Sztochasztikus rendszerek matematikája

2. gyakorlat

2017. szeptember 26. kedd: 2-4; 4-6 és szeptember 28. csütörtök: 2-4

[A] Lineáris regresszió egy magyarázó változó esetén
[B] Legkisebb négyzetek módszere - több magyarázó változó esete

[A] Lineáris regresszió egy magyarázó változó esetén

Jelölje $x = (x_1, ..., x_n)$ a magyarázó változó értékeiből álló vektort.
Jelölje $y = (y_1, ..., y_n)$ a magyarázott azaz a célváltozó értékeiből álló vektort.
Jelölje $\overline{x} = \sum_{k=1}^{n} x_k/n$, $\overline{y} = \sum_{k=1}^{n} y_k/n$ az átlagértékeket.

Az ÖSSZESÍTŐ TÁBLA értelmezése

`r értéke` a tapasztalati korreláció az x és az y közt
$r = \sum_{k=1}^{n}(y_k - \overline{y})(x_k - \overline{x})/\sqrt{\sum_{k=1}^{n}(y_k - \overline{y})^2 \sum_{k=1}^{n}(x_k - \overline{x})^2}$
`r négyzet` a tapasztalati korreláció négyzete: r^2
`Korrigált r-négyzet`: $1 - (1 - r^2)(n-1)/(n-2)$
A modell becsült paraméterei
$\widehat{\beta}_1 = \sum_{k=1}^{n}(y_k - \overline{y})(x_k - \overline{x})/\sum_{k=1}^{n}(x_k - \overline{x})^2$
$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$
A célváltozó modell szerint becsült értéke: $\widehat{y}_k = \widehat{\beta}_0 + \widehat{\beta}_1 x_k$, $k = 1, ..., n$
Mindig érvényes: $\overline{y} = \overline{\widehat{y}} = \sum_{k=1}^{n} \widehat{y}_k/n$
`Standard hiba` a modell hiba korrigált tapasztalati szórása: $s = \sqrt{\sum_{k=1}^{n}(\widehat{y}_k - y_k)^2/(n-p)}$
`Megfigyelések` a megfigyelésszám n

A VARIANCIA ANALÍZIS táblázat értelmezése

A sorelnevezések és fejléc:

`Regresszió`: a modell négyzetes kitérése

`Maradék`: a hiba négyzetes kitérése

`Összesen`: az y megfigyelések négyzetes kitérése

`df` – "degree of freedom", magyarul: szf – szabadságfok

`SS` – "Sum of Squares", magyarul: négyzetösszeg

`MS` – "mean SS", magyarul: az átlagos négyzetösszegek

`F` – azaz "F value", magyarul: az F próba értéke

`F szignifikanciája`, másként mondva: az F próba p-értéke

A táblázat adatai

A szabadságfokok: modell $df_R = p - 1$, hiba $df_{\mathtt{M}} = n - p$, mfi $df_{\mathtt{Ö}} = n - 1$

A négyzetösszegek: $\mathtt{SS}_{\mathtt{R}} = \sum_{k=1}^{n}(\widehat{y}_k - \overline{y})^2$, $\mathtt{SS}_{\mathtt{M}} = \sum_{k=1}^{n}(\widehat{y}_k - y)^2$, $\mathtt{SS}_{\mathtt{Ö}} = \sum_{k=1}^{n}(y_k - \overline{y})^2$

Az átlagos négyzetösszegek: $\mathtt{MS}_{\mathtt{R}} = \mathtt{SS}_{\mathtt{R}}/\mathtt{df}_{\mathtt{R}}$, $\mathtt{MS}_{\mathtt{M}} = \mathtt{SS}_{\mathtt{M}}/\mathtt{df}_{\mathtt{M}}$

Az F próba értéke: $\mathtt{F} = \mathtt{MS}_{\mathtt{R}}/\mathtt{MS}_{\mathtt{M}}$,

Az F érték szignifikanciája (p-értéke): $1 - pf(\mathtt{F}, df_{\mathtt{R}}, df_{\mathtt{M}})$,

Az együttható statisztikák

A sorelnevezések és fejléc:

