Úvod do analýzy časových řad

Statistika II

Jiří Neubauer

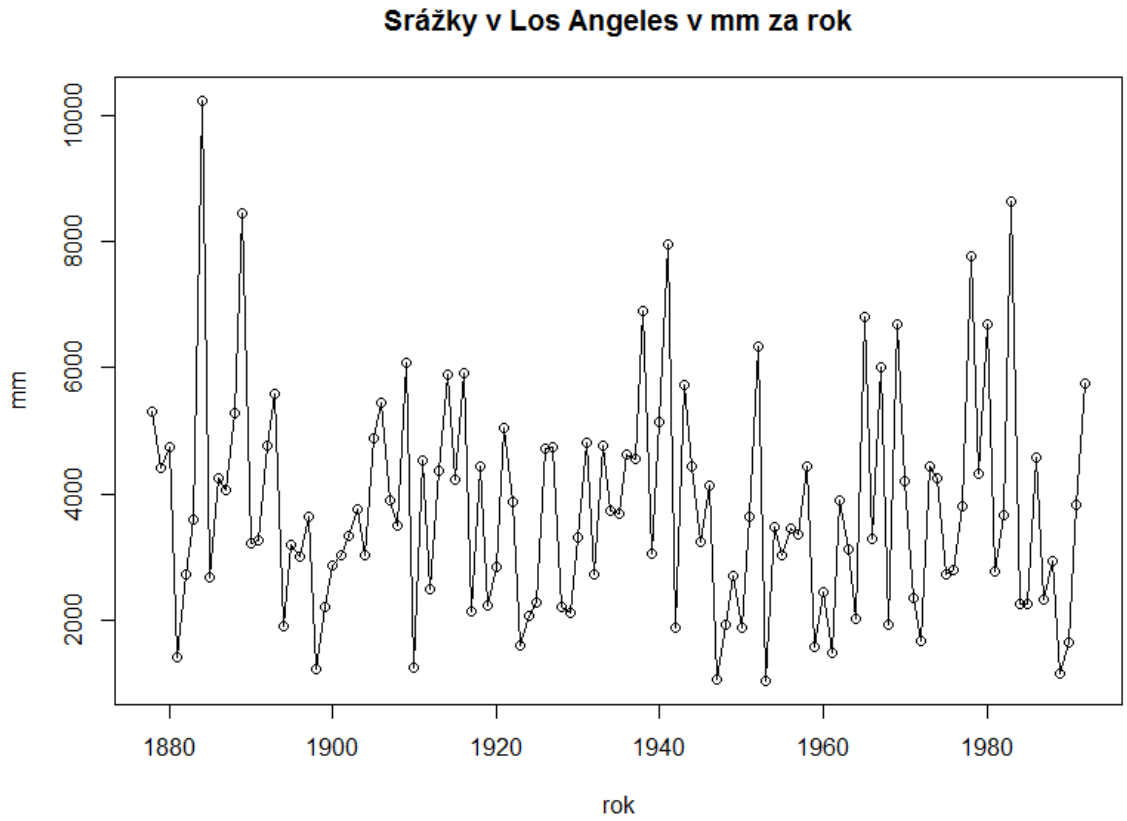
Katedra ekonometrie FVL UO Brno
kancelář 69a, tel. 973 442029
email:Jiri.Neubauer@unob.cz

