LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKĀ TICAMĪBAS FUNKCIJA LINEĀRAI REGRESIJAI

BAKALAURA DARBS

Autors: **Līga Mangule**
Studenta apliecības Nr.: lm11074
Darba vadītājs: asoc.prof. Jānis Valeinis

RĪGA 2015

Anotācija

Bakalaura darbā tiek aplūkota empīriskās ticamības (EL) metode lineārai regresijai ar mērķi konstruēt empīriskās ticamības apgabalu vienfaktora lineārās regresijas koeficientiem. Konstruētie EL ticamības apgabali tiek salīdzināti ar parametriskās ticamības apgabaliem. Darbā tiek arī veikta jaudas analīze simulētiem datiem ar dažādi sadalītiem atlikumiem, kur EL metodes sniegums tiek salīdzināts ar t-testu lineārai un robustai regresijai. Tiek secināts, ka empīriskās ticamības metode ir strādā labi, taču tā ir jūtīga pret izlecēju ietekmi. Empīriskiem ticamības apgabaliem pārsvarā gadījumu nav elipses forma un viennozīmīgs novietojums attiecībā pret parametrisko ticamības apgabalu.

Atslēgas vārdi: empīriskā ticamības funkcija, lineāra regresija, robusta regresija, empīriskais ticamības apgabals, parametriskais ticamības apgabals

Abstract

Bachelor's Thesis outlines the empirical likelihood (EL) method for linear regression and it is used to construct likelihood ratio confidence regions for regression parameters. EL confidence regions are compared with parametric confidence regions. In thesis also statistical power of EL method for linear regression with differently distributed errors has been analyzed and the results have been compared with results of t-test for linear and robust regression. It was concluded that EL method works very well however method is sensitive to outliers. In most cases EL confidence regions are not ellipsoid and have ambiguous position in relation to the parametric confidence regions.

Keywords: empirical likelihood, linear regression, robust regression, empirical likelihood confidence regions, parametric confidence regions

Saturs

Ievads

Viens no praksē visplašāk lietotajiem statistikas rīkiem, kas tiek izmantots, lai pētītu viena mainīgā atkarību no cita, ir lineārā regresija, taču lai varētu pielietot lineārās regresijas modeli, ir nepieciešams, lai izpildās pieņēmumi par atlikumu neatkarību un normalitāti. Bieži vien praksē aplūkotajiem datiem šie pieņēmumi neizpildās, jo tiek aplūkotas izlases ar pārāk mazu apjomu vai datos ir *izlecēji* jeb novērojumi, kuri nepakļaujas kādai datu struktūrai.

Arvien vairāk datu analīzei tiek izmantoti neparametriskās statistikas rīki. Viens no tiem ir empīriskās ticamības metode, kuru 20. gadsimta beigās izstrādāja Owen [1], [2], [3]. Līdzīgi kā maksimālās ticamības metodi, arī empīrisko ticamības metodi var pielietot parametru novērtēšanai, hipotēžu pārbaudei, ticamības intervālu un ticamības apgabalu konstruēšanai. Chen un Keilegom [11] 2009. gadā publicēja apkopojumu par empīriskās ticamības metodes pielietošanu dažādu regresiju veidiem, tai skaitā lineārai regresijai.

Bakalaura darba galvenais mērķis ir konstruēt empīriskās ticamības metodes ticamības apgabalu vienfaktora lineārās regresijas koeficientiem. Papildus tiek veikta jaudas analīze, lai secinātu, kādiem datiem empīriskās ticamības metode strādā vislabāk, kā arī lai salīdzinātu metodes sniegumu ar t-testu lineārām un robustām regresijām.

Darba izstrādes laikā tika veikti šādi uzdevumi:

1. iepazīties ar empīriskās ticamības attiecības testu;
2. iepazīties ar robustas regresijas jēdzienu;
2. veicot simulācijas, analizēt empīriskās ticamības metodes efektivitāti;
3. atrisināt nelineāru vienādojumu sistēmu programmā R;
4. salīdzināt parametriskās un empīriskās ticamības apgabalus reālās un simulētās datu problēmās.

Darbs sastāv no ievada, 6 nodaļām, secinājumiem, izmantotās literatūras saraksta un pielikuma. 1.nodaļā apskatīta teorija par lineāru regresiju, 2. nodaļā izklāstīts par empīrisko ticamības funkciju, 3. nodaļā definēta empīriskā ticamības funkcija lineārai metodei, 4. nodaļā aprakstīti M-novērtējumi, 5. nodaļā apskatīta teorija par robustu regresiju, 6. nodaļā konstruēti ticamības apgabali un apkopoti jaudas analīzes rezultāti. Visi aprēķinu rezultāti tika iegūti, izmantojot programmu R.

1. Lineārā regresija

Lineārai regresijai plaši tiek izmantoti divu veidu izlases modeļi. Vienā gadījumā skaidrojošais mainīgais X_i ir gadījuma lielums, otrā gadījumā $X_i = x_i$ ir fiksēts. Praktiski bieži tiek novēroti gadījuma lielumu pāri (X_i, Y_i), taču tie tiek analizēti, it kā X_i būtu fiksēti. Turpmāk tiks aplūkots modelis, kad X_i ir gadījuma lielums. Šīs nodaļas teorētiskais materiāls sagatavots, balstoties uz materiālu [4].

Pieņemsim, ka mums ir n novērojumi (X_i,Y_i), $X \in \mathbb{R}^k, Y \in \mathbb{R}$. Daudzfaktoru lineārās regresijas modelī pieņem, ka mainīgais Y ir atkarīgs no k neatkarīgiem skaidrojošajiem mainīgajiem:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.1}$$

Tiek pieņemts, ka ε_i ir neatkarīgi un vienādi sadalīti, $\varepsilon_i \sim N(0, \sigma^2)$.

Lai pierakstītu lineārās regresijas vienādojumu matricu formā, ieviesīsim četras matricas:

- atkarīgā mainīgā Y_i novērojumu matricu Y ar dimensiju $n \times 1$;
- skaidrojošo mainīgo novērojumu matricu X ar dimensiju $n \times (k+1)$;
- parametru matricu β ar dimensiju $(k+1) \times 1$;
- gadījuma kļūdu matricu ε ar dimensiju $n \times 1$.

Modelis matricu formā pierakstāms šādi:

$$Y = X\beta + \varepsilon,$$

kur

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Lai novērtētu β, tiek izmantota mazāko kvadrātu metode.

1. Definīcija. [7] Par mazāko kvadrātu metodes novērtējumu parametram β sauc $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_n)^T$, kas minimizē atlikumu kvadrātu summu

$$RSS = (Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} X_{ij}\beta_j \right)^2.$$

Ja $X^T X$ ir pilna ranga matrica, tad mazāko kvadrātu metodes novērtējums ir

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T Y.$$

2. Definīcija. Hipotēžu pārbaudei $H_0 : \beta_j = \beta_j^0$ statistika ir

$$\frac{\beta_j - \beta_j^0}{SE(\beta_j)}, \tag{1.2}$$

kura ir sadalīta pēc t_{n-k-1} sadalījuma, ja H_0 ir spēkā.

3. Definīcija. $100(1-\alpha)$ ticamības apgabals visiem parametriem vektorā $\beta \in \mathbb{R}^{k+1}$ tiek iegūts $(k+1)-$dimensionālā telpā no nevienādības

$$\frac{(\hat{\beta} - \beta) X^T X (\hat{\beta} - \beta)}{(k+1)\hat{\sigma}^2} \leqslant F_{p,n-k-1,1-\alpha}, \tag{1.3}$$

kur $F_{p,n-k-1,1-\alpha}$ ir $F_{p,n-k-1}$ sadalījuma $1-\alpha$ kvantile.

1.1. Vienfaktora lineārā regresija

Vienfaktora lineārās regresijas gadījumā modeļa vienādojums ir formā

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1.4}$$

un mazāko kvadrātu metodes novērtējums parametriem β_0 un β_1 ir

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}, \tag{1.5}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{1.6}$$

Var pierādīt, ka parametru β_0 un β_1 standartkļūdu novērtējumi ir

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}}, \tag{1.7}$$

$$\widehat{SE}(\hat{\beta}_0) = \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \overline{X})^2}}, \tag{1.8}$$

kur nenovirzīts σ^2 novērtējums ir

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}{n-2}. \tag{1.9}$$

1.2. Vienfaktora lineārā regresija caur sākumpunktu

Pieņemot, ka $\beta_0 = 0$, iegūstam regresijas caur sākumpunktu vienādojumu (*Regression through the origin*, RTO):

$$Y_i = \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.10}$$

Ja $\beta_0 = 0$, tad parametru novērtējumi atšķirās:

$$\hat{\beta}_{RTO} = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\,, \tag{1.11}$$

$$\widehat{SE}(\hat{\beta}_{RTO}) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} X_i^2}}\,, \tag{1.12}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}_1 X_i)^2}{n-1}\,. \tag{1.13}$$

Ievērosim, ka šajā gadījumā brīvības pakāpju skaits ir $n-1$, nevis $n-2$.

2. Empīriskā ticamības funkcija

Šajā nodaļā definēsim empīriskās ticamības funkciju un empīriskās ticamības metodi vispārīgā formā.

4. Definīcija. Ja $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti, tad empīriskā sadalījuma funkcija tiek definēta kā

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leqslant x\}}, \tag{2.1}$$

kur

$$I_{\{X_i \leqslant x\}} = \begin{cases} 1, & X_i \leqslant x \\ 0, & X_i > x \end{cases} \quad ; \ -\infty < x < \infty.$$

5. Definīcija. Ja $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti, $X_1 \sim F$, tad funkcijas F empīriskā (neparametriskā) ticamības funkcija ir

$$L(F) = \prod_{i=1}^{n} p_i, \tag{2.2}$$

kur $p_i = P(X = X_i)$.

1. Teorēma. Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti, $X_1 \sim F$, F_n ir to empīriskā sadalījuma funkcija. Ja $F \neq F_n$, tad $L(F) < L(F_n)$.

