Pořadové testy v regresi při rušivé heteroskedasticitě

Radim Navrátil, Jana Jurečková

Katedra pravděpodobnosti a matematické statistiky, MFF UK, Praha

Robust 2014, Jetřichovice
22.1.2014

Úvod

Homoskedasticita je často předpokládána při klasické analýze lineárního modelu.

Heteroskedasticita:

- Můžeme testovat její přítomnost před statistickou inferencí.
- Najít přístup, který je invariantní vůči heteroskedasticitě.

Cíl:

- Testy o regresi za přítomnosti rušivé heteroskedasticity.
- Testy homoskedasticity za přítomnosti rušivé regrese.

Oba typy testů jsou založeny na vhodných ancilárních statistikách, není tedy nutné odhadovat rušivé parametry modelu.

Historie

- Lineární model s možnou heteroskedasticitou

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \sigma_i U_i, \quad i = 1, \ldots, n.$$

- Požadavky na znalost struktury σ_i – mnoho modelů.
- Testy heteroskedasticity: Breusch and Pagan (1979), Carroll and Ruppert (1981), Koenker and Bassett (1982), Dette and Munk (1999).
- Pořadové testy: Akritas and Albers (1993), Gutenbrunner (1994).
- Odhady parametrů: Dixon and McKean (1996).

Model

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \exp\{\mathbf{z}_i^\top \boldsymbol{\gamma}\} U_i, \quad i = 1, \ldots, n,$$
$$\beta_0 \in \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{\gamma} \in \mathbb{R}^q,$$

- $\mathbf{x}_i$ je p-rozměrný vektor regresorů.
- $\mathbf{z}_i$ je q-rozměrný vektor regresorů.
- U_i jsou i.i.d. náhodné veličiny s distribuční funkcí F, resp. absolutně spojitou hustotou f a konečnými nenulovými Fisherovými informacemi vzhledem k posunutí i měřítku.

$$\mathbf{H}_1 : \boldsymbol{\gamma} = \mathbf{0},\ \text{proti alternativě} \quad \mathbf{K}_1 : \boldsymbol{\gamma} \neq \mathbf{0},$$
$$\mathbf{H}_2 : \boldsymbol{\beta} = \mathbf{0},\ \text{proti alternativě} \quad \mathbf{K}_2 : \boldsymbol{\beta} \neq \mathbf{0}.$$

Test $\mathbf{H}_1$

Za $\mathbf{H}_1 : \gamma = \mathbf{0}$ máme model

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + U_i,\ i = 1, \ldots, n.$$

Zkonstruujeme vektor regresních pořadových skórů $\widehat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n,1}(\alpha), \ldots, \hat{a}_{n,n}(\alpha))^\top$, optimální řešení úlohy lineárního programování:

$$\widehat{\mathbf{a}}_n(\alpha) = \arg\max\{\mathbf{Y}_n^\top \mathbf{a} \mid \mathbf{X}_n^{*\top} \mathbf{a} = (1-\alpha)\mathbf{X}_n^{*\top} \mathbf{1}_n, \mathbf{a} \in [0,1]^n\}, \quad 0 < \alpha < 1,$$

kde $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top, \mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ a
$\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top, \quad \mathbf{X}_n^* = (\mathbf{1}_n, \mathbf{X}_n)$.

Regresní kvantily

Koenker and Bassett (1978) zobecnili pojem kvantilu pro lineární regresní model, α-regresní kvantil:

$$\widehat{\boldsymbol{\beta}}_n(\alpha) = \operatorname{argmin}\left\{\sum_{i=1}^{n} \rho_\alpha\left(Y_i - \mathbf{x}_i^{*\top}\mathbf{t}\right), \mathbf{t} \in \mathbb{R}^{p+1}\right\},$$

