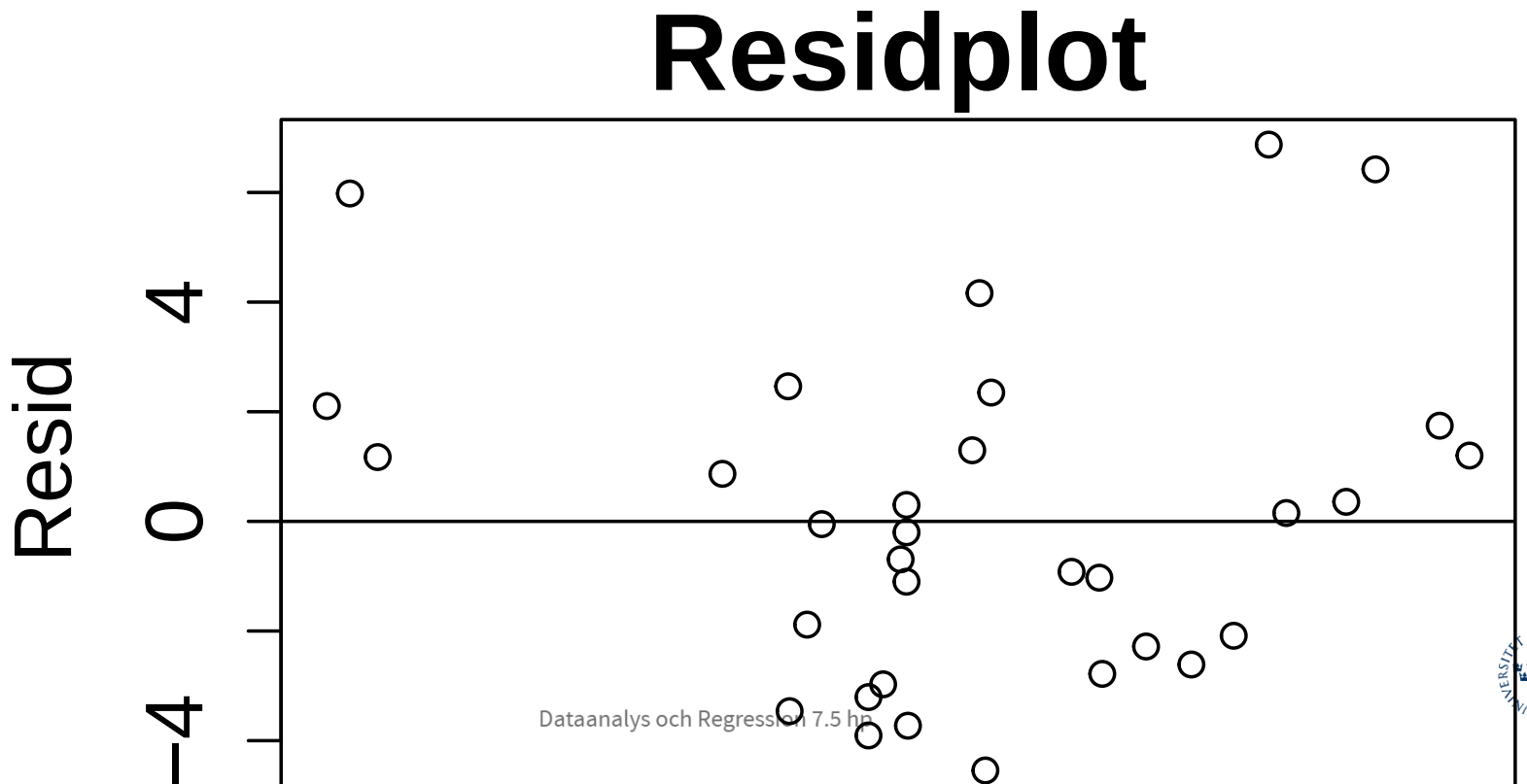


Stockholm University

Enkel linjär regression - residualanalys i R

Så här kan vi skapa en residualgraf med R.

