Sovellettu todennäköisyyslaskenta B

Antti Rasila

30. marraskuuta 2007

1. Lineaarinen regressiomalli ja suurimman uskottavuuden menetelmä
   - Minimin löytäminen lineaarisessa tapauksessa
2. Epälineaarisista malleista
3. Linearisoituvat mallit
4. Usean muuttujan lineaarinen regressiomalli
   - Polynomimalli

Lineaarinen regressiomalli ja suurimman uskottavuuden menetelmä 2/2

- Tutkitaan havaintoja $(x_1, y_1), \ldots, (x_n, y_n)$. Oletetaan, että

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

  missä β_0, β_1 ovat regeressiomallin kertoimet ja ε_i:t ovat riippumattomia, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.
- Tällöin kullakin x:n arvolla havainto y on satunnaismuuttujan

$$Y \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$$

  saama arvo.

Lineaarinen regressiomalli ja suurimman uskottavuuden menetelmä 1/2

- Voidaan määrittää otoksen likelihood-funktio

$$L(\beta_0, \beta_1) = f_Y(y_1; \beta_0, \beta 1) \cdot \ldots \cdot f_Y(y_n; \beta_0, \beta 1),$$

missä f_Y on normaalijakauman tiheysfunktio:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(y - \beta_0 + \beta_1 x)^2}{2\sigma^2} \Big).$$

- Saadaan

$$L(\beta_0, \beta_1) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big(- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \Big).$$

Likelihood-funktio ja PNS-menetelmä

- Koska eksponenttifunktio on kaikkialla kasvava, likelihood-funktio $L(\beta_0, \beta_1)$ saavuttaa maksiminsa parametrien β_0, β_1 suhteen, kun neliösumma

$$F(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  saavuttaa miniminsä.
- **Seuraus:** Pienimmän neliösumman estimaatti parametreille β_0, β_1 on suurimma uskottavuuden estimaatti ko. parametreille.

Minimin löytäminen lineaarisessa tapauksessa 1/2

- Tarkastellaan funktiota

$$F(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Etsitään piste (β_0, β_1) siten, että

$$\nabla F(\beta_0, \beta_1) = 0.$$

- Lasketaan osittaisderivaatta

$$\frac{\partial}{\partial \beta_0} F(\beta_0, \beta_1) = 2\left(\beta_1 \sum x_i + n\beta_0 - \sum y_i\right).$$

- Ratkaistaan nollakohta

$$\beta_0 = \frac{1}{n} \sum y_i - \frac{\beta_1}{n} \sum x_i = \bar{y} - \beta_1 \bar{x}.$$

Minimin löytäminen lineaarisessa tapauksessa 2/2

- Lasketaan seuraavaksi osittaisderivaatta

$$\frac{\partial}{\partial \beta_1} F(\beta_0, \beta_1) = 2\big(\beta_0 \sum x_i + \beta_1 \sum x_i^2 - \sum x_i y_i\big).$$

- Sijoittamalla β_0:n lauseke saadaan

$$n\bar{x}\bar{y} - n\beta_1 \bar{x}^2 + \beta_1 \sum x_i^2 - \sum x_i y_i = 0.$$

- Ratkaistaa nollakohta

$$\beta_1 = \frac{n\bar{x}\bar{y} - \sum x_i y_i}{n\bar{x}^2 - \sum x_i^2} = \frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})(y_i - \bar{y})}.$$

Esimerkki: 2. asteen sovitus 1/3

- Tutkitaan lisäaineen määrän x vaikutusta kuivumisaikaan y. Eri lisäaineen määrillä x_i (grammaa) saatiin kuivumisajat y_i (tuntia), $i = 1, \ldots, 9$:

x_i	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0
y_i	11.0	9.4	9.1	7.0	6.2	7.1	6.6	7.5	8.2

- Huomataan, että kuivumisajan riippuvuus lisäaineen määrästä on epälineaarista. Minimikohdan estimoimiseksi sovitetaan havaintoihin paraabeli $y = \beta_0 + \beta_1 x + \beta_2 x^2$.
- Pienimmän neliösumman yhtälöryhmä mallille on

$$\frac{\partial}{\partial \beta_k} \sum (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2 = 0, \qquad k = 0, 1, 2.$$

Esimerkki: 2. asteen sovitus 2/3

- Näistä saadaan yhtälöryhmä

$$\begin{cases} n\beta_0 + \beta_1 \sum x_i + \beta_2 \sum x_i^2 &= \sum y_i, \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 &= \sum x_i y_i \\ \beta_0 \sum x_i^2 + \beta_1 \sum x_i^3 + \beta_2 \sum x_i^4 &= \sum x_i^2 y_i. \end{cases}$$

- Laskemalla yhtälöryhmän kertoimet havainnoista saadaan

$$\begin{cases} 9\beta_0 + 36\beta_1 + 204\beta_2 &= 72.1 \\ 36\beta_0 + 204\beta_1 + 1296\beta_2 &= 266.6 \\ 204\beta_0 + 1296\beta_1 + 8772\beta_2 &= 1515.4 \end{cases}$$

