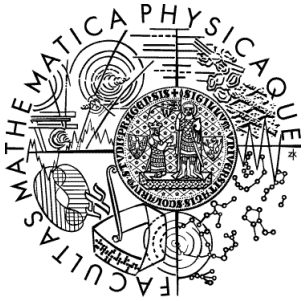
Univerzita Karlova v Prahe
Matematicko-fyzikálna fakulta

BAKALÁRSKA PRÁCA



Ján Krnáč

Odhady v lineárnej regresii

Katedra pravdepodobnosti a matematickej štatistiky

Mgr. Alena Černíková

Študijný program: Matematika

Študijný obor: Finančná a poistná matematika

2008

Ďakujem vedúcej bakalárskej práce Mgr. Aleně Černíkovej za veľmi cenné pripomienky a rady poskytnuté pri jej vypracovaní.

Prehlasujem, že som svoju bakalársku prácu napísal samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičaním práce a jej zverejňovaním.

V Prahe dňa 25.5.2008 Ján Krnáč

Obsah

Názov práce: Odhady v lineárnej regresii
Autor: Ján Krnáč
Katedra (ústav): Katedra pravdepodobnosti a matematickej štatistiky
Vedúci diplomovej práce: Mgr. Alena Černíková
E-mail vedúceho: cernikov@natur.cuni.cz

Abstrakt: Hlavným cieľom tejto práce je popísať teoretické vlastnosti odhadov regresných koeficientov, konkrétne odhadu metódou najmenších štvorcov - L2 odhad a odhadu metódou najmenšej absolútnej odchýlky - L1 odhad a porovnať ich s konkrétnymi výsledkami, ktoré dostaneme pri náhodnom generovaní dát v štatistickom programe R. Takisto porovnávame vhodnosť jednotlivých odhadov v závislosti na type dát, tj. v závislosti na rozdelení náhodných chýb e. Rozdelenie náhodných chýb má významný vplyv na výber typu odhadu regresných koeficientov, a k účelu správneho rozhodnutia ktorý odhad je pre daný model najlepší používame kritériá: rozdiel priemeru odhadov a reálnej hodnoty parametrov, rozptyl odhadov, AIC, BIC, RSS, reziduálnu sumu absolútnych hodnôt a intervalové odhady jednotlivých odhadov. Jednotlivé výpočty sú doplnené o histogramy odhadov a jeden ukážkový graf reálnej závislosti dát x a y, a odhadnutej závislosti pre obidva odhady.

Kľúčové slová: Lineárna regresia, L1 odhad, L2 odhad, LAD

Title: Estimators in linear regression
Author: Ján Krnáč
Department: Department of probability and mathematical statistics
Supervisor: Mgr. Alena Černíková
Supervisor's e-mail address: cernikov@natur.cuni.cz

Abstract: The main intention of this work is to describe theoretical characteristics of estimators of regression coeficients, specifically the least squares method estimator - L2 estimator and estimator using a least absolute deviation method - L1 (LAD) estimator. I comapre these characteristics with particular results of tests, that we get by generating data for these tests in statistical program R. I also compare the suitability of particular estimators in dependence on the distribution of errors in regression model. This distribution of random errors in linear regression model has significant influence on the choice of particular estimation method. For this purpose I used several criterions: The difference between average values of the estimated coefficients and their real values, variance of estimated parameters, AIC, BIC, RSS, residual sum of absolute values, and interval estimations of particular estimators. Numerical results of tests are supplied by histograms of estimated parameters and one instance of graph of real dependance between x and y, and estimated dependance for both estimators.

Keywords: Linear regression, L1 estimate, L2 estimate, LAD

Úvod

Vo svojej práci sa zaoberám asi najčastejšie používaným spôsobom modelovania príčinnej závislosti medzi dvomi premennými - lineárnou regresiou. Modelovanie tejto závislosti je spojené s odhadom regresných koeficientov, čiže koeficientov, ktoré udávajú polohu priamky, ktorou sa snažíme túto lineárnu závislosť popísať. Počet týchto koeficientov je daný počtom premenných, medzi ktorými túto závislosť skúmame. V tejto práci sa budeme zaoberať najjednoduchším modelom lineárnej regresie a to modelom s jednou nezávisle premennou. Veľkú časť tejto práce tvorí porovnávanie metód odhadov týchto regresných koeficientov, konkrétne je to, odhad metódou najmenších štvorcov, a odhad metódou najmenšej absolútnej odchýlky. Asi najznámejším odhadom regresných koeficientov, je práve odhad metódou najmenších štvorcov. Ten, ako sa ukáže, nemusí byť vždy tou najlepšou voľbou pre odhad koeficientov, kedže minimalizuje štvorec odchýlky bodu od priamky, a tým robí L2 odhad veľmi citlivý na prípadné odľahlé pozorovania. Preto je tento odhad vhodný len pre lineárny regresný model s normálnym rozdelením chýb, avšak v praxi sa metóda najmenších štvorcov často používa, aj keď tento predpoklad splnený nie je a táto práca okrem iného poukazuje na to, k akým chybám môže táto ignorácia nesplnenia predpokladu normality náhodných chýb viesť.

Javí sa ako dobrá alternatíva, práve ako akási "obrana" voči hrubým chybám v pozorovaniach, metóda ktorá neminimalizuje štvorec, ale iba dĺžku úsečky bodu od priamky. Táto metóda je zastúpená práve odhadom L1.

Rôznu pravdepodobnosť, že pri pozorovaniach urobíme hrubú chybu, čím získame odľahlé pozorovanie, môžme vyjadriť rôznym rozdelením náhodných chýb v lineárnom regresnom modeli, a tým zásadne ovplyvniť odpoveď na otázku: "ktorou metódou by som mal odhadnúť regresné koeficienty pre dané data?".

V prvej kapitole sa zaoberám základnou definíciou regresného modelu. Od všeobecného modelu som prešiel k modelu, ktorý budeme potrebovať a používať v ďalších kapitolách.

V druhej kapitole, sa pozrieme na skúmané odhady a na ich teoretické vlastnosti, takisto ako na spôsoby ich výpočtu.

V tretej kapitole si pripomenieme niektoré základné spojité rozdelenia, ktoré budeme používať hlavne pri generovaní dát-náhodných chýb a ich vlastnosti sa nám budú hodiť pri skúmaní reálneho správania sa jednotlivých odhadov.

V štvrtej kapitole sa oboznámime s kritériami vhodnosti odhadu pre daný regresný model, budeme skúmať ktoré kritérium je, alebo nie je smerodatné pri porovnávaní kvality odhadov.

V piatej kapitole sú už moje vlastné výpočty a výsledky. Číselné údaje sú doplnené o rôzne grafy, aby sme mohli lepšie porovnať, či sa teoretické poznatky o obidvoch odhadoch zhodujú s realitou.

Kapitola 1

Regresný model

Nech $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$ je náhodný vektor, $\boldsymbol{X} = (x_{ij})$ je matica známych konštánt (nenáhodných čísel) typu $n \times k$ kde $k < n$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)^T$ je vektor neznámych parametrov a $\boldsymbol{e} = (e_1, ..., e_n)^T$ je náhodný vektor (*vektor chýb*) spĺňajúci podmienky $\mathsf{E}\boldsymbol{e} = \boldsymbol{0}$, $\mathsf{var}\,\boldsymbol{e} = \sigma^2\boldsymbol{I}$. Potom model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

nazývame *lineárny regresný model* alebo tiež *lineárna regresia.*

V definícii lineárneho regresného modelu vyššie, je matica $\boldsymbol{X}$ typu $n \times k$ kde n je počet pozorovaní, ktoré predstavujú riadky matice a k je počet *regresorov* alebo *nezávisle premenných* ktoré predstavujú stĺpce matice. Tzn

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ x_{21} & \cdots & x_{2k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix} . \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} .$$

Poznámka. Pre náhodný vektor chýb $\boldsymbol{e}$ budeme predpokladať, že jeho zložky $e_1, \ldots, e_n$ sú nezávislé, rovnako rozdelené náhodné veličiny. Regresor $\mathbf{x}_i$, čiže i-tý stĺpec matice je **nenáhodná** veličina, ktorej hodnoty sú presné, čiže nie sú zaťažené náhodnou chybou.

Často sa používa aj tento zápis

$$\boldsymbol{Y} = \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + , \ldots , + \beta_k\mathbf{x}_k + \boldsymbol{e},$$

kde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ sú pozorovania, $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})^T \in \mathbb{R}^n$ sú regresory, $\beta_1, \ldots, \beta_k \in \mathbb{R}$ sú neznáme parametre a $\boldsymbol{e} = (e_1, \ldots, e_n)'$ je vektor chýb.

Pri štúdiu lineárneho regresného modelu sa niekedy prvá zložka vektoru regresných koeficientov uvažuje spoločná všetkým pozorovaniam a preto sa nenásobí žiadnou nezávisle premennou. Preto v našom výklade model rozšírime o absolútny člen β_0(*intercept*):

$$\boldsymbol{Y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + , \ldots , + \beta_k\mathbf{x}_k + \boldsymbol{e}.$$

V tejto práci prejdeme od všeobecnej definície lineárnej regresie, k jednoduchšiemu modelu, kde budeme uvažovať len jednu nezávisle premennú. Tzn., že budeme pracovať s modelom $\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x}_1 + \boldsymbol{e}$ a kedže máme len jeden regresor označíme ho jednoducho $\mathbf{x}$, čiže výsledný model bude mať tvar

$$\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{e},$$

ktorý ak zapíšeme v tvare $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ má nasledujúcu štruktúru

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Zložky náhodného vektoru $\boldsymbol{Y}$, náhodné veličiny $Y_1, ..., Y_n$ sú *pozorovania* a nazývajú sa tiež *závisle premenné.*

Kapitola 2

Odhady regresných koeficientov

2.1 Metóda najmenších štvorcov (L2 Odhad)

Budeme naďalej uvažovať model

$$\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{e}. \tag{2.1}$$

Najznámejší odhad koeficientov v lineárnom regresnom modeli je odhad metódou najmenších štvorcov. Názorne povedané, ide vlastne o preloženie dát (x_i, Y_i) priamkou, tak aby súčet druhých mocnín odchýliek pôvodných dát a predpovedaných (odhadnutých) dát bol minimálny. Takáto priamka (množina odhadnutých bodov) sa nazýva *regresná priamka.* Ak si prepíšeme tento model na tvar

$$Y_i = \beta_0 + \beta_1 x_i + e_i,$$

potom odhad parametrov β_0, β_1 metódou najmenších štvorcov sú také hodnoty $\hat{\beta}_0, \hat{\beta}_1$, pre ktoré platí

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta \in \mathbb{R}^2} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2, \tag{2.2}$$

kde $\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} = (\hat{\beta}_0, \hat{\beta}_1)$, $\boldsymbol{X}_i^T = (1, x_i)$ a $\boldsymbol{\beta}^{\boldsymbol{T}} = (\beta_0, \beta_1)$.

Poznámka. Pri odhadovaní regresných koeficientov (ľubovoľnou metódou) dostaneme pomocou odhadnutých koeficientov vyrovnané hodnoty náhodnej veličiny $\boldsymbol{Y}$ čiže $\boldsymbol{y} = a + b\mathbf{x}$, kde a a b sú nejaké bodové odhady β_0 a β_1. Špeciálne, odhad metódou najmenších štvorcov, spočíva v podstate v minimalizovaní L_2 normy reziduí, čiže rozdielu medzi pôvodnými hodnotami $\boldsymbol{Y}$ a teoretickými hodnotami $\boldsymbol{y}$. Máme $\mathbf{res} = \boldsymbol{Y} - \boldsymbol{y} = \boldsymbol{Y} - (a + b\mathbf{x})$, a minimalizujeme $\|\boldsymbol{Y} - (a + b\mathbf{x})\|_2 = \sum_{i=1}^{n} (Y_i - a - bx_i)^2$ a po menšej úprave dostávame práve vzťah (2.2)

Praktický výpočet odhadov regresných koeficientov, vychádza s maticového zápisu lineárnej regresie, $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, a z podmienky, že výraz $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$ ako funkcia $\boldsymbol{\beta}$ má byť minimálny.

