4. cvičení

Verifikace LRM

4.1 Úvod do problematiky

4.1.1 Verifikace (ověření) lineárního regresního modelu

Odhadnutý ekonometrický model je nutné před jeho aplikací verifikovat, tzn. ověřit, zda jsou odhadnuté parametry v souladu s výchozími hypotézami a zda mají požadované ekonomické, statistické, ekonometrické či matematické charakteristiky.

Verifikaci modelu lze rozdělit do čtyř kroků, a to podle toho, co je ověřováno.

(i) *Ekonomická verifikace*
V rámci ekonomické verifikace se posuzuje zejména směr a intenzita působení vysvětlujících proměnných na proměnnou vysvětlovanou. Ověřuje se zde správnost znamének a velikost číselných hodnot odhadnutých parametrů. Pokud získané parametry nejsou v souladu s předpoklady, je zpravidla nutné ověřit správnost specifikace modelu.

(ii) *Statistická verifikace*
Statistická verifikace slouží k posouzení statistické významnosti odhadnutých parametrů, jednotlivých rovnic i celého modelu.
V rámci statistické verifikace se hodnotí:
a. shoda odhadnutého modelu s daty;
b. statistická významnost odhadnutých parametrů.

(iii) *Ekonometrická verifikace*
V rámci ekonometrické verifikace se ověřují podmínky nutné pro aplikaci konkrétních ekonometrických metod, testů a technik, tj. **předpoklady ekonometrického modelu** specifikované předchozím cvičením.

(iv) *Matematická verifikace*
Matematická verifikace slouží k posouzení správnosti výpočtu parametrů. Správnost výpočtu je ověřena, pokud se průměrná hodnota vysvětlované proměnné rovná teoretické hodnotě, získané dosazením průměrných hodnot vysvětlujících proměnných modelu do odhadnuté rovnice.

Add ii.a. Shoda odhadnutého modelu s daty

Shoda odhadnutého modelu s daty se v případě lineární funkce posuzuje pomocí koeficientu vícenásobné determinace R^2. Tento ukazatel je založen na rozkladu celkového rozptylu vysvětlované proměnné (S_y^2) na rozptyl teoretický, někdy též označován za regresní ($S_{\hat{y}}^2$) a reziduální (S_u^2):

$$S_y^2 = S_{\hat{y}}^2 + S_u^2 \qquad (4.1)$$

$$S_y^2 = \frac{\sum_{t=1}^{n}(y_t - \bar{y})^2}{n} \quad , \qquad (4.2)$$

kde y_t ....jsou skutečné hodnoty vysvětlované proměnné v jednotlivých letech pozorování,
$\bar{y}$.......... je průměr skutečných hodnot vysvětlované proměnné,
n ....je délka časové řady.

$$S_{\hat{y}}^2 = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{n} \qquad (4.3)$$

kde $\hat{y}_t$.......... jsou teoretické hodnoty vysvětlované proměnné v jednotlivých letech pozorování.

$$S_u^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n} \qquad (4.4)$$

Koeficient vícenásobné determinace je dán vztahem:

$$R^2 = 1 - \frac{S_u^2}{S_y^2} \qquad (3.5)$$

Vyjadřuje se obvykle v % a udává, **z kolika % jsou změny závisle proměnné vysvětleny změnami nezávisle proměnných.**

Hodnota R^2 se pohybuje od 0 % do 100 %. Pokud $R^2 = 0\ \%$, všechny odhadnuté koeficienty jsou nulové, celkový rozptyl je roven reziduálnímu a daná funkce nevysvětluje vůbec zkoumaný vztah. Naopak $R^2 = 100\ \%$ nastane, když všechna rezidua jsou nulová, tudíž také reziduální rozptyl je nulový a daná funkce plně vystihuje zkoumaný vztah.

Protože hodnota R^2 nikdy neklesne (zpravidla vždy vzroste) přidáním dalších vysvětlujících proměnných do modelu, je často používán korigovaný koeficient vícenásobné determinace:

$$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-p} \qquad (4.6)$$

kde p .... je počet odhadovaných parametrů v dané rovnici.

