

Predikce v normálním lineárním regresním modelu

Část 1 – testy stability

Opět přijímáme tyto vlastnosti o veličinách lineárního regresního (jednorovnicového) modelu :

1. **Centrovanost náhodných složek** $E(\varepsilon) = 0$.
2. **Diagonalita kovarianční matice náhod.složek:** $Cov(\varepsilon, \varepsilon) = \sigma^2 \cdot I_T$, tj.
(diagonální nestochastická matice),
 - 2a) **homoskedasticita náhodných složek**
 - 2b) **neautokorelovanost náhodných složek**
3. **Nekorelovanost náhodných složek s nezávisle proměnnými** $E(X' \varepsilon) = 0$.
4. **Plná hodnota matice vysvětlujících proměnných** $h(X) = k$
5. **Normalita T-rozměrného vektoru náhodných složek** s nulovým vektorem středních hodnot a s diagonální kovarianční maticí

$$\Sigma = \sigma^2 I_T, \quad \text{neboli} \quad \varepsilon \approx N(0; \sigma^2 I_T)$$

Výše uvedené předpoklady jsou nezbytné k tomu, abychom uchovali platnost všech dříve získaných poznatků o veličinách lineárního regresního modelu a vztazích mezi nimi, zejména o vývodech, které jsme uvedli ve větách 1 a 2 .

Abychom mohli využít znalosti o chování vysvětlujících veličin modelu k predikcím vysvětlované proměnné, je třeba vždy :

- a) **Ověřit, zda vlivy působení vysvětlujících proměnných na závisle proměnnou signalizované v pozorovaném období se přenášejí (stejným způsobem jako dosud) do budoucího (predikovaného) období.** Jinými slovy to znamená **posoudit, zda se podstatným způsobem nemění hodnoty modelových parametrů.**
- b) **Stanovit pro předpovědní období vývoj vysvětlujících veličin**, tzn. určit pro každý sloupec matice X jeho „pokračování“ pro období $T+1, T+2, \dots, T+m$. Predikované hodnoty vysvětlujících proměnných získáme pomocí odhadnutých regresních parametrů a předpovědi vysvětlujících veličin v budoucím období.
- c) **Uvážít, zda v budoucím období nedojde ke změně specifikace modelu**, tzn. zda na závisle proměnnou nebudou působit jiné vlivy než dosud (včetně dalších dosud neuvažovaných proměnných), popř. zda se vliv stávajících proměnných neprojeví jiným způsobem než dosud (např. v nelineárním tvaru působení).
- d) **O náhodné složce předpokládáme, že se její vlastnosti v budoucím období (žádným zásadním způsobem) nezmění.** Tzn. budeme předpokládat konstantnost rozptylu náhodných složek a jejich nekorelovanost s nezávisle proměnnými během celého predikovaného období.

Poznámka 1

Ad a) Porušení podmínky a) znamená, že vliv působení vysvětlujících veličin na závisle proměnnou se podstatným způsobem mění a že znalost (jakkoliv dobře) odhadnutého vektoru b nemá pro předpovídání příliš velký význam.

Ad b) Znalost predikcí vysvětlujících veličin je podstatná, neboť (při pevném vektoru parametrů b) na přesnostech těchto předpovědí závisí přesnost předpovědí závisle proměnné. Proto – ve vztahu k podmínce b) – usilujeme o co možná nejvěrohodnější předpovědi sloupců matice X : k jejich určení můžeme uplatnit několik způsobů : prostou trendovou extrapolaci, jiný regresní vztah, kde tato proměnná vystupuje jako vysvětlovaná nebo prostě uvážený expertní odhad vývoje.

Ad c) Pokud v budoucím období začnou na vysvětlovanou proměnnou působit další vlivové faktory, které se neprojeví v pozorovaném období, znamená to, že dochází ke specifikační chybě, která má vliv na výsledné hodnoty predikcí (Obdobná, byť méně vážná situace, nastane, pokud některá z dosud uvažovaných veličin přestane působit vůbec nebo začne působit s jinou intenzitou vlivu.). Jestliže se změní původně lineární model směrem k nelineárním závislostem, jsou nám jen málo platné dosud získané informace o spočtených hodnotách regresních koeficientů.