Tvrdenie 2.1 Odhady parametrov lineárnej regresie metódou najmenších štvorcov sú $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Dôkaz. Pretože hľadáme takú lineárnu kombináciu $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ stĺpcov matice $\boldsymbol{X}$, ktorá sa rovná odhadu $\hat{\boldsymbol{Y}}$ vektoru $\boldsymbol{X}\boldsymbol{\beta}$ dostávame, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Pretože s ukáže, že $\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{0}$ dostaneme pre odhad $\hat{\boldsymbol{\beta}}$

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{Y}},$$
$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\hat{\boldsymbol{Y}},$$
$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y},$$

čiže dostávame

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad \square$$

Ak si tento vzťah ďalej upravíme, za predpokladu, že pracujeme s modelom $\boldsymbol{Y} = \beta_0 + \beta_1\mathbf{x} + \boldsymbol{e}$, a označíme si $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ a $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ máme

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}^{-1}, \qquad \boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\bar{Y} \\ \sum x_iY_i \end{pmatrix},$$

a z toho dostávame *sústavu normálnych rovníc*

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} n\bar{Y} \\ \sum x_iY_i \end{pmatrix} =$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum x_iY_i \end{pmatrix} =$$

$$= \frac{1}{\sum(x_i - \bar{x})^2} \begin{pmatrix} \bar{Y}\sum x_i^2 - \bar{x}\sum x_iY_i + n\bar{x}^2\bar{Y} - n\bar{x}^2\bar{Y} \\ \sum x_iY_i - n\bar{x}\bar{Y} \end{pmatrix} =$$

$$= \begin{pmatrix} \bar{Y} - \dfrac{\sum(x_i - \bar{x})(Y_i - \bar{x})}{\sum(x_i - \bar{x})^2}\bar{x} \\ \\ \dfrac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2} \end{pmatrix}$$

Zvyčajne sa podľa týchto vzorcov počíta len odhad $\hat{\beta}_1$ a odhad $\hat{\beta}_0$ sa ľahko dopočíta zo vzťahu $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$.

Definícia. Nech $\Omega \in \mathbb{R}_m$ je parametrický priestor. Hovoríme, že odhad T parametru θ je *nestranný*, ak pre každé $\theta \in \Omega$ platí $\mathsf{E}T = \theta$.

Nestrannosť je jeden z najčastejších požiadavkov na odhad parametru. Avšak niektoré odhady samozrejme nestranné nie sú. Pre takéto odhady platí, že $\mathsf{E}T = \theta + g(\theta)$ kde vektor $g(\theta)$ sa nazýva *vychýlenie* odhadu T v bode θ.

Tvrdenie 2.2 (Nestrannosť L2 odhadu)
Odhad metódou najmenších štvorcov $\hat{\boldsymbol{\beta}}$ je nestranný odhad regresných koeficientov $\boldsymbol{\beta}$.

Dôkaz. Pre strednú hodnotu odhadu $\hat{\boldsymbol{\beta}}$ platí:

$$\mathsf{E}\hat{\boldsymbol{\beta}} = \mathsf{E}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathsf{E}\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \qquad \square$$

Odhad regresných koeficientov metódou najmenších štvorcov je veľmi nerobustný a citlivý k odľahlým pozorovaniam Y_i (*outliers*) takisto ako k odľahlým pozorovaniam elementom vektoru $\mathbf{x}$ (*leverage points*).

2.2 Metóda najmenšej absolútnej odchýlky (L1 odhad)

Znovu budeme uvažovať lineárny regresný model (2.1) a zoznámime sa s ďalším veľmi dôležitým odhadom regresných parametrov a to je L1 odhad. Ako už názov tohto odstavcu napovedá ide o preloženie dát (x_i,Y_i) priamkou, tak aby súčet absolútnych hodnôt odchýliek predpovedaných (odhadnutých) hodnôt a pôvodných dát bol minimálny.

To znamená, že odhad parametrov β_0, β_1 metódou najmenšej absolútnej odchýlky sú také hodnoty $\tilde{\beta}_0, \tilde{\beta}_1$, pre ktoré platí

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\beta\in\mathbb{R}^2} \sum_{i=1}^{n} \mid Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} \mid,$$

kde $\tilde{\boldsymbol{\beta}}^{\boldsymbol{T}} = (\tilde{\beta}_0, \tilde{\beta}_1)$, $\boldsymbol{X}_i^T = (1, x_i)$ a $\boldsymbol{\beta}^{\boldsymbol{T}} = (\beta_0,\beta_1)$.

Poznámka. Ekvivalentne môžme túto úlohu zapísať ako minimalizáciu funkcie $L(\beta_0, \beta_1) = \sum_{i=1}^{n} \mid Y_i - \beta_0 - \beta_1 x_i \mid$.

Túto úlohu môžme riešiť ako úlohu lineárneho programovania, pretože už Gauss dokázal (znenie vety + dôkaz je uvedený v knihe Anděl [2], str 259), že za predpokladu , že z čísel $x_1, ..., x_n$ sú aspoň dve čísla rôzne, existuje takáto optimálna priamka, ktorá prechádza najmenej dvoma bodmi (x_i, Y_i) a (x_j, Y_j) kde $i \neq j$.

Tvrdenie 2.3
Minimalizácia výrazu $\sum_{i=1}^{n} \mid Y_i - \beta_0 - \beta_1 x_i \mid$ je ekvivalentná s úlohou lineárneho programovania minimalizovať

$$\sum_{i=1}^{n} r_i$$

za podmienok,

$$r_i + \beta_0 + \beta_1 x_i \geq Y_i, \quad i = 1, \ldots, n,$$
$$r_i - \beta_0 - \beta_1 x_i \geq -Y_i, \quad i = 1, \ldots, n.$$

Dôkaz. Dôkaz tohto tvrdenia v obecnej podobe, je uvedený v knihe Anděl [2], str 259. □

Z tvrdenia 2.3 vyplýva, že úloha minimalizovať $\sum_{i=1}^{n} \mid Y_i - \beta_0 - \beta_1 x_i \mid$ je úloha lineárneho programovania, ktorá má $n+2$ premenných a $2n$ obmedzujúcich podmienok.

Túto úlohu si môžeme prepísať do tvaru

$$\text{minimalizovať } \boldsymbol{c}^T\boldsymbol{\lambda} \quad \text{za podmienok } \boldsymbol{A}\boldsymbol{\lambda} \geq \boldsymbol{b},$$

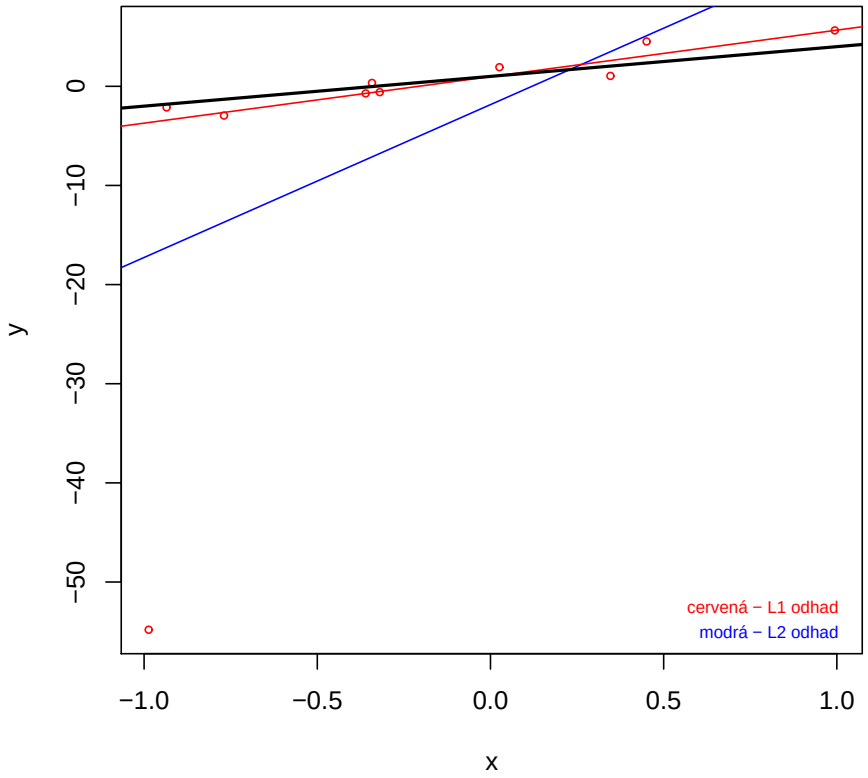
kde

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & x_1 \\ 0 & 1 & \cdots & 0 & 1 & x_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 1 & x_n \\ 1 & 0 & \cdots & 0 & -1 & -x_1 \\ 0 & 1 & \cdots & 0 & -1 & -x_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 & -x_n \end{pmatrix}, \qquad \boldsymbol{b} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \\ -Y_1 \\ -Y_2 \\ \vdots \\ -Y_n \end{pmatrix}, \qquad \boldsymbol{\lambda} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \\ \beta_0 \\ \beta_1 \end{pmatrix},$$

$$\boldsymbol{c}^T = (\underbrace{1, 1, 1, \ldots 1, 1}_{n}, 0, 0).$$

Metóda najmenšej absolútnej odchýlky (L1 odhad) ako jeden z mnohých postupov ako odhadnúť koeficienty v lineárnom regresnom modeli je viac robustný voči odľahlým pozorovaniam Y_i, čiže voči závisle premennej, ako metóda najmenších štvorcov, ale je relatívne nerobustná (citlivá) k odľahlým elementom vektoru $\mathbf{x}$.

Pre porovnanie, ako môže jeden odľahlý bod, ovplyvniť regresnú priamku počítanú metódou najmenších štvorcov, si uvedieme ilustračný graf:



regresné priamky L1 a L2 odhadu + skutočná lineárna závislosť dát (čierna priamka)

Kapitola 3

Dôležité rozdelenia

Úlohou tejto práce je popísať, vhodnosť jednotlivých odhadov pre rôzne rozdelenia náhodných chýb $\boldsymbol{e}$. Preto si pripomenieme niektoré základné vlastnosti najdôležitejších spojitých rozdelení.

3.1 Špeciálne funkcie

Pred tým ako pristúpime k samotným charakteristikám spojitých rozdelení, pripomeňme si niektoré dôležité funkcie, ktoré budeme potrebovať k zadefinovaniu niektorých týchto rozdelení. Zavedieme si tzv. gamma funkciu $\Gamma(a)$ ktorú budeme potrebovať pri definícii t-rozdelenia. Máme

$$\Gamma(a) = \int_0^\infty x^{a-1}\mathrm{e}^{-x}dx, \quad a > 0.$$

Dôležité vzťahy ktoré pre tieto funkcie platia

$$\Gamma(a+1) = a\Gamma(a), \qquad \Gamma(\frac{1}{2}) = \sqrt{\pi}.$$

3.2 Normálne rozdelenie

Hovoríme, že náhodná veličina X má normálne rozdelenie (nazývané aj *Gaussovo* rozdelenie) s parametrami μ a σ^2, ak hustota tejto náhodnej veličiny má tvar:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R},\ \mu \in \mathbb{R},\ \sigma > 0.$$

Normálne rozdelenie s parametrami μ a σ^2 označujeme symbolom $\mathsf{N}(\mu, \sigma^2)$. Pre náhodnú veličinu $X \sim \mathsf{N}(\mu, \sigma^2)$ platí $\mathsf{E}X = \mu$, $\mathsf{var}\, X = \sigma^2$.

Veľmi dôležitým špeciálnym prípadom normálneho rozdelenia, je *normované normálne rozdelenie.* Čiže normálne rozdelenie s nulovou strednou hodnotou a jednotkovým rozptylom. Hustota normovaného normáleno rozdelenia, ktoré značíme $\mathsf{N}(0,1)$, má tvar

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad x \in \mathbb{R}.$$

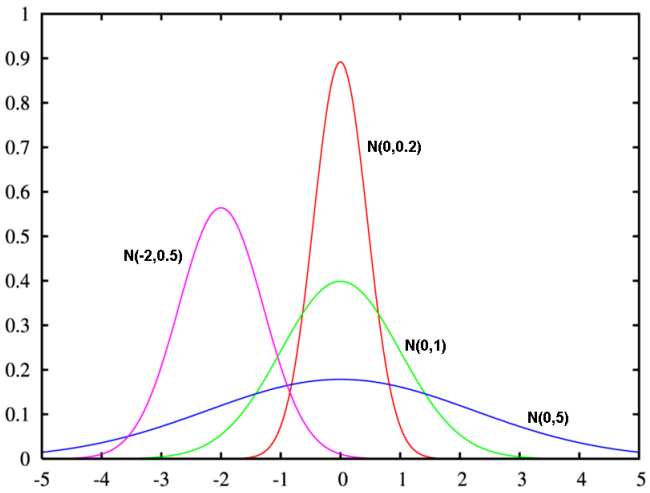
Distribučná funkcia normovaného normálenho rozdelenia

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\} \mathrm{d}u,$$

sa nedá vyjadriť pomocou elementárnych funkcií, jej hodnoty sú ale podrobne zaznamenané v tabuľkách. A pretože, medzi distribučnou funkciou $F(x)$ obecného normálneho rozdelenia $\mathsf{N}(\mu, \sigma^2)$ a distribučnou funkciou $\Phi(x)$ normovaného normálenho rozdelenia $\mathsf{N}(0, 1)$ platí vzťah $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$, môžeme veľmi ľahko určiť hodnoty distribučnej funkcie rozdelenia $\mathsf{N}(\mu, \sigma^2)$.

