M5VM05 Statistické modelování

6. Ověřování předpokladů v klasickém modelu lineární regrese – II

Jan Koláček (kolacek@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

Multikolinearita

Multikolinearitou se rozumí vzájemná lineární závislost vysvětlujících proměnných. Přesnou multikolinearitou se rozumí případ, kdy jednotlivé sloupce $\mathbf{x}_j,\ j = 1, \dots, k$ matice plánu $\mathbf{X}$ jsou lineárně závislé, takže pro aspoň jednu nenulovou konstantu c_j platí

$$c_1\mathbf{x}_1 + \cdots + c_k\mathbf{x}_k - \mathbf{0}.$$

V praxi bychom se s tímto případem neměli setkávat, neboť při rozumně sestaveném regresním modelu využijeme lineární kombinaci a zmenšíme počet vysvětlujících proměnných. Podobně nereálný je v praxi případ ortogonálních vysvětlujících proměnných, kdy matice $\mathbf{X}$ je ortogonální a platí, že $\mathbf{X}'\mathbf{X} = \mathbf{I}_k$.
V praxi se tedy **multikolinearitou** rozumí případ, kdy **přibližně** platí rovnice vyjadřující lineární kombinaci vysvětlujících proměnných. V případě silné multikolinearity je determinant informační matice $\mathbf{X}'\mathbf{X}$ blízký nule, nejmenší vlastní číslo je rovněž blízké nule a matice $\mathbf{X}'\mathbf{X}$ je „skoro singulární“. O multikolinearitě svědčí i vysoké hodnoty poměru největšího a nejmenšího vlastního čísla.

Důvody multikolinearity

- Multikolinearitu způsobuje regresní rovnice obsahující **nadbytečné** vysvětlující proměnné. Statistickými technikami můžeme přebytečné proměnné identifikovat a vyloučit z regresní rovnice.
- Multikolinearitu jen ztěží odstraníme v úlohách, kdy vzájemná spřaženost hodnot vysvětlujících proměnných je způsobena neuvažovanými veličinami nebo **formou statistického zjišťování**. Jde-li např. o údaje z časových řad, je podobný vývoj sledovaných veličin dostatečným důvodem vzniku multikolinearity. Vzhledem k tomu, že multikolinearitu hodnotíme výhradně na základě určitého souboru pozorování, stačí nesprávný výběr kombinací hodnot vysvětlujících proměnných, nereprezentujících obor možných hodnot, k existenci významné multikolinearity.
- Závažným důvodem multikolinearity je **skutečný vztah** vysvětlujících proměnných v rámci sledovaného jevu, procesu nebo systému. V tomto případě je třeba využít všechny informace nevýběrového charakteru k zlepšení kvality regresních odhadů.

Důsledky multikolinearity

V případě přesné multikolinearity je matice $\mathbf{X}'\mathbf{X}$ singulární a běžnou inverzí nepořídíme odhad neznámých parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců. Pro přibližnou multikolinearitu jsme sice schopni matici $\mathbf{X}'\mathbf{X}$ invertovat, ale kvalita pořízených odhadů je poměrně nízká.
Snížení kvality se projeví

- v kovarianční matici $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- v přesnosti prováděných výpočtů

neboť důsledkem vysokých rozptylů odhadů jsou příliš široké intervaly spolehlivosti, a tedy malá přesnost odhadu.
Logickým důsledkem multikolinearity je obtížné vyjádření individuálního vlivu jednotlivých vysvětlujících proměnných. Projeví se to nízkými hodnotami testových kritérií v t-testech nedovolujícími potvrdit závažnost jednotlivých regresorů v regresní funkci. Závažným důsledkem je značná výpočetní nespolehlivost a nestabilní hodnoty regresních odhadů. Stačí malý zásah do statistických údajů a výsledné odhady jsou odlišné.

