MA012 Statistika II

8. Autokorelace a multikolinearita v lineárním modelu

Ondřej Pokora (pokora@math.muni.cz)

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

(podzim 2015)

Klasický lineární regresní model

$$\boldsymbol{Y} \sim \mathcal{L}(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{I}_n)$$

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

- nesystematické chyby:

$$\mathrm{E}\varepsilon_i = 0,\ i = 1, \ldots, n, \qquad \text{tzn. } \mathrm{E}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta},$$

- homogenita rozptylu chyb (měření):

$$\mathrm{D}\varepsilon_i = \sigma^2 > 0,\ i = 1, \ldots, n,$$

- nekorelovanost chyb (měření):

$$\mathrm{C}(\varepsilon_i, \varepsilon_j) = 0,\ i, j = 1, \ldots, n,\ i \neq j, \qquad \text{tzn. } \mathrm{D}\boldsymbol{\varepsilon} = \mathrm{D}\boldsymbol{Y} = \sigma^2 \boldsymbol{I}_n$$

Autokorelace

V některých případech (často v časových řadách) hodnoty náhodné chyby ε_i závisí i na předchozích hodnotách ε_{i-k}, $k = 1, 2, \ldots$, což má za následek, že efekt náhodných chyb není jen okamžitý, ale je pociťován i v budoucnosti. Tento případ se nazývá **autokorelace**.

Častým typem autokorelace je tzv. **autoregrese 1. řádu**, označovaná AR(1):

Definice 1 (AR(1))

$$\varepsilon_i = \theta \varepsilon_{i-1} + u_i,$$

kde θ je neznámý parametr, $|\theta| < 1$,

$$\mathrm{E}u_i = 0, \quad \mathrm{D}u_i = \sigma_u^2, \quad C(u_i, u_j) = 0 \text{ pro } i \neq j.$$

AR(1)

Postupným aplikováním AR(1) formule na náhodné chyby obdržíme vyjádření

$$\varepsilon_i = \theta\varepsilon_{i-1} + u_i = \theta(\theta\varepsilon_{i-2} + u_{i-1}) + u_i = \theta^2\varepsilon_{i-2} + \theta u_{i-1} + u_i =$$

$$= \theta^2(\theta\varepsilon_{i-3} + u_{i-1}) + \theta u_{i-1} + u_i = \theta^3 u_{i-3} + \theta^2 u_{i-2} + \theta u_{i-1} + u_i =$$

$$\cdots = \sum_{l=0}^{\infty} \theta^l u_{i-l}.$$

Počítejme střední hodnotu náhodných chyb:

$$\mathrm{E}\varepsilon_i = \mathrm{E}\left(\sum_{l=0}^{\infty} \theta^l u_{i-l}\right) = \sum_{l=0}^{\infty} \underbrace{\theta^l}_{\to 0} \underbrace{\mathrm{E}u_{i-l}}_{0} = 0.$$

Při výpočtu rozptylu náhodných chyb využijeme nekorelovanosti veličin $\boldsymbol{u}$:

$$\mathrm{D}\varepsilon_i = \mathrm{D}\left(\sum_{l=0}^{\infty} \theta^l u_{i-l}\right) = \sum_{l=0}^{\infty} \underbrace{\theta^{2\,l}}_{\to 0} \underbrace{\mathrm{D}u_{i-l}}_{\sigma_u^2} = \sigma_u^2 \underbrace{\sum_{l=0}^{\infty} (\theta^2)^l}_{\frac{1}{1-\theta^2}} = \frac{\sigma_u^2}{1-\theta^2} = \sigma^2.$$

AR(1)

Podobně spočítáme kovariance náhodných chyb (pro $l = 1, 2, \ldots$):

$$\mathrm{C}(\varepsilon_i, \varepsilon_{i-l}) = \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \theta^r \theta^s \mathrm{C}(u_{i-r}, u_{i-l-s}) = \theta^l \sigma_u^2 \sum_{r=0}^{\infty} \theta^{2r} = \frac{\theta^l \sigma_u^2}{1-\theta^2} = \theta^l \, \sigma^2.$$