Úvod do analýzy časových řad

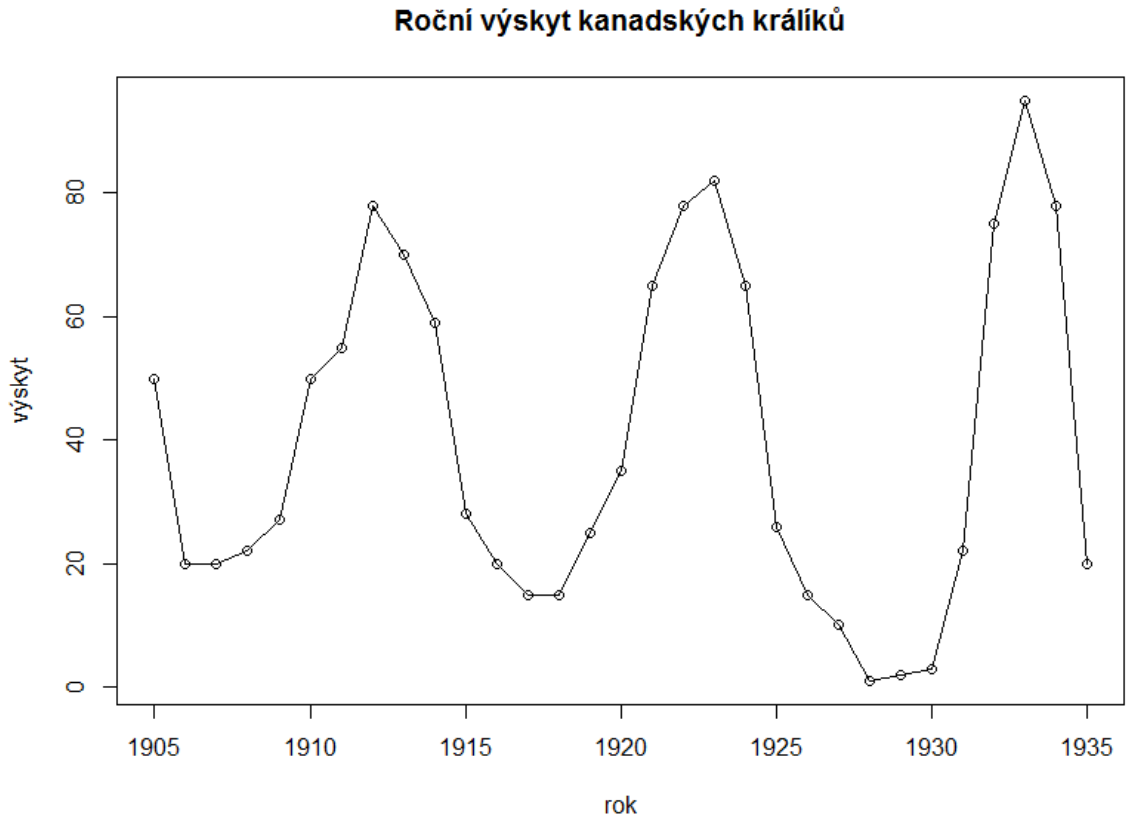
Časová řada – konečná posloupnost reálných hodnot určitého sledovaného ukazatele měřeného v určitých časových intervalech

- okamžikové – např. kurs dolaru k určitému datu, ...
- intervalové – např. objem výroby za měsíc, ...