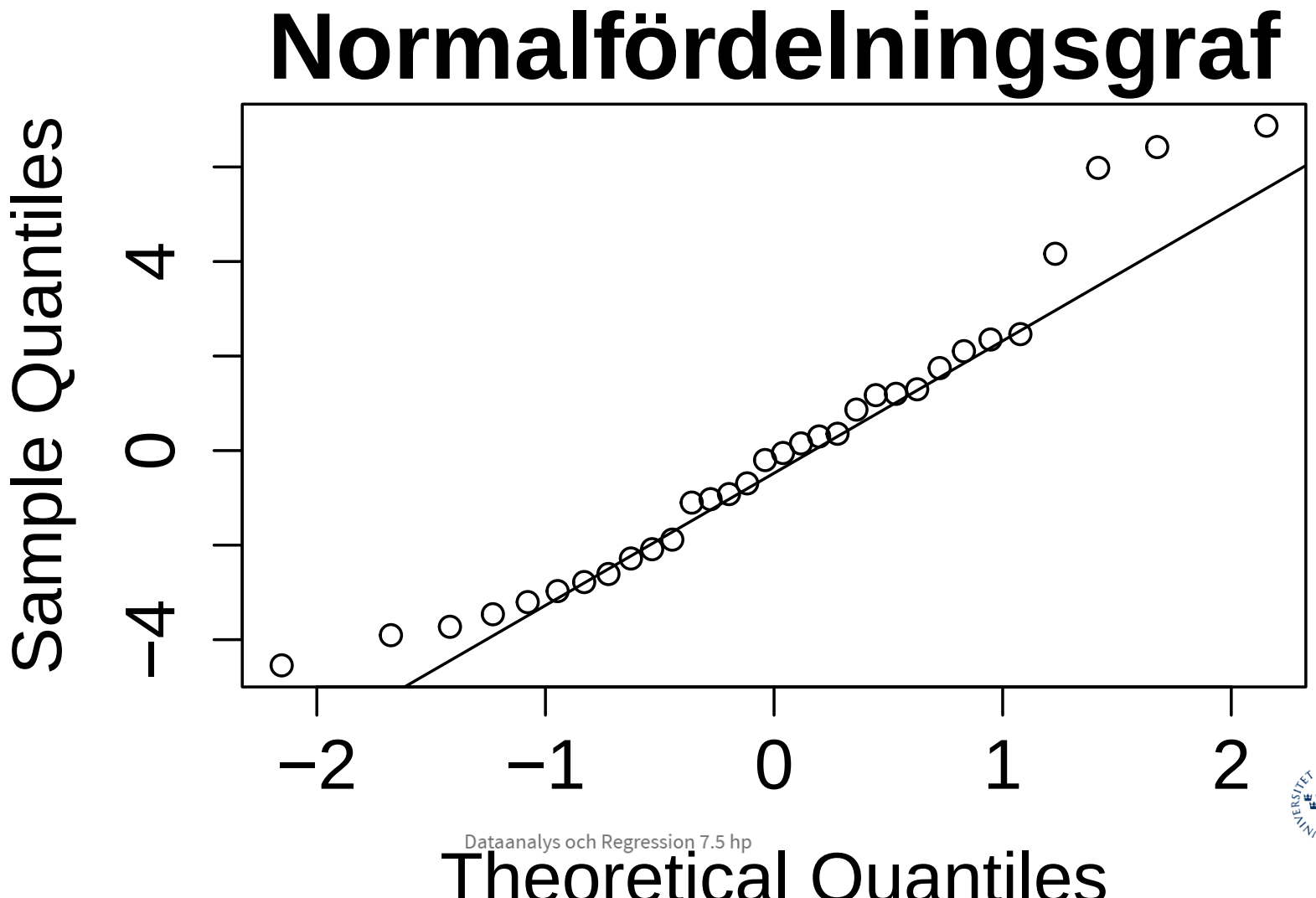
```
lmod2 <- lm(mpg ~ wt, data=mtcars) #Skapa en modell
mtcars$res <- resid(lmod2) # Skapa en vektor med residualerna
mtcars$y_hatt <- fitted(lmod2) # Estimerade y-värden
plot(mtcars$res ~ mtcars$y_hatt, ylab="Resid", xlab="y-hatt",
     main="Residplot")
abline(h=0) #Dra en linje genom residualgrafen vid 0
```



