

Statistinė analizė socialiniuose tyrimuose

Seminaras „**Geroji LiDA mokymų praktika: tyrėjų metodologinės kompetencijos tobulinimo link**“

dr. Vytautas Janilionis,
Kauno technologijos universitetas
Kaunas, 2011 10 12

- **Metodologiniai savarankiškų studijų paketai** skirti praktinių įgūdžių lavinimui:
 - Statistinė kiekybinių duomenų analizė su SPSS ir STATA
 - Kompiuterizuota kokybinių duomenų analizė su NVivo ir Text Analysis Suite
 - Lyginamieji tyrimai su Tosmana ir fs/QCA
 - Taikomoji regresinė analizė socialiniuose tyrimuose
- **Mokymų seminarų medžiaga** apima medžiagą, kuri yra skirta bendrosioms žinioms apie kiekybinę, kokybinę analizę [gyti, o taip pat sustiprina bendrąjį informacinį archyvo vartotojo pasiruošimą:
 - Kompiuterinis pasirengimas duomenų archyvo vartojimui
 - Informacinis raštingumas duomenų archyvo vartojimui
 - Statistinė analizė humanitarinių ir socialinių mokslų tyrimuose
 - Kokybinių duomenų analizė humanitarinių ir socialinių mokslų tyrimuose
 - Kokybinė lyginamoji analizė (QCA): principai ir programinės priemonės
 - Aprašomoji kiekybinių duomenų analizė
 - Antrinė kiekybinių duomenų analizė
 - Kiekybinių duomenų internetiniuose archyvuose analizė
 - Apklausų duomenų analizė
 - Inferencinė statistika socialiniuose moksluose
 - Koreliacinės ir regresinės analizės pagrindai
 - Daugialypės regresinės analizės taikymas socialiniuose tyrimuose
 - Logistinė regresija socialiniuose tyrimuose
 - Kokybinė lyginamoji analizė ir neryškiųjų aibių metodas
- Tarptautinių mokymo seminarų medžiaga (*anglų kalba*):
 - Tarptautinių apklausų duomenų analizė (*anglų kalba*)
 - Apklausų tyrimų problemos: metodologija ir kokybė (*anglų kalba*)

Mokymo seminarų medžiaga

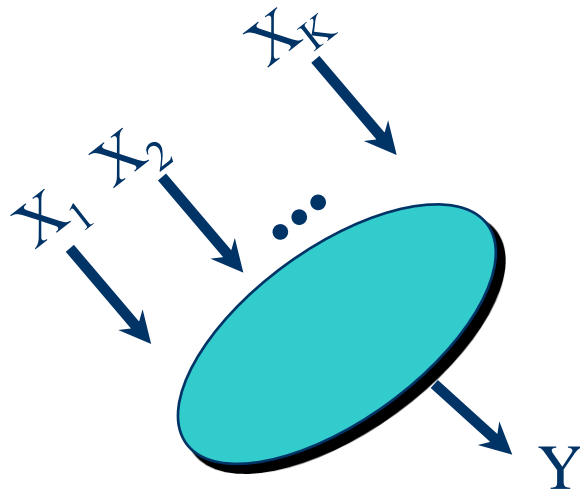
Koreliacinės ir regresinės analizės pagrindai

Daugialypės regresinės analizės taikymas socialiniuose tyrimuose

Mokymo kurso turinys:

- 1. Daugialypės tiesinės regresijos modelis.**
- 2. Regresinio modelio sudarymas kai skirtingose skalėse išmatuoti nepriklausomi kintamieji.**
- 3. Daugialypės tiesinės regresinės analizės modelio prielaidų tikrinimas ir neatitikimų šalinimo strategijos.**
- 4. Daugialypės regresinės analizės išvados.**

DAUGIALYPĖS TIESINĖS REGRESIJOS MODELIS



Kokia yra stochastinė priklausomybė tarp nepriklausomų kintamųjų

X_1, X_2, \dots, X_K

ir priklausomo kintamojo Y ?

