MVE051/MSG810 2016 Föreläsning 14

Petter Mostad

Chalmers

December 14, 2016

Beroende och oberoende variabler

- Hittills i kursen har vi tittat på modeller där alla observationer representeras av stokastiska variable. Variablerna antas vara stickprov från modeller med vissa parametrar. Vi estimerar dessa parametrar, och kan sen göra prediktioner från modellerna när parametrarna är lika med estimaterna.
- Ett alternativ är att vi bara modellerar hur vissa variabler (de *beroende variablerna*) beror på andra variabler (de *oberoende variablerna*). Denna typ modeller kallas *regressionsmodeller*.
- Oberoende variablar kan även kallas *förklarande* variablar, eller *prediktorer*. De är inte stokastiska variabler, men vanliga matematiska variabler. Vid experiment är de värden som fastställs av experimenteraren oberoende variabler.
- Beroende variabler kan även kallas *responsvariabler*. De är stokastiska variable med en sannolikhetsfördelning för varje kombination av värden för de oberoende variablen.
- Märk: Det är inte alltid uppenbart vilka variable som skall vara oberoende eller beroende. Men valet gör skillnad i analysen!

Enkel regression

- I denna kursen tittar vi bara på situationen där vi har *en* oberoende variabel x och *en* beroende variabel Y: Detta kallas *enkel regression (simple regression)*.
- I denna kursen tittar vi bara på situationen där *väntevärdet* för Y givet x, $\mu_{Y|x}$, är en *linjär* funktion av x, alltså där det finns konstanter β_0 och β_1 så att för alla värden av x har vi

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

  Detta kallas *enkel linjär regression (simple linear regression)*.
- Det finns metoder för att undersöka om en sådan modell är rimlig för data: För oss gäller att vi behöver göra ett *scatterplot* för att värdera om linjär regression kan vara en rimlig modell.
- Kanske kan en *transformation* av data göra att en linje passar bättre?

Minimum residualkvadratsumma

- Anta vi har observerade data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ och vill hitta *interceptet* β_0 och *lutningskvoten* β_1 så att linjen $\beta_0 + \beta_1 x$ passar till data.
- För fixerade värden β_0 och β_1 så kan vi för $i = 1, 2 \ldots, n$ skriva

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

  där $e_1, e_2, \ldots, e_n$ är *residualerna*.
- En standard metod för att estimera β_0 och β_1 är att välja de som minimerar residulalkvadratsumman $\sum_{i=1}^{n} e_i^2$ (kallas även SSE).
- Minimum kan hittas genom att sätta lika med noll de partiella deriverade med hänsyn till β_0 och β_1 av residualkvadratsumman.

β_0 och β_1 som minimerar residualkvadratsumman

- Om vi definerar

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

$$\begin{aligned} \hat{\beta}_1 &= S_{xy}/S_{xx} \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned}$$

  så minimerar $\hat{\beta}_0$ och $\hat{\beta}_1$ residualkvadratsumman.
- Med dessa $\hat{\beta}_0$ och $\hat{\beta}_1$ får vi tex. att $\sum_{i=1}^{n} e_i = 0$.
- Märk: Om vi *innan beräkningarna* centrerar data, dvs. drar av $\overline{x}$ från alla värden x_i och motsvarande för y, så blir formlerna mycket enkla:

$$\begin{aligned} \hat{\beta}_1 &= \sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 \\ \hat{\beta}_0 &= 0 \end{aligned}$$

Beräkningsformler för $\hat{\beta}_0$ och $\hat{\beta}_1$

- Det snabbaste sättet att beräkna $\hat{\beta}_0$ och $\hat{\beta}_1$ direkt från data är följande formler:

$$\begin{aligned} \hat{\beta}_1 &= \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned}$$

- I praktiken:
  1. Beräkna $\sum_{i=1}^n x_i$, $\sum_{i=1}^n x_i^2$, $\sum_{i=1}^n y_i$, och $\sum_{i=1}^n x_i y_i$ från data.
  2. Sätt in i formlerna över.

