LMD mokslo darbai, III tomas, 392–398

Apie tiesinės regresijos identifikavimo metodus

Romualdas SALĖTIS* (VGTU, MII)
el. paštas: romas@goda.vtu.lt

1. Įvadas

Mažiausių kvadratų metodas yra populiariausias taikomasis statistinis metodas identifikuojant tiesinės regresijos modelius. Galiojant tam tikroms sąlygoms, mažiausių kvadratų metodu gautas įvertis yra efektyvus (Gauso-Markovo teorema). Tačiau, praktikoje jos nėra pilnai patenkinamos, todėl yra tikslinga naudoti robastines įvertinimo procedūras, kurios yra nejautrios nedideliems nuokrypiams nuo prielaidų apie modelį, o jų pagalba gautas įvertis yra beveik efektyvus.

Šiame darbe yra tiriami mažiausių kvadratų ir medianinio įvertinimo metodai bei įvairios jų modifikacijos, kurios yra nejautrios nedidelės dalies duomenų užteršimui grubiai klaidingais duomenimis – išsiskiriančiais stebėjimais, bei kurie įvertina ir eliminuoja modelio liekanų koreliuotumo ir heteroskedastiškumo (dispersijos priklausomybės nuo laiko) įtaką.

Siekiant palyginti nagrinėjamų metodų įvertinimo tikslumą, esant aukščiau išvardintoms situacijoms, imitacinio modeliavimo būdu buvo atlikta serija eksperimentų, kurių aprašymas ir rezultatai yra pateikiami šio darbo eksperimentinėje dalyje.

2. Tiesinės regresijos modelis ir jo identifikavimo metodai

Tarkime, kad laike stebime atsitiktinius dydžius $Y(t)$, $t = 1, 2, \ldots, n$, kurie yra aprašomi regresijos lygtimi:

$$Y(t) = \boldsymbol{\theta}^T \boldsymbol{x}(t) + \xi(t), \quad t = 1, 2, \ldots, n. \tag{1}$$

čia: $\boldsymbol{\theta}$ – nežinomas daugiamatis parametras, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^T$;
$\boldsymbol{x}(t)$ – determinuotas vektorius, $\boldsymbol{x}(t) = \big(x_1(t), \ldots, x_d(t)\big)^T$, kur $x_i(t)$, $i = 1, 2, \ldots, d$ – žinomo pavidalo funkcijos;
$\xi(t)$ – atsitiktiniai dydžiai, modelyje vadinami liekanomis.
Tiesinės regresijos modelyje (1) dažnai yra priimamos šios prielaidos.
1. Liekanų vidurkis yra lygus nuliui: $\mathbf{E}\xi(t) = 0$.
2. Liekanų $\xi(t)$ pasiskirstymas nepriklauso nuo $\boldsymbol{x}(t)$.
3. Liekanų dispersija yra pastovi: $\mathbf{D}\xi(t) = \sigma^2$.
4. Liekanos yra nekoreliuotų atsitiktinių dydžių seka: $\operatorname{cov}\big(\xi(t), \xi(s)\big) = 0$, kai $t \neq s$.

*Darbą remia Lietuvos valstybinis mokslo ir studijų fondas, Grantas K-014.

5. Liekanos $\xi(t)$ yra pasiskirstę pagal Gauso dėsnį.

Parametro θ mažiausių kvadratų metodo įvertis yra gaunamas minimizuojant liekanų kvadratų sumą:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{t=1}^{n} \left(Y(t) - \theta^T x(t)\right)^2 = \arg\min_{\theta} (\boldsymbol{Y} - \boldsymbol{X}\theta)^T (\boldsymbol{Y} - \boldsymbol{X}\theta) \tag{2}$$

čia: $\boldsymbol{Y}$ – Y stebėjimų vektorius, $\boldsymbol{Y} = \left(Y(1), Y(2), \ldots, Y(n)\right)^T$;

$\boldsymbol{X}$ – $x_j(t)$ funkcijų reikšmių matrica, $j = 1, 2, \ldots, d$,

$$\boldsymbol{X} = \begin{pmatrix} x_1(1) & x_2(1) & \ldots & x_d(1) \\ x_1(2) & x_2(1) & \ldots & x_d(2) \\ \ldots & \ldots & \ldots & \ldots \\ x_1(n) & x_2(n) & \ldots & x_d(n) \end{pmatrix}.$$

