Vilniaus Universitetas
Matematikos ir informatikos fakultetas
Matematinės statistikos katedra

Anastasija Pereskokova

Magistro baigiamasis darbas

Išsiskiriančių reikšmių nustatymas duomenyse apie Lietuvos užsienio prekybą

Darbo vadovas:
Doc. Vytautas Kazakevičius

2006
Vilnius

Turinys:

Įvadas

Lietuvos užsienio prekybos statistinės informacijos vartotojai turi aukštus reikalavimus šių duomenų kokybei. Todėl iš respondentų gauti duomenys prieš juos skelbiant yra tikrinami Statistikos departamento Užsienio prekybos skyriaus darbuotojais. Kadangi informacija iš respondentų yra gaunama kas mėnesį dideliais kiekiais, neįmanoma patikrinti kiekvieno įrašo. Taigi yra įdiegta nemažai kontrolių įvairių rūšių klaidų nustatymui. Siekiant gerinti duomenų kokybę ir užtikrinti kuo didesnį klaidų nustatymo procentą, kontrolių skaičius yra didinamas.
Šio darbo tikslas – išbandyti išsiskiriančių reikšmių nustatymo metodą duomenims apie Lietuvos užsienio prekybą, išanalizuoti rezultatus ir ateityje gal būt paversti tai dar viena kontrole.
Išsiskiriančių reikšmių nustatymas yra vienas iš metodų, tikrinant statistinių duomenų patikimumą. Šiuo metu jis jau yra įdiegtas keliose Europos Sąjungos valstybėse, tikrinant užsienio prekybos statistikos duomenis, tačiau Lietuvoje dar nenaudojamas. Eurostatas yra parengęs metodiką nustatant išsiskiriančias reikšmes duomenyse apie užsienio prekybą. Ji yra išbandyta šiame darbe.

Bendra informacija apie duomenis

Užsienio prekybos statistika iki 2004 m. gegužės mėnesio buvo rengiama remiantis Muitinės deklaracijos (Bendrojo administracinio dokumento) duomenimis. Muitinės departamentas kiekvieną mėnesį Statistikos departamentui pateikia Bendrojo administracinio dokumento pirminius duomenis, gautus iš eksportuotojų ir importuotojų.
Įstojus į Europos Sąjunga (ES), muitų sienos su ES šalimis buvo panaikintos. Lietuvos užsienio prekybos duomenys pasidalijo į dvi dalis – Ekstrastatą ir Intrastatą. Ekstrastato sistema apima Lietuvos prekybos su ne ES šalimis duomenis, kurių šaltinis ir toliau yra Muitinės deklaracijos (Bendrojo administracinio dokumento) informacija. Intrastatas – tai duomenų surinkimo sistema iš įmonių, kurios prekiauja su ES šalimis.
Lietuvos užsienio prekybą apibūdina keturi rodikliai: eksportas ir importas, apibrėžiantys prekybos apimtis su ne ES šalimis bei išvežimas ir įvežimas, apibūdinantys prekybą su ES šalimis.
Užsienio prekybos statistikoje prekės klasifikuojamos ir koduojamos remiantis Kombinuotąja nomenklatūra (toliau - KN). Iš respondentų yra surenkama informacija apie prekių, identifikuojamų pagal aštuonių Kombinuotosios nomenklatūros skaitmenų kodą, gabenimą. Prekių neto masė apskaičiuojamas kilogramais, tačiau daliai prekių kodų suteikiami papildomi mato vienetai, pagal kuriuos yra renkama informacija apie šių eksportuojamų arba importuojamų prekių kiekį.
Dar vienas iš rodiklių, pagal kurį iš respondentų yra renkama informacija apie eksportuojamas ir importuojamas prekes yra statistinė prekių vertė, kuri apskaičiuojama sumuojant prekės vertę su transporto ir draudimo išlaidomis iki Lietuvos sienos.

Išsiskiriančių reikšmių nustatymui, buvo analizuojami eksporto neagreguoti duomenys nuo 2005 m. kovo mėnesio iki 2006 m. kovo mėnesio. Tai yra informacija apie visas eksportuotas iš Lietuvos Respublikos per nagrinėjamą laikotarpį prekes. Iš Kombinuotosios nomenklatūros aštuonženklių kodų, kurių viso yra 9848, analizei buvo pasirinkti tik pirmi 8 skirsniai (1065 kodai), t.y. žemės ūkio produktai, kur stebėjimų skaičius buvo ne mažesnis nei trisdešimt stebėjimų. Tokiu būdu į analizę pakliuvo tik 168 KN kodai.

Patikimumo kontrolei svarbu nustatyti sąryšius tarp kintamųjų. Fiksuojant srautą, KN8 kodą ir metus buvo analizuojamos trys laiko eilutės, išvardintos lentelėje:

Priklausomas kintamasis y	Nepriklausomas kintamasis x	Modelio pavadinimas	Laiko eilučių skaičius
Eksportuojamų prekių statistinė vertė	Eksportuojamų prekių neto masė	MODELIS 1	168
Eksportuojamų prekių statistinė vertė	Eksportuojamų prekių kiekis papildomais mato vienetais	MODELIS 2	7
Eksportuojamų prekių neto masė	Eksportuojamų prekių kiekis papildomais mato vienetais	MODELIS 3	7

Antrame ir trečiame modelyje, laiko eilučių skaičius mažesnis dėl tos priežasties, kad ne visiems Kombinuotosios nomenklatūros kodams yra priskiriami papildomi mato vienetai.

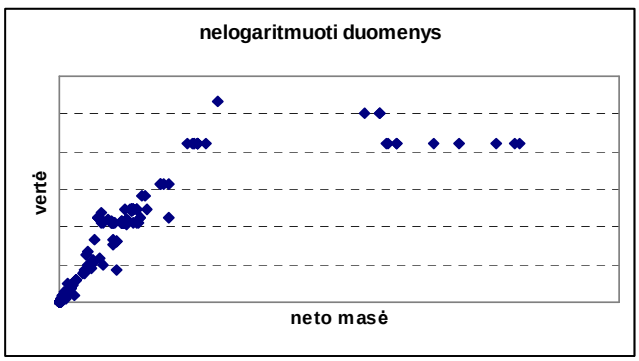
Trumpas metodo aprašymas

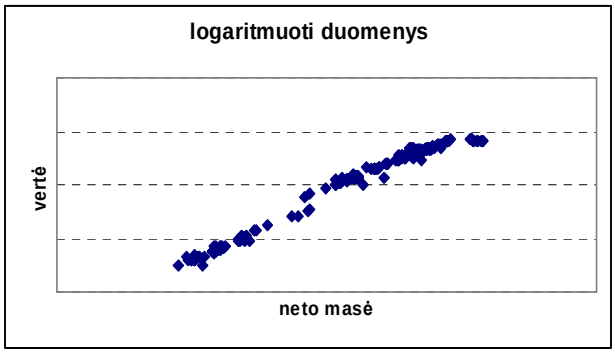
Pasirinkto metodo išsiskiriančioms reikšmėms nustatyti esminė idėja – parinkti regresijos modelį naudojantis Mažiausių kvadratų metodu (toliau - MKM), apskačiuoti prognozės intervalą ir pažymėti reikšmes, kurios į jį nepatenka. Kadangi MKM reikalauja atsižvelgti į tam tikras prielaidas, duomenys buvo tikrinami testais, ir, jeigu prielaidos nebuvo tenkinamos, išsiskiriančių reikšmių nustatymui buvo rekomenduojama naudoti alternatyvųjį metodą.

Duomenų analizė

Duomenų paruošimas ir galimų išsiskiriančių reikšmių pašalinimas, įvertinant modelį (trijų pakopų filtracija)

Kadangi analizuojant Y ir X kintamųjų priklausomybę, buvo nustatyta, kad ji ne visada tiesinė, prieš nustatant modelį, duomenims buvo pritaikyta logaritmo transformacija.





Taigi buvo nagrinėjamas modelis $Ln(Y_i) = \alpha + \beta \cdot Ln(X_i) + \varepsilon_i$, kuriam reikėjo apskaičiuoti α ir β (čia Y_i, X_i yra vektoriai, indeksas i reiškia nagrinėjamos laiko eilutės numerį, $Y_i = \{y_{i1},...,y_{it},...,y_{iN_i}\}, X_i = \{x_{i1},...,x_{it},...,x_{iN_i}\}, N_i -$ stebėjimų skaičius i-tojoje laiko eilutėje). Pažymėkime $\tilde{Y}_i = Ln(Y_i), \tilde{X}_i = Ln(X_i)$, ir toliau šiame darbe bus naudojamas šis pažymėjimas.

Kadangi MKM yra jautrus išsiskiriančioms reikšmėms, norėta pašalinti iš analizės galimas išskirtis prieš įvertinant modelio regresijos parametrus. Tam buvo naudojamas Pasvertų mažiausių kvadratų metodas *(Weighted Least Squares)*.

Buvo atliekami tokie žingsniai:

1) Atliekama tiesinė regresija, taikant MKM.
2) Kiekvienam stebėjimui nustatomas nuokrypis MKM įvertintos $\tilde{y}_{ij}$ reikšmės nuo tikrosios reikšmės: $d_{it} = \tilde{y}_{it} - \hat{\tilde{y}}_{it}$, (čia *i* – fiksuotas nagrinėjamos laiko eilutės numeris, o *t* galima traktuoti kaip laiko momentą, arba stebėjimo laiko eilutėje numerį, $t = 1,...N_i$.);
3) Kiekvienam stebėjimui skaičiuojamas $z_{it} = \frac{d_{it}}{3 * IQR}$, kur *IQR = Q3 – Q1*, o *Q1* ir *Q3* yra atitinkamai apatinis ir viršutinis d_i kvartiliai, kur $d_i = \{d_{it}, t = 1,..., N_i\}$.
4) Paskaičiuojami svoriai: $w_{it} = \begin{cases} (1 - z_{it}^{\ 2})^2, kai |z_{it}| \le 1 \\ 0, kai |z_{it}| \succ 1 \end{cases}$.
5) Atliekama regresija, taikant Pasvertų mažiausių kvadratų metodą – minimizuojant pasvertų liekanų skirtumų kvadratų sumą *(weighted residual sum of squares)*, $\sum_{t=1,...,N_i} w_{it} \cdot (\tilde{y}_{it} - \hat{\tilde{y}}_{it}) \to \min$.

Pasvertoji regresija atliekama tris kartus, kiekvieną kartą iš naujo perskaičiuojant svorius w_{it}, kaip aprašyta (2)-(4) žingsniuose. Paskutinį kartą apskaičiavus svorius, iš laiko eilučių pašalinami stebėjimai, kur $w_{it} = 0$.

Autokoreliacijos atžvilgiu ištaisytas regresijos modelis

Kadangi analizuojami duomenys yra laiko eilutės, Eurostato metodikoje rekomenduojama atsižvelgti į jų galimą statistinę priklausomybę įvairiais laiko momentais. Buvo tikrinama pirmos eilės autokoreliacija.

