



MVE051/MSG810 - Statistik och diskret matematik (LP4 2022-2023)

Chalmers och Göteborgs universitet

Föreläsning 14 – Linjär regression

May 11, 2023

Dagens föreläsning

- Introduktion till regression
- Mer om korrelation

Linjär regression

Beskriver en stokastisk motsvarighet till ett linjärt samband mellan två variabler, t ex kan temperatur av havsvattnet y antas vara en linjär funktion av djupet x . Man kan skriva då att

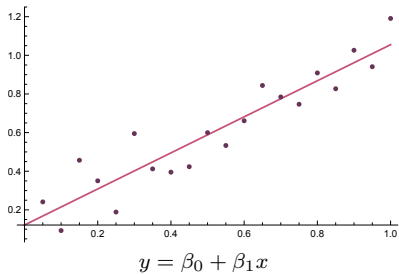
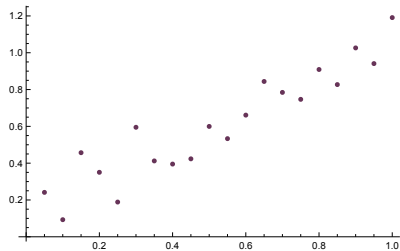
$$y = \beta_0 + \beta_1 x,$$

där β_0 är värdet på y då linjen skär y -axeln och β_1 linjens lutning (ändringen av y då x ökar med en enhet). Det motsvarande stokastiska sambandet är

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma).$$

Y kallas responsvariabel och x prediktorvariabel.

Exempel



Linjär regression

Modell

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

$$\varepsilon_j \sim N(0, \sigma) \quad \text{eller} \quad \mathbb{E}[\varepsilon_j] = 0, \quad \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \text{ oberoende.}$$

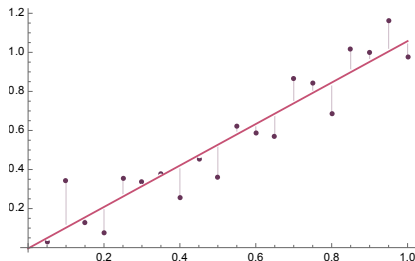
och oberoende.

Hur väljer man $x_1, \dots, x_n$?

- Fixera $x_1, x_2, \dots, x_n$ först (kontrollerad studie)
- Man kan inte alltid fixera värdena på förhand och måste använda de värden som finns observationsstudie.

Linjär regression

Hur skattar man β_0 och β_1 i modellen $Y = \beta_0 + \beta_1 x + \varepsilon$?

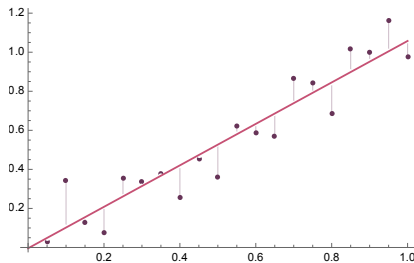


Residualer

Vi vill välja linjen $y = \hat{\beta}_0 + \hat{\beta}_1 x$ så att de vertikala avstånden $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ blir så små som vanligt. Talen $e_1, \dots, e_n$ kallas för residualer.

Linjär regression

Hur skattar man β_0 och β_1 i modellen $Y = \beta_0 + \beta_1 x + \varepsilon$?



Skattare

Vi vill att residualerna ska vara små. Ett sätt att se till att alla residualer är små är att minimera

$$L(\beta_0, \beta_1) = \sum e_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

med avseende på β_0 och β_1 .