Jei matrica $\boldsymbol{X}^T\boldsymbol{X}$ yra neišsigimusi ($\det \boldsymbol{X}^T\boldsymbol{X} \neq 0$), tai parametro θ įvertis gaunamas iš formulės:

$$\widehat{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

Praktikoje trečioji ir ketvirtoji modelio prielaidos dažnai yra netenkinamos ir tuomet mažiausių kvadratų metodu gautas įvertis nėra efektyvus. Šiuo atveju efektyvus įvertis yra gaunamas apibendrintu mažiausių kvadratų metodu:

$$\widehat{\theta} = (\boldsymbol{X}^T\boldsymbol{Q}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Q}^{-1}\boldsymbol{Y}.$$

kur $\boldsymbol{Q}$ – liekanų kovariacijos matrica, (prie sąlygos kad $\det \boldsymbol{X}^T\boldsymbol{Q}^{-1}\boldsymbol{X} \neq 0$).

Suprantama, jei liekanų seka būtų stacionari plačiąja prasme, tai kovariacijos matricą $\boldsymbol{Q}$ vienareikšmiškai apibrėžtų šių liekanų kovariacijos funkcija. Tačiau dėl liekanų dispersijos kitimo laike – heteroskedastiškumo $\left(\mathbf{D}\xi(t) = \sigma^2(t)\right)$, dažnai taip nėra. Šioje situacijoje paprastai daroma prielaida apie normuotų liekanų stacionarumą. Tegul seka $\varepsilon(t) = \xi(t)/\sigma(t)$ yra stacionari atsitiktinių dydžių seka su nuliniu vidurkiu ir pastovia dispersija ($\mathbf{E}\varepsilon(t) = 0$, $\mathbf{D}\varepsilon(t) = \sigma_0^2$). Šiuo atveju θ įvertis gaunamas iš formulės:

$$\widehat{\theta} = (\boldsymbol{X}\boldsymbol{S}^{1/2}\boldsymbol{R}^{-1}\boldsymbol{S}^{-1/2}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{S}^{-1/2}\boldsymbol{R}^{-1}\boldsymbol{S}^{-1/2}\boldsymbol{Y}. \tag{3}$$

Čia: $\boldsymbol{S} = \operatorname{diag}\left(\sigma^2(1), \sigma 2(2), \ldots, \sigma^2(n)\right)$; $\boldsymbol{R}$ – stacionarių liekanų $\varepsilon(t)$ kovariacijos matrica. Praktikoje matricos $\boldsymbol{S}$ ir $\boldsymbol{R}$, dažnai yra nežinomos, todėl formulėje (3) keičiamos statistiniais įverčiais.

Praktikoje dažnos yra situacijos, kai imtis yra užteršta išsiskiriančiais stebėjimais. (Išsiskiriančiais stebėjimais šiuo atveju yra vadinami stebėjimai tam tikrais laiko momentais, kai aiškiai

negalioja priimtas modelis.) Todėl modelį (3) reikia apibendrinti, stacionarioms liekanoms $\xi(t)$ suteikiant svorius $0 \leqslant \psi(\xi(t)) \leqslant 1$:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T \widehat{\Psi}^{1/2} \widehat{\boldsymbol{S}}^{-1/2} \widehat{\boldsymbol{R}}^{-1} \widehat{\boldsymbol{S}}^{-1/2} \widehat{\Psi}^{1/2} \boldsymbol{X})^{-1} \boldsymbol{X} \widehat{\Psi}^{1/2} \widehat{\boldsymbol{S}}^{-1/2} \widehat{\boldsymbol{R}}^{-1} \widehat{\boldsymbol{S}}^{-1/2} \widehat{\Psi}^{1/2} \boldsymbol{Y}, \tag{4}$$

kur $\Psi = \operatorname{diag}(\psi(\varepsilon(1)), \psi(\varepsilon(2)), \ldots, \psi(\varepsilon(n)))$. $\widehat{\Psi}$ gauname, nestebimą seką $\xi(t)$ keičiant jos įverčiais: $\widehat{\varepsilon}(t) = Y(t) - \widetilde{\boldsymbol{\theta}}^T \boldsymbol{x}(t)$, kur $\widetilde{\boldsymbol{\theta}}$ – pradinis θ įvertis.