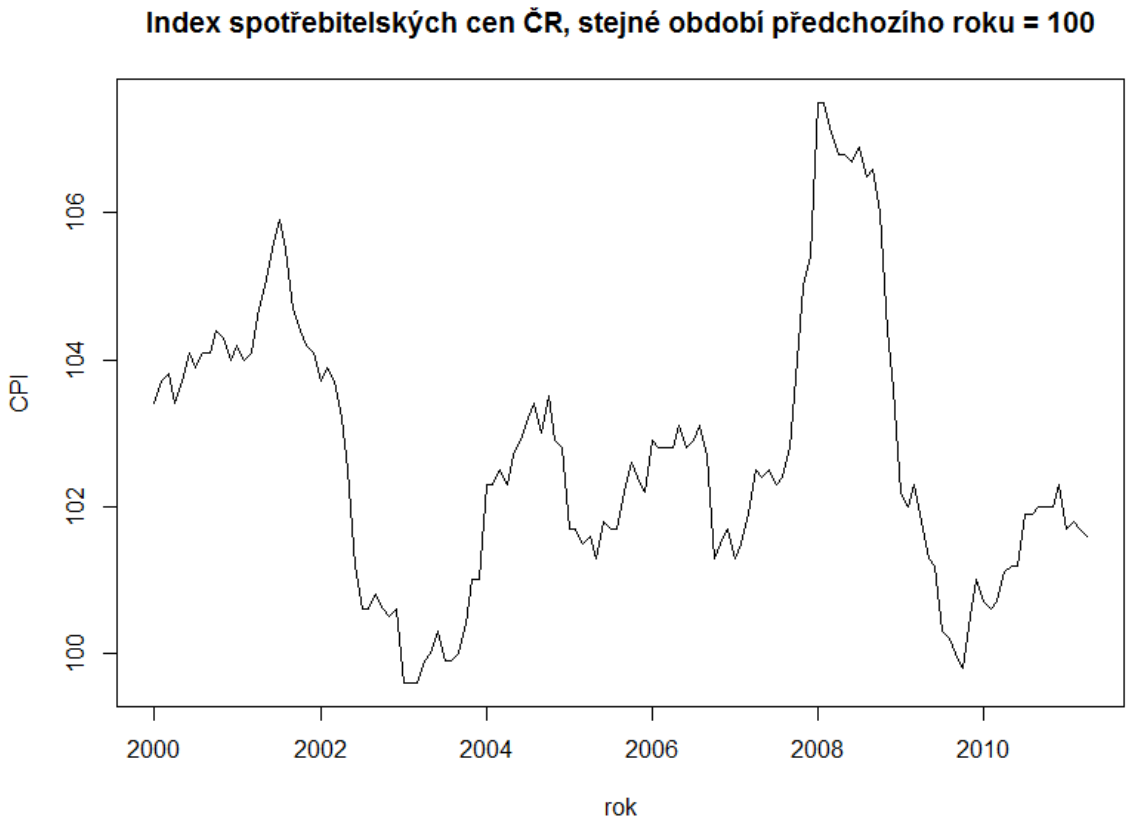
Úvod do analýzy časových řad



Úvod do analýzy časových řad



Úvod do analýzy časových řad



Úvod do analýzy časových řad

Cíl analýzy časové řady: porozumět mechanismu, který určuje hodnoty sledované veličiny a předpovědět její vývoj.

K pochopení vývoje sledované veličiny slouží **model časové řady**, matematicky vyjádřený vztah mezi vysvětlovanou proměnnou a vysvětlujícími proměnnými (většinou má model podobu jedné nebo více stochastických rovnic).

Úvod do analýzy časových řad

Prakticky se používají různé metody – volba použité metody závisí na účelu a cíli analýzy, typu časové řady, zkušenosti statistika, dostupném softwaru, teoretickém východisku apod.

- **expertní metody** – patří do kategorie kvalitativních metod, uplatní se tam, kde není rozumné nebo možné využívat kvantitativní metody, např. dotazování zákazníků, prodejců
- **grafická analýza** – představuje jen jednoduchou metodu analýzy časové řady, která se opírá o grafické zobrazení vývoje sledované veličiny, má subjektivní charakter (nejsnadněji lze odhadnout trend řady, užitečné bývá srovnání grafů různých časových řad mezi sebou)
- **dekompozice časových řad** – vychází z předpokladu, že hodnota sledované veličiny závisí pouze na čase ? časovou řadu rozložíme na několik nezávislých složek: trend, sezónní, cyklickou a náhodnou složku