Ad d) Pokud by se chování náhodné složky významným způsobem změnilo (např. realizace náhodné složky by pocházely z jiného základního statistického rozdělení, než tomu bylo v minulosti) nemohli bychom uplatnit (aspoň ne přesně) např. konstrukce intervalů spolehlivosti. Pokud by došlo ke korelovanosti s některou z vysvětlujících proměnných v budoucnosti, došlo by opět ke zkreslení výsledných odhadů hodnot vysvětlované proměnné v důsledku dopadů na věrohodnost odhadů regresních parametrů, které k predikcím používáme.

V podstatě všechny „pasivní“ předpovědi chování závisle proměnné do budoucnosti (tzn. predikce, které jsou založeny na nezměněných regresních koeficientech) jsou **PODMÍNĚNÝMI PŘEDPOVĚĎMI**, které závisí na tom, jak dobře určíme hodnoty vysvětlujících veličin v předpovídaném období. Tento náhled na **PODMÍNĚNOST PREDIKCÍ** se nazývá podmínkou „**CETERIS PARIBUS**“ (tzn. podmínkou o výpovědích učiněných „za jinak nezměněných okolností“).

Předpověď EX ANTE je tedy vždy PODMÍNĚNOU předpovědí.

Předpověď EX POST může být výjimečně NEPODMÍNĚNOU předpovědí, pokud bychom s jistotou znali všechny budoucí hodnoty vysvětlujících proměnných.

Poznámka 2 Pokud bychom tuto podmínku opustili (mj. připustili proměnlivost vývoje modelových parametrů v čase), získali bychom sice obecnější a patrně i věrohodnější modelové zobrazení ekonomické reality, na druhé straně bychom však potřebovali podstatně bohatší informace o chování modelových proměnných a jejich očekávaných změnách, abychom mohli tento zvětšený počet parametrů modelu statisticky korektně odhadnout .

Pro signalizaci intenzity změn modelových parametrů lze použít několik postupů. Nejčastější je tzv. postupná regrese (výstižnější by byl přívlastek „klouzavá“), kdy celé minulé pozorované období o délce T „prokládáme“ dílčími regresními vztahy, při kterých bereme „klouzavým způsobem“ vždy období o délce $n < T$. Přitom si všímáme změn modelových parametrů, které postupně získáváme při provádění těchto dílčích regresí.

Test stability 1

Nyní formálně vyložíme postup, kterým je možno ověřit, zda jsou parametry β, β^* vystupující v regresním modelu spočtené na základě dvou různě dlouhých časových vzorků v obdobích T_1 a $T_1 + T_2$ v čase stabilní.

V dalším přijmeme následující značení :

Počet pozorování získaných v pozorovaném období	T_1
Počet pozorování rozšířeného datového vzorku	T_2
Počet vysvětlujících proměnných regresního vztahu	k

Počet vysvětlujících proměnných k nezávisí na počtu pozorování, předpokládáme nicméně, že $k < T_1, k < T_2$

Na úlohu se můžeme dívat tak, jako bychom k původnímu vzorku o délce T_1 připojili dodatečných T_2 pozorování, o kterých předpokládáme, že pocházejí z téhož základního souboru: Jako T označíme součet $T_1 + T_2$.

Lineární regresní model nejprve vyjádříme zvlášť pro původní a rozšířený výběr pozorování. Nejdříve dostaneme pro původní model T_1 hodnot vyjádření

(1A)
$$y_I = X_I \beta + \varepsilon_I$$

se součtem čtverců reziduí
$$e_I' \cdot e_I = (y_I - X_I b)' (y_I - X_I b)$$

Dále přejdeme k rozšířenému modelu zahrnujícímu $T_1 + T_2$ pozorování

(1B)
$$y = X \cdot \beta^* + \varepsilon$$

se součtem čtverců reziduí
$$e' \cdot e = (y - X \cdot b^*)' (y - X \cdot b^*)$$

V modelu **(1B)** jsou vektor závisle proměnné, matice vysvětlujících proměnných a vektor náhodných složek sestaveny jako

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{T_1} \\ y_{T_1+1} \\ \dots \\ y_T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{T_1,1} & x_{T_1,2} & x_{T_1,3} & \dots & x_{T_1,k} \\ x_{T_1+1,1} & x_{T_1+1,2} & x_{T_1+1,3} & \dots & x_{T_1+1,k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{T1} & x_{T2} & x_{T3} & \dots & x_{Tk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_{T_1} \\ \varepsilon_{T_1+1} \\ \dots \\ \varepsilon_{T1} \end{pmatrix}$$

V původním i v rozšířeném vzorku samostatně odhadneme vektor parametrů b .