Pierādījums. Pieņemsim, ka $z_1 < z_2 < \ldots < z_m$ ir atšķirīgas $\{X_1, \ldots, X_n\}$ vērtības un $n_j \geqslant 1$ ir tādu X_i skaits, kad X_i ir vienāds ar z_j, $p_j = P(z_j = X_i)$ un $\hat{p}_j = n_j/n$. Ja $p_j = 0$ katram $j = 1, \ldots, m$, tad $L(F) = 0 < L(F_n)$, tādēļ pieņemsim, ka visi $p_j > 0$ un ka vismaz vienam j izpildās $p_j \neq \hat{p}_j$. Tā kā $\ln(x) \leqslant x - 1$ katram $x > 0$ ar vienādību tikai, kad $x = 1$, tad

$$\ln\left(\frac{L(F)}{L(F_n)}\right) = \sum_{j=1}^{m} n_j \ln\left(\frac{p_j}{\hat{p}_j}\right) = n \sum_{j=1}^{m} \hat{p}_j \ln\left(\frac{p_j}{\hat{p}_j}\right) \leqslant n \sum_{j=1}^{m} \hat{p}_j \left(\frac{p_j}{\hat{p}_j} - 1\right) < 0.$$

Tātad $L(F) < L(F_n)$. □

Šī teorēma pierāda to, ka empīrisko ticamības funkciju maksimizē empīriskā sadalījuma funkcija. Līdz ar to var definēt empīrisko ticamības funkcijas attiecību

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^{n} np_i. \tag{2.3}$$

Aplūkosim empīrisko ticamības metodi vispārīgā formā, kas ir aprakstīta Qin un Lawless publikācijā [6]. Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti d-dimensionāli gadījuma lielumi ar nezināmu sadalījuma funkciju F un $\theta \in \mathbb{R}^p$. Pieņemsim, ka informācija par F un θ ir dota $r \geqslant p$ neatkarīgu nenovirzītu funkciju $g_j(X,\theta)$ formā, $j = 1, 2, \ldots, r$, tādas, ka $E_F\{g_j(X,\theta)\} = 0$. Pārrakstot vektorformā,

$$g(X,\theta) = (g_1(X,\theta), \ldots, g_r(X,\theta))^T, \qquad E_F\{g(X,\theta)\} = 0.$$

Definēsim profila empīrisko ticamības funkciju:

$$L(\theta) = \max\Big(\prod_{i=1}^{n} p_i \,|\, \sum_{i=1}^{n} p_i g(X_i,\theta) = 0,\ p_i \geqslant 0,\ \sum_{i=1}^{n} p_i = 1\Big). \tag{2.4}$$

Šai funkcijai eksistē viens vienīgs maksimums pie nosacījuma, ka 0 atrodas izliektas čaulas, kuru veido punkti $g(X_1,\theta), g(X_2,\theta), \ldots, g(X_n,\theta)$, iekšienē. Maksimumu var atrast ar Lagranža reizinātāju palīdzību:

$$H = \sum_{i=1}^{n} \ln(p_i) + \lambda_0 \left(1 - \sum_{i=1}^{n} p_i\right) - n\lambda^T \sum_{i=1}^{n} p_i g(X_i,\theta)\,, \tag{2.5}$$

kur λ_0 un $\lambda = (\lambda_1, \ldots, \lambda_r)^T$ ir Lagranža reizinātāji. Atvasinot H pēc p_i, iegūst

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \lambda_0 - n\lambda^T g(X_i,\theta) = 0,$$
$$\sum_{i=1}^{n} p_i \frac{\partial H}{\partial p_i} = n - \lambda_0 = 0 \Rightarrow \lambda_0 = n$$

un

$$p_i = \left(\frac{1}{n}\right) \frac{1}{1 + \lambda^T g(X_i,\theta)}. \tag{2.6}$$

No ierobežojuma $\sum_{i=1}^{n} p_i g(X_i,\theta) = 0$ seko, ka

$$\frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i,\theta)}{1 + \lambda^T g(X_i,\theta)} = 0. \tag{2.7}$$

Lagranža reizinātājs λ ir nosakāms kā funkcija no θ. Tā kā $0 \leqslant p_i \leqslant 1$, parametriem λ un θ jāapmierina $1 + \lambda^T g(X_i,\theta) \geqslant 1/n$ katram i. Fiksētam θ aplūkosim kopu $D_\theta = \{\lambda : 1 + \lambda^T g(X_i,\theta) \geqslant 1/n\}$. D_θ ir izliekta, slēgta un ierobežota, ja 0 pieder punktu $g(X_i,\theta)$ veidotās

izliektās čaulas iekšienē. Turklāt,

$$\frac{\partial}{\partial\lambda}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i,\theta)}{1+\lambda^T g(X_i,\theta)}\right\}=-\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i,\theta)g^T(X_i,\theta)}{(1+\lambda^T g(X_i,\theta))^2}$$

vienmēr ir negatīvs argumentam $\lambda \in D_\theta$. Saskaņā ar inversās funkcijas teorēmu, $\lambda = \lambda(\theta)$ ir nepārtraukti diferencējama pēc θ.

Empīriskā ticamības funkcija parametram θ ir

$$L(\theta)=\prod_{i=1}^{n}\left\{\frac{1}{n}\frac{1}{1+\lambda^T(\theta)g(X_i,\theta)}\right\} \tag{2.8}$$

un empīriskā ticamības attiecības funkcija parametram θ ir

$$R(\theta)=\prod_{i=1}^{n}\frac{1}{1+\lambda^T(\theta)g(X_i,\theta)}. \tag{2.9}$$

Qin un Lawless [6] savā darbā parāda, ka pie zināmiem nosacījumiem empīriskās ticamības attiecībai ir spēkā neparametriskā Vilksa teorēma.

2. Teorēma (Neparametriskā Vilksa teorēma [6])**.** Empīriskās ticamības attiecības statistika hipotēžu pārbaudei $H_0 : \theta = \theta_0$ ir

$$W(\theta_0) = -2\ln R(\theta_0) \xrightarrow{d} \chi^2_p, \tag{2.10}$$

kad H_0 ir spēkā.

Izmantojot šo rezultātu, ir iespējams konstruēt ticamības apgabalu parametram θ.

6. Definīcija. Empīriskās ticamības metodes $100(1-\alpha)\%$ ticamības apgabals parametram θ ir

$$I_{100(1-\alpha)\%} = \{\theta : W(\theta) \leqslant \chi^2_{p,1-\alpha}\}, \tag{2.11}$$

kur $\chi^2_{p,1-\alpha}$ ir χ^2_p sadalījuma $1-\alpha$ kvantile.

1. Piemērs. Pieņemsim, ka $X_1,\ldots,X_n$ ir neatkarīgi un vienādi sadalīti ar nezināmu sadalījuma funkciju F un $\theta = \mu$. Tad $g(X_i,\theta) = X_i - \mu$ un ar Lagranža reizinātāju palīdzību iegūstam

$$R(\mu)=\prod_{i=1}^{n}\frac{1}{1+\lambda(X_i-\mu)},$$

kur $\lambda = \lambda(\mu)$ apmierina nosacījumu

$$\frac{1}{n}\prod_{i=1}^{n}\frac{(X_i-\mu)}{1+\lambda(X_i-\mu)}=0.$$

Logaritmiskā empīriskās ticamības attiecības statistika ir

$$W(\mu) = -2\ln R(\mu) = 2\sum_{i=1}^{n}\ln(1+\lambda(X_i-\mu)).$$

3. Empīriskā ticamības funkcija lineārai regresijai

Šajā nodaļā aplūkota empīriskās ticamības metode daudzfaktoru lineārai regresijai, balstoties uz Chen un Keilegom darbu [11].

Aplūkosim regresijas modeli formā

$$Y_i = m(X_i;\beta) + \varepsilon_i\,, \quad i = 1,\dots,n, \tag{3.1}$$

kur $m(x;\beta) = x\beta$, $\beta \in \mathbb{R}^p$ $(p < n)$ un ε_i - neatkarīgi gadījuma lielumi, kam $E(\varepsilon_i|X_i) = 0$ un $D(\varepsilon_i|X_i) = \sigma^2(X_i)$.

Parametra β mazāko kvadrātu metodes novērtējumu iegūst, minimizējot atlikumu kvadrātu summu

$$S_n(\beta) := \sum_{i=1}^{n} \left(Y_i - m(X_i;\beta)\right)^2.$$

Novērtējums $\hat{\beta}_{LS} = \arg\inf_\beta S_n(\beta)$ ir sekojoša vienādojuma atrisinājums:

$$\sum_{i=1}^{n} \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = 0. \tag{3.2}$$

Profila empīriskā ticamības funkcija parametram β ir

$$L(\beta) = \max\Big(\prod_{i=1}^{n} p_i \,\Big|\, \sum_{i=1}^{n} p_i \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = 0,\ p_i \geqslant 0,\ \sum_{i=1}^{n} p_i = 1\Big). \tag{3.3}$$

Šai funkcijai eksistē viens vienīgs maksimums pie nosacījuma, ka $0 \in \mathbb{R}^p$ atrodas izliektas čaulas, kuru veido telpas $\mathbb{R}^p$ punkti $\{\frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))\}_{i=1}^n$, iekšienē. Maksimumu var atrast ar Lagranža reizinātāju palīdzību:

$$H(p,\lambda_0,\lambda) = \sum_{i=1}^{n} \ln(p_i) + \lambda_0\left(1 - \sum_{i=1}^{n} p_i\right) - n\lambda^T \sum_{i=1}^{n} p_i \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))\,, \tag{3.4}$$

kur $\lambda_0 \in \mathbb{R}$ un $\lambda = (\lambda_1,\dots,\lambda_p)^T$ ir Lagranža reizinātāji un $p = (p_1,\dots,p_n)^T$.

Atvasinot $H(p,\lambda_0,\lambda)$ pēc p_i, var iegūt, ka $\lambda_0 = n$ un

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))}, \tag{3.5}$$

pie tam no ierobežojuma (3.2) seko, ka λ apmierina

$$\sum_{i=1}^{n} \frac{\frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))} = 0. \tag{3.6}$$

Empīriskā ticamības funkcija parametram β ir

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{n} \frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))} \tag{3.7}$$

un empīriskā ticamības attiecības funkcija parametram β ir

$$R(\beta) = \prod_{i=1}^{n} \frac{1}{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))}. \tag{3.8}$$

Logaritmiskā empīriskā ticamības attiecības funkcija ir

$$W(\beta) = -2 \ln R(\beta) = 2 \sum_{i=1}^{n} \ln\{1 + \lambda^T \frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta))\}. \tag{3.9}$$

Pie zināmiem nosacījumiem hipotēžu pārbaudei $H_0 : \beta = \beta^0$ ir

$$W(\beta^0) \xrightarrow{d} \chi^2_p. \tag{3.10}$$

Izmantojot šo rezultātu, ir iespējams konstruēt ticamības apgabalu parametram β^0.

7. Definīcija. Empīriskās ticamības metodes $100(1-\alpha)\%$ ticamības apgabals parametram β ir

$$I_{100(1-\alpha)\%} = \{\beta : W(\beta) \leqslant \chi^2_{p,1-\alpha}\}, \tag{3.11}$$

kur $\chi^2_{p,1-\alpha}$ ir χ^2_p sadalījuma $1-\alpha$ kvantile.