$$Y_t = T_t + S_t + C_t + \epsilon_t$$

Úvod do analýzy časových řad

- **ekonometrické modely** – kauzální modely, které vysvětlují hodnotu vysvětlované proměnné pomocí jedné nebo více vysvětlujících proměnné. Cílem je tedy odhalit příčinné vazby mezi ekonomickými veličinami; např. při modelování inflace je vysvětlovanou proměnou cenová hladina, vysvětlujícími proměnnými mohou být reálný HDP, množství peněz v oběhu, vývoz a dovoz zboží, příjmy obyvatel.
- **Box–Jenkinsonova metodologie** – je založena na důkladném modelování náhodné složky a snaží se identifikovat vzájemnou závislost jednotlivých prvků časové řady s různým zpožděním, případně jejich závislost na různém zpoždění
- **spektrální analýza** – vychází z předpokladu, že si časovou řadu můžeme představit jako směs sinusových a kosinusových křivek s různými frekvencemi a amplitudami, a snaží se vyšetřit intenzitu zastoupení jednotlivých frekvencí; lze tak posuzovat např. zpoždění ve vývoji mezi dvěma veličinami.

Lineární dynamické modely

Příkladem jednoduchého modelu je např.

$$C_t = \alpha + \beta C_{t-1} + \gamma X_t + \delta P_t + \epsilon_t,$$

kde výdaje obyvatelstva C_t na nákup spotřebního zboží v roce t jsou vysvětlovány pomocí minulé hodnoty C_{t-1} a navíc pomocí disponibilních peněžních příjmů X_t obyvatelstva a cenového indexu P_t spotřebního zboží (α, β, γ a δ jsou parametry, ϵ_t označuje bílý šum)

Lineární dynamické modely

Uvažujme jen **jednorovnicové lineární modely** vyjádřené jedinou rovnicí ve tvaru

$$Y_t = \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_k X_{tk} + \epsilon_t, \text{ kde } t = 1, 2, \ldots, n.$$

Y_t představuje v rovnici hodnotu vysvětlované veličiny Y v čase t, $X_{t1}, \ldots, X_{tk}$ jsou hodnoty vysvětlujících veličin $X_1, \ldots, X_k$ v čase t, $\beta_1, \ldots, \beta_k$ představují neznámé parametry modelu (viz LRM). Obvykle první vysvětlující proměnná $X_1 = 1$ představuje konstantu, ϵ_t představuje chybovou (náhodnou) složku.

Lineární dynamické modely

Příklad. Při těžbě dřeva v ČR se předpokládá vliv čtyřech faktorů: zalesňování, hnojení lesních porostů, lesní požáry a škody zvěří. Na základě roční časové řady z období 1995–2004 posuďte skutečný podíl těchto faktorů a sestrojte ekonometrický model, na základě kterého by bylo možné provést odhady těžby dřeva za různých podmínek.

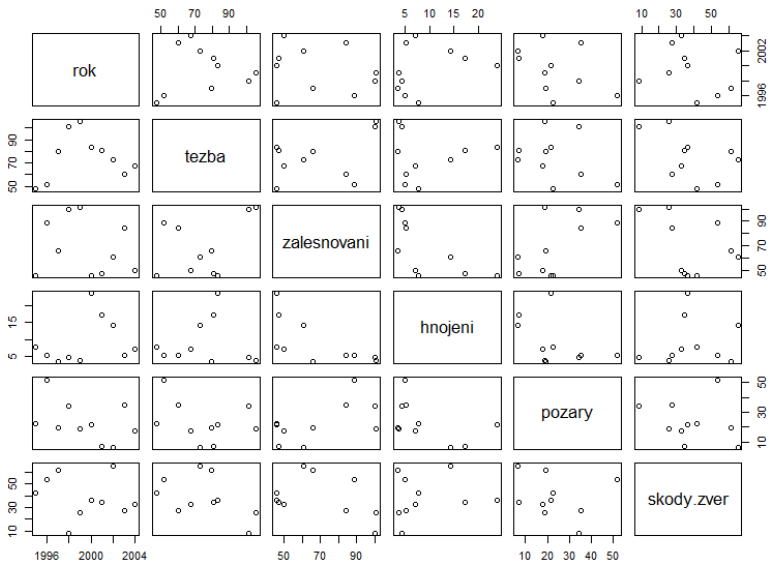
	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
Těžba (tis. m^3)	47,52	51,64	79,65	101,42	105,35	83,45	80,54	72,79	60,45	67,62
Zalesňování (ha)	46	89	66	100	101	46	47	61	84	50
Hnojení porostů (tis. ha)	7,86	5,13	3,49	4,48	3,79	23,67	17,23	14,31	5,25	7,11
Lesní požáry (des. ha)	22,7	51,9	19,5	34,2	18,9	21,5	6,8	6,6	35,1	17,7
Škody zvěří (mil. Kč)	41,8	53,8	61,1	8,2	25,8	36,4	34,5	65,3	27,4	33,0

