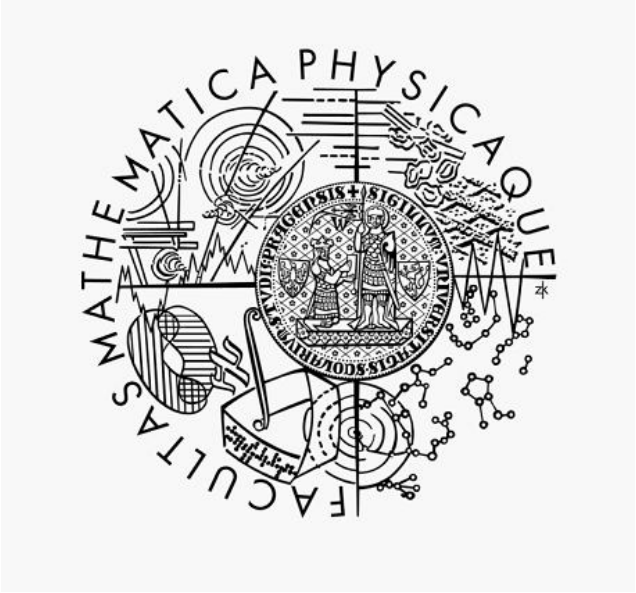
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Peter Rusnák

Regresní kvantily

Katedra pravděpodobnosti a matematické statistiky MFF UK

Vedoucí bakalářské práce: RNDr. Jan Kalina, Ph.D.
Ústav informatiky AV ČR

Studijní program: matematika
Studijní obor: finanční matematika

Praha 2011

Ďakujem pánovi RNDr. Janu Kalinovi, Ph.D. za trpezlivé vedenie bakalárskej práce. Ďálej ďakujem svojej rodine, blízkym a každému, kto mi s touto prácou akokoľvek pomohol.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V …..... dne..... podpis

Názov práce: Regresní kvantily

Autor: Peter Rusnák

Katedra / Ústav: Katedra pravděpodobnosti a matematické statistiky MFF UK

Vedúci bakalárskej práce: RNDr. Jan Kalina, Ph.D. ,Ústav informatiky AV ČR

Abstrakt: Kvantilová regresia je štatistická metóda slúžiaca na určovanie závislostí medzi premennými, ktorá bola navrhnutá už v článku Koenker a Bassett (1978). Od tej doby prešla veľkým rozvojom, keď boli študované jej teoretické vlastnosti, a zároveň si našla radu praktických aplikácii pri spracovaní reálnych dát v najrôznejších oboroch. Kým bežná metóda najmenších štvorcov popisuje vzťah medzi jedným respektíve viacerými kovariátmi X a podmieneným priemerom odpovedajúcej premennej Y daným $X = x$, kvantilová regresia popisuje vzťah medzi X a podmienenými kvantilmi Y danými $X = x$. Táto práca obsahuje teóriu nevyhnutnú pre pochopenie vzťahu medzi štandardnou a kvantilovou regresiou a umožňujúcu začlenenie takto získaných odhadov do väčšej skupiny M-odhadov. Výpočet koeficientu pre jednotlivé kovariáty je prevedený Frisch-Newtnovým algoritmom, ktorý patrí k metódam lineárneho programovania. Taktiež si ukážeme, ako vedľajší produkt tohto algoritmu, takzvané regresné poradie je vypočítané a ako ho použiť pre testovanie hypotéz. V druhej časti budeme ilustrovať numerický výpočet pre kvantilovú regresiu ako na vygenerovaných dátach tak na dátach reálnych.

Kľúčové slová: kvantilová regresia, M-odhady, regresné poradie, regresné poradové skóry

Title: Regression Quantiles

Author: Peter Rusnák

Department: Department of Probabilty and Mathematical Statistics

Supervisor: RNDr. Jan Kalina, Ph.D.,Institute of Computer Science, AS CR

Abstract: Quantile regression is a statistical method for specifying dependencies among variables, which was introduced by Koenker a Bassett in 1978. Since that time it has gone through a big development, when its theoretical properties have been under study, and it also has found many practical applications for data processing in variety of fields.

While ordinary least-squares regression describes the relationship between one or more covariates X and the conditional mean of a response variable Y given $X = x$, quantile regression describes the relationship between X and the conditional quantiles of variable Y given $X = x$. This work contains the theory necessary for understanding relationship between standard and quantile regression and enabling include so received estimates to bigger group of M-estimates. The computation of coefficients for particular covariates is made by using Frisch-Newton algorithm belonging to methods of linear programming. The so-called regression ranks are also obtained as a by-product of this algorithm and we discuss their computational aspects and usage for hypothesis testing.In the second part, we are illustrating numerical computation for quantile regression on theoretical data sets as well as data sets from real world.

Keywords: Quantile regression, M-estimates, regression ranks, regression rank scores

Obsah

Úvod

Hlavnou úlohou štatistiky je priniesť poriadok do spleti zdanlivo chaotických údajov, k čomu nám slúži analýza dát. Veľmi efektívnym spôsobom, ako analyzovať dáta, je skúmať ich metódami klasickej lineárnej regresie, akou je napríklad najčastejšie používaná metóda najmenších štvorcov. Táto metóda slúži na odhadnutie parametrov lineárneho regresného modelu, prostredníctvom ktorého získame odhad podmienenej strednej hodnoty závislej na hodnotách vysvetľujúcich premenných (regresorov). Stredná hodnota je síce dôležitým ukazovateľom, ktorý predstavuje "stredové" chovanie náhodnej veličiny, avšak nedáva nám žiadnu predstavu o tom, ako sa správa podmienená hustota na krajoch rozdelenia. To priviedlo pánov G. Basseta a R. Koenkera k snahe o zobecnenie obyčajných výberových kvantilov na lineárny regresný model.

Kvantilovú regresiu môžme použiť na preskúkamie správania sa nezávislých premenných (kovariátov) v každom kvantile skúmanej premennej, čo nám dáva príležitosť pre ucelenejší pohľad na priebeh a vzťahy medzi stochastickými premennými, ako použitie metód klasickej lineárnej regresie.

Táto práca si kladie za úlohu sprostretkovať výklad základných vlastností regresných kvantiov a demonštráciu takto získaných teoretických vedomostí na vhodne zvolených príkladoch. Pred odvodením teórie regresných kvantilov bude nutné zoznámiť sa taktiež so základmi regresie, robustnej štatistiky a optimalízácie.

V prvej kapitole si popíšeme teoretické základy stojace za kvantilmi, regresiou a v krátkosti si teoreticky zhrnieme postup pri odhade metódou najmenších štvorcov. V druhej kapitole si popíšeme základné vlastnosti robustných odhadov, špeciálne sa budeme venovať M-odhadom. V tretej kapitole si zavedieme odhad regresných kvantilov, popíšeme základy lineárneho programovania, ukážeme si ako previesť úlohu odhadnutia regresných koeficientov na úlohu lineárneho programovania a popíšeme si ako ju teoreticky riešiť prostredníctvom Frisch-Newtnovho algoritmu. Taktiež poukážeme na riešenie duálnej úlohy, ktoré nám vznikne pri riešení úlohy lineárneho programovania, toto riešenie nazveme regresné poradové skóre a ukážeme si jeho využitie pri testovaní hypotéz. V poslednej štvrtej kapitole si ukážeme využitie nadobudnutých znalostí na dvoch príkladoch, kde prvý bude demonštratívny na vhodne vygenerovaných dátach a druhý bude predstavovať použitie kvantilovej regresie na reálnych dátach.

1. Úvod do kvantilov a regresnej analýzy

1.1. Kvantil pravdepodobnostného rozdelenia

Definícia kvantilov vychádza z jednoduchej optimalizačnej úlohy, ktorá pojednáva o teoretickom probléme bodového odhadu, ktorý je požadovaný pre náhodnú veličinu s distribučnou funkciou F. Nech strata je popísaná lineárnou funkciou v závislosti od objemu financovaných prostredkov x:

$$\rho_\tau(x) = x(\tau - I(x < 0)) \text{ pre nejaké } \tau \in (0, 1).$$

Našou úlohou je nájsť $\tilde{x}$ minimalizujúce očakávanú stratu, a teda budeme sa snažiť minimalizovať

$$E\,\rho_\tau(X - \tilde{x}) = (\tau - 1) \cdot \int_{-\infty}^{\tilde{x}} (x - \tilde{x})\, dF(x) + \tau \cdot \int_{\tilde{x}}^{\infty} (x - \tilde{x})\, dF(x).$$

Derivovaním v premennej $\tilde{x}$ dostávame

$$(1 - \tau) \cdot \int_{-\infty}^{\tilde{x}} dF(x) - \tau \cdot \int_{\tilde{x}}^{\infty} dF(x) = F(\tilde{x}) - \tau.$$

Keďže F je neklesajúca funkcia, tak každý prvok množiny $\{x : F(x) = \tau\}$ minimalizuje očakávanú stratu. Ak je táto množina jednoprvková potom $\tilde{x} = F^{-1}(\tau)$ v opačnom prípade zvolíme za $\tilde{x}$ najmenší prvok množiny $\{x : F(x) = \tau\}$, aby bola dodržaná dohoda, že empirická kvantilová funkcia je zľava spojitá.

Definícia : Nech $F(y) = P(Y \le y)$ je distribučná funkcia daného rozdelenia pravdepodobnosti náhodnej veličiny Y a $\alpha \in \langle 0,1 \rangle$. Potom kvantilovú funkciu náhahodnej veličiny Y definujeme ako

$$Q(\alpha) = F^{-1}(\alpha) = \inf\{y \in \mathbb{R} : F(y) \ge \alpha\}$$

a číslo $y_\alpha = Q(\alpha)$ sa nazýva α-kvantil rozdelenia s distribučnou funkciou $F(y)$.

α-kvantil y_α rozdeľuje definičný obor náhodnej veličiny Y na dve časti tak, že platí

$$P(Y < y_\alpha) = \alpha \qquad \& \qquad P(Y \ge y_\alpha) = 1 - \alpha.$$

Kvantily pre niektoré špeciálne hodnoty α sú označené špeciálnymi názvami a ich hodnoty pre najdôležitejšie rozdelenia sú uvedené v tabuľkách:

- $F^{-1}\left(\frac{1}{4}\right)$ sa nazýva dolný kvartil a oddeľuje zo štatistického súboru jednu štvrtinu.

- $F^{-1}\left(\frac{1}{2}\right)$ sa nazýva medián a rozdeľuje súbor na dve rovnako početné množiny.
- $F^{-1}\left(\frac{3}{4}\right)$ sa nazýva horný kvantil.

1.2. Regresná analýza

Účelom regresnej analýzi je popísať vzťah medzi skúmanou náhodnou veličinou (označme ju Y) a známym náhodným vektorom $\mathbb{Z} = (Z_1, Z_2, \ldots Z_p)$, ktorý sa na ňu viaže. Našim cieľom bude určiť, na ktorých premenných z $Z_1, Z_2, \ldots, Z_p$ skúmaná premenná Y závisí, popísať vzťah ako na nich závisí a predpovedať hodnoty Y, ak máme k dispozícii pozorovania $\mathbb{Z}$. Porovanie Y_i nemusí priamo závisieť na $\mathbb{Z}_i$, ale na nejakej jeho transformácii $\mathbb{X}_i = f(\mathbb{Z}_i)$, $\mathbb{X}_i \in \mathbb{R}^k$. Náš cieľ dosiahneme prostredníctvom vhodne zvolenej funkcie (označme ju g), ktorá je funkciou "vysvetľujúcich premenných $X_1, X_2, \ldots, X_k$ a vhodne aproximuje skúmanú veličinu . t.j. "$Y = g(\mathbb{X})$". V tejto práci sa budeme zaoberať najmä lineárnou regresnou analýzou a teda g bude mať tvar

$$g(\mathbb{X}) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k,$$

kde konštanty $\beta_1, \ldots, \beta_k \in \mathbb{R}$ predstavujú regresné koeficienty.

Model prisluchajúci k takto zvolenej funkcii g nazveme lineárny regresný model a jeho tvar pre počet pozorovaní $n > k$ odvodíme ako:

Nech $x_{1_1}, \ldots x_{1_n}$ ... predstavuje konkrétnych n hodnôt náhodnej veličiny X_1.

$\vdots$

Nech $x_{k_1}, \ldots x_{k_n}$ ... predstavuje konkrétnych n hodnôt náhodnej veličiny X_k.

Nech $y_1, \ldots, y_n$ ... predstavuje konkrétnych n hodnôt náhodnej veličiny Y.

Keďže g ($\mathbb{X}$) je "len vhodnou aproximáciou" Y tak existujú $\epsilon_1,\ldots,\epsilon_n$ také ,že

$$y_1 = \beta_1 x_{1_1} + \beta_2 x_{2_1} + \ldots + \beta_k x_{k_1} + \epsilon_1,$$

$$y_2 = \beta_1 x_{1_2} + \beta_2 x_{2_2} + \ldots + \beta_k x_{k_2} + \epsilon_2,$$

$$\vdots$$

$$y_n = \beta_1 x_{1_n} + \beta_2 x_{2_n} + \ldots + \beta_k x_{k_n} + \epsilon_n.$$

$\epsilon_1,\ldots,\epsilon_n$ predstavujú "chyby aproximácie" a pomocou regresnej analýzi sa ich budeme snažiť zmenšovať prostredníctvom vhodných volení funkcie g($\mathbb{X}$).