Ďalej si zavedieme tzv. *kritickú hodnotu* normovaného normálneho rozdelenia, čiže číslo $u(a)$, ktoré náhodná veličina $X \sim \mathsf{N}(0, 1)$ prekročí s pravdepodobnosťou α. Toto číslo môžme vypočítať z rovnice $\mathsf{P}[X \geq u(\alpha)] = \alpha$, čiže $1 - \Phi[u(\alpha)] = \alpha$, z čoho nám vyjde, že $u(\alpha) = \Phi^{-1}(1 - \alpha)$, kde Φ^{-1} je tzv *kvantilová funkcia*. Dôležitý vzťah, ktorý platí pre kvantily normálneho normovaného rozdelenia je $\Phi^{-1}(1 - \alpha) = -\Phi^{-1}(\alpha)$ čiže $u(1 - \alpha) = -u(\alpha)$.

Poznámka. Ďalší veľmi dôležitý vzťah je $\Phi(-x) = 1 - \Phi(x)$.



Obr 3.1 Hustota normálneho rozdelenia pre rôzne hodnoty parametrov μ, σ^2

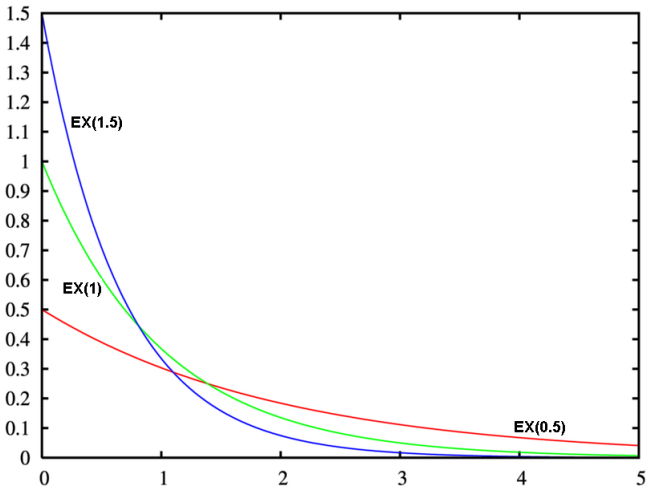
3.3 Exponenciálne rozdelenie

Toto rozdelenie nebudeme pri porovnávaní používať, uvádzam ho len pre predstavu, z čoho vychádza rozdelenie dvojito-exponenciálne, na ktoré sa bližšie pozrieme v ďalšom odstavci, a ktoré takisto budeme potrebovať, pri generovaní náhodných chýb.

Hovoríme, že náhodná veličina X má exponenciálne rozdelenie s parametrom λ, ak hustota tejto náhodnej veličiny má tvar:

$$f(x) = \lambda \mathrm{e}^{-\lambda x}, \quad x > 0.$$

Exponenciálne rozdelenie s parametrom λ označujeme symbolom $\mathsf{Ex}(\lambda)$. Pre náhodnú veličinu $X \sim \mathsf{Ex}(\lambda)$ platí $\mathsf{E}X = 1/\lambda$, $\mathsf{var}\, X = 1/\lambda^2$.



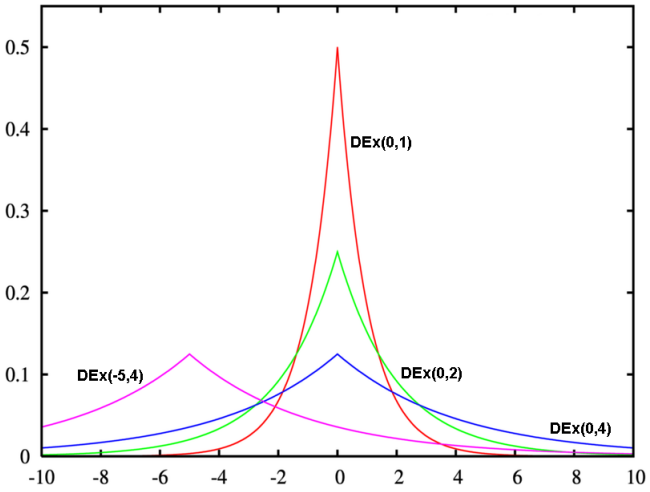
Obr 3.2 Hustota exponenciálneho rozdelenia pre rôzne hodnoty parametru λ

3.4 Dvojito exponenciálne rozdelenie

Hovoríme, že náhodná veličina X má dvojito exponenciálne rozdelenie (nazývané aj *Laplaceovo* rozdelenie) s parametrami a a b, ak hustota tejto náhodnej veličiny má tvar:

$$f(x) = \frac{1}{2b} \exp\left\{-\frac{|x-a|}{b}\right\}, \quad a \in \mathbb{R},\ b > 0.$$

Exponenciálne rozdelenie s parametrami a a b označujeme symbolom $\mathsf{DEx}(a, b)$. Pre náhodnú veličinu $X \sim \mathsf{DEx}(a, b)$ platí $\mathsf{E}X = a$, $\mathsf{var}\, X = 2b^2$.



Obr 3.3 Hustota Laplaceovho rozdelenia pre rôzne hodnoty parametrov a a b

3.5 t-rozdelenie

Hovoríme, že náhodná veličina X má t rozdelenie o k stupňoch voľnosti (nazývané aj *Studentovo* rozdelenie), ak hustota tejto náhodnej veličiny má tvar:

$$f(x) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{\pi k}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad x \in \mathbb{R},\ k \geq 1.$$

Studentovo rozdelenie o k stupňoch voľnosti označujeme symbolom t_k. Ak je $k > 1$ potom pre náhodnú veličinu $X \sim \mathsf{t}_k$ existuje stredná hodnota $\mathsf{E}X = 0$, a ak naviac platí, že $k > 2$ tak existuje aj konečný rozptyl $\mathsf{var}\, X = \frac{k}{k-2}$.

Poznámka. Pre k=1 dostávame *Cauchyho* rozdelenie, ktoré ako sme uviedli strednú hodnotu nemá.

Pre hodnoty $k > 500$ je rozdelenie t veľmi blízke normovanému normálnemu rozdeleniu. To nám potvrdí aj nasledujúce tvrdenie.

Tvrdenie 3.1 Nech $X \sim \mathsf{t}_k$ je náhodná veličina s hustotou $f_k(x)$. Potom pre každé $x \in \mathbb{R}$ platí

$$\lim_{k\to\infty} f_k(x) = \varphi(x).$$

Dôkaz. Uvedený v knihe Anděl [1], str 336 □

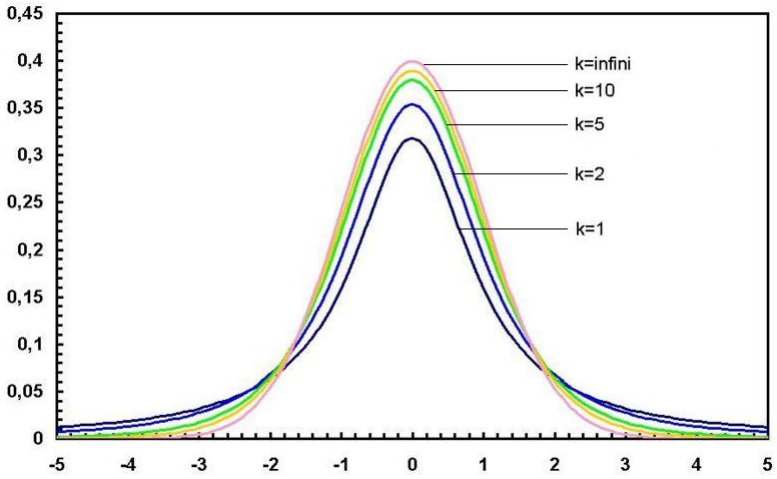
Veľmi užitočné je aj nasledujúce tvrdenie, ktoré popisuje vzťah medzi normovaným normálnym rozdelením a Studentovým rozdelením:

Tvrdenie 3.2 Nech $X \sim \mathsf{N}(0,1)$ a $Y \sim \chi^2_k$ sú nezávislé náhodné veličiny, kde χ^2_k značí chí-kvadrát rozdelenie o k stupňoch voľnosti. Potom náhodná veličina definovaná ako

$$T = \frac{X}{\sqrt{\frac{Y}{n}}},$$

má Studentovo t rozdelenie o k stupňoch voľnosti.

Dôkaz. Uvedený v knihe Anděl [1], str 74 □



Obr 3.4 Hustota t-rozdelenia pre rôzne počty stupňov voľnosti

Kapitola 4

Kritériá vhodnosti modelu

Výber vhodného spôsobu vyhodnotenia súboru údajov a spôsobu odhadu regresných koeficientov je veľa krát závislý iba na základe osobného názoru. To znamená, že výsledky získané touto zvolenou metódou, aj keď štatisticky zhodnotené, nemusia vždy odpovedať realite. Je preto dôležité venovať pozornosť výberu a štatistickému posúdeniu vhodného modelu. Termín výber vhodného modelu môžme takisto použiť k označeniu situácie, keď sa snažíme pre rôzne metódy odhadu regresných koeficientov posúdiť, ktorý z týchto výberov je lepší. Pojem lepší odhad sa dá posudzovať na základe rôznych kritérií, z ktorých najznámejšie si uvedieme v nasledujúcom odstavci.

4.1 Vierohodnostná funkcia

V štatistickej analýze je vierohodnotstná funkcia (*angl. "Likelihood function"*), funkciou parametrov štatistického modelu. Ak si preložíme slovo "Likelihood" do slovenčiny tak, v obecnej rovine, je to vlastne "možnosť" alebo tiež synonymum - "pravdepodobnosť", s technického hľadiska budeme v tomto texte používať, aj keď trochu nepresne, pojem "vierohodnosť". Veľmi zjednodušene povedané, "pravdepodobnosť" nám umožňuje predpovedať, odhadovať neznáme výsledky založené na známych parametroch, zatiaľ čo "vierohodnosť" nám umožňuje odhadovať neznáme parametre na základe známych výsledkov.

V podstate, vierohodnosť pracuje opačne ako pravdepodobnosť: ak máme daný jav B, používame podmienenú pravdepodobnosť $P(A \mid B)$ na zistenie štatistických informácií o jave A, naopak pre dané A, používame vierohodnostnú funkciu na zistenie štatistických informácií o jave B. Tento postup si môžme odvodiť z Bayesovho vzorca:

$$P(B \mid A) = \frac{P(A \mid B).P(B)}{P(A)}.$$

Vierohodnostná funkcia je vlastne podmienená pravdepodobnosť uvažovaná ako funkcia jej druhého argumentu s daným prvým argumentom, čiže

$$L(b \mid A) = P(A \mid B = b).$$

4.1.1 Vierohodnostná funkcia v parametrickom modeli

Predpokladajme, že máme parametrický štatistický model, v ktorom vektor pozorovaní $\mathbf{x}$ má distribučnú funkciu F o ktorej vieme, že je prvkom rodiny (parametrickej množiny) $\{F_\theta; \theta \in \Theta\}$, kde $\Theta \in \mathcal{B}^k$ je borelovská množina v $\mathbb{R}^k$. Prvky tejto rodiny sa od seba líšia iba hodnotou parametru θ tzn, že toto rozdelenie je dané hustotou $f(\mathbf{x} \mid \theta)$. Potom vierohodnostná funkcia je

$$L(\theta \mid \mathbf{x}) = f(\mathbf{x} \mid \theta).$$

Inak povedané, ak $f(\mathbf{x} \mid \theta)$ je funkcia $\mathbf{x}$ s daným pevným θ, je to hustota rozdelenia v tomto modeli, a naopak, ak je $f(\mathbf{x} \mid \theta)$ funkcia θ s pevným x je to vierohodnostná funkcia.

4.2 RSS - Reziduálna suma štvorcov

Reziduálna suma štvorcov - skrátene RSS je suma druhých mocnín reziduií v lineárnom modeli. Vyjadruje základný obraz o odchýlke medzi reálnymi datami a odhadnutými datami. Samozrejme platí, že čím menšia RSS, tým je náš zvolený odhad lepší.

V našom lineárnom regresnom modeli

$$\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{e},$$

pri značení $\mathbf{x} = (x_1, \ldots, x_n)^T$, odhadu vektoru $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ ako $\hat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^T$, reziduá ako $\text{res}_i = Y_i - \hat{Y}_i$, odhad regresných parametrov $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ ako $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$, počte pozorovaní n a kde

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

platí

$$RSS = \sum_{i=1}^{n} \text{res}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Poznámka. Všimnime si, že RSS je vlastne minimalizovaná funkcia z ktorej sa počíta odhad regresných parametrov metódou najmenších štvorcov. To ale znamená, že RSS ako jeden z kritérií vhodnosti použitia daného odhadu je vždy najmenší pre L2 odhad. Ale ako už vieme, L2 odhad, pri určitom type rozdelenia vektoru náhodných chýb, nie je vždy najlepšou alternatívou pre odhad regresných koeficientov. Preto sa za viac smerodatné kritérium považuje kritérium ktoré si uvedieme o dva odstavce nižšie.