Linjär regression

$$L(\beta_0, \beta_1) = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

Flervariabelanalys

$$\begin{cases} \frac{dL}{d\beta_0} = -\sum 2(y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \frac{dL}{d\beta_1} = -\sum 2(y_i - (\beta_0 + \beta_1 x_i))\beta_1 x_i = 0 \end{cases}$$

$$\begin{cases} \sum y_i - n\beta_0 - \beta_1 \sum x_i = 0 \\ \sum x_i y_i - \beta_0 \sum x_i - \sum \beta_1 x_i^2 = 0 \end{cases}$$

Linjär algebra

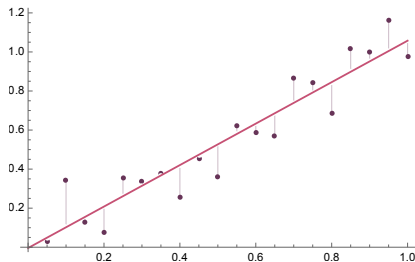
$$\underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_A \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}}_{\mathbf{x}} = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{b}}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \hat{\mathbf{x}} = (A^T A)^{-1} A \mathbf{b}$$

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \quad \text{och} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Linjär regression

Hur skattar man β_0 och β_1 i modellen $Y = \beta_0 + \beta_1 x + \varepsilon$?



Skattare

Residualerna är så små som möjligt om vi minimerar

$$L(\beta_0, \beta_1) = \sum e_i^2 = \sum (y_i - (\beta_0 + \beta_1 x_i))^2.$$

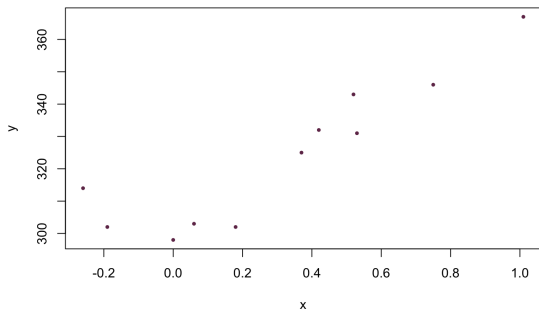
Ett unikt minimum finns i

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{och} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Exempel

Kan man beskriva tiden det tar att köra en algoritm y som en linjär funktion av medelvärdet x i den lista som används som indata?

x	y
0.00	298
0.75	346
0.06	303
-0.26	314
0.18	302
0.42	332
0.19	302
-0.52	343
1.01	367
0.37	325
0.53	331



Exempel

Vi har

- $n = 11$,
- $\bar{x} = 0.3082$,
- $\bar{y} = 323.9$,
- $\sum (x_i - \bar{x})(y_i - \bar{y}) = 79.12$,
- $\sum (x_i - \bar{x})^2 = 1.542$

och får därför

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{79.12}{1.542} = 51.3$$

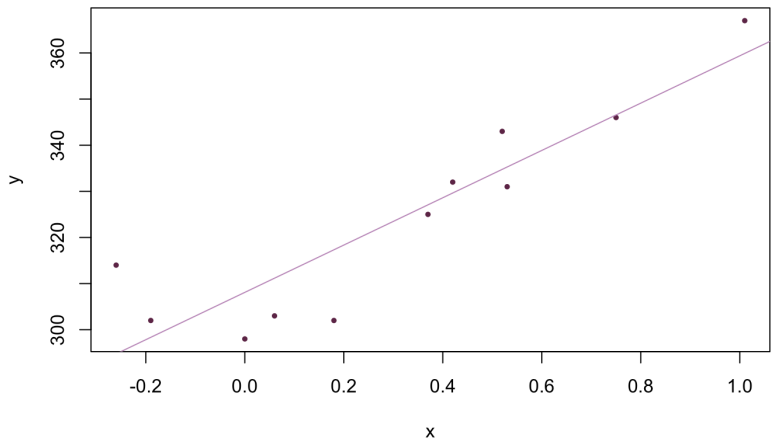
och

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 = 323.9 - 51.3 \cdot 0.3082 = 308.1$$

dvs. linjen

$$y = 308.1 + 51.3x$$

Exempel



Fördelningen för β_0 och β_1

Modell

Givet att

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n,$$

och att

$$\varepsilon_j \sim N(0, \sigma)$$

och oberoende. Vi antar att (x_j) inte är slumpvariabler.

Fördelning för Y_j

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \sim \beta_0 + \beta_1 x_j + N(0, \sigma) = N(\beta_0 + \beta_1 x_j, \sigma), \quad j = 1, 2, \dots, n$$

Eftersom att $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ är oberoende så blir $Y_1, Y_2, \dots, Y_n$ oberoende.