Ak o chybovom vektore $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ budeme predpokladať, že jeho zložky sú normálne rozdelené, vzájomne nezávislé náhodné veličiny s nulovou strednou hodno-

tou a rovnakým rozptylom σ^2 a teda $\epsilon \sim N_n(0_n, \sigma^2 \cdot I_n)$. Potom hovoríme o klasickom lineárnom regresnom modely, ktorého tvar môžeme preformulovať ako:

$$E[Y \mid x_{1_1}, \ ..., x_{k_i}] = \beta_1 x_{1_i} + \beta_2 x_{2_i} + ... + \beta_k x_{k_1}$$

$$var[Y \mid x_{1_1}, \ ..., x_{k_i}] = \sigma^2$$

Úlohou regresie je získať nejaký odhad $\hat{\beta}$ vektoru parametrov β, za účelom spočítania $\hat{y}_i = \mathbb{x}^T.\hat{\beta}$. Pomocou dobrého odhadu $\hat{\beta}$ by sme mali získať $\hat{y}_i$, čo najbližšie k y_i. Teraz si ukážeme ako sa odhadúvajú regresné koeficienty v klasickom lineárnom modely pomocou najčastejšie používanej metódy najmenších štvorcov.

Definícia: Odhad $\hat{\beta}$ vektoru parametrov β definovaný vzťahom

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \arg\min_{\beta \in \mathbb{R}^k} (\mathbb{y} - \mathbb{x}\beta)^T \cdot (\mathbb{y} - \mathbb{x}\beta)$$

sa nazýva *odhad metódou najmenších štvorcov*.

Minimalizovaná funkia sa nazýva reziduálny súčet štvorcov a keďže nie je ohraničená zhora tak $\hat{\beta}$, ktoré získame ako riešenie sústavy normálnych rovníc (1), musí byť minimom.

$$\frac{\partial}{\partial \beta} (\mathbb{y} - \mathbb{x}\beta)^T \cdot (\mathbb{y} - \mathbb{x}\beta) = 0 \tag{1}$$

Riešenie sústavy bude mať tvar

$$\hat{\beta} = \left(\mathbb{x}^T \cdot \mathbb{x}\right)^{-1} \mathbb{x}^T \mathbb{y}.$$

Veta: Vlastnosti odadu $\hat{\beta}$ získaného metódou najmenších štvorcov:

1) $E\hat{\beta} = \beta$

2) $var\,\hat{\beta} = \sigma^2\left(\mathbb{x}^T \mathbb{x}\right)^{-1}$

3) $\hat{\beta}$ *je konzistentný odhad*

4) Ak $\epsilon \sim N_n(0_n, \sigma^2 \cdot I_n)$, *potom* $\sqrt{n} \cdot (\hat{\beta} - \beta) \xrightarrow{d} N_k\left(0, \sigma^2\left(\mathbb{x}^T \mathbb{x}\right)^{-1}\right)$.

Dôkaz:viď [7] kapitola 13.

Veta: *Nech platia vyššie uvedené predpoklady a nech navyše platí, že* $\mathbb{Y} \sim N_n\left(\mathbb{x}\beta, \sigma^2.I_n\right)$ *potom pre reziduálny sučet štvorcov* SS_e *platí, že* $\frac{SS_e}{\sigma^2} \sim \chi^2_{n-k}$.

Dôkaz: viď [7] Veta 13.6.

2. M-odhady

Odhady získané lineárnou regresiou sú špeciálnym prípadom *M-odhadov*, ktoré patria k robustným odhadom. Samotná metóda najmenších štvorcov je príliš citlivá k predpokladom normality i k prítomnosti odľahlých pozorování, a preto je žiadúce študovať aj iné (robustné) odhady, ktoré by neboli natoľko citlivé. Jednou z možností sú práve *M-odhady*. Zároveň je pravda, že sám odhad metódou nejmenších štvorcov patrí do triedy *M-odhadov*. V tejto kapitole uvedieme základy teórie *M-odhadov* umožňujúce začlenenie nami počítaných odhadov do väčšej štatistickej rodiny a ukážeme si dôležité vlastnosti, ktoré na nich môžme aplikovať.

2.1. Obecný M-odhad

Majme štatistický model s náhodným výberom $y_1,y_2,...,y_n$ z neznámej náhodnej veličiny Y. Označme θ ako funkciu distribučnej funkcie náhodnej premennej Y (štatistický funkcionál) $\theta = T(P)$, kde P je rozdelenie Y.

M -odhad parametru θ získame ako riešenie minimalizačnej úlohy

$$\sum_{i=1}^{n}\rho(y_i, \theta) := \min \ , \text{kde}\, \theta \in \Theta\,.$$

V tejto bakalárskej práci sa budeme zaoberať príkladom, kde *M-odhad* je model s posunutím θ t.j. budeme hľadať riešenie minimalizačnej úlohy

$$\sum_{i=1}^{n}\rho(y_i - \theta) := \min \ , \text{kde}\, \theta \in \Theta. \tag{2}$$

V prípade ak je funkcia $\rho(\cdot)$ diferenciovateľná s absolútne spojitou deriváciou, tak môžme túto minimalizačnú úlohu prepísať do tvaru hľadania riešenia rovnice

$$\sum_{i=1}^{n}\psi(y_i - \theta) = 0\,, \text{kde}\ \psi(\,.\,) = \frac{\partial}{\partial\theta}\rho(.). \tag{3}$$

Predpokladajme, že ρ je konvexná funkcia, z toho vyplýva, že aj $\sum_{i=1}^{n}\rho(y_i - \theta)$ je konvexná funkcia a teda existuje riešenie minimalizačnej úlohy (3). Avšak toto riešenie nemusí byť určené jednoznačne. Napríklad ak ρ je lineárna funkcia, potom jej derivácia ψ je konštantná funkcia a teda rovnica (3) nemá riešenie.

Potom riešenie určíme ako

$$\theta := \frac{1}{2}\,(\theta^{+} + \theta^{-})\,, \text{kde}\quad \theta^{+} = \inf\left\{\theta \;\middle|\; \sum_{i=1}^{n}\psi(y_i - \theta) < 0\right\}\ \ \theta^{-} = \sup\left\{\theta \;\middle|\; \sum_{i=1}^{n}\psi(y_i - \theta) > 0\right\}.$$

2.2.Príklady M-odhadov

Z rovností (2) a (3) vyplýva, že *M-odhad* je určený voľbou funkcie ρ, alebo jej derivácie ψ, ktorých voľba závisí od vlastností parametra, ktorý chceme odhadnúť. Napríklad, ak má byť parameter zároveň stredom symetrie s okolím v nule, tak je vhodné voliť párnu funkciu ρ a teda ψ je tak nepárna funkcia. Tým, že požadujeme aby odhad bol rozbustný, dostávame ohraničenosť influenčnej funkcie a z toho vyplývajúcu ohraničenosť funkcie ψ. Uveďme si konkrétne príklady takýchto funkcií ρ.

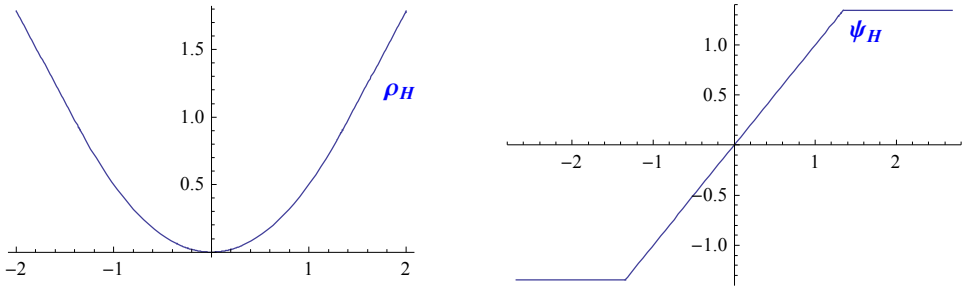
- **Huberova funkcia**

$$\rho_H(x) = \begin{cases} \frac{x^2}{2} & \dots |x| < k \\ k \cdot \left(|x| - \frac{k}{2}\right) & \dots |x| \geq k \end{cases}$$

kde $k = c\sigma$ a c je nejaká konštanta. Empiricky bolo určené, že najvyššiu efektivitu Huberovej funkcie dosiahneme pre c=1.345.

Príslušná funkcia ψ_H bude mať tvar:

$$\psi_H(x) = \begin{cases} x & \dots |x| < k \\ k \cdot \mathrm{sgn}(x) & \dots |x| \geq k \end{cases}$$



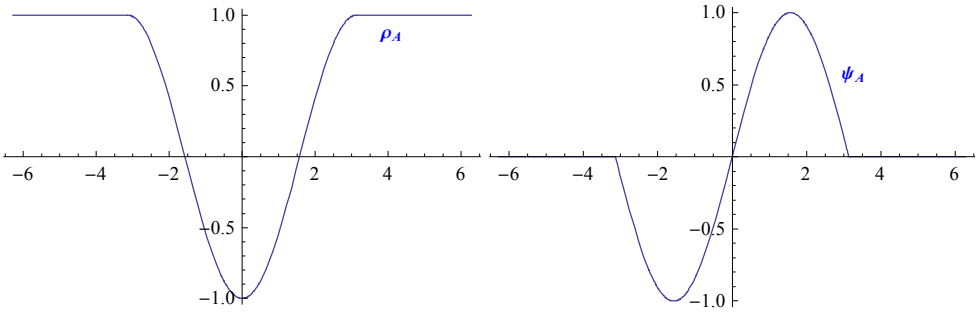
Obrázok 1. Graf Huberovej funkcie a k nej príslušnej funkcie ψ_H

- **Andrewsova funkcia**

Túto funkciu si uvedieme ako príklad toho, že nie všetky voľby funkcie ρ sú konvexné, čo spôsobuje, že $\sum_{i=1}^{n} \rho(y_i - \theta)$ môže nadobúdať okrem globálnych extrémov aj extrémy lokálne.

$$\rho_A(x) = \begin{cases} -k \cdot \cos\frac{x}{k} & \dots |x| \leq k \cdot \pi \\ k & \dots |x| > k \cdot \pi \end{cases}$$

$$\psi_A(x) = \begin{cases} \sin\frac{x}{k} & \dots |x| \leq k \cdot \pi \\ 0 & \dots |x| > k \cdot \pi \end{cases}$$



Obrázok 2. Graf Andrewsonovej funkcie a k nej príslušnej funkcie ψ_A

2.3. M-odhad regresných parametrov

M-odhad regresných parametrov je definovaný ako riešenie $\beta = (\beta_1, \ldots, \beta_k)^T$ minimalizujúce súčet funkcií reziduí t.j. rieši minimalizačnú úlohu

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{k} x_{j_i}\,\beta_j}{\hat{\sigma}}\right) := \min,$$

kde $\hat{\sigma}$ predstavuje odhad smerodatnej odchýlky.

V prípade ak je funkcia $\rho(\cdot)$ diferenciovateľná s absolútne spojitou deriváciou, tak môžme túto minimalizačnú úlohu prepísať do tvaru hľadania riešenia rovnice

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \sum_{j=1}^{k} x_{j_i}\,\beta_j}{\hat{\sigma}}\right) = 0.$$

Existuje veľa používaných funkcií ψ , ktoré vedú k rôznym M-odhadom napríklad tie uvedené v 2.2.

M- odhad s monotónnou funkciou ψ je doporučený k použitiu, ak náhodné chyby v regresnom modely majú rozdelenie s chvostami, ktoré sú ťažšie ako chvosty normálneho rozdelenia, ale nie sú ťažsie ako chvosty dvojite exponenciálneho rozdelenia.

2.4. Vlastnosti M- odhadov

Napriek tomu, že M-odhady vznikli ako robustné odhady klasických metód, sú robustné len voči odľahlým hodnotám odozvy. Pritom sú veľmi citlivé (nerobustné) voči prítomnosti odľahlých hodnôt v regresoroch (viď [5]), takže má zmysel ich používať len v prípade kontaminácie vysvetľovanej premennej. Vlastnosťami *M-odhadov* budeme rozumieť popísanie vplyvov týchto kontaminácii na hodnotu odhadov.

Definícia: Nech $\mathcal{P}$ je systém všetkých pravdepodobnostných mier na priestore $(\mathcal{X},\mathcal{B})$, kde $\mathcal{X}$ je úplný, separabilný metrický priestor a $\mathcal{B}$ je σ-algebra borelovských podmnožín $\mathcal{X}$ a nech $P\in\mathcal{P}$. Povieme, že štatistický funkcionál T je diferencovateľný v Gateauxovom zmysle podľa P v smere Q, ak existuje limita

$$T_Q(P) = lim_{t\to 0}\frac{T((1-t)\,P + t\cdot Q) - T(P)}{t}$$

hodnota tejto limity predstavuje Gateauxovu deriváciu podľa P v smere Q.

