Testování hypotéz o parametrech regresního modelu

Statistika II

Jiří Neubauer

Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Lineární regresní model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

kde

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Odhady neznámých parametrů metodou nejmenších čtverců jsou dány

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y},$$

reziduální součet čtverců je

$$S_e = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}.$$

Lineární regresní model

Předpokládejme nyní, že náhodné chyby $e_i, i = 1 \ldots, n$ v lineárním regresním modelu mají normální rozdělení s nulovou střední hodnotou a rozptylem σ^2. Potom mají odhady $\widehat{\beta}_j, j = 1, \ldots, k$ regresní koeficientů β_j normální rozdělení, tedy platí $\widehat{\beta}_j \sim N(\beta_j, D(\widehat{\beta}_j))$, kde rozptyly $D(\widehat{\beta}_j)$ jsou dány:

$$D(\widehat{\beta}_1) = \sigma^2 v_{11}, D(\widehat{\beta}_2) = \sigma^2 v_{22}, \ldots, D(\widehat{\beta}_k) = \sigma^2 v_{kk},$$

přičemž $v_{11}, v_{22}, \ldots, v_{kk}$ jsou prvky na hlavní diagonále matice $(\mathbf{X}'\mathbf{X})^{-1}$. Rozptyly odhadů regresních parametrů odhadneme $\widehat{D}(\widehat{\beta}_j) = s_e^2 v_{jj}$, druhé odmocniny těchto odhadů

$$s(\widehat{\beta}_j) = \sqrt{s_e^2 v_{jj}}$$

se nazývají **směrodatné chyby** odhadů regresních parametrů.

Testy významnosti parametrů β_j, $j = 1, \ldots, k$ (jejich nenulovosti) jsou založeny na statistikách

$$t = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)},$$

které mají Studentovo rozdělení s $n - k$ stupni volnosti.
Budeme testovat nulovou hypotézu
H: $\beta_j = 0$ proti alternativní hypotéze A: $\beta_j \neq 0$. Při platnosti nulové hypotézy má statistika

$$t = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} = \frac{\widehat{\beta}_j - 0}{s(\widehat{\beta}_j)} = \frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}$$

Studentovo rozdělení s $n - k$ stupni volnosti. Kritickou hodnotou odpovídající hladině významnosti α je tedy kvantil $t_{1-\frac{\alpha}{2}}(n - k)$.

Test významnosti regresního modelu

Zřejmě platí, že $y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})$. Lze ukázat, že také platí

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \rightarrow S_Y = S_e + S_T,$$

kde

- **celkový součet čtverců** $S_Y = \mathbf{Y}'\mathbf{Y} - n\overline{y}^2$

$$S_Y = \sum_{i=1}^{n}(y_i - \overline{y})^2 = n \cdot s_n^2(y), \text{ kde } s_n^2(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

- **reziduální součet čtverců** $S_e = \mathbf{Y}'\mathbf{Y} - \widehat{\beta}'\mathbf{X}'\mathbf{Y}$

$$S_e = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (n-k) \cdot s_e^2, \text{ kde } s_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

- **teoretický součet čtverců** $S_T = \widehat{\beta}'\mathbf{X}'\mathbf{Y} - n\overline{y}^2$

$$S_T = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = n \cdot s_n^2(\hat{y}), \text{ kde } s_n^2(\widehat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

Test významnosti regresního modelu

Pro regresní přímku $\widehat{y} = \widehat{\beta}_1 + \widehat{\beta}_2 x$ dostáváme

$$S_e = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{\beta}_1 - \widehat{\beta}_2 x_i)^2 = \cdots =$$

$$= \sum_{i=1}^{n} y_i^2 - \widehat{\beta}_1 \sum_{i=1}^{n} y_i - \widehat{\beta}_2 \sum_{i=1}^{n} x_i y_i$$

$$S_T = \sum_{i=1}^{n} (\widehat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(\widehat{\beta}_1 + \widehat{\beta}_2 x_i - \frac{1}{n} \sum_{i=1}^{n} y_i^2 \right) = \cdots =$$

$$= \widehat{\beta}_1 \sum_{i=1}^{n} y_i + \widehat{\beta}_2 \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2$$

$$S_Y = S_R + S_T = \cdots = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2$$