Fördelningen för β_0 och β_1

Fördelning för β_1

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2} - \overbrace{\frac{\sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2}}^{=0} \bar{Y} \\ &= \frac{\sum (x_i - \bar{x})Y_i}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} Y_i \\ &\sim \sum \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} N(\beta_0 + \beta_1 x_i, \sigma) = N\left(\beta_1, \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}\right)\end{aligned}$$

Fördelning för β_0

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} = N(\beta_0 + \beta_1 \bar{x}, \sigma/\sqrt{n}) - N\left(\beta_1, \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}\right) \bar{x} \\ &= N\left(\beta_0, \frac{\sigma \sqrt{\sum x_i^2}}{\sqrt{n \sum (x_i - \bar{x})^2}}\right)\end{aligned}$$

Fördelningen för β_0 och β_1

Skattare för σ

Eftersom att σ ofta är okänd vill vi nästan alltid skatta σ . Man kan visa att

$$s^2 = \frac{\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n - 2} = \frac{\sum e_i^2}{n - 2}$$

är en väntevärdesriktig skattare för variansen, och att

$$\frac{(n - 2)S^2}{\sigma^2} \sim \chi_{n-2}^2$$

Fördelning för β_1 vid okänd varians

$$\frac{\beta_1 - \hat{\beta}_1}{S \sqrt{\sum (x_i - \bar{x})^2}} = \frac{\beta_1 - \hat{\beta}_1}{\sqrt{n - 2} \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$$

Fördelning för β_0 vid okänd varians

$$\frac{\beta_0 - \hat{\beta}_0}{S \sqrt{\sum x_i^2} / \sqrt{\sum (x_i - \bar{x})^2}} \sim t_{n-2}$$

Inferens för β_1

Finns det något samband mellan x och y , dvs är $\beta_1 \neq 0$?

Hypotestest för linjärt beroende

$$H_0: \beta_1 = \beta_1^0 \quad \text{mot} \quad H_1: \beta_1 \neq \beta_1^0$$

med teststatistikan

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{\sqrt{n-2} \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2}}} \sim t_{n-2}$$

Förkasta H_0 om $|t| > t_{n-2, (1+\alpha)/2}$.

Konfidensintervall för β_1

$$CI_\alpha = \hat{\beta}_1 \pm t_{n-2, (1+\alpha)/2} \sqrt{n-2} \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2}}$$