4.3 Reziduálna suma absolútnych hodnôt

Pre úplnosť si uvedieme ešte jedno kritérium, ktoré si uvádzame len pre lepšie pochopenie, prečo RSS a reziduálna suma absolútnych hodnôt (označme si ju RSAV) nie je smerodatné pri porovnávaní L1 a L2 odhadu.

$$RSAV = \sum_{i=1}^{n} \mid \mathrm{res}_i \mid = \sum_{i=1}^{n} \mid Y_i - \hat{Y}_i \mid = \sum_{i=1}^{n} \mid Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \mid .$$

Obe tieto kritériá, RSS aj RSAV sú obsiahnuté v definícii L2 resp L1 odhadu, ako funkcie parametrov β_0 a β_1, ktoré sa za účelom získania odhadov pre parametre β_0 a β_1 minimalizujú. To znamená, že RSS bude vždy menšia pre L2 odhad, a naopak RSAV bude vždy menšia pre L1 odhad. RSS sa avšak veľmi často používa ako kritérium vhodnosti odhadu, preto ho pre úplnosť k výpočtom uvádzam v tejto práci spolu s "variantou" pre L1 odhad - RSAV.

4.4 AIC

AIC (Akaike's information criterion) popísané matematikom menom *Hirotsugu Akaike* je kritérium vhodnosti použitia odhadu (regresného modelu) pre dané data. Presný postup výpočtu si uvedieme nižšie, ale je dôležité vedieť, že je súčtom dvoch členov, z ktorých prvý je úmerný logaritmu reziduálnej sumy štvorcov, a druhý člen je úmerný zložitosti modelu (počet členov v modeli). Môžme povedať, že AIC meria *úspornosť*, resp *účelnú úspornosť* modelu, čiže takú vlastnosť modelu, ktorá váži schopnosť modelu predpovedať hodnoty vysvetlovanej premennej proti jeho zložitosti.

Čiže z poznatkov ktoré sme uviedli vyššie vyplýva, že čím nižšia hodnota AIC pre daný model, tým úspornejší=lepší model sme vybrali. Alebo naopak, aspekty, ktoré zvyšujú hodnotu AIC je počet parametrov a reziduálna suma štvorcov.

Pre ilustráciu si uveďme aspoň základný výpočet AIC:

$$AIC = p.k - 2\ln(L),$$

kde k je počet parametrov v štatistickom modeli, $p = 2$ pre klasický AIC a L je vierohodnostná funkcia.

Ak budeme predpokladať, že n je počet pozorovaní a $RSS = \sum_{i=1}^{n} \mathrm{res}_i^2$ je reziduálna suma štvorcov a vektor chýb má **normálne rozdelenie**, potom AIC môžeme prepísať na tvar

$$AIC = 2k - n\ln\left(\frac{RSS}{n}\right).$$

Poznámka. Z poslednej rovnice a z poznámky ktorú sme uviedli v odstavci o RSS, vyplýva, že pokiaľ majú náhodné chyby normálne rozdelenie, potom AIC má vždy najnižšiu hodnotu pre L2 odhad, čo hovorí v prospech faktu, že keď máme v regresnom modeli náhodné chyby s normálnym rozdelením, je L2 odhad najlepším možným odhadom.

4.5 BIC

BIC (Bayesian information criterion) je štatistické kritérium pre výber modelu. BIC je taktiež známe pod názvom Schwarzovo kritérium, alebo "Schwarz information criterion" (SIC). BIC je v podstate modifikované AIC, kde za p dosadíme $\ln(n)$, kde n je počet pozorovaní. Čiže dostávame

$$BIC = k\ln(n) - 2\ln(L),$$

alebo, za predpokladu, že vektor chýb má normálne rozdelenie,

$$BIC = k\ln(n) - n\ln\left(\frac{RSS}{n}\right).$$

Existujú samozrejme aj ďalšie kritériá na posúdenie vhodnosti použitého odhadu, ale najčastejšie sa jedná o myšlienku použitia RSS + penalizácia, ktorá je obvykle nejaká monotónna funkcia vysvetľujúcich premenných.

Kapitola 5

Výsledky testov

Po oboznámení s teóriou ktorú budeme potrebovať, aby sme správne pochopili a interpretovali výsledky, pristúpime k praktickej časti. Budeme uvažovať lineárny regresný model

$$\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{e}$$

V jednotlivých sekciách tejto kapitoly, budeme meniť rozdelenie vektoru náhodných chýb $\boldsymbol{e}$ a v rámci sekcií naviac uvažovať rôzne možnosti, ako napr. počet pozorovaní, parametre rozdelenia vektoru náhodných chýb $\boldsymbol{e}$ atď. K účelu testovania používame štatistický program R 2.6.1, v ktorom používame hlavne funkcie na generovanie náhodných vektorov s požadovaným rozdelením, odhad pomocou metódy najmenších štvorcov (funkcia *lm*) a odhad pomocou metódy najmenšej absolútnej odchýlky (doinštalovaná funkcia *rq*). Výpočty prebiehajú tým spôsobom, že si zvolíme nejakú hodnotu β_0 a β_1, náhodne si vygenerujeme data vektoru $\mathbf{x}$ (náhodné čísla v intervale $< -1, 1 >$), vygenerujeme náhodné chyby (vektor $\boldsymbol{e}$) tak aby mali požadované rozdelenie a dopočítame zložky vektoru $\boldsymbol{Y}$ z rovnice $\boldsymbol{Y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{e}$ a generujeme 1000 opakovaní - 1000 krát odhadneme regresné koeficienty L2 odhadom, výsledky si zapíšeme a pre tie isté data opakujeme postup, s tým rozdielom, že tento krát odhadujeme regresné koeficienty L1 odhadom. Pre jednoduchosť označíme odhady parametrov β_0 a β_1, a a b pre obidva odhady. S nameraných hodnôt odhadov regresných koeficientov potom vypočítame jednotlivé zrovnávacie kritériá. Tieto kritériá sú, stredná hodnota a rozptyl jednotlivých odhadov regresných parametrov nameraných pri jednotlivých opakovaniach testu, AIC, BIC, RSS, RSAV a intervalové odhady regresných parametrov. Chceme, aby hodnoty týchto kritérií boli čo najmenšie. To znamená, že pri skúmaní kvality odhadnutých parametrov požadujeme, aby boli čo **najpresnejšie**, čiže minimálny rozdiel skutočnej hodnoty parametru a priemeru jeho odhadov, a zároveň chceme, aby rozptyl odhadov bol čo najmenší preto, aby sme v odhadoch nezískávali príliš často veľmi chybné (odľahlé) hodnoty. Takisto sú výpočty doplnené grafmi, či už sú to histogramy jednotlivých odhadov parametrov pre obidva typy odhadov, vývoj jedného kritéria v závislosti napr na počte stupňov voľnosti v prípade t-rozdelenia a pod. V grafoch sú hodnoty vypočítané L1 odhadom vždy červené a hodnoty vypočítané L2 odhadom sú vždy modré.

Značenie ktoré budeme používať je:

- $AIC = (AIC_1 \ldots, AIC_n)$ kde AIC_i je AIC kritérium modelu s príslušným odhadom regresných koeficientov pre i-té opakovanie merania.
- $\boldsymbol{a} = (a_1 \ldots, a_n)$ kde a_i je odhad regresného koeficientu β_0 pre i-té opakovanie merania.
- $\boldsymbol{b} = (b_1 \ldots, b_n)$ kde b_i je odhad regresného koeficientu β_1 pre i-té opakovanie merania.
- n je počet pozorovaní.
- m je počet opakovaní testu pre jednotlivé rozdelenie náhodných chýb.

Platí:

- $\beta_0 = 1$
- $\beta_1 = 3$

Teraz pristúpime k prvému modelu, a to k lineárnej regresii s normálnym rozdelením náhodných chýb.

5.1 Normálne rozdelenie náhodných chýb

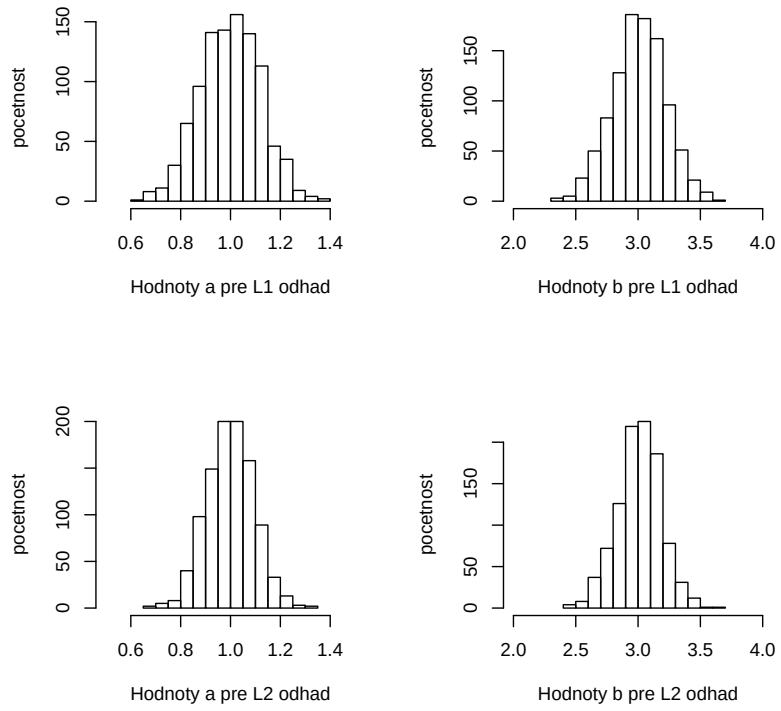
Lineárny regresný model, kde náhodné chyby majú normálne rozdelenie, je najčastejšie používaným modelom v matematickej štatistike. Vzhľadom na teoretické poznatky o obidvoch odhadoch, môžme očakávať, že L2 odhad bude pri normálnom rozdelení náhodných chýb vo všetkých kritériách lepším odhadom.

5.1.1 $\boldsymbol{e} \sim \mathsf{N}(0,1)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

Budeme uvažovať $n = 100$ čiže 100 pozorovaní. Náhodné chyby majú normálne rozdelenie so strednou hodnotou 0 a rozptylom 1.

	$\mid \bar{\boldsymbol{a}}\text{-}\beta_0 \mid$	$\mid \bar{\boldsymbol{b}}\text{-}\beta_1 \mid$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0009**	**0.0079**	0.0147	0.0453	293.9247	299.1351	99.3805	**78.6196**
L2 odhad	0.0011	0.0094	**0.0090**	**0.0309**	**287.0033**	**294.8188**	**98.2510**	79.0563

Tabuľka 5.1 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{N}(0,1)$



Obr 5.1 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{N}(0,1)$

Z tabuľky 5.1 vidíme, že L2 odhad bol vo väčšine meraných kritérií lepším odhadom ako L1 odhad. Rozdiely v hodnotách kritérií u oboch odhadov sú však relatívne malé. V hlavných kritériách, čiže AIC a BIC dosahoval lepších výsledkov L2 odhad, hodnoty RSS a RSAV zodpovedajú teórii, čiže RSS preferuje L2 odhad, naopak RSAV preferuje L1 odhad. Čo sa týka vychýlenosti odhadov, teda rozdielu $\mid \bar{\boldsymbol{a}}\text{-}\beta_0 \mid$ a ich rozptylov, tak ako absolútny člen β_0 tak lineárny člen β_1

odhaduje s menším vychýlením L1 odhad, ale zároveň má väčší rozptyl, to znamená, že L1 odhad je častejšie viac nepresný ako L2 odhad.

Posledným kritériom, ktoré si uvedieme sú intervalové odhady, budeme ich počítať v tvare:

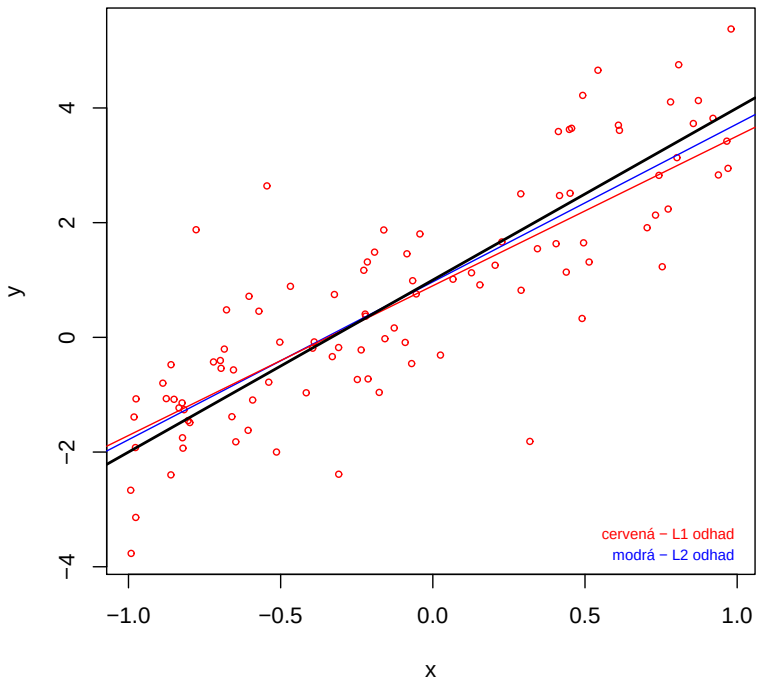
$$\bar{\boldsymbol{a}} \pm \mathsf{t}_{1-\frac{\alpha}{2}}(df)\sqrt{\mathsf{var}\boldsymbol{a}}\frac{1}{\sqrt{m}} \quad \text{resp} \quad \bar{\boldsymbol{b}} \pm \mathsf{t}_{1-\frac{\alpha}{2}}(df)\sqrt{\mathsf{var}\boldsymbol{b}}\frac{1}{\sqrt{m}}.$$

kde $\mathsf{t}_{1-\frac{\alpha}{2}}(df)$ je kvantil Studentovho t-rozdelenia od df stupňoch voľnosti, my budeme uvažovať 95% intervalové odhady, čiže $\alpha = 0.05$. Pre intervalové odhady, požadujeme aby mali skutočnú hodnotu parametrov čo najviac uprostred intervalu, to znamená minimálne, že požadujeme minimálne vychýlenie odhadov a čo najkratšiu dĺžku intervalu to znamená malý rozptyl odhadov.