Enkel linjär regression - residualanalys i R

Så här kan vi skapa en normalfördelningsgraf med R.

```
qqnorm(resid(lmod2), main="Normalfördelningsgraf") #Rita graf
qqline(resid(lmod2)) #Lägg till en linje i grafen
```

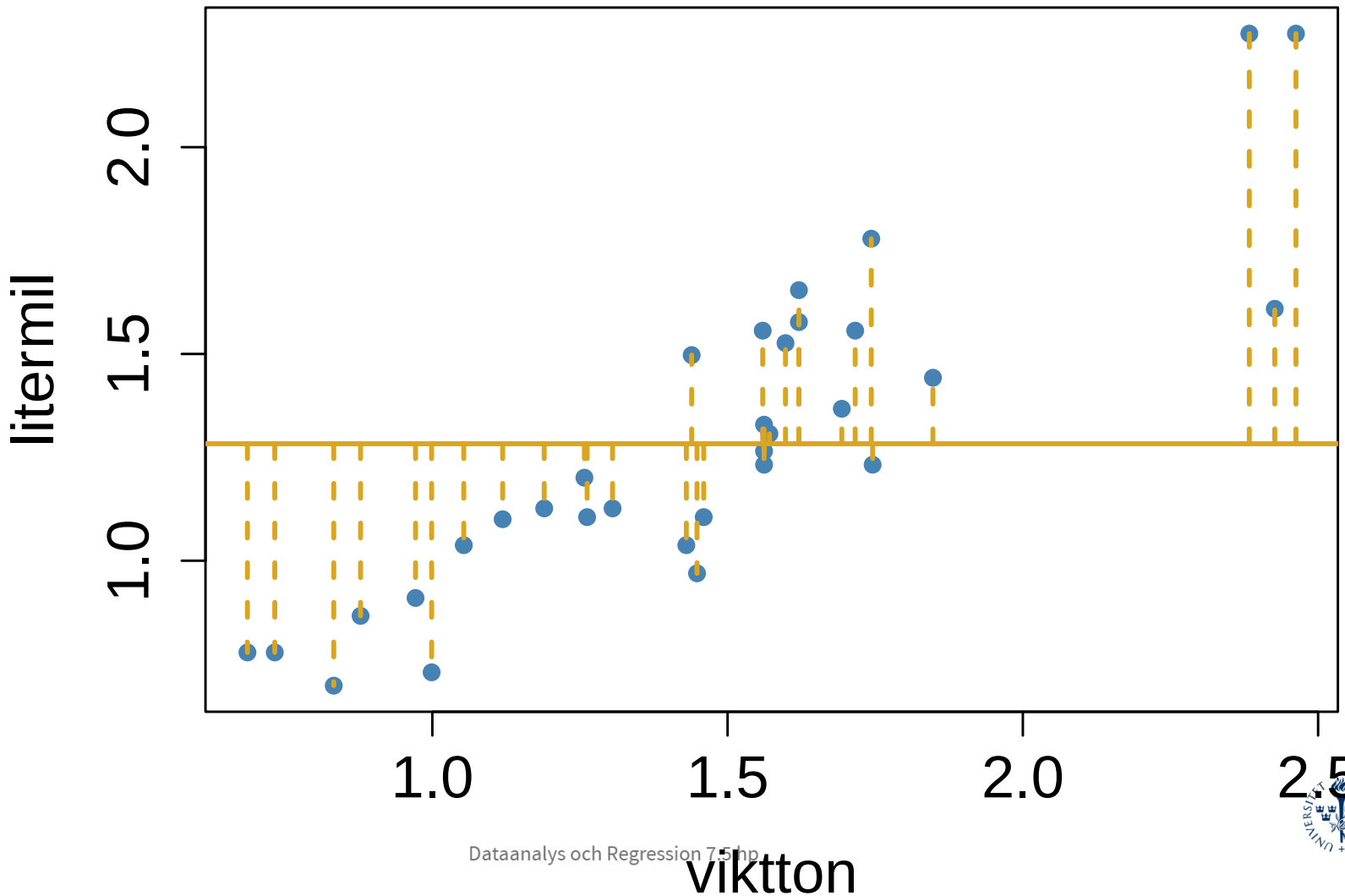


Enkel linjär regression - R-kvadrat

- R^2, som uttalas **R-kvadrat (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 att *förklara variationen*, börja med att föreställa dig den **enklaste tänkbara modell** som inte innehåller någon förklaringsvariabel. Den enklaste tänkbara modellen är $\hat{y} = b_0$.
- I en regressionsmodell med bara interceptet blir $b_0 = \bar{y}$, dvs estimatet blir alltid detsamma som variabelns medelvärde.