Tiesinė $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon$

Netiesinė $Y = f(X_1, X_2, \dots, X_K, \beta_1, \beta_2, \dots) + \varepsilon$

Regresijos modelio kintamųjų matavimo skalės

<p>Intervalų</p>	<p>A10. Kaip jūs manote, ar žmonės dažniausiai stengiasi padėti kitiems, ar rūpinasi tik savimi? Atsakydami naudokitės šia kortele. <i>/PAŽYMĖTI TIK VIENĄ VARIANTĄ/</i></p> <table style="width: 100%; text-align: center;"> <tr> <td>00</td><td>01</td><td>02</td><td>03</td><td>04</td><td>05</td><td>06</td><td>07</td><td>08</td><td>09</td><td>10</td><td style="border-left: 1px solid black;">88</td> </tr> <tr> <td colspan="10">Žmonės dažniausiai rūpinasi tik savimi</td> <td colspan="2">Žmonės dažniausiai stengiasi padėti kitiems</td> </tr> <tr> <td colspan="11"></td> <td>(Nežino)</td> </tr> </table>	00	01	02	03	04	05	06	07	08	09	10	88	Žmonės dažniausiai rūpinasi tik savimi										Žmonės dažniausiai stengiasi padėti kitiems													(Nežino)
00	01	02	03	04	05	06	07	08	09	10	88																										
Žmonės dažniausiai rūpinasi tik savimi										Žmonės dažniausiai stengiasi padėti kitiems																											
											(Nežino)																										
<p>Dvireikšmis kintamasis</p>	<p>C3. Ar jūs turite su kuo aptarti asmeninius ir intymius reikalus?</p> <table style="margin-left: auto; margin-right: auto;"> <tr> <td>Taip</td><td>1</td> </tr> <tr> <td>Ne</td><td>2</td> </tr> <tr> <td>(Nežino)</td><td>8</td> </tr> </table>	Taip	1	Ne	2	(Nežino)	8																														
Taip	1																																				
Ne	2																																				
(Nežino)	8																																				
<p>Santykių</p>	<p>F1. Kiek žmonių nuolatos gyvena šiame gyvenamame būste (jūsų namų ūkyje), įskaitant jus patį ir vaikus.</p> <p style="text-align: center;">ĮRAŠYTI SKAIČIŲ: <input style="width: 40px; height: 20px; border: 1px solid black;" type="text"/> <input style="width: 40px; height: 20px; border: 1px solid black;" type="text"/></p> <p style="text-align: center;">(Nežino) 88</p>																																				

REGRESIJOS MODELIO SUDARYMAS KAI SKIRTINGOSE SKALĖSE IŠMATUOTI NEPRIKLAUSOMI KINTAMIEJI

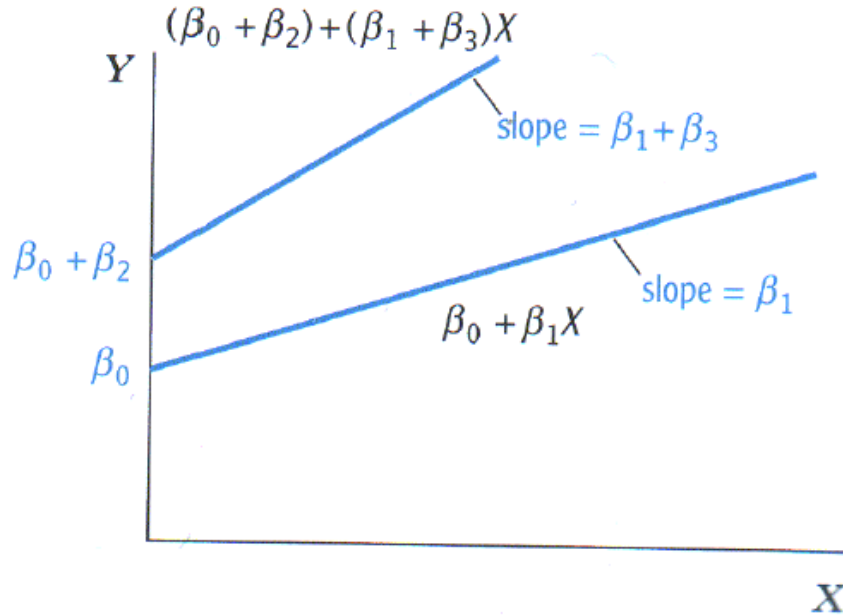
Klasikiniame regresijos modelyje kintamieji Y, X_1, X_2, \dots, X_K yra kiekybiniai, išmatuoti intervalų arba santykių skalėje. Be klasikinio regresijos modelio yra daug kitų regresijos modelių, kuriuose dalis kintamųjų yra kokybiniai ir išmatuoti „silpnesnėse“ (vardų ir/arba tvarkos) skalėse. Šiame kurse apsiribosime tik modeliu, kai klasikinis regresijos modelis papildomas dvireikšmiais kokybiniais kintamaisiais, kurie dar vadinami pseudokintamaisiais. Visi dvireikšmiai kintamieji gali įgyti tik dvi reikšmes: **0** ir **1**.