VIF – Variance Inflarion Factors

Diagonální prvky matice $(\mathbf{X}'\mathbf{X})^{-1}$, tj.

$$\mathbf{a} = diag(\mathbf{X}'\mathbf{X})^{-1}$$

označované v literatuře jako **VIF – variance inflarion factors** úzce souvisí s **vícenásobnými korelačními koeficienty**, vyjadřující vztah j-té vysvětlující proměnné a lineární funkce ostatních vysvětlujících proměnných. Lze je zapsat jako

$$a_j = \frac{1}{(1 - r_j^2)\mathbf{x}_j'\mathbf{x}_j},$$

kde $r_j = r_{x_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_k}$ je koeficient mnohonásobné korelace.
Vysoký stupeň multikolinearity se projevuje vysokými hodnotami korelačních koeficientů r_j, ale i vysokými hodnotami některých jednoduchých korelačních koeficientů.

Detekce multikolinearity

Testujeme hypotézu $H_0 : \mathbf{R} = \mathbf{I}_k$ proti $H_1 : \mathbf{R} \neq \mathbf{I}_k$, kde $\mathbf{R}$ je korelační matice proměnných. Platí-li nulová hypotéza, pak

$$K = -\left[n - 1 - \frac{1}{6}(2k + 7)\right] \ln |\mathbf{R}| \sim \chi^2\left(\frac{k(k-1)}{2}\right).$$

Hypotézu H_0 tedy na hladině významnosti α zamítáme, pokud $K > \chi^2_{1-\alpha}\left(\frac{k(k-1)}{2}\right)$.

Identifikace proměnných způsobujících multikolinearitu

Pro identifikaci proměnných způsobujících multikolinearitu se doporučují statistiky

$$F_j = \frac{n-k}{k-1}(d_{jj} - 1),$$

kde d_{jj} jsou diagonální prvky matice $\mathbf{D} = \mathbf{R}^{-1}$. V případě, že proměnná X_j nezpůsobuje multikolinearitu, má veličina F_j Fisherovo – Snedecorovo rozdělení $F(k-1, n-k)$.

Odstranění multikolinearity

Do modelu zařadíme jen ty regresory, které významně přispívají ke zlepšení kvality odhadu $\boldsymbol{\beta}$. K výběru nejlepší podmnožiny regresorů použijeme **metodu postupné regrese**.

Algoritmus:

1. Spočteme korelační matici $\mathbf{R}$ a provedeme test hypotézy $H_0 : \mathbf{R} = \mathbf{I}_k$. Je-li korelace prokázána, pokračujeme dalším krokem.
2. Spočteme korelační koeficienty $r_{Y,X_1}, \ldots, r_{Y,X_k}$ a vybereme ten regresor X_i, jehož r_{Y,X_i} je v absolutní hodnotě největší.
3. Sestavíme model $Y = \beta_0 + \beta_1 X_i$ a odhadneme jeho parametry. Vypočteme hodnotu statistiky $F = \frac{(n-2)s^2_{\hat{Y}}}{s^2} = \frac{(n-2)ID}{1-ID}$, kde ID značí index determinace. Pokud $F > F_{1-\alpha}(1, n-2)$, ponecháme regresor X_i v modelu.
4. Spočteme parciální korelační koeficienty $r_{Y,X_1 \cdot X_i}, \ldots, r_{Y,X_{i-1} \cdot X_i}, r_{Y,X_{i+1} \cdot X_i}, \ldots, r_{Y,X_k \cdot X_i}$ a vybereme ten regresor X_j, jehož $r_{Y,X_j \cdot X_i}$ je v absolutní hodnotě největší.

Odstranění multikolinearity

5. Sestavíme model $Y = \beta_0 + \beta_1 X_i + \beta_2 X_j$ a odhadneme jeho parametry. Vypočteme hodnotu statistiky $F = \frac{(n-3)\Delta ID}{1-ID}$, kde ΔID je přírůstek indexu determinace při zařazení X_j do modelu a ID je index determinace pro model $Y = \beta_0 + \beta_1 X_i + \beta_2 X_j$. Pokud $F > F_{1-\alpha}(2, n-3)$, ponecháme regresor X_j v modelu.
6. Spočteme parciální korelační koeficienty $r_{Y,X_1\cdot(X_i,X_j)}, \dots, r_{Y,X_k\cdot(X_i,X_j)}$ a vybereme ten regresor X_l, jehož $r_{Y,X_l\cdot(X_i,X_j)}$ je v absolutní hodnotě největší a tak pokračujeme dále.