3.1. Empīriskā ticamības funkcija vienfaktora lineārai regresijai

Gadījumā, kad aplūkojam vienfaktora lineāru regresiju $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ i = 1, \ldots, n$, iegūstam, ka

$$\frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = \begin{pmatrix} Y_i - \beta_0 - \beta_1 X_i \\ X_i(Y_i - \beta_0 - \beta_1 X_i) \end{pmatrix}$$

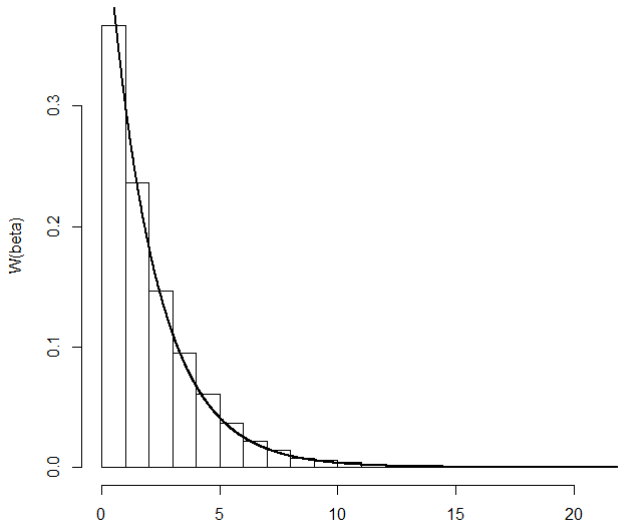
un no nosacījuma (3.6) seko, ka $\lambda = (\lambda_1,\lambda_2)^T$ apmierina šādu nelineāru sistēmu:

$$\begin{cases} \sum_{i=1}^{n} \frac{Y_i - \beta_0 - \beta_1 X_i}{1 + \lambda_1(Y_i - \beta_0 - \beta_1 X_i) + \lambda_2 X_i(Y_i - \beta_0 - \beta_1 X_i)} = 0 \\ \sum_{i=1}^{n} \frac{X_i(Y_i - \beta_0 - \beta_1 X_i)}{1 + \lambda_1(Y_i - \beta_0 - \beta_1 X_i) + \lambda_2 X_i(Y_i - \beta_0 - \beta_1 X_i)} = 0 \end{cases} . \tag{3.12}$$

Empīriskā ticamības attiecības statistika ir

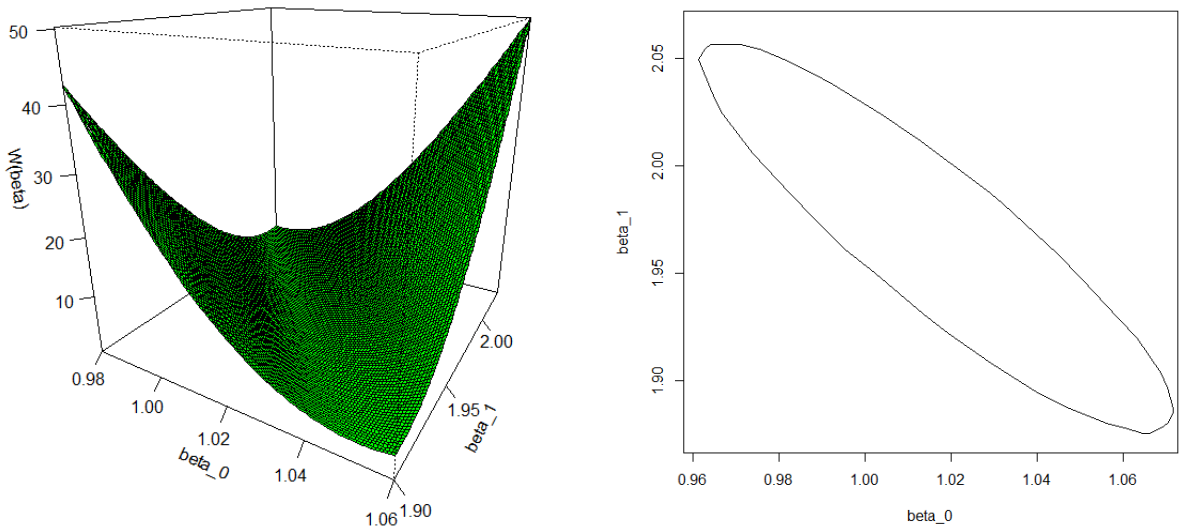
$$W(\beta) = 2\sum_{i=1}^{n} \ln\{1 + \lambda_1(Y_i - \beta_0 - \beta_1 X_i) + \lambda_2 X_i(Y_i - \beta_0 - \beta_1 X_i)\} \xrightarrow{d} \chi^2_2. \tag{3.13}$$

2. Piemērs. Tiek aplūkota vienfaktora lineāra regresija $Y_i = 1 + 2X_i + \varepsilon_i, i = 1,\dots,100$, $X_i \sim U(0,1), \varepsilon_i \sim N(0,0.1)$. Vispirms ar simulāciju palīdzību tiek pārbaudīts, vai empīriskā ticamības attiecības statistika konverģē pēc sadalījuma uz χ^2_2.



3.1. att. Histogramma statistikai $W(\beta)$ $N = 10000$ reižu ģenerētiem (X_i,Y_i) datiem, $n = 100$; χ^2_2 blīvuma funkcija

3.1. attēlā redzams, ka χ^2_2 labi aproksimē histogrammu. Minimizējot logaritmisko empīrisko ticamības attiecības funkciju, tiek iegūts vektora $\beta = (1,2)^T$ novērtējums $\hat{\beta} = (1.015,1.969)^T$. 3.2. (a) attēlā redzams funkcijas $W(\beta)$ grafiks minimuma apkārtnē. 3.2. (b) attēlā redzams, ka īstās β vērtības atrodas empīriskā 95% ticamības apgabala iekšpusē.



(a) (b)

3.2. att. (a) Logaritmiskā empīriskā ticamības attiecības funkcija (b) Empīriskās ticamības metodes 95% ticamības apgabals regresijas koeficientiem

3.2. Empīriskā ticamības funkcija lineārai regresijai caur sākumpunktu

Pieņemot, ka $\beta_0 = 0$, iegūstam regresijas caur sākumpunktu (RTO) vienādojumu:

$$Y_i = \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{3.14}$$

Owen [1] norāda, ka ievietojot $\beta_0 = 0$ vienādojumā (3.2), iegūst

$$\sum_{i=1}^{n} p_i (Y_i - \beta_1 X_i) = 0 \quad \text{un} \tag{3.15}$$

$$\sum_{i=1}^{n} p_i X_i (Y_i - \beta_1 X_i) = 0. \tag{3.16}$$

Mazāko kvadrātu metodes novērtējums $\sum_{i=1}^{n} X_i Y_i / \sum_{i=1}^{n} X_i^2$ atbilst vienādojumam (3.15), tomēr šajā metodē rezultējošo atlikumu summa nav vienāda ar nulli.

Profila empīriskā ticamības attiecības funkcija ir

$$R(\beta_1) = \max\Big(\prod_{i=1}^{n} np_i \mid \sum_{i=1}^{n} p_i (Y_i - \beta_1 X_i) = 0,\ \sum_{i=1}^{n} p_i X_i (Y_i - \beta_1 X_i) = 0,\ p_i \geqslant 0,\ \sum_{i=1}^{n} p_i = 1 \Big). \tag{3.17}$$

Vērtību, kura maksimizē profila empīrisko ticamības attiecības funkciju, apzīmēsim ar $\hat{\beta}_1$. Pie zināmiem nosacījumiem $-2\ln(R(\beta_{LS,1})/R(\hat{\beta}_1)) \xrightarrow{d} \chi_1^2$ un $\beta_{LS,1}$ ir īstā β_1 vērtība.

Tomēr, ja balstāmies uz Chen un Keilegom publikācijā doto aprēķinu gaitu, tad iegūstam,

ka

$$\frac{\partial m(X_i;\beta)}{\partial \beta}(Y_i - m(X_i;\beta)) = X_i(Y_i - \beta_1 X_i)$$

un $\lambda \in \mathbb{R}$ apmierina

$$\sum_{i=1}^{n} \frac{X_i(Y_i - \beta_1 X_i)}{1 + \lambda X_i(Y_i - \beta_1 X_i)} = 0. \tag{3.18}$$

Empīriskā ticamības attiecības statistika ir

$$W(\beta) = 2\sum_{i=1}^{n} \ln\{1 + \lambda X_i(Y_i - \beta_1 X_i)\} \xrightarrow{d} \chi_1^2. \tag{3.19}$$

Tātad, Owen norāda, ka, lai definētu empīrisko ticamības attiecības funkciju RTO gadījumā, nepieciešami 4 ierobežojumi, taču balstoties uz Chen un Keilegom publikāciju, pietiek ar 3 ierobežojumiem. Simulētu datu analīzē tiks aplūkotas abas šīs metodes.

4. M-novērtējums

Robustas regresijas definēšanai ir nepieciešams iepazīt M-novērtējuma jēdzienu [10].

Lai definētu lokācijas M-novērtējumu, iepazīsimies ar lokācijas modeļa jēdzienu. Pieņem, ka

$$x_i = \mu + \varepsilon_i,\ i = 1,...,n,$$

kur kļūdas $\varepsilon_1,...,\varepsilon_n$ ir neatkarīgi gadījuma lielumi ar sadalījuma funkciju F_0. Šādu konstrukciju sauksim par lokālo modeli. No tā seko, ka $x_1,...,x_n$ ir neatkarīgi, vienādi sadalīti gadījuma lielumi ar sadalījuma funkciju

$$F(x) = F_0(x - \mu).$$

Pieņem, ka ε_i blīvuma funkcija ir $f_0 = F_0'$. Tad kopējā blīvuma funkcija novērojumiem jeb ticamības funkcija ir

$$L(x_1,...,x_n;\mu) = \prod_{i=1}^{n} f_0(x_i - \mu).$$

Maksimālās ticamības funkcijas novērtējums (*MLE*) parametram μ ir

$$\hat{\mu} = \hat{\mu}(x_1,...,x_n) = \arg\max_{\mu} L(x_1,...,x_n;\mu). \tag{4.1}$$

Ja f_0 ir visur pozitīva, tad (4.1) var pārrakstīt kā

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu), \tag{4.2}$$

kur $\rho = -\ln f_0$. Piemēram, ja $F_0 = N(0,1)$, tad

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

un neņemot konstanti, $\rho(x) = \frac{x^2}{2}$. Tad

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Ja ρ ir diferencējama funkcija, tad (4.2) var pārrakstīt kā

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0, \tag{4.3}$$

kur $\psi = \rho'$.

Piemēram, ja $\rho(x) = \dfrac{x^2}{2}$, tad $\psi(x) = x$ un (4.3) var uzrakstīt kā

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0,$$

kur $\hat{\mu} = \overline{x}$ ir atrisinājums.

Dotai funkcijai ρ, par lokācijas M-novērtējumu sauc novērtējumu (4.2).

Ja ψ ir monotoni nedilstoša funkcija un $\psi(-\infty) < 0 < \psi(\infty)$, tad (4.3) un līdz ar to arī (4.2) vienmēr būs atrisinājums. Ja ψ ir nepārtrauka un augoša, tad atrisinājums ir viens vienīgs.

Viens no visplašāk lietotajiem lokācijas M-novērtējumiem ir Hūbera novērtējums.

8. Definīcija. Par Hūbera novērtējumu $\hat{\mu}$ sauc lokācijas parametra μ novērtējumu, kam izpildās (4.3) un

$$\psi(x) = \begin{cases} k & , x > k \\ x & , |x| \leqslant k \\ -k & , x < -k \end{cases} \qquad , k \in (0, \infty).$$

Hūbera novērtējums ir viduscelš starp diviem visbiežāk pielietotajiem lokācijas parametru novērtējumiem - vidējo vērtību un mediānu, tas dod kompromisu starp vidējās vērtības efektivitāti un mediānas robustumu. Konstantes k robežgadījumi:

- Ja $k \to \infty$, tad iegūst vidējo vērtību,
- Ja $k \to 0$, tad iegūst mediānu.

Parametru k var interpretēt kā saskaņošanas konstanti, kas nosaka novērtējuma robustuma pakāpi.

Aplūkosim vēl vienu M-novērtējumu.