Kovarianční matice náhodných chyb je tedy tvaru

$$\mathrm{D}\boldsymbol{\varepsilon} = \underbrace{\frac{\sigma_u^2}{1-\theta^2}}_{\sigma^2} \underbrace{\begin{pmatrix} 1 & \theta & \theta^2 & \ldots & \theta^{n-1} \\ \theta & 1 & \theta & \ldots & \theta^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \theta \\ \theta^{n-1} & \ldots & \theta^2 & \theta & 1 \end{pmatrix}}_{\boldsymbol{W}} = \sigma^2 \, \boldsymbol{W}.$$

Rozšířený lineární model

Máme tedy lineární regresní model model $\boldsymbol{Y} \sim \mathcal{L}(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{W})$:

Definice 2 (Rozšířený lineární model)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}\boldsymbol{\varepsilon} = \boldsymbol{0}, \quad \mathrm{D}\boldsymbol{\varepsilon} = \sigma^2 \boldsymbol{W},$$

kde kovariační matice již není diagonální.

Věta 3 (Aitkenův odhad, zobecněná metoda nejmenších čtverců)

Mějme regresní model $\boldsymbol{Y} \sim \mathcal{L}(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{W})$ *plné hodnosti, kde* $|\boldsymbol{W}| > 0$.
Pak odhad vektoru parametrů zobecněnou metodou nejmenších čtverců je roven

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

Z věty tedy plyne, že pokud známe parametr θ, dokážeme najít odhady $\widehat{\boldsymbol{\beta}}$
i v linárním modelu s autokorelovanými náhodnými chybami.

Detekce autokorelace

Graficky je možno autokorelaci detekovat tak, že vynášíme hodnoty reziduí r_i klasického lineárního modelu v závislosti na r_i. Je-li z grafu patrná přibližná lineární závislost, svědčí to o autoregresi 1. řádu (AR(1)) nebo o špatné volbě modelu.

Věta 4 (Asymptotický test)

Statistika U má asymptoticky standardizované normální rozdělení pravděpodobnosti,

$$U = \frac{\widehat{\theta} - \theta}{\sqrt{\frac{1-\theta^2}{n}}} \overset{as.}{\sim} N(0,1).$$

Při testování hypotézy $H_0 : \theta = 0$ proti $H_1 : \theta \neq 0$ tedy pro dostatečně velká n $(n \geq 30)$ platí

$$U = \widehat{\theta}\sqrt{n} \overset{as.}{\sim} N(0,1),$$

a H_0 zamítáme na hladině významnosti α, pokud

$$|\widehat{\theta}\sqrt{n}| > u_{1-\alpha/2}.$$

Durbinův-Watsonův test

Definice 5 (Durbinova-Watsonova statistika)

Durbinova-Watsonova statistika je definována pomocí reziduí v lineárním modelu:

$$D = \frac{\sum_{i=2}^{n} (r_i - r_{i-1})^2}{\sum_{i=1}^{n} r_i^2}.$$

Věta 6 (Durbinův-Watsonův test)

Vždy platí $0 \leq D \leq 4$.

Při nekorelovanosti náhodných chyb má statistika hodnotu $D = 2$.
Kladná korelace chyb se projeví hodnotou $D \to 0$,
záporná korelace chyb se projeví hodnotou $D \to 4$.
Hodnoty kritických hodnot pro Durbinův-Watsonův test závisí na počtu pozorování, počtu parametrů a na hladině významnosti a jsou tabelovány.

Odhad parametru θ

Parametr θ v AR(1) modelu náhodných chyb lze odhadnout více způsoby:

- Odhadujeme jako regresní koeficient v modelu pro rezidua

$$r_i = \theta\, r_{i-1} + u_i, \quad i = 2, \ldots, n$$

metodou nejmenších čtverců. Odtud pak

$$\widehat{\theta} = \frac{\sum_{i=2}^{n} r_i\, r_{i-1}}{\sum_{i=2}^{n} r_{i-1}^2}.$$

- Pomocí Durbin – Watsonovy statistiky:

$$\widehat{\theta} = 1 - \frac{D}{2}.$$

Model bez autokorelace 1. řádu

1. Nalezneme odhad $\widehat{\theta}$.
2. Vytvoříme nový model

$$Y_i^* = Y_{i+1} - \widehat{\theta}\, Y_i, \qquad x_{i,j}^* = x_{i+1,j} - \widehat{\theta}\, x_{i,j}, \qquad i = 1, \ldots, n-1,\ j = 1, \ldots, k,$$

tj. vznikne model

$$\boldsymbol{Y}^* = \boldsymbol{X}^* \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*, \qquad \mathrm{E}\boldsymbol{\varepsilon}^* = \boldsymbol{0}, \qquad \mathrm{D}\boldsymbol{\varepsilon}^* = \sigma_{\varepsilon^*}^2\, \boldsymbol{I}_n$$

v němž počítáme odhady $\widehat{\boldsymbol{\beta}}^*$ standardním způsobem.