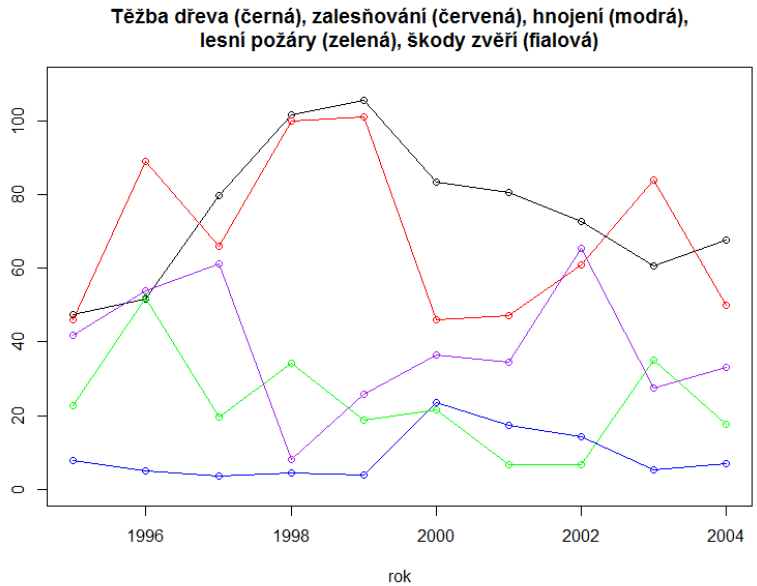
Dostáváme model

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \beta_5 X_{t5} + \epsilon_t, \text{ kde } t = 1, 2, \ldots, 10,$$

Y je těžba dřeva, X_2 zalesňování, X_3 hnojení lesních porostů, X_4 lesní požáry a X_5 jsou škody zvěří.

Lineární dynamické modely





Lineární dynamické modely

Podobně jako u LRM lze celý model zapsat v maticovém tvaru

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

kde

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Lineární dynamické modely

Odhady metodou nejmenších čtverců za předpokladu, že matice $\mathbf{X}'\mathbf{X}$ je regulární a tedy existuje inverzní matice $(\mathbf{X}'\mathbf{X})^{-1}$, jsou

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

Pozn. U LRM se předpokládá, že

(P1) Střední hodnota $E\epsilon_t = 0$, $t = 1, \ldots, n$, tj. náhodné chyby jsou **nesystematické**.

(P2) Rozptyl $D\epsilon_t = \sigma^2$, $t = 1, \ldots, n$, tj. náhodné chyby jsou **homogenní** se stejným neznámým rozptylem σ^2.

(P3) Kovariance $C(\epsilon_i, \epsilon_l) = 0$, $i \neq l$, $i, l = 1, \ldots, n$, tj. náhodné chyby jsou **nekorelované**.

Navíc se předpokládá, že hodnoty vysvětlujících proměnných nejsou náhodné, ale jsou pevně dané. Lze však ukázat, že v případě náhodnosti vysvětlujících proměnných (viz např. Hamilton[1]) je možné vlastnosti odhadů LRM zobecnit.