Pieņem, ka

$$x_i = \sigma\varepsilon_i,\ i = 1, \ldots, n,$$

kļūdas ε_1,...,ε_n ir neatkarīgi gadījuma lielumi ar blīvuma funkciju f_0 un $\sigma > 0$ ir nezināms parametrs. Novērojumi x_1,...,x_n ir ar blīvuma funkciju

$$\frac{1}{\sigma} f_0\left(\frac{x}{\sigma}\right).$$

Maksimālās ticamības funkcijas novērtējums parametram σ ir

$$\hat{\sigma} = \arg\max_{\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0\left(\frac{x}{\sigma}\right). \tag{4.4}$$

Logaritmējot un atvasinot pēc σ, iegūst, ka

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i}{\hat{\sigma}}\right) = 1,$$

kur $\rho(t) = t\psi(t), \psi = -f_0'/f_0$.

Vispārējā gadījumā, ikvienu novērtējumu, kas apmierina vienādību

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{x_i}{\hat{\sigma}}\right) = \delta, \tag{4.5}$$

kur δ ir pozitīva konstante, sauc par mēroga M-novērtējumu.

Ja $0 < \delta < \rho(\infty)$, tad (4.5) ir atrisinājums.

Visbiežāk lietotais mēroga M-novērtējums ir mediānas absolūtā novirze (angl. *median absolute deviation*)

$$MAD = median(|x_i - median(x_i)|).$$

5. Robusta regresija

Bieži vien pieņēmums, ka regresijas atlikumi ir normāli sadalīti, neizpildās, jo datos ir sastopami izlecēji. Viens vienīgs izlecējs spēj ļoti ietekmēt klasiskos novērtējumus. Šādā gadījumā statistisko datu apstrādei tiek izmantotas robustas statistikas. Šajā nodaļā tiek definēts robusts regresijas M-novērtējums [10].

Aplūkojam modeli

$$y_i = x_i\beta + \varepsilon_i,$$

kur $x_i = (1,x_{i1},...,x_{ik})^T, \beta = (\beta_0, \beta_1,...,\beta_k)$, un ε_i blīvuma funkcija ir

$$\frac{1}{\sigma} f_0 \left(\frac{\varepsilon}{\sigma} \right),$$

σ ir mēroga parametrs. Šim modelim y_i ir neatkarīgi, bet ne vienādi sadalīti, y_i blīvuma funkcija ir

$$\frac{1}{\sigma} f_0 \left(\frac{y_i - x_i\beta}{\sigma} \right)$$

un ticamības funkcija parametram β, kad σ ir fiksēts, ir

$$L(\beta) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{y_i - x_i\beta}{\sigma} \right).$$

Maksimālās ticamības funkcijas novērtējumu $\hat{\beta}$ iegūst, maksimizējot $L(\beta)$, kas ir ekvivalenti, minimizējot funkciju

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{y_i - x_i\beta}{\sigma} \right) + \ln \sigma, \tag{5.1}$$

kur $\rho_0 = -\ln f_0$.

9. Definīcija. Par regresijas M-novērtējumu sauc

$$\hat{\beta}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho \left(\frac{y_i - x_i\beta}{\hat{\sigma}} \right), \tag{5.2}$$

kur $\hat{\sigma}$ ir kļūdas mēroga novērtējums.

Diferencējot (5.2), iegūstam

$$\sum_{i=1}^{n} \psi \left(\frac{y_i - x_i\beta}{\hat{\sigma}} \right) x_i = 0, \tag{5.3}$$

kur $\psi = \rho'$. Vienādojuma (5.3) atrisinājumu sauc par monotonu regresijas M-novērtējumu.

Monotonas regresijas M-novērtējuma priekšrocība ir tā, ka ikviens vienādojuma (5.3) atrisinājums ir (5.2) atrisinājums. Pie tam, ja funkcija ψ ir augoša, tad atrisinājums ir viens vienīgs.

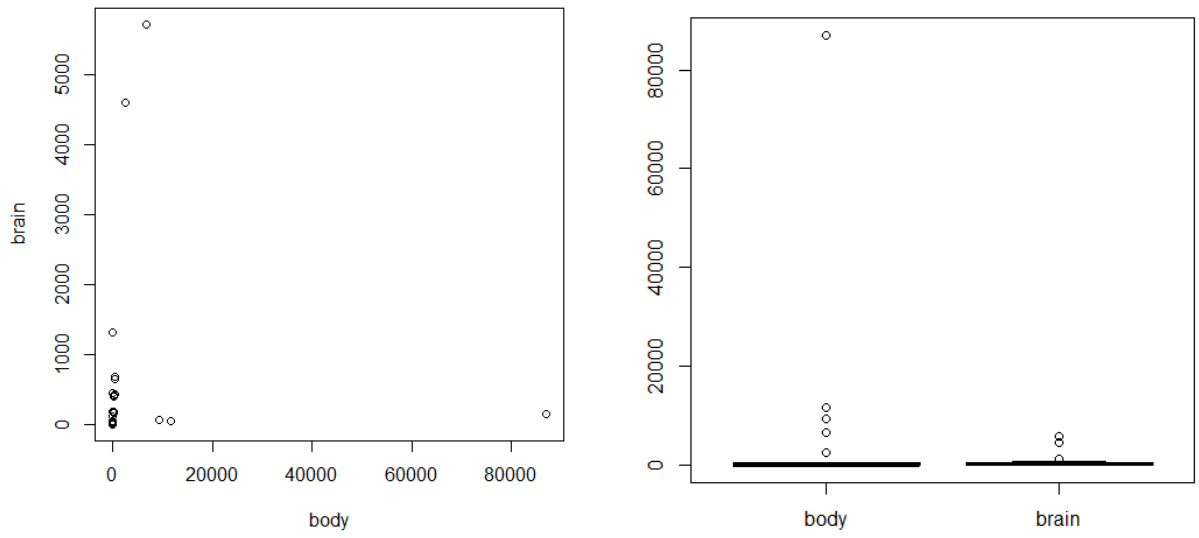
6. Rezultāti

Darba galvenais mērķis ir konstruēt un salīdzināt parametrisko ticamības apgabalu ar empīrisko ticamības apgabalu vienfaktora lineārās regresijas koeficientiem β_0 un β_1. Lai to veiktu, bija nepieciešams atrisināt nelineāru sistēmu (3.12), kas tika darīts, izmantojot programmā R iebūvētu paketi *nleqslv*. Empīrisko ticamības apgabalu atrod kā logaritmiskās empīriskās ticamības attiecības funkcijas grafika un χ^2_2 sadalījuma $1-\alpha$ kvantiles šķēluma projekciju uz β_0 un β_1 plakni. Empīriskā ticamības apgabala līnija konstruēta ar Z. Hu izstrādāta R koda [5] palīdzību.

6.1. Ticamības apgabalu konstruēšana

Šajā apakšnodaļā tiks aplūkoti vairāki reālu un simulētu datu piemēri, kuriem tiks konstruēti ticamības apgabali.

Dzīvnieku smadzeņu un ķermeņa masas dati. [8] Tiek aplūkota smadzeņu masa (g) un ķermeņa masa (kg) 28 dzīvniekiem ar nolūku noskaidrot, vai lielāka ķermeņa masa nosaka, ka ir lielākas smadzenes.



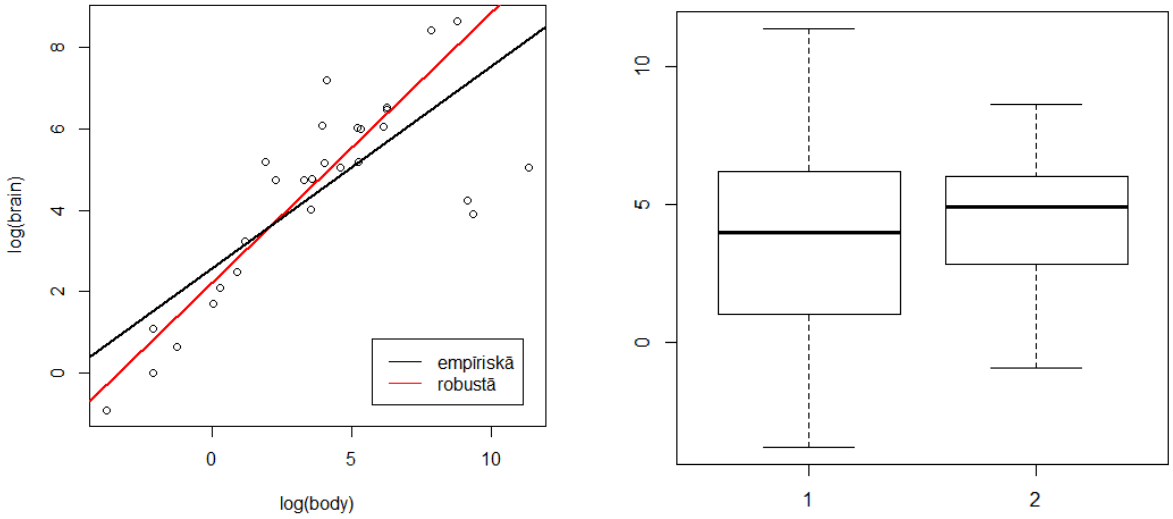
(a) (b)

6.1. att. (a) Smadzeņu masa atkarībā no ķermeņa masas 28 dzīvniekiem (b) Kastu grafiki ķermeņu masai un smadzeņu masai

Kā redzams 6.1. attēlā, datos ir vairāki izlecēji. Lai labāk reprezentētu mērījumus, tiek aplūkoti logaritmizēti dati.

6.2. attēlā var redzēt, ka pastāv lineāra atkarība starp logaritmizētu smadzeņu masu un logaritmizētu ķermeņa masu, kā arī var redzēt, ka robustās regresijas taisne labāk aproksimē

datus. 6.3. (b) attēlā redzams, ka empīriskās ticamības apgabalam nav elipses forma.