Hodnota korigovaného koeficientu determinace je zpravidla nižší, než hodnota R^2. Odchylka těchto dvou koeficientů se snižuje s růstem počtu stupňů volnosti *(n-p)*. Při velkém počtu stupňů volnosti se R^2 a $\overline{R^2}$ liší velice málo. Při malém počtu stupňů volnosti může nabývat $\overline{R^2}$ i záporných hodnot. V takovém případě se hodnota korigovaného koeficientu vícenásobné determinace interpretuje jako nulová.

Statistickou významnost modelu jako celku lze testovat pomocí F-testu, v jehož rámci se porovnává *F poměr* s tabulkovou hodnotou F^*. Je-li *F poměr* větší než tabulková hodnota na zvolené hladině významnosti a při daném počtu stupňů volnosti, zamítá se nulová hypotéza o statistické nevýznamnosti R^2, a tedy shoda odhadnutého modelu s daty je statisticky významná.

U nelineární funkce je jako míra těsnosti závislosti používán index determinace I^2, jeho výpočet i interpretace se však shodují s R^2.

Add ii.b. Testování statistické významnosti odhadnutých parametrů

Statistická významnost jednotlivých strukturálních parametrů se hodnotí *t-testem*. Při výpočtu testovacího kritéria, *t-hodnoty*, je používán korigovaný reziduální rozptyl. Korekce se provádí opět počtem stupňů volnosti v daném vztahu.

Korigovaný reziduální rozptyl je tedy určen jako:

$$\overline{S_u^2} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n - p} . \qquad (4.7)$$

Ověření statistické významnosti strukturálních parametrů se liší v závislosti na použité metodě, kterou byly parametry odhadnuty.

Doporučený postup statistické verifikace strukturálních parametrů LRM odhadnutých BMNČ:

(i) Výpočet matice pro ověření statistické významnosti parametrů: $\left(X^T X\right)^{-1}$.

(ii) Výpočet korigovaného reziduálního rozptylu.

(iii) Výpočet rozptylu odhadnutých parametrů:

$$S_{ii} = \overline{S_u^2}(X^T X)^{-1} = \begin{pmatrix} S_{11} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & S_{jj} \end{pmatrix} \qquad (4.8)$$

.

Prvky na hlavní diagonále matice vzniklé vynásobením korigovaného reziduálního rozptylu $\overline{S_u^2}$ a matice $\left(X^T X\right)^{-1}$ jsou rozptyly odhadnutých parametrů S_{ii}.

(iv) Výpočet standardní chyby odhadnutých parametrů: $S_{bi} = \sqrt{S_{ii}}$. $\qquad$ *(4.9)*

Vyčíslení standardních chyb jednotlivých parametrů jako odmocniny prvků z hlavní diagonály výše uvedené matice S_{ii}.

(v) Výpočet testovacího kriteria:

$$t - hodnota = \frac{hodnota\ parametru}{chyba\ odhadu} = \frac{|\gamma_{it}|}{S_{bi}} \qquad (4.10)$$

(vi) Zjištění statistické významnosti odhadnutých parametrů: porovnání vypočtené *t-hodnoty* s tabulkovou hodnotou t-testu na zvolené hladině významnosti s přihlédnutím k příslušnému počtu stupňů volnosti t_α.

Je-li $t > t_\alpha$, zamítá se nulová hypotéza o statistické nevýznamnosti parametrů. Vysvětlující proměnná je z hlediska svého vlivu na vysvětlovanou proměnnou na hladině významnosti α a při n-p stupních volnosti významnou proměnnou.

Je-li $t < t_\alpha$, s pravděpodobností $100(1\text{-}\alpha)\%$ není parametr statisticky významný, tj. statisticky významně odlišný od nuly.

Zamítnutí nulové hypotézy ještě neznamená, že bodové odhady parametrů jsou přesnými odhady jejich skutečných hodnot. Pro určení stupně shody skutečné hodnoty parametru s odhadem se stanovuje interval spolehlivosti, tzv. konfidenční interval. Neboli hledají se meze, v nichž se bude skutečná hodnota parametru při opakovaných výběrech nacházet s určitým stupněm spolehlivosti, tj. s určitou zvolenou pravděpodobností.

Intervalový odhad parametrů se stanovuje pomocí vztahu:

$$\gamma_{ii\ interval} = \gamma_{ii} \pm t_\alpha S_{bi}. \qquad (4.11)$$

Odhadnutý parametr se významně liší od nuly, pokud tento interval nulu neobsahuje. Obsahuje-li konfidenční interval nulu, je parametr statisticky nevýznamný.