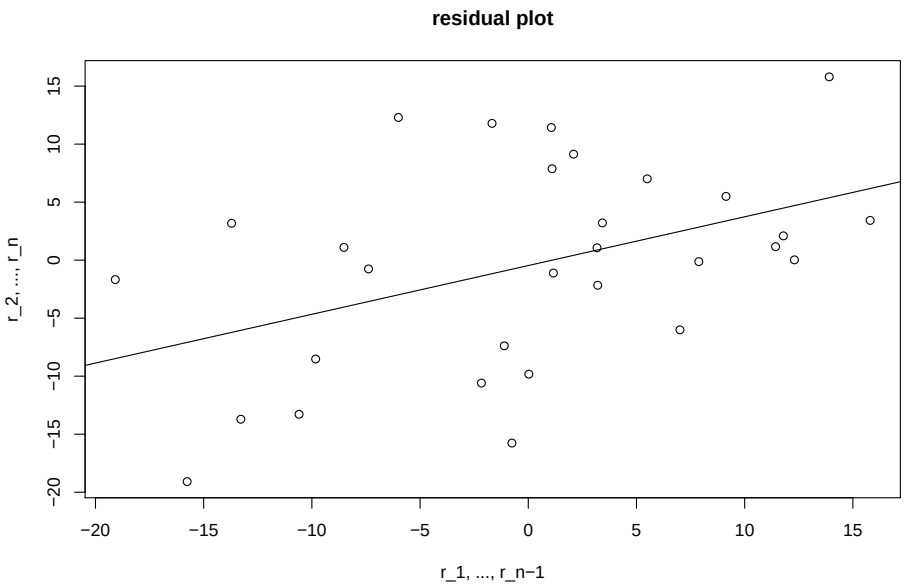
Příklad

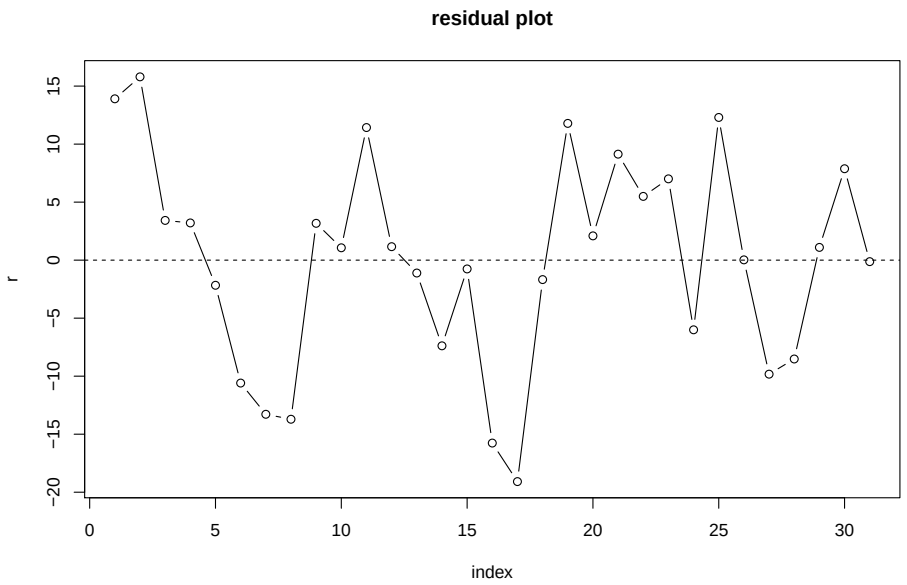
Příklad 1

V letech 1953 – 1983 byly měřeny ztráty vody při distribuci do domácností. Výsledky měření jsou uloženy v souboru `voda.csv`*. Proměnná x označuje množství vyrobené vody, proměnná Y ztrátu. Ověřte, zda se v datech vyskytuje autokorelace 1. řádu a případně ji odstraňte.*

Řešení

Graficky: je patrná lineární závislost mezi rezidui





Řešení

- **Asymptotický test:**

$$U = |\widehat{\theta}\sqrt{n}| = 2{,}339.$$

  Nulovou hypotézu tedy **zamítáme**, neboť $|U| > u_{1-\alpha/2} = 1{,}96$.