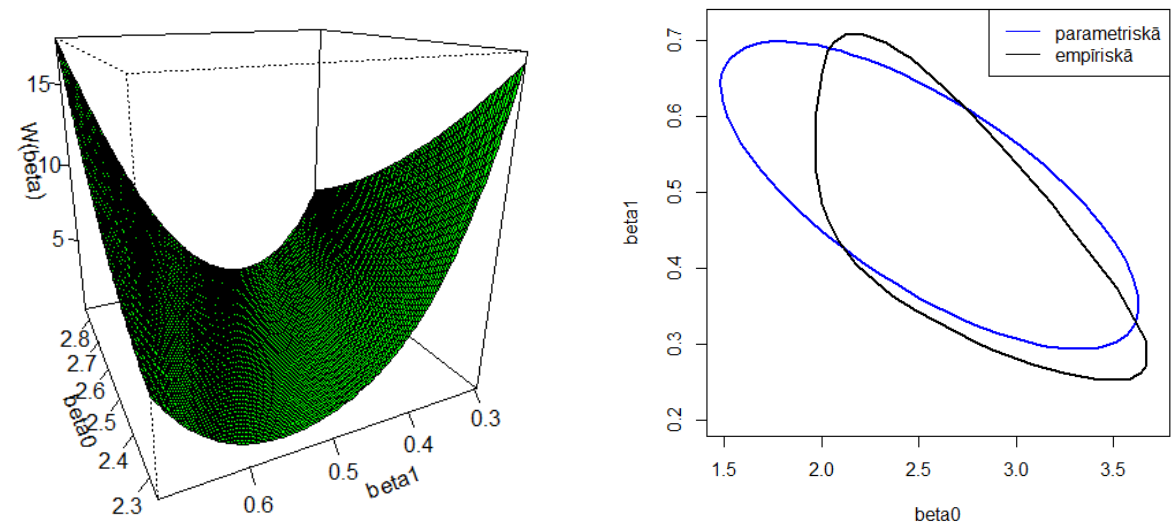


(a) (b)

6.2. att. (a)Logaritmizētā smadzeņu masa atkarībā no logaritmizētās ķermeņa masas 28 dzīvniekiem (b) Kastu grafiki logaritmizētai ķermeņu masai un logaritmizētai smadzeņu masai

6.1. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi(LS), empīriskās ticamības metodi(EL) un M-novērtējums (M)

	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	2.5548981	0.4959947
EL	2.5548900	0.4959990
M	2.2112186	0.6636215

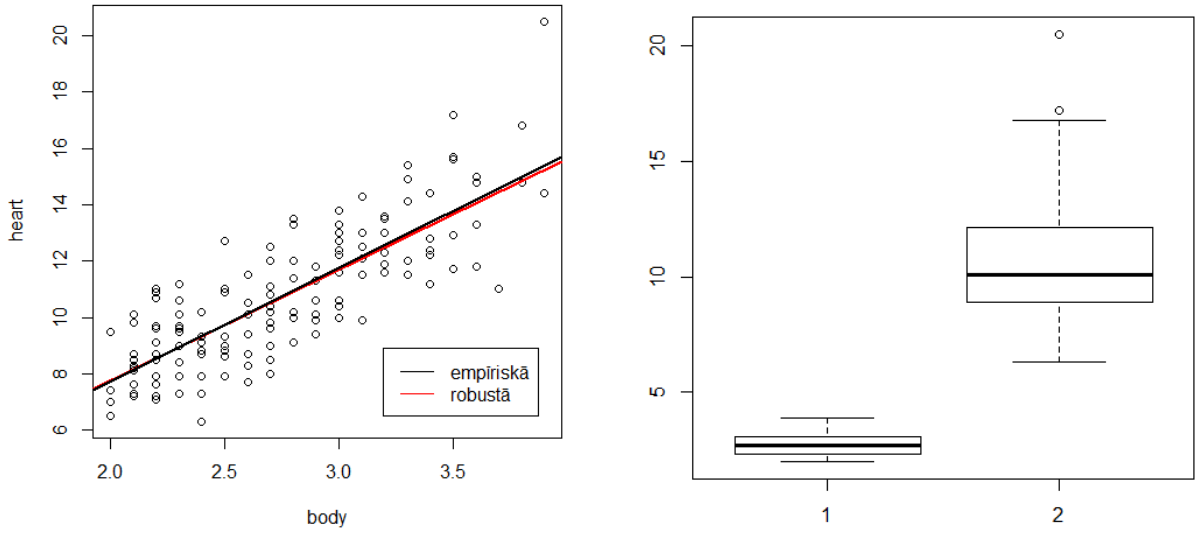


(a) (b)

6.3. att. (a)Logaritmiskās empīriskās ticamības attiecības funkcijas grafiks (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

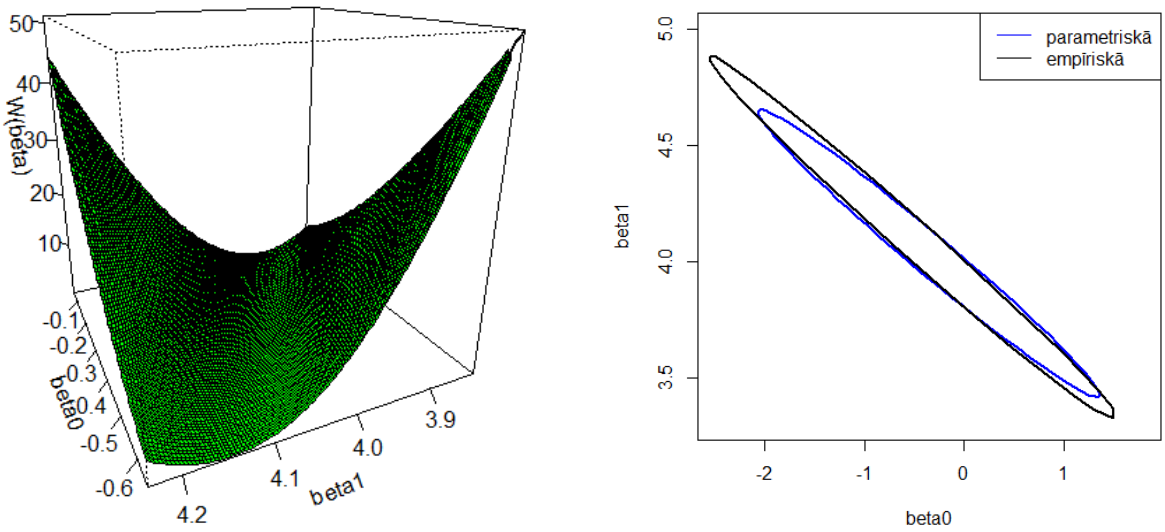
Kaķu sirds un ķermeņa masas dati. [9] Datu kopa satur informāciju par 144 pieaugušu kaķu(sver virs 2 kg) sirds un ķermeņa masu.

Lai gan aplūkotajos datos ir izlecēji (ko parāda arī 6.4. (b) attēlā redzamais kastu grafiks), robustās regresijas koeficientu M-novērtējumi minimāli atšķiras no mazāko kvadrātu metodes un empīriskās ticamības metodes novērtējumiem. Salīdzinājumā ar parametrisko ticamības apgabalu, empīriskais ticamības apgabals ir garāks.



(a) (b)

6.4. att. (a) Sirds masa atkarībā no ķermeņa masas 144 kaķiem (b) Kastu grafiki ķermeņu masai un sirds masai



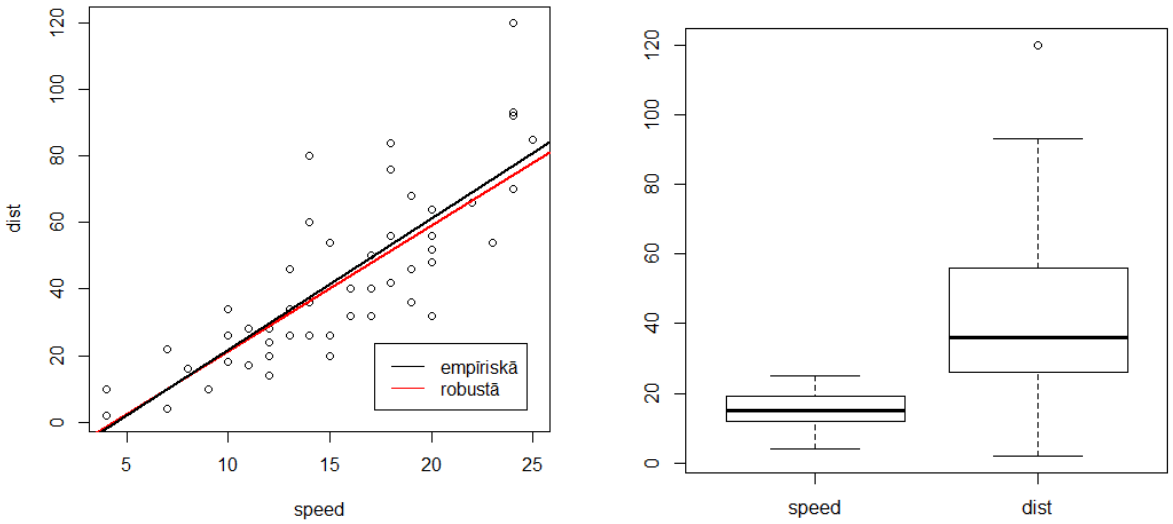
(a) (b)

6.5. att. (a) Logaritmiskās empīriskās ticamības attiecības funkcijas grafiks (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

6.2. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi(LS), empīriskās ticamības metodi(EL) un M-novērtējums (M)

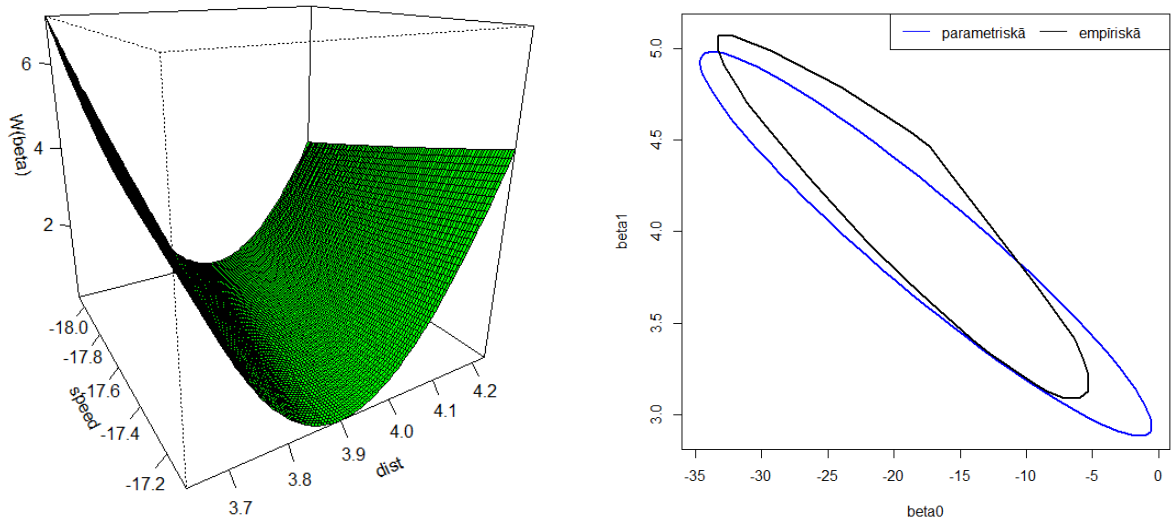
	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	-0.3566624	4.0340627
EL	-0.3567772	4.0341118
M	-0.1361777	3.9380535

Cars dati. Tiek aplūkoti programmā R iebūvēti *cars* dati, kas apraksta saistību starp 50 mašīnu ātrumu un bremzēšanas distanci.