Definícia: Nech $\epsilon_1...,\epsilon_n$ je náhodný výber z náhodnej veličny Y. Potom empirickou (výberovou) distribučnou funkciou pre $y \in \mathbb{R}$ nazývame funkciu

$$F_n(y) = \frac{1}{n}\sum_{k=1}^{n}\mathcal{I}_{(-\infty,y)}(y_k).$$

Označme P_n ako empirickú distribučnú funkciu prisluchajúcu vektoru $(y_1...,y_n)$:

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{y_i}.$$

Pre ďalšie počítanie v tejto práci budeme predpokladať, že P_n konverguje k P, resp. je v istom okolí P a platí

$$T(P_n) - T(P) = \frac{1}{n}\sum_{i=1}^{n}T_{\delta_{y_i}}(P) + o_P\left(n^{\frac{-1}{2}}\right) \text{ pre } n\to\infty.$$

Táto rovnosť nám umožňuje $\frac{1}{n}\sum_{i=1}^{n}T_{\delta_{y_i}}(P)$ chápať ako chybu odhadu $T(P)$ pomocou $T(P_n)$ a $T_{\delta_{y_i}}$ ako príspevok i-teho pozorovania k tejto chybe. To nás privádza k myšlienke definície influenčnej funkcie .

Definícia: Gateauxovú deriváciu T podľa rozdelenia P v smere δ_y, $y \in \mathbb{Y}$ nazývame influenčná funkcia funkcionálu T v rozdelení pravdepodobnosti P a píšeme

$$IF(y,\,T,\,P) = T_{\delta_y}(P) = lim_{\epsilon\to 0}\frac{T((1-\epsilon)\,P + \epsilon\delta_y) - T(P)}{\epsilon}.$$

Pri kvantitatívnych charakteristikách robustnosti nás bude zaujímať, aký veľký vplyv na funkcionál T má "kontaminácia" rozdelenia P, ktorá vznikne pridaním nových pozorovaní y_{n+1}, y_{n+2},... . Pri globálnej citlivosti nás bude zaujímať aký maximálny vplyv by mohlo mať nové pozorovanie y_{n+1} na funkcionál T.

Definícia: Nech T je štatistický funkcionál a P je rozdelenie pravdepodobnosti na $(\mathcal{Y},\mathcal{B})$. Potom globálnou citlivosťou funkcionálu T pre rozdelenie pravdepodobnosti P nazývame hodnotu

$$\gamma = sup_{x\in\mathcal{Y}} \|IF(y,\ T,\ P)\|\,.$$

Pre lepšie pochopenie si uveďme i diskretizovanú formu influenčnej funkcie. Pridajme dodatočné pozorovanie y_{n+1} k pozorovaniam $(y_1...,y_n)$. Potom vplyv tohoto pozorovania môžme vyjadriť ako

$$I(T_n,\ y_{n+1}) = T_{n+1}(y_1,\ y_2,\ \ ...,\ y_n,\ y_{n+1}) - T_n(y_1,\ y_2,\ \ ...,\ y_n). \qquad (4)$$

Citlivosťou funkcionálu T_n na pridanie ďalšieho pozorovania, tu rozumieme číslo

$$S(T_n) = \sup_y |I(T_n,\ y)|.$$

Predveďme si to na výpočte globálnej citlivosti strednej hodnoty

$$T(P) = \mathbb{E}_P\, X\ \ ,\ T_n = \overline{X}\ ,\ T_{n+1} = \overline{X}_{n+1},$$

potom T_{n+1} môžme vyjadriť ako

$$T_{n+1} = \frac{1}{n+1}\left(n\cdot\overline{X}\ + y\right)$$

a z (4) tak dostávame

$$I(T_n,\ y) = \frac{1}{n+1}\left(n\cdot\overline{X}\ + y\right) - \overline{X}\ = \frac{1}{n+1}\left(n\cdot\overline{X}\ - (n+1)\cdot\overline{X}\ + y\right) = \frac{1}{n+1}\cdot\left(y - \overline{X}\right),$$

prechodom k suprému cez všetky Y získame

$$S(\overline{X}) = \frac{1}{n+1}\cdot\sup_y\left|y - \overline{X}\right| = \infty.$$

A teda globálna citlivosť je rovná nekonečnu, čo znamená, že priemer nie je robustným odhadom strednej hodnoty. Tu sa ukazuje, že M-odhady nemusia mať vždy výhodné robustné vlastnosti. A teda triedu M-odhadov tvoria nielen odhady, ktoré su robustné, ale patria tam aj iné odhady (napr. priemer), ktoré robustné vlastnosti nemajú. Vo všeobecnosti platí, že ak influenčná funkcia nie je obmedzená, potom odhad príslušného parametru nie je robustný.

Pri lokálnej citlivosti nás bude zaujímať vplyv relatívne malých zmien, pri nahradení pozorovania x pozorovaním y, na hodnotu funkcionálu T.

Definícia: Nech T je štatistický funkcionál a P je rozdelenie pravdepodobnosti na $(\mathcal{Y},\mathcal{B})$. Potom lokálnou citlivosťou funkcionálu T pre rozdelenie pravdepodobnosti P nazývame hodnotu

$$\lambda = sup_{x,y\in\mathcal{Y},x\neq y}\frac{\|IF(x,T,P)-IF(y,t,p)\|}{\|x-y\|}.$$

Rovnako ako pri globálnej citlivosti sa pri stanovení odhadu snažíme docieliť, čo najmenšiu hodnotu lokálnej citlivosti. Avšak pri dosiahnutí veľkých hodnôt (resp. nekonečna) si musíme uvedomiť, že zmena influenčnej funkcie je meraná vzhľadom k rozdielu pozorovaní x a y a tak samotná zmena influenčnej funkcie môže byť nepatrná.

Ďalej by nás mohlo zaujímať, koľko minimálne pozorovaní v našom náhodnom výbere by sa muselo zmeniť, aby nový odhad funkcionálu založený na takto upravenom náhodnom výbere bol ľubovoľne vzdialený od nášho pôvodného odhadu.

Definícia: Nech T je štatistický funkcionál a P je rozdelenie pravdepodobnosti na $(\mathcal{Y},\mathcal{B})$. Označme $Y^0=(y_1, ...,y_n)$ náhodný výber z náhodnej veličiny Y a k nemu príslušnú hodnotu funkcionálu $T(Y^0)$. Nech H^m je množina všetkých m-prvkových podmnožín množiny {1,...,n}. Označme:

$\mathcal{Y}_n^m = \{Y_n^m = (\hat{y}_1, ..., \hat{y}_n) \mid \hat{y}_i = y_i,$ pre $i \in \{1, ..., n\}\backslash h$, $\hat{y}_i \in \mathcal{Y}$ je ľub. pre $i \in h$, $h \in H^m\}$

Ďalej označme $m(T, Y^0)$ najmenšie celé číslo splňujúce

$$sup_{Y_n^m\in\mathcal{Y}_n^m}\left\|T(Y_n^m) - T(Y^0)\right\| = 0.$$

Potom bodom zlyhania funkcionálu T vo výbere Y^0 nazývame číslo

$$\epsilon_n(T, Y^0) = \left(\frac{m(T, Y^0)}{n}\right). \qquad (5)$$

Ak číslo $m(T, Y^0)$ nezávisí na počiatočnom náhodnom výbere Y^0 potom bodom zlyhania budeme rozumieť limitu

$$\epsilon(T) = \lim_{n\to\infty}\epsilon_n(T).$$

Bod zlyhania nám tak určuje minimálny podiel potrebných nových pozorovaní. Vo všeobecnosti požadujeme, aby tento pomer bol čo najväčší (ale maximálne $\frac{1}{2}$). Z príkladu o strednej hodnote a z rovnosti (5) dostávame, že ak je influenčná funkcia neobmedzená, tak bod zlyhania je rovný nule.

Existujú aj iné robustné odhady parametrov v lineárnom regresnom modely ako len *M-odhady*. Ako príklad uveďme *R-odhady* (viď [6]), ktoré sa získajú inverziou testov hypotéz o regresných parametroch založených na poradí (regression ranks). Bližšie sa im budem venovať v kapitole 3.3.

3. Regresné kvantily

3.1.Regresný kvantil ako M-odhad

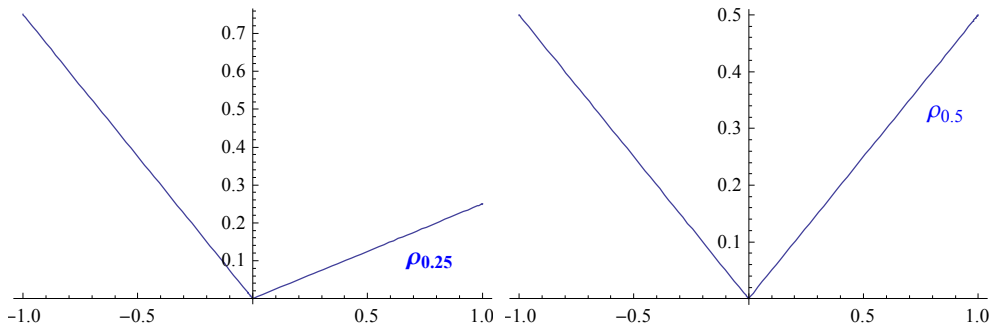
Regresný α-kvantil pre $\alpha \in (0,1)$ potom môžme definovať ako jedno* z riešení minimalizačnej úlohy, kde za stratovú funckiu ρ vezmeme tzv. check function:

$$\rho_\alpha(u) = \begin{cases} \alpha \cdot u & \dots u \geq 0 \\ (\alpha - 1) \cdot u & \dots u < 0' \end{cases}$$

respektíve si ju môžme vyjadriť ako

$$\rho_\alpha(u) = (\alpha - 1) \cdot u \cdot \mathcal{I}_{(-\infty,0)}(u) + \alpha \cdot u \cdot \mathcal{I}_{[0,\infty)}(u) = u \cdot \left(\alpha - \mathcal{I}_{(-\infty,0)}(u)\right).$$

Pre grafické znázornenie si môžme vykresliť stratové funkcie pre časté voľby α:



Obrázok 3. Grafy stratových funkcií ρ pre α=0.25 resp. α=0.5

Minimalizačná úloha tak bude mať tvar

$$\min_{\beta \in R^k} \sum_{i=1}^{n} \rho_\alpha\left(y_i - x_i{}^T \beta\right), \tag{6}$$

čo si môžme charakterizovať ako súčet reziduí prenásobených α v prípade, že sú kladné a 1-α ak sú záporné. Táto minimalizačná úloha má aspoň jedno riešenie, pretože platí

$$\lim_{(y_i - x_i{}^T \beta) \to \infty} \sum_{i=1}^{n} \rho_\alpha\left(y_i - x_i{}^T \beta\right) = \lim_{(y_i - x_i{}^T \beta) \to -\infty^p} \sum_{i=1}^{n} \rho_\alpha\left(y_i - x_i{}^T \beta\right) = \infty.$$

* minimalizačná úloha môže mať viac riešení ako len jedno, keďže minimalizujeme súčet po častiach lineárnych, konvexných, spojitých funkcií a teda tento súčet je po častiach lineárna, konvexná a spojitá funkcia .

3.2.Regresný kvantil ako riešenie úlohy lineárneho programovania

Lineárne programovanie sa zaoberá riešením úloh na viazané extrémy lineárnych funkcií na množinách, ktoré sú určené viazobnými podmienkami v podobe lineárnych rovníc a nerovníc.

Našou úlohou bude nájsť

$$\hat{\beta} = \arg\min_{\beta \in R^k} \sum_{i=1}^{n} \rho_\alpha\left(y_i - x_i{}^T \beta\right).$$

Majme teda mimimalizačnú úlohu

$$\min_{\beta \in R^k} \sum_{i=1}^{n} \rho_\alpha\left(y_i - x_i{}^T \beta\right).$$

Túto úlohu si môžme previesť na úlohu lineárneho programovania

$$\min_{(\beta,\epsilon^-,\epsilon^+)} \left\{ \alpha \cdot 1_n^T \cdot \epsilon^{+T} + (1-\alpha) \cdot 1_n^T \cdot \epsilon^{-T} \;\middle|\; \mathbb{X}\beta + \epsilon^+ - \epsilon^- = \mathbb{y}\,,\ (\epsilon^+, \epsilon^-) \in \mathbb{R}_+^{2n} \right\}. \qquad (7)$$

Zaveďme si označenia:

ϵ^- pre zápornú časť vektoru residua $\mathbb{y}$-$\mathbb{X}\cdot\beta$ a ϵ^+ pre kladnú časť residua $\mathbb{y}$-$\mathbb{X}\cdot\beta$ t.j.

$$\epsilon_i^- := \begin{cases} -\epsilon_i & \dots \epsilon_i < 0 \\ 0 & \dots \epsilon_i > 0 \end{cases};\ \epsilon_i^+ := \begin{cases} \epsilon_i & \dots \epsilon_i > 0 \\ 0 & \dots \epsilon_i < 0 \end{cases} \quad \text{pre } i = 1 \dots n$$

β^- pre zápornú časť vektoru β a β^+ pre kladnú časť vektoru β t.j.

$$\beta_i^- := \begin{cases} -\beta_i & \dots \beta_i < 0 \\ 0 & \dots \beta_i > 0 \end{cases};\ \beta_i^+ := \begin{cases} \beta_i & \dots \beta_i > 0 \\ 0 & \dots \beta_i < 0 \end{cases} \quad \text{pre } i = 1 \dots p$$

