Pořadové testy v lineárním regresním modelu

a) model regresní přímky: $Y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \dots, n$

chyby modelu e_i jsou nezávislé, stejně rozdělené s distribuční funkcí F a hustotou f.
$x_1, \dots, x_n$ jsou pevná čísla

$H_0: \beta_1 = 0$ (je tam vůbec nějaká regrese?)

$H_1: \beta_1 \neq 0$ (případně $\beta_1 > 0$ nebo $\beta_1 < 0$)

Konstrukce pořadových testů

Pokud bychom znali F, resp. f, pak lokálně nejsilnější pořadový test H_0 proti H_1 je založen na testové statistice $S_n^f = \sum_{i=1}^{n} (x_i - \bar{x})\, a_n(R_i, f)$, kde R_i je pořadí Y_i mezi $Y_1, \dots, Y_n$.

Skóry $a_n(i, f)$ můžeme použít buďto přesné $E\varphi(U_{(i)}, f)$, nebo aproximativní $\varphi(\frac{i}{n+1}, f)$.

$$\varphi(u, f) = -\frac{f'(F^{-1}(u))}{f(F^{-1}(u))}, \quad 0 < u < 1.$$

V praxi ale f neznáme, proto zvolíme funkci $\varphi : [0,1] \to \mathbb{R}$ neklesající, nekonstantní a integrovatelná se čtvercem a definujeme $S_n = \sum_{i=1}^{n} (x_i - \bar{x})\, a_n(R_i) = \sum_{i=1}^{n} (x_i - \bar{x})\, \varphi\left(\frac{R_i}{n+1}\right)$,

kde $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$.

Přesné rozdělení S_n za platnosti H_0: Vektor pořadí $(R_1, \dots, R_n)^T$ má rovnoměrné rozdělení na množině všech permutací čísel $\{1, \dots, n\}$ (podmíněně při daných $x_1, \dots, x_n$)

permutační princip (věta o transformaci)
spočítáme hodnoty S_n pro všechny permutace čísel $\{1, 2, \dots, n\}$.

Pořadí jsou invariantní vůči posunutí $\Rightarrow$ rušivý parametr β_0 nám nevadí (za H_0 máme model: $Y_i = \beta_0 + e_i$)

Pro n středně výpočetně náročné $\to$ asymptotická aproximace

Poznámka

Za platnosti H_0 platí: $ES_n = 0$ a $DS_n = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\cdot\left(\sum_{j=1}^{n}(a(j)-\bar{a})^2\right)$.

člen $\left(\sum_{j=1}^{n}(a(j)-\bar{a})^2\right)$ můžeme opět pro n velké aproximovat pomocí $(n-1)\int_0^1(\varphi(u)-\bar{\varphi})^2\,du$.

Věta

Za platnosti H_0 a mírných předpokladů na $X_1,\dots,X_n$ platí $S_n \overset{H_0}{\approx} N(0, DS_n)$ pro $n\to\infty$.

b) model vícenásobné regrese: $Y_i = \beta_0 + x_i^T\beta + e_i$, $i=1,\dots,n$ $\qquad$ $\beta\in\mathbb{R}^p$ je p-rozměrný parametr

$H_0: \beta = 0$

$H_1: \beta \neq 0$

(je tam vůbec nějaká regrese?)

Pro $j=1,\dots,p$ definujme $S_{nj} = \sum_{i=1}^{n}(X_{ij}-\bar{X}_j)\,a_n(R_i)$, kde $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$ pro nějakou nekonstantní, neklesající a integrovatelnou se čtvercem

R_i je pořadí Y_i mezi $Y_1,\dots,Y_n$.

$$\mathbf{S}_n = (S_{n1},\dots,S_{np})^T = \sum_{i=1}^{n}(x_i-\bar{x})\cdot a_n(R_i) = \sum_{i=1}^{n}(x_i-\bar{x})\,\varphi\left(\frac{R_i}{n+1}\right).$$