Enkel linjär regression - R-kvadrat

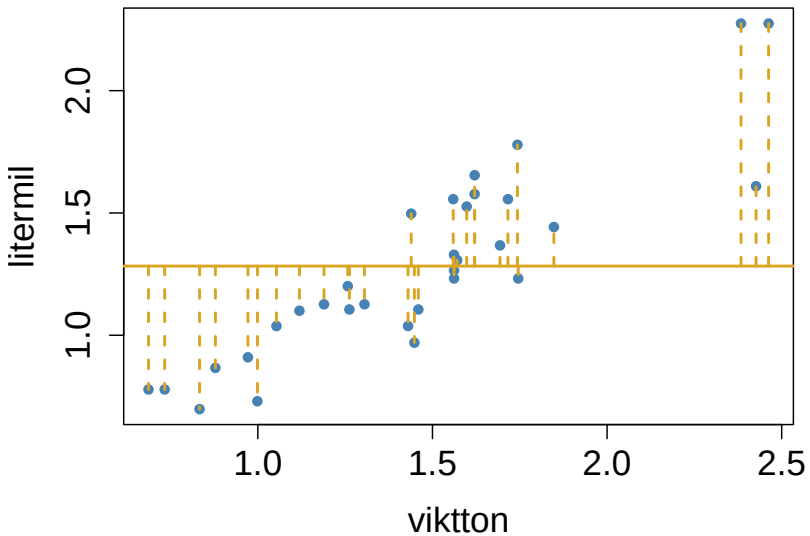
- Modellen $\hat{y} = b_0$ representeras av den gula linjen i bilden.
- Modellen predikterar att alla bilar har samma bensinförbrukning. Därmed ger den **ingen** förklaring till varför bensinförbrukningen varierar mellan de olika bilmodellerna.

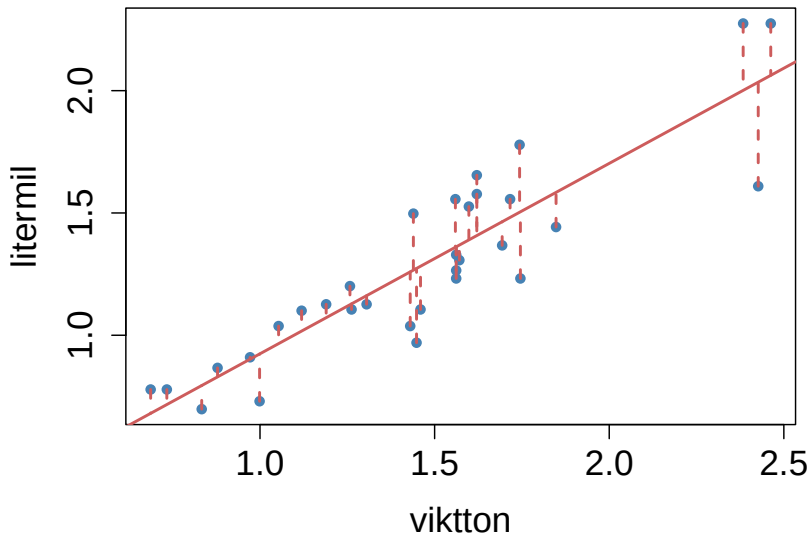


Stockholm University

Enkel linjär regression - R-kvadrat

- Om vi lägger till förklaringsvariabeln *vikt*, får vi modellen $\hat{y} = b_0 + b_1 x$, som i bilden till höger. 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 röda streckade linjer.