- **Durbinův-Watsonův test:**

$$D = 1{,}082$$

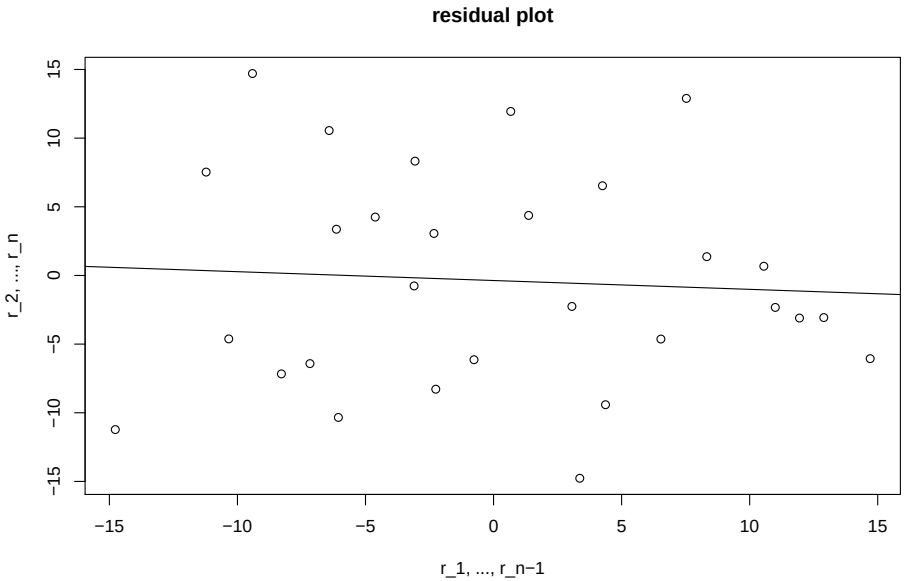
  a p-hodnota testu je $0{,}0016$, takže také **zamítáme** nulovou hypotézu.

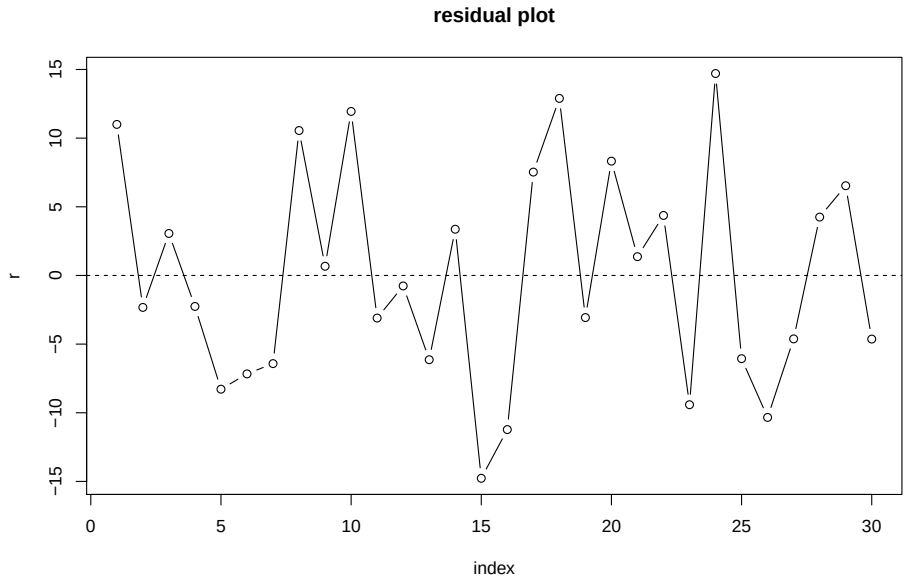
Řešení

Odstranění autokorelace:

Odhad metodou nejmenších čtverců: $\widehat{\theta} = 0{,}42$, z D-W statistiky: $\widehat{\theta} = 0{,}459$.

