Kvantilová regresia

Oborový seminár - NMSA401

Lukáš Račko

4.1.2022

UNIVERZITA KARLOVA
Matematicko-fyzikální
fakulta

Základné značenie

- $\mathrm{E}\,(A|B)$ - Podmienená stredná hodnota A za podmienky B;
- $\boldsymbol{Z}$ - Vektor vysvetľujúcich premenných;
- Y - Závislá premenná;
- $\mathbb{X}$ - Regresná matica zodpovedajúca vhodnej transformácií vysvetľujúcich premenných;
- $\boldsymbol{\beta}$ - Vektor parametrov modelu;
- $\hat{\theta}$ - Odhad parametru θ;
- I_n - Matica identického zobrazenia na $\mathbb{R}^n$;
- 1_n - Vektor jednotiek dimenzie n;
- $||.||_2$ - ℓ^2-norma vektoru.

Lineárny model

Nech $\begin{pmatrix} Y_1 \\ \boldsymbol{Z}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \boldsymbol{Z}_n \end{pmatrix}$ je náhodný výber. Označíme $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$:

- Predpokladáme $\boldsymbol{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n)$;
- Modelujeme $\mathrm{E}\,(Y|\boldsymbol{Z})$;
- Riešime $\arg\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{Y} - \hat{\boldsymbol{Y}}(\beta)||_2 = \arg\min_{\beta \in \mathbb{R}^p} ||\boldsymbol{Y} - \mathbb{X}\beta||_2$.

Teoretické problémy modelu

Príliš silné predpoklady modelu.

- Normalita - väčšinou nie až tak dôležitá;
- Homoskedasticita.

Aplikačné problémy modelu

Jednoduchý model vytvára problémy pri interpretácií zložitejších dát a môže byť absolútne nevhodný pre riešenie niektorých problémov.

- Riešenie heteroskedasticity - Box-Coxové transformácie;
- Zanedbáva "rizikovú štruktúru" problému.

Motivácia pre kvantilovú regresiu

Chceme vytvoriť model, ktorý bude modelovať distribúciu $Y|\boldsymbol{Z}$. Dôvody:

- Vyriešíme tak problém s interpretáciou heteroskedastických dát;
- Budeme "poznať" riziko v probléme. Dôležité pri modelovaní vo financiách.

Strátová funkcia modelu

- Pre $\tau \in (0,1)$ definujeme stratovú funkciu

$$\rho_\tau(x) = \tau x \mathbb{I}\{x > 0\} - (1-\tau) x \mathbb{I}\{x \leq 0\}$$

- a jej "deriváciu" ako

$$\psi_\tau(x) = \rho'_\tau(x) = \tau \mathbb{I}\{x > 0\} - (1-\tau)\mathbb{I}\{x < 0\},\ x \neq 0$$

a $\psi_\tau(0) := 0$.

Kvantilová regresia

Nech náhodná veličina $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ má distribučnú funkciu F, potom

$$\forall \tau \in (0,1) : F^{-1}(\tau) = \arg\min_{\theta \in \mathbb{R}} \mathrm{E}\left[\rho_\tau(X - \theta) - \rho_\tau(X)\right].$$

Pre náhodný výber $X_1, \ldots, X_n \sim X$ potom zjavne platí

$$\forall \tau \in (0,1) : \hat{F}_n^{-1}(\tau) = \arg\min_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(X_i - \theta).$$

Regresné kvantily

Majme náhodný výber

$$\begin{pmatrix} \boldsymbol{Z_1} \\ Y_1 \end{pmatrix}, \ldots, \begin{pmatrix} \boldsymbol{Z_n} \\ Y_n \end{pmatrix}$$

z nejakého rozdelenia náhodného vektoru $\begin{pmatrix} \boldsymbol{Z} \\ Y \end{pmatrix}$. Pre $\tau \in (0,1)$ definujeme τ-tý regresný kvantil ako

$$\beta_{\boldsymbol{Z}}(\tau) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \, \mathrm{E}\left(\rho_\tau(Y - \boldsymbol{X}^T \boldsymbol{b}) - \rho_\tau(Y)\right)$$

a jeho odhad ako

$$\hat{\beta}_n(\tau) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \, \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^T \boldsymbol{b})$$