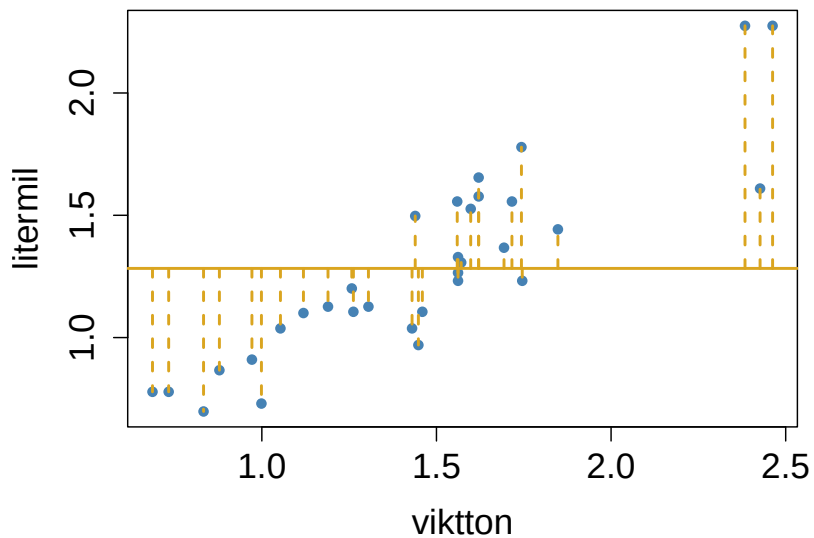
Enkel linjär regression - R-kvadrat

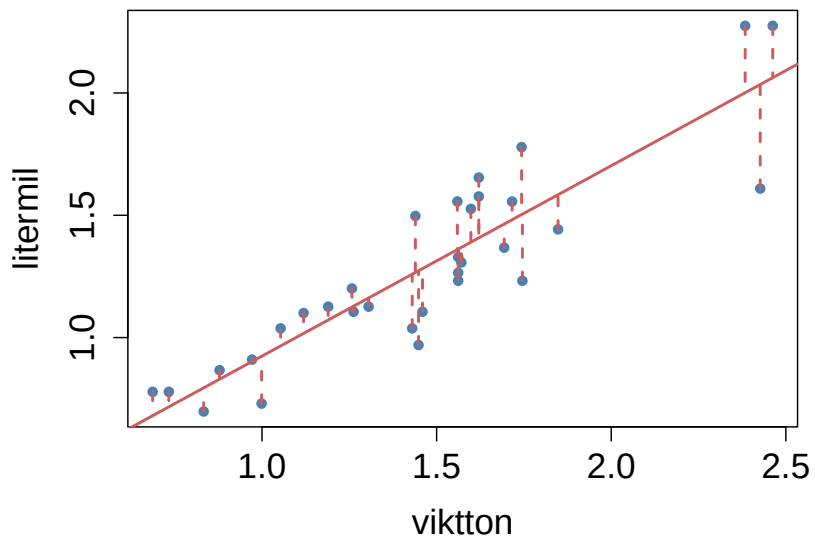
- R^2 anger hur stor **andel** av variationen som en modell förklarar. Om exempelvis $R^2 = 0.6$, då betyder det att modellen förklarar 60% av den totala variationen.
- För att räkna ut R^2 för en modell behöver vi två nya begrepp
  - **SST (Sum of Squares Total):** Den totala variationen i responsvariabeln.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

Enkel linjär regression - R-kvadrat

Vi kan räkna ut SST och SSE med formlerna

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





Enkel linjär regression - R-kvadrat

När vi har räknat ut SST och SSE kan vi räkna ut r-squared som

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

- Uttrycket $\frac{SSE}{SST}$ är andelen av den totala variationen som modellen lämnar *oförklarad*, så $R^2 = 1 - \frac{SSE}{SST}$ blir då den andel som *är* förklarad.
- $SSE \leq SST$, vilket innebär att $0 \leq R^2 \leq 1$
- Notera att i den enklast tänkbara modellen är $\hat{y} = \bar{y}$. Det innebär att för den modellen är SSE = SST, vilket ger $R^2 = 0$.
- Om $R^2 = 1$ för en modell ligger alla observationer exakt på regressionslinjen.