Potom môžme úlohu (6) previesť do tvaru

$$\min_{\mathbb{x}} \left\{ \mathbb{c}^T \cdot \mathbb{x} \;\middle|\; \mathbb{A} \cdot \mathbb{x} = \mathbb{y},\ \mathbb{x} \geq 0 \right\},$$

kde:

$$\mathbb{x} = \left(\beta^{+T},\ \beta^{-T},\ \epsilon^{+T},\ \epsilon^{-T}\right),$$
$$\mathbb{c} = \left(0_p^T,\ 0_p^T,\ \alpha \cdot 1_n^T,\ (1-\alpha) \cdot 1_n^T\right),$$
$$\mathbb{A} = (\mathbb{X} \quad -\mathbb{X} \quad I_n \quad -I_n).$$

Majme dvojicu úloh

$$\min \left\{ \mathbb{c}^T \cdot \mathbb{x} \;\middle|\; \mathbb{A} \cdot \mathbb{x} = \mathbb{y},\ \mathbb{x} \geq 0 \right\},$$
$$\max \left\{ \mathbb{b}^T . \mathbb{y} \;\middle|\; \mathbb{A} \cdot \mathbb{y} \leq \mathbb{c} \right\}.$$

Prvú z týchto úloh nazveme primárna, kde lineárna funkcia $\mathbb{c}^T.\mathbb{x}$ predstavuje účelovú funkciu. Každý vektor $\mathbb{x}$ splňujúci obmedzenia primárnej úlohy je prípustným riešením minimalizačnej úlohy. Prípustné riešenie $\mathbb{x}^*$ splňujúce pre každé prípustné riešenie $\mathbb{x}$: $\mathbb{c}^T.\mathbb{x}^* \leq c^T.\mathbb{x}$ nazveme optimálnym riešením minimalizačnej úlohy. Druhá úloha predstavuje duálnu úlohu k primárnej a platí:

***Veta o dualite**: Majme dvojicu vzájomne duálnych úloh. Potom platí práve jedno z nasledujúcich troch tvrdení:*

1. Obe úlohy majú optimálne riešenie a optimálne hodnoty účelových funkcií sú si rovné.

2. Jedna z úloh nemá žiadne prípustné riešenie, druhá úloha má prípustné riešenie, ale nemá optimálne, pretože jej účelová funkcia je na množine prípustných riešení neohraničená.

3. Ani jedna z úloh nemá prípustné riešenie.

Dôkaz: viď [9] Tvrdenie 1.7.

Duálna úloha k našej minimalizačnej úlohe bude mať tvar

$$\max_{\mathbb{b}} \left\{ \mathbb{b}^T \cdot \mathbb{y} \;\middle|\; \mathbb{b}^T \cdot \mathbb{X} = \mathbb{0}, \mathbb{b} \in [\alpha - 1, \alpha]^n \right\} \quad (\mathbb{b}_i = \alpha \vee \mathbb{b}_i = 1 - \alpha),$$

čo si môžme prepísať do tvaru :

$$\max_{\mathbb{a}} \left\{ \mathbb{a}^T \cdot \mathbb{y} \;\middle|\; \mathbb{a}^T \cdot \mathbb{X} = (1-\alpha)\, \mathbb{1}_n^T \cdot \mathbb{X}, \mathbb{a} \in [0, 1]^n \right\}, \text{ kde } \mathbb{a} = \mathbb{b} + (1-\alpha) \cdot \mathbb{1}_n^T. \qquad (8)$$

Takýto tvar duálnej úlohy zodpovedá tvaru úlohy lineárneho programovania pre použitie metódy vnútorných bodov. Špeciálne vhodnou metódou riešenia je použitie Frisch-Newtnovho algoritmu.

Zaveďme si premennú $\mathbb{s}$ splňujúcu :

$$\mathbb{s} = \mathbb{1}_n - \mathbb{a} \quad t.j\, \mathbb{s}_i = 1 - \mathbb{a}_i \text{ pre } i = 1, \ ..., n$$

Teraz preformujeme zadanie našej maximalizačnej úlohy (7), pomocou bariérovej funkcie

$$B(\mathbb{a}, \mathbb{s}, \sigma) = \mathbb{a}^T.\mathbb{y} + \sigma \cdot \sum_{i=1}^{n} (\ln \mathbb{a}_i + \ln \mathbb{s}_i).$$

Získame zadanie v tvare

$$\max_{\mathbb{a}} \left\{ B(\mathbb{a}, \mathbb{s}, \sigma) \;\middle|\; \mathbb{a}^T \cdot \mathbb{X} = (1-\alpha)\, \mathbb{1}_n^T \cdot \mathbb{X}, \mathbb{s} = \mathbb{1}_n - \mathbb{a} \right\}. \qquad (9)$$

Vyriešením tohto problému a následným vypočítaním limity pre $\sigma \to 0$ dostaneme riešenie úlohy (7).

Definícia: Majme funkciu n premenných $f(\mathbb{x}) = f(x_1,...,x_n)$. Gradient ∇f definujeme ako vektor parciálnych derivácii funkcie f vzhľadom k jednotlivým premenným. Ďalej definujeme Hessián $\nabla^2 f$ ako maticu druhých parciálnych derivácii.

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ ..., \ \frac{\partial f}{\partial x_n} \right)$$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Vyjadrením gradientu a Hessiánu Bariérovej funkcie dostaneme:

$$\frac{\partial B}{\partial \mathbb{a}_i} = \mathbb{y}_i + \sigma\left(\frac{1}{\mathbb{a}_i} - \left(\frac{1}{1-\mathbb{a}_i}\right)\right) \qquad \Longrightarrow \nabla B(\mathbb{a}, \mathbb{s}, \sigma) = \mathbb{y} + \sigma\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n$$

kde $\mathbb{A} = \mathrm{diag}(\mathbb{a}_1, \ldots \mathbb{a}_n)$; $\mathbb{S} = \mathrm{diag}(\mathbb{s}_1, \ldots \mathbb{s}_n)$

$$\frac{\partial^2 B}{\partial \mathbb{a}_i \partial \mathbb{a}_j} = \begin{cases} \sigma\left(-\frac{1}{\mathbb{a}_i^2} - \left(\frac{1}{1-\mathbb{a}_i}\right)^2\right) & \text{pre } i = j \\ 0 & \text{pre } i \neq j \end{cases} \Longrightarrow \nabla^2 B(\mathbb{a}, \mathbb{s}, \sigma) = -\sigma\left(\mathbb{A}^{-2} - \mathbb{S}^{-2}\right)$$

V každom iteračnom kroku budeme chcieť nájsť také $\mathbb{p} = \mathbb{a}^{(k+1)} - \mathbb{a}^{(k)}$, ktoré maximalizuje hodnotu

$$B\left(\mathbb{a}^{(k)} + \mathbb{p}, \mathbb{s}^{(k)} - \mathbb{p}, \sigma\right).$$

Pre zjednodušenie budeme odteraz označovať $\mathbb{a}^{(k)}$ len ako $\mathbb{a}$. Vyjadríme Taylorov rozvoj druhého rádu barierovej funkcie v premennej $\mathbb{p}$ ako

$$B(\mathbb{a} + \mathbb{p}, \mathbb{s} - \mathbb{p}, \sigma) = B(\mathbb{a}, \mathbb{s}, \sigma) + \mathbb{p}^T \cdot \nabla B(\mathbb{a}, \mathbb{s}, \sigma) + \frac{1}{2} \cdot \mathbb{p}^T \cdot \nabla^2 B(\mathbb{a}, \mathbb{s}, \sigma) \cdot \mathbb{p}.$$

Podmienku z (9) v tvare

$$(\mathbb{a} + \mathbb{p})^{\mathrm{T}} \cdot \mathbb{X} = \mathbb{a}^{\mathrm{T}} . \mathbb{X} + \mathbb{p}^{\mathrm{T}} \cdot \mathbb{X} = (1 - \alpha)\, 1_{\mathrm{n}}^{\mathrm{T}} \cdot \mathbb{X},$$

prevedieme do tvaru

$$\mathbb{p}^{\mathrm{T}} \cdot \mathbb{X} = 0.$$

Vďaka predpokladu, že $\mathbb{a}$ je prípustné riešenie, tak má platiť

$$\mathbb{a}^T . \mathbb{X} = (1 - \alpha)\, 1_n^T \cdot \mathbb{X}.$$

Z Taylorovho rozvoja, prevedenej podmienky a úlohy (9) tak dostaneme úlohu v tvare

$$\max_{\mathbb{p}} \left\{ B(\mathbb{a}, \mathbb{s}, \sigma) + \mathbb{p}^T \cdot \mathbb{y} + \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n - \frac{1}{2} \cdot \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-2} - \mathbb{S}^{-2}\right) . \mathbb{p} \;\middle|\; \mathbb{p}^T \mathbb{X} = 0 \right\}.$$

Keďže $B(\mathbb{a}, \mathbb{s}, \sigma)$ nezávisí na $\mathbb{p}$ dostávame tvar (nás totiž zaujíma argmax_p)

$$\max_{\mathbb{p}} \left\{ \mathbb{p}^T \cdot \mathbb{y} + \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) . 1_n - \frac{1}{2} \cdot \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-2} - \mathbb{S}^{-2}\right) . \mathbb{p} \;\middle|\; \mathbb{p}^T \cdot \mathbb{X} = 0 \right\}.$$

Vyriešením takéhoto problému a posunutím z $\mathbb{a}^{(k)}$ (naše $\mathbb{a}$) vo výslednom smere $\mathbb{p}$, predstavujúcim smer k hranici množiny prípustných riešení, obdržíme opäť prípustné riešenie a hodnota účelovej funkcie sa nezmenší. Túto úlohu budeme ďalej riešiť prostredníctvom Lagrangerovej metódy neurčitých koeficientov. Nech $\mathbb{d}$ predstavuje vektor Lagrangerových multiplikátorov, potom Lagrangerova funkcia bude mať tvar

$$L(\mathbb{p}, \mathbb{d}) = \mathbb{p}^T \cdot \mathbb{y} + \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n - \frac{1}{2} \cdot \mathbb{p}^T \cdot \sigma\left(\mathbb{A}^{-2} - \mathbb{S}^{-2}\right).\mathbb{p} - \mathbb{p}^T \cdot \mathbb{X} \cdot \mathbb{d}.$$

Položíme parciálne derivácie podľa $\mathbb{p}$ a $\mathbb{d}$ rovné nule a dostávame:

$$\mathbb{y} + \sigma\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n - \sigma\left(\mathbb{A}^{-2} - \mathbb{S}^{-2}\right).\mathbb{p} - \mathbb{X} \cdot \mathbb{d} = 0$$

$$\mathbb{p}^T.\mathbb{X} = 0$$

Predpokladáme, že $\sigma \to 0$ a teda z prvej rovnice dostávame $\mathbb{y} \to \mathbb{X} \cdot \mathbb{d}$ z čoho môžme predpokladať, že $\mathbb{d} \to \beta$.

Položme $\mathbb{W} = (\mathbb{A}^{-2} - \mathbb{S}^{-2})^{-1}$, potom prenásobením prvej rovnice výrazom $\mathbb{X}^T \cdot \mathbb{W}$ a použitím druhej rovnosti za účelom eleminácie $\mathbb{p}$ dostaneme

$$\mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{y} + \sigma \cdot \mathbb{X}^T \cdot \mathbb{W}\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n = \mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{X} \cdot \mathbb{d}.$$

Odkiaľ môžme explicitne vyjadriť vektor Lagrangerových multiplikátorov $\mathbb{d}$ ako

$$\mathbb{d} = \left(\mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{X}\right)^{-1} \cdot \left[\mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{y} + \sigma \cdot \mathbb{X}^T \cdot \mathbb{W}\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n\right]$$

a teda:

$$\hat{\beta} = \lim_{\sigma \to 0} \left(\left(\mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{X}\right)^{-1} \cdot \left[\mathbb{X}^T \cdot \mathbb{W} \cdot \mathbb{y} + \sigma \cdot \mathbb{X}^T \cdot \mathbb{W}\left(\mathbb{A}^{-1} - \mathbb{S}^{-1}\right) \cdot 1_n\right]\right).$$

Čo už sme schopný aritmeticky dopočítať a získať tak $\hat{\beta}$ odhad parametra β lineárneho modelu $Y = \beta^T \cdot \mathbb{X}$ pre α-regresný kvantil.

3.3.Regresné poradové skóre

Koenker a Basset (1978) charakterizovali regresné kvantily ako riešenie parametrickej úlohy lineárneho programovania. Zodpovedajúca množina duálnych úloh pre nich predstavovala len technický nástroj využitý pri výpočte regresných kvantilov. Teraz ukážeme, že duálne riešenia, ktoré budeme nazývať regresné poradia (regression ranks) majú štatistický význam a môžu byť, veľmi prirodzeným spôsobom, použité na definovanie regresných poradových skór (regression rank scores) v lineárnych modeloch, umožňujúc nám tak mnoho aplikácii medzi, ktoré patrí vytvorenie intervalu spoľahlivosti pre β_i a testovanie hypotéz pre β_i (viď [2]).