(a) (b)

6.6. att. (a) Mašīnas bremzēšanas distance atkarībā no mašīnas ātruma (b) Kastu grafiki mašīnas ātrumam un bremzēšanas distancei



(a) (b)

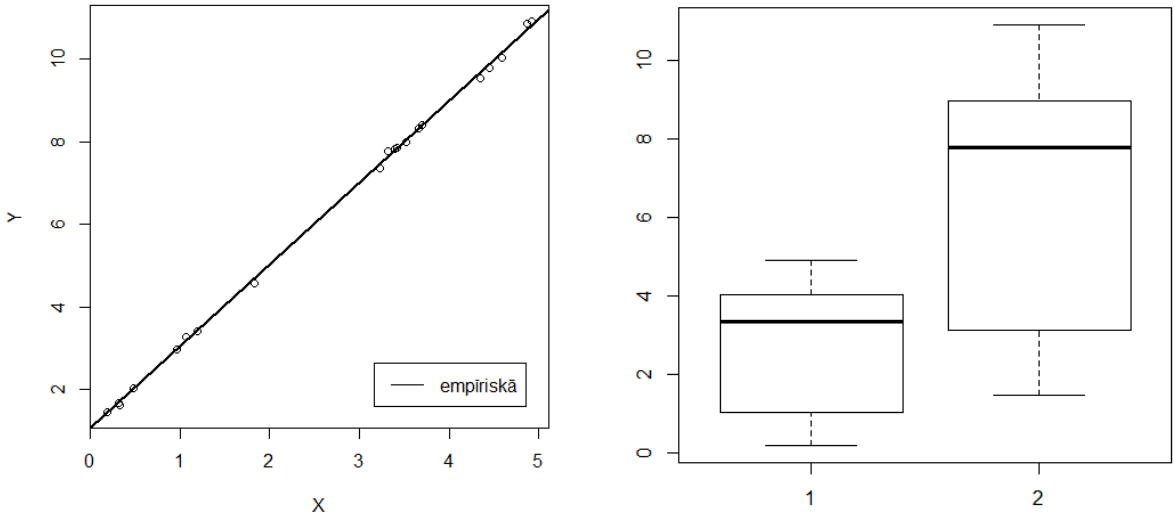
6.7. att. (a) Logaritmiskā empīriskā ticamības attiecības funkcija (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

6.3. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi(LS), empīriskās ticamības metodi(EL) un M-novērtējums(M)

	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	-17.579095	3.932409
EL	-17.579028	3.932407
M	-16.527535	3.773429

Lai gan datos ir 1 izlecējs, tas būtiski neietekmē vektora β novērtējumu. Kā redzams 6.3. tabulā, regresijas koeficientu novērtējumi ar 3 metodēm minimāli atšķiras. 6.7. (b) attēlā var redzēt, ka empīriskais ticamības apgabals nav elipse.

Simulēti dati I. Tiek simulēta vienfaktora lineārā regresija $Y_i = 1 + 2X_i + \varepsilon_i,\ i = 1,\ldots,20$, $X_i \sim U(0,5), \varepsilon_i \sim N(0,0.1)$.

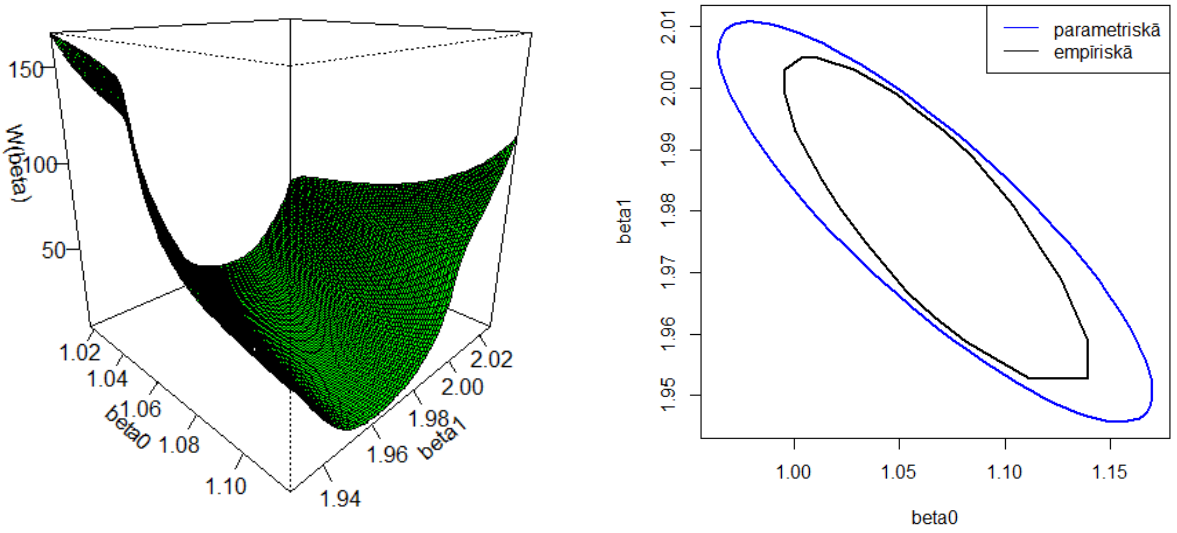


(a) (b)

6.8. att. (a) Izkliedes grafiks kopā ar EL metodes regresijas taisni (b) Kastu grafiki mainīgajiem X un Y

6.4. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi (LS), empīriskās ticamības metodi (EL) un M-novērtējums (M)

	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	1.067182	1.978292
EL	1.067146	1.978314
M	1.067233	1.978256

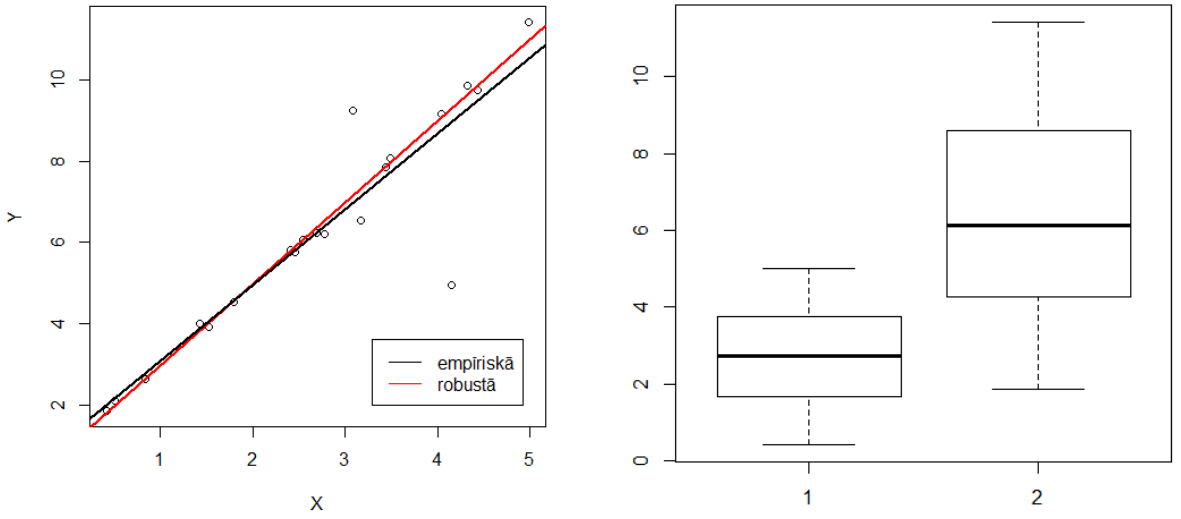


(a) (b)

6.9. att. (a) Logaritmiskās empīriskās ticamības attiecības funkcijas grafiks (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

Tiek simulēti dati, kuriem ir spēkā pieņēmums par atlikumu normalitāti. Kā redzams 6.4. tabulā, parametru novērtējumi sakrīt visām 3 metodēm ar precizitāti līdz 10^{-3}. Iegūtais empīriskais ticamības apgabals atrodas parametriskā ticamības apgabala iekšpusē, kas liecina par to, ka EL metode strādā labāk.

Simulēti dati II. Tiek simulēta vienfaktora lineārā regresija $Y_i = 1 + 2X_i + \varepsilon_i,\ i = 1,\ldots,20$, $X_i \sim U(0{,}5), \varepsilon_i \sim Cauchy(0{,}0.1)$.

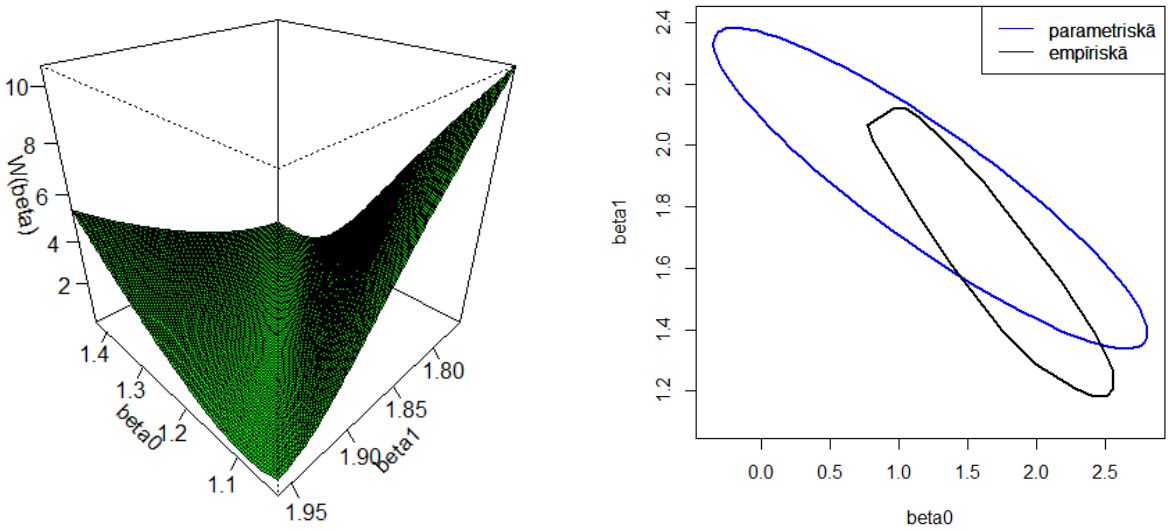


(a) (b)

6.10. att. (a) Izkliedes grafiks kopā ar regresijas taisnēm (b) Kastu grafiki mainīgajiem X un Y

6.5. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi (LS), empīriskās ticamības metodi (EL) un M-novērtējums (M)

	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	1.231676	1.860518
EL	1.231706	1.860505
M	0.9636586	2.0080305



(a) (b)

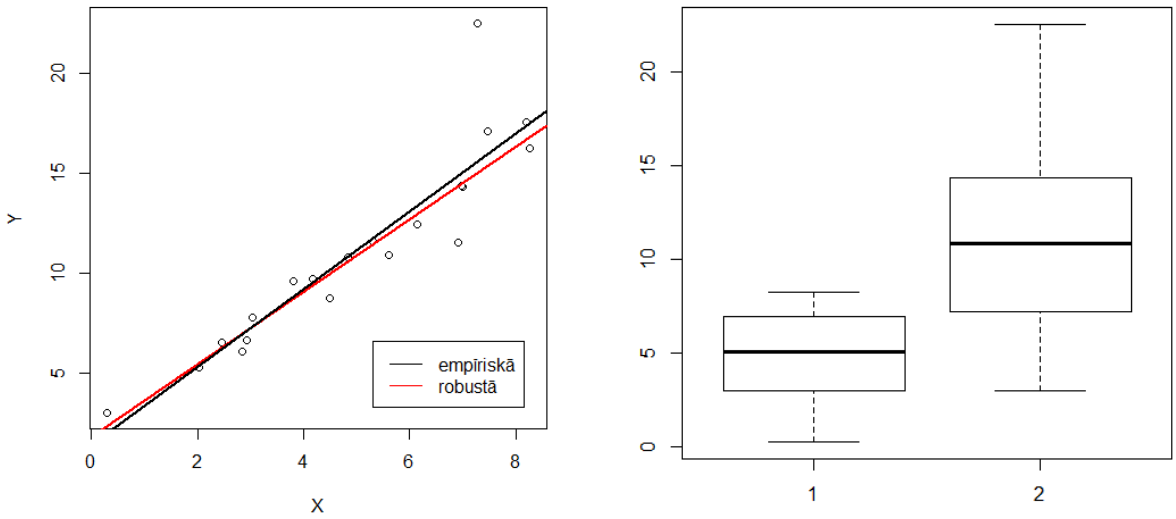
6.11. att. (a) Logaritmiskās empīriskās ticamības attiecības funkcijas grafiks (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

Balstoties uz kastu grafikiem, šajos simulētos datos nav izlecēju, lai gan izkliedes grafikā ir redzams, ka punkts (4.1,4.9) ietekmē robustās regresijas koeficientu novērtējumus. 6.13. attēlā redzams, ka empīriskais ticamības apgabals ir šaurāks par parametrisko apgabalu, un atšķirībā no iepriekš aplūkotiem gadījumiem, šeit empīriskais ticamības apgabals salīdzinoši mazāk pārklāj parametrisko ticamības apgabalu.