Exempel

```
1 model <- lm(y~x)
2 summary(model)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.3329	-6.2328	-0.5758	5.3673	19.2405

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	308.098	3.978	77.453	5.05e-14 ***
x	51.303	8.203	6.254	0.000149 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.19 on 9 degrees of freedom
Multiple R-squared: 0.813, Adjusted R-squared: 0.7922
F-statistic: 39.12 on 1 and 9 DF, p-value: 0.0001488

Förklaringsgrad

Man brukar beskriva andelen av variansen av Y som förklaras av sambandet med X genom

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2},$$

där $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Residualanalys

Residualer

Residualerna ges av

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1), \quad i = 1, 2, \dots, n$$

Kom ihåg att

$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_1) \sim N(0, \sigma)$, $i = 1, 2, \dots, n$ antogs vara oberoende

Residualgraf

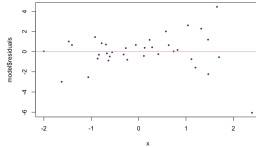
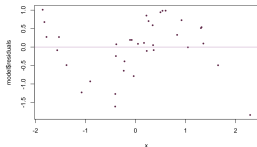
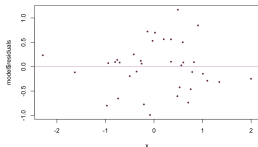
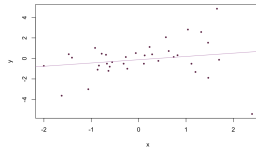
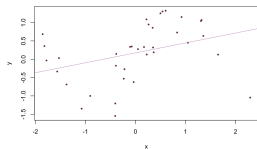
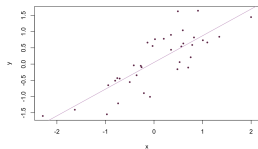
När man plottar (x_i, e_i) får man en residualgraf. Om alla antaganden stämmer borde följande stämma:

- Väntevärdet av bruset är noll (punkterna borde ligga omkring $x = 0$).
- Variansen är konstant (punkterna sprider sig lika mycket för alla värden av x)

Om sakerna ovan inte verkar stämma behöver vi fundera på om våra antaganden stämmer.

Residualanalys

Data och linje (övre) och motsvarande residualplottar (nedre)

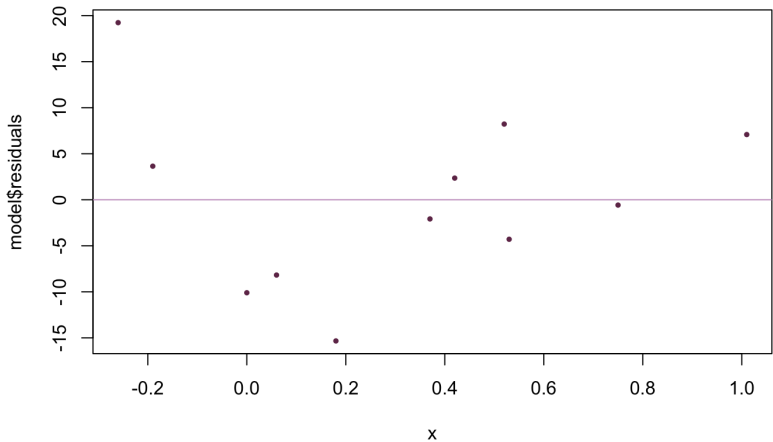


Alla antaganden ok

Linjär modell fel

Variansen inte konstant
(beror på x).

Residualanalysis



Residualanalys

qq-plot

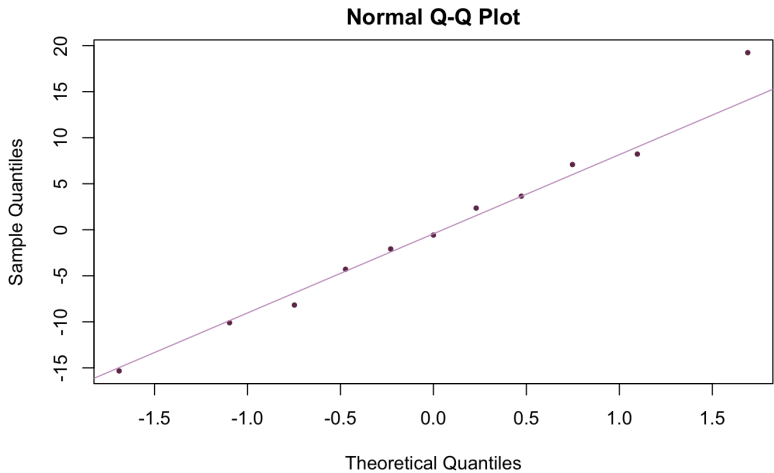
Ett annat sätt att titta på residualerna är med hjälp av en så kallad qq-plot.

För alla test och konfidensintervall har vi antagit antingen