Test významnosti regresního modelu

- teoretický součet čtverců S_T je ta část celkového součtu čtverců S_Y, která je vysvětlená zvolenou regresní funkcí
- reziduální součet čtverců S_e je ta část celkového součtu čtverců S_Y, která zvolenou regresní funkcí vysvětlená není

Test významnosti regresního modelu

Při ověřování významnosti regresního modelu (model s konstantou β_1) se testuje nulová hypotéza H: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$ proti alternativní hypotéze A: $\beta_j \neq 0$ pro alespoň jedno $j = 2, 3, \ldots, k$. Testové kritérium je statistika

$$F = \frac{S_T(y)}{k-1} : \frac{S_e(y)}{n-k},$$

která má při platnosti nulové hypotézy Fisher-Snedecorovo rozdělení F s $k-1$ a $n-k$ stupni volnosti, Kritickou hodnotou je kvantil $F_{1-\alpha}(k-1, n-k)$ daného F rozdělení.

Test významnosti regresního modelu

- Jsou-li celkový F-test i všechny t-testy jsou statisticky významné, model se považuje za vhodný k vystižení variability proměnné Y (to však ještě neznamená, že je model správně navržen).
- Jsou-li celkový F-test i všechny t-testy jsou statisticky nevýznamné, model se považuje za nevhodný, protože nevystihuje variabilitu proměnné Y.
- Je-li celkový F-test statisticky významný, ale některé t-testy vychází nevýznamné, model se považuje za vhodný, ale provádí se zpravidla vypuštění nevýznamných parametrů.
- Je-li celkový F-test statisticky významný, ale všechny t-testy vychází nevýznamné – paradox: formálně model jako celek vyhovuje, ale žádný člen modelu sám o sobě významný není – jde o důsledek tzv. **multikolinearity**, tj. lineární závislosti mezi jednotlivými regresory.

Koeficient (index) determinace

Vhodnost zvoleného modelu lze vyjádřit pomocí tzv. indexu (koeficientu) determinace, který je definován jako podíl variability, kterou je schopen popsat regresní model, ku celkové variabilitě vysvětlované proměnné $S_c(y) = \sum_{i=1}^{n}(y_i - \overline{y_i})^2$

$$R^2 = \frac{S_T(y)}{S_c(y)} = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}.$$

Toto číslo nabývá hodnot z intervalu $\langle 0, 1 \rangle$. Čím více se R^2 blíží k 1, tím považujeme danou závislost za silnější, a tedy dobře vystiženou použitým regresním modelem; naopak čím více se bude blížit k 0, tím považujeme danou závislost za slabší a regresní funkci za méně výstižnou. Nízká hodnota R^2 ještě nemusí znamenat nízký stupeň závislosti mezi proměnnými, ale může signalizovat chybnou volbu regresního modelu.

Koeficient (index) determinace

R^2 představuje výběrový index determinace, který lze použít jako odhad teoretického indexu determinace. Tento odhad je asymptoticky nestranný, nicméně pro malé výběry nadhodnocuje skutečnou těsnost závislosti a je závislý na počtu parametrů regresního modelu. Lze provést jeho korekci

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k},$$

čím získáme odhad nestranný.