Regresné kvantily

Tak ako v lineárnom modele, opäť predpokladáme linearnu závislosť modelovanej kvantity na regresoroch. T.j. predpokladáme

$$F^{-1}_{Y|\boldsymbol{Z}}(\tau) = \boldsymbol{X}^T \beta_{\boldsymbol{X}}(\tau).$$

V lineárnom modele sme predpokladali

$$\mathrm{E}\,(Y|\boldsymbol{Z}) = \boldsymbol{X}^T \beta.$$

Interpretácia regresných kvantilov

Za platnosti modelu $F^{-1}_{Y|\boldsymbol{Z}}(\tau) = \boldsymbol{X}^T\beta$ máme "štandardnú" interpretáciu hodnôt $\hat{\beta}_{nk}(\tau)$ ako odhad zmeny hodnoty τ-kvantilu podmienenej distribúcie $Y|\boldsymbol{Z}$ pri zmene k-tého regresoru o 1. Špeciálne v prípade jednoduchého modelu s interceptom to znamená zmenu k-tej vysvetľujúcej premennej o 1.

Regresný medián

Špeciálne sa pozrime na prípad $\tau = 0.5$. Vtedy dostávame $\rho_\tau(.) = 0.5|.|$ a preto zjavne

$$\beta_{\boldsymbol{X}} = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \, \mathrm{E}\left(|Y - \boldsymbol{X}^T \boldsymbol{b}| - |Y|\right).$$

Takže $\mathbb{L}_1$-regresia je špeciálnym prípadom kvantilovej regresie a zodpovedá regresnému mediánu.

Vlastnosti regresných kvantilov

Uvažujeme homoskedastický model s interceptom

$$Y \sim \beta_0 + \boldsymbol{X}^T \boldsymbol{\beta} + \varepsilon,$$

kde $\varepsilon \perp \boldsymbol{X}$. Teda pre $\tau \in (0,1)$ pevné platí
$F^{-1}_{Y|\boldsymbol{Z}}(\tau) = \beta_0 + F^{-1}_{\varepsilon}(\tau) + \boldsymbol{X}^T \boldsymbol{\beta}$ a

$$\boldsymbol{\beta}_{\boldsymbol{X}}(\tau) = \begin{pmatrix} \beta_0 + F^{-1}_{\varepsilon}(\tau) \\ \boldsymbol{\beta} \end{pmatrix}.$$

Teda jednotlivé kvantily sa budú líšiť iba v intercepte. Čo by sme čakali, pretože v tomto prípade sa budú marginálne distribúcie líšiť iba v posunutí.

Vlastnosti regresných kvantilov

Nech h je rastúca merateľná funkcia. Potom platí

$$F^{-1}_{h(Y)|\boldsymbol{Z}}(\tau) = \boldsymbol{X}^T\beta \iff F^{-1}_{Y|\boldsymbol{Z}}(\tau) = h^{-1}(\boldsymbol{X}^T\beta).$$

Pre nás je hlavne zaujímavá implikácia z ľava do prava. Zaujímavý príklad pre $h = \log$ potom nech

$$F^{-1}_{\log(Y)|\boldsymbol{Z}}(\tau) = \boldsymbol{X}^T\beta$$

a máme teda

$$F^{-1}_{Y|\boldsymbol{Z}}(\tau) = \exp(\boldsymbol{X}^T\beta).$$

Výpočet regresných kvantilov

Riešime úlohu

$$\min_{\boldsymbol{b} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \boldsymbol{X}_i^T \boldsymbol{b}).$$

Je konvexná?

Výpočet regresných kvantilov

Tento problém vieme previesť do tvaru úlohy lineárneho programovania:

$$\min_{\boldsymbol{b}\in\mathbb{R}^p, \boldsymbol{r}^+\geq\mathbf{0}_n, \boldsymbol{r}^-\geq\mathbf{0}_n} \tau\sum_{i=1}^{n} r_i^+ + (1-\tau)\sum_{i=1}^{n} r_i^-,$$

$$\text{s.t.: } \sum_{j=1}^{p} X_{ij} b_j + r_i^+ - r_i^- = Y_i,\ i = 1, \ldots, n.$$

Túto úlohu už vieme ďalej riešiť štandardnými metódami LP.