Nagrinėkime modelį $\tilde{Y}_i = \alpha + \beta \cdot \tilde{X}_i + \varepsilon$, kur vektoriaus $\tilde{Y}_i$ *t*-toji komponentė – tai priklausomo kintamojo reikšmė laiko momentu *t*, $t = 1,..., N_i$. Laiko momentu *t* lygtis atrodys taip: $\tilde{y}_{it} = \alpha + \beta \cdot \tilde{x}_{it} + \varepsilon_t$. Tai, kad atsitiktinė seka $\{\varepsilon_t, t = 1,..., n\}$ sudaro autoregresinį pirmos eilės procesą reiškia, kad klaidos gali būti užrašomos tokiomis rekurentiškomis lygybėmis:

$$\varepsilon_t = \rho \cdot \varepsilon_{t-1} + \nu_t, \qquad (1.1)$$

kur $\{\nu_t, t = 1,..., n\}$- nepriklausomų normalių atsitiktinių dydžių su nuliniu vidurkiu ir pastovia dispersija σ_ν^2 seka, o ρ - vadinamasis autoregresijos koeficientas ($|\rho| \le 1$). Griežtai kalbant, pilnam modelio aprašymui reikia nustatyti ε_0. Laikysime, kad ε_0 - normalus atsitiktinis dydis su nuliniu vidurkiu ir dispersija $\sigma_\varepsilon^2 = \sigma_\nu^2 \big/ (1 - \rho^2)$, kuri nepriklauso nuo $\{\nu_t, t = 1,..., n\}$. Paskaičiavę abiejų lygybės (1.1) pusių vidurkį, gauname $E\varepsilon_t = \rho \cdot E\varepsilon_{t-1}$, iš kur seka, kad $E\varepsilon_t = 0, t = 1,..., N_i$. Kadangi ε_{t-1} išreiškiamas per $\nu_1,..., \nu_{t-1}$, vadinasi ε_{t-1} ir ν_t nepriklausomi. Todėl $E(\varepsilon_t^2) = E(\rho \cdot \varepsilon_{t-1} + \nu_t)^2 = \rho^2 E(\varepsilon_{t-1}^2) + \sigma_\nu^2$.

Jeigu $E(\varepsilon_0^2) = \sigma_\nu^2 \big/ (1 - \rho^2)$, tai $\sigma_\varepsilon^2 = E(\varepsilon_t^2) = V(\varepsilon_t) = \sigma_\nu^2 \big/ (1 - \rho^2)$, $t = 1,..., N_i$ (1.2).

Padauginę (1.1) iš ε_{t-1} ir pasinaudoję ε_{t-1} ir ν_t nepriklausomumu, gauname

$$E(\varepsilon_t, \varepsilon_{t-1}) = Cov(\varepsilon_t, \varepsilon_{t-1}) = \rho \cdot V(\varepsilon_{t-1}) = \rho \cdot \sigma_\varepsilon^2 \quad (1.3).$$

Analogiškai, $Cov(\varepsilon_t, \varepsilon_{t-2}) = \rho^2 \cdot \sigma_\varepsilon^2$ ir

$$Cov(\varepsilon_t, \varepsilon_{t-m}) = \rho^m \cdot \sigma_\varepsilon^2 \quad (1.4).$$

Taigi seka $\{\varepsilon_t\}$- sudaro stacionarų atsitiktinį procesą - $P(\varepsilon_t \geq x) = P(\varepsilon_0 \geq x)$. Iš (1.3) išplaukia, kad

$$\rho = Cov(\varepsilon_t, \varepsilon_{t-1}) / \sigma_\varepsilon^2 = Cov(\varepsilon_t, \varepsilon_{t-1}) / (V(\varepsilon_t)^{1/2} V(\varepsilon_{t-1})^{1/2}),$$

t.y. ρ yra koreliacijos koeficientas tarp dviejų kaimyninių paklaidų. Pasinaudoję (1.4) išrašysime atsitiktinio vektoriaus ε kovariacijos matricą:

$$\Omega = \frac{\sigma_\nu^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

Modeliai su autoregresija buvo įvertinami, taikant Durbino procedūrą. Kadangi tokie atvejai, kai autoregresijos parametras ρ yra žinomas pasitaiko labai retai, atsiranda būtinybė įvertinti ρ.
Užrašysim $\tilde{y}_{it} = \beta \cdot \tilde{x}_{it} + \varepsilon_t$ (1.5) laiko momentui *t-1* ($t \geq 2$)

$$\tilde{y}_{it-1} = \alpha + \beta \cdot \tilde{x}_{it-1} + \varepsilon_{t-1}$$

padauginsime abi puses iš ρ ir atimsime panariui iš (1.5). Tada atsižvelgę į (1.1), gausime

$$\tilde{y}_{it} - \rho \cdot \tilde{y}_{it-1} = \alpha(1-\rho) + \beta \cdot (\tilde{x}_{it} - \rho \cdot \tilde{x}_{it-1}) + \nu_t \quad (1.6).$$

Perrašom (1.6):

$$\tilde{y}_{it} = \alpha(1-\rho) + \rho \cdot \tilde{y}_{it-1} + \beta(\tilde{x}_{it} - \rho \cdot \tilde{x}_{it-1}) + \nu_t,$$

t.y. $\tilde{y}_{it-1}$ prijungiamas prie regresorių, o ρ - vertinamų parametrų skaičius. Šiai sistemai sudaromi įprasti MKM parametrų ρ, $\rho\alpha$ ir $\rho\beta$ įverčiai - atitinkamai *r*, $\hat{\theta}_1$ ir $\hat{\theta}_2$. $\hat{\alpha}, \hat{\beta}$ įverčiams įmami santykiai atitinkamai $\hat{\theta}_1 / r, \hat{\theta}_2 / r$.

Įvertinus modelį, kiekvienai laiko eilutei buvo skaičiuojamos tokios statistikos:

- stebėjimų skaičius *N*;
- determinacijos koeficientas $RSQR_i = \dfrac{\sum_t \left(\hat{\tilde{y}}_{it} - \bar{\bar{Y}}_i\right)^2}{\sum_t \left(\tilde{y}_{it} - \bar{\bar{Y}}_i\right)^2}$, (kur $\bar{\bar{Y}}_i = \dfrac{1}{n}\sum_t \tilde{y}_{it}$);
- nuokrypių kvadratų suma $SSD_i = \dfrac{\sum_t \left(\tilde{x}_{it}^2 - \bar{\bar{X}}_i\right)^2}{n-1}$, (kur $\bar{\bar{X}}_i = \dfrac{1}{N_i}\sum_t \tilde{x}_{it}$);
- standartinis nuokrypis $SY_i = \sqrt{\sum_t \left(\tilde{y}_{it}^2 - \bar{\bar{Y}}\right)^2}$;
- vidurkis $XMOY_i = \bar{\bar{X}}_i$.

MKM prielaidų tikrinimas

Toliau yra tikrinamos MKM prielaidos. Lentelėje pateikiami naudojami testai:

Prielaida	**Testas**
Homoskedastiškumas	White testas
Nepriklausomumas	Durbin *t*-testas
Normališkumas	Bera-Jarque testas
Tiesiškumas	Determinacijos koeficientas R^2

White testas

Jeigu modelis yra heteroskedastiškas, tai dažnai susiję su tuo, kad paklaidų dispersijos kažkokiu būdu priklauso nuo regresorių, o heteroskedastiškumas turi kažkaip atsispindėti regresijos paklaidose. Tikrinama hipotezė H_0, kad modelis homoskedastiškas.

Iš pradžių pritaikius modeliui $\tilde{Y}_i = \tilde{X}_i\beta + \varepsilon$ MKM metodą randamos regresijos liekanos e_{it}, $t = 1,..., N_i$.

Tada skaičiuojama šių paklaidų kvadratų e_{it}^2 regresija su regresoriais X_i, jų kvadratais, tarpusavio sandaugomis ir konstanta – šiame darbe tai buvo regresija $e_{it}^2 = \alpha + \beta_1\tilde{x}_{it} + \beta_2\tilde{x}_{it}^2 + \varepsilon$. Tada esant hipotezei H_0 dydis $n \cdot R_i^2$ asimptotiškai pasiskirstęs pagal $\chi^2(n-1)$, kur R_i^2- determinacijos koeficientas, o *n* – skaičius regresorių antroje regresijoje.

Durbin *t*-testas

Tikrinant regresijos liekanų nepriklausomumą, atliekamas Durbino *t*-testas. Skaičiuojama regresija $\hat{e}_{it}$ ant $\hat{e}_{it-1}$ ir nepriklausomų regresorių ir tikrinama, ar reikšmingas koeficientas prie $\hat{e}_{it-1}$. Jeigu koeficientas reikšmingas, sakoma, kad liekanos priklausomos.

Šiame darbe analizuojamiems duomenims šis testas aprašomas taip.

- gaunamas regresijos modelis: $\hat{e}_{it} = \alpha + \beta_1 \cdot \hat{e}_{it-1} + \beta_2 \cdot \tilde{x}_{it} + \varepsilon$.
- Tuomet *Stjudento* kriterijumi tikrinama statistinė hipotezė: $H_0 : \beta_1 = 0$.
- Apskaičiuojama kriterijaus statistika $t = \dfrac{\hat{\beta}_1}{\sqrt{MSE/SSE_1}}$, kur SSE_1- regresijos modelio $\hat{e}_{it-1} = \alpha + \beta \cdot \tilde{x}_{it} + \varepsilon$ liekamųjų paklaidų kvadratų suma, o $MSE = \dfrac{SSE}{n-k-1}$, (k – nepriklausomų kintamųjų skaičius regresijoje, šiuo atveju *k = 2*).
- Tegul reikšmingumo lygmuo lygus α. Hipotezė H_0 atmetama, jeigu $|t| > t_{\alpha/2}(n-3)$. Čia $t_{\alpha/2}(n-3)$ - yra Stjudento skirstinio su (*n – 3)* laisvės laipsinių $\alpha/2$ lygmens kritinė reikšmė.

Bera-Jarque testas

Šiuo testu tikrinamas normališkumas. Skaičiuojama Bera-Jarque (1982) statistika:

$$T_{N_i} = \left[\frac{N_i}{6} b_{1_i}^2 + \frac{N_i}{24} (b_{2i} - 3)^2 \right], \text{ kur}$$

$$b_{1i} = \frac{\sqrt{N_i} \sum_{t=1}^{N_i} \hat{u}_{it}^3}{\left(\sum_{t=1}^{N_i} \hat{u}_{it}^2 \right)^{\frac{3}{2}}}$$

$$b_{2_i} = \frac{\sqrt{N} \sum_{t=1}^{N_i} \hat{u}_{it}^4}{\left(\sum_{t=1}^{N_i} \hat{u}_{it}^2 \right)^2},$$

kur $\hat{u}_{it}$ - MKM liekanų vektorius. Esant teisingai H_0, statistika T_{N_i} aproksimuojama $\chi^2(2)$ pasiskirstymu.

Homoskedastiškumo, nepriklausomumo ir normališkumo testams nustatomas reikšmingumo lygmuo $\alpha = 0.01$. Laikoma, kad tenkinama tiesiškumo prielaida, kai $R_i^2 > 0{,}65$ ir $\beta_i > 0$.

Kai visos MKM prielaidos tenkinamos, buvo rekomenduojama išsiskiriančių reikšmių nustatymui taikyti regresijos metodą, priešingu atveju – alternatyvųjį metodą.

Alternatyvusis metodas

Kai MKM prielaidos nėra tenkinamos, išsiskiriančių reikšmių nustatymui buvo rekomenduojama naudoti alternatyvųjį metodą. Imami stebėjimai, kurie liko tris kartus minimizavus pasvertų liekanų kvadratų sumą, skaičiuojamas santykis $Ln\frac{Y_i}{X_i} = Ln(Y_i) - Ln(X_i)$, šiam santykiui apskaičiuojami apatinis ir viršutinis kvartiliai *Q1, Q3* ir apatinis bei viršutinis pasikliovimo lygmenys: $UL = Q3 + 4 * IQR,\ LL = Q1 - 4 * IQR$. Jeigu santykis nepakliūna tarp viršutinio ir apatinio pasikliovimo lygmens, laikoma, kad tai yra išsiskirianti reikšmė.

Regresijos metodas

Išsiskiriančių reikšmių nustatymui pagal regresijos metodą buvo taikomas prognozės intervalas, apskaičiuojamas pagal tokią formulę: $(\alpha + \beta \cdot \tilde{x}_{it}) \pm t_{0.99} \cdot sy \sqrt{\frac{1}{n} + \frac{(\tilde{x}_{it} - xmoy)^2}{ssd} + 1}$, kur $t_{0.99}$ randamas Stjudento p-reikšmių lentelėje (imamas laisvės laipsnių skaičius *df = n-2*).

