Uporabna statistika

Gregor Dolinar

Fakulteta za elektrotehniko
Univerza v Ljubljani

17. december 2013

Preverjanje hipotez

Hočemo preveriti hipotezo, da je smerni koeficient premice enak $\beta_{1,0}$, torej

$$H_0 : \beta_1 = \beta_{1,0}$$

$$H_1 : \beta_1 \neq \beta_{1,0}$$

Upoštevamo, da je

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}^2 / S_{xx}}$$

t-porazdeljena slučajna spremenljivka z $n-2$ prostostnimi stopnjami. (Vemo, da je $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$.)

Poseben primer.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Če ne moremo zavrniti hipoteze $H_0 : \beta_1 = 0$, potem je bodisi slučajna spremenljivka Y neodvisna od x in je $\hat{y} = \overline{Y}$, bodisi odvisnost ni linearna.
Če pa zavrnemo hipotezo $H_0 : \beta_1 = 0$, potem je bodisi slučajna spremenljivka Y linearno odvisna od x bodisi odvisnost ni linearna, vendar linearni člen delno vpliva na odvisnost.

Opomba

S preverjanjem domneve $H_0 : \beta_1 = 0$ preverjamo moč (signifikanco) regresije.
Moč (signifikanco) regresije lahko preverjamo tudi s pomočjo analize variance (analysis of variance - ANOVA).

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$$

Delimo obe strani enačbe z n.
Celotna varianca je enaka vsoti pojasnjene variance zaradi linearnosti pojava in nepojasnjene variance glede na linearnost pojava (ki je enaka vsoti napak).

Interval zaupanja

Interval zaupanja za povprečno vrednost Y pri dani vrednosti x_0.

$$E(Y|x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0.$$

Ocena

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Varianca za $\hat{\mu}_{Y|x_0}$

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right)$$

Potem je

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)}}$$

t-porazdeljena slučajna spremenljivka z $n-2$ prostostnimi stopnjami.

Pri določanju regresijske premice smo predpostavili:

- napake ϵ so nekorelirane slučajne spremenljivke s povprečjem 0 in konstantno varianco
- za preverjanje domnev in računanje intervalov zaupanja potrebujemo predpostavko, da so ε normalno porazdeljene slučajne spremenljivke
- utemeljeno domnevamo, da je odvisnost v modelu prvega reda (linearna)

Vedno se moramo vprašati, ali so naše predpostavke upravičene. Upravičenost predpostavk moramo preverjati.
Na primer:

- graf
- determinacijski koeficient:

$$R^2 = \frac{SS_R}{SS_T} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$0 \leq R^2 \leq 1$$

Delež pojasnjene variance v celotni varianci.

Včasih lahko nelinearne funkcije pretvorimo v linearne. Na primer:

- $$Y = \beta_0 e^{\beta_1 x} \epsilon$$

  $$\log Y = \log \beta_0 + \beta_1 x + \log \epsilon$$

- $$Y = \beta_0 + \beta_1 \frac{1}{x} + \epsilon$$

  $$Y = \beta_0 + \beta_1 z + \epsilon$$

Definicija

Model, pri katerem nastopa več neodvisnih (pojasnjevalnih) spremenljivk (regresorjev), imenujemo model multiple regresije

$$Y = Y(X_1, \ldots, X_n).$$

Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon,$$

imenujemo model linearne multiple regresije.

Opomba

- Graf $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ je ravnina.
- Za predpisano vrednost $E(Y) = y_0$ ležijo vrednosti pojasnjevalnih spremenljivk x_1 in x_2 na premici.

Tudi bolj zapletene nelinearne modele lahko do določene mere obravnavamo kot modele multiple regresije.

Primer

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

$x_3 = x_1 x_2$,

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Pozor - pojasnjevalne spremenljivke potem niso več neodvisne!
Graf ni več ravnina.