Add iii. Ekonometrická verifikace – specifikační předpoklady – Multikolinearita

Multikolinearita vyjadřuje závislost mezi dvěma či více vysvětlujícími proměnnými v rovnici. Perfektní multikolinearita nastává v případech, kdy závislost mezi dvěma či více vysvětlujícími proměnnými je deterministická, tj. párový korelační koeficient nebo koeficient vícenásobné korelace je roven 1. V případě, že je v modelu přítomna perfektní multikolinearita, nelze takovýto model odhadnout.
Při výskytu vysoké multikolinearity není možné separovat vlivy jednotlivých vysvětlujících proměnných na vysvětlovanou proměnou, a proto je vysoká multikolinearita nežádoucí.
Vysoká multikolinearita se zpravidla vyskytuje tehdy, když hodnoty vysvětlujících proměnných mají nízkou variabilitu. Z toho plyne, že vyvarování se problému přítomnosti vysoké multikolinearity lze dosáhnout zajištěním dostatečné variability vysvětlujících proměnných. Avšak určitá výše multikolinearity je v modelu vždy přítomna.
Přítomnost vysoké multikolinearity neumožňuje dosáhnout přesného odhadu parametrů vysvětlujících proměnných, které multikolinearitu způsobují. Tato skutečnost působí problémy při aplikaci modelu ve strukturální analýze, kde co nejlepší znalost velikosti parametrů je nezbytností.

Přítomnost vysoké multikolinearity lze identifikovat vyčíslením korelační matice. Korelační matice obsahuje párové korelační koeficienty jednotlivých vysvětlujících proměnných a lze ji vyčíslit z následujícího vztahu:

$$\underline{X'}^{T}\underline{X'} \qquad (4.12)$$

,

kde X´ je matice normalizovaných vektorů, které lze získat podle *(4.12)*

$$x'_{it} = \frac{x_{it} - \bar{x}_i}{\sqrt{n}.\sigma_{x_i}} \qquad \begin{matrix} i = (1.......k) \\ t = (1.......n) \end{matrix} \qquad (4.13)$$

,

kde x_{it} je hodnota i-té vysvětlující proměnné v čase t, $\bar{x}_i$ je její průměr a σ_{x_i} směrodatná odchylka n je počet pozorování.

Z konstrukce korelační matice je zřejmé, že tato matice je symetrická podle hlavní diagonály. Vysoká multikolinearita je přítomna tehdy, jestliže některý z párových korelačních koeficientů dosahuje vysokých hodnot. Obvykle je za vysokou úroveň považována hodnota 0,8 a vyšší (uvedené však nemusí být vždy pravidlem).

Multikolinearita může být snížena použitím speciálních umělých proměnných, tzv. dummy proměnných (viz předcházející cvičení) nebo vhodnou transformací podkladových údajů (např. vyjádřením proměnné(ých) v postupných diferencích nebo relativně). V krajním případě lze vysokou multikolinearitu odstranit tím, že proměnnou způsobující vysokou multikolinearitu z modelu vypustíme. Multikolinearita může být také ignorována, zejména pokud jsou parametry vysvětlujících proměnných statisticky významné. Pro účely analýzy je však nutné zdůraznit, že u proměnných, u nichž byla detekována vysoká multikolinearita, nelze interpretovat parametry jednotlivě (separovaně), protože působí na endogenní proměnnou společně do dané míry viz párové korelační koeficienty.

4.2 Praktická cvičení

Úkoly:

1. Otestujte v Excelu statistickou významnost odhadnutých parametrů pomocí t-testu a vypočítejte koeficient determinace včetně jeho korigované formy. Kritické hodnoty t-testu jsou uvedeny v příloze č. 2.