Pavyzdžiui,

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon)$$

$$D = \begin{cases} 1, & \text{moteris,} \\ 0, & \text{vyras.} \end{cases}$$

Įrašę į regresijos lygtį $D=1$, gausime regresijos lygtį moterims, o įrašę 0 gausime regresijos lygtį vyrams. Taigi gavome kiekvieną pseudokintamojo reikšmę atitinkančią regresijos funkciją. Pseudokintamuosius tikslinga taikyti kai regresijos tiesės abiejų kategorijų yra lygiagrečios. Dažnai į regresijos modelį įtraukiami kintamieji lytis, gyvenamoji vieta, politinės pažiūros, tautybė ir pan. Kai kategorinis kintamasis turi $m > 2$ kategorijų, jis keičiamas $(m-1)$ dvireikšmiu kintamuoju. Galima sudaryti regresijos modelį su tolydžiujų ir pseudokintamųjų sąveikomis.



Skirtingos atkarpos,
skirtingi nuolydžiai.

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + \varepsilon$$

Pavyzdžiui: Y–atlyginimas, X–darbo stažas.

$$D = \begin{cases} 1, & \text{baigęs aukštąją mokyklą,} \\ 0, & \text{priešingu atveju.} \end{cases}$$

„**B24**|Pasitenkinimas dabartiniu gyvenimu apskritai“ =

$\beta_0 + \beta_1 * „\mathbf{B23}$ |Pozicija kairės-dešinės skalėje× **C3d**|Ar turi su kuo aptarti

asmeninius ir intymius reikalus“ +

$\beta_2 * „\mathbf{B25}$ |Pasitenkinimas dabartine Lietuvos ekonomine situacija“ +

$\beta_3 * „\mathbf{B27}$ |Pasitenkinimas tuo, kaip demokratija veikia Lietuvoje“ +

$\beta_4 * „\mathbf{A10}$ |Žmonės dažniausiai stengiasi padėti kitiems, ar rūpinasi tik savimi“ +

$\beta_5 * „\mathbf{F1}$ |Kiek žmonių nuolatos gyvena namų ūkyje, įskaitant respondentą ir vaikus“ +

$\beta_6 * „\mathbf{C3d}$ |Ar turi su kuo aptarti asmeninius ir intymius reikalus“ + ε

$$Y = \beta_0 + \beta_1 (D \times X_1) + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 D + \varepsilon$$

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	2,846	,163		17,420	,000	2,525	3,166		
	B25 Pasitenkinimas dabartine Lietuvos ekonomine situacija	,313	,035	,239	8,853	,000	,244	,382	,605	1,653
	B27 Pasitenkinimas tuo, kaip demokratija veikia Lietuvoje	,238	,026	,241	9,029	,000	,186	,289	,620	1,613
	A10 Žmonės dažniausiai stengiasi padėti kitiems, ar rūpinasi tik savimi	,116	,024	,107	4,766	,000	,068	,164	,870	1,150
	F1 Kiek žmonių nuolatos gyvena namų ūkyje, įskaitant respondentą ir vaikus	,123	,043	,060	2,847	,004	,038	,207	,981	1,019
	C3d Ar turi su kuo aptarti asmeninius ir intymius reikalus	-2,321	,437	-,312	-5,307	,000	-3,178	-1,463	,128	7,826
	C3d_x_B23	,273	,082	,195	3,321	,001	,112	,435	,128	7,830