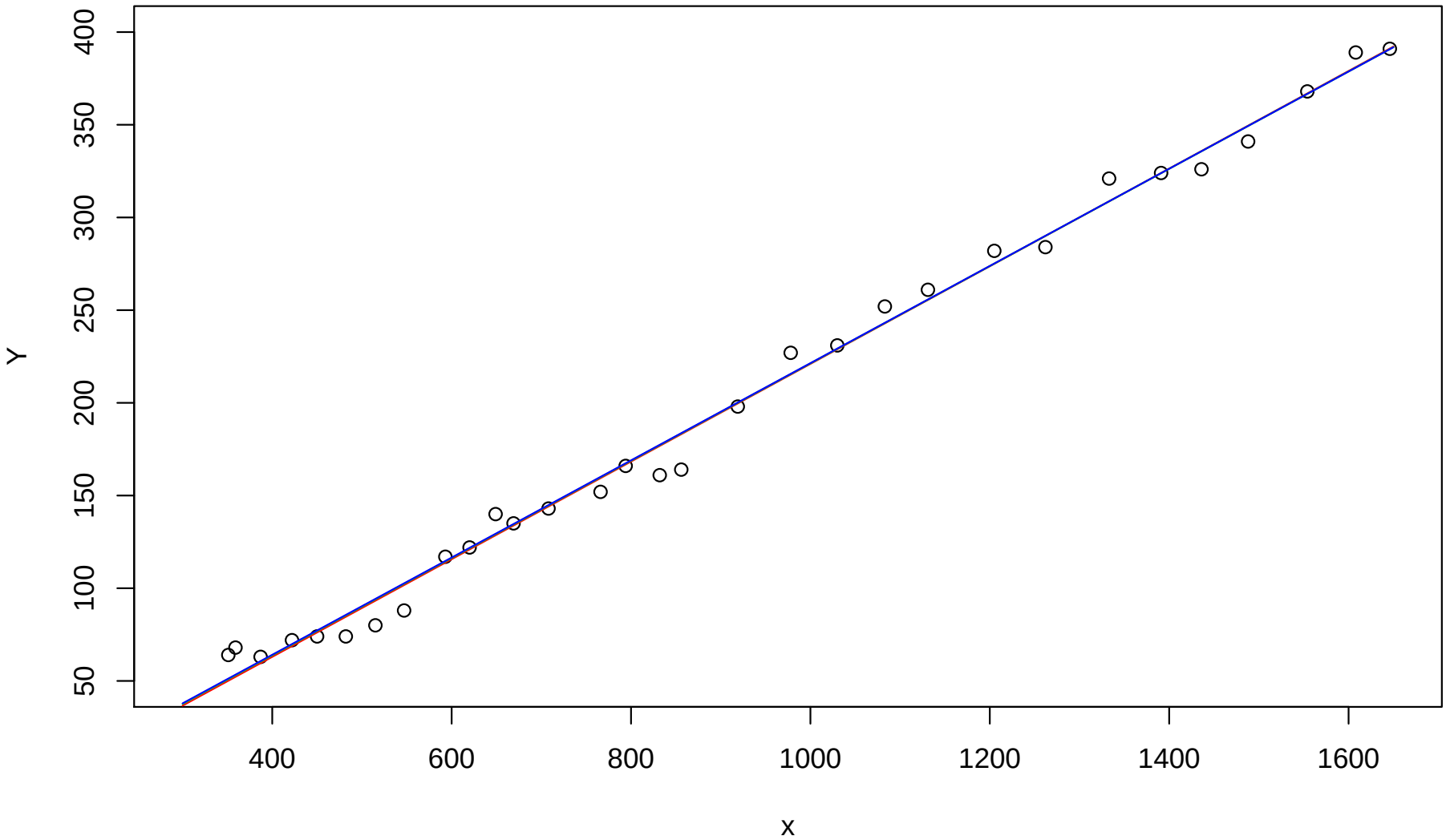
V nově vzniklém modelu vykreslíme rezidua:



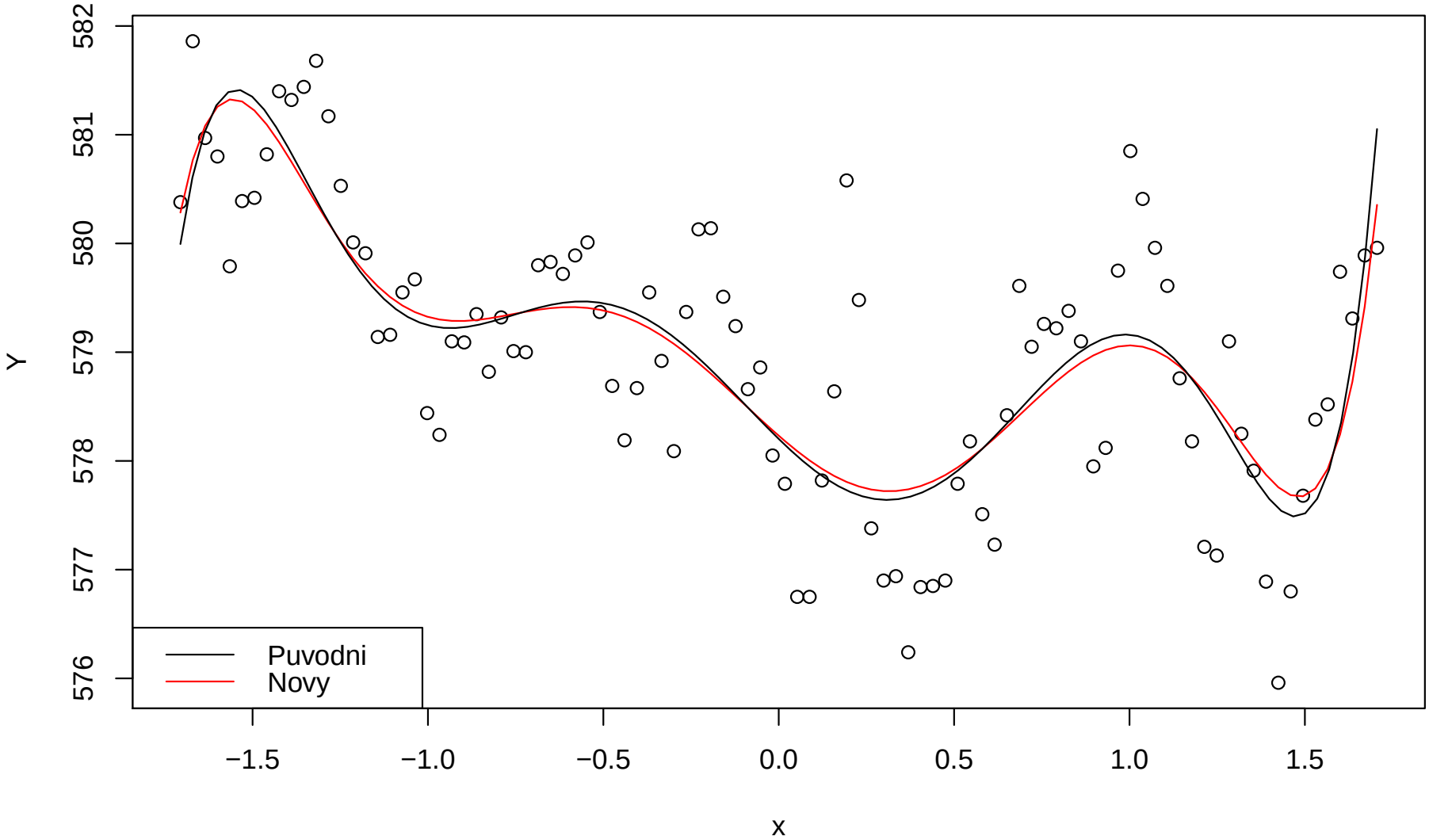


Také D-W test již **nezamítá** nulovou hypotézu (p-hodnota je 0,4).