[1] Hamilton, J., D. *Time series analysis.* Princeton, 1994

Lineární dynamické modely

	Odhad	Sm. chyba	t-test	p-hodnota
konst.	46,7804	29,4463	1,59	0,1730
zalesnovani	0,8147	0,2876	2,83	0,0365
hnojeni	1,0182	0,8280	1,23	0,2735
pozary	−1,0373	0,3749	−2,77	0,0395
skody.zver	−0,3353	0,2605	−1,29	0,2544

	Odhad	Sm. chyba	t-test	p-hodnota
konst.	50,5821	14,8608	3,40	0,0114
zalesnovani	0,7372	0,2523	2,92	0,0223
pozary	−1,1242	0,4149	−2,71	0,0302

$$\widehat{Y}_t = 50{,}5821 + 0{,}7372 X_{2t} - 1{,}1242 X_{4t}$$

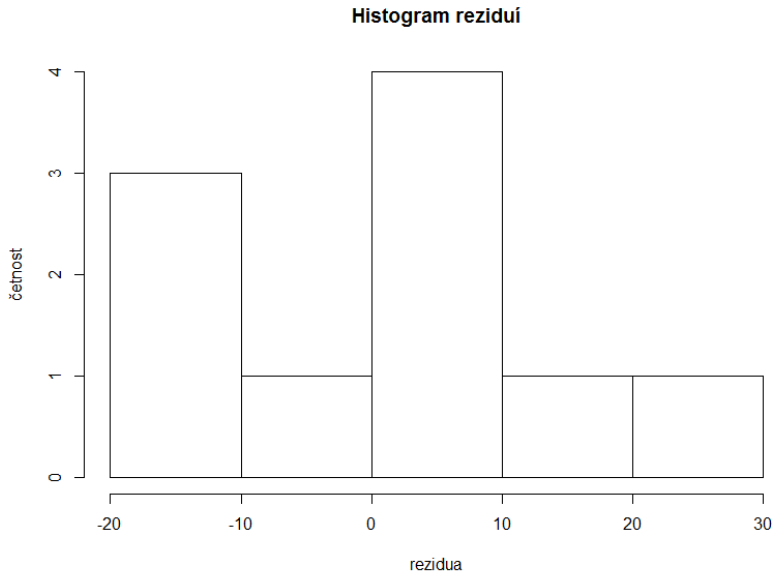
Ověřování modelu – normalita reziduí

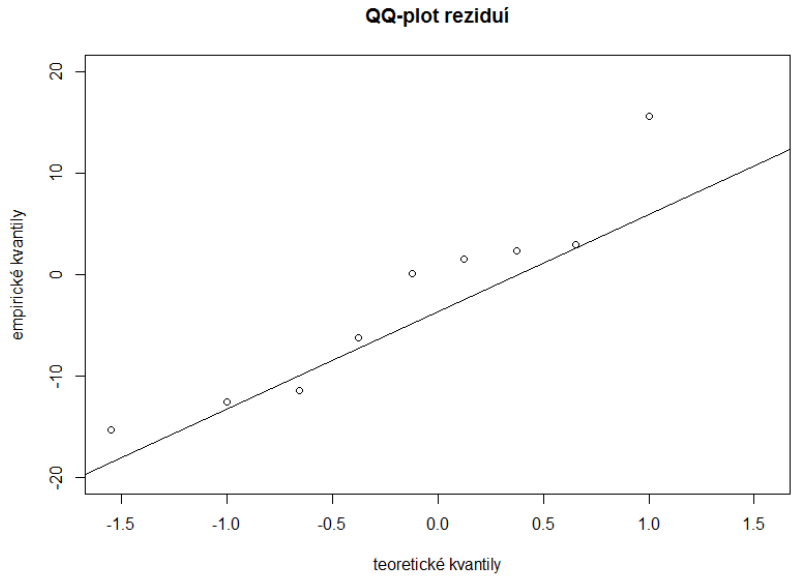
Pro test normality reziduí lze použít grafických metod jako jsou

- histogram,
- QQ plot,

nebo použít některý z testů normality: test nulové šikmosti a špičatosti, Shapiro-Wilkův test (`shapiro.test`), Lillieforsův test (`lillie.test`), Jarque-Bera test (`jarque.bera.test`) apod.

