

Matematisk Statistik och Diskret Matematik, MVE051/MSG810, VT21

Timo Vilkas

Chalmers - Göteborgs Universitet

Föreläsning 14
20 maj 2021



Regression

- **Regression** är en teknik som används för att undersöka samband mellan olika variabler.
- Regressionen kallas **linjär** när man vill etablera ett linjärt samband mellan variablerna.
- Man utgår ifrån att den ena variabeln X (**förklarande variabel**, eng. *independent variable*) ligger till grund för (fördelningen av) den andra variabeln Y (**responsvariabel**, eng. *dependent variable*). Utfallen på X anses ofta som exakta (eller t.o.m. förbestämda) värden (som inte är slump).
- För ett givet värde x av X är Y fortfarande en slumpvariabel, dock beror dess fördelning på x . Vi skriver $Y|x$ för Y givet att $X = x$. Väntevärdet av $Y|x$ betecknas med $\mu_{Y|x}$.

Linjär Regression

- I allmänhet pratar man om en **regressionskurva** (för sambandet hur Y beror på X) som för specialfallet av linjär regression ges av linjen

$$\mu_{Y|x} = \beta_0 + \beta_1 x,$$

där β_0 och β_1 är okända parameter som måste bestämmas/skattas.

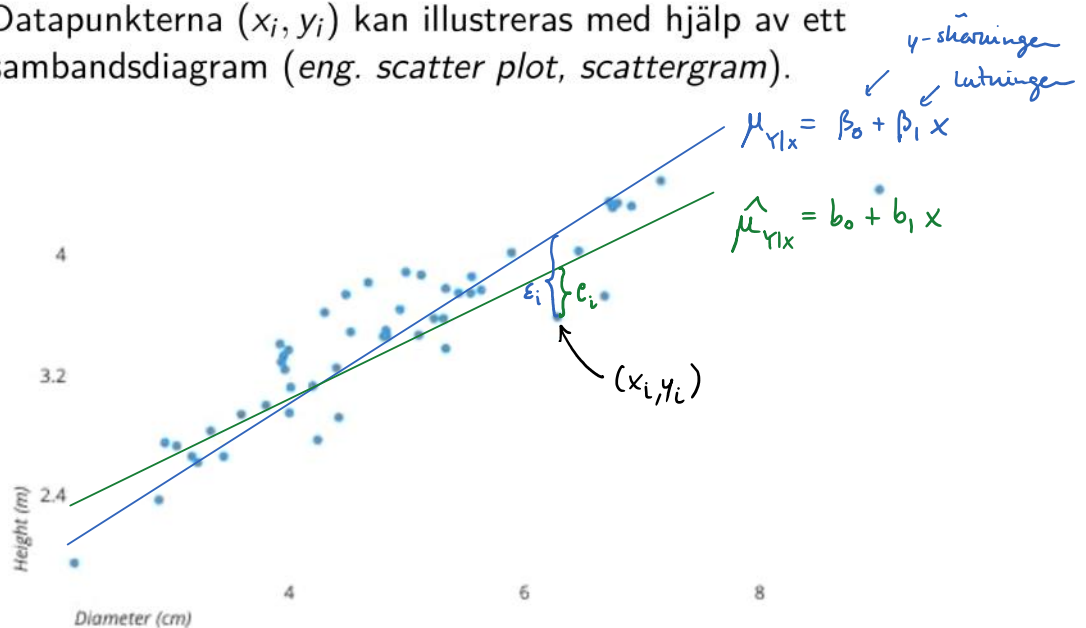
- För en datapunkt (x_i, y_i) betraktar vi x_i som mätvärde på X och y_i som utfall av $Y|x_i$ (som vanligtvis skrivs Y_i för enkelhetens skull), där $\{(x_i, y_i), i = 1, \dots, n\}$ är vår datamängd. Den enkla modellen för linjär regression blir då

$$Y_i = \beta_0 + \beta_1 x_i + E_i \quad \text{resp.} \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

där ε_i är utfallet på s.v. E_i (som antas ha väntevärde 0).

Sambandsdiagram

- Datapunkterna (x_i, y_i) kan illustreras med hjälp av ett sambandsdiagram (eng. *scatter plot*, *scattergram*).