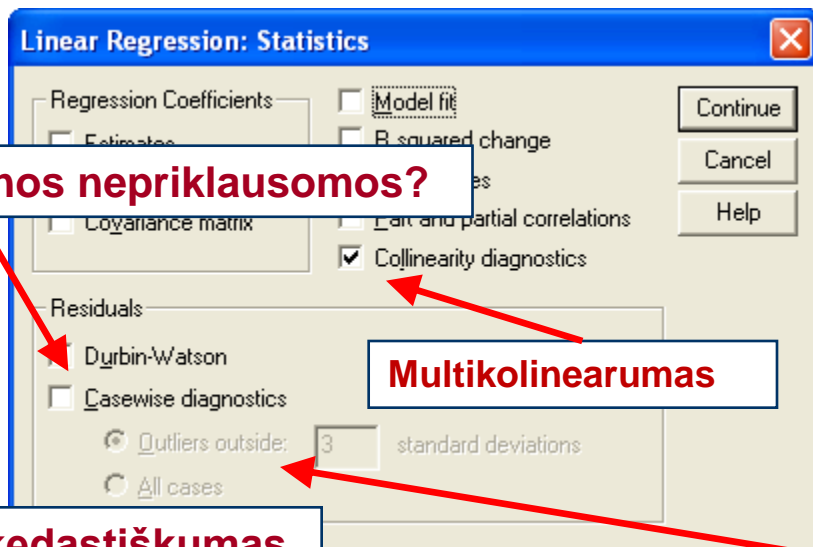
a. Dependent Variable: B24|Pasitenkinimas dabartiniu gyvenimu apskritai

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,507 ^a	,257	,254	2,089	1,530

DAUGIALYPĖS TIESINĖS REGRESINĖS ANALIZĖS MODELIO PRIELAUDŲ TIKRINIMAS IR NEATITIKIMŲ ŠALINIMO STRATEGIJOS

Tiesinės regresijos modelio prielaidų tikrinimas (SPSS)

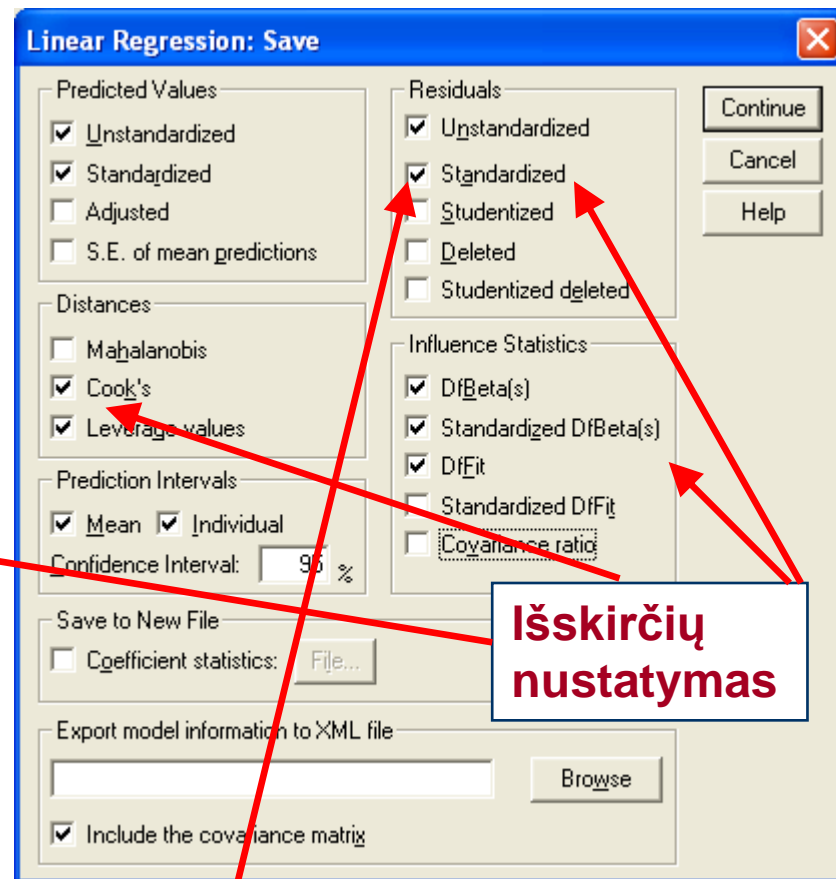


Linear Regression: Statistics

Regression Coefficients: Model fit, Estimates, R squared change, Covariance matrix, Variance-covariance matrix, Variance inflation factors, Partial and partial correlations, Collinearity diagnostics