Svoriai $\psi(\varepsilon(t))$ gali būti suteikti remiantis taip vadinama „3σ“ taisykle. Ši taisyklė yra ypač taikoma Gauso sekoms su nuliniu vidurkiu ir pastovia dispersija. Pagal ją iš imties yra pašalinamas stebėjimas, kuris moduliu yra didesnis už tris vidutinius standartinius nuokrypius $|\varepsilon(t)| > 3\sigma_0$.

Įvertį (4) galima dar labiau apibendrinti, aprioriai įvedant papildomus svorius $0 \leqslant w(t) \leqslant 1$, dėl kitų priežasčių, pvz. dėl modelio parametrų lėto kitimo laike, t.y. dėl duomenų senėjimo. Šiuo atveju $\boldsymbol{\theta}$ įvertis gaunamas iš formulės:

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T \Lambda^{1/2} \widehat{\boldsymbol{R}}^{-1} \widehat{\Lambda}^{1/2} X)^{-1} X^T \Lambda^{1/2} \widehat{\boldsymbol{R}}^{-1} \Lambda^{1/2} \boldsymbol{Y}, \tag{5}$$

kur $\Lambda = \boldsymbol{W} \Psi \boldsymbol{S}^{-1}$; $\boldsymbol{W} = \operatorname{diag}(w(1), w(2), \ldots, w(n))$.

Metodas (5) yra mažiausių kvadratų metodo apibendrinimas susijęs su regresijos modelio sąlygų silpninimu, tačiau šį metodą galima apibendrinti ir kitaip – nagrinėjant bendresnę nei kvadratinę nuostolių funkciją. Priminsime, kad klasikinio mažiausių kvadratų metodo atveju įverčiai gaunami minimizuojant liekanų kvadratų sumą (2), o siekiant robastiškumo galima minimizuoti sumą $\sum_{t=1}^{n} \rho\big((Y(t) - \boldsymbol{\theta}^T \boldsymbol{x}(t)\big)$, kur ρ yra funkcija auganti lėčiau nei kvadratinė. Bendru atveju šio metodo modifikacija yra pavidalo:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \rho\big(Y(t) - \boldsymbol{\theta}^T \boldsymbol{x}(t)\big)/\sigma(t)\big)\psi(\varepsilon(t))w(t).$$

Tokio robastinio metodo pavyzdžiu būtų modifikuotas medianinis įvertinimo metodas:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \Big|Y(t) - \boldsymbol{\theta}^T \boldsymbol{x}(t)\Big|\psi(\varepsilon(t))w(t)/\sigma(t).$$

Atveju $\psi(\cdot) = w(\cdot) = s(\cdot) = 1$ gautume įprastinį medianinį parametro θ įvertį. Vienamačiu atveju ($d = 1$) šis įvertis reikštų, kad liekanų $\varepsilon(t)$ empirinė mediana prilyginama nuliui.

3. Eksperimentų aprašymas

Tiriant metodų įvertinimo efektyvumą eksperimentiškai yra generuojamos atsitiktinės sekos su žinomais parametrais, o po to šių parametrų įverčiai lyginami su jų tikromis reikšmėmis. Metodo tikslumo matu yra naudojamas parametro įvertinimo paklaidos modulio vidurkis:

$$\Delta = \Delta(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = \mathbf{E}\sqrt{(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})}.$$

Kadangi matematinį paklaidos modulio vidurkį sunku apskaičiuoti, šiame darbe jis keičiamas aritmetiniu:

$$\overline{\Delta} = \frac{1}{N}\sum_{i=1}^{N}\Delta_i,$$

čia N bandymų skaičius, Δ_i žymi paklaidos modulį i-tos sekos atveju.