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9916, 1.0066)	(2.9947, 3.0211)
L2 odhad	(0.9930, 1.0047)	(2.9985, 3.0203)

Tabuľka 5.2. Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{N}(0, 1)$

Pre ďalšiu predstavu ako sa odhady odchýlili od skutočnej závislosti medzi y a x si uvedieme grafy, v ktorých si znázorníme regresnú priamku vypočítanú L1 a L2 odhadom spolu s priamkou skutočnej závislosti. Vyberieme vždy jeden názorný prípad (jedno opakovanie testu). Čierna priamka značí skutočnú závislosť medzi y a x.

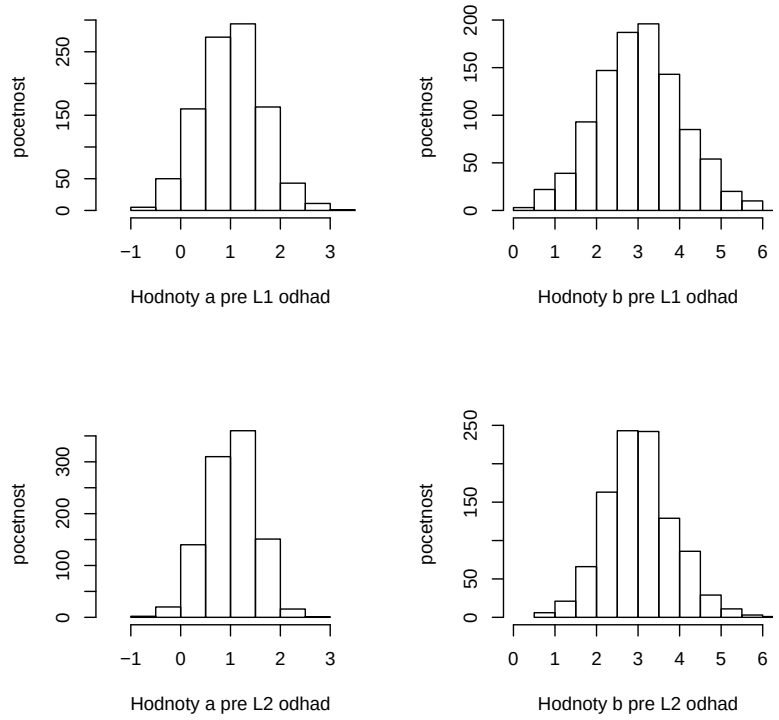


Obr.5.2 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

5.1.2 $\boldsymbol{e} \sim \mathsf{N}(0,5)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert\ \bar{\boldsymbol{a}}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{\boldsymbol{b}}\text{-}\beta_1\ \rvert$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0032**	0.0337	0.3977	1.0780	615.8447	621.0550	2486.2747	**393.1105**
L2 odhad	0.0147	**0.0310**	**0.2494**	**0.6910**	**608.9743**	**616.7899**	**2458.0884**	395.3534

Tabuľka 5.3 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{N}(0,5)$



Obr 5.3 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{N}(0,5)$

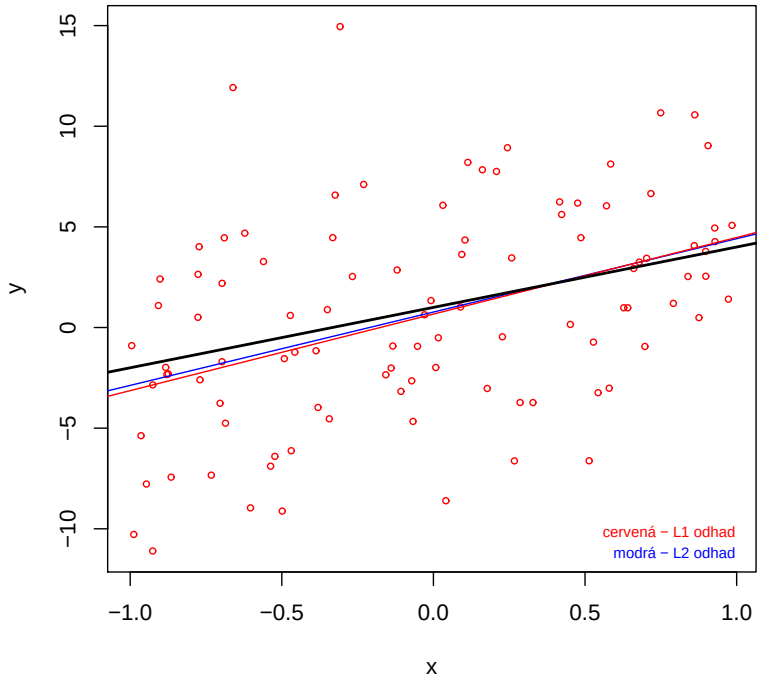
Z tabuľky vidíme, že lepším odhadom je znova L2 odhad. Môžme si všimnúť, že rozdiel priemeru odhadov parametru β_0 a skutočnej hodnoty parametru β_0, je menší pre L1 odhad, takisto ako v predošlom teste. Táto situácia môže byť ale úplne odlišná keby sme tento test urobili znova. Tento fakt vyplýva z toho, že obidva odhady sú nestranné a raz z tohto pohľadu vyjde lepšie L1 odhad, inokedy zasa L2 odhad.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9641, 1.0423)	(2.9693, 3.0980)
L2 odhad	(0.9837, 1.0456)	(2.9795, 3.0826)

Tabuľka 5.4. Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{N}(0,5)$

Vidíme, že 95% interval spoľahlivosti pre obidva odhadované parametry, je celkovo užší pre L2 odhad.

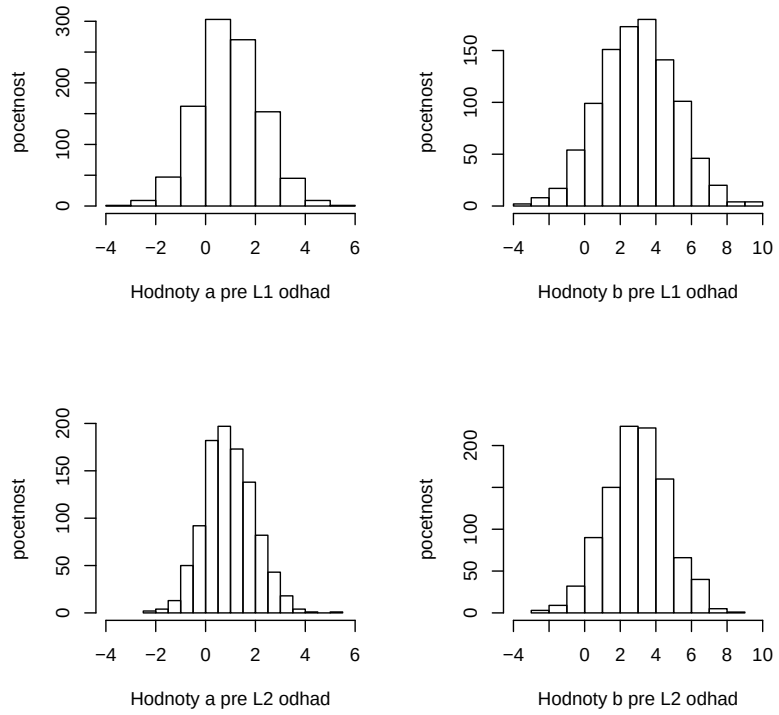


Obr. 5.4 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

5.1.3 $\boldsymbol{e} \sim \mathsf{N}(0, 10)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

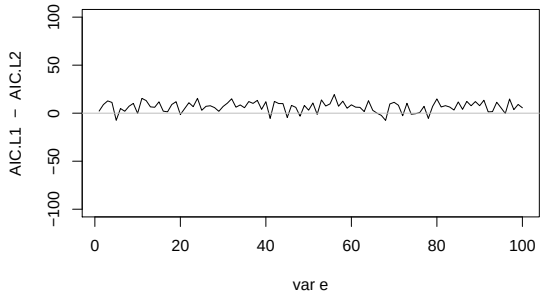
	$\lvert\ \bar{a}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{b}\text{-}\beta_1\ \rvert$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0486**	0.0306	1.5962	4.5037	753.2687	758.4791	9828.0281	**781.3949**
L2 odhad	0.0507	**0.0199**	**1.0097**	**2.9717**	**746.4711**	**754.2866**	**9716.5579**	786.0296

Tabuľka 5.5 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{N}(0, 10)$



Obr 5.5 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{N}(0, 10)$

Pri normálnom rozdelení s takto veľkým rozptylom, sú hodnoty AIC, RSS a pod. relatívne veľké. Na vhodnosť odhadu, ale hodnota rozptylu rozdelenia náhodných chýb, nemá vplyv. Hodnoty kritérií sa úmerne zväčšujú pre obidva odhady v rovnakej miere, to znamená že rozdiel medzi hodnotami AIC kritéria pre L1 a L2 odhad, nemá so zväčšujúcim sa rozptylom rozdelenia náhodných chýb, ani klesajúci ani rastúci charakter, ako ukazuje nasledujúci graf 5.6.

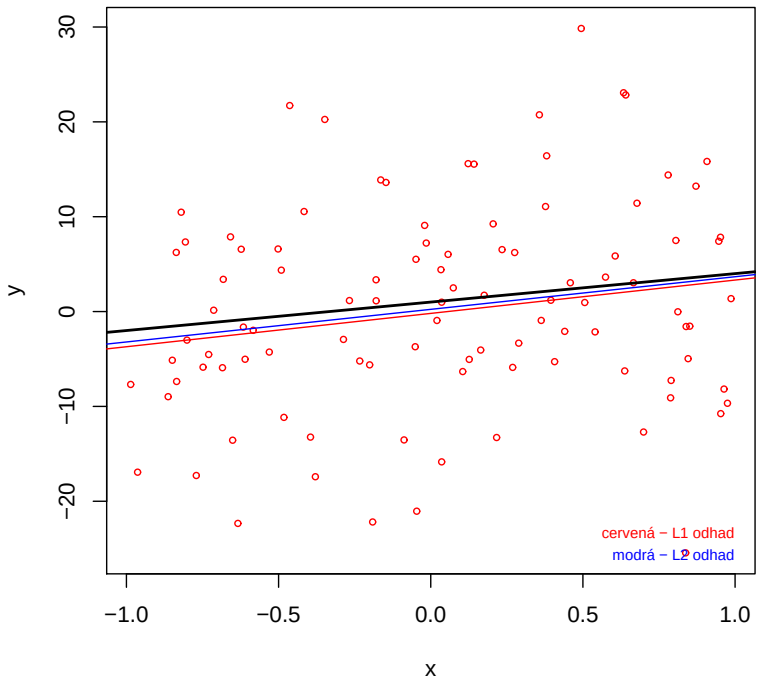


Obr 5.6 Závislosť kritéria AIC na rozptyle rozdelenia náhodného vektoru$\boldsymbol{e}$

Intervalové odhady:

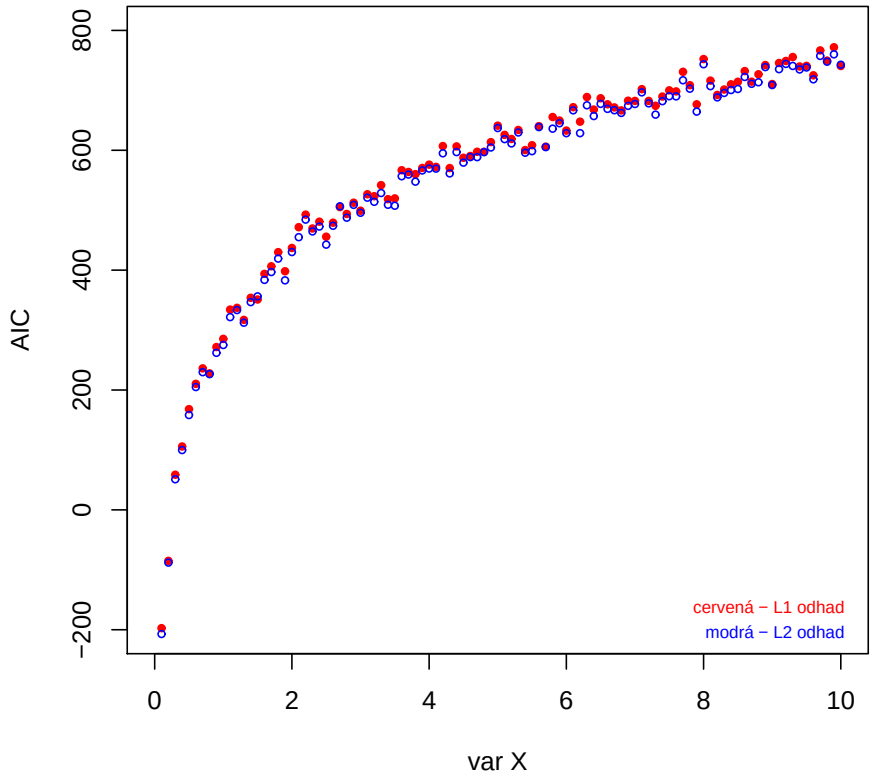
	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.8731, 1.0297)	(2.8378, 3.1009)
L2 odhad	(0.8870, 1.0115)	(2.8733, 3.0870)

Tabuľka 5.6 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{N}(0, 10)$



Obr 5.7. Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

Pre, porovnanie ako sa jednotlivé odhady správajú pri rôznej hodnote rozptylu pre normálne rozdelenie, si uvedieme graf závislosti AIC na hodnote rozptylu rozdelenia náhodných chýb.