Modellantagningar

- Vi önskar göra som vi brukar: Titta på egenskaperna till *estimatorerna* över för β_0 och β_1 som stokastiska variabler.
- Därmed måste vi beskriva exakt de stokastiska variablerna $Y_1, Y_2, \ldots, Y_n$: Vi *antar* $Y_1, Y_2, \ldots, Y_n$ är oberoende och, för $i = 1, \ldots, n$,

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 x_i, \sigma)$$

  där σ är en tredje parameter i modellen.
- En (väntevärdesrätt) estimator för σ^2 är

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\text{SSE}}{n-2}.$$

Estimatorernas fördelningar

- Estimatorerna $\hat{\beta}_1$, $\hat{\beta}_0$ och $\hat{\sigma}^2$ har med antagningarna över oberoende fördelningar som kan specificeras med

$$\hat{\beta}_1 \sim \text{Normal}\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right)$$

$$\hat{\beta}_0 \sim \text{Normal}\left(\beta_0, \frac{\sigma\sqrt{\sum_{i=1}^n x_i^2/n}}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right)$$

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

- Genom att kombinera dessa fördelningar kan vi ta fram formler för konfidensintervaller och tester för β_0, β_1 och andra värden.

Hypotestest och konfidensintervall för lutningskvoten β_1

- Antag vi vill jämföra $H_0 : \beta_1 = b_1$ med $H_1 : \beta_1 \neq b_1$ för någon känd b_1 (oftast $b_1 = 0$).
- Test statistikan är $(\hat{\beta}_1 - b_1)/(\hat{\sigma}/\sqrt{S_{xx}})$ och den har fördelning $T(n-2)$ när H_0 är sann. Förkastningsområdet ligger därför utanför ett intervall $[-t_{\alpha/2}, t_{\alpha/2}]$ på samma sätt som liknande hypotestest.
- Vi får även ett konfidensintervall för β_1 på formen

$$\hat{\beta}_1 \pm t_{\alpha/2}\hat{\sigma}/\sqrt{S_{xx}}$$

- Märk: Nollhypotesen $H_0 : \beta_1 = 0$ motsvarar att det inte finns något (linjärt) samband mellan x och Y!

Hypotestest och konfidensintervall för interseptet β_0

- Antag vi vill jämföra $H_0 : \beta_0 = b_0$ med $H_1 : \beta_0 \neq b_0$ för någon känd b_0 (oftast $b_0 = 0$).
- Test statistikan är $(\hat{\beta}_0 - b_0)/\left(\hat{\sigma}\sqrt{\sum_{i=1}^n x_i^2}/\sqrt{nS_{xx}}\right)$, och den har fördelning $T(n-2)$ när H_0 är sann. Förkastningsområdet ligger därför utanför ett intervall $[-t_{\alpha/2}, t_{\alpha/2}]$ på samma sätt som liknande hypotestest.
- Vi får även ett konfidensintervall för β_0 på formen

$$\hat{\beta}_0 \pm t_{\alpha/2}\frac{\hat{\sigma}\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}$$

- Märk: För vissa tillämpningar är det naturligt med en modell $\beta_0 = 0$.

Konfidensintervall för väntevärdet $\mu_{Y|x}$ för ny observation

- Antag vi vill studera *väntevärdet* $\mu_{Y|x}$ för en ny observation Y givet ett (nytt) värde av x.
- Vi har estimatorn $\hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$.
- Vi får ett konfidensintervall på formen

$$\hat{\mu}_{Y|x} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}}$$

  där $t_{\alpha/2}$ är kvantil i en $T(n-2)$ fördelning.
- Innan man använder modellen som över måste man diskutera om den är giltig för den givna x: Det är lättare att argumentera för *interpolation* än *extrapolation*!

Konfidensintervall för ny observation Y

- Antag vi vill studera *värdet* för en ny observation Y givet ett (nytt) värde av x.
- Estimatorn har samma väntevärde som över: $\hat{Y} \mid x = \hat{\mu}_{Y|x} = \hat{\beta}_0 + \hat{\beta}_1 x$. Men variansen är större!
- Vi får ett konfidensintervall på formen

$$\hat{Y} \mid x \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}}}$$

  där $t_{\alpha/2}$ är kvantil i en $T(n-2)$ fördelning.
- Innan man använder modellen som över måste man diskutera om den är giltig för den givna x: Det är lättare att argumentera för *interpolation* än *extrapolation*!