V **(1A)** půjde o odhad ve tvaru

$$b = (X_I' X_I)^{-1} \cdot X_I' \cdot y_I \quad \text{s reziduy} \quad e_I = y_I - X_I \cdot b$$

V **(1B)** půjde o odhad ve tvaru

$$b^* = (X'X)^{-1} \cdot X' \cdot y \quad \text{s reziduy} \quad e = y - X \cdot b^*$$

Přidržíme se značení b jako OLS-odhad β , b^* jako OLS-odhad β^* .

Připomeňme, že všechny vektory b , b^* , β , β^* mají shodnou délku k .

Pokud nedojde ke zřetelným rozdílům v získaných hodnotách b , b^* , lze soudit na stabilní regresní vztah, v němž přidáním dalších pozorování nedojde ke změně modelové struktury.

Nulová hypotéza o shodě vektorů regresních koeficientů spočtených z původního i rozšířeného vzorku má tedy tvar:

$$H_0 : \beta^* = \beta$$

Test této hypotézy založíme, přirozeně, na rozdílu $b^* - b$, resp. na porovnání součtu čtverců reziduí odvozených na základě odhadnutých vektorů b , b^* .

Pro součty čtverců reziduí získané z původního a rozšířeného regresního modelu platí následující vztahy:

$$(2A) \quad e_1' e_1 = \varepsilon_1' M_{T1} \varepsilon_1, \quad \text{kde} \quad M_{T1} = I_{T1} - X_1 (X_1' X_1)^{-1} X_1'$$

$$(2B) \quad e' e = \varepsilon' M_T \varepsilon, \quad \text{kde} \quad M_T = I_T - X (X' X)^{-1} X'$$

Na základě dřívějších poznatků lze vyvodit, že kvadratické formy $e_1' e_1$ a $e' e$ mají

$$e_1' e_1 / \sigma^2 : \chi^2 - \text{rozdělení o } T_1 - k \text{ stupních volnosti.}$$

$$e' e / \sigma^2 : \chi^2 - \text{rozdělení o } T - k \text{ stupních volnosti.}$$

Dimenze matice M_{T1} je T_1 , dimenze matice M_T je rovna T . Matici M_{T1} "rozšíříme" na stejnou dimenzi jako má matice M_T přidáním nulových prvků¹. Takto vzniklou matici označíme M_{T1}^* . Matici M_T současně rozdělíme na bloky synchronně s maticí M_{T1} :

$$(3) \quad M_{T1}^* = \begin{pmatrix} M_{T1} & 0 \\ 0 & 0 \end{pmatrix} \quad M_T = \begin{pmatrix} M_{T1} & M_{T12} \\ M_{T21} & M_{T2} \end{pmatrix}$$

přičemž jednotlivé bloky matice M_{T1}^* jsou obsazeny těmito maticovými útvary:

$$M_{T1} = I_{T1} - X_1 (X' X)^{-1} X_1'$$

$$M_{T12} = -X_1 (X' X)^{-1} X_2'$$

$$M_{T21} = -X_2 (X' X)^{-1} X_1'$$

$$M_{T2} = I_{T2} - X_2 (X' X)^{-1} X_2'$$

¹ Rozšíření je nutné, protože dále budeme porovnávat kvadratické formy, které mají shodné proměnné (prvky vektoru e) a tedy i délku vektoru těchto proměnných $T_1 + T_2$.