Eksperimentiškai prie įvairių sąlygų buvo tiriamas tiesinio trendo parametrų įverčių tikslumas. Tegul stebima seka tenkina lygybę: $Y(t) = a + bt + \xi(t)$, $t = 1, 2, \ldots, n$, kur $\mathbf{E}\xi(t) = 0$. Visuose toliau aprašytuose eksperimentuose buvo generuojama po 100 skirtingų nepriklausomų atsitiktinių imčių realizacijų, o parametrai $(a, b) = \boldsymbol{\theta}^T$ įvertinti šiais dešimčia metodų:

1. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ – mažiausių kvadratų metodo įvertis.
2. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\widehat{\boldsymbol{S}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{S}}^{-1}\boldsymbol{Y}$.
3. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\widehat{\Psi}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\Psi}\boldsymbol{Y}$.
4. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\widehat{\Psi}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\Psi}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{Y}$.
5. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\widehat{\boldsymbol{R}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{R}}^{-1}\boldsymbol{Y}$.
6. $\widehat{\boldsymbol{\theta}} = (\boldsymbol{X}^T\widehat{\Psi}^{1/2}\widehat{\boldsymbol{S}}^{1/2}\widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{S}}^{-1/2}\widehat{\Psi}^{1/2}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\Psi}^{1/2}\widehat{\Psi}^{1/2}\widehat{\boldsymbol{S}}^{-1/2}\widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{S}}^{-1/2}\widehat{\Psi}^{1/2}\boldsymbol{Y}$.
7. $\widehat{\boldsymbol{\theta}} = \arg\min_{\theta}\sum_{t=1}^{n}\left|Y(t) - \boldsymbol{\theta}^T\boldsymbol{x}(t)\right|$ – medianinio įvertinimo metodo įvertis.
8. $\widehat{\boldsymbol{\theta}} = \arg\min_{\theta}\sum_{t=1}^{n}\left|Y(t) - \boldsymbol{\theta}^T\boldsymbol{x}(t)\right|/\widehat{\sigma}(t)$.
9. $\widehat{\boldsymbol{\theta}} = \arg\min_{\theta}\sum_{t=1}^{n}\left|Y(t) - \boldsymbol{\theta}^T\boldsymbol{x}(t)\right|\Psi(\widehat{\varepsilon}(t))$.
10. $\widehat{\boldsymbol{\theta}} = \arg\min_{\theta}\sum_{t=1}^{n}\left|Y(t) - \boldsymbol{\theta}^T\boldsymbol{x}(t)\right|\Psi(\widehat{\varepsilon}(t))/\widehat{\sigma}(t)$.

Pastaba. Metoduose 2, 4, 6, 8, 10 kintanti dispersija yra aprašoma funkcija: $\sigma^2(t) = \exp(\beta_0 + \beta_1 t)$, kur parametrai β_0 ir β_1 yra vertinami mažiausių kvadratų metodu. Metoduose 3, 4, 5, 8, 9 svoriai $\Psi(\widehat{\varepsilon}(t))$ yra suteikiami remiantis „3σ“ taisykle. O metoduse 5 ir 6 liekanos $\varepsilon(t)$ yra laikomos $ARMA(1, 1)$ procesu.

1 eksperimentas. Šio eksperimento metu buvo gauti parametro $\boldsymbol{\theta}$ įverčiai, kai visos modelio (1) prielaidos buvo tenkinamos. Atsitiktinės imtys buvo generuojamos pagal formulę $Y(t) = a + bt + \xi(t)$, kur $a = 1$, $b = 1$.

2 eksperimentas. Ši eksperimentą aprašysime detaliau.

Tarkime, kad $\xi(t)$ yra stacionarus baltas triukšmas, bet prielaida apie liekanų normalųjį pasiskirstymą yra netenkinama, t.y. liekanos $\xi(t)$ yra pasiskirstę ne pagal Gauso, o pagal simetrinį Pareto dėsnį:

$$F_\xi(y) = \mathbf{P}\{\xi < y\} = \begin{cases} 1 - \frac{1}{2}\left(\frac{c_0}{c_0 + y}\right)^{\alpha}, & \text{kai } y>0; \\ \frac{1}{2}\left(\frac{c_0}{c_0 - y}\right)^{\alpha}, & \text{kai } y<0. \end{cases}$$

Čia $\alpha > 0$ ir c_0 yra konstantos.