Residuer

- En skattning av parametrarna med hjälp av estimatorer $\hat{\beta}_0 = b_0$ och $\hat{\beta}_1 = b_1$ leder sedan till en (approximativ) regressions linje i form av

$$\hat{\mu}_{Y|X} = b_0 + b_1 X.$$

- Avvikelsen per datapunkt (x_i, y_i) ges då genom

$$y_i = b_0 + b_1 x_i + e_i.$$

Dessa "fel" e_i , $i = 1, \dots, n$ kallas **residuer** (eng. *residuals*).

- Observera: $\varepsilon_i = \mu_{Y|X_i} - y_i = \beta_0 + \beta_1 x_i - y_i$
och $e_i = \hat{\mu}_{Y|X_i} - y_i = b_0 + b_1 x_i - y_i$.

Minsta-kvadrat-skattning

- Den vanligaste metoden för att skatta β_0 och β_1 i linjär regression är de så kallade **minsta-kvadrat estimatorer**, som minimerar det sammanlagda kvadrerade felet

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

- En enkel extrempunktsundersökning av SSE ger

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

och

$$b_0 = \bar{y} - b_1 \bar{x}.$$

Minsta-kvadrat-skattning

Skiss: $SSE = \sum_i (y_i - b_0 - b_1 x_i)^2$; $\frac{\partial SSE}{\partial b_0} = -2 \sum_i (y_i - b_0 - b_1 x_i) = 0 \Leftrightarrow n b_0 + b_1 \sum_i x_i = \sum_i y_i$

$\frac{\partial SSE}{\partial b_1} = -2 \sum_i (y_i - b_0 - b_1 x_i) x_i = 0 \Leftrightarrow b_0 \sum_i x_i + b_1 \sum_i x_i^2 = \sum_i x_i y_i$

$\begin{pmatrix} \sum_i x_i & \sum_i x_i^2 \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \Rightarrow \frac{1}{n \sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$

$b_1 = \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i x_i^2 - (\sum_i x_i)^2}$, $b_0 = \frac{1}{n} \sum_i y_i - b_1 \cdot \frac{1}{n} \sum_i x_i = \bar{y} - b_1 \cdot \bar{x} \quad \square.$

Anmärkning

- Att man kvadrerar felen e_i har två anledningar: Å den ena sidan vill man undvika att positiva och negativa fel tar ut varandra, å den andra ska större fel bestraffas "hårdare".
- Vi kommer fram till en skattning av $\mu_{Y|X}$, inte själva $Y|X$. Det duger dock ofta redan som approximation (eller man kan anta att spridningen kring väntevärdet är normalfördelad).

Exempel

Låt X stå för antalet rader i en SAS kod som körs, och Y vara motsvarande tid det tar att utföra koden. Vi får följande information om datapunkterna:

$$n = 10, \quad \sum_{i=1}^{10} x_i = 16.75, \quad \sum_{i=1}^{10} y_i = 170$$

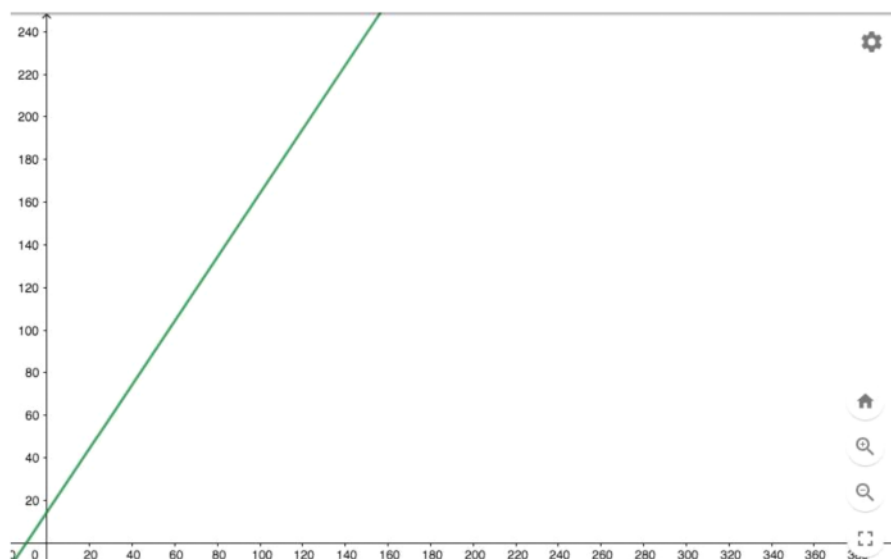
$$\sum_{i=1}^{10} x_i^2 = 28.64, \quad \sum_{i=1}^{10} y_i^2 = 2898, \quad \sum_{i=1}^{10} x_i y_i = 285.625$$