Řešení


Y
50
100
150
200
250
300
350
400
400
600
800
1000
1200
1400
1600
x

Příklad LakeHuron: řešení



Multikolinearita

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

$$c_1\boldsymbol{x}_1 + \cdots + c_p\boldsymbol{x}_p = \boldsymbol{0}.$$

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 $\boldsymbol{X}$ je ortogonální a platí, že $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_p$. V praxi se tedy **multikolinearitou** rozumí případ, kdy **přibližně** platí rovnice vyjadřující lineární kombinaci vysvětlujících proměnných. V případě silné multikolinearity je determinant informační matice $\boldsymbol{X}'\boldsymbol{X}$ blízký nule, nejmenší vlastní číslo je rovněž blízké nule a matice $\boldsymbol{X}'\boldsymbol{X}$ 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 $\boldsymbol{X}'\boldsymbol{X}$ singulární a běžnou inverzí nepořídíme odhad neznámých parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců. Pro přibližnou multikolinearitu jsme sice schopni matici $\boldsymbol{X}'\boldsymbol{X}$ invertovat, ale kvalita pořízených odhadů je poměrně nízká.
Snížení kvality se projeví

- v kovarianční matici $\mathrm{D}\widehat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$
- 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 Inflation Factors

Definice 7 (VIF = variance inflation factors)

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

$$(a_1, \ldots, a_p) = \operatorname{diag}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

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

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

kde $r_j = r_{x_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p}$ je koeficient mnohonásobné korelace mezi j-tou proměnnou a všemi ostatními proměnnými (kromě j).

Vysoký stupeň multikolinearity se projevuje vysokými hodnotami korelačních koeficientů r_j, ale i vysokými hodnotami některých jednoduchých korelačních koeficientů.

Detekce multikolinearity

Test multikolinearity je založen na testování, že korelační matice nezávisle proměnných je jednotková,

$$H_0 : \boldsymbol{R} = \boldsymbol{I}_p, \qquad H_1 : \boldsymbol{R} \neq \boldsymbol{I}_p,$$

kde $\boldsymbol{R}$ je korelační matice nezávisle proměnných rozměru $p \times p$.

Věta 8 (Test multikolinearity)

Za platnosti H_0 má statistika

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

χ^2 rozdělení pravděpodobnosti.
Hypotézu H_0 tedy na hladině významnosti α zamítáme, pokud

$$K > \chi^2_{1-\alpha}\left(\frac{1}{2}p(p-1)\right).$$

Multikolinearita: identifikace proměnných

Pro identifikaci proměnných způsobujících multikolinearitu se doporučují statistiky

$$F_j = \frac{n-p}{p-1}\,(d_j - 1),$$

kde d_k jsou diagonální prvky inverzní matice ke korelační matici nezávisle proměnných, $(d_1, \ldots, d_p) = \operatorname{diag} \boldsymbol{R}^{-1}$.

Věta 9

V případě, že proměnná X_j nezpůsobuje multikolinearitu, má veličina F_j Fisherovo-Snedecorovo rozdělení,

$$F_j \sim F(p-1,\ n-p).$$

Pokud tedy $F_j > F_{1-\alpha}(p-1,\ n-p)$, přispívá j-tá proměnná k multikolinearitě.

Metoda postupné regrese

Následující algortimus stavby lineárního modelu bez multikolinearity je založen na tom, že do modelu postupně zařazujeme jen ty regresory (nezávisle proměnné), které významně přispívají ke zlepšení kvality odhadu $\widehat{\boldsymbol{\beta}}$. K výběru nejlepší podmnožiny regresorů používá tzv. **metodu postupné regrese**:

1. Spočteme korelační matici $\boldsymbol{R}$ regresorů a provedeme test hypotézy $H_0 : \boldsymbol{R} = \boldsymbol{I}_p$. Pokud test H_0 nezamítne, pracujeme s klasickým lineárním modelem, do něhož můžeme zařadit všechny regresory. Je-li korelace mezi regresory prokázána, pokračujeme dalším krokem.
2. Spočteme korelační koeficienty $r_{Y,X_1}, \ldots, r_{Y,X_k}$ a pro další krok vybereme regresor X_j s největší absolutní hodnotou korelačního koeficientu $|r_{Y,X_j}|$.
3. Sestavíme model $Y = \beta_0 + \beta_1 X_j$ a odhadneme jeho parametry metodou nejmenších čtverců. Vypočteme hodnotu statistiky F-testu modelu, $F = (n-2)\dfrac{R^2_{Y\cdot X_j}}{1-R^2_{Y\cdot X_j}}$, kde $R^2_{Y\cdot X_j}$ je koeficient determinace tohoto modelu.
   Pokud $F > F_{1-\alpha}(1,\ n-2)$, ponecháme regresor X_j v modelu.