Asymptotické rozdelenie parametrov

Za platnosti modelu vieme ukázať, že platí

$$\sqrt{n}(\hat{\beta}_n(\tau) - \beta_{\boldsymbol{X}}(\tau)) \xrightarrow{D} N_p(\mathbf{0}, \mathbb{V}),$$

kde

$$\mathbb{V} = (\mathrm{E}\,[\boldsymbol{X}\boldsymbol{X}^T f_{Y|\boldsymbol{Z}}(F^{-1}_{Y|\boldsymbol{Z}}(\tau))])^{-1}\tau(1-\tau)\,\mathrm{E}\,\boldsymbol{X}\boldsymbol{X}^T(\mathrm{E}\,[\boldsymbol{X}\boldsymbol{X}^T f_{Y|\boldsymbol{Z}}(F^{-1}_{Y|\boldsymbol{Z}}(\tau))])^{-1}.$$

V prípade homoskedastického modelu vieme značne zjednodušiť na tvar

$$\mathbb{V} = \frac{\tau(1-\tau)}{[f_\varepsilon(F_\varepsilon^{-1}(\tau))]^2}(\mathrm{E}\,\boldsymbol{X}\boldsymbol{X}^T)^{-1}.$$

Odhad variančnej matice za predpokladu homoskedasticity

Hlavný problém je v odhade $s(\tau) = 1/f_\varepsilon(F_\varepsilon^{-1}(\tau))$. Koenker navrhuje použiť

$$\hat{s}_n = \frac{\hat{F}_{n\hat{\varepsilon}}^{-1}(\tau + h_n) - \hat{F}_{n\hat{\varepsilon}}^{-1}(\tau - h_n)}{2h_n},$$

kde h_n je ľubovoľná postupnosť konvergujúca k 0 a

$$\hat{F}_{n\hat{\varepsilon}}^{-1}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\hat{\varepsilon}_i(\tau) \leq x\} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{Y_i - \boldsymbol{X}_i^T \hat{\beta}_n(\tau)\}.$$

Za predpokladu normálne rozdelených reziduí, Koenker ďalej navrhuje použiť

$$h_n = n^{-1/3} u_{1-\tau/2}^{2/3} \left(\frac{1.5\phi^2(u_\tau)}{2u_\tau^2 + 1} \right)^{1/3},$$

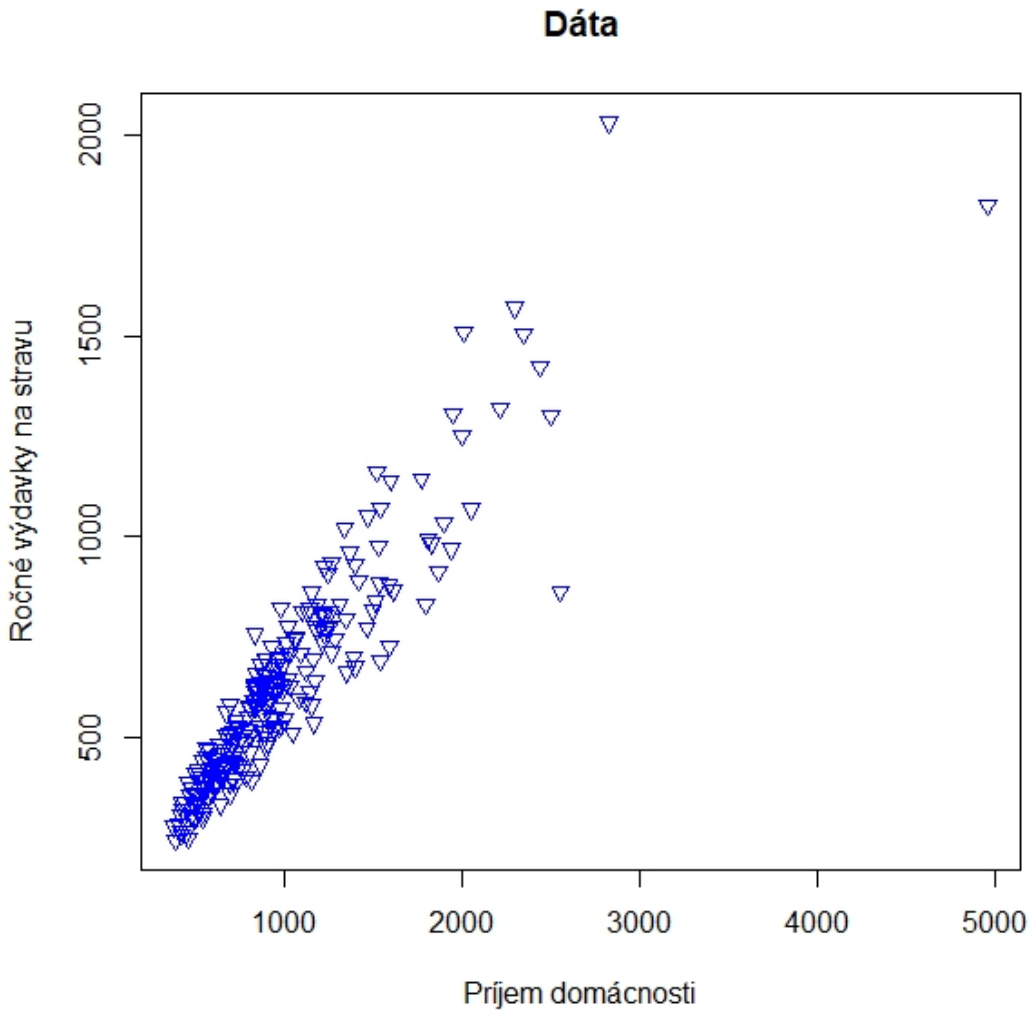
kde ϕ značí hustotu $N(0, 1)$.