   Počet pozorování =

   Počet stupňů volnosti =

   Korigovaný reziduální rozptyl ($\overline{S_u^2}$) =

   Matice pro ověření statistické významnosti parametrů: $\left(X^T X\right)^{-1}$

$$\begin{vmatrix} 6.639017 & & & & \\ & 0.001055 & & & \\ & & 0.001079 & & \\ & & & 0.00168 & \\ & & & & 0.022937 \end{vmatrix}$$

2. Vypočítejte interval, ve kterém se budou odhadnuté parametry nacházet s pravděpodobností 95 %.

3. Srovnejte vypočítané výsledky s výstupy SW Gretl.

4. Proveďte ekonometrickou verifikaci.

Výstup ze SW Gretl:

```
Model 1: OLS, za použití pozorování 1993-2018 (T = 26)
Závisle proměnná: SpDM

             koeficient   směr. chyba   t-podíl   p-hodnota
  ---------------------------------------------------------
  const      16,6144      5,38530        3,085    0,0056    ***
  SpCDM      -0,265937    0,0856577     -3,105    0,0054    ***
  SpCHM      -0,0466104   0,0686389     -0,6791   0,5045
  SpCVM       0,0909532   0,0678853      1,340    0,1946
  Prij        0,992597    0,316539       3,136    0,0050    ***

  Střední hodnota závisle proměnné     22,32346
  Sm. odchylka závisle proměnné        5,088549
  Součet čtverců reziduí               91,73501
  Sm. chyba regrese                    2,090056
  Koeficient determinace               0,858288
  Adjustovaný koeficient determinace   0,831295
  F(4, 21)                             31,79693
  P-hodnota(F)                         1,23e-08
  Logaritmus věrohodnosti             -53,28290
  Akaikovo kritérium                   116,5658
  Schwarzovo kritérium                 122,8563
  Hannan-Quinnovo kritétium            118,3772
  rho (koeficient autokorelace)        0,605317
  Durbin-Watsonova statistika          0,771180
zde je poznámka o zkratkách statistik modelu

Pomine-li se konstanta, p-hodnota byla nejvyšší pro proměnnou 3 (SpCHM)
   Breusch-Paganův test heteroskedasticity -
     Nulová hypotéza: není zde heteroskedasticita
     Testovací statistika: LM = 5,79731
     s p-hodnotou = P(Chí-kvadrát(4) > 5,79731) = 0,214805

   LM test pro autokorelaci až do řádu 1 -
     Nulová hypotéza: žádná autokorelace
     Testovací statistika: LMF = 18,5123
     s p-hodnotou = P(F(1, 20) > 18,5123) = 0,000346757

   Test normality reziduí -
     Nulová hypotéza: chyby jsou normálně rozdělené
     Testovací statistika: Chí-kvadrát(2) = 0,922766
     s p-hodnotou = 0,630411

   Test RESET pro specifikaci -
     Nulová hypotéza: specifikace je adekvátní
     Testovací statistika: F(2, 19) = 8,37232
     s p-hodnotou = P(F(2, 19) > 8,37232) = 0,00246977

   Chowův test pro strukturální zlom při pozorování 2005 -
     Nulová hypotéza: žádný strukturální zlom
     Testovací statistika: F(5, 16) = 11,1596
     s p-hodnotou = P(F(5, 16) > 11,1596) = 9,23738e-005

   Durbin-Watsonova statistika = 0,77118
   p-hodnota = 1,51806e-005
```

Korelační matice

z původních dat

Korelační koeficienty, za použití pozorování 1993 - 2018
5% kritická hodnota (oboustranná) = 0,3882 pro n = 26

SpDM	SpCDM	SpCHM	SpCVM	Prij	
1,0000	0,3090	0,8260	-0,1856	0,8822	SpDM
	1,0000	0,6667	0,4161	0,5890	SpCDM
		1,0000	0,0472	0,9701	SpCHM
			1,0000	-0,1180	SpCVM
				1,0000	Prij

z upravených dat

Korelační koeficienty, za použití pozorování 1994 - 2018
5% kritická hodnota (oboustranná) = 0,3961 pro n = 25

SpDM	SpCDM	SpCHM	SpCVM	d_Prij	
1,0000	0,2520	0,7845	-0,4331	-0,0069	SpDM
	1,0000	0,6599	0,3737	0,0253	SpCDM
		1,0000	-0,1702	0,0186	SpCHM
			1,0000	0,2069	SpCVM
				1,0000	d_Prij

Úkoly k samostatnému procvičení

1. Na základě odhadu modelu z úkolů k samostatnému procvičení z minulého cvičení otestujte statistickou významnost odhadnutých parametrů a vypočítejte R^2.
2. Odhadněte model v SW Gretl.
3. Proveďte ekonometrickou verifikaci.