Za platnosti H_0 platí: $E S_n = 0$, $DS_n = E S_n S_n^T = \underbrace{\frac{1}{n-1}\left(\sum_{i=1}^{n} (x_i-\bar{x})(x_i-\bar{x})^T\right)}_{Q_n} \cdot \underbrace{\left(\sum_{j=1}^{n} (a_n(j)-\bar{a})^2\right)}_{A_n^2}$.

a za mírných předpokladů na $x_1, \dots, x_n$ platí $S_n \overset{H_0}{\approx} N_p(0, DS_n)$.

__Poznámka__

A_n^2 můžeme pro n velké aproximovat pomocí $(n-1)\int_0^1 (\varphi(u)-\bar{\varphi})^2 \, du$.

Testová statistika: $T_n = S_n^T (DS_n)^{-1} S_n = \frac{1}{A_n^2} S_n^T Q_n^{-1} S_n$.

Opět umíme určit přesné rozdělení T_n (podmíněně při $x_1, \dots, x_n$) za platnosti H_0 pomocí permutačního principu.

Pro n středně použijeme aproximaci: za H_0 a mírných předpokladů na $x_1, \dots, x_n$ platí $T_n \overset{H_0}{\approx} \chi^2_p$.

__Poznámka__

Konkrétní příklady testů jsou analogické těm v předchozích kapitolách.

– Wilcoxonův, van der Waerdenův apod.

R-odhady v lineárním regresním modelu

a) model regresní přímky: $Y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, \dots, n$ e_i jsou i.i.d. chyby modelu s distr. fcí F a hustotou f

Zajímá nás odhad směrnice β_1.

Vyjdeme z testu hypotézy $H_0: \beta_1 = b$, kde b je známé číslo

Za platnosti H_0 máme model: $Y_i = \beta_0 + b x_i + e_i$

$$\underbrace{Y_i - b x_i}_{Y_i(b) = Y_i^*} = \beta_0 + e_i$$ v tomto modelu testujeme $H_0^*: \beta = 0$.

Testová statistika: $S_n(b) = \sum_{i=1}^{n} (x_i - \bar{x}) a_n(R_i(b))$, kde $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$ jsou skóry generované neklesající, nekonstantní, integrovatelnou se čtvercem a antisymetrickou kolem $\frac{1}{2}$ (tj. $\varphi(u) = -\varphi(1-u)$) funkcí $\varphi: [0,1] \to \mathbb{R}$.

$R_i(b)$ je pořadí $\underbrace{Y_i - bx_i}_{Y_i^*}$ mezi $\underbrace{Y_1 - bx_1}_{Y_1^*}, \dots, \underbrace{Y_n - bx_n}_{Y_n^*}$.

$S_n(b)$ jakožto funkce $b \in \mathbb{R}$ je nerostoucí, po částech konstantní funkce

Je-li skutečná hodnota parametru rovna β_1 (za platnosti H_0), pak $ES_n(b) = 0$. $\Rightarrow$ hledáme b takové, pro které $S_n(b) \doteq 0$. obecně nemusí mít řešení

R-odhad parametru β_1 definujeme jako $\hat{\beta}_1 = \frac{1}{2}(\hat{\beta}_1^+ + \hat{\beta}_1^-)$, kde $\hat{\beta}_1^+ = \inf\{b : S_n(b) < 0\}$ a $\hat{\beta}_1^- = \sup\{b : S_n(b) > 0\}$.

Poznámka

$\hat{\beta}_1$ je asymptoticky nestranný, konzistentní a asymptoticky normální odhad β_1 (as. rozptyl závisí na neznámé hustotě f).

b) model vícenásobné regrese: $Y_i = \beta_0 + x_i^T \beta + e_i$, $i = 1, \dots, n$

zajímá nás odhad p-rozměrného parametru β.