1. våra fel ε_i är normalfördelade, eller
2. vi har så många punkter att vi kan använda centrala gränsvärdessatsen

För att testa om data verkar vara normalfördelad kan jämföra de kvantiler vi kan räkna ut från datan med motsvarande kvantiler från en normalfördelning. Med andra ord, plotta alla par $(z_{i/n}, e^{(i)})$, där $e^{(i)}$ är det i :e minsta felet. Om datan är normalfördelad ligger punkterna nära en rät linje.

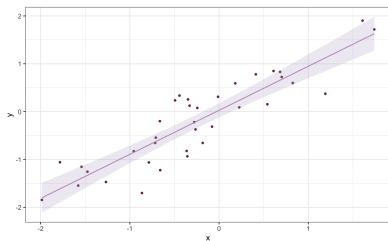
Exempel



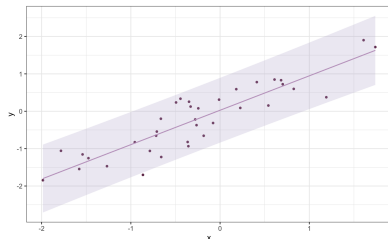
Konfidensintervall och prediktionsintervall

Regressionslinjen ger till varje värde på x en skattning $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ av det förväntade värdet av Y i x .

- Osäkerheten i skattningen i en given punkt kan beskrivas med ett konfidensintervall för $\mathbb{E}[Y | x] = \beta_0 + \beta_1 x$.
- Osäkerheten kring en ny observation Y vid x är typiskt större än osäkerheten av väntevärdet och beskrivs av ett prediktionsintervall (höger).



Konfidensintervall för $\mathbb{E}[Y|x]$



Prediktionsintervall

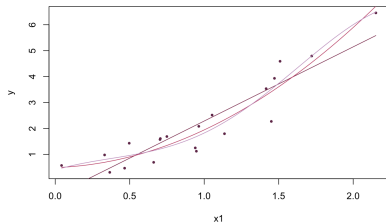
Obs! I båda fallen gäller konfidensintervallen bara för en punkt åt gången.

Obs!

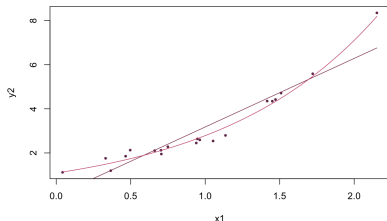
- Att ha bekräftat ett samband mellan två variabler betyder inte att man kan säga någonting om orsakssamband (kausalitet).
- Man bör inte dra slutsatser utanför det undersökta området (extrapolation).

Linjär regression med mer komplicerade funktioner

Man kan använda linjär regression även med fler variabler, t ex om $y = \beta_0 + \beta_1 \log x + \beta_2 x^2$, så länge som sambandet är en linjär funktion i $\beta_0, \beta_1, \beta_2, \dots$



$$\begin{aligned}y &= \beta_0 + \beta_1 x \\y &= \beta_0 + \beta_1 x + \beta_2 x^2 \\y &= \beta_0 + \beta_1 \cos(x) + \beta_2 \sin(x) \\&\quad + \beta_3 \cos(2x) + \beta_4 \sin(2x)\end{aligned}$$

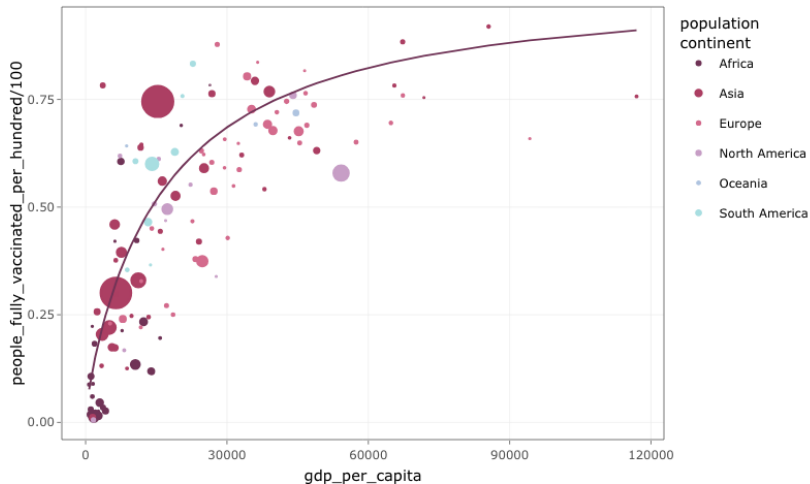