Buvo naudojama Studento skirstinio *P*-osios kritinių reikšmių lentelės iš J. Kruopio knygos „Matematinė statistika“ ir iš *Thesaurus* internetinės enciklopedijos.
Kadangi šiose lentelėse *P*-osios kritinės reikšmės nurodytos ne visiems n, ieškant $t_{0.99}(n-2)$ reikšmių atitinkančių tarpines argumento *n* reikšmes, kurių nėra lentelėse, buvo taikoma tiesinio interpoliavimo formulė argumento $\frac{1}{n-2}$ atžvilgiu: $t_{0.99}(n-2) = (1-u)t_{0.99}(n_0) + ut_p(n_1)$, čia n_0 ir

n_1 ($n_0 <$ (n - 2) $< n_1$) – gretimos argumento reikšmės iš lentelės, o

$$u = \left(\frac{1}{n_0} - \frac{1}{(n-2)}\right) \Big/ \left(\frac{1}{n_0} - \frac{1}{n_1}\right) = \frac{(n-n_0)n_1}{(n_1-n_0)n}.$$

Kai duotam $\tilde{x}_{it}$ tikra $\tilde{y}_{it}$ reikšmė nepatenka į prognozės intervalą, toks stebėjimas laikomas išsiskiriančiu.

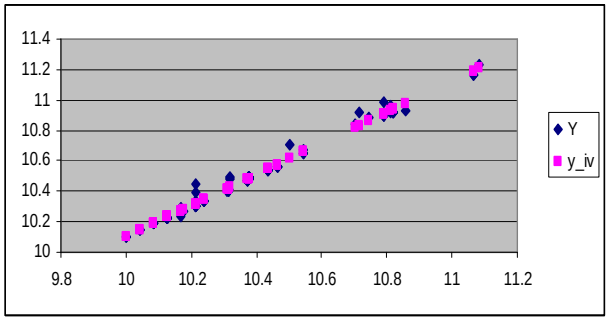
Rezultatai

Duomenų analizė buvo atliekama su SAS 8.02 versija. Buvo naudojamos dvi stebėjimų imtys. Viena imtis – tai jau pataisyti 2005 m. 1-12 mėnesių ir 2006 m. 1-2 mėnesių duomenys. Antra - pradiniai to paties periodo plius 2006 m. trečio mėnesio duomenys. Viso 33759 stebėjimų pirmoje imtyje ir 38878 stebėjimų antroje imtyje. Siekiant kuo mažesnių nukrypimų, modelis buvo įvertinamas ir pasikliovimo lygmenys buvo nustatomi pagal ištaisytos imties duomenis. Išsiskiriančios reikšmės buvo nustatomos imtyje su neištaisytais duomenimis. Norėta patikrinti ar aukščiau aprašytais metodais bus nustatytos tikros išsiskiriančios reikšmės, ar jos atitiks anksčiau nustatytas išsiskiriančias reikšmes ir ar papildomos nustatytos išsiskiriančios reikšmės pasiteisina. Kitaip sakant, norėta patikrinti aukščiau aprašytų metodų efektyvumą ir visų jų žingsnių naudingumą.

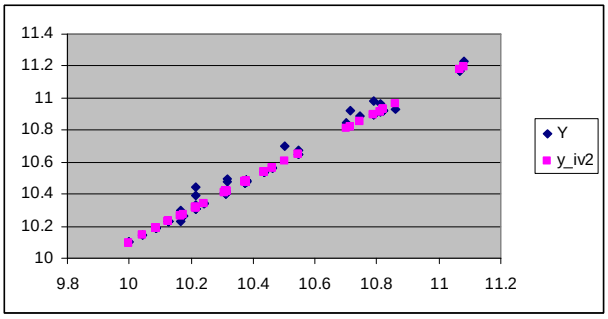
1 žingsnis (trijų pakopų filtracija)

Paimkim tik tas imtis, kuriose buvo išfiltruoti stebėjimai, kaip galimos išsiskiriančios reikšmės. Žemiau pateiktuose grafikuose, matosi, kad atsitiktinai parinktoje imtyje, besikeičiant svoriams *w*, *pasvertosios regresijos* prognozės $\hat{\tilde{Y}}_i$ kito labai nežymiai. Grafike 1.f parodyta, kaip autoregresijos prognozuojamos $\tilde{Y}_i$ reikšmės atitinka tikras, o grafike 1.e – palyginamos *pasvertosios regresijos* ir *autoregresijos* $\tilde{Y}_i$ įverčiai.

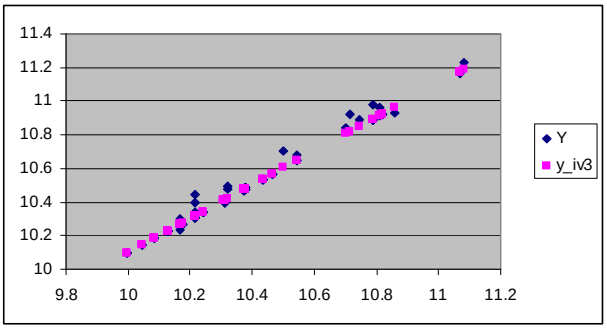
1.a



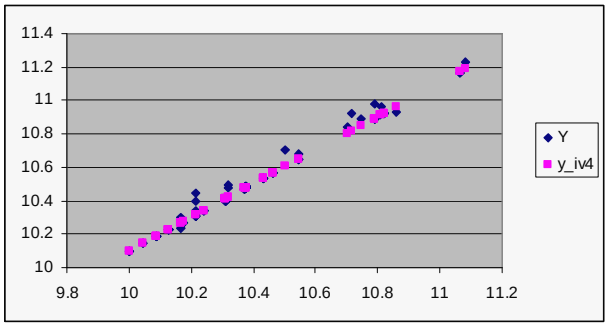
1.b



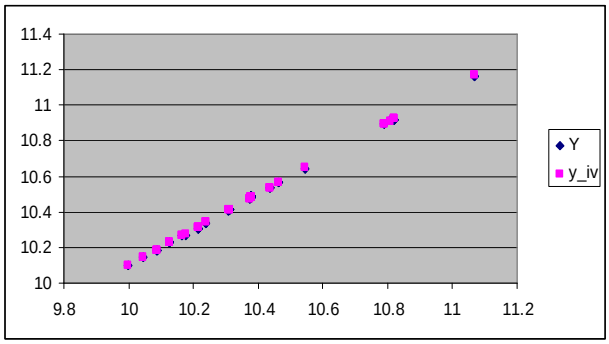
1.c



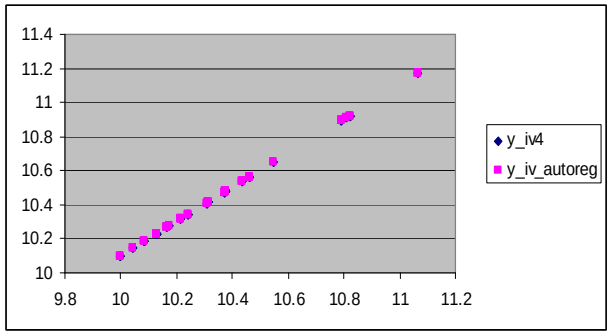
1.d



1 f.



1 e.



Taigi galima padaryti išvadą, kad galimų išsiskiriančių reikšmių „išfiltravimas“ ir suteikimas joms nulinių svorių w_{it}, labai nežymiai įtakoja regresijos parametrų įverčius $\hat{\alpha}$ ir $\hat{\beta}$. Tačiau suteikus svorius stebėjimams, pasikeičia statistikos $SSD_i, XMOY_i$ ir SY_i, ir tai įtakoja prognozės intervalo nustatymą ir, atitinkamai, išsiskiriančių reikšmių, nustatytų, taikant regresijos metodą, skaičių. Kadangi, taikant alternatyvų metodą apatinio ir viršutinio pasikliovimo lygmens nustatymui imamos jau „išfiltruotos“ imtys, šis žingsnis įtakoja ir alternatyvaus metodo nustatytų išsiskiriančių reikšmių skaičių.
Šio žingsnio metu šios tris laiko eilutės buvo visiškai pašalintos iš analizės:

Prekės KN kodas	Modelis
01059200	MODELIS 1
01059200	MODELIS 2
01059200	MODELIS 3

Šiose laiko eilutėse X_i, Y_i ir d_i įgyja tik tris reikšmes, todėl apatinis ir viršutinis kvartiliai nustatomi vienodi ir dydis z_i tampa neapibrėžtas.

2 žingsnis (regresijos modelio ištaisymas autokoreliacijos atžvilgiu)

Analizuotose laiko eilutėse buvo tikrinama pirmos eilės autokoreliacija. Esant pirmos eilės autokoreliacijai MKM nustato nepaslinktus regresijos parametrų įverčius, tačiau galima parodyti, jog gaunamas dispersijos įvertis yra paslinktas žemyn, kas gali turėti įtakos tikrinant hipotezes apie

regresijos parametrų reikšmingumą. Tačiau šiame darbe tokie testai nebuvo atliekami, kadangi kiekviename modelyje tėra tik vienas regresorius.
Atlikus Durbin-Watson *t*-testą, autokoreliacija buvo nustatyta tik devyniolikai laiko eilučių. Autoregresijos paskaičiuoti regresijos parametrų įverčiai ir gauti priklausomų kintamųjų įverčiai nuo anksčiau gautų įverčių skyrėsi labai nereikšmingai.
Taigi galima padaryti išvadą, jog šis žingsnis neturėjo įtakos tolimesniems skaičiavimams.
Be to, tikrinant autokoreliaciją, stebėjimai turi būti surikiuoti didėjančia tvarka pagal laiko momentus. Kitaip tampa neaišku, kokios būtent eilės autokoreliacija tikrinama. Duomenims, kurie buvo analizuojami šiame darbe, neįmanoma nustatyti eiliškumo pagal laiką, be to tą daryti nėra prasmės, nes duomenys gaunami iš visos Lietuvos skirtingų įmonių ir negrupuojami pagal įmones šioje analizėje. Dėl šių priežasčių autokoreliacijos tikrinime ir modelio ištaisyme autokoreliacijos atžvilgiu nėra prasmės.

3 žingsnis (MKM prielaidų tikrinimas)

Patikrinus MKM prielaidas gauti tokie rezultatai:

Modelis	Laiko eilutės, kuriose buvo nustatytas heteroskedastiškumas		Laiko eilutės, su autokoreliacija		Laiko eilutės, kuriose paklaidos netenkina normališkumo prielaidos		Laiko eilutės, kuriose determinacijos koeficientas $R^2 < 0,65$		Laiko eilutės, kuriose regresijos parametro įvertis $\hat{\beta} < 0$	
	Skaičius	Dalis procentais (%)	Skaičius	Dalis procentais (%)	Skaičius	Dalis procentais (%)	Skaičius	Dalis procentais (%)	Skaičius	Dalis procentais (%)
Modelis 1	138	82,63	18	10,71	52	30,95	13	7,26	1	0,56
Modelis 2	6	100,00	0	0,00	2	33,33	0	0,00	0	0,00
Modelis 3	4	66,67	1	16,67	1	16,67	1	16,67	0	0,00

Pagal gautus rezultatus regresijos metodas išsiskiriančių reikšmių nustatymui turėtų būti taikomas tik aštuoniolikai laiko eilučių, kas sudaro 10,05 % nuo visų laiko eilučių skaičiaus. Tačiau, norėta palyginti alternatyvaus ir regresijos metodo, nustatant išsiskiriančias reikšmes rezultatus ir todėl buvo nusižengta MKM reikalavimams, ir išsiskiriančios reikšmės buvo nustatomos abiem metodais visiems stebėjimams. Vis dėl to ši informacija nėra prarandama, ir žemiau bus pateikta su ja susijusi statistika.