Simulēti dati III. Tiek simulēta vienfaktora lineārā regresija $Y_i = 1 + 2X_i + \varepsilon_i,\ i = 1,\ldots,20$, $X_i \sim U(0{,}10)$, $\varepsilon_i \sim t(2)$.

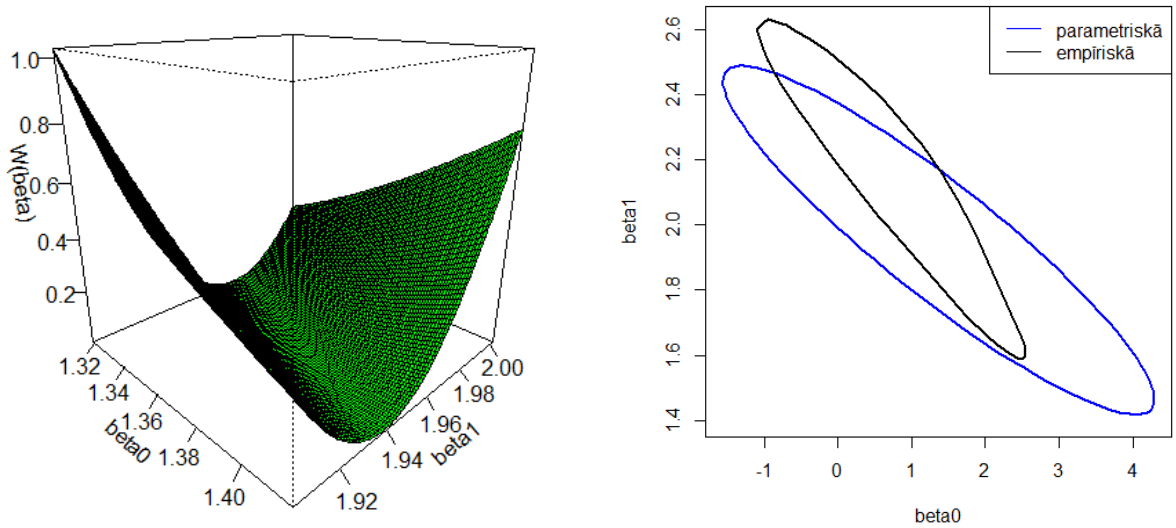
6.6. tabula. Regresijas koeficientu novērtējumi ar mazāko kvadrātu metodi (LS), empīriskās ticamības metodi (EL) un M-novērtējums (M)

	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	1.368554	1.952852
EL	1.368588	1.952842
M	1.778134	1.819144



(a) (b)

6.12. att. (a) Izkliedes grafiks kopā ar regresijas taisnēm (b) Kastu grafiki mainīgajiem X un Y



(a) (b)

6.13. att. (a) Logaritmiskās empīriskās ticamības attiecības funkcijas grafiks (b) Parametriskais un empīriskais 95% ticamības apgabals regresijas koeficientiem

6.2. Jaudas analīze

Analīzes mērķis ir salīdzināt jaudas analīzes rezultātus empīriskās ticamības (EL) metodēm (Chen un Keilegom piedāvātā, Owen piedāvātā) un t-testam (lineārai un robustai regresijai). Tiek simulēta vienfaktora lineāra regresija

$$Y_i = 2X_i + \varepsilon_i, i = 1, \ldots, n$$

dažādi sadalītiem atlikumiem. Tiek veikta hipotēžu pārbaude $H_0 : \beta_1 = 2$ pret $H_1 : \beta_1 \neq 2$.

6.7. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,1)$, $\varepsilon_i \sim N(0, 0.1)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.0874	0.0584	0.0529	0.0517
Owen	0.1054	0.0631	0.0549	0.0520
Lineāra	0.0518	0.0472	0.0491	0.0519
Robusta	0.0546	0.0502	0.0511	0.0476

6.8. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,1)$, $\varepsilon_i \sim Logistic(0, 0.1)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.0893	0.0659	0.0570	0.0490
Owen	0.1132	0.0717	0.0574	0.0490
Lineāra	0.0471	0.0501	0.0473	0.0476
Robusta	0.0513	0.0535	0.0520	0.0496

6.9. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,10)$, $\varepsilon_i \sim t(3)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.1118	0.0773	0.0669	0.0538
Owen	0.1329	0.0868	0.0723	0.0540
Lineāra	0.0476	0.0455	0.0479	0.0457
Robusta	0.0553	0.0509	0.0498	0.0516

6.7., 6.8. un 6.9. tabulā apkopoti jaudas analīzes rezultāti regresijas modeļiem, kuros atlikumi ir ar simetrisku sadalījumu. Redzams, ka pie nelieliem izlašu apjomiem EL metodes strādā nedaudz sliktāk par pārējām abām.

6.10., 6.11., 6.12. un 6.13. tabulās tiek aplūkoti piesārņoti dati ar piesārņojuma līmeni $\epsilon = 0.05$. 6.11., 6.12. un 6.13. tabulās piesārņojums sadalīts ar vidējo vērtību, kas nav 0, līdz ar to tiek aplūkota varbūtība noraidīt nulles hipotēzi H_0, kad H_0 nav spēkā.

6.10. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,10)$, $\epsilon = 0.05$, $\varepsilon_i \sim (1 - \epsilon)N(0, 0.3) + \epsilon N(0,3)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.1531	0.1200	0.1154	0.0653
Owen	0.1395	0.1143	0.1231	0.0732
Lineāra	0.0314	0.0412	0.0470	0.0473
Robusta	0.0536	0.0474	0.0521	0.0492

6.11. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,10)$, $\epsilon = 0.05$, $\varepsilon_i \sim (1-\epsilon)N(0,0.3) + \epsilon N(5,3)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.2289	0.3724	0.8062	1.0000
Owen	0.1522	0.2755	0.7125	0.9996
Lineāra	0.0168	0.0559	0.2908	0.9983
Robusta	0.0531	0.0538	0.0831	0.2727

6.12. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,10)$, $\epsilon = 0.05$, $\varepsilon_i \sim (1-\epsilon)N(0,0.1) + \epsilon\, Noncentral\ t(3,0.5)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.1854	0.1991	0.2741	0.5350
Owen	0.1461	0.1551	0.2587	0.5403
Lineāra	0.0309	0.0386	0.0704	0.3769
Robusta	0.0567	0.0538	0.0539	0.0869

6.13. tabula. Jaudas analīzes rezultāti, $N = 10000$, $\alpha = 0.05$, $X_i \sim U(0,10)$, $\epsilon = 0.05$, $\varepsilon_i \sim (1-\epsilon)N(0,0.1) + \epsilon\, Gamma(0.5,2)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.1626	0.2128	0.4807	0.9907
Owen	0.1308	0.1537	0.3839	0.9715
Lineāra	0.0355	0.0494	0.1258	0.8231
Robusta	0.0542	0.0587	0.076	0.2343

6.14. tabula. Jaudas analīzes rezultāti, $N = 10000, \alpha = 0.05, X_i \sim U(0,1), \varepsilon_i \sim Cauchy(0,0.1)$

	$n = 20$	$n = 50$	$n = 100$	$n = 500$
Chen, Keilegom	0.2353	0.2267	0.2229	0.2141
Owen	0.2184	0.1984	0.2021	0.2127
Lineāra	0.0323	0.0330	0.0320	0.0305
Robusta	0.0486	0.0478	0.0477	0.0498

6.14. tabulā parādīti jaudas analīzes rezultāti, kad aplūkotajos datos izlecēju ietekme mēdz būt liela. Var redzēt, ka gan t-tests, gan EL metode ir jūtīgi pret izlecējiem.

Kopumā var secināt, ka empīriskās ticamības metode uzrāda labu sniegumu. Chen un Keilegom publicētā metode strādā labāk kā Owen norādītā EL metode.

Secinājumi

Bakalaura darbā tika aplūkots empīriskās ticamības funkcijas pielietojums lineārai regresijai - empīriskās ticamības apgabala konstruēšanai vienfaktora lineārās regresijas koeficientiem. Papildus tika pētīts, cik labi strādā empīriskās ticamības metode salīdzinājumā ar t-testu lineārai un robustai regresijai, veicot jaudas analīzi.

Simulāciju piemēros tika iegūts, ka gadījumā, kad tiek aplūkota regresija caur sākumpunktu, Chen un Keilegom publicētā aprēķinu gaita EL metodei strādā labāk kā Owen publicētā, citiem vārdiem, definējot profila empīrisko ticamības attiecību, nav nepieciešams ierobežojums (3.15).

Balstoties uz jaudas analīzes rezultātiem, empīriskās ticamības metode strādā labi, vislabākos rezultātus uzrādot pie lieliem izlašu apjomiem. Tomēr EL metode ir jūtīga pret izlecēju ietekmi.

Aplūkojot reālu datu piemērus, tika iegūts, ka parametra β novērtējumi būtiski neatšķiras EL metodes novērtējumam no mazāko kvadrātu metodes novērtējuma. Bieži vien gadījumos, kad datos ir izlecēji, empīriskiem ticamības apgabaliem nav elipses forma. Ja izpildās pieņēmums par atlikumu normalitāti, tad empīriskās ticamības apgabals atrodas parametriskā ticamības apgabala iekšpusē, taču pārsvarā gadījumu EL apgabalam nav viennozīmīgs novietojums attiecībā pret parametrisko ticamības apgabalu. Turpmākiem pētījumiem būtu interesanti salīdzināt empīriskos ticamības apgabalus ar robustiem un citu neparametriskās statistikas metožu iegūtiem ticamības apgabaliem.

Literatūras saraksts

[1] A.B. Owen. Empirical likelihood. CRC press, 2001.

[2] A. B. Owen. Empirical likelihood confidence regions. The Annals of Statistics, 18:90-120, 1990.

[3] A. B. Owen. Empirical likelihood for linear models. The Annals of Statistics, (19):1725-1747, 1991.

[4] D.A.Dickey,S.G.Pantula,J.O.Rawlings. Applied Regression Analysis: A Research Tool. Springer,New York, 1998

[5] http://statgen.ualberta.ca/download/software/convex.hull.R, skatīts 17.05.2015.

[6] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. The Annals of Statistics, 22(1):300-325, 1994.

[7] L.Wasserman. All of Nonparametric Statistics. Springer, New York, 2006.