Další možnosti

Zlepšování podmíněnosti matice $\mathbf{X}'\mathbf{X}$

- **Model standardizovaných proměnných**. Místo původních proměnných y_i a x_{ij} pracujeme s proměnnými ve tvaru

$$q_i = \frac{y_i - \bar{y}}{s_y}, \quad z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}},$$

kde s_y a s_{x_j} jsou směrodatné odchylky jednotlivých proměnných. Standardizací vysvětlujících proměnných dostáváme při použití metody nejmenších čtverců místo matice $\mathbf{X}'\mathbf{X}$ korelační matici $\mathbf{R} = \mathbf{Z}'\mathbf{Z}/n$. Vektor $\mathbf{Z}'\mathbf{q}/n$ obsahuje jednoduché korelační koeficienty r_{yx_j}. Standardizací proměnných se zmenšují zaokrouhlovací chyby a zlepšují se možnosti hodnocení individuálního vlivu proměnných pomocí regresních parametrů.

Další možnosti

- **Model v kanonickém tvaru**. Místo modelu ve tvaru

$$\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pracujeme s modelem

$$\mathbf{Y} = \mathbf{U}'\gamma + \boldsymbol{\varepsilon},$$

kde matice $\mathbf{U} = \mathbf{X}\mathbf{V}$, vektor $\gamma = \mathbf{V}'\boldsymbol{\beta}$ a $\mathbf{V}$ je matice standardizovaných vlastních vektorů odpovídajících vlastním číslům matice $\mathbf{X}'\mathbf{X}$. Odhady parametrů v kanonickém tvaru:

$$\hat{\gamma} = \mathbf{L}^{-1}\mathbf{U}'\mathbf{Y},$$

kde $\mathbf{L}$ je diagonální matice s vlastními čísly matice $\mathbf{X}'\mathbf{X}$. Kovarianční matice odhadů $var(\hat{\gamma}) = \sigma^2\mathbf{L}^{-1}$ ukazuje, že i v tomto případě jsou odhady nezávislé. Residuální součet čtverců se transformací nemění.

- **Hřebenová regrese** (ridge regression) – `lm.ridge` v balíku `MASS`

Příklad

V souboru „`vydaje.Rdata`“ jsou uložena data o 20 náhodně vybraných domácnostech. Sloupce proměnné „`domacnosti`“ obsahují postupně tyto údaje: výdaje za potraviny a nápoje (Y), počet členů domácnosti (X_1), počet dětí (X_2), průměrný věk výdělečně činných (X_3) a příjem domácnosti (X_4). Metodou postupné regrese zkonstruujte model s nejlepší podmíněností regresorů.

Řešení Uvažujme nejdřív model se všemi regresory. Spočtěme nejprve pro ilustraci $det((\mathbf{X}'\mathbf{X})^{-1}) = 4,65 \times 10^{-16}$. Také hodnoty VIF jsou pro první dva regresory vysoké:

clenu	deti	vek	prijem
$21,23$	$16,18$	$1,31$	$3,4$

Testujeme-li hypotézu $H_0 : \mathbf{R} = \mathbf{I}_4$, hodnota testové statistiky

$$K = -\left[19 - \frac{15}{6}\right] \ln |\mathbf{R}| = 64,94$$

výrazně převyšuje kritickou hodnotu $\chi^2_{0,95}(6) = 12,59$. Hypotézu H_0 tedy na hladině významnosti $0,05$ zamítáme.

Řešení

Pro identifikaci proměnných způsobujících multikolinearitu můžeme spočítat dílčí statistiky F_j

clenu	deti	vek	prijem
$107,88$	$80,94$	$1,66$	$12,83$

které porovnáme s kritickou hodnotou $F_{0,95}(3,16) = 3,24$.