Príklad

- Na ilustráciu použijeme dáta "engel" z knižnice "quantreg". Tieto dáta obsahujú ročný príjem (income) a ročné výdavky na jedlo (foodexp) v 235 belgických domácnostiach.
- Táto knižnica taktiež obsahuje funkciu "rq", ktorá je veľmi podobná nám známej funkcií "lm". Na vyrovnanie dát 10%-ným regresným kvantilom použijeme príkaz '>rq(foodexp∼income, tau = 0.1, data = engel)'.
- "rq" má ako defaultnú hodnotu $\tau = 0.5$.
- Jedná z možností je "method", ktorá umožňuje výber výpočetného algoritmu. Pri dátach väčšieho rozsahu sa odporúča zmeniť defailtnú metódu pre urýchlenie výpočtu.

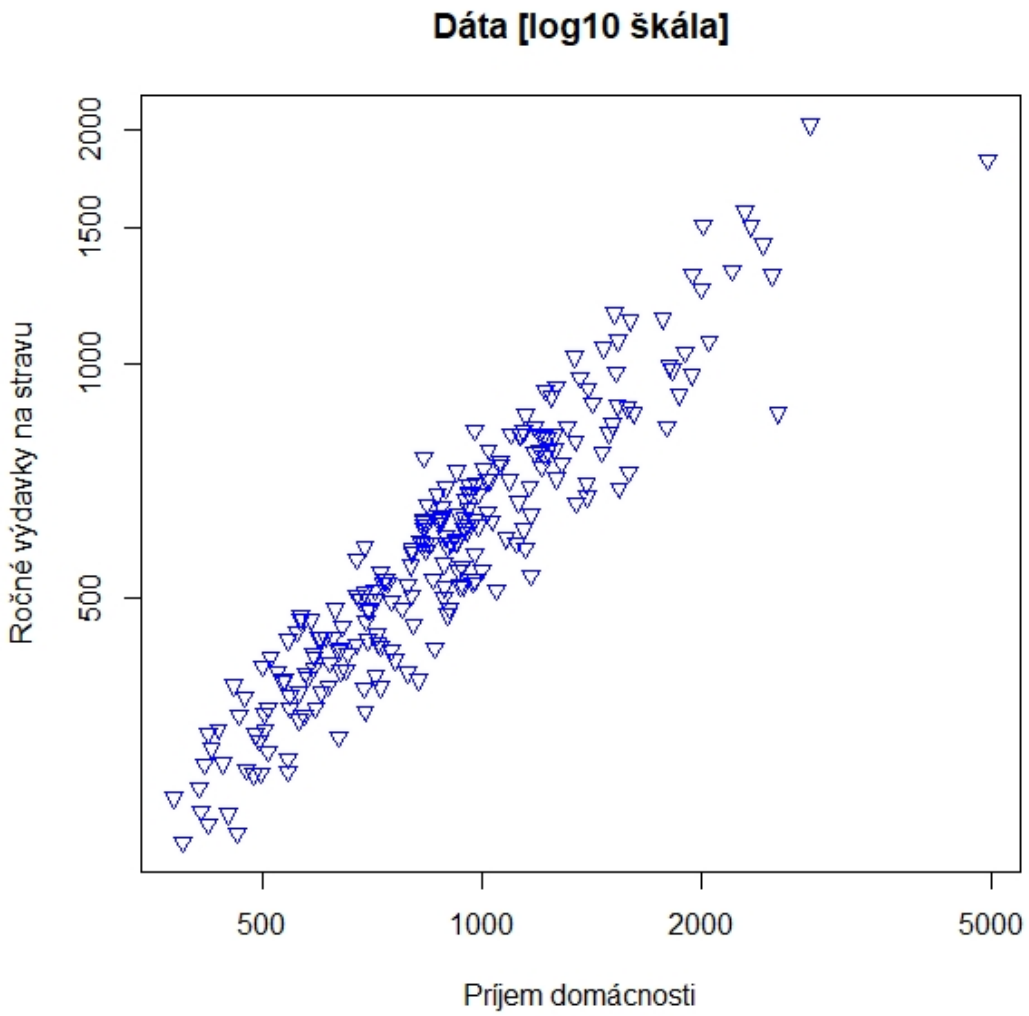
Dáta

Zrejme homoskedasticita sa nedá predpokladať.



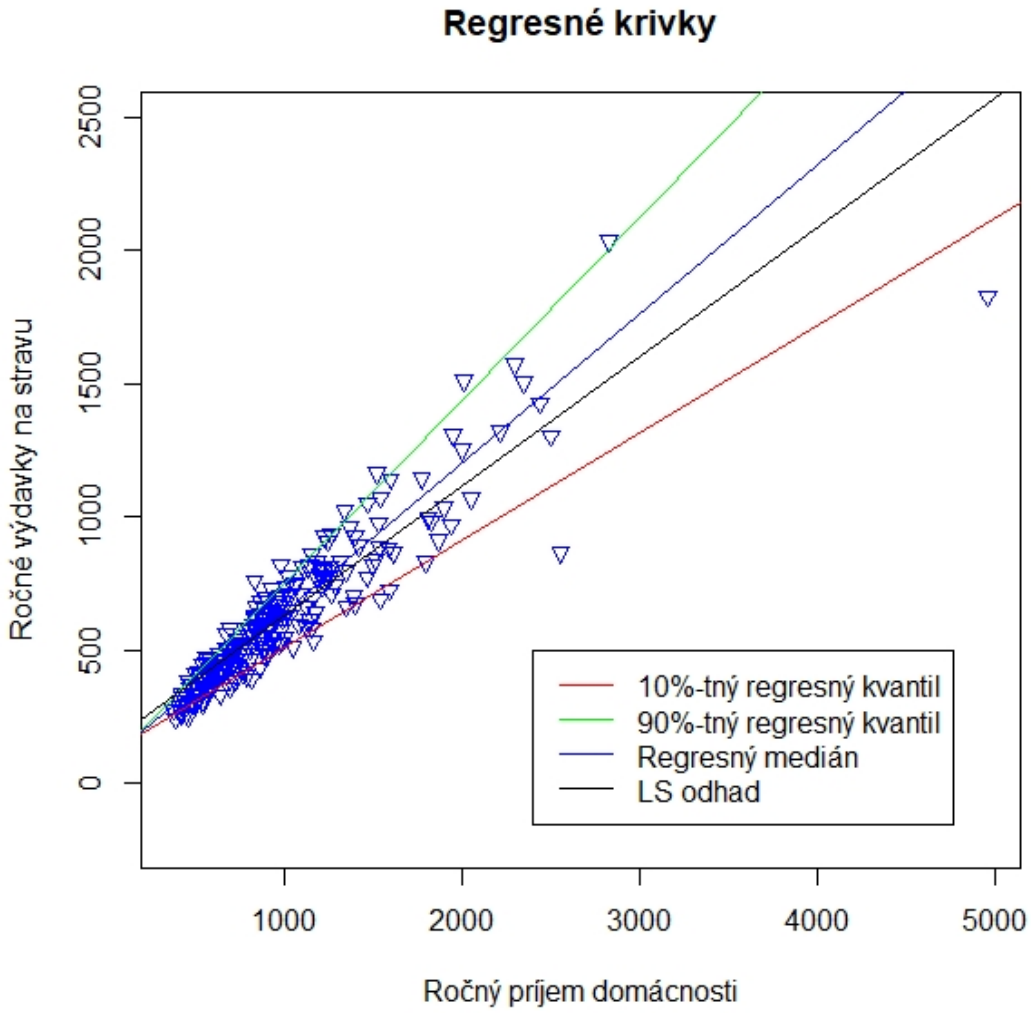
Dáta

Vieme sa "zbaviť" heteroskedasticity transformáciou odpovede aj vysvetľujúcej premennej.