$$y = e^{\beta_0 + \beta_1 x} \Leftrightarrow \log y = \beta_0 + \beta_1 x$$

Man hittar då parametrar genom att ex. hitta minstekvadratlösningen till

$$\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

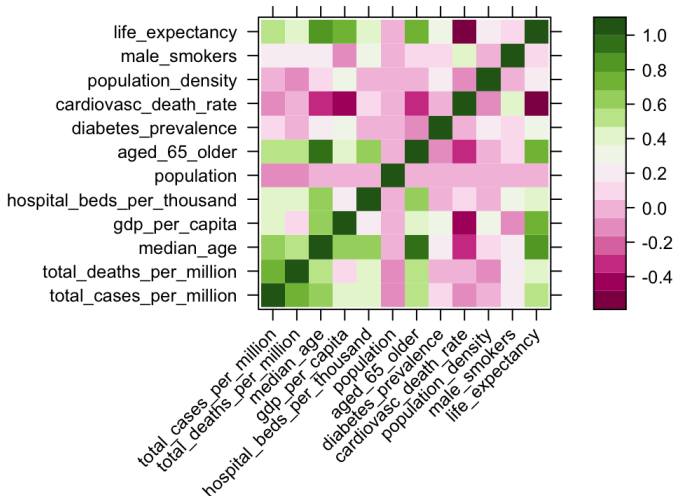
Linjär regression med mer komplicerade funktioner



Multipel regression

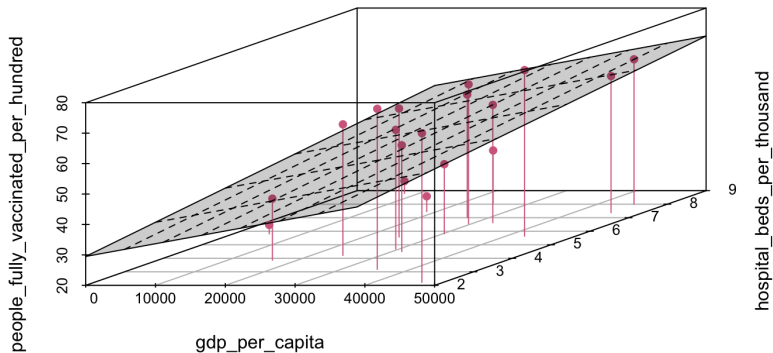
Om man har fler än en prediktor x pratar man om multipel regression,

ex. $y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)}$



Multipel regression

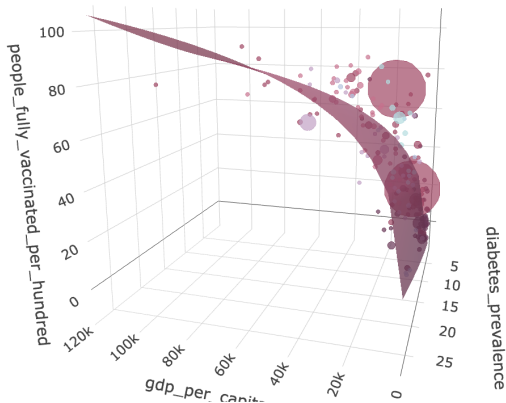
Om man har fler än en prediktor x pratar man om multipel regression,
ex. $y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)}$



Multipel regression

Om man har fler än en prediktor x pratar man om multipel regression,

ex. $y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)}$



Korrelation vs linjärt samband

Korrelation

När vi hade två stokastiska variabler X och Y var korrelationen mellan X och Y

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\mathbb{E}[(X - \bar{X})(Y - \bar{Y})]}{\sqrt{\text{Var}(X) \text{Var}(Y)}}$$

Skattning av korrelation

Korrelationen ρ i data kan skattas med hjälp av

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Korrelationskoefficienten hänger samman med lutningen β_1 och R^2 genom

$$r = \frac{\hat{\beta}_1 \sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{\sum (y_i - \bar{y})^2}} \quad \text{och} \quad R^2 = r^2$$

Nollhypoteserna $H_0: \beta_1 = 0$ och $H_0: \rho = 0$ är därför ekvivalenta.

Test av signifikant korrelation

Finns signifikant linjärt samband mellan två variabler?

Hypotestest för signifikant korrelation

Man kan testa $H_0: \rho = 0$ genom att använda teststatistikan

$$t = \frac{\hat{\beta}_1}{\sqrt{n-2} \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Obs! Detta är exakt samma test som det för nollhypotesen $H_0: \beta_1 = 0$ i regressionsanalys. Man får därför exakt samma p -värde.