Jak matice M_T tak $M_{T_1}^*$ jsou idempotentní matice hodností $(T - k)$ resp. $(T_1 - k)$. Platí tedy vlastnost

$$(4) \quad M_T = M_T \cdot M_T \quad M_{T_1}^* = M_{T_1}^* \cdot M_{T_1}^*$$

Dále lze snadno dokázat, že platí

$$(M_T - M_{T_1}^*) \cdot M_{T_1}^* = 0^2$$

a že matice $(M_T - M_{T_1}^*)$ je rovněž idempotentní. Proto (připomeňme, že hodnost idempotentní matice je rovna její stopě) platí

$$\text{Tr}(M_T - M_{T_1}^*) = T - T_1 = T_2.$$

Obě kvadratické formy lze nyní vyjádřit ve stejných proměnných ε (přirozeně však s různými maticemi – byť stejných dimenzí - těchto forem):

$$(5A) \quad e'e - e_1'e_1 = \varepsilon'(M_T - M_{T_1}^*)\varepsilon$$

$$(5B) \quad e_1'e_1 = \varepsilon' M_{T_1}^* \varepsilon$$

Obě tyto kvadratické formy mají $\sigma^2 \cdot \chi^2$ - rozdělení:

Kvadratická forma (5A) má $\sigma^2 \cdot \chi^2$ - rozdělení o $(T - k) - (T_1 - k) = T_2$ stupních volnosti.

Kvadratická forma (5B) má $\sigma^2 \cdot \chi^2$ - rozdělení o $(T_1 - k)$ stupních volnosti.

Rozdělení obou těchto kvadratických forem jsou v důsledku platnosti vztahu

$$(M_T - M_{T_1}^*) M_{T_1}^* = 0 \quad \text{vzájemně nezávislá.}$$

Odtud plyne, že příslušná **podílová testová statistika**

$$F_{ee_1} = \frac{(e'e - e_1'e_1) / T_2}{e_1'e_1 / (T_1 - k)} = \frac{\frac{(e'e - e_1'e_1)}{\sigma^2 T_2}}{\frac{e_1'e_1}{\sigma^2 (T_1 - k)}}$$

má Fisher-Snedecorovo F - rozdělení o T_2 a $T_1 - k$ stupních volnosti.

Uvedené zjištění lze nyní užít pro testování hypotézy, že regresní koeficienty jsou pro obě situace (původní i rozšířený výběr) shodné. Reziduální hodnoty, na jejichž základě konstruujeme testovou statistiku, jsou totiž určeny právě v závislosti na odhadnutých regresních koeficientech :

- Pokud spočtená testová statistika překročí teoretickou kritickou hodnotu $F^*_{(T_2, T_1 - k)}$ na zvolené hladině významnosti α , mluví to v neprospěch totožnosti odhadnutých parametrů.

² I odtud vyplývá nutnost rozšíření matice M_{T_1} na matici $M_{T_1}^*$. Jinak by násobení a odečtení nešlo provést.

- Pokud naopak spočtená statistika kritické hodnoty $F^*_{(T_2, T_1 - k)}$ nedosáhne, lze s pravděpodobností $100 * (1 - \alpha)$ usuzovat na shodu modelové podoby pro původní i rozšířený vzorek.

Výše popsaný postup je dobrým indikátorem toho, zda si model ponechává i po rozšíření datového vzorku o dalších T_2 pozorování původní modelovou strukturu.

Zřetelná rozdílnost v hodnotách parametrů ${}_{T_1}b$, ${}_T b^*$ ³, kterou zaznamenáváme nepřímo právě přes různé chování reziduálních hodnot svědčí o znatelné změně modelové struktury po připojení dalších T_2 pozorování.

Obdobný postup lze zobecnit také pro případ, že bychom pracovali s více než dvěma náhodnými výběry.

Alternativní test stability 2

Tentokrát přistoupíme k **testování na základě dvou nepřekrývajících se vzorků pozorování**, jednoho o délce T_1 , druhého o délce T_2 . Regrese mají tvar

$$(11A) \quad y_1 = X_1 \cdot \beta_1 + \varepsilon_1 \quad \text{pro } t = 1, 2, \dots, T_1$$

se součtem čtverců reziduí $e_1' e_1 = (y_1 - X_1 b_1)' (y_1 - X_1 b_1)$, resp.

$$(11B) \quad y_2 = X_2 \cdot \beta_2 + \varepsilon_2 \quad \text{pro } t = 1, 2, \dots, T_2$$

se součtem čtverců reziduí $e_2' e_2 = (y_2 - X_2 b_2)' (y_2 - X_2 b_2)$

Testovaná hypotéza nyní bude mít tvar $H_0 : \beta_1 = \beta_2$

příčemž pokud nebude tato hypotéza zamítnuta, bude to svědčit o statisticky nevýznamných rozdílech mezi oběma vektory.