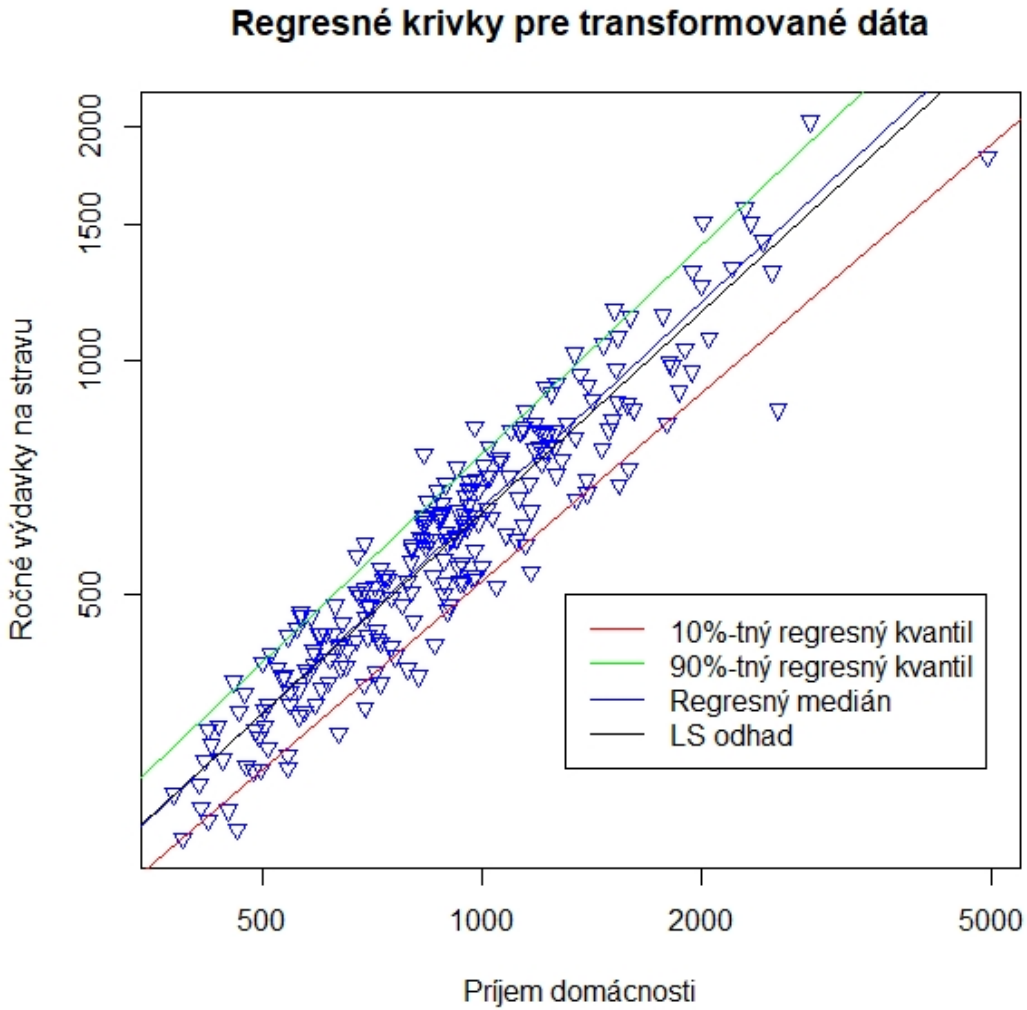
Odhadnuté hodnoty

- Bez použitia transformácií;
- Priesečník odhadnutých kvantilov - môže byť znamením zlého modelu.



Odhadnuté hodnoty

- Logaritmická škála;
- Bez priesečníka odhadnutých kvantilov.



Zhrnutie

- Kvantilová regresia môže byť vhodnejšia z povahy problému;
- Máme radi homoskedastické dáta;
- Tradeoff interpretácia vs. výpočetná zložitosť pri heteroskedastických dátach.

Zdroje

[1] Koenker, R.
Quantile regression
Cambridge University Press, 2005.

[2] Omelka, M.
NMST434 - Modern statistical methods, Course notes
MFF UK, September, 2021.