kde $\rho_\alpha(x) = |x| \cdot (\alpha\mathbb{I}\{x > 0\} + (1-\alpha)\mathbb{I}\{x < 0\})$.
Koenker and Bassett (1978): $\widehat{\boldsymbol{\beta}}_n(\alpha)$ lze počítat jako komponentu $\boldsymbol{\beta}$ optimálního řešení $(\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-)$ úlohy lineárního programování

$$\min\ \alpha\mathbf{1}_n^\top\mathbf{r}^+ + (1-\alpha)\mathbf{1}_n^\top\mathbf{r}^-$$

vzhledem k

$$\begin{aligned} &\mathbf{X}_n^*\boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{Y}_n \\ &(\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}^{p+1} \times \mathbb{R}_+^{2n}, \end{aligned}$$

kde $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top, \mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ a
$\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top, \quad \mathbf{X}_n^* = (\mathbf{1}_n, \mathbf{X}_n)$.

Regresní pořadové skóry

Regresní kvantily – úloha lineárního programování:

$$\min \ \alpha \mathbf{1}_n^\top \mathbf{r}^+ + (1-\alpha)\mathbf{1}_n^\top \mathbf{r}^-$$

vzhledem k

$$\begin{aligned} \mathbf{X}_n^* \boldsymbol{\beta} + \mathbf{r}^+ - \mathbf{r}^- &= \mathbf{Y}_n \\ (\boldsymbol{\beta}, \mathbf{r}^+, \mathbf{r}^-) &\in \mathbb{R}^{p+1} \times \mathbb{R}_+^{2n}. \end{aligned}$$

Vektor regresních pořadových skórů $\widehat{\mathbf{a}}_n(\alpha)$ je optimálním řešením duální úlohy:

$$\widehat{\mathbf{a}}_n(\alpha) = \arg\max\{\mathbf{Y}_n^\top \mathbf{a} \mid \mathbf{X}_n^{*\top}\mathbf{a} = (1-\alpha)\mathbf{X}_n^{*\top}\mathbf{1}_n, \mathbf{a} \in [0,1]^n\}, \quad 0 < \alpha < 1.$$

Klíčová vlastnost: $\widehat{\mathbf{a}}_n(\alpha)$ je invariantní vůči regresi $\mathbf{X}_n^*$, t.j. $\widehat{\mathbf{a}}_n(\alpha)$ se nezmění pokud místo $\mathbf{Y}_n$ pozorujeme $\mathbf{Y}_n + \mathbf{X}_n^*\boldsymbol{\delta}$ pro všechna $\boldsymbol{\delta} \in \mathbb{R}^{p+1}$.

Testová statistika

- Zvolíme funkci $\varphi : (0,1) \mapsto \mathbb{R}$, integrovatelnou se čtvercem, "tvaru U".
- $$\hat{b}_{n,i} = -\int_0^1 (\varphi(t) - \bar{\varphi}) d\hat{a}_{n,i}(t).$$
- $$\mathbf{S}_n = n^{-1/2} \sum_{i=1}^{n} (\mathbf{z}_i - \widehat{\mathbf{z}}_i)\hat{b}_{n,i}, \quad \mathcal{T}_n^2 = \frac{1}{A^2(\varphi)}\, \mathbf{S}_n^\top \widehat{\mathbf{D}}_n^{-1} \mathbf{S}_n,$$
- $$A^2(\varphi) = \int_0^1 (\varphi(t) - \overline{\varphi})^2 dt, \qquad \overline{\varphi} = \int_0^1 \varphi(t) dt.$$
- $$\widehat{\mathbf{D}}_n = n^{-1}(\mathbf{Z}_n - \widehat{\mathbf{Z}}_n)^\top (\mathbf{Z}_n - \widehat{\mathbf{Z}}_n)$$

  a $\widehat{\mathbf{Z}}_n = \mathbf{X}_n^*(\mathbf{X}_n^{*\top}\mathbf{X}_n^*)^{-1}\mathbf{X}_n^{*\top}\mathbf{Z}_n$ je projekce $\mathbf{Z}_n$ na prostor sloupců matice $\mathbf{X}_n^*$, resp. $\widehat{\mathbf{z}}_i^\top$ je i-tý řádek matice $\widehat{\mathbf{Z}}_n$.