Ověřování modelu – normalita reziduí





Ověřování modelu – autokorelace reziduí

Rezidua $e_t = y_t - \widehat{y}_t$ by měla být podle předpokladů nekorelovaná. To lze ověřit např.

- Durbin-Watsonovým testem, který založený na statistice

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}.$$

  Hodnoty Durbin-Watsonovy statistiky se pohybují v intervalu $\langle 0, 4\rangle$, pokud je tato statistika rovna číslu 2, rezidua nevykazují žádnou autokorelaci, hodnoty menší než 2 značí kladnou – přímou autokorelaci a hodnoty větší než 2 značí zápornou – nepřímou autokorelaci (`dwtest`).

- pomocí autokorelační a parciální autokorelační funkce, portmanteau testu – viz později

Ověřování modelu – homoskedasticita

Homoskedasticita náhodné složky – náhodná složka modelu ϵ_t má v čase konstantní rozptyl. Pokud tomu tak není, mluvíme o **heteroskedasticitě**. Ta zpravidla také nemá vliv na odhad parametrů modelu, avšak odhady směrodatných odchylek parametrů β_j jsou už vychýlené.

Heteroskedasticitu lze ověřit vizuálně z grafu reziduí nebo testovat např. Goldfeld-Quandtovým testem (`gqtest`), Breusch-Paganovým testem (`bptest`, `ncvTest`) apod.

Ověřování modelu – multikolinearita

Pro použití metody nejmenších čtverců je důležitý předpoklad lineární nezávislosti matice $\mathbf{X}$. Jsou-li sloupce této matice lineárně závislé, potom je hodnost matice plánu $\mathbf{X}$ menší než počet odhadovaných parametrů modelu, determinant $\det(\mathbf{X}'\mathbf{X}) = 0$ a matici $\mathbf{X}'\mathbf{X}$ neexistuje matice inverzní. Hovoříme potom o **multikolinearitě** (přesné).
Problémem může být i silná korelace mezi jednotlivými vysvětlujícími proměnnými (přibližná multikolinearita). Čím je multikolinearita silnější, tím více se determinant $\det(\mathbf{X}'\mathbf{X})$ se blíží k nule.

Ověřování modelu – multikolinearita

Multikolinearita má za následek

- nadhodnocení součtu čtverců regresních koeficientů, což lze prakticky vnímat tak, že některé vysvětlující proměnné jsou důležitější, než ve skutečnosti jsou,
- zvýšení rozptylu odhadů parametrů modelu, což znamená pokles spolehlivost jejich odhadu, neboť rostou hodnoty směrodatných odchylek parametrů β_j – širší intervaly spolehlivosti resp. menší hodnoty testových kritérií pro individuální t-testy,
- zdánlivý rozpor mezi nevýznamnými výsledky t-testů a významným výsledkem celkového F-testu modelu,
- numerické problémy, které úzce souvisí s malou stabilitou odhadů některých regresních koeficientů,
- komplikace v rozumné interpretaci individuálního vlivu jednotlivých vysvětlujících proměnných na proměnnou vysvětlovanou.

Ověřování modelu – multikolinearita

Pro testování multikolinearity existuje celá řada různých kritérií. Jedno z jednoduchých kritérií vychází z párových korelačních koeficientů r_{ij}, které vyjadřují míru závislosti mezi dvěma vysvětlujícími proměnnými x_{ti} a x_{tj}, $i, j = 1, 2, \ldots, k$ a $i \neq j$. Hodnoty blízké ± 1 naznačují možnost existence multikolinearity. Vzhledem ke vzájemným vztahům jednoduchých korelačních koeficientů s **koeficientem mnohonásobné korelace** je vhodné používat pro identifikaci multikolinearity jejich kombinaci. Párové koeficienty korelace r_{ij} nemají překročit hodnotu 0,8 a žádný z nich nesmí být větší než koeficient mnohonásobné korelace.