Obr 5.8 Závislosť kritéria AIC na rozptyle rozdelenia náhodného vektoru$\boldsymbol{e}$

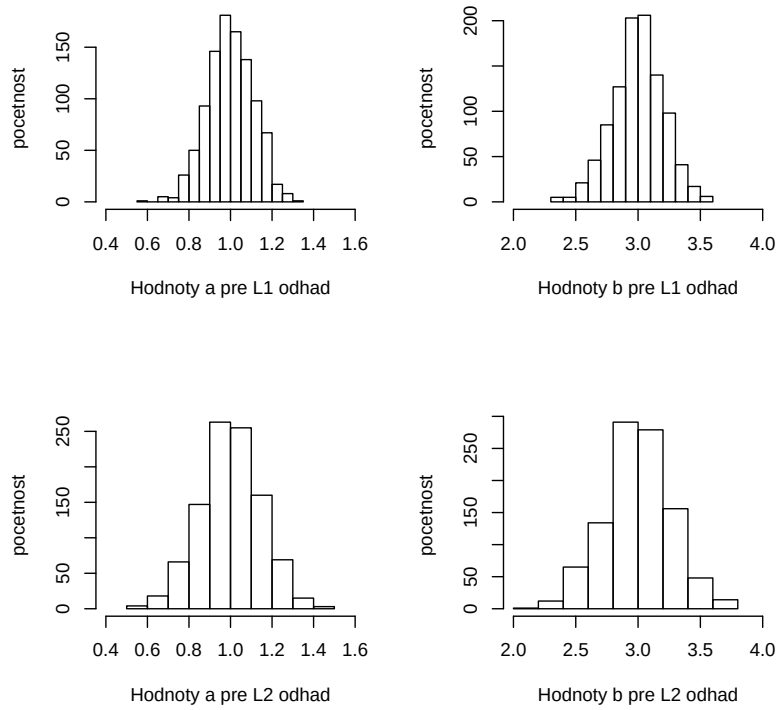
5.2 Laplaceovo rozdelenie náhodných chýb

Pri Laplaceovom, alebo tiež dvojito-exponenciálnom rozdelení, je väčšia pravdepodobnosť, že pri generovaní dát, vygenerujeme odľahlé pozorovanie. Z toho vyplýva, kedže L1 odhad je robustnejší voči odľahlým pozorovaniam, že práve L1 odhad by mal byť lepším odhadom pri tomto type rozdelenia náhodných chýb. Poďme sa pozrieť, ako obidva odhady dopadli v testoch.

5.2.1 $\boldsymbol{e} \sim \mathsf{DEx}(0,1)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert\ \bar{\boldsymbol{a}}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{\boldsymbol{b}}\text{-}\beta_1\ \rvert$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	0.0012	**0.0007**	**0.0122**	**0.0409**	**339.9358**	**345.1461**	199.7115	**99.1567**
L2 odhad	**0.0002**	0.0022	0.0210	0.0693	355.5395	363.3550	**197.6409**	100.0587

Tabuľka 5.7 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{DEx}(0,1)$



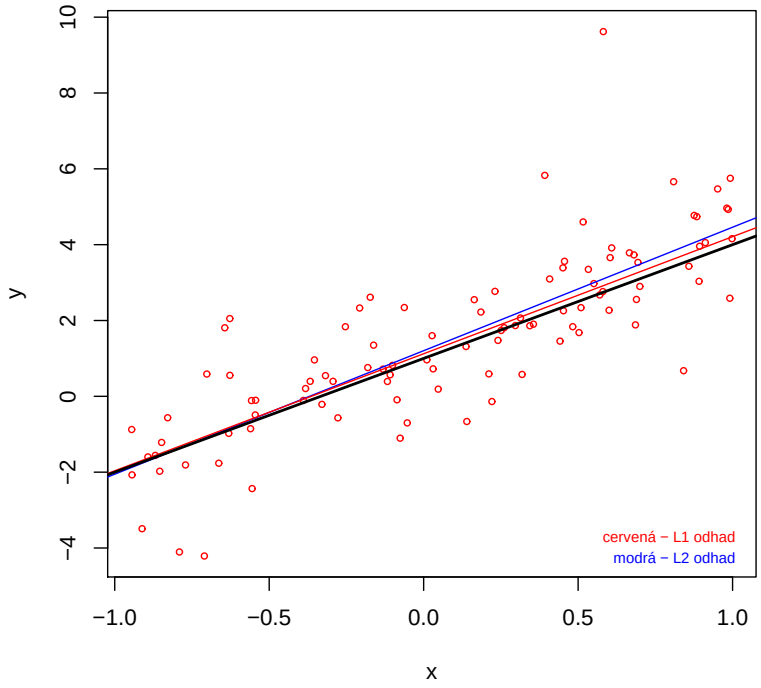
Obr 5.9 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{DEx}(0,1)$

Lepší odhad, ako vidíme z tabuľky a nakoniec aj z histogramov, je L1 odhad. Má menší rozptyl odhadov, takisto ako hlavné kritériá AIC a BIC.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9943, 1.0081)	(2.9867, 3.0118)
L2 odhad	(0.9912, 1.0092)	(2.9814, 3.0141)

Tabuľka 5.8 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{DEx}(0,1)$



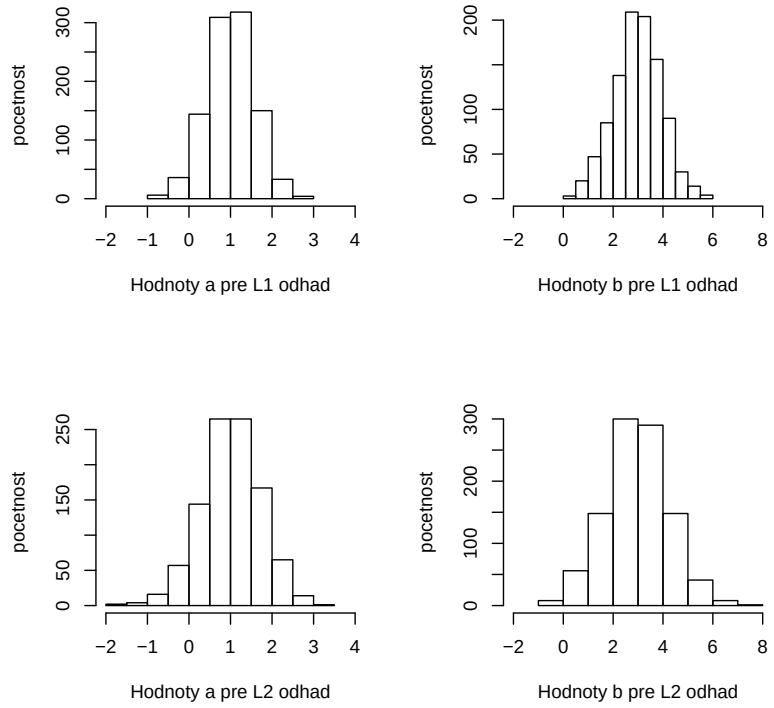
Obr 5.10 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

Z grafu môžme vidieť, ako sa jednotlivé odhady pre dané data odchýlili od skutočnej závislosti medzi vektormi $\mathbf{x}$ a $\boldsymbol{Y}$. Situácia sa môže zmeniť v prospech odhadu L2 pri ďalšom opakovaní testu, ja som vybral príklad ktorý vystihuje situáciu, ktorá bola najčastejšia pri tomto type rozdelenia náhodných chýb.

5.2.2 $\boldsymbol{e} \sim \mathsf{DEx}(0,5)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert\ \bar{\boldsymbol{a}}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{\boldsymbol{b}}\text{-}\beta_1\ \rvert$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0027**	**0.0129**	**0.3261**	**0.8941**	**660.7493**	**665.9597**	4915.9784	**493.0728**
L2 odhad	0.0047	0.0415	0.5188	1.5530	675.8875	683.7030	**4864.9243**	497.5087

Tabuľka 5.9 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{DEx}(0,5)$



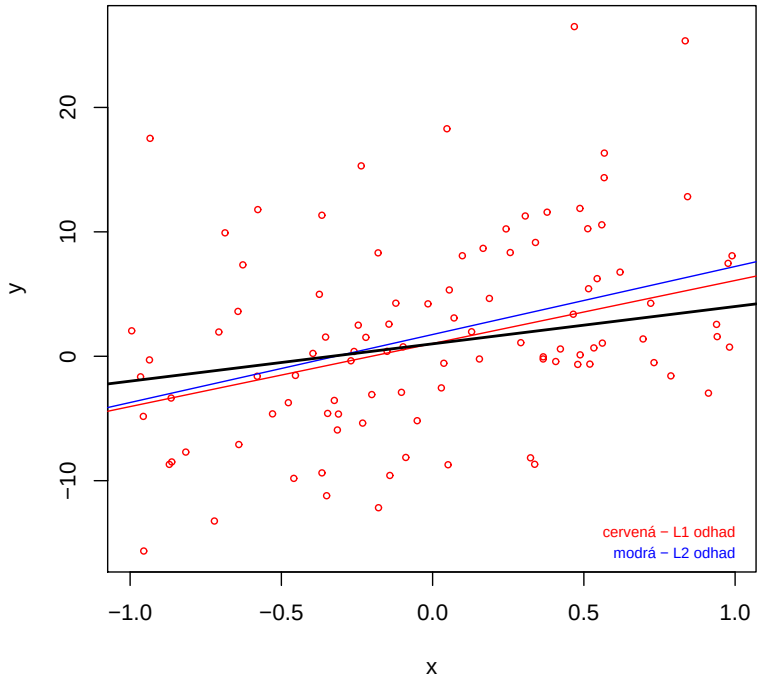
Obr 5.11 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{DEx}(0,5)$

Situácia je rovnaká ako pri zväčšovaní rozptylu pri normálnom rozdelení náhodných chýb. Môžme si všimnúť, že kritériá sú väčšie ako pri $\boldsymbol{e} \sim \mathsf{N}(0,5)$, čo podporuje fakt, že odľahlé pozorovania "zhoršujú" kvalitu a presnosť oboch odhadov. Alebo inými slovami, pri prechode od normálneho rozdelenia k laplaceovmu rozdeleniu náhodných chýb, sa kvalita odhadov zhorší, konkrétne sa zväčší rozptyl odhadov, AIC, BIC, RSS a $RSAV$.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9619, 1.0327)	(2.9285, 3.0457)
L2 odhad	(0.9600, 1.0493)	(2.8812, 3.0357)