Předpoklady

(F.1) $\left|\frac{f'(x)}{f(x)}\right| \leq c|x|$, pro $|x| \geq K \geq 0$, $c > 0$.

(F.2) $f(x) > 0$ na (A, B) a absolutně spojitá, omezená a klesající pro $x \to A+$ a $x \to B-$, kde

$$-\infty \leq A = \sup\{x : \; F(x) = 0\},$$

$$\infty \geq B = \inf\{x : \; F(x) = 1\},$$

$$\sup_{0<u<1} u(1-u)\frac{|f'(F^{-1}(u))|}{f^2(F^{-1}(u))} = \alpha$$

pro $1 \leq \alpha \leq 1 + \frac{1}{4} - \varepsilon$, $\varepsilon > 0$.

Nechť existují pozitivně definitní matice $\widehat{\mathbf{D}}$, $\mathbf{M}$ takové, že

$$\lim_{n\to\infty} \widehat{\mathbf{D}}_n = \widehat{\mathbf{D}}, \; \lim_{n\to\infty} n^{-1}\mathbf{X}_n^{*\top}\mathbf{X}_n^* = \mathbf{M}.$$

$$\max_{1\leq i\leq n} \|\mathbf{x}_i^*\| = o\left(n^{\frac{1}{4}-\eta}\right) \text{ pro nějaké } \eta > 0, \quad \text{pro } \; n \to \infty.$$

Asymptotické rozdělení

Věta

Za výše uvedených předpokladů $\mathcal{T}_n^2$ má za hypotézy $\mathbf{H}_1$ asymptoticky χ^2 rozdělení o q stupních volnosti a za lokální alternativy

$$\mathbf{K}_{1,n}: \; \boldsymbol{\gamma} = \boldsymbol{\gamma}_n = n^{-1/2}\boldsymbol{\gamma}^*, \quad \boldsymbol{\gamma}^* \in \mathbb{R}^q \textit{ pevné}$$

má asymptoticky χ^2 rozdělení o q stupních volnosti a parametrem necentrality

$$\eta^2 = \frac{\tau_1^2(\varphi, f)}{A^2(\varphi)} \cdot \boldsymbol{\gamma}^{*\top}\widehat{\mathbf{D}}\boldsymbol{\gamma}^*,$$

$$\tau_1(\varphi, f) = \int_0^1 \varphi(t)\left(-1 - F^{-1}(t)\,\frac{f'(F^{-1}(t))}{f(F^{-1}(t))}\right) dt.$$

Poznámky

- Asymptotické rozdělení za $\mathbf{H}_1$ nezávisí na rozdělení chyb modelu.
- Asymptotické rozdělení za $\mathbf{K}_1$ nezávisí na hodnotě rušivého parametru $\boldsymbol{\beta}$.
- Asymptotická síla testu je stejná jako u klasického pořadového testu, kde parametr $\boldsymbol{\beta}$ je známý.
- Test nerozliší alternativy s chybami $\exp\{\mathbf{x}_i^\top \gamma\} U_i$ kvůli ancilaritě regresních pořadových skórů.