Vyjdeme z testu hypotézy $H_0 : \beta = b$, kde b je známý p-rozměrný vektor

za platnosti H_0 máme model: $Y_i = \beta_0 + x_i^T b + e_i$

$$\underbrace{Y_i - x_i^T b}_{Y_i(b) = Y_i^*} = \beta_0 + e_i$$ v tomto modelu testujeme $H_0^* : \beta = 0$.

testová statistika: $S_n(b) = \sum_{i=1}^{n} (x_i - \bar{x})\, a_n(R_i(b))$, kde $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$ jsou skóry generované neklesající nekonstantní, int. se čtvercem a antisymetrickou funkcí $\varphi : [0,1] \to \mathbb{R}$.

$R_i(b)$ je pořadí $Y_i - x_i^T b$ mezi $Y_1 - x_1^T b, \dots, Y_n - x_n^T b$.

Je-li skutečná hodnota parametru rovna β, pak $E S_n(b) = 0$. ($S_n : \mathbb{R}^p \mapsto \mathbb{R}^p$ je p-rozměrná funkce)

$T_n(b) = \| S_n(b) \|$, kde $\| . \|$ je libovolná norma v $\mathbb{R}^p$ (volba normy neovlivní vlastnosti výsledného odhadu).

R-odhad parametru β definujeme jako $\hat{\beta} = \underset{b \in \mathbb{R}^p}{\arg\min}\, T_n(b) = \underset{b \in \mathbb{R}^p}{\arg\min}\, \| S_n(b) \|$.

Poznámka

$\hat{\beta}$ je as. nestranný, konzistentní a as. normální odhad β (as. rozptyl závisí na neznámé hustotě f).

jiná definice R-odhadů: Jaeckel (1972)

rank measure of dispersion: $D_n(\mathbb{b}) = \sum_{i=1}^{n} \left(Y_i - (\mathbb{x}_i - \bar{\mathbb{x}})^T \mathbb{b}\right) a_n(R_i(\mathbb{b}))$, kde $a_n(i)$ a $R_i(\mathbb{b})$ jsou definované výše.

$D_n(\mathbb{b})$ je konvexní, po částech lineární, se subgradientem $-\sum_{i=1}^{n} (\mathbb{x}_i - \bar{\mathbb{x}}) a_n(R_i(\mathbb{b}))$.

$$\hat{\beta}^J = \underset{\mathbb{b} \in \mathbb{R}^p}{\operatorname{argmin}} \; D_n(\mathbb{b}).$$

Tvrzení

Odhady $\hat{\beta}$ a $\hat{\beta}^J$ jsou asymptoticky ekvivalentní.

Testy hypotéz o části parametru β

$Y_i = \beta_0 + \mathbb{x}_i^T \beta + \mathbb{z}_i^T \gamma + e_i$, $i = 1, \dots, n$

$(\underset{\in \mathbb{R}^p}{\mathbb{x}_i^T}, \underset{\in \mathbb{R}^q}{\mathbb{z}_i^T})^T$ ... vektor regresorů dimenze $p+q$

$(\beta_0, \underset{\in \mathbb{R}^p}{\beta^T}, \underset{\in \mathbb{R}^q}{\gamma^T})$ ... vektor neznámých parametrů dimenze $1+p+q$

$H_0: \gamma = \mathbb{0}$

β_0, β jsou rušivé parametry

$H_1: \gamma \neq \mathbb{0}$

a) aligned pořadový test o parametru γ

- pořadové testy jsou invariantní vůči posunutí => parametr β_0 nám „nevadí".

- odhadneme parametr $\boldsymbol{\beta}$ pomocí R-odhadu, uděláme rezidua a na ně aplikujeme pořadový test o parametru γ.