Išsiskiriančios reikšmės

Šiame skyrelyje pateikiami rezultatai, gauti nustačius išsiskiriančias reikšmes abiem metodais. Taikytų metodų efektyvumui nustatyti buvo pasinaudota jau turima patikrinta informacija apie duomenis. Analizuotoje imtyje Užsienio prekybos statistikos skyriaus darbuotojais per visą periodą viso buvo nustatyta 61 išsiskirianti reikšmė, kas sudaro 0,16 procentų nuo visų stebėjimų, iš jų 2006 m. kovo mėnesiui priklauso tik keturi stebėjimai. Tai yra informacija apie jau patikrintas ir pataisytas išskirtis, tačiau jos negalima laikyti absoliučia ir teigti, kad visi kiti duomenys tikrai geri, todėl žemiau pateiktose lentelėse žodis geri paimtas į kabutes.

Nustatytos išskirtys, taikant alternatyvų metodą, naudojant svorius

		Alternatyvus metodas									
		Visas periodas				2006 m. Kovo mėnuo					
		„Geri“ duomenys		Nustatytos išskirtys		„Geri“ duomenys		Nustatytos išskirtys		Pasiteisinusios išskirtys	
		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*	
		A	R	A	R	A	R	A	R	A	R
Praktikoje patikrinti duomenys	„Geri“ duomenys	36503	818	1330	64	4821	66	220	8	11	0
	Išskirtys	0	0	61	0	0	0	3	0	3	0

Nustatytos išskirtys, taikant alternatyvų metodą, nenaudojant svorių

		Alternatyvus metodas									
		Visas periodas				2006 m. Kovo mėnuo					
		„Geri“ duomenys		Nustatytos išskirtys		„Geri“ duomenys		Nustatytos išskirtys		Pasiteisinusios išskirtys	
		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*	
		A	R	A	R	A	R	A	R	A	R
Praktikoje patikrinti duomenys	„Geri“ duomenys	37546	421	721	27	4997	32	83	3	11	0
	Išskirtys	0	0	61	0	0	0	3	0	3	0

Nustatytos išskirtys, taikant regresijos metodą, naudojant svorius, kai P = 1%

		Regresijos metodas									
		Visas periodas				2006 m. Kovo mėnuo					
		„Geri“ duomenys		Nustatytos išskirtys		„Geri“ duomenys		Nustatytos išskirtys		Pasiteisinusios išskirtys	
		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*	
		A	R	A	R	A	R	A	R	A	R
Praktikoje patikrinti duomenys	„Geri“ duomenys	37807	874	26	8	5038	73	2	1	1	0
	Išskirtys	7	0	54	0	3	0	1	0	1	0

Nustatytos išskirtys, taikant regresijos metodą, nenaudojant svorių, kai P = 1%

		Regresijos metodas									
		Visas periodas				2006 m. Kovo mėnuo					
		„Geri“ duomenys		Nustatytos išskirtys		„Geri“ duomenys		Nustatytos išskirtys		Pasiteisinusios išskirtys	
		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*	
		A	R	A	R	A	R	A	R	A	R
Praktikoje patikrinti duomenys	„Geri“ duomenys	38256	448	11	0	5078	35	3	0	1	0
	Išskirtys	7	0	54	0	3	0	1	0	1	0

* Čia ir toliau indikatorius R – reiškia, kad nustatant išsiskiriančias reikšmes šioms laiko eilutėms buvo rekomenduojama taikyti regresijos metodą, A – kad buvo rekomenduojama taikyti alternatyvųjį metodą.

Nustatytos išskirtys, taikant regresijos metodą, nenaudojant svorių, kai P = 5 %

		Regresijos metodas									
		Visas periodas				2006 m. Kovo mėnuo					
		„Geri“ duomenys		Nustatytos išskirtys		„Geri“ duomenys		Nustatytos išskirtys		Pasiteisinusios išskirtys	
		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*		Indikatorius*	
		A	R	A	R	A	R	A	R	A	R
Praktikoje patikrinti duomenys	„Geri“ duomenys	38208	448	59	0	5063	35	16	0	2	0
	Išskirtys	2	0	59	0	0	0	4	0	4	0

Iš aukščiau pateiktų lentelių matosi, jog alternatyvaus metodo rezultatai yra žymiai geresni, kai neatsižvelgiama į nustatytus svorius. Panašias išvadas galima padaryti ir taikant regresijos metodą. Nustatant tikras išsiskiriančias reikšmes, nenaudojant svorių efektyvumas šiek tiek sumažėja, tačiau atrandama mažiau papildomų išskirčių, kurios vėliau nepasiteisina.

Nepaisant to, kad netenkinamos MKM prielaidos, regresijos metodo efektyvumas, nustatant išsiskiriančias reikšmes yra didesnis, nei alternatyvaus metodo. Pabandžius susiaurinti prognozės intervalo ribas, nustatant P = 5%, taikant regresijos metodą, gauti dar geresni rezultatai – tikrų nustatytų išsiskiriančių reikšmių skaičius padidėjo, o papildomų nustatytų išskirčių skaičius vis dar nėra labai didelis. Nepaisant to, jog 2006 m. trečio mėnesio duomenyse iš visų rastų regresijos metodo išskirčių pasiteisino šiek tiek mažiau nei trečdalis, vis dar negalima daryti išvadų, jog dauguma nustatomų išskirčių nepasiteisina. Remiantis istoriniais duomenimis, nustatyta, kad iš viso periodo regresijos metodo rastų išskirčių, pasiteisino 50,85%. Tačiau likusiųjų 41,15% nustatytų išskirčių vis dar negalima vadinti „gerais“ duomenimis. Tam, kad sužinoti, kokia dalis iš šių duomenų buvo tikrai geri duomenys, reikalinga papildoma analizė. Nustačius, kad iš tiesų pernelyg didelė dalis nustatomų išskirčių nepasiteisina, galima siaurinti prognozės intervalą.

Taip pat iš rezultatų matome, jog labai mažai išskirčių randama tose laiko eilutėse, kurios tenkino visas MKM prielaidas. Analizuojant duomenis buvo pastebėta, kad daugiausiai išskirčių randama laiko eilutėse su didele variacija. Faktiškai per visą laikotarpį nagrinėjamoje KN dalyje, išskirtys buvo randamos tik šešiuose koduose. Tai, žinoma, galima paaiškinti ir tuo, kad tuose koduose buvo labiausiai tikimasi rasti išskirčių ir juose jų atidžiai ieškoma. Naudojant alternatyvųjį ir regresijos metodus buvo rastos išskirtys ir kitose laiko eilutėse, kuriose jų nebuvo randama anksčiau. Gauti rezultatai pateikiami lentelėje:

Laiko eilutės, kuriose buvo nustatytos išskirtys

	Regresijos metodas		Alternatyvus metodas	
	eilučių skaičius	dalis (%)	eilučių skaičius	dalis (%)
Viso	***39***	***100***	***76***	***100***
Laiko eilutės, iš kurių buvo pašalinti stebėjimai dėl pernelyg didelės variacijos	29	*74,36*	70	*92,11*
Laiko eilutės, kurios visų „filtracijos“ žingsnių metu turėjo stebėjimų su nuliniais svoriais	28	*71,79*	66	*86,84*
Kitos	10	*25,64*	6	*7,89*

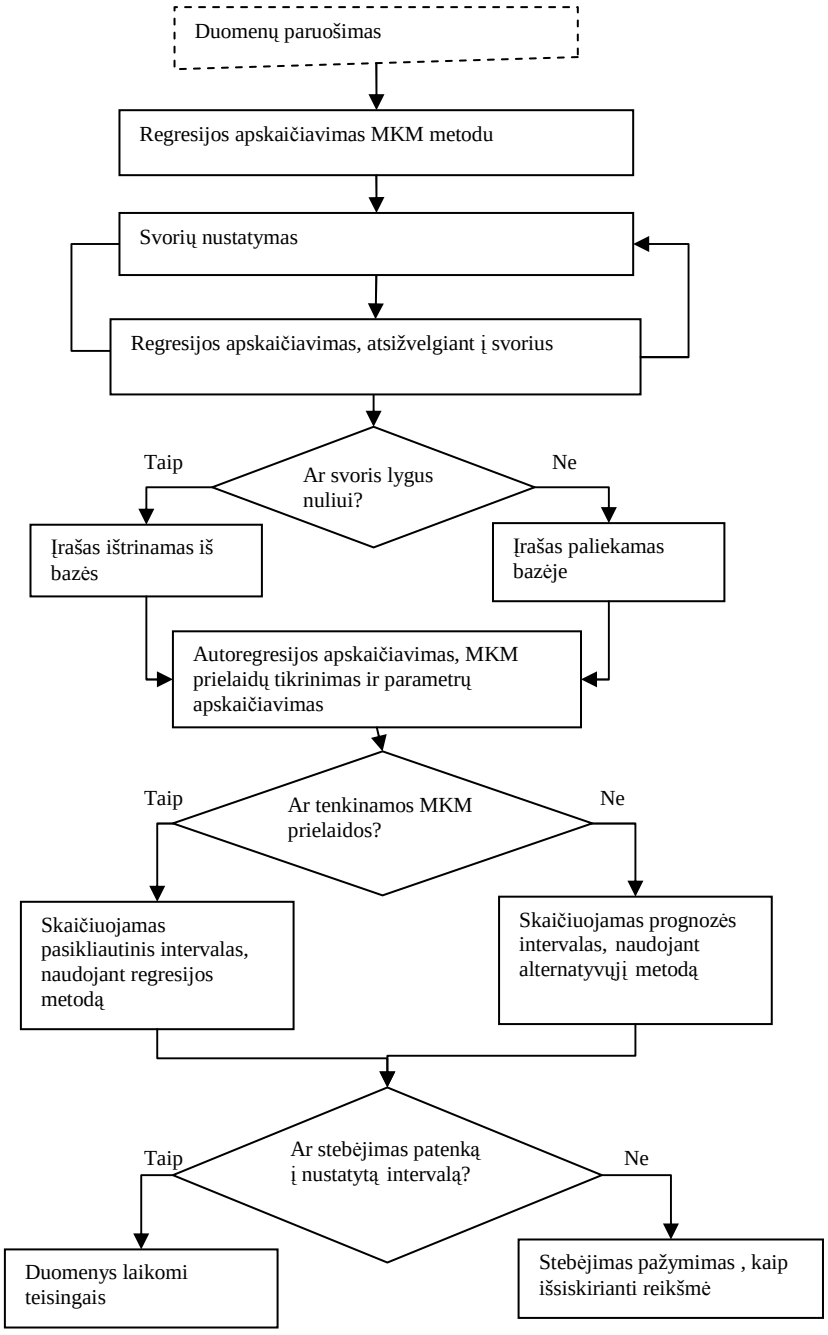
Išvados

Iš aukščiau pateiktos analizės matome, jog siūlomos išsiskiriančių reikšmių nustatymo metodikos supaprastinimas tik pagerintų gaunamus rezultatus.
Suformuluosiu pagrindines išvadas:

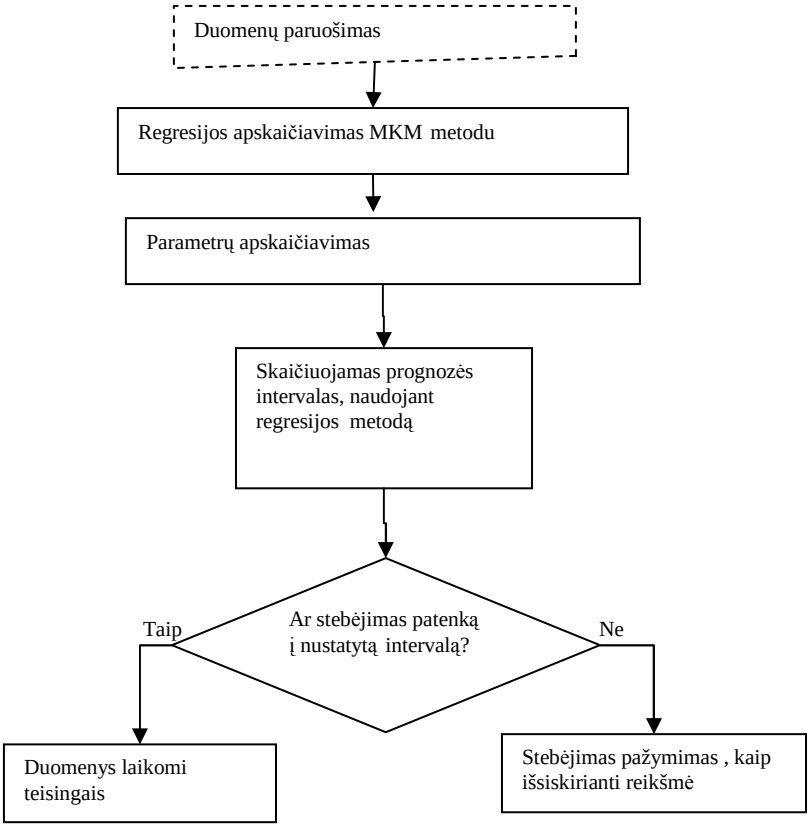
- Išsiskiriančios reikšmės yra nustatomos geriau, kai nenaudojami svoriai.
- Modelio pataisymas autokoreliacijos atžvilgiu neįtakoja rezultatų ir be to neturi prasmės.
- Laiko eilutėse, kurios tenkina MKM prielaidas beveik nerandama išsiskiriančių reikšmių.
- MKM prielaidų nepaisymas neįtakoja rezultatų, taikant regresijos metodą.
- Regresijos metodas yra geresnis už alternatyvųjį.
- Didžiausia dalis išsiskiriančių reikšmių randama laiko eilutėse su didele variacija, tačiau ne visos nustatytos išskirtys pasiteisina, kadangi nors ir daroma prielaida apie laiko eilučių homogeniškumą, tai neatitinka realybės.

Remiantis šiomis išvadomis siūlau pakeisti išsiskiriančių reikšmių nustatymo pradinę blokinę schemą Nr. 1 į schemą Nr. 2.

Schema Nr. 1


Duomenų paruošimas
Regresijos apskaičiavimas MKM metodu
Svorių nustatymas
Regresijos apskaičiavimas, atsižvelgiant į svorius
Taip
Ar svoris lygus nuliui?
Ne
Įrašas ištrinamas iš bazės
Įrašas paliekamas bazėje
Autoregresijos apskaičiavimas, MKM prielaidų tikrinimas ir parametrų apskaičiavimas
Taip
Ar tenkinamos MKM prielaidos?
Ne
Skaičiuojamas pasikliautinis intervalas, naudojant regresijos metodą
Skaičiuojamas prognozės intervalas, naudojant alternatyvųjį metodą
Taip
Ar stebėjimas patenką į nustatytą intervalą?
Ne
Duomenys laikomi teisingais
Stebėjimas pažymimas , kaip išsiskirianti reikšmė

Schema Nr. 2


Duomenų paruošimas
Regresijos apskaičiavimas MKM metodu
Parametrų apskaičiavimas
Skaičiuojamas prognozės intervalas, naudojant regresijos metodą
Ar stebėjimas patenką į nustatytą intervalą?
Taip
Ne
Duomenys laikomi teisingais
Stebėjimas pažymimas , kaip išsiskirianti reikšmė

Summary

Most of the users of foreign trade statistics have high requirements concerning data quality and speediness of deliver. Unfortunately these two requirements often conflict and it is impossible to check every observation. That‘s why there are number of quality controls of the data implemented concerning credibility and validity checks in order to focus on the most outlying observation and revice them.
The objective of this work was to test the outliers’ detection metodology which was developed by Eurostat, to analize rezults and perhaps to turn it a new control in the Foreign Trade Statistics Division of Statistics Lithuania.
The initial idea of Eurostat methodology to detect ouliers was to calculate a regression line by using the ordinary least squares (OLS) method to calculate by inference a prediction interval and to flag all observations that lie outside this interval. As the OLS method requires respecting some assumtions if these assumptions were violated, the OLS method was violated it was proposed to switch to an alternative method which was applied to those time series for which the OLS method was not possible. Before calculating the upper and lower limits for detecting outliers it was proposed to use a weighted least squares regression as well in order to get rid of the possible outlying observations that could appear in the data. The step with recalculating the weights and the regression model had to be repeated three times. Afterwards it was proposed to use an autoregressive error model in order to avoid the problem of autocorelation as the data are time series.
After implementing the methodology described above and analysing the results the following conclusions were made:

- The detection of outliers is more effective when the weights are not used;
- Correction of the the regression line concerning autocorrelation does not impact the results and is meanless;
- The most of the outliers are found in the time series that violate the OLS method assumptions and unrespecting OLS assumptions still does not impact efficiency of results. So regression methodology can be used even for those time series that violate the assumptions;
- Regression methodology performs more effectively than the alternative one.
- The majority of outliers are found in the time series with big variation. However not all of the detected „outliers” are real ones. It happens because the assumption of the homogenity of CN8 code is only theoretical and it does not stand in the practice.

Concerning the conclusions described above we propose to change the structure of procedure of detection of outliers.

LITERATŪRA:

1. EDICOM 2002, Action 2b(3) – Reduction of outliers, Final Report
2. **J. R. Magnus, P.K. Katyšev, A. A. Pereseckij** Ekonometrika. V.: Delo, 2000 m.
3. **J. Kruopis** Matematinė statistika. V.: Mokslo ir enciklopedijų leidykla, 1993 m.
4. SAS/ETS® 9.1 User‘s Guide. SAS Institute Inc., 2004 m.
5. SAS/STAT® 9.1 User‘s Guide. SAS Institute Inc., 2004 m.
6. Base SAS® 9.1 Procedures Guide. SAS Institute Inc., 2004 m.
7. SAS ® 9.1 SQL Procedure User‘s guide. SAS Institute Inc., 2004 m.
8. **V. Čekanavičius, G. Murauskas** Statistika ir jos taikymai, II dalis, TEV, 2004 m.

PRIEDAS

Čia pateikiu SAS programą, su kuria buvo atlikti skaičiavimai.

Duomenų paruošimas:

```
LIBNAME OUTLIER_ ('C:\Documents and Settings\OUTLIER\DUOMENYS\05_11');
RUN;
/* GERA IMTIS */
DATA OUTLIER_.EXP_01_24;
SET OUTLIER_.EXP_01_24;
VERTLIT = VERTLIT*1000;
run;
PROC SQL;
alter table OUTLIER_.EXP_01_24
 add X num format= best12.;
alter table OUTLIER_.EXP_01_24
 add Y num format= best12.;
alter table OUTLIER_.EXP_01_24
 add MODEL char format=$32.;
CREATE TABLE OUTLIER_.EXP_01_24_M3
 AS SELECT * FROM OUTLIER_.EXP_01_24;
CREATE TABLE OUTLIER_.EXP_01_24_M5
 AS SELECT * FROM OUTLIER_.EXP_01_24;
quit;
data OUTLIER_.EXP_01_24;
set OUTLIER_.EXP_01_24;
Model = 'MODEL1';
Y = VERTLIT;
X = NETTO;
RUN;
data OUTLIER_.EXP_01_24_M3;
set OUTLIER_.EXP_01_24_M3;
Model = 'MODEL3';
Y = VERTLIT;
X = KIEKIS;
RUN;
data OUTLIER_.EXP_01_24_M5;
set OUTLIER_.EXP_01_24_M5;
Model = 'MODEL5';
Y = NETTO;
X = KIEKIS;
RUN;
PROC SQL;
INSERT INTO OUTLIER_.EXP_01_24
    SELECT *
    FROM OUTLIER_.EXP_01_24_M3;
INSERT INTO OUTLIER_.EXP_01_24
    SELECT *
    FROM OUTLIER_.EXP_01_24_M5;
DROP TABLE OUTLIER_.EXP_01_24_M3;
DROP TABLE OUTLIER_.EXP_01_24_M5;

CREATE TABLE OUTLIER_.EXP_01_08 AS
  SELECT DKLR_NR, PRKOD, MODEL, PRNR, X, Y, MET, MENUO
   FROM OUTLIER_.EXP_01_24;
    QUIT;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08
 ADD SU CHAR FORMAT = $32.;
UPDATE OUTLIER_.EXP_01_08 AS A
 SET SU = (SELECT SU FROM OUTLIER_.MATAVIMAS AS B
           WHERE B.CN8 = A.PRKOD AND A.MET = B.YEAR);
DELETE FROM OUTLIER_.EXP_01_08
  WHERE SU = '-' AND (MODEL = 'MODEL3' OR MODEL = 'MODEL5');
ALTER TABLE OUTLIER_.EXP_01_08
  DROP SU;
create table outlier_.skirtumas
```

```
 as select dklr_nr, prkod, model, x, y, prnr, MET, MENUO
  from outlier_.exp_01_08;

CREATE TABLE OUTLIER_.EXP_01_08_KA
    AS SELECT DKLR_NR, PRKOD, MODEL, PRNR, X, Y, MET, MENUO
    FROM OUTLIER_.EXP_01_08
    ORDER BY PRKOD, MODEL;

UPDATE OUTLIER_.EXP_01_08_KA
    SET X = LOG(X);
  UPDATE OUTLIER_.EXP_01_08_KA
    SET Y = LOG(Y);

 QUIT;
data outlier_.exp_01_08_ka;
set outlier_.exp_01_08_ka;
retain rakt(0);
rakt = rakt+1;
run;
proc sql;
 CREATE TABLE OUTLIER_.EXP_01_08_K
    AS SELECT RAKT, PRKOD, MODEL, PRNR, X, Y, MET, MENUO
    FROM outlier_.exp_01_08_KA
  ORDER BY PRKOD, MODEL;

create table outlier_.exp_01_08_RAKTAS
    AS SELECT RAKT, PRKOD, MODEL, PRNR, DKLR_NR FROM OUTLIER_.EXP_01_08_KA;
DROP TABLE OUTLIER_.EXP_01_08_KA;
QUIT;
DATA OUTLIER_.EXP_01_08_K;
SET OUTLIER_.EXP_01_08_K;
DKLR_NR = RAKT;
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_K
DROP RAKT;
QUIT;
PROC SQL;
alter table OUTLIER_.EXP_01_24B
 add X num format= best12.;
alter table OUTLIER_.EXP_01_24B
 add Y num format= best12.;
alter table OUTLIER_.EXP_01_24B
 add MODEL char format=$32.;
CREATE TABLE OUTLIER_.EXP_01_24_M3
 AS SELECT * FROM OUTLIER_.EXP_01_24B;
CREATE TABLE OUTLIER_.EXP_01_24_M5
 AS SELECT * FROM OUTLIER_.EXP_01_24B;
quit;


data OUTLIER_.EXP_01_24B;
set OUTLIER_.EXP_01_24B;
Model = 'MODEL1';
Y = VERTLIT;
X = NETTO;
RUN;
data OUTLIER_.EXP_01_24_M3;
set OUTLIER_.EXP_01_24_M3;
Model = 'MODEL3';
Y = VERTLIT;
X = KIEKIS;
RUN;
data OUTLIER_.EXP_01_24_M5;
set OUTLIER_.EXP_01_24_M5;
Model = 'MODEL5';
Y = NETTO;
X = KIEKIS;
RUN;
PROC SQL;
```

```
INSERT INTO OUTLIER_.EXP_01_24B
    SELECT *
    FROM OUTLIER_.EXP_01_24_M3;
INSERT INTO OUTLIER_.EXP_01_24B
    SELECT *
    FROM OUTLIER_.EXP_01_24_M5;
DROP TABLE OUTLIER_.EXP_01_24_M3;
DROP TABLE OUTLIER_.EXP_01_24_M5;

CREATE TABLE OUTLIER_.EXP_01_08B AS
  SELECT DKLR_NR, PRKOD, PRNR, MODEL, X, Y, MET, MENUO
   FROM OUTLIER_.EXP_01_24B;
QUIT;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08B
 ADD SU CHAR FORMAT = $32.;
UPDATE OUTLIER_.EXP_01_08B AS A
 SET SU = (SELECT SU FROM OUTLIER_.MATAVIMAS AS B
           WHERE B.CN8 = A.PRKOD AND A.MET = B.YEAR);
DELETE FROM OUTLIER_.EXP_01_08B
  WHERE SU = '-' AND (MODEL = 'MODEL3' /*OR MODEL = 'MODEL4' */OR MODEL = 'MODEL5');
ALTER TABLE OUTLIER_.EXP_01_08B
  DROP SU;
CREATE TABLE OUTLIER_.EXP_01_08_KA
    AS SELECT DKLR_NR, PRKOD, MODEL, PRNR, X, Y, MET, MENUO
    FROM OUTLIER_.EXP_01_08B
    ORDER BY PRKOD, MODEL;
/*JEIGU "BLOGA" IR "GERA" IMTIS PALYGINAMOS*/
ALTER TABLE OUTLIER_.EXP_01_08_KA
 ADD RAKT NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_KA AS A
 SET RAKT = (SELECT RAKT FROM OUTLIER_.EXP_01_08_RAKTAS B
   WHERE A.PRKOD = B.PRKOD AND A.PRNR = B.PRNR AND
                  A.DKLR_NR = B.DKLR_NR AND A.MODEL = B.MODEL);
UPDATE OUTLIER_.EXP_01_08_KA
    SET X = LOG(X);
UPDATE OUTLIER_.EXP_01_08_KA
    SET Y = LOG(Y);
 QUIT;

proc sql;
 CREATE TABLE OUTLIER_.EXP_01_08_KB
    AS SELECT RAKT, PRKOD, MODEL, PRNR, X, Y, MET, MENUO
    FROM outlier_.exp_01_08_KA
  ORDER BY PRKOD, MODEL;


DROP TABLE OUTLIER_.EXP_01_08_KA;
QUIT;
DATA OUTLIER_.EXP_01_08_KB;
SET OUTLIER_.EXP_01_08_KB;
DKLR_NR = RAKT;
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_KB
DROP RAKT;
QUIT;
/***************istrinu_nereikalinga_menesi is geru!!!!!!!!!!!!!!!!*****************/
proc sql;
delete from outlier_.exp_01_08_k
 where (MET = 2006 AND MENUO = '03');

delete from outlier_.exp_01_08
 where (MET = 2006 AND MENUO = '03');
quit;
```