Test $\mathbf{H}_2$

Model:

$$Y_i = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \exp\{\mathbf{z}_i^\top \gamma\} U_i, \quad i = 1, \ldots, n.$$

Nechť F splňuje všechny předpoklady z minulého odstavce a navíc je symetrická a $\beta_0 = 0$.
Za platnosti $\mathbf{H}_2 : \boldsymbol{\beta} = \mathbf{0}$ můžeme psát:

$$W_i = \mathbf{z}_i^\top \gamma + V_i, \ i = 1, \ldots, n, \tag{1}$$

kde $W_i = \ln|Y_i|, \ V_i = \ln|U_i|, \ i = 1, \ldots, n.$

Test $\mathbf{H}_2$

Zkonstruujeme vektor regresních pořadových skórů $\widehat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n,1}(\alpha), \ldots, \hat{a}_{n,n}(\alpha))^\top$ v modelu

$$W_i = \mathbf{z}_i^\top \gamma + V_i,\ i = 1, \ldots, n,$$

t.j. optimální řešení úlohy lineárního programování:

$$\widehat{\mathbf{a}}_n(\alpha) = \arg\max\{\mathbf{W}_n^\top \mathbf{a} \mid \mathbf{Z}_n^\top \mathbf{a} = (1-\alpha)\mathbf{Z}_n^\top \mathbf{1}_n, \mathbf{a} \in [0,1]^n\}, \quad 0 < \alpha < 1,$$

kde $\mathbf{W}_n = (W_1, \ldots, W_n)^\top, \mathbf{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ a $\mathbf{Z}_n = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$.

Předpoklady

- Nechť F je symetrická s konečnými nenulovými Fisherovými informacemi vzhledem k posunutí i měřítku a splňuje (F.1) a (F.2).
- $z_{i,1} = 1,\ i = 1, \ldots, n$ a $\max_{1 \le i \le n} \|\mathbf{z}_i\| = \mathcal{O}(1)$.
- $\widetilde{\mathbf{D}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^\top \to \widetilde{\mathbf{D}}$ kde $\widetilde{\mathbf{D}}$ je pozitivně definitní.
- $\widetilde{\mathbf{Q}}_n(\gamma) = n^{-1} \sum_{i=1}^{n} \mathrm{e}^{-\mathbf{z}_i^\top \gamma} (\mathbf{x}_i - \widehat{\mathbf{x}}_i)(\mathbf{x}_i - \widehat{\mathbf{x}}_i)^\top \to \widetilde{\mathbf{Q}}(\gamma), \quad \forall \gamma \in \mathbb{R}^q$, kde $\widetilde{\mathbf{Q}}(\gamma)$ je pozitivně definitní a $\widehat{\mathbf{x}}_i^\top$ je i-tý řádek $\widehat{\mathbf{X}}_n = \mathbf{Z}_n (\mathbf{Z}_n^\top \mathbf{Z}_n)^{-1} \mathbf{Z}_n^\top \mathbf{X}_n$. Dále pro jednoduchost označme $\widetilde{\mathbf{Q}}_n(\mathbf{0}) \equiv \widetilde{\mathbf{Q}}_n$ and $\widetilde{\mathbf{Q}}(\mathbf{0}) \equiv \widetilde{\mathbf{Q}}$.

Testová statistika

- Zvolíme funkci $\varphi : (0,1) \mapsto \mathbb{R}$, integrovatelnou se čtvercem, neklesající a nekonstantní, pro jejíž derivaci platí, že pro všechna $0 < u < \alpha_0,\ 1 - \alpha_0 < u < 1$ je
$$|\varphi'(u)| \leq c(u(1-u))^{-1-\delta} \quad \text{pro nějaké} \ \ \delta > 0.$$
- $$\hat{b}^+_{n,i} = -\int_0^1 (\varphi^+(t) - \bar{\varphi}^+) \, d\hat{a}_{n,i}(t), \quad \varphi^+(u) = \varphi\left(\frac{u+1}{2}\right),\ 0 < u < 1.$$
- $$\widetilde{\mathbf{S}}_n = n^{-1/2} \sum_{i=1}^{n} (\mathbf{x}_i - \widehat{\mathbf{x}}_i) \ \mathrm{sign} \ Y_i \ \hat{b}^+_{n,i},$$
- $$\widetilde{\mathcal{T}}^2_n = \frac{1}{A^2(\varphi)} \widetilde{\mathbf{S}}_n^\top \widetilde{\mathbf{Q}}_n^{-1} \widetilde{\mathbf{S}}_n$$