- Ratkaisuna ovat estimaatit $\hat{\beta}_0 = 11.15$, $\hat{\beta}_1 = -1.806$ ja $\hat{\beta}_2 = 0.1803$. Pienimmän neliösumman mielessä parhaiten havaintoihin liittyvä paraabeli on siis

$$y = 11.15 - 1.806x + 0.1803x^2.$$

Esimerkki: 2. asteen sovitus 3/3

- Havainnoista voidaan edelleen laskea kokonaisneliösumma

$$SST = \sum (y_i - \bar{y})^2 = 19.469,$$

jäännösneliösumma

$$SSE = \sum (y_i - 11.15 + 1.806x_i - 0.1803x_i^2)^2 = 1.525,$$

ja mallineliösumma

$$SSM = SST - SSE = 17.944.$$

- Selitysasteeksi saadaan

$$R^2 = \frac{SSM}{SST} = \frac{17.944}{19.469} = 0.922.$$

Linearisoituvat mallit

- Toisinaan funktio, joka ei ole lineaarinen parametrien suhteen, voidaan saattaa sopivalla muunnoksella lineaariseksi.
- Esimerkki tällaisesta tapauksesta on funktio

$$y = \beta_0 e^{\beta_1 x},$$

joka linearisoituu lograritmin avulla:

$$\ln(y) = \ln(\beta_0) + \beta_1 x.$$

- Nyt voidaan sovittaa suora aineistoon, jossa arvot on logaritmoitu.

Esimerkki (Laininen) 1/2

- Vesialtaan happipitoisuus DO (mg/l) riippuu altaan lämpötilasta ($^\circ C$) yhtälön

$$DO = ae^{-bT}$$

  mukaisesti, missä $a, b > 0$.

- Parametrien estimoimiseksi mitattiin erilaisilla T:n arvoilla t_i vastaavat DO:n arvot d_i:

t_i	23.1	24.4	25.0	26.5	27.1	28.0
d_i	4.7	3.2	3.2	2.8	2.8	2.5

- Ottamalla logaritmi saadaan

$$\ln(DO) = \ln(a) - bT.$$

Esimerkki (Laininen) 2/2

- Merkitään $t_i = x_i$ ja $\ln(d_i) = y_i$. Saadaan

$$\hat{\beta}_1 = -b = \frac{-1.853}{16.828} = -0.110.$$

ja

$$\hat{\beta}_0 = \ln(a) = 1.141 - (-0.110) \cdot 25.683 = 3.966.$$

- Saadaan siis

$$\ln(DO) = 3.966 - 0.110 \cdot T,$$

eli

$$DO = 52.77 \cdot e^{-0.110 \cdot T}.$$

Usean muuttujan lineaarinen regressiomalli 1/4

- Oletetaan suure y riippuu useasta muuttujasta x_i, $i = 1, 2, \ldots, m$ ja riippuvuus on lineaarista.
- Voidaan kirjoittaa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m.$$

- Oletetaan edelleen, että käytettävissä on tilastoaineisto, jossa suureesta y ja jokaisesta selittävästä muuttujasta x_i on n havaintoa.
- Käytännön syistä on oletettavas myös, että $n \geq m + 1$ (ja mieluiten tietysti n on vieläkin suurempi).
- Merkitään havaintoja y_k, x_{ik}, $i = 1, \ldots, m$, $k = 1, \ldots, n$,

Usean muuttujan lineaarinen regressiomalli 2/4

- Kuten yhden muuttujan tapauksessa, malli saadaan minimoimalla funktio

$$f(\beta_0, \ldots, \beta_m) = \sum_{k=1}^{n} \left(y_k - (\beta_0 + \beta_1 x_{1k} + \ldots + \beta_m x_{mk})\right)^2.$$

- Minimi saadaan etsimällä piste, jossa $\nabla F(\beta_0, \ldots, \beta_m) = 0$. Nollakohdat ratkaisemalla päädytään yhtälöryhmään, jotka kutsutaan normaaliryhmäksi. Kertoimet saadaan ratkaisemalla normaaliryhmä.

Usean muuttujan lineaarinen regressiomalli 3/4

- Kirjoitetaan nyt regressiomalli muotoon

$$y = X\beta, \tag{1}$$

missä X on kerrointen x_{ik} muodostama $n \times m$-matriisi:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{m1} \\ 1 & x_{12} & x_{22} & \dots & x_{m2} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix} \text{ ja } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_m \end{pmatrix}.$$

Usean muuttujan lineaarinen regressiomalli 4/4

- Valitettavasti yleensä $n \neq m+1$, joten X ei ole neliömatriisi. Kertomalla matriisiyhtälö (1) puolittain X:n transpoosilla X^T, saadaan normaaliyhtälö muotoon

$$X^T y = X^T X \beta,$$

  missä $X^T X$ on $(m+1) \times (m+1)$-neliömatriisi.
- Yhtälön ratkaisu on

$$\beta = (X^T X)^{-1} X^T y,$$

  olettaen tietenkin, että käänteismatriisi $(X^T X)^{-1}$ on olemassa.

Esimerkki

- Etsitään aineistoon

x_k	-2.2	0	3.5	4.2
y_k	-8.8	0.2	3.9	5.2

- Erityisesti yhden muuttujan lineaarinen regressiomalli on erikoistapaus:

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

- Ratkaistaan

$$\beta = (X^T X)^{-1} X^T y = \begin{pmatrix} -2.616 \\ 1.994 \end{pmatrix},$$

missä

$$y = \begin{pmatrix} -8.8 \\ 0.2 \\ 3.9 \\ 5.2 \end{pmatrix} \text{ ja } X = \begin{pmatrix} 1 & -2.2 \\ 1 & 0 \\ 1 & 3.5 \\ 1 & 4.2 \end{pmatrix}.$$

Polynomimalli

- Polynomimallissa sovitetaan tilastoaineistoon astetta n oleva polynomi

$$y = c_0 + c_1 x + c_2 x^2 + \ldots + c_n x^n$$

.

- Malli palautuu usean muuttujan lineaariseksi malliksi sijoittamalla $x_1 = x$, $x_2 = x^2$, jne.