Cvičenia z ekonometrie
Letný semester 2008/2009

Lineárna regresia 1 - základné pojmy

Vo všetkých úlohách uvažujeme regresný model s k parametrami

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_{k-1} x_{k-1,i]} + \varepsilon_i, \tag{1}$$

pričom náhodné chyby ε_i sú nezávislé a majú normálne rozdelenie $N(0, \sigma^2)$. Model odhadujeme z n dát.

1. Odhadujme parametre metódou maximálnej vierohodnosti. Dokážte, že odhadovanie parametrov β_i vedie k metóde najmenších štvorcov (t.j. minimalizácii súčtov druhých mocnín odchýliek skutočných a odhadnutých hodnôt Y). Ako sa odhadne parameter σ^2?

2. Nech $\hat{\beta}_i$ sú odhady parametrov β_i získané metódou najmenších štvorcov. Definujme TSS (total sum of squares, celková suma štvorcov), ESS (explained sum of squares, vysvetlená suma štvorcov), RSS (total sum of squares, celková suma štvorcov) nasledovne:

   $$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \; ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2, \;\; RSS = \sum_{i=1}^{n} (Y_i - \hat{Y})^2,$$

   kde $\bar{Y}$ je aritmetický priemer a $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_{k-1} x_{k-1,i}$.

   - Dokážte, že ak model (1) obsahuje konštantný člen β_0, tak platí

     $$TSS = ESS + RSS.$$

   - Na základe predchádzajúceho výsledku ukážte, že nasledovné definície koeficientu determinácie sú ekvivalentné:

     $$R^2 = \frac{ESS}{TSS}, \quad R^2 = 1 - \frac{RSS}{TSS}.$$

     Dokážte, že R^2 je z intervalu $[0, 1]$.
   - Ukážte na príklade, že predchádzajúce tvrdenia neplatia, ak model (1) neobsahuje konštantný člen (napr. ak máme model $Y_i = \beta x_i + \varepsilon_i$).
   - Dokážte, že po pridaní ďalšej premennej do modelu (1) sa nemôže koeficient determinácie zmenšiť.

3. Nech ℓ je logaritmus funkcie vierohodnosti po dosadenení maximálne vierohodných odhadov β_i (pre $i = 0, 1, \ldots k-1$) a σ^2 . Definujme Akaikeho informačné kritérium (AIC) a Schwarzovo informačné kritérium (SIC):

   $$AIC = -\frac{2\ell}{n} + \frac{2k}{n}, SIC = -\frac{2\ell}{n} + \frac{k \, \ln n}{n}.$$

   Ukážte, že sa dajú napísať v tvare

   $$AIC = 1 + \ln(2\pi) + \ln\left(\frac{RSS}{n}\right) + \frac{2k}{n}, SIC = 1 + \ln(2\pi) + \ln\left(\frac{RSS}{n}\right) + \frac{k \, \ln n}{n}.$$

   Na základe týchto vzťahov vysvetlite, ako sa AIC a SIC dajú použiť na porovnávanie kvality dvoch modelov pre tie isté dáta.

4. Zapíšme model (1) v maticovom tvare: $Y = (Y_1, \dots Y_n)^T$ $\varepsilon = (\varepsilon_1, \dots \varepsilon_n)^T$, vektor parametrov $\beta = (\beta_0, \beta_1, \dots \beta_{k-1})^T$, X je matica rozmeru $n \times k$ s vysvetľujúcimi premennými (prvý stĺpec sú jednotky zodpovedajúce konštantnému členu, druhý stĺpec sú pozorovanie premennej x_1 atď). To znamená, že

$$Y = X\beta + \varepsilon,$$

pričom ε má normálne rozdelenie $N(0, \sigma^2 I)$. Predpokladajme, že stĺpce matice X sú lineárne nezávislé.

Dokážte, že metódou najmenších štvorcov dostaneme odhad

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Načo potrebujeme predpoklad o lineárnej nezávislosti stĺpcov? Čo sa stane, ak sú závislé?

5. Uvažujme regresnú priamku

$$Y_i = \alpha + \beta x_y + \varepsilon_i.$$

Dokážte, že MNŠ odhady $\hat{\alpha}$, $\hat{\beta}$ sa dajú zapísať v tvare

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$