Testovací statistiku odvodíme následovně: sloučíme oba vzorky dohromady, přičemž model odhadneme jednou s ohledem na omezení $\beta_1 = \beta_2 = \beta$ (tj. ve znění hypotézy H_0), podruhé bez něho. Z obou dílčích modelů vytvoříme jeden společný o T pozorováních:

$$(12) \quad y = X \cdot \beta + \varepsilon,$$

$$\text{kde } y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

[Vektor y má délku T , matice X má rozměry $T \times 2k$, vektor β délku $2k$, vektor ε má délku T]

Odhad vektoru β pořízený obyčejnou metodou nejmenších čtverců na základě vzorku o rozsahu $T = T_1 + T_2$ pozorování má tvar $b = (b_1, b_2)'$, kde

³ Index vlevo dole značí rozsah vzorků, ze kterých je ten-který odhad spočten. Oba vektory odhadů parametrů přitom mají stejnou délku k .

$$b_1 = (X_1' X_1)^{-1} X_1' y_1 \quad \text{s reziduy} \quad e_1 = y_1 - X_1 b_1$$

V (12) půjde o odhad ve tvaru

$$b_2 = (X_2' X_2)^{-1} X_2' y_2 \quad \text{s reziduy} \quad e_2 = y_2 - X_2 b_2 .$$

Za těchto okolností lze rozdělit celkový součet čtverců reziduí na dvě „disjunktní“ části:

neomezený SSE:

$$e^{*'} e^* = e_1' e_1 + e_2' e_2 =$$

$$= (y_1 - X_1 b_1)' (y_1 - X_1 b_1) + (y_2 - X_2 b_2)' (y_2 - X_2 b_2) = (y - Xb)' (y - Xb) .$$

Tento výraz (*neomezený součet čtverců reziduí*) má χ^2 - rozdělení s $T - 2k$ stupni volnosti. (První skalární součin má $T_1 - k$, druhý $T_2 - k$ stupňů volnosti) .

Dále se odhadne stejný model z téhož počtu pozorování $T = T_1 + T_2$ pozorování, avšak **při respektování omezení** $\beta_1 = \beta_2 = \beta$, kde vektor β má k parametrů.⁴ Příslušný (omezený) součet čtverců reziduí má pak tvar

omezený SSE:

$$e' e = (y - Xb)' (y - Xb)$$

Tento omezený SSE má tentokrát opět χ^2 - rozdělení, ale pouze s $T - k$ stupni volnosti. Počet stupňů volnosti je oproti předchozí situaci větší právě o oněch k omezujících podmínek tvaru $b_1(j) = b_2(j) ; j = 1, 2, \dots, k$.⁵

Do čitatele výrazu, který tvoří testovou statistiku, musíme nyní dosadit kvadratickou formu, která je nezávislá na **omezeném SSE**: Tuto podmínku splňuje kvadratická forma tvaru

$$(e' e - e^{*'} e^*) / \sigma^2 \quad , \text{ která má } \chi^2\text{-kvadrát-rozdělení o } k$$

stupních volnosti ($e' e$ má $T - k$ stupňů volnosti, $e^{*'} e^*$ má $T - 2k$ stupňů volnosti)

Testovací statistiku získáme tedy **jako podíl** dvou χ^2 - rozdělení dělených příslušnými stupni volnosti:

$$F_{ee_1} = \frac{\frac{e' e - e^{*'} e^*}{k \cdot \sigma^2}}{\frac{e^{*'} e^*}{(T - 2k) \cdot \sigma^2}} = \frac{(e' e - e^{*'} e^*) / k}{e^{*'} e^* / (T - 2k)}$$

která **má** opět **F - rozdělení s počtem stupňů volnosti** $k, T - 2k$. Síla tohoto testu je větší než předchozího testu, nutnou podmínkou je však, aby $k < T_1, k < T_2$. Způsob posuzování testové statistiky je stejný jako u předchozího testu.

^{4/} Výpočet je ovšem nutno provést jiným způsobem než pomocí OLS , a to prostou metodou nejmenších čtverců s dodatečnou informací OLS-AI . Stručný výklad této metody uvedeme v samostatné části.

^{5/} Tento způsob testování s porovnáváním omezeného a neomezeného SSE je zpravidla spojen s testy založenými na tzv. **věrohodnostním poměru** (*likelihood ratio*)