Postupná regrese

1. Spočteme korelační koeficienty $r_{Y,X_1},\ldots,r_{Y,X_4} = (0,77; 0,67; 0,18; 0,73)$. Vybereme regresor X_1, neboť jeho korelace je v absolutní hodnotě největší.
2. Sestavíme model $Y = \beta_0 + \beta_1 X_1$. Vypočteme hodnotu statistiky $F = \frac{(n-2)ID}{1-ID} = \frac{18\cdot 0,5897}{1-0,5897} = 25,87$. Tato hodnota je větší než $F_{0,95}(1,18) = 4,41$, takže regresor X_1 ponecháme v modelu.
3. Spočteme parciální korelační koeficienty $r_{Y,X_2\cdot X_1}, r_{Y,X_3\cdot X_1}, r_{Y,X_4\cdot X_1} = (0,36; 0,499; 0,32)$. Vybereme regresor X_3, jehož parciální korelační koeficient je v absolutní hodnotě největší.

Řešení

4. Sestavíme model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3$. Vypočteme hodnotu statistiky $F = \frac{(n-3)\Delta ID}{1-ID} = \frac{17 \cdot 0{,}102}{1-0{,}69} = 5,64$. Tato hodnota je větší než $F_{0{,}95}(2,17) = 3,59$, tedy ponecháme regresor X_3 v modelu.
5. Spočteme parciální korelační koeficienty $r_{Y,X_2 \cdot (X_1,X_3)}, r_{Y,X_4 \cdot (X_1,X_3)} = (0,19; 0,17)$. Vybereme regresor X_2, jehož parciální korelační koeficient je v absolutní hodnotě největší.
6. Sestavíme model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_2$. Vypočteme hodnotu statistiky $F = \frac{(n-4)\Delta ID}{1-ID} = \frac{16 \cdot 0{,}012}{1-0{,}704} = 0,63$. Tato hodnota je menší než $F_{0{,}95}(3,16) = 3,24$, a tedy regresor X_2 již nezahrneme do modelu.

Výsledný model je tedy tvaru

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3.$$

Úlohy k procvičení

Příklad 1.1

V souboru „`cement.RData`" jsou uloženy údaje, které se týkají chemického složení portlandského cementu:

y	*množství tepla v kaloriích na gram cementu*
x_1	*Tricalcium aluminate 3CaO.Al2O3 v %*
x_2	*Tricalciam silicate 3CaO.SiO2 v %*
x_3	*Tetracalcium alumino ferrite 4CaO.Al2O3.Fe2O3 v %*
x_4	*Dicalcium silicate 2CaO.SiO2 v %*

Testujte multikolinearitu v daném modelu. Metodou postupné regrese nalezněte vhodný model. Poté ověřte normalitu residuí.

[Multikolinearita se nezamítá, vhodný model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4$, normalita residuí se nezamítá.]

Úlohy k procvičení

Příklad 1.2

V proměnné „`mtcars`“[a] jsou uložena data pro modelování závislosti spotřeby paliva osobních automobilů (proměnná `mpg`, počet mil/galon) na vlastnostech motoru, které jsou popsány následujícími proměnnými:

`cyl`	*počet válců*
`disp`	*objem válců (kubické palce)*
`hp`	*výkon (počet koní)*
`drat`	*převodový poměr zadní nápravy*
`wt`	*hmotnost vozidla (kilolibry)*
`qsec`	*zrychlení (počet sekund z 0 na 1/4 míle)*
`vs`	*uspořádání válců (1 – „V“, 0 – za sebou)*
`am`	*převodovka (0 – automat, 1 – manuál)*
`gear`	*počet převodových stupňů*
`carb`	*počet karburátorů*

Testujte multikolinearitu v daném modelu. Metodou postupné regrese nalezněte vhodný model. Ověřte také normalitu residuí.

[a] datový soubor implementovaný v jazyce R

Úlohy k procvičení

[Multikolinearita se nezamítá, vhodný model: $\texttt{mpg} = \beta_0 + \beta_1 \texttt{wt} + \beta_2 \texttt{cyl}$, normalita residuí se nezamítá.]

Příklad 1.3 (pro náročné)

Naprogramujte funkci „`multicol.R`“, která pro zadaný model zjistí přítomnost multikolinearity v datech. V případě, že je multikolinearita přítomna, metodou postupné regrese nalezne vhodný model.