Majme duálnu úlohu

$$\operatorname{argmax}_{\mathbb{a}} \left\{\mathbb{a}^T \cdot \mathbb{y} \;\middle|\; \mathbb{a}^T \cdot \mathbb{X} = (1 - \alpha)\, 1_n^T \cdot \mathbb{X}, \mathbb{a} \in [0, 1]^n\right\}.$$

Našou úlohou bude skúmať štatistické vlasnosti riešenia $\hat{\mathbb{a}}(\alpha) = \left(\hat{\mathbb{a}}_1(\alpha), \ldots, \hat{\mathbb{a}}_n(\alpha)\right)$ tejto úlohy, ktoré si teraz charakterizujeme prostredníctvom teórie lineárneho programovania. Nech $M = \left\{i \in \{1, \ldots, n\} \mid \mathbb{y}_i = \mathbb{x}_i^T \beta(\alpha)\right\}$, potom môžme $\mathbb{a}(\alpha)$ určiť prostredníctvom nerovníc:

$$\hat{a}_i(\alpha) = \begin{cases} 1 & \text{ak } y_i > x_i^T \beta(\alpha) \\ 0 & \text{ak } y_i < x_i^T \beta(\alpha) \end{cases} \text{ pre } i = 1, \ ..., n,$$

a pomocou lineárnych rovníc:

$$\sum_{i \in M} \hat{a}_i(\alpha) \cdot x_i = (1-\alpha) \cdot \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} I\left[Y_i > x_i^T \beta(\alpha)\right] \cdot x_i.$$

Pripomeňme si, že takéto riešenie existuje na základe *Vety o dualite* vždy, ak existuje riešenie úlohy (6). Ak má navyše "chyba aproximácie" ϵ spojitú distribučnú funkciu, tak toto riešenie je jednoznačné pre každé $\alpha \in (0,1)$. Výpočtu sme sa venovali v predchádzajúcej kapitole.

Teraz si zavedieme tzv. skórovú funkciu (score function) $\varphi:(0,1) \to \mathbb{R}$, o ktorej budeme predpokladať, že má konečnú variáciu a je konštantná mimo intervalu $[\varepsilon, 1-\varepsilon]$ pre nejaké $\varepsilon \in (0,1)$. Pomocou tejto skórovej funkcie budeme definovať skóry $\hat{b}_i$ (regression rank scores) ako

$$\hat{b}_i = -\int_0^1 \varphi(s)\, d\hat{a}_i(s) \qquad i = 1, \ ..., n.$$

kde $\hat{a}_i(s)$, $s \in (0,1)$ je regresné poradie.

Ako príklad si uvedieme klasickú Wilcoxonovu skórovú funkciu centrovanú v 0:

$$\varphi(\alpha) = \alpha - \frac{1}{2},$$

ktorej integrovanie po častiach nám dáva:

$$\begin{aligned} \hat{b}_i &= -\int_0^1 \left(s - \tfrac{1}{2}\right) d\hat{a}_i(s) \\ &= \int_0^1 \hat{a}_i(s)\, s - \tfrac{1}{2} \\ &= \sum_{i=1}^{J} \tfrac{1}{2} \left(\hat{a}_i\left(\alpha_{j+1}\right) - \hat{a}_i\left(\alpha_j\right)\right) - \tfrac{1}{2}, \end{aligned}$$

kde $\alpha_1, \ ... \ \alpha_J$ sú body zlomu funkcie $\hat{a}_i(\alpha)$ *a* $\alpha_{J+1} = 1$.

Pre takúto a podobné voľby skórovej funkcie φ môže byť skonštruovaná testová štatistika T_n, ktorú vieme použiť na skúmanie globálneho efektu regresorov v matici $\mathbb{Z}$ na pozorovanú veličinu Y.

Rozdelíme regresory do dvoch skupín na regresory v matici $\mathbb{X}$ prostredníctvom, ktorých budeme odhadovať pozorovanú veličinu Y a regresory v matici $\mathbb{Z}$, u ktorých budeme skúmať či ich pridanie do regresného modelu malo efekt na odhad pozorovanej veličiny.

Budeme uvažovať model

$$Y = \mathbb{X} \cdot \beta(\alpha) + \mathbb{Z} \cdot \zeta(\alpha) + \epsilon.$$

Splňujúci:

$$Y_i = \beta_1(\alpha) + x_{i_2}\,\beta_2\,(\alpha) + \ldots + x_{i_{k-q}}\,\beta_{k-q} + z_{i_1}\,\zeta_1\,(\alpha) + \ldots + z_{i_q}\,\zeta_q\,(\alpha) + \epsilon_i,$$

$\mathbb{Z} \in \mathbb{R}^{n\times q}$ je matica , ktorej riadky tvoria trojuholníkové pole,

na ktorom budeme testovať nulovú hypotézu $H_0 : \zeta(\alpha) = o$.

Spočítame regresné poradové skóry z "restricted form" (použijeme len maticu $\mathbb{X}$):

$$\text{argmax}_{\mathbb{a}} \left\{ \mathbb{a}^T \cdot \mathbb{y} \;\middle|\; \mathbb{a}^T \cdot \mathbb{X} = (1-\alpha)\,\mathbb{1}_n^T \cdot \mathbb{X},\, \mathbb{a} \in [0,\,1]^n \right\}.$$

Skúmaná testová štatistika má tvar

$$T_n = S_n{}^T\, M_n{}^{-1}\, S_n \,/\, A^2(\varphi), \tag{10}$$

kde:

$$S_n = n^{-1/2}\left(\mathbb{Z} - \hat{\mathbb{Z}}\right)\hat{b},$$

$$\mathbb{Z} - \hat{\mathbb{Z}} = \mathbb{Z} - \mathbb{X}\left(\mathbb{X}^T\,\mathbb{X}\right)^{-1}\,\mathbb{X}^T\,\mathbb{Z},$$

$$M_n = \left(\mathbb{Z} - \hat{\mathbb{Z}}\right)\left(\mathbb{Z} - \hat{\mathbb{Z}}\right)^T \Big/ n,$$

$$\hat{b} = \int_0^1 \hat{\mathbb{a}}(s)\,d\varphi(s) \;\; = \left(\hat{b}_1,\; \ldots\, \hat{b}_n\right),$$

$$A^2(\varphi) = \int_0^1 \left(\varphi(s)^2 - \int_0^1 \varphi(s)\,ds\right)^2 ds.$$

$\mathbb{Z} - \hat{\mathbb{Z}}$ predstavuje reziduá vzniklé odhadom $\mathbb{Z}$ prostredníctvom metódy najmenších štvorcov vykonanej na dátach $\mathbb{X}$.

Pri platnosti vyššie uvedených predpokladov a nulovej hypotézy má T_n limitné rozdelenie χ^2_q, kde q predstavuje počet zložiek parametra ζ. Všimnime si, že takto navrhnutý test za platnosti nulovej hypotézy nie je vôbec závislý na distribučnej funkcii F "chyby aproximácie" ϵ a ani na hodnotách β.

Dôležitou aplikáciou teórie testovania založenej na regresných poradových skóroch je vytváranie intervalov spoľahlivostu pre jednotlivé parametre kvantilového regresného modelu s konkrétnym kvantilom α. Štatistiku T_n budeme vytvárať pomocou skórovej funkcie φ, ktorá sa bude exklusívne sústrediť na jednu hodnotu α (chceme otestovať vplyv regresorov na hodnotu pozorovaní Y len v časti populácie zodpovedajúcej α).

Budeme uvažovať model

$$Q_{Y_i}(\alpha \mid x_i, z_i) = x_i{}^T\,\beta(\alpha) + z_i{}^T\,\zeta(\alpha).$$

Našim cieľom bude určiť, či vektor regresorov z_i má vplyv na sledovanú náhodnú veličinu Y pre zvolený α-regresný kvantil.

Testujeme hypotézu $H_0 : \zeta = \zeta_0$ (špeciálny prípad, ktorý nás najčastejšie bude zaujímať je ζ=0), tento test bude založený na regresných poradiach, ktoré môžme získať ako riešenie úlohy

$$\text{argmax}_{\mathfrak{a}} \left\{ \mathfrak{a}^T \cdot (\mathfrak{y} - \mathfrak{z}\zeta_0) \;\middle|\; \mathfrak{a}^T \cdot \mathbb{X} = (1-\alpha)\, I_n^T \cdot \mathbb{X},\, \mathfrak{a} \in [0,\, 1]^n \right\}.$$

Skórová funkcia

$$\varphi_\alpha(s) = \alpha - I(s < \alpha) \tag{11}$$

nám umožňuje zamerať sa počas testu výhradne na α-kvantil.

Špeciálny prípad, pre medián majúci tvar

$$\varphi_{1/2}(t) = \frac{1}{2}\,\text{sgn}\left(t - \frac{1}{2}\right),$$

je dôležitou aplikáciou, ktorá sa často označuje ako sgn-skóre.

Použitím skórovej funkcie (11) vypočítame skóre ako

$$\hat{b}_i = -\int_0^1 \varphi(s)\, d\hat{\mathfrak{a}}_i(s) = \hat{\mathfrak{a}}_i(\alpha) - (1-\alpha) \qquad i = 1,\ \ldots,\, n.$$

Pri platnosti H_0 platí

$$S_n(\zeta_0) = n^{-1/2}\left(Z - \hat{Z}\right)^T \hat{b} \xrightarrow{\mathcal{P}} S \sim N\left(0,\, A^2(\varphi_\alpha) \cdot q_n^2\right),$$

kde:

$$q_n^2 = n^{-1}\, Z^T \left(I - X\left(X^T X\right)^{-1} X^T\right) Z,$$

$$\hat{b} = \left(\hat{b}_1,\ \ldots,\, \hat{b}_n\right),$$

$$A^2(\varphi_\alpha) = \int_0^1 \left(\varphi_\alpha(s) - \int_0^1 \varphi_\alpha(s)\, ds\right)^2 ds = \alpha \cdot (1-\alpha).$$

Potom platí ,že

$$T_n(\zeta_0) = S_n(\zeta_0) \Big/ \left(A(\varphi_\alpha)\, q_n\right) \xrightarrow{\mathcal{P}} T \sim N\left(0,\, A^2(\varphi_\alpha) \cdot q_n^2\right).$$

Hypotézu H_0 zamietame, ak $|T_n(\zeta_0)| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$.

4. Aplikácie kvantilovej regresie

4.1. Aplikácia prevedená na vygenerovaných dátach

Pre lepšie pochopenie, toho ako funguje kvantilová regresia, si ukážeme použitie štatistických nájstrojov z nej vychádzajúcich na jednoduchých štatistických dátach, ktoré si sami vygenerujeme takým spôsobom, aby výsledky mohli byť jednoducho graficky prezentované. Výpočty prevedieme v štatistickom programe ℝ (viď [10]), niektoré pomocné výpočty si predpripravíme v matematickom programe *Mathematica*.

Zvolíme si regresný model s dvoma regresormi:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i \quad , \text{ pre } i = 1, \ ..., n. \tag{12}$$

Teraz si vhodne zvolíme hodnoty skutočných regresných koeficientov a to takým spôsobom, aby sme si na nich mohli prezentovať rôzne výsledky pri testovaní hypotéz významnosti jednotlivých regresorov (viď predchádzajúcu kapitolu).

$$(\beta_0, \beta_1, \beta_2) = (1, 2, 0.01)$$

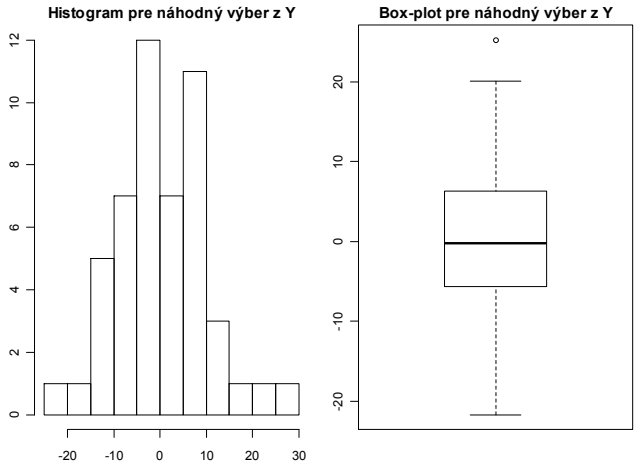
Počet pozorovaní n si zvolíme 50 a následne si vhodne vygenerujeme hodnoty regresorov a chýb aproximácie:

- hodnoty regresora x_i budú predstavovať náhodný výber z normálneho rozdelenia $\mathcal{N}(0,5)$
- hodnoty regresora z_i budú predstavovať náhodný výber z rovnomerného rozdelenia $\mathcal{R}(-1,1)$
- hodnoty chýb aproximácie ϵ_i predstavujú náhodný výber z normálneho rozdelenia $\mathcal{N}(0,1)$

Teraz si prostredníctvom takto vygenerovaných hodnôt a ich vložením do modelu (12) vypočítame príslušné hodnoty $y_1, \ ..., y_n$ skúmanej veličiny Y.