Skaičiavimai

```
/* 1 ZINGSNIS (DARYTI IKI ANTRO ZINGSNIO, PO TO UZKOMENTUOTI)*/
PROC MEANS DATA = OUTLIER_.EXP_01_08_K MEAN STD MAX MIN;
 BY PRKOD MODEL;
 VAR X Y;
 OUTPUT OUT = OUTLIER_.EXP_01_08_MEANS;
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_MEANS
    DROP _TYPE_;
DELETE
FROM OUTLIER_.EXP_01_08_MEANS
    WHERE _STAT_ = 'N';
ALTER TABLE OUTLIER_.EXP_01_08_K
ADD N NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_K AS A
 SET N = (SELECT _FREQ_ FROM OUTLIER_.EXP_01_08_MEANS AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND _STAT_ = 'MIN');
DELETE FROM OUTLIER_.EXP_01_08_K
 WHERE N < 30;
ALTER TABLE OUTLIER_.EXP_01_08_K
 DROP N;

ALTER TABLE OUTLIER_.EXP_01_08
ADD N NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08 AS A
 SET N = (SELECT _FREQ_ FROM OUTLIER_.EXP_01_08_MEANS AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND _STAT_ = 'MIN');

create table outlier_.smulkus
 as select prkod, model, sum(y), N from outlier_.exp_01_08 where
   N<30
group by prkod, model;

DELETE FROM OUTLIER_.EXP_01_08
 WHERE N < 30;
ALTER TABLE OUTLIER_.EXP_01_08
 DROP N;

QUIT;
/*PASKAICIOJAM CHARAKTERISTIKAS "BLOGAI" IMCIAI*/
PROC MEANS DATA = OUTLIER_.EXP_01_08_KB MEAN STD MAX MIN;
 BY PRKOD MODEL;
 VAR X Y;
 OUTPUT OUT = OUTLIER_.EXP_01_08_MEANSB;
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_MEANSB
    DROP _TYPE_;
  DELETE
    FROM OUTLIER_.EXP_01_08_MEANSB
    WHERE _STAT_ = 'N';
ALTER TABLE OUTLIER_.EXP_01_08_KB
ADD N NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_KB AS A
 SET N = (SELECT _FREQ_ FROM OUTLIER_.EXP_01_08_MEANS AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND _STAT_ = 'MIN');

DELETE FROM OUTLIER_.EXP_01_08_KB
 WHERE N < 30;
ALTER TABLE OUTLIER_.EXP_01_08_KB
 DROP N;

 ALTER TABLE OUTLIER_.EXP_01_08B
ADD N NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08B AS A
 SET N = (SELECT _FREQ_ FROM OUTLIER_.EXP_01_08_MEANS AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND _STAT_ = 'MIN');
DELETE FROM OUTLIER_.EXP_01_08B
```

```
 WHERE N < 30;
ALTER TABLE OUTLIER_.EXP_01_08B
 DROP N;
QUIT;
/*LEIDZIAM REGRESIJA "GERAI" IMCIAI*/
PROC REG DATA = OUTLIER_.EXP_01_08_K OUTEST = OUTLIER_.EXP_01_08_STEP1;
  BY PRKOD MODEL;
  MODEL Y = X;
  output out = outlier_.exp_01_08_skirtumas1
        p = y_iv r = d;
 RUN;
/*sutvarkome lentele su parametrais*/
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_STEP1
 DROP _MODEL_, _TYPE_, _DEPVAR_, Y ;
QUIT;
PROC SORT DATA = OUTLIER_.EXP_01_08_STEP1;
  BY PRKOD MODEL;
RUN;
/*2 ZINGSNIS*/
PROC UNIVARIATE DATA = OUTLIER_.EXP_01_08_SKIRTUMAS1;
  BY PRKOD MODEL;
  VAR D;
  OUTPUT OUT = OUTLIER_.EXP_01_08_QUART1 q1=Q1_D q3=q3_D;
RUN;
PROC SQL;
  CREATE TABLE OUTLIER_.EXP_01_08_SVORIAI1
    AS SELECT DKLR_NR, A.PRKOD, A.MODEL, PRNR, X, Y, D, Q1_D, Q3_D
    FROM OUTLIER_.EXP_01_08_SKIRTUMAS1 A, OUTLIER_.EXP_01_08_QUART1 B
      WHERE A.PRKOD = B.PRKOD AND A.MODEL=B.MODEL;

  DROP TABLE OUTLIER_.EXP_01_08_QUART1;
quit;
data OUTLIER_.EXP_01_08_SVORIAI1;
 set OUTLIER_.EXP_01_08_SVORIAI1;
   Z = D/(3*(Q3_D - Q1_D));
 if abs(z) <=1 then W = (1-Z*Z)*(1-Z*Z);
  else W = 0;
run;
proc sql;
CREATE TABLE OUTLIER_.EXP_01_08_DROP1
    AS SELECT *  FROM OUTLIER_.EXP_01_08_SVORIAI1
            WHERE W =0 OR W IS NULL;
delete from outlier_.exp_01_08_SVORIAI1
   WHERE W IS NULL;
QUIT;
/***<><><><><><><><><><><><><>2 FILTRAS<><><><><><><><><><><><><>******/
PROC SORT DATA = OUTLIER_.EXP_01_08_SVORIAI1;
BY PRKOD MODEL;
RUN;
PROC REG DATA = OUTLIER_.EXP_01_08_SVORIAI1;
 MODEL Y = X;
 BY PRKOD MODEL;
 WEIGHT W;
 OUTPUT OUT = OUTLIER_.EXP_01_08_SKIRTUMAS2 P = y_iv R = D2;
 RUN;
 /*2 ZINGSNIS*/
PROC UNIVARIATE DATA = OUTLIER_.EXP_01_08_SKIRTUMAS2;
  BY PRKOD MODEL;
  VAR D2;
OUTPUT OUT = OUTLIER_.EXP_01_08_QUART2 q1=Q1_D q3=q3_D;
RUN;
PROC SQL;
  CREATE TABLE OUTLIER_.EXP_01_08_SVORIAI2
    AS SELECT DKLR_NR, A.PRKOD, A.MODEL, PRNR, X, Y, D2, B.Q1_D, B.Q3_D
    FROM OUTLIER_.EXP_01_08_SKIRTUMAS2 A, OUTLIER_.EXP_01_08_QUART2 B
      WHERE A.PRKOD = B.PRKOD AND A.MODEL=B.MODEL;

  DROP TABLE OUTLIER_.EXP_01_08_QUART2;
quit;
```

```
data OUTLIER_.EXP_01_08_SVORIAI2;
 set OUTLIER_.EXP_01_08_SVORIAI2;
   Z = D2/(3*(Q3_D - Q1_D));
 if abs(z) <=1 then W = (1-Z*Z)*(1-Z*Z);
  else W = 0;
run;
proc sql;
CREATE TABLE OUTLIER_.EXP_01_08_DROP2
    AS SELECT *  FROM OUTLIER_.EXP_01_08_SVORIAI2
            WHERE W =0 OR W IS NULL;
delete from outlier_.exp_01_08_SVORIAI2
   WHERE W IS NULL;

CREATE TABLE OUTLIER_.EXP_01_08_DROP1_2
 AS SELECT A.DKLR_NR, A.PRKOD, B.MODEL, B.X, B.Y
  FROM OUTLIER_.EXP_01_08_DROP1 A, OUTLIER_.EXP_01_08_DROP2 B
   WHERE A.DKLR_NR =B.DKLR_NR;
QUIT;
/***<><><><><><><><><><><><>3 FILTRAS<><><><><><><><><><><><>******/
PROC SORT DATA = OUTLIER_.EXP_01_08_SVORIAI2;
BY PRKOD MODEL;
RUN;
PROC REG DATA = OUTLIER_.EXP_01_08_SVORIAI2;
 MODEL Y = X;
 BY PRKOD MODEL;
 WEIGHT W;
 OUTPUT OUT = OUTLIER_.EXP_01_08_SKIRTUMAS3 P = y_iv R = D3;
 RUN;
 /*2 ZINGSNIS*/
PROC UNIVARIATE DATA = OUTLIER_.EXP_01_08_SKIRTUMAS3;
  BY PRKOD MODEL;
  VAR D3;
OUTPUT OUT = OUTLIER_.EXP_01_08_QUART3 q1=Q1_D q3=q3_D;
RUN;
PROC SQL;
  CREATE TABLE OUTLIER_.EXP_01_08_SVORIAI3
    AS SELECT DKLR_NR, A.PRKOD, A.MODEL, PRNR, X, Y, D3, B.Q1_D, B.Q3_D
    FROM OUTLIER_.EXP_01_08_SKIRTUMAS3 A, OUTLIER_.EXP_01_08_QUART3 B
      WHERE A.PRKOD = B.PRKOD AND A.MODEL=B.MODEL;

  DROP TABLE OUTLIER_.EXP_01_08_QUART3;
quit;
data OUTLIER_.EXP_01_08_SVORIAI3;
 set OUTLIER_.EXP_01_08_SVORIAI3;
   Z = D3/(3*(Q3_D - Q1_D));
 if abs(z) <=1 then W = (1-Z*Z)*(1-Z*Z);
  else W = 0;
run;
proc sql;
CREATE TABLE OUTLIER_.EXP_01_08_DROP3
    AS SELECT *  FROM OUTLIER_.EXP_01_08_SVORIAI3
            WHERE W =0 OR W IS NULL;
delete from outlier_.exp_01_08_SVORIAI3
   WHERE W IS NULL;

CREATE TABLE OUTLIER_.EXP_01_08_DROP2_3
 AS SELECT A.DKLR_NR, A.PRKOD, B.MODEL, B.X, B.Y
  FROM OUTLIER_.EXP_01_08_DROP2 A, OUTLIER_.EXP_01_08_DROP3 B
   WHERE A.DKLR_NR =B.DKLR_NR;
QUIT;
/***<><><><><><><><><><><><>4 FILTRAS<><><><><><><><><><><><>******/
PROC SORT DATA = OUTLIER_.EXP_01_08_SVORIAI3;
BY PRKOD MODEL;
RUN;
PROC REG DATA = OUTLIER_.EXP_01_08_SVORIAI3 OUTEST=OUTLIER_.EXP_01_08_STEP4;
 MODEL Y = X/EDF SPEC;
 BY PRKOD MODEL;
 WEIGHT W;
 OUTPUT OUT = OUTLIER_.EXP_01_08_SKIRTUMAS4 P = y_iv R = D4;
 ODS OUTPUT SPECTEST = OUTLIER_.REGWHITE;
```

```
 RUN;
 PROC SQL;
  ALTER TABLE OUTLIER_.EXP_01_08_STEP4
   DROP _MODEL_,_DEPVAR_,_TYPE_,Y,_IN_,_P_,_EDF_;
  ALTER TABLE OUTLIER_.REGWHITE
   DROP _MODEL_,DEPENDENT,DF;
  QUIT;
 /*2 ZINGSNIS*/
PROC UNIVARIATE DATA = OUTLIER_.EXP_01_08_SKIRTUMAS4;
  BY PRKOD MODEL;
  VAR D4;
OUTPUT OUT = OUTLIER_.EXP_01_08_QUART4 q1=Q1_D q3=q3_D;
RUN;
PROC SQL;
  CREATE TABLE OUTLIER_.EXP_01_08_SVORIAI4
    AS SELECT DKLR_NR, A.PRKOD, A.MODEL, PRNR, X, Y, D4, B.Q1_D, B.Q3_D
    FROM OUTLIER_.EXP_01_08_SKIRTUMAS4 A, OUTLIER_.EXP_01_08_QUART4 B
      WHERE A.PRKOD = B.PRKOD AND A.MODEL=B.MODEL;

  DROP TABLE OUTLIER_.EXP_01_08_QUART4;
quit;
data OUTLIER_.EXP_01_08_SVORIAI4;
 set OUTLIER_.EXP_01_08_SVORIAI4;
   Z = D4/(3*(Q3_D - Q1_D));
 if abs(z) <=1 then W = (1-Z*Z)*(1-Z*Z);
  else W = 0;
run;
proc sql;
CREATE TABLE OUTLIER_.EXP_01_08_DROP4
    AS SELECT *  FROM OUTLIER_.EXP_01_08_SVORIAI4
            WHERE W =0 OR W IS NULL;
delete from outlier_.exp_01_08_SVORIAI4
   WHERE W IS NULL OR W=0;

CREATE TABLE OUTLIER_.EXP_01_08_DROP3_4
 AS SELECT A.DKLR_NR, A.PRKOD, B.MODEL, B.X, B.Y
  FROM OUTLIER_.EXP_01_08_DROP3 A, OUTLIER_.EXP_01_08_DROP4 B
   WHERE A.DKLR_NR =B.DKLR_NR;
QUIT;
```