`Koefficiensek`: a modell becsült paraméterei

`Standard hiba`: becsült paraméterek becsült szórása

`t érték`: t érték a becsült paraméterre 0 értéket feltételezve

`p-érték`: p-értéke

`alsó 95%`: a paraméter 95%-os konfidencia tartományának alsó határa

`felső 95%`: a paraméter 95%-os konfidencia tartományának felső határa

`alsó 97%`: a 97%-ra átállított konfidencia tartomány alsó határa

`felső 97%`: a 97%-ra átállított konfidencia tartomány felső határa

`Tengelymetszet`: a konstans, a β_0 értékének statisztikái

`X változó 1`: a meredekség, a β_1 értékének statisztikái

A táblázat adatai

a modell becsült paraméterei: $\widehat{\beta}_0$ és $\widehat{\beta}_1$

a becsült paraméterek becsült szórása:

$s_{\beta 0} = s\sqrt{\sum_{k=1}^{n} x^2/n \sum_{k=1}^{n}(x - \overline{x})^2}$, $s_{\beta 1} = s\sqrt{1/\sum_{k=1}^{n}(x - \overline{x})^2}$

a t-érték: $t_{\beta 0} = \widehat{\beta}_0/s_{\beta 0}$, $t_{\beta 1} = \widehat{\beta}_1/s_{\beta 1}$

a p-érték $2 * pt(-abs(t_{\beta 1}), n - p)$ és $2 * pt(-abs(t_{\beta 1}), n - p)$

a 95%-os konfidencia tartomány, ha $q_{5\%} = qt(.975, n - p)$

akkor a konstansra $(\widehat{\beta}_0 \mp q_{5\%} * se.b0,)$; a meredekségre $(\widehat{\beta}_1 \mp q_{5\%} * se.b1,)$

a 97%-os konfidencia tartomány esetén $q_{5\%}$ helyett $q_{3\%} = qt(.985, n - p)$

[B] Legkisebb négyzetek módszere - több magyarázó változó esete

Tartalmazza az $n \times p$ méretű X mátrix

az első oszlopában a végig 1-es vektort, és

az utolsó $p-1$ oszlopában a $p-1$ darab magyarázó változó mért értékeit!

Tartalmazza az $n \times 1$ méretű Y mátrix a magyarázott (cél) változó értékeit!

Ha a cél változó adatait az

$$y_k = \beta_0 + \beta_1 x_{k,2} + \beta_2 x_{k,3} + ...\beta_{p-1} x_{k,p} + \varepsilon_k, \qquad k = 1, ..., n$$

modell alapján a legkisebb négyzetek módszerével,

azaz a $\sum_{k=1}^{n} \varepsilon_k^2$ összeget minimalizálva akarjuk közelíteni,

akkor a $\beta_0, \beta_1, ..., \beta_{p-1}$ paraméterek b vektorba foglalt optimális értékeit a

$$b = (X^T X)^{-1} X^T Y$$

képlettel nyerjük. A ε korrigált tapasztalati szórása: $s_\varepsilon = \sqrt{||Y - Xb||/(n-p)}$

A b együtthatóbecslések variancia becsléseit a $s_b^2 = s_\varepsilon^2 \cdot diag((X^T X)^{-1})$ vektor elemei adják.

Ha feltételezzük, hogy a ε hibák

független 0 várhatóértékű, azonos szórású normális eloszlású mennyiségek, akkor a:

$$d_{\beta j} = \frac{b_j - \beta_j}{s_{bj}} \qquad j = 0, ..., p \text{ mennyiségek } t_{n-p} \text{ eloszlásúak.}$$

Ha a képletben a β_j helyett egy tetszőleges m értéket (akár nullát) írunk,

akkor a kapott d_m statisztika a $H_0 = m$ hipotézis

tetszőleges ellenhipotézissel szembeni, t-próbával való tesztelésére alkalmazható.

A regresszió modellnek a minimál modellel szembeni érvényessége

az F statisztikával vizsgálható.

Az F statisztika értéke, alkalmazva a $\widehat{Y} = Xb$ jelölést és figyelembe véve, hogy $\overline{Y} = \overline{\widehat{Y}}$:

$$d = \frac{||\widehat{Y} - \overline{\widehat{y}}||/(p-1)}{||Y - \widehat{Y}||/(n-p)}$$

Ha a regresszió modell érvényes, akkor a d eloszlása $F_{p-1,n-p}$.

A regresszió modell elutasítandó, ha az F értéke az eloszlást figyelembe véve nagy.