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:*** $Cov(\varepsilon,\varepsilon) = \sigma^2 . I_T$, 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 hodnost matice vysvětlujících proměnných*** $h(X) = k$
5. ***Normalita T-rozměrného vektoru náhodných složek*** *s nulovým vektorem středních hodnot a s diagonální kovarianční maticí*

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

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, …T+m. Predikované hodnoty vysvětlujících proměnných získáme pomocí odhadnutých regresních parametrů a předpovědí vysvětlujících veličin v budoucím období.*

c) ***Uvážit, 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 neprojevily 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 předpověd**í.*

***Předpověď EX POST může být** výjimečně **NEPODMÍNĚNOU předpověd**í, 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 pozorovaní 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) $$y_1 = X_1\beta + \varepsilon_1$$

se součtem čtverců reziduí $$e_1'.e_1 = (y_1 - X_1 b)'(y_1 - X_1 b)$$

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

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

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

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

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

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

V (1A) půjde o odhad ve tvaru

$$b = (X_1'X_1)^{-1}.X_1'.y_1 \qquad \text{s reziduy} \qquad e_1 = y_1 - X_1.b$$

V (1B) půjde o odhad ve tvaru

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

Přidržujeme 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) $e_1'e_1 = \varepsilon_1' M_{T1} \varepsilon_1$, kde $M_{T1} = I_{T1} - X_1(X_1'X_1)^{-1}X_1'$

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

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

$e_1'e_1/\sigma^2$: χ^2 - rozdělení o $T_1 - k$ stupních volnosti.

$e'e/\sigma^2$: χ^2 - rozdělení o $T - k$ 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 prvků[1]. Takto vzniklou matici označíme M^*_{T1}. Matici M_T současně rozdělíme na bloky synchronně s maticí M_{T1}:

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

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

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

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

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

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

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

Jak matice M_T tak M^*_{T1} jsou idempotentní matice hodností $(T-k)$ resp. (T_1-k). Platí tedy

(4) $$M_T = M_T.M_T \qquad M_{T1}^* = M_{T1}^*.M_{T1}^*$$

Dále lze snadno dokázat, že platí

$$(M_T - M_{T1}^*).M_{T1}^* = 0^2$$

a že matice $(M_T - M_{T1}^*)$ je rovněž idempotentní. Proto (připomeňme, že hodnost idempotentní matice je rovna její stopě) platí

$$Tr(M_T - M_{T1}^*) = T - T_1 = T_2 .$$

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

(5A) $$e'e - e_1'e_1 = \varepsilon'(M_T - M_{T1}^*).\varepsilon$$

(5B) $$e_1'e_1 = \varepsilon' M_{T1}^*.\varepsilon$$

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

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

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

Rozdělení obou těchto kvadratických forem jsou v důsledku platnosti vztahu

$$(M_T - M_{T1}^*)M_{T1}^* = 0$$ vzájemně nezávislá.

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

$$F_{ee_1} = \frac{(e\acute{e} - e_1\acute{e}_1)/T_2}{e'_1e_1/(T_1-k)} = \frac{\dfrac{(e\acute{e} - e_1\acute{e}_1)}{\sigma^2 T_2}}{\dfrac{e'_1e_1}{\sigma^2(T_1-k)}}$$

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 :

[2] I odtud vyplývá nutnost rozšíření matice M_{T1} na matici M^*_{T1}. Jinak by násobení a odečtení nešlo provést.

- 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.
- Pokud naopak spočtená statistika kritické hodnoty $F^*_{(T_2,T_1-k)}$ nedosáhne, lze s pravděpodobností $1-\alpha$ 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 ${}_{T1}b$, ${}_{T}b^*$, 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) $y_1 = X_1.\beta_1 + \varepsilon_1$ pro $t = 1,2,...,T_1$

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

(11B) $y_2 = X_2.\beta_2 + \varepsilon_2$ pro $t = 1,2,...,T_2$

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 $H_0 : \beta_1 = \beta_2$

přičemž 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) $y = X.\beta + \varepsilon$,

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

[Vektor y má délku T , matice X má rozměry $Tx2k$, 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 $b = (b_1, b_1)'$, kde

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

V (12) půjde o odhad ve tvaru

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

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.[3] Příslušný (omezený) součet čtverců reziduí má pak tvar

omezený **SSE**: $e'e = (y - Xb)'(y - Xb)$

Tento omezený SSE má tentokrát χ^2 - rozdělení (jen) s $T - k$ stupni volnosti. Počet stupňů volnosti je oproti předchozí situaci větší právě o oněch k omezujících podmínek tvaru $b_1(j) = b_2(j)\ ;\ j = 1,2,\ldots,k.$ [4]

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

$(e'e - e^{*}{}'e^{*})/\sigma^2$, která má χ^2 –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:

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

---

[3]/ 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.

[4] / Tento způsob testování s porovnáváním omezeného a neomezeného SSE je zpravidla spojen s testy založenými na tzv. **věrohodnostním poměru** (***likelihood ratio***)

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.