1 lentelė.
Parametro įvertinimo paklaidos modulio vidurkiai

Įvertinimo metodo numeris	$\overline{\Delta}$, kai $\alpha = 1$	$\overline{\Delta}$, kai $\alpha = 2$	$\overline{\Delta}$, kai $\alpha = 3$	$\overline{\Delta}$, kai $\alpha = 4$
1.	25.49342	1.33624	0.552668	0.339779
2.	21.62688	1.358871	0.567551	0.344082
3.	5.566214	0.620278	0.359143	0.257478
4.	6.541385	0.610019	0.361619	0.25801
5.	26.07149	1.331635	0.554122	0.340496
6.	7.225453	0.865068	0.42345	0.286761
7.	0.817534	0.385431	0.251884	0.187065
8.	0.708726	0.370205	0.249438	0.199592
9.	0.811271	0.369532	0.244961	0.18019
10.	0.914457	0.379111	0.253962	0.207839

Atsitiktinės sekos buvo generuojamos su parametrais $a = 1$, $b = 1$, $c_0 = 1$, $\alpha = 1, 2, 3, 4$.

3 eksperimentas. Šio eksperimento metu buvo gauti parametro θ įverčiai, kai imitys buvo užterštos išsiskiriančiais stebėjimais. Atsitiktinės imtys buvo generuojamos pagal formulę $Y(t) = a + bt + \xi(t) + \eta(t)$, kur $a = 1$, $b = 1$, $\xi(t) \sim N(0,1)$, $\eta(t) = \begin{cases} \gamma_k, & \text{jei } t = \tau_k \\ 0, & \text{jei } t \notin \{\tau_1, \tau_2, \ldots\} \end{cases}$, $\gamma_1, \gamma_2, \ldots \sim N(0, 49)$, laiko tarpai $\tau_1, \tau_2 - \tau_1, \tau_3 - \tau_2, \ldots$ yra pasiskirstę pagal Puasono dėsnį su parametrais $\lambda = 5, 10, 15$.

4 eksperimentas. Šio eksperimento metu buvo gauti parametro θ įverčiai, kai yra netenkinama modelio (1) trečioji prielaida, t.y. kai liekanų dispersija kinta laike. Atsitiktinės imtys buvo generuojamos pagal formulę: $Y(t) = a + bt + \sigma(t)\varepsilon(t)$, kur $a = 1$, $b = 1$, $\varepsilon(t) \sim N(0, 0.25)$, $\sigma^2(t) = \exp(1 + 0,02t)$.

5 eksperimentas. Šio eksperimento metu buvo gauti parametro θ įverčiai, kai yra netenkinama modelio (1) ketvirtoji prielaida, t.y. kai liekanos $\xi(t)$ yra tarpusavyje koreliuotos. Atsitiktinės imtys buvo generuojamos pagal formulę: $Y(t) = a + bt + \xi(t)$, kur $a = 1$, $b = 1$, $\xi(t)$ yra $ARMA(1,1)$ procesas: $\xi(t) + \alpha\xi(t-1) = \widetilde{\varepsilon}(t) + \beta\widetilde{\varepsilon}(t-1)$, $\alpha = 0.95$, $\beta = 0.5$.

4. Išvados ir pasiūlymai

Atliktų eksperimentų rezultai trumpai yra pateiliami šiose išvadose.

1. Pirmojo eksperimento rezultatai rodo, kad tuo atveju kai visos modelio (1) prielaidos yra tenkinamos, tiksliausi įverčiai yra gauti mažiausių kvadratų metodu, nedaug skiriasi ir kitų šio metodo modifikacijų tikslumas. Netiksliausias šiuo atveju yra šeštasis metodas, kurio parametro įvertinimo modulio vidurkis yra apie 10% didesnis negu mažiausių kvadratų metodo. Medianinio

įvertinimo metodas bei įvairios jo modifikacijos šiame eksperimente pasirodė esą mažiau tikslūs už mažiausių kvadratų metodą. Jų visų tikslumas yra žymiai mažesnis.