Modelio pataisymas autokoreliacijos atžvilgiu ir testų tikrinimas

```
PROC SORT DATA = OUTLIER_.EXP_01_08_SVORIAI1;
BY PRKOD MODEL;
RUN;
PROC AUTOREG DATA = OUTLIER_.EXP_01_08_SVORIAI1 OUTEST =
OUTLIER_.EXP_01_08_STEP4_AUTOREG;
  BY PRKOD MODEL;
  MODEL Y = X/BACKSTEP DWPROB NORMAL LAGDEP;
  OUTPUT OUT = OUTLIER_.EXP_01_08_SKIRTUMAS_AUTOREG P = y_iv R = D_A
                                      UCL = UCL LCL = LCL ALPHACLI = .01;
 ODS OUTPUT FITSUMMARY = OUTLIER_.FITSUMMARY_AUTOREG;
 RUN;
PROC SQL;
  ALTER TABLE OUTLIER_.EXP_01_08_STEP4_AUTOREG
     DROP _MODEL_,_DEPVAR_, _TYPE_ , _STATUS_, _METHOD_, _NAME_, Y;
ALTER TABLE OUTLIER_.FITSUMMARY_AUTOREG
  DROP CVALUE1, CVALUE2;
DELETE FROM OUTLIER_.FITSUMMARY_AUTOREG
 WHERE LABEL1 = 'SSE' OR LABEL1 = 'MSE' OR LABEL1 = 'SBC';
QUIT;
PROC SORT DATA = OUTLIER_.EXP_01_08_STEP4_AUTOREG;
  BY PRKOD MODEL;
RUN;
PROC SORT DATA = OUTLIER_.FITSUMMARY_AUTOREG;
 BY PRKOD MODEL;
RUN;
```

Parametrų apskaičiavimas

```
PROC MEANS DATA = OUTLIER_.EXP_01_08_SVORIAI1 MEAN STD MAX MIN;
 BY PRKOD MODEL;
 VAR X Y;
 OUTPUT OUT = OUTLIER_.EXP_01_08_MEANS4;
 RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_MEANS4
    DROP _TYPE_;
DELETE
FROM OUTLIER_.EXP_01_08_MEANS4
    WHERE _STAT_ = 'N';
QUIT;
PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_PARAMETRAI
 AS SELECT a.PRKOD, a.MODEL, B._FREQ_
 FROM OUTLIER_.EXP_01_08_STEP4_AUTOREG a,OUTLIER_.EXP_01_08_MEANS4 B
   WHERE a.PRKOD = b.PRKOD AND A.MODEL = B.MODEL AND B._STAT_ = 'MIN';

ALTER TABLE OUTLIER_.EXP_01_08_PARAMETRAI
 ADD N NUM FORMAT = BEST12.;
alter table OUTLIER_.EXP_01_08_PARAMETRAI
  add SY num label = 'SY (STANDART DEVIATION)' format=BEST12.;
alter table OUTLIER_.EXP_01_08_PARAMETRAI
  ADD XMOY NUM LABEL = 'AVERAGE_X' FORMAT = BEST12.;
 alter table OUTLIER_.EXP_01_08_PARAMETRAI
  ADD SSD NUM LABEL = 'SUM OF SQUARED DEVIATION (X)' FORMAT = BEST12.;
 alter table OUTLIER_.EXP_01_08_PARAMETRAI
  ADD STD NUM LABEL = 'STANDART DEVIATION (X)' FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_PARAMETRAI AS A
  SET STD = (SELECT X FROM OUTLIER_.EXP_01_08_MEANS4 AS B
             WHERE B._STAT_='STD' AND B.PRKOD = A.PRKOD AND A.MODEL = B.MODEL);
ALTER TABLE OUTLIER_.EXP_01_08_PARAMETRAI
  ADD RSQR NUM LABEL = 'DETERMINATION COEFFICIENT' FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_PARAMETRAI AS A
 SET N = (SELECT _FREQ_ FROM OUTLIER_.EXP_01_08_MEANS4 AS B
    WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND B._STAT_ = 'MIN');
UPDATE OUTLIER_.EXP_01_08_PARAMETRAI AS A
  SET SY = (SELECT Y FROM OUTLIER_.EXP_01_08_MEANS4 AS B
             WHERE B._STAT_='STD' AND B.PRKOD = A.PRKOD AND A.MODEL = B.MODEL);
UPDATE OUTLIER_.EXP_01_08_PARAMETRAI AS A
  SET XMOY = (SELECT X FROM OUTLIER_.EXP_01_08_MEANS4 AS B
              WHERE B._STAT_='MEAN' AND B.PRKOD = A.PRKOD AND A.MODEL = B.MODEL);
UPDATE OUTLIER_.EXP_01_08_PARAMETRAI AS A
  SET RSQR = (SELECT NVALUE1 FROM OUTLIER_.FITSUMMARY_AUTOREG AS B
      WHERE B.PRKOD = A.PRKOD AND B.MODEL = A.MODEL AND LABEL1 = 'Regress R-Square');
QUIT;
DATA OUTLIER_.EXP_01_08_PARAMETRAI;
SET OUTLIER_.EXP_01_08_PARAMETRAI;
 SSD = STD*STD*(_FREQ_-1);
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_PARAMETRAI
 DROP STD;
QUIT;


PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_RESULTS AS SELECT *
 FROM OUTLIER_.EXP_01_08_PARAMETRAI;
ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 DROP _FREQ_;

ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 ADD AUTOCORR1 NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RESULTS AS A
 SET AUTOCORR1 = (SELECT NVALUE2 FROM OUTLIER_.FITSUMMARY_AUTOREG AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND LABEL1 = "Durbin's t");
```

```
ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 ADD NON_NORMAL1 NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RESULTS AS A
 SET NON_NORMAL1 = (SELECT NVALUE2 FROM OUTLIER_.FITSUMMARY_AUTOREG AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL AND LABEL1 = 'Normal Test');
ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 ADD HETERO1 NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RESULTS AS A
 SET HETERO1 = (SELECT PROBCHISQ FROM OUTLIER_.REGWHITE AS B
   WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL);

ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 ADD A NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RESULTS AS A
  SET A = (SELECT INTERCEPT FROM OUTLIER_.EXP_01_08_STEP4_AUTOREG AS B
    WHERE B.PRKOD = A.PRKOD AND B.MODEL=A.MODEL);

ALTER TABLE OUTLIER_.EXP_01_08_RESULTS
 ADD B NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RESULTS AS D
    SET B = (SELECT X FROM OUTLIER_.EXP_01_08_STEP4_AUTOREG AS E
               WHERE D.PRKOD = E.PRKOD AND D.MODEL=E.MODEL);
QUIT;
DATA OUTLIER_.EXP_01_08_RESULTS;
 SET OUTLIER_.EXP_01_08_RESULTS;
 IF AUTOCORR1 < 0.01 THEN AUTOCORR = 'Y';
  ELSE AUTOCORR = 'N';
 IF HETERO1 < 0.01 THEN HETERO = 'Y';
  ELSE HETERO = 'N';
 IF NON_NORMAL1 < 0.01 THEN NON_NORMAL = 'Y';
  ELSE NON_NORMAL = 'N';
IF RSQR > 0.65 AND B > 0 AND AUTOCORR = 'N' AND HETERO = 'N' AND NON_NORMAL = 'N'
   THEN INDIKATOR = 'A';
      ELSE INDIKATOR = 'R';
RUN;
PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_REZULT
AS SELECT PRKOD, MODEL, INDIKATOR, AUTOCORR, HETERO, NON_NORMAL, N, RSQR, A, B, SSD, SY,
XMOY
FROM OUTLIER_.EXP_01_08_RESULTS;

DROP TABLE OUTLIER_.EXP_01_08_RESULTS;
QUIT;
```