Bližší náhľad na získané hodnoty y_i:

```
   Min.   1st Qu.    Median      Mean   3rd Qu.    Max.
-21.67   -5.5250   -0.2492    0.3470    6.2730   25.2100
```



Obrázok 4. Histogram a boxplot vygenerovaných hodnôt y_i

Ešte pred samotnou regresnou analýzou si zvolíme, čo najvhodnejší model, na ktorý ju budeme aplikovať. Na začiatku uvažujeme náš pôvodný regresný model (12)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i\,, \text{ pre } i = 1,\ ..., 50.$$

Následne sa budeme snažiť určiť na ktorých regresoroch je skúmaná veličina závislá a naopak, ktoré regresory môžme za istých okolností z modelu vylúčiť. To môžeme docieliť viacerými metódami:

- **Procedúra stepwise**

Táto metóda minimalizuje súčet reziduálnych štvorcov prostredníctom pridávania, alebo odoberania premenných spôsobom, ktorý zohľadňuje štatistickú významnosť všetkých premenných v modely. Táto stepwise procedúra bude kombinovať prvky napredovacieho (forward) výberu a spätnej (backward) eliminácie, keď po každom eliminačnom štádiu bude nasledovať kontrola pre možné pridanie. Ako počiatočný štatistiský objekt reprezentujúci model, ku ktorému procedúra začne pridávať regresory, zvolíme objekt vzniklý kvantilovou regresnou analýzou, ktorá skúmanú veličinu modeluje len pomocou konštanty. Výpočet regresných koeficientov prebieha prostredníctvom Frisch-Newtnovho algoritmu, ktorého priebeh sme si ukázali v predchádzajúcej kapitole. Procedúre umožníme pridávať regresory obsiahnuté v našom modely (12).

```
model0<-y~x+z
model<-step(rq(y~1,method="fn"),scope=model0,trace=0)$formula
```

procedúra nám vráti:

```
Call:
rq(formula = y ~ x, method = "fn")

Coefficients:
(Intercept)           x
   1.108684    2.025940

Degrees of freedom: 50 total; 48 residual
```

Poznamenajme, že program ℝ vo výstupnom modely neuvádza konštantu, ktorá je však v modely obsiahnutá. Procedúra navrhuje vynechať z pôvodného modelu regresor *Z*. Model navrhnutý stepwise procedúrou, ako vhodný na regresnú analýzu, tak má tvar

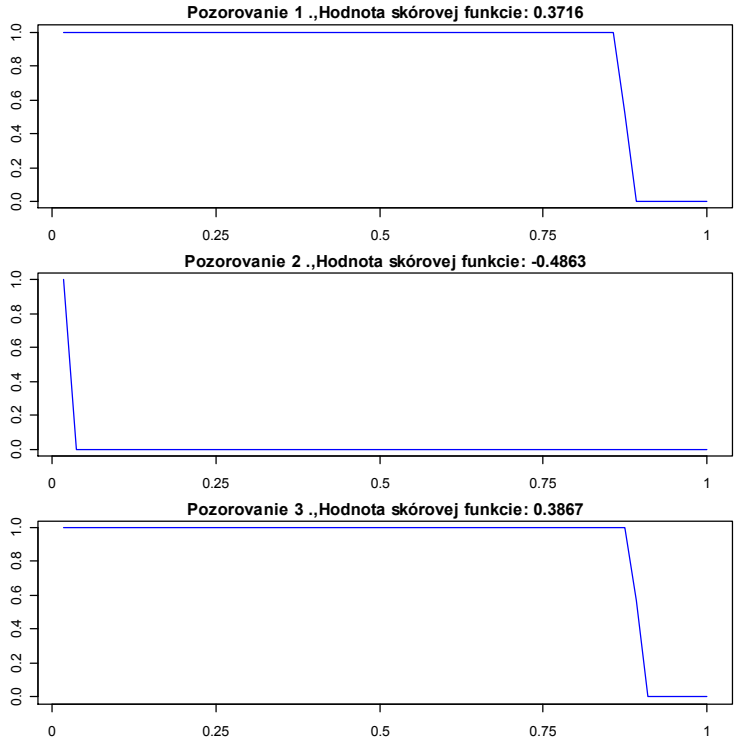
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \text{ pre } i = 1,\ ..., 50. \tag{13}$$

- **Testovanie hypotéz významnosti**

Tento test bude založený na regresných poradiach a pre príslušné "restricted form" nášho modelu ($Y \sim X$ resp. $Y \sim Z$). Pre lepšie pochopenie toho, čo vlastne sú regresné poradia si ich môžme v programe ℝ vyjadriť pre "restricted form" $Y \sim X$ použitím príkazu:

```
model<-y~x
regres2<-rq(model,tau=-1,method="br")$sol
```

a následne graficky znázorniť :



Obrázok 5. Grafy regresných poradových skór prvých troch pozorovaní pre $y \sim x$

Regresné poradové skóry $\hat{b}$ z príslušnej "restricted form" $Y \sim Z$, využité pri výpočte významnosti regresora X, vypočítame v programe ℝ použitím príkazu

```
skore1<-ranks(rq(y~z,tau=-1,method="br"),score="wilcoxon").
```

Analogicky vypočítame regresné skóry pre "restricted form" $Y \sim X$:

```
skore2<-ranks(rq(y~x,tau=-1,method="br"),score="wilcoxon")
```

Ako skórovú funkciu sme zvolili Wilcoxonovu funkciu, ktorá má tvar

$$\varphi(t) = 2\,t - 1 \quad ; \; 0 < t < 1.$$

Teraz použijeme testovú štatistiku (10) založenú na regresných poradových skórach

$$T_n = S_n{}^T \, M_n{}^{-1} \, S_n \,/\, A^2(\varphi)\,, \text{ pre } n = 1, 2,$$

o ktorej na základe teórie uvedenej v kapitole 3.3 predpokladáme, že má χ^2_1 rozdelenie.

Hodnoty testovej štatistiky a ich *p-hodnoty* si uvedieme v tabuľke:

Regresor	*Testová štatistika* T_n	*p-hodnota*
X	25.1642	$5.265054\,\mathrm{e}-07$
Z	0.00372309	0.9513456

Tabuľka 1. Testy závislosti na jednom regresore pre TP

Z tabuľky vidíme, že na 5% hladine významnosti nezamietame nevýznamnosť regresora Z. Rovnako ako stepwise procedúra teda aj testovanie hypotéz významnosti navrhuje vynechať z pôvodného modelu regresor Z.

Ako optimálny model pre regresnú analýzu nám tak v oboch prípadoch vyšiel model v tvare

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \text{ pre } i = 1, \ ..., 50.$$

Na tento model (13) teda budeme aplikovať kvantilovú regresiu prevedenú v kvartiloch použitím príkazu

```
regresory <- summary(rq(model, tau = c(0.25, 0.5, 0.75))).
```

Výsledky odhadov regresných koeficientov spolu s 95 % intervalmi spoľahlivosti si uvedieme v tabuľke:

Regr. koeficient	*0.25*	*0.5*	*0.75*
β_0	0.25781 $(-0.01760, 0.58221)$	1.10868 (0.66501,1.41321)	2.24806 (1.58247,2.38872)
β_1	1.98903 (1.89188,2.07814)	2.02594 (1.90321,2.08260)	1.96722 (1.92740,2.09957)

Tabuľka 2. Výsledky odhadov regresných koeficientov kvantilovou regresiou pre TP

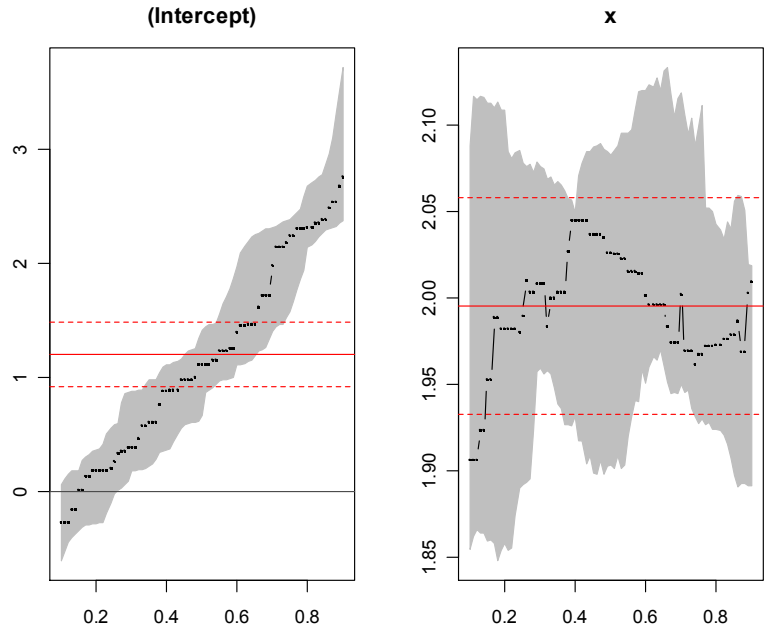
Poznamenajme, že regresný koeficient β_2 volíme na základe testov o významnosti regresorov nulový vo všetkých kvantiloch.

Pre úplnosť si ešte uvedieme odhady regresných koeficientov získané metódou najmenších štvorcov:

Regr. koeficient	*MNČ*
β_0	1.20793
β_1	1.99619
β_2	-0.04609

Tabuľka 3. Výsledky odhadov regresných koeficientov MNČ pre TP

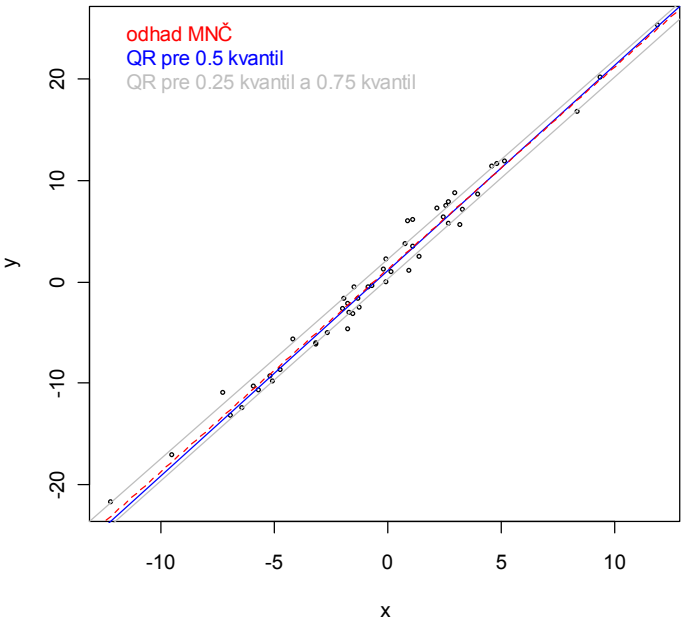
Zmeny regresných koeficientov v závislosti na jednotlivých kvantiloch si znázornime graficky:



Obrázok 6. Odhady regresných koeficientov pre TP

Červené konštantné priamky predstavujú odhad koeficientov metódou najmenších štvorcov a 90% interval spoľahlivosti. Šedá oblasť predstavuje 90 % intervaly spoľahlivosti pre jednotlivé regresné koeficienty v závislosti od kvantilov.

Následne si vykreslíme odhady pre regresné kvartily a porovnáme ich s odhadom získaným metódou najmenších štvorcov.



Obrázok 7. Kvantilové priamky pre TP

Teraz predvedieme niektoré aplikácie regresných kvantilov pri testovaní hypotéz. Prirodzenou otázkou je, či odhadnuté kvantilovo regresné vzťahy potvrdzujú tzv. hypotézu o posunutí polohy (location shift hypothesis), ktorá predpokladá, že všetky kvantilové funkcie majú rovnaké stĺpcové parametre. Jedná sa o test zhody smerníc a teda testujeme platnosť

$$H_0 : \beta_i(\tau) = \alpha_i \text{ pre } i = 1, 2 ..., n.$$

Dôsledkom platnosti tejto hypotézy by bola lineárna závislosť premennej Y na X.

Ukážeme si ako si overiť platnosť tejto hypotézy v kvartiloch. Prevedieme tri procesy odhadovania (fits):

```
fit1 <- rq(y ~ x, tau = 0.25),
fit2 <- rq(y ~ x, tau = 0.5),
fit3 <- rq(y ~ x, tau = 0.75),
```

na ktoré použijeme funkciu anova (viď [10] help pre funkciu anova.rq) :

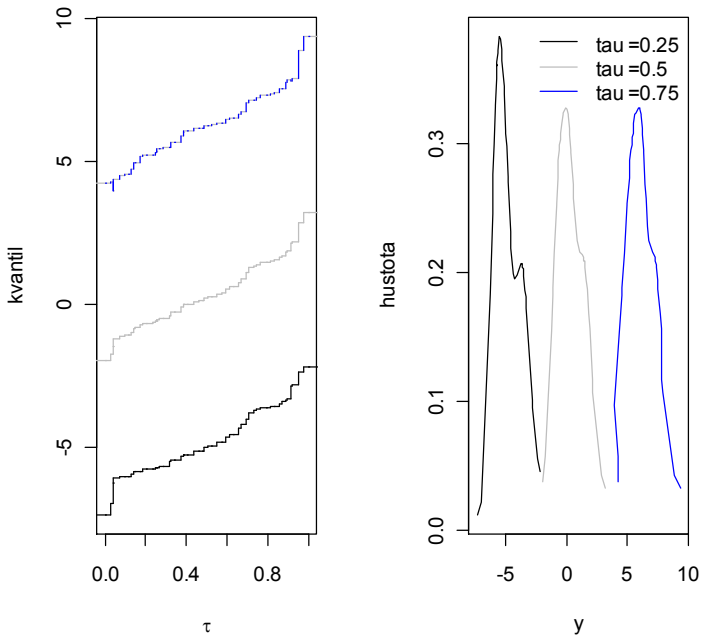
```
anova(fit1,fit2,fit3),
```

ktorá nám vráti tabuľku analýzy rozptylov :

```
Quantile Regression Analysis of Deviance Table
Model: y ~ x
Joint Test of Equality of Slopes: tau in {  0.25 0.5 0.75  }
  Df Resid Df F value Pr(>F)
1  2      148  0.6272 0.5355
```

Na základe výsledkov tohto testu nemôžeme zamietnuť platnosť hypotézy H_0 a teda ani hypotézu o lineárnej závislosti skúmanej premennej Y na X.