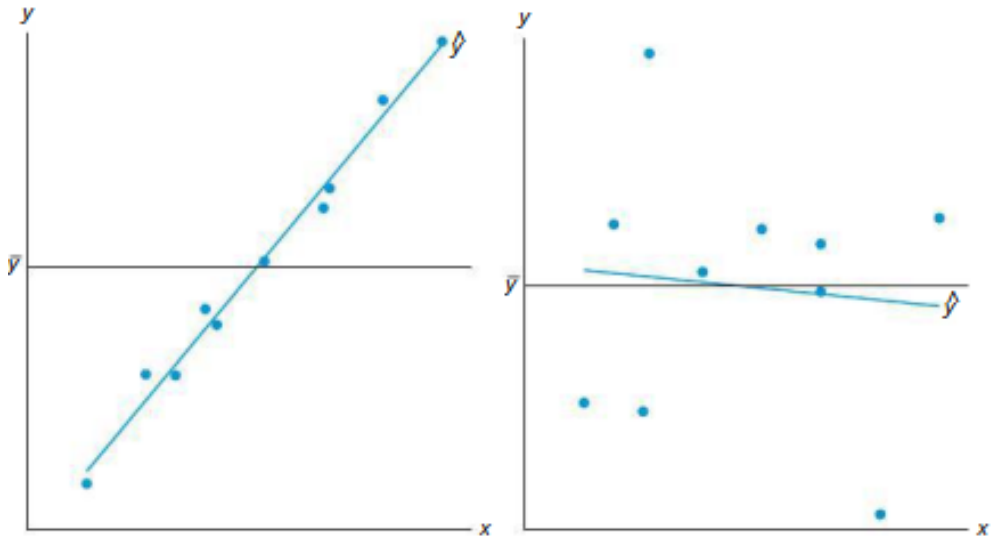
Enkel linjär regression - R-kvadrat

- För enkel linjär regression gäller att $R^2 = r^2$, där r är korrelationskoefficienten för sambandet mellan responsvariebeln och förklaringsvariabeln.
- Det går inte att säga generellt vad ett som är ett bra värde på R^2.
- Inom vissa naturvetenskaper kan det förekomma att R^2 är nära 1.
- Inom 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, och att det är omöjligt att inkludera mer än ett fåtal av dem i en statistisk modell.

Enkel linjär regression - R-kvadrat

Här ser vi två figurer ur Walpole et al. (2016).