Beräkna regressionslinjen.

$$b_1 = \frac{10(285.625) - (16.75)(170)}{10(28.64) - (16.75)^2} = 1.499$$

samt

$$b_0 = \frac{170}{10} - 1.499 \frac{16.75}{10} = 14.489$$



Modellantaganden

- Då utfallen b_0 , b_1 samt e_i beror på datamängden, är det rimligt att definiera motsvarande slumpvariabler B_0 , B_1 och E_i :

$$B_1 = \frac{n \sum_{i=1}^n x_i Y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n Y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

och $B_0 = \bar{Y} - B_1 \bar{x}$.

- E_i antas vara normalfördelad med väntevärde 0 och varians σ^2 samt att avvikelserna är oberoende, dvs.
 - Y_i är oberoende och normalfördelade.
 - Väntevärdet av Y_i ges av $\beta_0 + \beta_1 x_i$.
 - Variansen av Y_i är σ^2 ,
 alltså $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i \in \{1, \dots, n\}$, oberoende.

Egenskaper hos minsta-kvadrat-estimatorer

- Vi vill undersöka B_0 och B_1 (alltså deras fördelning, med hjälp av konfidensintervall och hypotesprövning).
- För det visar sig följande lätt verifierade räkneregler hjälpsamma:

- $\sum_{i=1}^n (x_i - \bar{x}) = (\sum_{i=1}^n x_i) - n\bar{x} = 0$

- $\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i - \bar{x})x_i$
 $= \frac{1}{n} \left(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right)$

- $\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^n (x_i - \bar{x})Y_i$
 $= \frac{1}{n} \left(n \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \sum_{i=1}^n Y_i \right)$

- $$B_1 = \sum_{j=1}^n c_j Y_j, \quad \text{där } c_j = \frac{x_j - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$E[B_1] = \beta_1 \quad \text{samst} \quad \text{Var}[B_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

- Sammantaget följer alltså att B_1 är **väntevärdesriktig** och $B_1 \sim \mathcal{N}(\beta_1, \sigma^2(\sum_{i=1}^n (x_i - \bar{x})^2)^{-1})$ (jfr. föreläsning 7).
- Analoga beräkningar baserat på $B_0 = \bar{Y} - B_1 \bar{x}$ leder till $B_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2})$ (jfr. s.389-392 i boken).

- $$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2}{n-2}$$

är en väntevärdesriktig estimator för variansen σ^2 (se uppgift 11.13 i boken). Dess utfall beräknas som $s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$.

Sammanfattning (jfr. s.393 i boken)

Låt oss skriva $s_{xx} := \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{yy} := \sum_{i=1}^n (Y_i - \bar{Y})^2$ och $S_{xy} := \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$.

- ① $B_1 = \frac{S_{xy}}{s_{xx}}$ är en väntevärdesriktig estimator för β_1 , som är normalfördelad och har varians $\frac{\sigma^2}{s_{xx}}$.
- ② $B_0 = \bar{Y} - B_1 \bar{x}$ är en väntevärdesriktig estimator för β_0 , som är normalfördelad och har varians $\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n s_{xx}}$.
- ③ $SSE = S_{yy} - B_1 S_{xy}$.
- ④ $S^2 = \frac{SSE}{n-2}$ är en väntevärdesriktig estimator för σ^2 .

Inferens på β_1 (lutningen)

- Då $B_1 \sim \mathcal{N}(\beta_1, \sigma^2/s_{xx})$, gäller $\frac{B_1 - \beta_1}{\sigma/\sqrt{s_{xx}}} \sim \mathcal{N}(0, 1)$.
- Man kan visa att $\frac{(n-2)S^2}{\sigma^2} = \frac{SSE}{\sigma^2}$ har en χ^2 fördelning med $n-2$ frihetsgrader, samt att B_1 och S^2 är oberoende under våra förutsättningar.
Därmed är $\frac{B_1 - \beta_1}{S/\sqrt{s_{xx}}}$ t -fördelad med $n - 2$ frihetsgrader.
(jfr. föreläsning 8)
- Ett $100(1 - \alpha)\%$ konfidensintervall för β_1 ges därför av

$$B_1 \pm t_{\alpha/2} \frac{S}{\sqrt{s_{xx}}},$$

där $t_{\alpha/2}$ som vanligt betecknar talet så att $\mathbb{P}(T > t_{\alpha/2}) = \frac{\alpha}{2}$.