Pre lepšie grafické pochopenie vykreslíme:



Obrázok 8. Odhadnuté podmienené kvantilové a hustotové funkcie pre Y

Tento test nám neumožnil zamietnuť hypotézu H_0 a preto skúsime previesť dôslednejší test, ktorý je založený na regresných poradiach a berie do úvahy hodnoty vo viacerých kvantiloch rozdelenia Y (nielen v kvartiloch):

```
KhmaladzeTest(y ~ x, taus = -1, nullH = "location")
```

výsledkom ktorého je:

```
$nullH
[1] "location"

$Tn
[1] 1.035798
```

Po dohľadaní v tabuľke obsahujúcej kritické hodnoty tohto testu dostávame, že 1% kritická hodnota pre naše dáta ($p = 2$,$\epsilon = 0.05$) je rovná 4,199 (viď [1] tab. B2) a teda ani týmto testom nezamietame hypotézu o zmene polohy a tak nemôžme vylúčiť, že Y je lineárne závislé na X a teda nami navrhovaná voľba modelu

$$Y = \beta_0 + \beta_1 \cdot X$$

sa javí ako vhodná.

4.2. Aplikácia prevedená na reálnych dátach

Praktické použitie metód kvantilovej regresie si ukážeme na príklade, v ktorom budeme na reálnych dátach zozbieraných v rokoch 1993-94 z 1283 amerických univerzít, určovať závislosť výšky školného príslušnej univerzity, na premenných uvedených v tabuľke 3. Túto závislosť sa pokúsime slovne interpretovať a zdôvodniť vývoj regresných koeficientov jednotlivých premenných v závislosti na kvantile τ. Taktiež overíme, či bolo vhodné voliť metódy lineárnej regresnej analýzi a zameriame sa na to ako sa mení závislosť školného na svojich regresoroch v závislosti od toho, či je škola štátna, alebo súkromná .

Datový súbor bol získaný z dvoch zdrojov a to :

1) U.S. News & World Report's 1995 Guide to Americas Best Colleges
2) AAUP's (American Association of University Professors) 1994 Salary Survey

Názov premennej	*stručná charakteristika*
tuition	výška vysokoškolského školného v $
pcttop25	percento novoprijatých študentov patriacich k 25% najlepších zo SŠ
sf_ratio	pomer členov fakulty k študentom
fac_comp	priemerná kompenzácia školného
accrate	podiel prijatých študentov k prihláseným
graduat	percento študentov, ktorí úšpešne doštudujú
pct_phd	percento vyučujúcih s Ph.D
fulltime	percento študujúcich na dennom štúdiu
alumni	percento absolventov univerzity, ktorí jej prispievajú
num_enrl	počet novoprijatých študentov
public/private	premenná určujúca, či je škola štátna , alebo súkromná (štátna=0 ,súkromná =1)

Tabuľka 4. Charakteristiky univerzít

Uvažujeme regresný model v tvare

$$tuition = \beta_0 + \beta_1\, pcttop25 + \beta_2\, sf\_ratio + \beta_3\, fac\_comp + \beta_4\, accrate + \beta_5\, graduat + \beta_6\, pct\_phd + \beta_7\, fulltime + \beta_8\, alumni + \beta_9\, num\_enrl + \beta_{10}\, public/private. \quad (14)$$

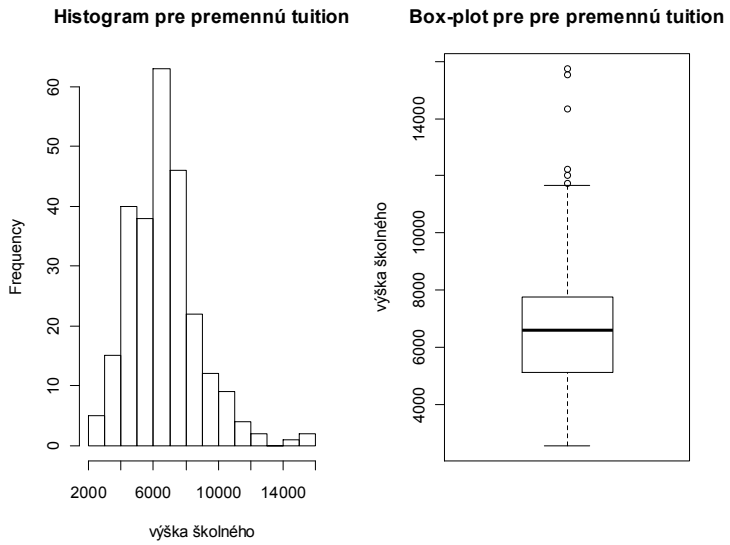
Náš datový súbor obsahuje informácie z 1283 amerických univerzít, avšak nie ku každej univerzite sú uvedené kompletné údaje . Použitím príkazu

```
data<- dataset[complete.cases(dataset), ]
```

vyradíme zo súboru univerzity, ku ktorým neboli dodané kompletné údaje pre všetky premenné. Takto vzniklý datový súbor obsahuje 804 univerzít s údajmi pre všetky premenné.

Bližší náhľad na hodnoty hodnoty výšky školného *tuition* :

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2340    7146    9842   10290   12850   25750
```



Obrázok 9. Histogram a boxplot hodnôt *tuition*

Z boxplotu vidíme, že náhodný výber obsahuje viaceré odľahlé pozorovania .

Ešte skôr, než posúdime odhady regresných koeficientov $\beta_i, i = 1, \ldots 10$, otestujeme jednotlivé regresory na nevýznamnosť. K tomuto účelu vyberieme jednu z metód, ktoré sme si predviedli v teoretickom príklade. Ako najvhodnejší sa pri väčšom počte regresorov javí postup pomocou stepwise procedúry, ktorej priebeh sme si popísali v 4.1. Na začiatku uvažujme náš pôvodný regresný model (14) pomocou stepwise procedúry sa budeme snažiť vybrať jeho pre regresnú analýzu čo najvhodnejší submodel.

Výpočet v programe R:

```
max.model<-as.formula(paste("tuition~",
paste(names(+ dataset[c(2:11)]),collapse="+")))
model<-step(rq(tuition~1,method="fn"),scope=max.model,trace=0)
```

Navrhnutý submodel má tvar:

$$tuition = \beta_0 + \beta_1\, sf\_ratio + \beta_2\, fac\_comp + \beta_3\, graduat + \beta_4\, pct\_phd + \beta_5\, fulltime + \beta_6\, alumni + \beta_7\, public\,/\,private$$

Na tento model aplikujeme príkaz:

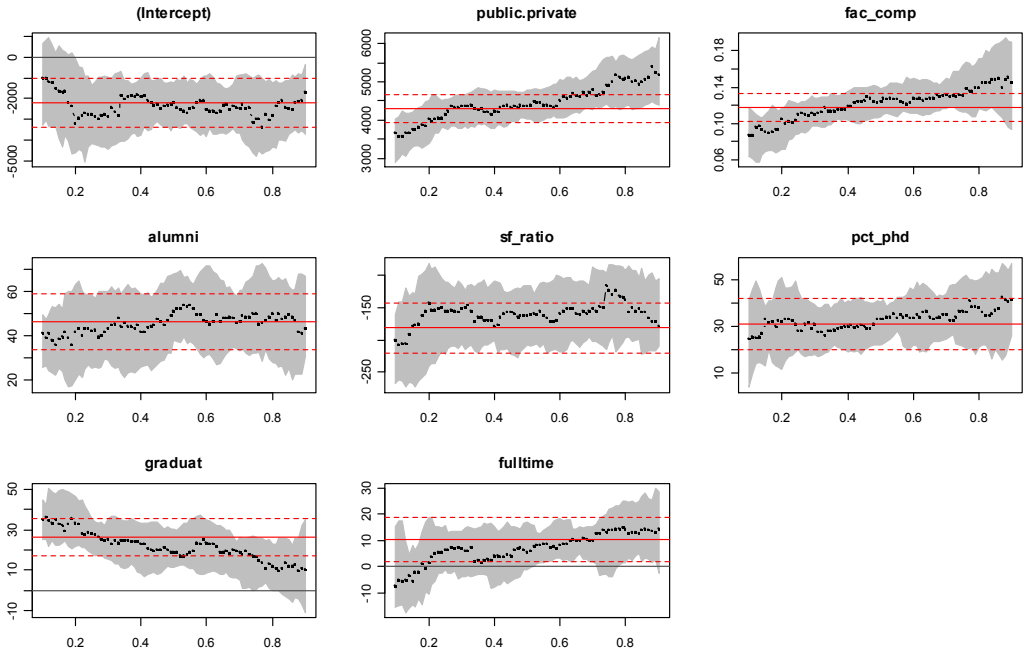
```
regresory <- summary(rq(model, tau = c(0.25, 0.5, 0.75)))
```

Výsledky odhadov regresných koeficientov pre najvýznamnejšie kvantily spolu s 95 % intervalmi spoľahlivosti si uvedieme v tabuľke:

Regresný koeficient	***0.25***	***0.5***	***0.75***
β_0	−2774.26811 (−4149.52193, −1351.57218)	−2221.79773 (−3337.04237,−1239.51208)	−2976.22398 (−4756.54820,−905.21631)
β_1	−155.36114 (−223.05887−111.13941)	−163.30405 (−208.00637,−117.28089)	−123.47423 (−204.37742,−84.28445)
β_2	0.10948 (0.08882,0.12202)	0.12382 (0.10803,0.13873)	0.13053 (0.11300,0.16339)
β_3	27.91620 (18.34451,35.18609)	18.60247 (11.29319,30.04818)	16.67539 (0.25042,26.30793)
β_4	27.93436 (19.98321,45.64387)	33.05932 (22.14588,40.67896)	36.45428 (21.41629,51.06758)
β_5	6.28444 (−1.42940 ,15.53886)	5.44838 (−1.42445 ,16.97730)	13.71562 (−3.87215,21.62930)
β_6	42.17818 (25.28324,58.31736)	51.31407 (37.96737,68.00181)	50.12455 (31.98825,71.62279)
β_7	4145.92048 (3785.38198,4546.09932)	4359.04528 (4033.89952,4773.58586)	4909.94529 (4371.58425,5629.10246)

Tabuľka 5. Výsledky odhadov regresných koeficientov pre PP

To ako sa menia regresné koeficienty v závislosti na jednotlivých kvantiloch pre výšku školného si znázorníme graficky :



Obrázok 10. Odhady regresných koeficientov pre príklad s reálnymi dátami

Teraz sa pokúsime slovne interpretovať výsledy kvantilovej regresnej analýzi založené na grafoch obsiahnutých v obrázku (14). Pripomeňme, že niektoré premenné boli pri vyberaní najvhodnejšieho submodelu z regresnej analýzi vyradené a tak ich regresné koeficienty uvažujeme rovné nule pre všetky kvantily.

- **Regresor : public.private , regresný koeficient : β_7**
  premenná určujúca,či je škola štátna,alebo súkromná štátna=0,súkromná=1

Školné v súkromných školách je, už z výsledkov získaných metódov najmenších štvorcov, očividne vyššie než v školách štátnych a to v priemere o 4359 dolárov. Regresná analýza nám však navyše poodhaľuje, že tento rozdiel je oveľa nižší pre nízke kvantily a podstatne vyšší na opačnom chvoste rozdelenia.

- **Regresor : fac_comp , regresný koeficient : β_2**
  priemerná kompenzácia školného

Regresor sa javí na základe porovnaní regresných koeficientov ako málo vyznamný, musíme však vziať do úvahy, že kým hodnoty ostatných regresorv predstavujú percentá prípadne nadobúdajú hodnoty {0,1}, tak tento regresor nadobúda rádovo vyššie hodnoty a tak je jeho vplyv aj pri nízkom regresnom koeficiente na výšku školného značný.Tento vplyv sa s narastajúcim školným postupne zvyšuje.

- **Regresor : alumni, regresný koeficient : β_6**
  percento absolventov univerzity, ktorí je prispievajú

Vplyv tohto regresora nám osciluje okolo priemeru získaného metódou najmenších štvorcov.

- **Regresor : sf_ratio , regresný koeficient : β_1**
  pomer členov fakulty *k* študentom

Pomer študentov k vyučujúcim má výrazný negatívny dopad na výšku školného. Tento negatívny dopad sa trochu prekvapujúco s rastúcim kvantilom školného znižuje, čo si ale môžme vysvetliť pomocou regresora udavajúceho percento vyučujúcich s Ph.D., ktorý v závislosti na školnom stúpa a tak môže aj menší počet členov fakulty zabezpečiť kvalitné vzdelanie. Pre vysoké kvantily sa tento trend obracia a negatívny dopad na školné sa zvyšuje.

- **Regresor : pct_phd, regresný koeficient : β_4**
  percento vyučujúcih *s* Ph.*D*

Vplyv tohto regresoru je kladný a postupne s rastúcim školným pomerne významne rastie. To môže byť spôsobené tým, že u škôl s vysokým pct_phd je nízke sf_ratio a výdaje na členov fakulky s Ph.D predstavujú vysoké náklady pre univerzitu, ktoré musia byť kompenzované výškou školného.

- **Regresor : graduat , regresný koeficient : β_3**
  percento študentov,ktorí úšpešne doštudujú

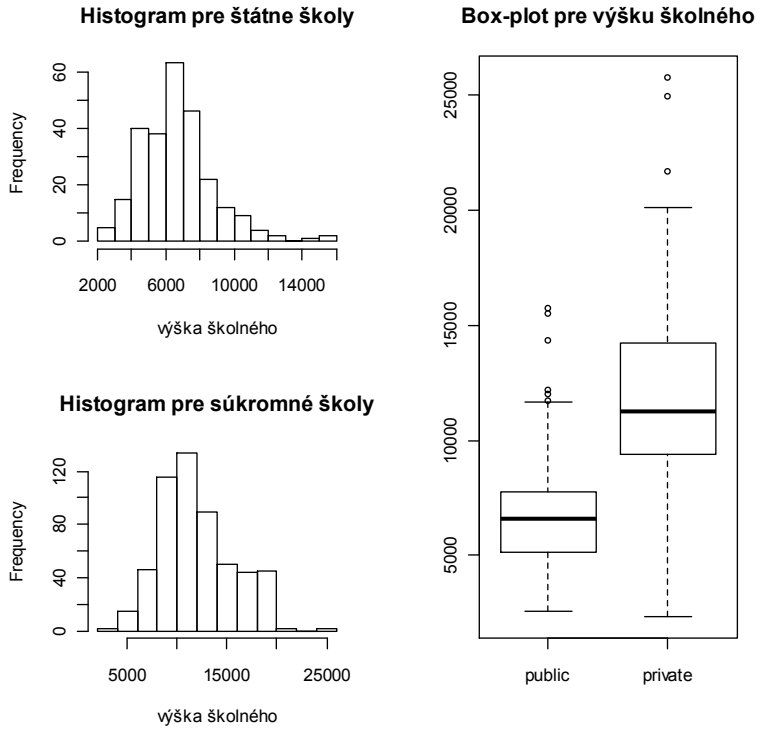
Vplyv tohto regresoru postupne s výškou školného klesá, čo si môžme vysvetliť tým, že školy poskytujúce kvalitnejšie vzdelanie, a tým pádom majúce vyššie školné, sú pre študentov náročnejšie, čo znižuje podiel študentov, ktorí školu úspešne doštudujú. Avšak to týchto študentov natoľko neovplyvňuje pretože táto náročnosť zvyšuje prestíž školy a takto získaného titulu a teda pozitívny vplyv podielu doštudovaných študentov je zmenšený úbytkom prestíže.

- **Regresor : fulltime, regresný koeficient : β_5**
  percento študujúcich na dennom štúdiu

Vplyv tohto regresoru nám postupne s výškou školného rastie , a kým u škôl z nízkych kvantilov môže byť tento vplyv dokonca záporný, pretože tieto školy s predpokladanou nižšou kvalitou sú zamerané na externých študentov, tak u vyšších kvantilov je vplyv podielu denných študentov na výšku školného už kladný.

Pri skúmaní výšky školného sa vynára otázka, či a o koľko sa lýši školné v súkromný a štátnych školách a či sa tento rozdiel odráža aj v hodnotách charakterizujúcich premenných. Našim cieľom bude určiť ako sa lýši vplyv charakterizujúcich premenných v jednotlivých kvantiloch v závislosti od toho, či sa jedná o školu štátnu alebo súkromnú.

Najprv si graficky znázornime hodnoty výšky školného rozdelené podľa toho, či sa jedná o štátne, alebo súkrómne školy:



Obrázok 11. Histogramy a boxploty pre hodnoty *tuition* rozdelenej podľa premennej public/private

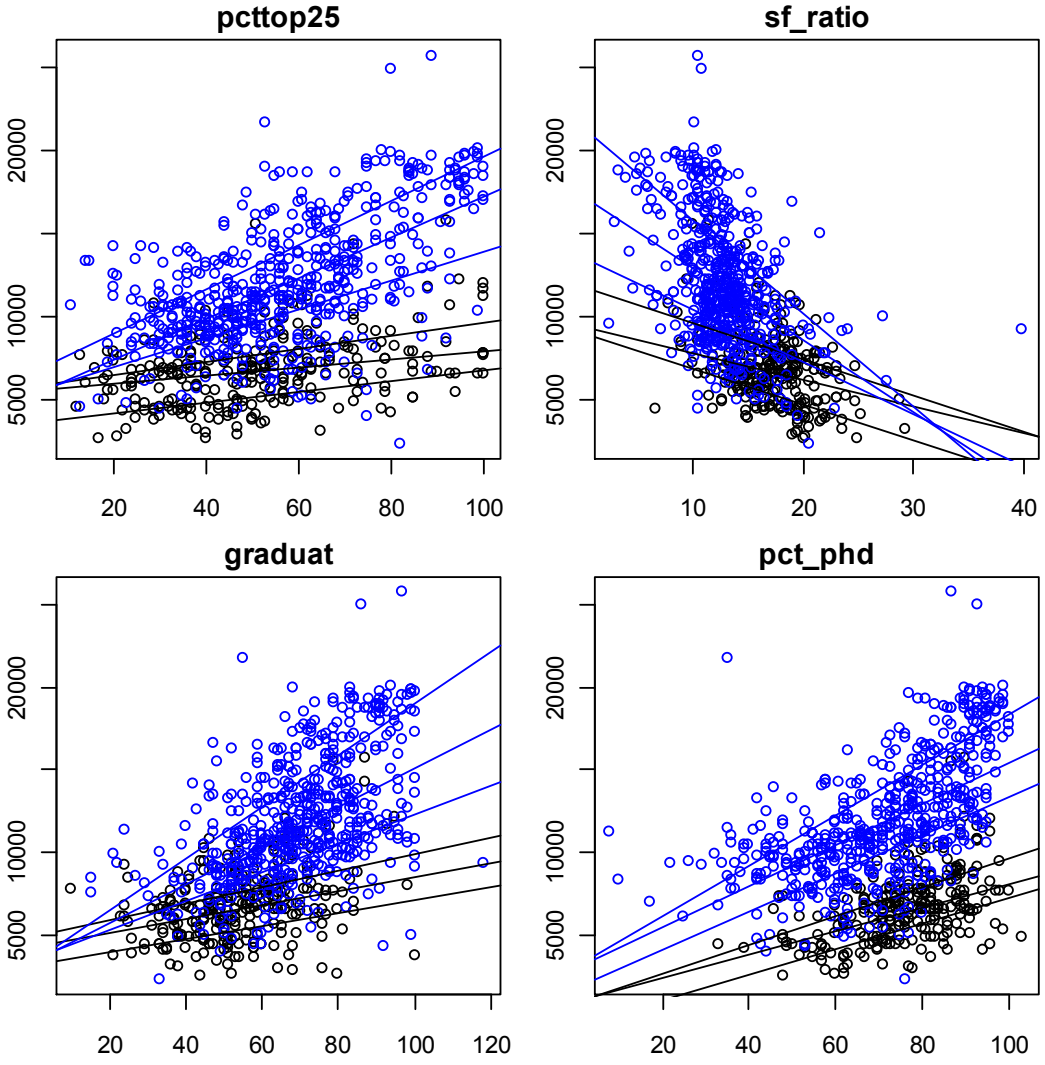
Z grafov vidíme ,že školné v súkromných školách:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2340    9400   11280   12000   14200   25750
```

je zvyčajne vyššie ako v školách štátnych:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2580    5130    6594    6698    7765   15730
```

Teraz si vykreslíme grafy závyslosti premennej *tuition* na vybraných charakterizujúcich premenných (*pcttop25*, *sf_ratio*,*graduat*, *pct_phd,*), od ktorých vo všeobecnosti očakávame, že popisujú kvalitu vzdelania poskytnutého vysokou školou. Data rozdelíme do dvoch skupín podľa toho, či sa jedná o školu štátnu (čierna farba), alebo školu súkromnú (modrá farba) a do grafov zakreslíme regresné kvantily v kvartiloch pre príslušné skupiny.



Obrázok 12. Vykreslené závyslosti *tuiotion* na vybraných regresoroch

Z grafov jednoznačne vyplýva, že v súkromných školách majú tieto regresory na stanovenie výšky školného väčší efekt ako v školách štátnych.

Na záver prevedieme opäť test hypotézy o zmene polohy (location shift), ktorou si overíme, či bolo vhodné modelovať výšku školného lineárnym regresným modelom.

Budeme overovať pre jednotlivé regresory ($i = 1, 2 ..., 7$) hypotézu

$$H_0 : \beta_i(\tau) = \alpha_i$$

a následne spoločnú hypotézu pre všetky regresory :

$$H_0 : \beta_i(\tau) = \alpha_i \text{ pre } i = 1, 2 ..., 7.$$

To dosiahneme použitím príkazu v ℝ

```
KhmaladzeTest(model, taus = -1, nullH = "location").
```

Výsledok testu si uvedieme v tabuľke:

Premenná	*T*
public.private	1.6879474
fac_comp	2.0102988
alumni	0.4734559
sf_ratio	1.0721061
pct_phd	0.5140671
graduat	0.5944883
fulltime	0.7793660
spoločný efekt	**6.523212**

Tabuľka 6. Testy hypotézy o zmene polohy

Po dohľadaní v tabuľke obsahujúcej kritické hodnoty tohto testu dostávame, že 5% kritická hodnota pre spoločný efekt ($p = 8$, $\epsilon = 0.05$) je rovná 9,552 (viď [1] tab. B2) a teda nezamietame hypotézu o zmene polohy a teda nemôžme vylúčiť, že Y je lineárne závislé na vektore $\mathbb{X}$ a teda voľba lineárnej regresie sa javí ako vhodná a ani štatistiky pre jednotlivé kovariáty nie sú zamietajúce na 5% hladine.

Záver

V tejto bakalárskej práci sú obsiahnuté teoretické základy, z ktorých bola odvodená teória regresných kvantilov. Snažili sme sa o ucelenejší pohľad na teóriu kvantilovej regresie, z pohľadu začlenenie odhadov v nej vzniklých do rámca M-ohdadov a ukázali sme jej súvis s metódami klasickej regresie, špeciálne s odhadmi prostredníctvom metódy najmenších štvorcov. Poukázali sme na jeden z prístupov v lineárnom programovaní pre výpočet regresných koeficientov a predviedli sme odvodenie jeho algoritmu (Frisch-Newtnov algoritmus). Ukázali sme si štatistické použitie regresných poradí, vzniklých pri výpočte duálnej úlohy k úlohe pre výpočet regresných koeficientov, a následne sme ho demonštrovali v praktickej časti. V tej sme ilustrovali znalosti nadobudnuté v teoretickej časti na dvoch príkladoch, ktoré sme volili takým spôsobom, aby z nich bolo jasne vyditelné, aký potenciál má použitie metód kvantilovej regresie a nakoľko nám skúmanie priebehu regresných koeficientov pre jednotlivé kvantily môže ozrejmiť závislosť medzi regresormi a skúmanou premennou. Teória kvantilovej regresie je oveľa obšírnejšia a prináša viac aplikácii ako je uvedené v tejto práci, účelom ktorej bolo priniesť čitateľovi základné znalosti potrebné pre ďalšie štúdium a prehĺbiť jeho záujem o ich ďalšie štúdium.

Zoznam použitej literatúry

[1] Koenker R.: Quantile Regression. Cambridge University Press, New York, 2005.

[2] Gutenbrunner C., Jurečková J.: Regression Rank Scores and Regression Quantiles. The Annals of Statistics, Volume 20, pp. 305–330,1992.

[3] Jurečková J.: Robustní statistické metody, Praha, 2001.

[4] Blatná D.: Robustní přístup v lineární regresi, Vysoká škola ekonomická v Praze, Praha, 2008 pp. 255–265.

[5] Hekimoglu S., Erenoglu R.C., Kalina J.: Outlier detection by means of robust regression estimators for engineering science. Journal of Zhejiang University – Science A (JZUS-A), 10 (6), pp. 909-921, 2009.

[6] Saleh A.K.Md.E., Picek J., Kalina J.: Nonparametric estimation of regression parameters in measurement error models. Metron 67 (2), pp. 177-200, 2009.

[7] Zvára K., Štepán J.: Pravděpodobnost a matematická statistika, Matfyzpress, Praha, 2006.

[8] Remo A.: Regresní kvantily, Diplomová práce, Masarykova univerzita v Brně, Přírodovědecká fakulta, Brno, 2008.

[9] Jurík T.: Vybrané algoritmy lineárneho programovania, Písomná práca k dizertačnej škúške, Univerzita Komenského v Bratislave, Fakulta Matematiky, Fyziky a Informatiky, Bratislava, 2007.

[10] Koenker R.: quantreg: Quantile Regression. R package version 4.67. http://cran.r-project.org/package=quantreg