Tabuľka 5.10 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{DEx}(0,5)$

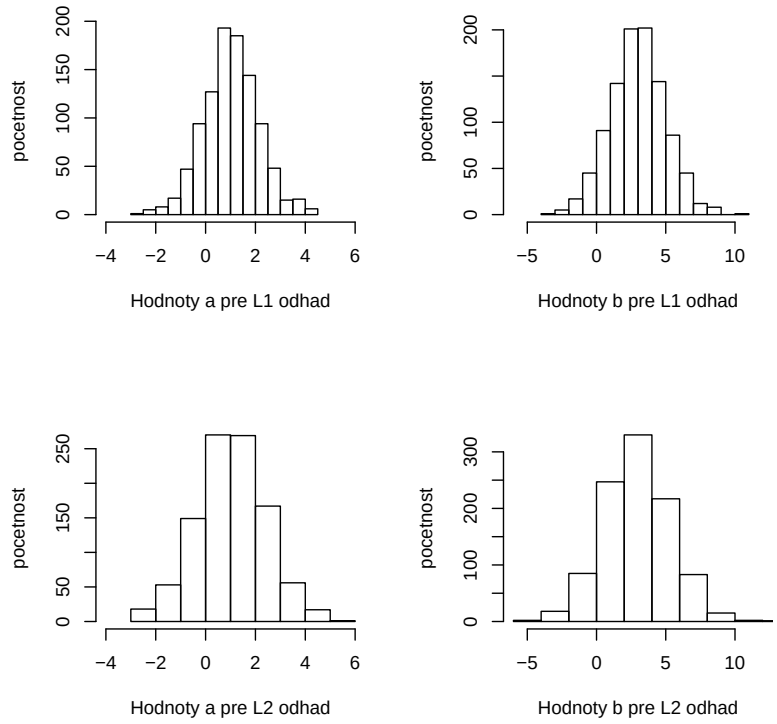


Obr 5.12 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

5.2.3 $\boldsymbol{e} \sim \mathsf{DEx}(0, 10)$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\mid \bar{\boldsymbol{a}}\text{-}\beta_0 \mid$	$\mid \bar{\boldsymbol{b}}\text{-}\beta_1 \mid$	var $\boldsymbol{a}$	var $\boldsymbol{b}$	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0315**	**0.0153**	**1.2235**	**3.9219**	**799.5083**	**804.7187**	19733.6562	**987.0818**
L2 odhad	0.0424	0.0856	1.9080	5.8301	814.7221	822.5376	**19537.4632**	995.5012

Tabuľka 5.11 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{DEx}(0, 10)$



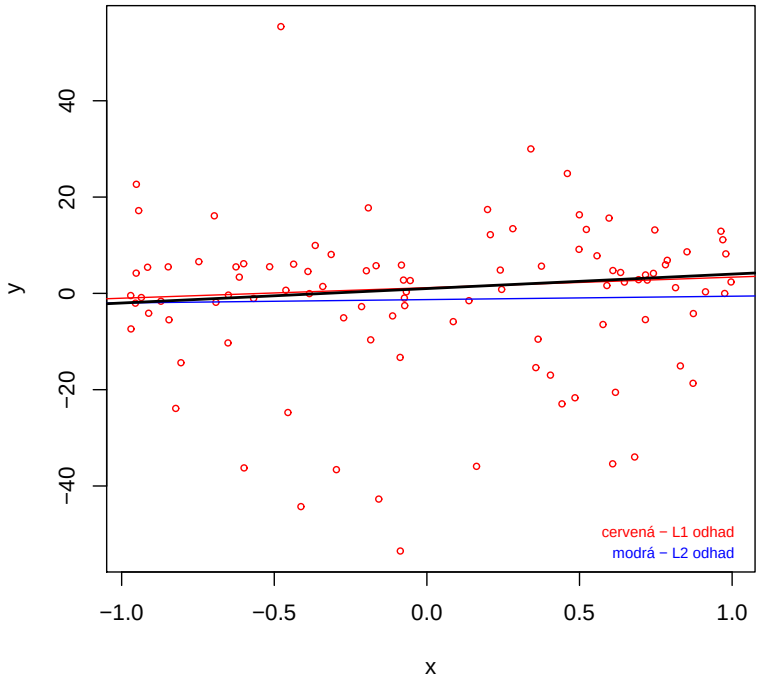
Obr 5.13 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{DEx}(0, 10)$

Všimnime si, že ako u normálneho tak u Laplaceovho rozdelenia, sa pri zvyšujúcom rozptyle náhodných chýb, zvyšuje aj rozptyl odhadov regresných koeficientov. A to ako pre L1 tak pre L2 odhad. Spolu s rozptylom sa zväčšuje aj hodnota výrazu $\mid \bar{\boldsymbol{a}}\text{-}\beta_0 \mid$pre obidva odhady.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9629, 1.1000)	(2.8619, 3.1074)
L2 odhad	(0.9568, 1.1280)	(2.7647, 3.0640)

Tabuľka 5.12 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{DEx}(0, 10)$



Obr 5.14 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

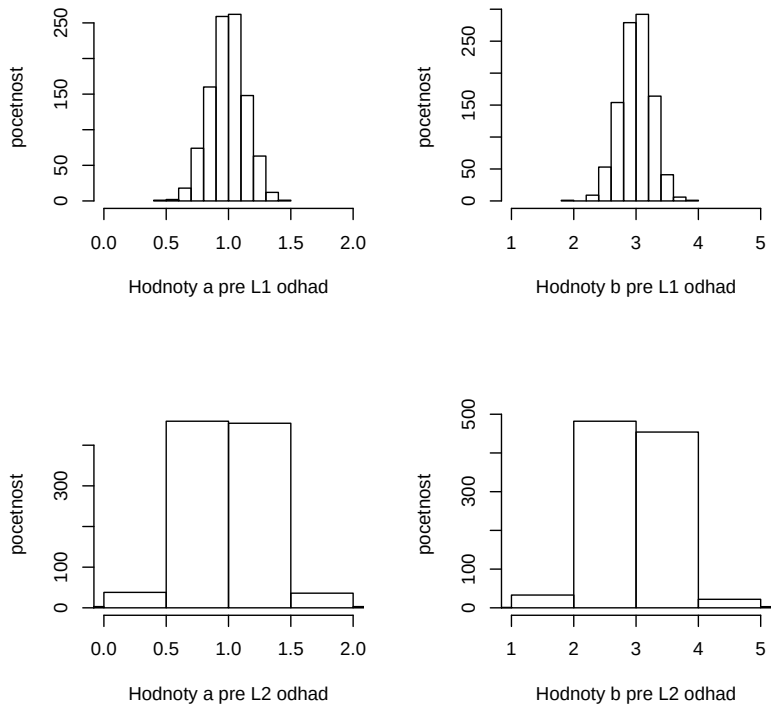
5.3 t-rozdelenie náhodných chýb

Studentovo t-rozdelenie, je rozdelenie, ktoré pri malých počtoch stupňov voľnosti, dáva veľkú pravdepodobnosť na vygenerovanie odľahlého pozorovania. Naopak so zvyšujúcim sa počtom stupňov voľnosti sa toto rozdelenie blíži normovanému normálnemu rozdeleniu, ako sme si uviedli v odstavci 3.5. To znamená, že pre malý počet stupňov voľnosti by mal byť výrazne lepším odhadom L1 odhad a čím sa bude počet stupňov voľnosti zvyšovať, bude mať L2 odhad postupne lepšie výsledky.

5.3.1 $\boldsymbol{e} \sim \mathrm{t}_2$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert \bar{a}\text{-}\beta_0 \rvert$	$\lvert \bar{b}\text{-}\beta_1 \rvert$	var a	var b	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	0.0053	**0.0027**	**0.0200**	**0.0643**	**404.8577**	**410.0681**	1187.7148	**139.0404**
L2 odhad	**0.0018**	0.0190	0.1217	0.3494	466.0474	473.8629	**1168.7558**	143.9567

Tabuľka 5.13 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathrm{t}_2$



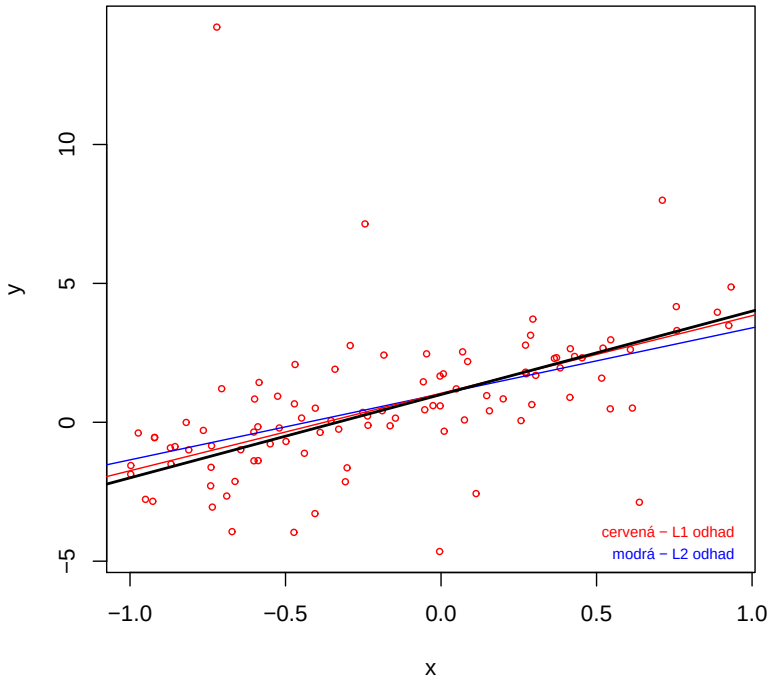
Obr 5.15 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathrm{t}_2$

Z tabuľky vidíme, že rozdiely medzi odhadmi sú veľké. Takisto histogramy ukazujú veľmi veľký rozptyl odhadnutých hodnôt pre L2 odhad, kde na druhej strane, histogram pre L1 odhad, resp hodnoty vypočítané pomocou L1 odhadu majú charakter normálneho rozdelenia.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9859, 1.0035)	(2.9816, 3.0131)
L2 odhad	(0.9802, 1.0234)	(2.9443, 3.0176)

Tabuľka 5.14 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathtt{t}_2$



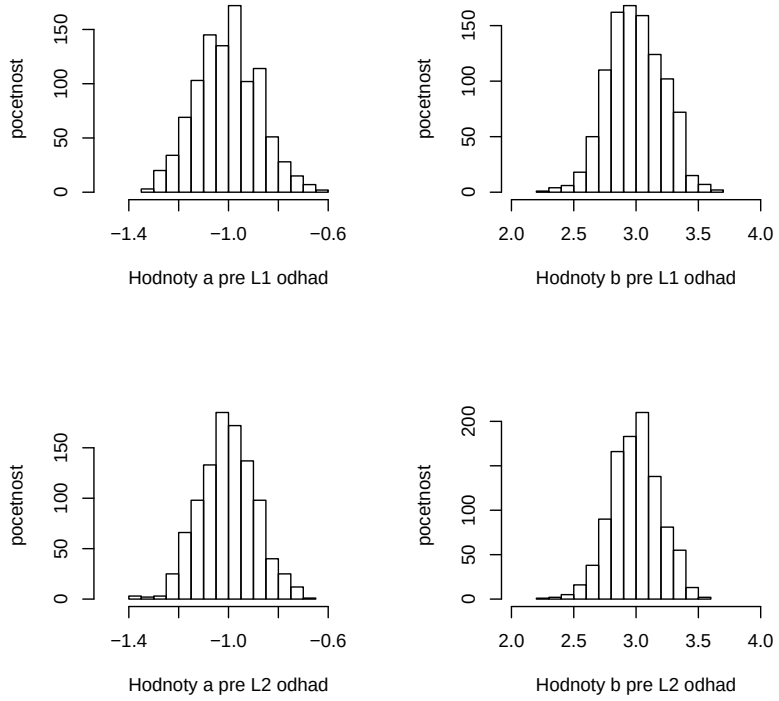
Obr 5.16 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

Priamo z tohto grafu je zrejmé, prečo rozptyl odhadov metódou L2 je o mnoho väčší ako metódou L1. Vidíme 3 odľahlé pozorovania v hornej časti grafu, ktoré regresnú priamku získanú L2 odhadom viac odchýlili od priamky skutočnej závislosti ako regresnú priamku získanú L1 odhadom.