[8] P. J. Rousseeuw and A. M. Leroy. Robust regression and outlier detection. Wiley, New York, 1987.

[9] R. A. Fisher. The Analysis of Covariance Method for the Relation between a Part and the Whole. Biometrics, 3:65-68, 1947.

[10] Ricardo A.Maronna, R.Douglas Martin, Victor J. Yohai. Robust Statistics: Theory and Methods. John Wiley and Sons, 2006.

[11] S.X. Chen and I. Van Keilegom. A review on empirical likelihood methods for regression. Test 18(3):415-447, 2009.

Pielikums

Programmas R kods pārbaudei par logaritmiskās EL attiecības statistikas koverģenci pēc sadalījuma uz χ^2_2

```
library(nleqslv)
n<-100
alpha<-0.05
b0<-1
b1<-2
koef<-c(1,2)
W<-c()
N<-10000
for (i in 1:N){
X<-runif(n,0,1)
eps<-rnorm(n,0,0.1)
Y<-b0+ b1*X+eps

Z<-function(beta){
dslnex <- function(lambda) {
z <- numeric(2)
a<- Y-beta[1]-beta[2]*X
z[1] <- sum(a/(1+lambda[1]*a+lambda[2]*X*a))
z[2] <- sum(X*a/(1+lambda[1]*a+lambda[2]*X*a))
z}
xstart <- c(0,0) #starta lambdas
fstart <- dslnex(xstart)
nleqslv(xstart, dslnex, control=list(btol=.01))$x
}
rez<-Z(koef)
W[i]<- 2*sum(log(1+rez[1]*(Y-koef[1]-koef[2]*X)+rez[2]*X*(Y-koef[1]-koef[2]*X)))
}

hist(W, prob=T, main="", xlab="", ylab="-2ln R(beta)")
```

```
x<-seq(0,60,by=0.01)
lines(x, dchisq(x,2),col="red", lwd=2)
```

Programmas R kods parametra β novērtēšanai ar EL metodi un logaritmiskās EL attiecības statistikas zīmēšanai

```
library(nleqslv)
n<-100
alpha<-0.05
b0<-1
b1<-2
X<-runif(n,0,1)
eps<-rnorm(n,0,0.1)
Y<-b0+b1*X+eps
plot(X,Y)
koef<-c(1,2)

f<- function (beta){
dslnex <- function(lambda) { #meklē lambdas
z <- numeric(2)
a<- Y-beta[1]-beta[2]*X
z[1] <- sum(a/(1+lambda[1]*a+lambda[2]*X*a))
z[2] <- sum(X*a/(1+lambda[1]*a+lambda[2]*X*a))
z}
xstart <- c(0,0) #starta lambdas
rez1<-nleqslv(xstart, dslnex, control=list(btol=.01))$x
2*sum(log(1+rez1[1]*(Y-beta[1]-beta[2]*X)+rez1[2]*X*(Y-beta[1]-beta[2]*X)))
}

optim(koef,f)$par
koef1<-c()
koef1[1]<- optim(koef,f)$par[1]
koef1[2]<- optim(koef,f)$par[2]
koef1 #EL metodes novērtējums  parametram beta
```

```
################ statistikas zīmēšana ######################

f2<- function (b1,b2){
r<-cbind(b1,b2)
dslnex <- function(lambda) {
z <- numeric(2)
a<- Y-r[,1]-r[,2]*X
z[1] <- sum(a/(1+lambda[1]*a+lambda[2]*X*a))
z[2] <- sum(X*a/(1+lambda[1]*a+lambda[2]*X*a))
z}
xstart <- c(0,0) #starta lambdas
rez1<-nleqslv(xstart, dslnex, control=list(btol=.01))$x #lambdas
2*sum(log(1+rez1[1]*(Y-r[,1]-r[,2]*X)+rez1[2]*X*(Y-r[,1]-r[,2]*X)))
}
VecFun <- Vectorize(f2)

x<- seq(koef1[1]-0.5,koef1[1]+0.5, length.out=100)
y<- seq(koef1[2]-0.3,koef1[2]+0.3, length.out=100)
z<-outer(x,y,VecFun)
persp(x,y,z,phi=20,theta=60,col="green",ticktype="detailed")
```

Programmas R kods EL ticamības apgabala līnijas zīmēšanai

```
#obtain angle between (0,0) and (x, y)
xytoangle<-function(x, y) {
  d=sqrt(x^2+y^2)
  ang.cos=acos(x/d)%%(2*pi)
  idx=which(y<0,arr.ind=T)
  ang.cos[idx]=2*pi-ang.cos[idx]
  ang.cos
}

#get index of outer points in x=data.frame(row, col)
border.points.idx<-function(x) {
  x=data.frame(row=x[,1], col=x[,2])
```

```
  k=order(x$row)[1]
  bord.points=k;
  ang0=pi;

  while(NROW(bord.points)==NROW(unique(bord.points))) {
    ang=xytoangle(x=x$row-x$row[k], y=x$col-x$col[k])
    len=(x$row-x$row[k])^2+(x$col-x$col[k])^2
    ang_diff=ang-ang0
    idx=which(ang_diff>pi,arr.ind=T)
    ang_diff=ang_diff%%(2*pi)
    minang=min(ang_diff,  na.rm = T)
    k=which((minang==(ang_diff))&(len>0), arr.ind=T)
    k=k[which(len[k]==max(len[k]))[1]]

    bord.points=c(bord.points, k)
    ang0=ang[k][1]%%(2*pi)
  }
  bord.points
}

#convex.hull peeling function
convex.hull<-function(x, alpha=0.05){
  n.alpha=(NROW(x)*alpha)

  n.out=0;
  apex.idx=border.points.idx(x)
  apex=x[apex.idx,];
  while (n.alpha>n.out) {
    apex0=apex
    n.out0=n.out

    x=x[-apex.idx,]
    apex.idx=border.points.idx(x)
```

```
    n.out=n.out+NROW(apex.idx)
    apex=x[apex.idx,];
  }

  if (abs(n.alpha-n.out0)<abs(n.alpha-n.out)) {
    apex=apex0
  }
  apex
}
#########
x<- seq(0.95,1.09, length.out=100) #beta0
y<- seq(1.85,2.08, length.out=100) #beta1
z<-outer(x,y,VecFun)
persp(x,y,z,phi=20,theta=60,col="green",ticktype="detailed")

c<-qchisq(1-alpha,2) #kritiskā kvantile
apg<- z-c
dat1<-which(apg<0,arr.ind=TRUE)
dat<-data.frame(x=x[dat1[,1]],y=y[dat1[,2]])
plot(dat)

#fit<-lm(Y~X)
#library(ellipse)
#plot(ellipse(fit), main="Pairwise confidence region")

polygon(convex.hull(dat, alpha=0.01)) #Līnija apkārt datu punktu kopai
```

Programmas R kods jaudas analīzei

```
library(emplik)
library(nleqslv)
library(MASS)
n<-100
alpha<-0.05
b0<-0
```

```
b<-2
b1<-2  #hipotēžu pārbaude H0: beta1=b1 pie pieņēmuma b0 = 0
delta<-rep(1,n)
N<-10000
W<-c()
t1<-c()
pv1<-c()
W2<- c()
WOwen<-c()
rob<-c()
pvrob<-c()
for (i in 1:N)
{
X1<-runif(n,0,10)
X<-X1
eps<-rnorm(n,0,0.1)
#eps <-rlogis(n,0,0.1)
#eps<-rt(n,3)
#eps <-rcauchy(n,0,0.1)
#eps<-c(rnorm(round(0.95*n),0,0.1),(rgamma(n-round(0.95*n),scale=2,shape=0.5)))
#eps<-c(rnorm(round(0.95*n),0,0.1),(rt(n-round(0.95*n),3,0.5)))
#eps<-c(rnorm(round(0.95*n),0,0.3),(rnorm(n-round(0.95*n),0,3)))
#eps<-c(rnorm(round(0.95*n),0,0.3),(rnorm(n-round(0.95*n),5,3)))
Y<-b*X1+eps

#####################
Z_B<-function(beta1){
dslnex <- function(lambda) {
z<- numeric(1)
a<- Y-beta1*X
z<- sum(X*a/(1+lambda*X*a))
z}
xstart <- 0 #starta lambda
fstart <- dslnex(xstart)
```

```
nleqslv(xstart, dslnex, control=list(btol=.01))$x
}

rez<-Z_B(b1)
W[i]<- 2*sum(log(1+rez*X*(Y-b1*X))) #Chen
######################

Z_B1<-function(beta1){
dslnex <- function(lambda) {
z<- numeric(2)
a<- Y-beta1*X
z[1] <- sum(a/(1+lambda[1]*a+lambda[2]*X*a))
z[2] <- sum(X*a/(1+lambda[1]*a+lambda[2]*X*a))
z}
xstart <- c(0,0) #starta lambda
fstart <- dslnex(xstart)
nleqslv(xstart, dslnex, control=list(btol=.01))$x
}

rez<-Z_B1(b1)
W2[i]<- 2*sum(log(1+rez[1]*(Y-b1*X)+rez[2]*X*(Y-b1*X)))

#######  LS, Fiksēts beta0=0 ###############
BetaRTO<-sum(X*Y)/sum(X^2)
sigma_kv<-sum((Y-BetaRTO*X)^2)/(n-1)
SE<-sqrt(sigma_kv/sum(X^2))
t1[i]<-(BetaRTO-b1)/SE
pv1[i]<- 2*(1-pt( abs(t1[i]) ,n-1))
##############
Betaprim<-BJnoint(X, Y, delta, beta0 = NA, maxiter=30, error = 0.00001)$beta
rez<-Z_B1(Betaprim)
WOwen[i]<- W2[i]- 2*sum(log(1+rez[1]*(Y-Betaprim*X)+rez[2]*X*(Y-Betaprim*X))) #Owen
###############
fm.rlm <- rlm(Y ~0+ X)
```

```
rob[i]<-(summary(fm.rlm)$coef[1]-b1)/summary(fm.rlm)$coef[1,"Std. Error"]
pvrob[i]<- 2*(1-pt( abs(rob[i]) ,n-1))
}

################ Jaudas analīzes rezultāti ####################
pvalue1<- 1-pchisq(W,1)
precizitate<-length(pvalue1[pvalue1<alpha]) # Chen
precizitate/N
pvalue5<- 1-pchisq(WOwen,1)
precizitate<-length(pvalue5[pvalue5<alpha]) # Owen prim
precizitate/N
ttest_precizitate<-length(pv1[pv1<alpha]) #t-tests fiksētam beta0
ttest_precizitate/N
rob_precizitate<-length(pvrob[pvrob<alpha]) #robust
rob_precizitate/N
```

Bakalaura darbs ”Empīriskā ticamības funkcija lineārai regresijai” izstrādāts LU Fizikas un matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Darba autore: _____ Līga Mangule
(Paraksts)

Rekomendēju darbu aizstāvēšanai

Darba vadītājs: _____ asoc.prof. Jānis Valeinis
(Paraksts)

Recenzents: pasniedzēja Leonora Pahirko

Darbs iesniegts Matemātikas nodaļā _____.06.2015.
Dekāna pilnvarotā persona: vecākā metodiķe Dzintra Holsta

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____
(datums)

Komisijas sekretārs:_____ _____
(vārds) (Paraksts)