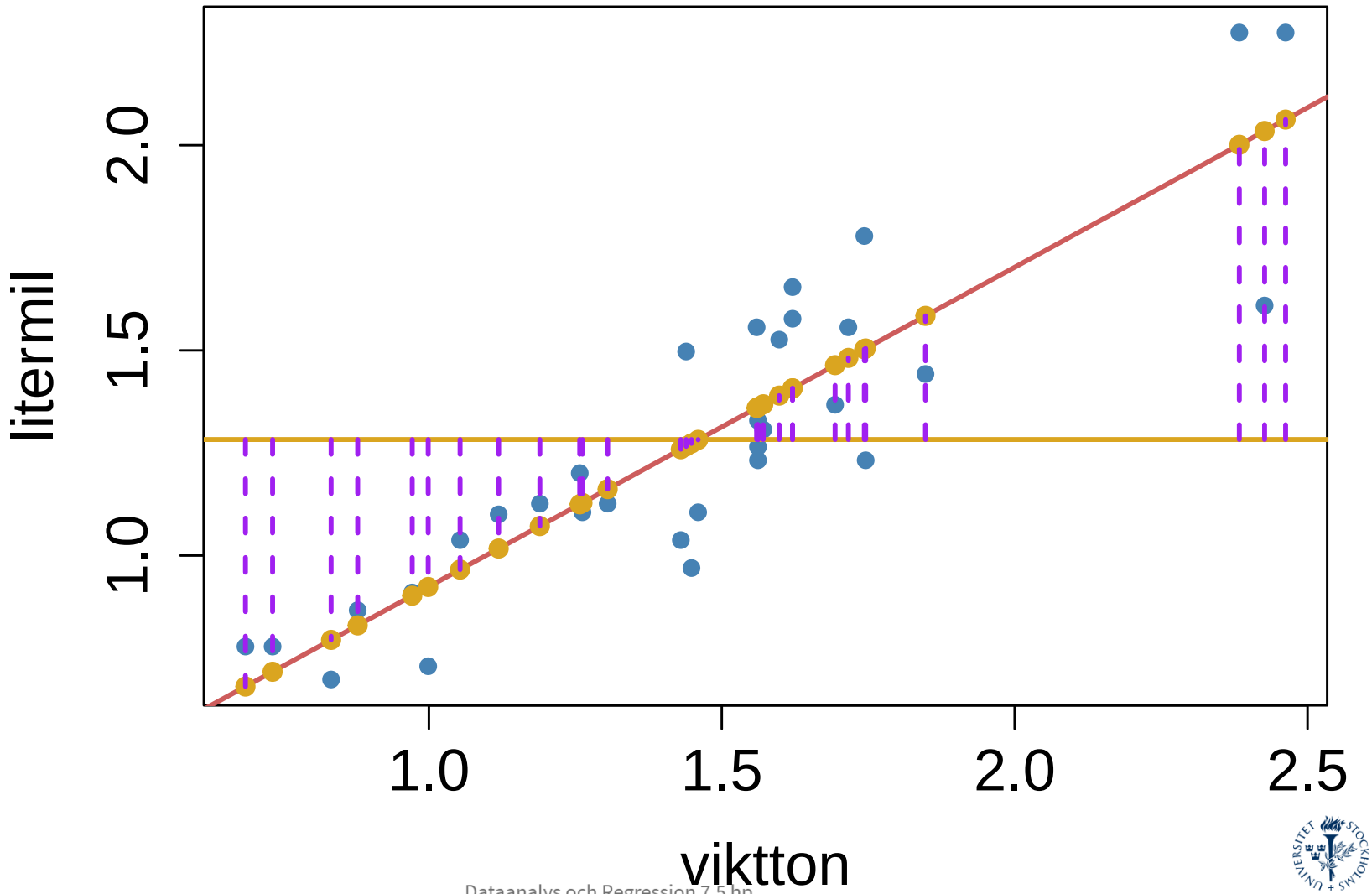
- Figuren till vänster förklarar nästan all variation.
- Figuren till höger nästan inte förklarar någon del alls av variationen.



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

Enkel linjär regression - R-kvadrat

- 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 $\bar{y}$.