Metoda postupné regrese

4. Spočteme parciální korelační koeficienty $r_{Y\,X_1\cdot X_j}$, …, $r_{Y\,X_{j-1}\cdot X_j}$, $r_{Y\,X_{j+1}\cdot X_j}$, …, $r_{Y\,X_p\cdot X_j}$ a pro další krok vybereme ten regresor X_i, jehož $|r_{Y\,X_i\cdot X_j}|$ je největší.
5. Sestavíme model $Y = \beta_0 + \beta_1 X_j + \beta_2 X_i$ a odhadneme jeho parametry metodou nejmenších čtverců. Vypočteme hodnotu statistiky $F = (n-3)\dfrac{R^2_{Y\cdot X_j X_i} - R^2_{Y\cdot X_j}}{1 - R^2_{Y\cdot X_j X_i}}$, kde ve jmenovateli je koeficient determinace nového modelu a v čitateli rozdíl koeficientů determinace nového a předchozího modelu. Připomeňme, že předchozí model je podmodelem nového modelu. Pokud $F > F_{1-\alpha}(2,\ n-3)$, ponecháme regresor X_i v modelu.
6. Spočteme parciální korelační koeficienty $r_{Y\,X_1\cdot X_j X_i}$, …, $r_{Y\,X_p\cdot X_j X_i}$ mezi Y a dosud nezařazenými regresory za vyloučení vlivu regresorů do modelu již zařazených. Ze skupiny dosud nezařazených regresorů vybereme regresor s největší absolutní hodnotou parciálního korelačního koeficientu, sestavíme s ním nový model a analogicky opakujeme kroky 5–6, dokud lze zařazovat další regresory.

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 $\boldsymbol{X}'\boldsymbol{X}$ korelační matici $\boldsymbol{R} = \boldsymbol{Z}'\boldsymbol{Z}/n$. Vektor $\boldsymbol{Z}'\boldsymbol{q}/n$ obsahuje jednoduché korelační koeficienty $r_{Y X_j}$. Standardizací proměnných se zmenšují zaokrouhlovací chyby a zlepšují se možnosti hodnocení individuálního vlivu proměnných pomocí regresních parametrů.

Model v kanonickém tvaru

Místo modelu ve tvaru

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

pracujeme s modelem

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

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

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

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

Hřebenová regrese (ridge regression)

Autoři tzv. **hřebenové regrese** vyšli z faktu, že vliv multikolinearity se silně projevuje při výpočtu inverzní matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, a to konkrétně tím, že diagonální prvky této inverzní matice jsou příliš velké.
Navrhli proto jiný odhad

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + m\,\boldsymbol{I}_k)^{-1}\boldsymbol{X}'\boldsymbol{Y},$$

závisející na parametru $m > 0$. Výhodou takového odhadu je především to, že vhodnou volbou m lze numerické vlastnosti potřebné pro výpočet inverní matice výrazně zlepšit a získat tak relativně přesné hodnoty. Nevýhodou je to, že tento odhad není nestranný a že neexistuje jednotný algoritmický postup pro určení vhodné hodnoty parametru m. Obvykle se vychází z modelu v kanonickém tvaru a postupně se zkouší různé volby hodnot m, dokud se nedosáhne „stabilizace“, kdy model dostává charakter ortogonálního systému.

Příklad

Příklad 2

V souboru `vydaje.csv` *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.*

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

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

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

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

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 $F = \frac{(n-2)R^2_{Y \cdot X_1}}{1-R^2_{Y \cdot X_1}} = \frac{18 \cdot 0,5897}{1-0,5897} = 25,87$. 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 $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3$. Vypočteme hodnotu statistiky $F = \frac{(n-3)(R^2_{Y\cdot X_1} - R^2_{Y\cdot X_1 X_3})}{1-R^2_{Y\cdot X_1 X_3}} = \frac{17\cdot 0{,}102}{1-0{,}69} = 5{,}64$. 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 $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_2$. Vypočteme hodnotu statistiky $F = \frac{(n-3)(R^2_{Y\cdot X_1 X_3} - R^2_{Y\cdot X_1 X_3 X_2})}{1-R^2_{Y\cdot X_1 X_3 X_2}} = \frac{16\cdot 0{,}012}{1-0{,}704} = 0{,}63$. 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

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3.$$