(i) za platnosti H_0 máme model: $Y_i = \beta_0 + x_i^T \boldsymbol{\beta} + e_i$ ← v tomto modelu odhadneme p-rozměrný parametr $\boldsymbol{\beta}$

$$S_n(\mathbf{b}) = \sum_{i=1}^{n} (x_i - \bar{x}) \varphi\left(\frac{R_i(\mathbf{b})}{n+1}\right)$$, kde φ je neklesající, nekonst. a int. ve čtverci a $\varphi(t) = -\varphi(1-t)$

$R_i(\mathbf{b})$ je pořadí $Y_i - x_i^T \mathbf{b}$ mezi $Y_1 - x_1^T \mathbf{b}, \ldots, Y_n - x_n^T \mathbf{b}$.

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \| S_n(\mathbf{b}) \|$$, kde $\| . \|$ je nějaká norma v $\mathbb{R}^p$.

(ii) definujeme rezidua $\hat{e}_i = Y_i - x_i^T \hat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$ ty ale nejsou ani nezávislé, ani stejně rozdělené (přesto „na ně" aplikujeme pořadový test o par. γ)

testová statistika $$S_n = \sum_{i=1}^{n} (z_i - \bar{z}) a_n(\hat{R}_i) = \sum_{i=1}^{n} (z_i - \bar{z}) \varphi\left(\frac{\hat{R}_i}{n+1}\right)$$, kde φ je nekles., nekonst. a int. ve čtverci

$\hat{R}_i$ je pořadí $\hat{e}_i$ mezi $\hat{e}_1, \ldots, \hat{e}_n$.

$$\mathbb{D}_n = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T, \quad A_n^2 = \sum_{j=1}^{n} (a_n(j) - \bar{a})^2$$

$$T_n = \frac{1}{A_n^2} S_n^T \mathbb{D}_n^{-1} S_n .$$

Nyní již nelze jednoduše vyjádřit přesné rozdělení T_n za platnosti H_0 ($\hat{e}_i$ jsou závislé) => nutnost asymptotické aproximace

za mírných předpokladů $T_n \overset{H_0}{\approx} \chi^2_q$ při $n \to \infty$.

2) test založený na regresních pořadových skórech

za platnosti H_0 máme model: $Y_i = \beta_0 + x_i^T \beta + e_i$ ← v tomto modelu najdeme regresní pořadové skóry

označme je $\hat{a}_i(\alpha)$, $i = 1, \dots, n$ pro $0 < \alpha < 1$ ty jsou invariantní vůči rušivé regresi v x_i (analogie pořadí) ⇐ rušivé parametry β_0 a β nevadí.

označme: $\mathbb{X} = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}$ a $\mathbb{Z} = \begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix}$.

Zvolíme neklesající, nekonstantní a int. na čtverec funkci $\varphi : [0,1] \to \mathbb{R}$ a definujeme skóry $\hat{b}_i = -\int_0^1 \varphi(t)\, d\hat{a}_i(t)$ pro $i = 1, \dots, n$ → $\hat{b} = (\hat{b}_1, \dots, \hat{b}_n)^T$.

testová statistika: $S_n = \sum_{i=1}^n (z_i - \hat{z}_i) \cdot \hat{b}_i$, kde $\hat{\mathbb{Z}} = \begin{pmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_n \end{pmatrix} = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Z}$ je projekce $\mathbb{Z}$ do prostoru sloupců matice $\mathbb{X}$.

$\hat{\mathbb{D}}_n = \frac{1}{n-1} \sum_{i=1}^n (z_i - \hat{z}_i)(z_i - \hat{z}_i)^T$, $A_n^2 = (n-1) \int_0^1 (\varphi(t) - \bar{\varphi})^2\, dt$.

$\tilde{T}_n = \frac{1}{A_n^2} \cdot S_n^T \hat{\mathbb{D}}_n^{-1} S_n$ má za mírných předpokladů asymptoticky χ^2_q.

Poznámka: Oba testy mají stejné asymptotické rozdělení (i za H_1), tedy i asymptoticky stejnou sílu.