Test obecné lineární hypotézy

Z předchozích výsledků lze odvodit, za předpokladu že LRM je plné hodnosti, že pro libovolnou reálnou matici $\mathbf{A}$ typu $m \times k$ a hodnosti $m \leq k$ má statistika

$$F = \frac{1}{ms_e^2}(\widehat{\beta} - \beta)'\mathbf{A}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}\mathbf{A}(\widehat{\beta} - \beta)$$

Fisher-Snedecorovo $F-$rozdělení o m a $n-k$ stupních volnosti. Tuto statistiku lze pak využít při testování obecné lineární hypotézy H_0, kterou zapíšeme ve tvaru

$$\mathbf{A}\beta = \mathbf{a}, \tag{1}$$

kde $\mathbf{a}$ je vhodný $m-$rozměrný reálný vektor, pro něž je rovnice (1) řešitelná. Odtud pak plyne, že testovací statistika

$$F = \frac{1}{ms_e^2}(\mathbf{A}\widehat{\beta} - \mathbf{a})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\widehat{\beta} - \mathbf{a}) \tag{2}$$

má za platnosti nulové hypotézy H_0 Fisher-Snedecorovo F rozdělení o m a $n-k$ stupních volnosti. Proto nulovou hypotézu H_0 zamítáme na hladině významnosti α, když $F \geq F_{1-\alpha}(m, n-k)$, kde $F_{1-\alpha}(m, n-k)$ kvantil $F-$rozdělení o m a $n-k$ stupních volnosti.

Test obecné lineární hypotézy

Speciální volbou matice **A** a vektoru **a** v (1) lze potom získat speciální hypotézy, které experimentátora zajímají. Po dosazení za **A** a **a** do statistiky F ve (2) lze získat odpovídající testovací statistiky pro testování hypotéz o parametrech $\beta_1, \ldots, \beta_k$.
Tímto způsobem lze konstruovat řadu běžných testů o neznámých parametrech včetně testů o parametrech LRM, které byly uvedeny dříve.

Test obecné lineární hypotézy

Příklad 1. Test hypotézy $\beta_j = 0$ lze získat při volbě $\mathbf{A} = \mathbf{u}_j$, kde $\mathbf{u}_j$ je $k-$rozměrný jednotkový vektor s jedničkou na $j-$tém místě a $\mathbf{a} = 0$. Jde o test, založený na testovací statistice t.

Příklad 2. Testy hypotéz o rovnosti parametrů, např. test hypotézy $\beta_1 = \beta_2 = \cdots = \beta_m$, $m \leq k$, lze získat volbou $\mathbf{a} = \mathbf{0}_{m-1}$ a

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 & \cdots & 0 \end{pmatrix}$$

je matice typu $(m-1) \times k$.

Příklad 3. Předpokládejme, že jsou dány dva nezávislé regresní modely, každý s jedním regresorem. První o rovnici $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ a druhý o rovnici $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{e}^*$, kde $X_{n\times 2}$ a $X^*_{n^*\times 2}$ jsou příslušné matice plánů s vektorem jedniček v prvním sloupci, $\mathbf{Y}$ a $\mathbf{Y}^*$ jsou vektory pozorování závisle proměnných v obou modelech, $\mathbf{e}$ a $\mathbf{e}^*$ jsou nezávislé normálně rozdělené vektory náhodných chyb a konečně $\boldsymbol{\beta}$ a $\boldsymbol{\beta}^*$ jsou dvourozměrné vektory neznámých parametrů. Tedy jsou dány dva nezávislé regresní modely a regresní funkce v obou modelech je dána přímkou. Když oba vektory náhodných chyb mají varianční matice $\sigma^2\mathbf{I}_n$ a $\sigma^2\mathbf{I}_{n^*}$, lze vytvořit spojení obou modelů a uvážit nový model tvaru

$$\begin{pmatrix}\mathbf{Y}\\ \mathbf{Y}^*\end{pmatrix} = \begin{pmatrix}\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{X}^*\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{\beta}^*\end{pmatrix} + \begin{pmatrix}\mathbf{e}\\ \mathbf{e}^*\end{pmatrix}$$

V tomto modelu lze např. testovací statistiku pro testování rovnoběžnosti obou regresních přímek snadno dostat ze vzorce (2) volbou $\mathbf{A} = (0, 1, 0, -1)$, $\mathbf{a} = 0$.