Išsiskiriančių reikšmių nustatymas

```
proc sql;
create table outlier_.exp_01_08_AR
 as select DKLR_NR, A.PRKOD, B.MODEL, PRNR, X, Y, B.N, b.indikator
  FROM OUTLIER_.Exp_01_08_SVORIAI1 A, OUTLIER_.EXP_01_08_REZULT B
   WHERE A.PRKOD = B.PRKOD /*AND B.INDIKATOR = 'R'*/ AND A.MODEL= B.MODEL;
quit;
data outlier_.exp_01_08_ar;
 set outlier_.exp_01_08_ar;
  Y = EXP(Y);
  X = EXP(X);
  sant = log(Y/X);
run;
PROC UNIVARIATE DATA = OUTLIER_.EXP_01_08_ar;
  BY PRKOD MODEL;
  VAR sant;
 OUTPUT OUT = OUTLIER_.EXP_01_08_arqu q1=Q1_sant q3=q3_sant;
RUN;
data outlier_.exp_01_08_arQU;
set outlier_.exp_01_08_arQU;
UL = Q3_SANT + 4*(Q3_SANT - Q1_SANT);
LL = Q1_SANT - 4*(Q3_SANT - Q1_SANT);
RUN;
```

```
/*SITOJE VIETOJE JAU GALIMA IDETI LENTELE, KURIAI REIKIA ISSKIRTI REIKSMES.
PRIES TAI PASKAICIAVUS SANTYKI Y/X.
TADA SUJUNGTI LENTELE SU OUTLIER_.EXP_01_08_ARQU PAGAL MODELI IR PREKES KODA.
REIKSME NEISSISKYRIANTI, KAI LL<= SANT<= UL. ISSISKIRIANCIOS REIKSMES BUS NUSTATYTOS TIK
TIEMS KODAMS IR MODELIAMS, KURIEMS NETIKO REGRESIJOS METODOLOGIJA*/
data outlier_.exp_01_08_kba;
set outlier_.exp_01_08_kb;
x = exp(X);
y = exp(Y);
run;
PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_ARB
 AS SELECT A.DKLR_NR, A.PRKOD, A.MODEL,A.PRNR, A.Y, A.X, B.UL, B.LL, c.indikator, Met,
menuo
  FROM OUTLIER_.EXP_01_08_kba A, OUTLIER_.EXP_01_08_ARQU B, outlier_.exp_01_08_rezult c
   WHERE A.PRKOD = B.PRKOD = c.prkod AND A.MODEL = B.MODEL = c.model;
QUIT;
DATA OUTLIER_.EXP_01_08_ARB;
SET OUTLIER_.EXP_01_08_ARB;
 SANT = log(Y/X);
 IF SANT<= UL AND SANT >= LL THEN OUTLIER ='N';
  ELSE OUTLIER = 'Y';
RUN;
PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_OUTLIER_ARB
 AS SELECT * FROM OUTLIER_.EXP_01_08_ARB
  WHERE OUTLIER = 'Y';
QUIT;
/*REGRESSION METHODOLOGY**/
PROC SQL;
create table outlier_.exp_01_08_RMB
 as select A.DKLR_NR, A.PRKOD, A.MODEL, A.X, A.Y, B.N, B.SSD, B.XMOY,
           B.SY, met, menuo, b.indikator
   FROM OUTLIER_.Exp_01_08_KB A, OUTLIER_.EXP_01_08_REZULT B
   WHERE A.PRKOD = B.PRKOD /*AND B.INDIKATOR = 'A' */AND A.MODEL= B.MODEL;

ALTER TABLE OUTLIER_.EXP_01_08_RMB
 ADD A NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RMB AS A
 SET A = (SELECT INTERCEPT FROM OUTLIER_.EXP_01_08_STEP4_AUTOREG B
          WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL);
ALTER TABLE OUTLIER_.EXP_01_08_RMB
 ADD B NUM FORMAT = BEST12.;
UPDATE OUTLIER_.EXP_01_08_RMB AS A
 SET B = (SELECT X FROM OUTLIER_.EXP_01_08_STEP4_AUTOREG B
          WHERE A.PRKOD = B.PRKOD AND A.MODEL = B.MODEL);
quit;
PROC SQL;
alter table outlier_.exp_01_08_rmB
 add NT_N1 num format = best12.;
update outlier_.exp_01_08_rmB as A
 set NT_N1 = case
              when (A.N - 2) >1000 THEN
                 (select B.NT_N1 FROM OUTLIER_.T_TABLE AS B WHERE B.N >1000)
                              ELSE (SELECT B.NT_N1 FROM OUTLIER_.T_TABLE AS B WHERE B.N =
(A.N - 2))
                                      END;

alter table outlier_.exp_01_08_rmB
 add T_T1 num format = best12.;
update outlier_.exp_01_08_rmB as A
 set T_T1 = case
             when (A.N - 2) >1000 THEN
                 (select B.T_T1 FROM OUTLIER_.T_TABLE AS B WHERE B.N >1000)
                              ELSE (SELECT B.T_T1 FROM OUTLIER_.T_TABLE AS B WHERE B.N =
(A.N - 2))
            END;

alter table outlier_.exp_01_08_rmB
 add N0_T num format = best12.;
```

```
update outlier_.exp_01_08_rmB as A
 set N0_T = case
             WHEN (A.N - 2) <=1000 THEN (SELECT B.N0_T FROM OUTLIER_.T_TABLE AS B WHERE
B.N = (A.N-2))
                                    END;

alter table outlier_.exp_01_08_rmB
 add T0 num format = best12.;
update outlier_.exp_01_08_rmB as A
 set T0 = case
             when (A.N - 2)<=1000 THEN (SELECT B.T0 FROM OUTLIER_.T_TABLE AS B WHERE B.N
= (A.N - 2))
                            END;
QUIT;
DATA OUTLIER_.EXP_01_08_RMB;
SET OUTLIER_.EXP_01_08_RMB;
IF ((N - 2) > 1000) OR  ((N - 2) = NT_N1)THEN TP = T_T1;
    ELSE
     TP = (1 - ((N - N0_T)*NT_N1)/((NT_N1 - N0_T)*N))*T0 +
                                  (((N - N0_T)*NT_N1)/((NT_N1 - N0_T)*N))*T_T1;
STUD = TP*SY*SQRT((1/N)+((X-XMOY)*(X-XMOY))/SSD+1);
YU = (A + B*X)+TP*SY*SQRT((1/N)+((X-XMOY)*(X-XMOY))/SSD+1);
YL = (A + B*X)-TP*SY*SQRT((1/N)+((X-XMOY)*(X-XMOY))/SSD+1);
IF Y <=YU AND Y >= YL THEN OUTLIER = 'N';
ELSE OUTLIER = 'Y';
RUN;
PROC SQL;
CREATE TABLE OUTLIER_.EXP_01_08_OUTLIER_RMB
 AS SELECT * FROM OUTLIER_.EXP_01_08_RMB
  WHERE OUTLIER = 'Y';
ALTER TABLE OUTLIER_.EXP_01_08_OUTLIER_RMB
 DROP NT_N1, T_T1, N0_T, T0, TP;
QUIT;
PROC SORT DATA = OUTLIER_.REAL_OUTLIERS;
BY PRKOD MODEL;
RUN;
proc sql;
CREATE TABLE OUTLIER_.EXP_01_08_SKIRT_AR
   AS SELECT A.RAKTAS, A.PRKOD, A.MODEL, A.X, A.Y, A.X_B, A.Y_B, X_XB, a.Met, a.menuo,
             Y_YB, A.SANT, A.SANTB, B.UL, B.LL, A.SKIRSANT, ABS, B.OUTLIER, b.indikator
    FROM OUTLIER_.EXP_01_08_OUTLIER_ARB B, OUTLIER_.REAL_OUTLIERS A
             WHERE A.RAKTAS = B.DKLR_NR and a.MET = b.met and a.menuo = b.menuo;

CREATE TABLE OUTLIER_.EXP_01_08_SKIRT_RM
   AS SELECT A.RAKTAS, A.PRKOD, A.MODEL, A.X, A.Y, A.X_B, A.Y_B, X_XB, a.met, a.menuo,
             Y_YB, A.SANT, A.SANTB, A.SKIRSANT, ABS, B.OUTLIER, b.indikator
    FROM OUTLIER_.EXP_01_08_OUTLIER_RMB B, OUTLIER_.REAL_OUTLIERS A
             WHERE A.RAKTAS = B.DKLR_NR and a.MET = b.met and a.menuo = b.menuo;

            create table outlier_.RM_skirt
             as select a.dklr_nr, a.prkod, b.model, b.x, b.y, b.x_b, b.y_b, outlier,
                        YL, YU, a.met, a.menuo, a.indikator
             from outlier_.exp_01_08_RMB A, outlier_.REAL_OUTLIERS B
               WHERE A.DKLR_NR = B.RAKTAS and a.MET = b.met and a.menuo = b.menuo;
            create table outlier_.AR_skirt
             as select a.dklr_nr, a.prkod, b.model, b.x, b.y, b.x_b, b.y_b, outlier,
               C.Q1_SANT, C.Q3_SANT, A.UL, A.LL, a.met, a.menuo, a.indikator
              from outlier_.exp_01_08_ARB A, outlier_.REAL_OUTLIERS B,
                   OUTLIER_.EXP_01_08_ARQU C
               WHERE A.DKLR_NR = B.RAKTAS AND A.MODEL = B.MODEL = C.MODEL AND A.PRKOD =
B.PRKOD = C.PRKOD
                and a.MET = b.met and a.menuo = b.menuo;
QUIT;
/***************NEPATIKRINTOS PASKUTINIO MENESIO KLAIDOS*****************************/
PROC SQL;
CREATE TABLE OUTLIER_.OUT0603_ARB
 AS SELECT * FROM OUTLIER_.EXP_01_08_OUTLIER_ARB
  WHERE MET = 2006 AND MENUO = '03';
  CREATE TABLE OUTLIER_.OUT0603_RMB
 AS SELECT * FROM OUTLIER_.EXP_01_08_OUTLIER_RMB
```

```
  WHERE MET = 2006 AND MENUO = '03';

ALTER TABLE OUTLIER_.OUT0603_ARB
 ADD DKLR_NR1 NUM FORMAT = BEST12. ;
ALTER TABLE OUTLIER_.OUT0603_RMB
 ADD DKLR_NR1 NUM FORMAT = BEST12. ;
QUIT;
DATA OUTLIER_.OUT0603_ARB;
SET OUTLIER_.OUT0603_ARB;
DKLR_NR1 = DKLR_NR;
RUN;
DATA OUTLIER_.OUT0603_RMB;
SET OUTLIER_.OUT0603_RMB;
DKLR_NR1 = DKLR_NR;
RUN;
DATA OUTLIER_.EXP_01_08_RMB;
SET OUTLIER_.EXP_01_08_RMB;
DKLR_NR1 = DKLR_NR;
RUN;
PROC SQL;
ALTER TABLE OUTLIER_.EXP_01_08_RMB
 DROP DKLR_NR;
CREATE TABLE OUTLIER_.EXP_01_08_RMB1
 AS SELECT A.DKLR_NR1, A.PRKOD, A.MODEL, B.PRNR, X, Y, N, SSD, XMOY, SY, MET, MENUO,
           INDIKATOR, A, B, NT_N1, T0, TP, STUD, YU, YL, OUTLIER, B.DKLR_NR
   FROM OUTLIER_.EXP_01_08_RMB A, OUTLIER_.EXP_01_08_RAKTAS B
 WHERE A.DKLR_NR1 = B.RAKT;
QUIT;
PROC SQL;
ALTER TABLE OUTLIER_.OUT0603_ARB
 DROP DKLR_NR;
ALTER TABLE OUTLIER_.OUT0603_RMB
 DROP DKLR_NR;

CREATE TABLE OUTLIER_.OUT0603_ARB1
 AS SELECT A.DKLR_NR1, A.PRKOD, A.MODEL, A.PRNR, Y, X, UL, LL, INDIKATOR, MENUO, SANT,
OUTLIER,
           B.DKLR_NR
  FROM OUTLIER_.OUT0603_ARB A, OUTLIER_.EXP_01_08_RAKTAS B
   WHERE A.DKLR_NR1 = B.RAKT;
PROC SQL;
CREATE TABLE OUTLIER_.OUT0603_RMB1
 AS SELECT A.DKLR_NR1, A.PRKOD, A.MODEL, B.PRNR, Y, X, N, SSD, XMOY, SY, MET, MENUO,
           B.DKLR_NR, YU, YL
  FROM OUTLIER_.OUT0603_RMB A, OUTLIER_.EXP_01_08_RAKTAS B
   WHERE A.DKLR_NR1 = B.RAKT;

QUIT;
```