Inferens på β_1 (lutningen)

- När det gäller hypotesprövning ($H_1 : \beta_1 \neq \beta_1^0$, $\beta_1 < \beta_1^0$ eller $\beta_1 > \beta_1^0$) är

$$T = \frac{B_1 - \beta_1^0}{S/\sqrt{s_{xx}}}$$

motsvarande teststatistik.

Anmärkning: Ofta tar man $\beta_1^0 = 0$ för att undersöka om det faktiskt finns ett linjärt samband mellan X and Y , som då resulterar i en icke-trivial lutning av regressionslinjen.

Exempel

Låt oss undersöka i föregående exempel om det faktiskt finns ett linjärt samband mellan X och Y (vilket motsvarar $\beta_1 \neq 0$, som blir vår alternativ hypotes) med signifikansnivå $\alpha = 0.05$.

Låt alltså nollhypotesen vara $H_0 : \beta_1 = 0$, vilket leder till ett tvåsidigt test. Vi beräknade $b_1 = 1.499$ och får dessutom

$$s_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \left(n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 \right) = 0.5838$$

$$s_{yy} = 8 \text{ och } s_{xy} = 0.875.$$

Med $SSE = s_{yy} - b_1 s_{xy} = 8 - 1.499 \cdot 0.875 = 6.688$ och

$$s^2 = \frac{SSE}{10-2} = 0.836 \text{ blir utfallet på teststatistiken } T:$$

$$t = \frac{b_1 - 0}{\sqrt{s^2/s_{xx}}} = \frac{1.498}{\sqrt{0.84/0.584}} = 1.37$$

Därför att $t_{0.025} = 2.306 > 1.37$ förkastar vi inte nollhypotesen.

Man kan inte dra slutsatsen att det finns ett linjärt samband mellan X och Y på signifikansnivån $\alpha = 0.05$.

Inferens på β_0 (y -skärningen)

- Då $B_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2}{ns_{xx}} \sum_{i=1}^n x_i^2)$, gäller

$$\frac{B_0 - \beta_0}{\sigma \sqrt{\sum_{i=1}^n x_i^2 / ns_{xx}}} \sim \mathcal{N}(0, 1).$$

- Under våra förutsättningar är också B_0 och S^2 oberoende, vilket innebär att

$$\frac{B_0 - \beta_0}{\frac{S}{\sqrt{ns_{xx}}} \cdot \sqrt{\sum_{i=1}^n x_i^2}}$$

också är t -fördelad med $n - 2$ frihetsgrader.

Inferens på β_0 (y -skärningen)

- Ett $100(1 - \alpha)\%$ konfidensintervall för β_0 ges då av

$$B_0 \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{ns_{xx}}} \sqrt{\sum_{i=1}^n x_i^2}.$$

- Motsvarande teststatistik för hypotesprövning är

$$T = \frac{B_0 - \beta_0^0}{\frac{S}{\sqrt{ns_{xx}}} \cdot \sqrt{\sum_{i=1}^n x_i^2}}$$

och återigen $\beta_0^0 = 0$ det vanligaste valet.

Exempel

Ett 95% KI för β_0 i exemplet ovan ges av

$$14.489 \pm 2.306 \sqrt{\frac{0.836}{10 \cdot 0.5838}} \sqrt{28.64}$$

alltså

$$(14.49 - 4.67, 14.49 + 4.67) = (9.82, 19.16)$$

Med andra ord är vi 95% säkra att den verkliga regressionslinjen skär y -axeln mellan $y = 9.82$ och $y = 19.16$.

Inferens på skattat väntevärde och enskilda prognosvärden

- Givet ett nytt värde x för X , vill vi gärna använda regressionen för att skatta värden $\mu_{Y|x}$ och $Y|x$.
- En punktskattning för $\mu_{Y|x}$ och $Y|x$ ges av

$$\hat{Y}|x = \hat{\mu}_{Y|x} = B_0 + B_1 x.$$

- Ett $100(1 - \alpha)\%$ KI för $\mu_{Y|x}$ ges av

$$\hat{\mu}_{Y|x} \pm t_{\alpha/2} \cdot S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}}$$

- Ett $100(1 - \alpha)\%$ "prognosintervall" för värdet $Y|x$ ges av

$$\hat{Y}|x \pm t_{\alpha/2} \cdot S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{s_{xx}}}$$

(härledning, se s.399-401 i boken).