5.3.2 $\boldsymbol{e} \sim \mathrm{t}_{10}$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert\ \bar{a}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{b}\text{-}\beta_1\ \rvert$	var a	var b	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	0.0022	0.0088	0.0173	0.0508	309.8105	**315.0209**	123.6378	**85.1375**
L2 odhad	**0.0019**	**0.0029**	**0.0128**	**0.0373**	**308.5502**	316.3657	**122.3364**	85.6343

Tabuľka 5.15 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathrm{t}_{10}$



Obr 5.17 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathrm{t}_{10}$

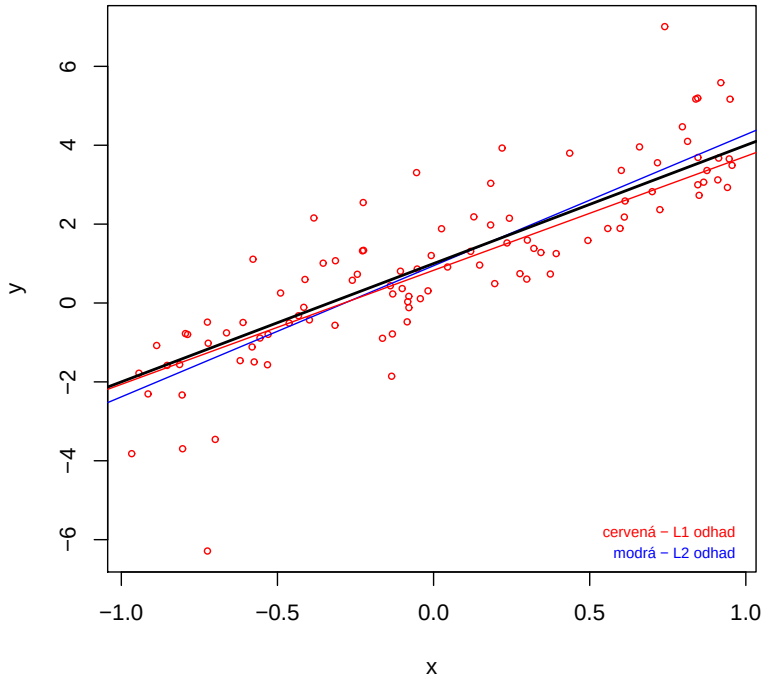
Môžeme si všimnúť, že $k = 10$ je práve tá hodnota počtu stupňov voľnosti, kedy sa hodnoty kritérií začínajú obracať v prospech L2 odhadu. Rozdiely v histogramoch sú zanedbateľné, rovnako ako rozdiely, medzi kritériami AIC a BIC. Zaujímavé je, že podľa kritéria AIC je lepším odhadom L2 odhad, naopak podľa kritéria BIC je lepším odhadom L1 odhad. To len potvrdzuje fakt, že v tomto prípade sú obidve metódy odhadu regresných koeficientov rovnako dobré.

Intervalové odhady:

	Pre odhad parametru β_0	**Pre odhad parametru** β_1
L1 odhad	(0.9941, 1.0104)	(2.9773, 3.0052)
L2 odhad	(0.9949, 1.0089)	(2.9852, 3.0091)

Tabuľka 5.16 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathrm{t}_{10}$

Aj z tabuľky 5.16 vyplýva, že kvalita odhadov pri $k = 10$ je rovnako dobrá. Dĺžka intervalov je nepatrne menšia pre L2 odhad.

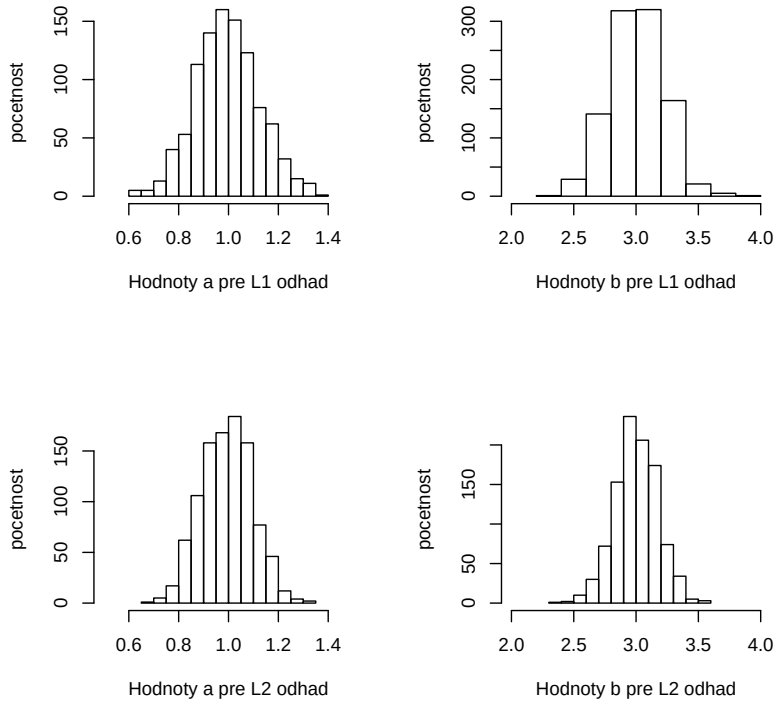


Obr 5.18 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

5.3.3 $\boldsymbol{e} \sim \mathsf{t}_{50}$, $\mathbf{x} = (x_1, \ldots, x_{100})$

	$\lvert\ \bar{a}\text{-}\beta_0\ \rvert$	$\lvert\ \bar{b}\text{-}\beta_1\ \rvert$	var a	var b	$\overline{AIC}$	$\overline{BIC}$	$\overline{RSS}$	$\overline{RSAV}$
L1 odhad	**0.0057**	0.0067	0.0161	0.0463	296.6636	301.8740	103.1007	**79.7174**
L2 odhad	0.0076	**0.0012**	**0.0105**	**0.0292**	**290.5347**	**298.3502**	**101.9258**	80.1923

Tabuľka 5.17 Kritériá porovnávajúce vhodnosť L1 a L2 odhadu, $\boldsymbol{e} \sim \mathsf{t}_{50}$



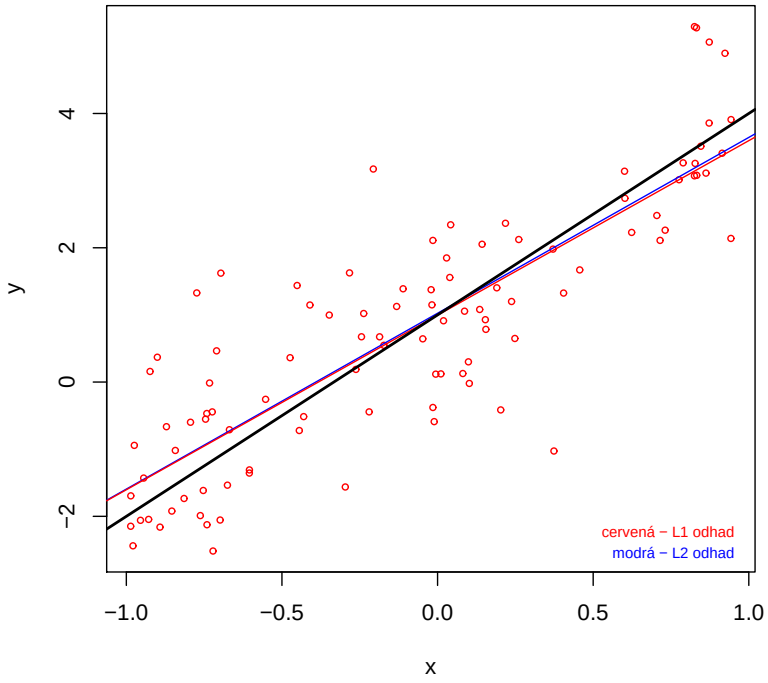
Obr 5.19 Histogramy odhadnutých parametrov pre L1 a L2 odhad, $\boldsymbol{e} \sim \mathsf{t}_{50}$

Pri takto vysokom počte stupňov voľnosti, sa už t- rozdelenie správa ako normované normálne rozdelenie. To potvrdzujú aj hodnoty jednotlivých kritérií, ktoré sú veľmi podobné ako pri prvom teste, kde náhodné chyby mali rozdelenie $N(0, 1)$. Čiže, z toho vyplýva, že L2 odhad by mal byť lepším odhadom. A to nám potvrdzujú aj jednotlivé kritériá. L1 odhad bol pre túto sériu testov síce menej vychýlený ako L2 odhad, ale hodnoty ostatných kritérií hovoria v prospech L2 odhadu.

Intervalové odhady:

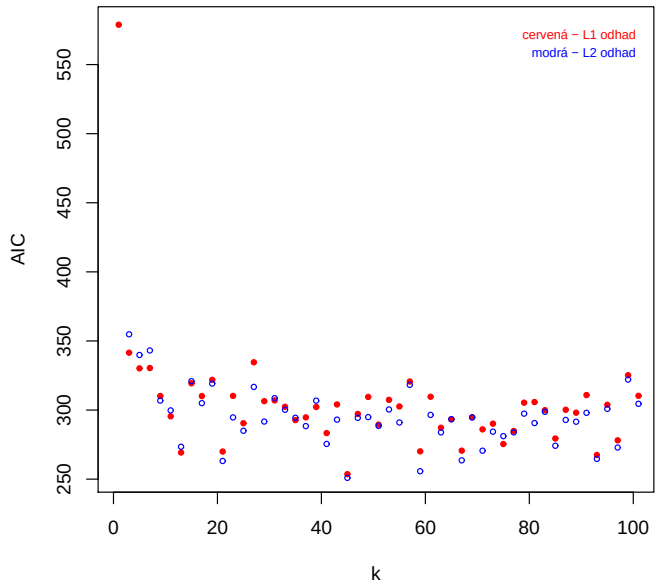
	Pre odhad parametru β_0	**Pre odhad parametru β_1**
L1 odhad	(0.9864, 1.0021)	(2.9933, 3.0200)
L2 odhad	(0.9860, 0.9987)	(2.9906, 3.0118)

Tabuľka 5.18 Intervalové odhady regresných koeficientov, $\boldsymbol{e} \sim \mathsf{t}_{50}$



Obr 5.20 Regresné priamky odhadov L1 a L2 + priamka skutočnej závislosti (čierna farba)

Pre, porovnanie ako sa jednotlivé odhady správajú pri rôznom počte stupňov voľnosti, si uvedieme graf závislosti AIC na počte stupňov voľnosti k, t-rozdelenia náhodných chýb $\boldsymbol{e}$.



Obr 5.21 Závislosť kritéria AIC na počte stupňov voľnosti k t-rozdelenia náhodných chýb $\boldsymbol{e}$

Kapitola 6

Zhrnutie výsledkov testov

Cieľom tejto práce bolo porovnanie vhodnosti metód odhadu regresných koeficientov v závislosti na rozdelení vektoru náhodných chýb $\boldsymbol{e}$ v regresnom modeli. Skúmali sme dva typy odhadov. Metódu najmenšej absolútnej odchýlky (LAD) a metódu najmenších štvorcov.

Testy som rozdelil do troch hlavných skupín, podľa typu rozdeleniu vektoru $\boldsymbol{e}$. Každú skupinu ďalej delím podľa veľkosti rozptylu jednotlivého rozdelenia vektoru $\boldsymbol{e}$. V prvej skupine testov, majú náhodné chyby normálne rozdelenie. Pri tomto type rozdelenia, dosahuje lepších výsledkov podľa očakávania L2 odhad. Normálne rozdelenie má nulovú špicatosť. To znamená, že hodnoty vektoru $\boldsymbol{e}$ ktoré sme generovali, sú rozdelené blízko strednej hodnoty s pravdepodobnosťou ktorá sa rovná povedzme p, takisto ako pravdepodobnosť, že vygenerujeme extrémnu hodnotu, označme si túto pravdepodobnosť q. L2 odhad bol lepším odhadom pre normálne rozdelenie náhodných chýb aj pri vysokom rozptyle. To je spôsobené tým, že data pri tomto type rozdelenia a pri tak veľkom rozptyle sú stále konzistentné, tzn., že vzdialenosť od strednej hodnoty je síce celkovo väčšia, ale pravdepodobnosť na vygenerovanie odľahlej hodnoty je stále rovnako malá.

Pre náhodné chyby s Laplaceovým rozdelením, ktoré má narozdiel od normálneho rozdelenia kladnú špicatosť, je L1 odhad lepším odhadom pre tento typ rozdelenia. Je to dané tým, že rozdelenia s kladnou špicatosťou majú síce pravdepodobnosť, že hodnoty budú bližšie strednej hodnote väčšiu ako p, ale zároveň pravdepodobnosť, že vygenerujeme extrémnu hodnotu je taktiež väčšia než q. To ale robí s L2 odhadu, nevhodnú a výrazne horšiu metódu pre odhad regresných koeficientov s Laplaceovým rozdelením náhodných chýb ako L1 odhad, pretože pre extrémne hodnoty pozorovaní, má táto odľahlá hodnota príliš veľký vplyv na polohu regresnej priamky a tým odchýli priamku v snahe znížiť štvorec vzdialenosti priamky od tohto bodu, a tým pádom negatívne ovplyvní (zväčší) niektoré z ďalších reziduí.

Posledným typom rozdelenia je Studentovo t-rozdelenie. Pre počet stupňov voľnosti ≤ 4 nie je špicatosť definovaná, ale ako môžeme vidieť na obrázku 5.8, pravdepodobnosť na vygenerovanie extrémneho pozorovania je veľmi vysoká, to znamená, že pre malý počet stupňov voľnosti je lepší L1 odhad. S rastúcim počtom stupňov voľnosti sa lepším odhadom stáva L2 odhad, kedže sa toto rozdelenie blíži

normálnemu rozdeleniu. Približný bod zlomu, kedy sa L2 odhad stáva lepším odhadom je pre $k = 10$.

Literatúra

[1] Anděl, J.: Základy matematické statistiky. Matfyzpress, Praha 2005.

[2] Anděl, J.: Matematika náhody. Matfyzpress, Praha 2007.

[3] Dupač, V., Hušková, M.: Pravděpodobnost a matematická statistika. Nakladatelství Karolinum, Praha 2005.