Asymptotické rozdělení

Věta

Za výše uvedených předpokladů $\widetilde{\mathcal{T}}_n^2$ má za hypotézy $\mathbf{H}_2$ asymptoticky χ^2 rozdělení o p stupních volnosti a za lokální alternativy

$$\mathbf{K}_{2,n}: \ \boldsymbol{\beta} = \boldsymbol{\beta}_n = n^{-1/2}\boldsymbol{\beta}^*, \quad \boldsymbol{\beta}^* \in \mathbb{R}^p \text{ pevné}$$

má asymptoticky χ^2 rozdělení o p stupních volnosti a parametrem necentrality

$$\tilde{\eta}^2(\boldsymbol{\gamma}) = \frac{\tau_1^2(\varphi, f)}{A^2(\varphi)} \cdot \boldsymbol{\beta}^{*\top}\widetilde{\mathbf{Q}}^\top(\boldsymbol{\gamma})\widetilde{\mathbf{Q}}^{-1}\widetilde{\mathbf{Q}}(\boldsymbol{\gamma})\boldsymbol{\beta}^*.$$

Poznámky

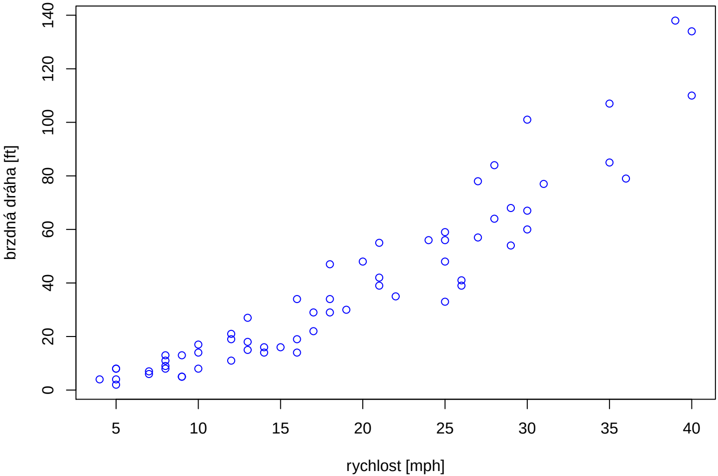
- Asymptotické rozdělení za $\mathbf{H}_2$ nezávisí na rozdělení chyb modelu.
- Asymptotické rozdělení za $\mathbf{K}_2$ závisí na hodnotě rušivého parametru γ.
- Asymptotická síla testu je stejná jako u klasického pořadového testu, kde parametr γ je známý.
- Předpoklad symetrie.

Příklad

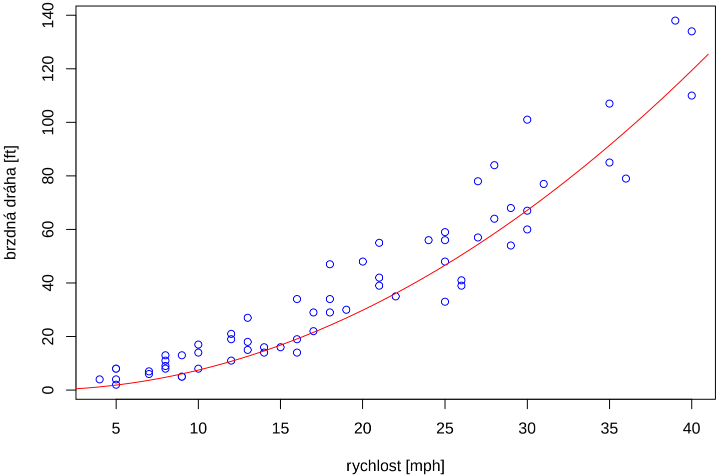
- Data: M. Ezekiel, K. A. Fox (1959). *Methods of correlation and regression analysis*. Wiley, New York.
- Údaje o $n = 63$ autech.
- Y_i brzdná dráha (ft) i-tého auta.
- z_i rychlost (mph) i-tého auta.
- Uvažovaný model kvadratické závislosti:

$$Y_i = \beta z_i^2 + e_i, \quad i = 1, \ldots, 63.$$

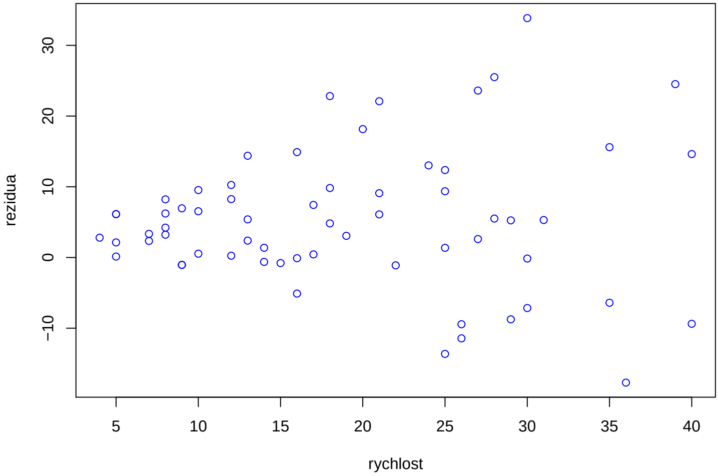
Data


brzdná dráha [ft]
rychlost [mph]
0
20
40
60
80
100
120
140
5
10
15
20
25
30
35
40

Data s odhadnutou kvadratickou závislostí



Závislost reziduí na rychlosti



Příklad - obecnější model

- Možná heteroskedasticita v datech.
- Obecnější model:

$$Y_i = \beta z_i^2 + \exp\{\gamma z_i\} U_i, \quad i = 1, \dots, 63.$$

Test $\mathbf{H}_1 : \gamma = 0$ proti $\gamma > 0$:

- Testová statistika $\mathcal{T}_n^2$ s Wilcoxonovými skóry pro měřítko, t.j. generovaná skórovou funkcí $\varphi(t) = -1 + (2t-1)\log\frac{t}{1-t}$.
- Odpovídající p-hodnota je 0.043 $\Longrightarrow$ heteroskedasticitu musíme připustit.

Test $\mathbf{H}_2 : \beta = 0$ proti $\beta > 0$:

- $\widetilde{\mathcal{T}}_n^2$ s Wilcoxonovými skóry generovanými funkcí $\varphi(t) = t - \frac{1}{2}$.
- Odpovídající p-hodnota je 0.0085 $\Longrightarrow$ regrese je skutečně přítomná.

Závěr

Testy homoskedasticity za přítomnosti rušivé regrese a testy významnosti regrese za přítomnosti rušivé heteroskedasticity.

- Konstruovány bez nutnosti odhadu rušivých odhadů.
- Výpočetně nenáročné, odpadá riziko užití špatného odhadu.
- Rozšíření klasických pořadových testů (regresní pořadové skóry místo pořadí).
- Asymptoticky ekvivalentní odpovídajícím testům v modelu bez rušivých parametrů, kde jejich hodnoty jsou známé.
- Blízkost potvrzena i numericky pro konečný počet pozorování.
- Testy homoskedasticity s rušivou regresí $\mathbf{X}\boldsymbol{\beta}$ nerozliší alternativy s chybami $\exp\{\mathbf{x}_i^\top \gamma\} U_i$ kvůli ancilaritě regresních pořadových skórů.
- Testy o regresi s rušivou heteroskedasticitou - předpoklad symetrie rozdělení chyb modelu.

Poděkování

Děkuji za pozornost.

Práce byla spolufinancována grantem SVV-2013-267 315 a projektem Klimatext reg.číslo CZ.1.07/2.3.00/20.0086.