2. Antrojo eksperimento rezultatai rodo, kad tuo atveju, kai yra netenkinama modelio (1) penktoji prielaida apie liekanų normalųjį pasiskirstymą, t.y. kai liekanos $\xi(t)$ yra pasiskirstę ne pagal Gauso, o pagal simetrinį Pareto dėsnį, geriausi rezultatai yra pasiekti taikant medianinį įvertinimo metodą bei įvairias jo modifikacijas. Tiksliausias šiuo atveju yra medianinis įvertinimo metodas su svoriais dėl išsiskiriančių stebėjimų (9 metodas). Mažiausių kvadratų metodu ir jo modifikacijų pagalba gauti rezultatai yra mažiau tikslūs.

3. Trečiojo eksperimento rezultatai rodo, kad tuoatveju, kai liekanos yra pasiskirstę pagal Gauso dėsnį, tačiau imtis yra užteršta išsiskiriančiais stebėjimais, tiksliausi įverčiai yra medianinio įvertinimo metodo bei įvairių jo modifikacijų. Čia tiksliausias yra medianinis įvertinimo metodas su svoriais dėl išsiskiriančių stebėjimų (9 metodas). Netiksliausias pasirodė esąs penktasis metodas, kurio pavyzdžiui parametro įvertinimo modulio vidurkis, kai imtis yra užteršta apie 10% išsiskiriančių stebėjimų, yra apie 100% didesnis negu aštuntojo metodo.

4. Iš ketvirtojo eksperimento rezultatų matyti, kad tuo atveju kai yra netenkinama modelio (1) trečioji prielaida, t.y. liekanų dispersija kinta laike, tiksliausi rezultatai yra gauti mažiausių kvadratų metodo su kintančia dispersija (2 metodas). Nedaug skiriasi ir kitų šio metodo modifikacijų tikslumas. Medianinio įvertinimo metodų grupės gauti rezultatai yra mažiau tikslūs. Netiksliausias pasirodė esąs mažiausių kvadratų metodas, kurio parametro įvertinimo modulio vidurkis yra apie 50% didesnis negu antrojo metodo.

5. Iš penktojo eksperimento rezultatų matyti, kad kai yra netenkinama modelio (1) ketvirtoji prielaida, t.y. kai liekanos $\xi(t)$ yra tarpusavyje koreliuotos ($\xi(t)$ yra $ARMA$ procesas), tiksliausi įverčiai yra šeštojo metodo, nors visų mažiausių kvadratų metodo modifikacijų tikslumas skiriasi nedaug. Medianinio įvertinimo metodo įverčiai šiuo atveju yra mažiau tikslūs.

6. Apskaičiuojant parametro $\boldsymbol{\theta}$ medianinį įvertį praktikoje galima naudoti rekurentinę procedūrą:

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\min_{\boldsymbol{\theta}} \sum_{t=1}^{n} \left(Y(t) - \boldsymbol{\theta}^{T(k)}\boldsymbol{x}(t)\right)^2 w^{(k)}(t),$$

$$\text{kur} \quad w^{(0)}(t) = 1, \quad w^{(k)}(t) = 1/\left|Y(t) - \boldsymbol{\theta}^{T(k)}\boldsymbol{x}(t)\right|, \quad k = 1, 2, \ldots.$$

Ši procedūra realizuojama panaudojant standartinę apibendrinto mažiausių kvadratų metodo procedūrą, kuri yra visuose didesniuose statistiniuose pajetuose. Jos pagalba gaunamas įvertis konverguoja į medianinį parametro įvertį, kai iteracijų skaičius k auga.

Literatūra

[1] R. Salėtis, *Apie tiesinės regresijos modelių identifikavimo metodus*, MII, Vilnius (1997).

[2] П. Хьюбер, *Робастность в статистике*, Мир, Москва (1984).

[3] С.А. Айвязян, В.С. Мхитарян, *Прикладная статистика и основы эконометрики*, Юнити, Москва (1998).

[4] Ю. Тюрин, Г. Симонова, *Знаковый анализ линейных моделей*, Обозренеие прикладной и ромышленной математики, 215–278 (1993).

On identification methods of linear regression models

R. Salėtis

The aim of this paper is to discus the efficiency of the least squares and least absolute values estimation methods, in such situations, then the sample is with outliers, the variance of the residuals changes in the time and the residuals are correlated. A range of different modifications of these methods in order to improve accuracy of the estimation is offend. All the discussed procedures are fulfilled in a PC by means of SAS (*Statistical Analysis System*).