Stockholm University

Enkel linjär regression - R-kvadrat

Vi räknar ut SSR med formeln

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

Sambandet mellan SST, SSR och SSE är

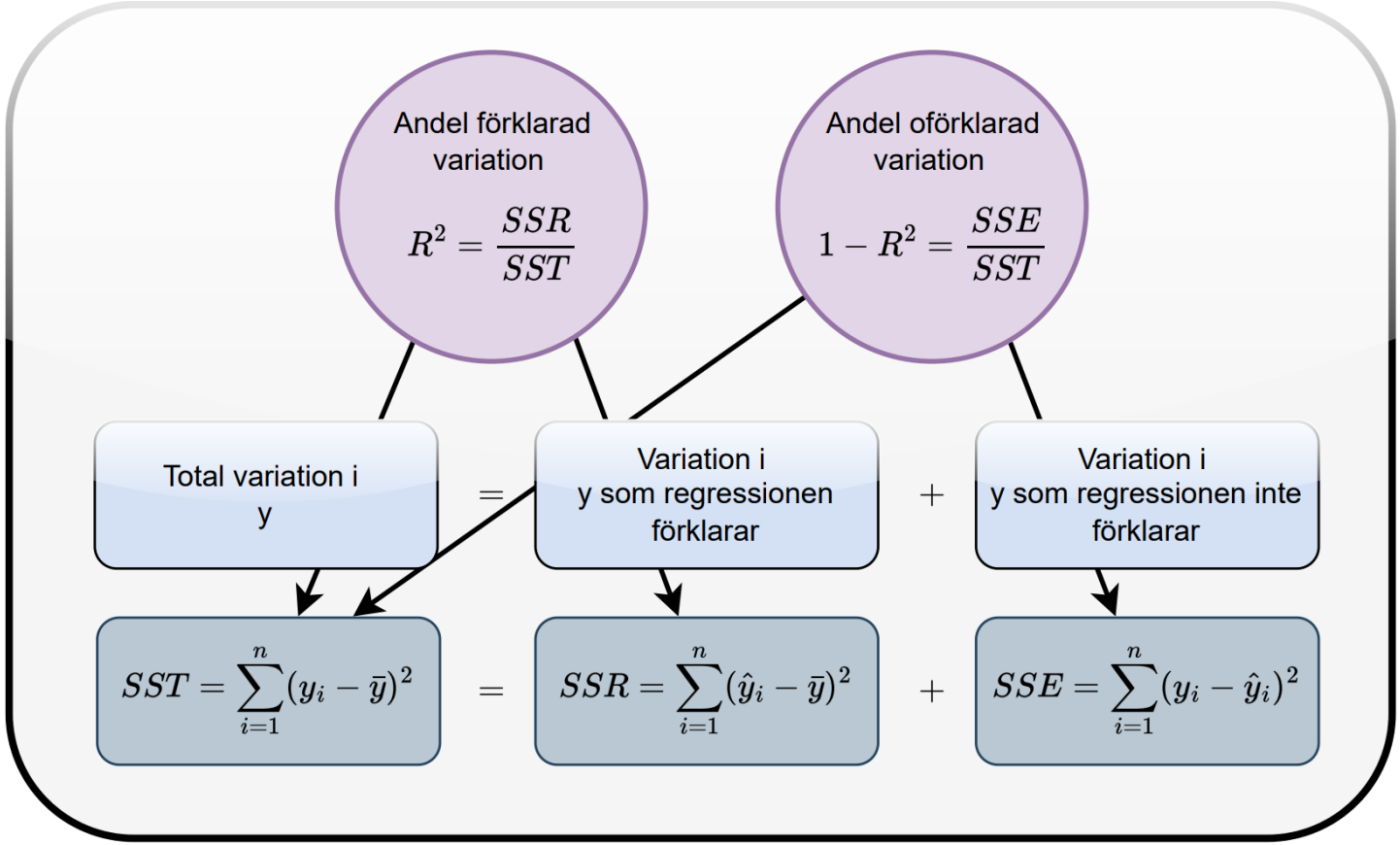
$$SST = SSR + SSE,$$

vilket betyder att r-sqaured också kan räknas ut som

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

Att dela upp variansen på det här sättet kallas **Analysis of variance (ANOVA)**.

Enkel linjär regression - R-kvadrat



Enkel linjär regression - R-kvadrat i R

- 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.

```
lmod <- lm(litermil~viktton, data=mtcars)
y <- mtcars$litermil
y_hatt <- fitted(lmod)

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

sprintf("SST = %.3f, SSR = %.3f, SSE = %.3f", SST, SSR, SSE)
```

```
[1] "SST = 4.680, SSR = 3.706, SSE = 0.974"
```

Enkel linjär regression - R-kvadrat i R

- Vi såg att SST $= 4.680$. Vi Adderar SSR och SSE för att bekräfta att $SSR + SSE = SST$

```
sprintf("SSR + SSE = %.3f", SSR + SSE)
```

```
[1] "SSR + SSE = 4.680"
```

- Sedan räknar vi ut R^2 med formeln $R^2 = 1 - \frac{SSE}{SST}$.

```
R2 <- 1 - SSE/SST
sprintf("With the formula 1 - SSE/SST we get R2 = %.4f", R2)
```

```
[1] "With the formula 1 - SSE/SST we get R2 = 0.7919"
```

Enkel linjär regression - R-kvadrat i R

Vi ser att *lm* räknar ut samma värde för R-squared.

```
summary(lmod)
```

```
 9 Coefficients:
10             Estimate Std. Error t value Pr(>|t|)
11 (Intercept)  0.14593    0.11106   1.314    0.199
12 viktton      0.77835    0.07285  10.685 9.57e-12 ***
13 ---
14 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
15
16 Residual standard error: 0.1802 on 30 degrees of freedom
17 Multiple R-squared:  0.7919,    Adjusted R-squared:  0.785
18 F-statistic: 114.2 on 1 and 30 DF,  p-value: 9.566e-12
```