Kursus 02403
Introduktion til Statistik

Klaus Kaae Andersen

Informatics and Mathematical Modelling
Building 321 - room 011
Technical University of Denmark
2800 Lyngby – Denmark
e-mail: kka@imm.dtu.dk

Regressionsanalyse (kap 11)

- Korrelation
- Simpel lineær regression
- Mindste kvadraters metode
- Inferens i en simpel lineær regressionsmodel

Korrelation

- Korrelationskoefficienten r angiver den lineære sammenhæng mellem variablerne x og y
- Korrelationskoefficienten mellem 2 variable x og y estimeres ved

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- Det antages her, at observationerne (x_i, y_i) er sammenhørende værdier. Der gælder $r \in [-1\ 1]$

Regressionsanalyse (kap 11)

- Antag at Y er en stokastisk variabel. Vi er interesseret i at modellere Y's afhængighed af en *forklarende variabel* x
- Vi undersøger en *lineær* sammenhæng mellem Y og x, dvs. ved en regressionsmodel på formen

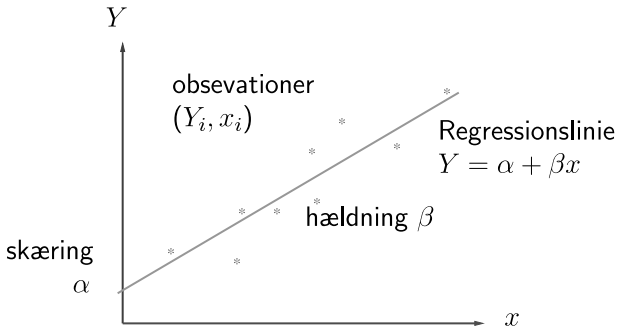
$$Y = \alpha + \beta x + \varepsilon$$

Simpel lineær regressionsmodel

$$Y = \underbrace{\alpha + \beta x}_{\text{model}} + \underbrace{\varepsilon}_{\text{residual}}$$

- Y afhængig variabel
- x uafhængig variabel
- α skæring med Y-akse
- β hældning
- ε residual (tilfældig fejl)

Simpel lineær model



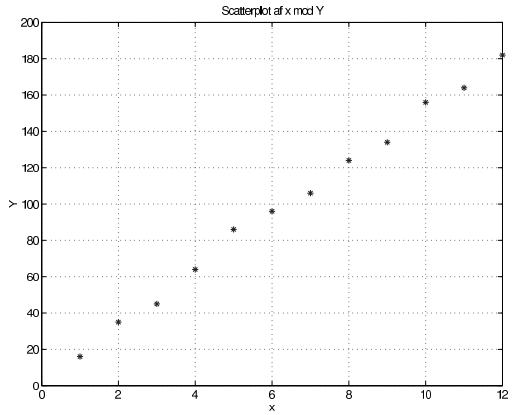
Mindste kvadraters metode

- Antag at vi har observationerne

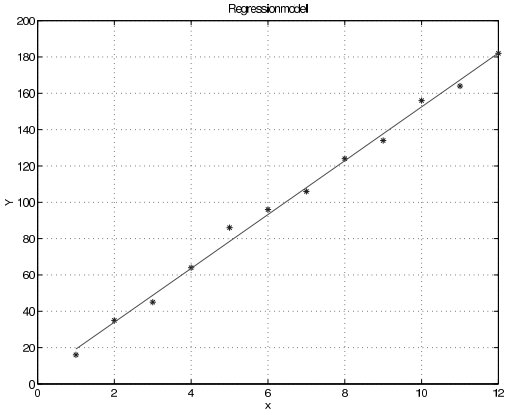
x	1	2	3	4	5	6	7	8	9	10	11	12
y	16	35	45	64	86	96	106	124	134	156	164	182

- Er der en sammenhæng mellem x og y?
- Vi foreslår en model på formen $\hat{y} = a + bx$
- Hvordan estimeres a og b?

Mindste kvadraters metode



Mindste kvadraters metode



Mindste kvadraters metode

- Vi definerer

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

Mindste kvadraters metode

- a og b bestemmes ved

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b \cdot \bar{x}$$

- a og b er nu de værdier, der giver den regressionslinie, der minimerer den kvadratiske afstand mellem punkter og linie
- a er et *estimat* for α og b er et *estimat* for β

Mindste kvadraters metode

- I eksemplet fås

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = 143$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 31533$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 2119$$

samt $\bar{x} = 6.50$ og $\bar{y} = 100.67$

Mindste kvadraters metode

- Estimater for α og β:

$$b = \frac{S_{xy}}{S_{xx}} = \frac{2119}{143} = 14.82$$

$$a = \bar{y} - b \cdot \bar{x} = 100.67 - 14.82 \cdot 6.50 = 4.34$$

Modellen bliver:

$$\hat{y} = 4.34 + 14.82 \cdot x$$

Inferens i regressionsmodel

- Vi antager at de observerede data (Y_i, x_i) kan beskrives ved modellen

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

hvor det antages at ε_i er uafhængige normalfordelte stokastiske variable med middelværdi 0 og konstant varians σ^2

- Et estimat af σ^2 bliver

$$s_e^2 = \frac{S_{yy} - (S_{xy})^2/S_{xx}}{n-2}$$

Inferens i regressionsmodel

- Antag at vi vil teste en hypotese om skæring med y-aksen

$$H_0: \quad a = \alpha$$
$$H_1: \quad a \neq \alpha$$

- Teststørrelsen bliver

$$t = \frac{(a-\alpha)}{s_e}\sqrt{\frac{nS_{xx}}{S_{xx} + n(\bar{x})^2}}$$

- Kritisk værdi findes i t-fordelingen, $t_{\alpha/2}(n-2)$

Inferens i regressionsmodel

- Antag at vi vil teste en hypotese om hældningen β

$$H_0: \quad b = \beta$$
$$H_1: \quad b \neq \beta$$

- Teststørrelsen bliver

$$t = \frac{(b-\beta)}{s_e}\sqrt{S_{xx}}$$

- Kritisk værdi findes i t-fordelingen, $t_{\alpha/2}(n-2)$

Konfidensintervaller for α og β

- Konfidensinterval for α

$$a \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

- Konfidensinterval for β

$$b \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

Konfidensinterval for $\alpha + \beta x_0$

- Konfidensinterval for $\alpha + \beta x_0$ svarer til et konfidensinterval for modellen i punktet x_0

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Prædiktionsinterval for $\alpha + \beta x_0$

- Prædiktionsinterval for $\alpha + \beta x_0$ svarer til et prædiktionsinterval for modellen i punktet x_0

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Et prædiktionsinterval bliver altså større end et konfidensinterval for fastholdt α

Regressionsanalyse (kap 11)

- Korrelation
- Simpel lineær regression
- Mindste kvadraters metode
- Inferens i en simpel lineær regressionsmodel