Residuals: Durbin-Watson, Casewise diagnostics, Outliers outside: 3 standard deviations, All cases

Buttons: Continue, Cancel, Help



Linear Regression: Save

Predicted Values: Unstandardized, Standardized, Adjusted, S.E. of mean predictions

Distances: Mahalanobis, Cook's, Leverage values

Prediction Intervals: Mean, Individual, Confidence Interval: 95 %

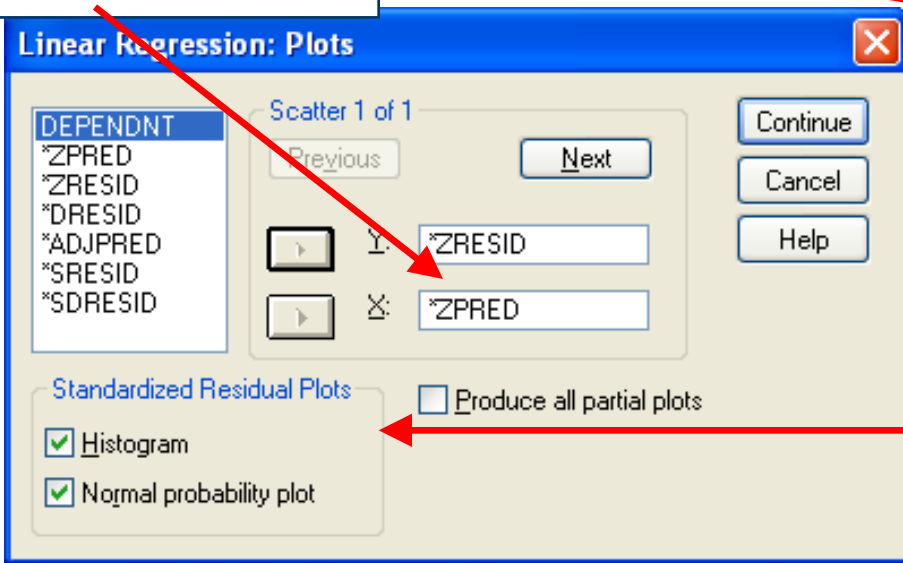
Save to New File: Coefficient statistics: File...

Export model information to XML file: Include the covariance matrix

Residuals: Unstandardized, Standardized, Studentized, Deleted, Studentized deleted

Influence Statistics: DfBeta(s), Standardized DfBeta(s), DfFit, Standardized DfFit, Covariance ratio

Buttons: Continue, Cancel, Help



Linear Regression: Plots

DEPENDNT: *ZPRED, *ZRESID, *DRESID, *ADJPRED, *SRESID, *SDRESID

Scatter 1 of 1: Y: *ZRESID, X: *ZPRED

Standardized Residual Plots: Histogram, Normal probability plot, Produce all partial plots

Buttons: Continue, Cancel, Help

Ar liekanos nepriklausomos?

Multikolinearumas

Homoskedastiškumas

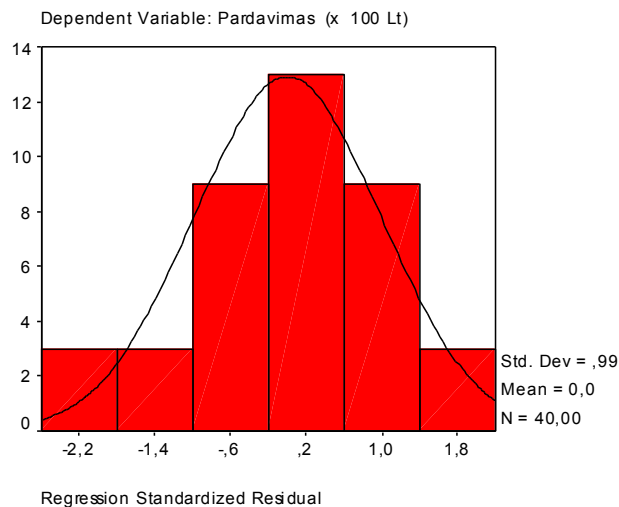
Išskirčių nustatymas

Standartizuotųjų liekanų normalumo tikrinimas

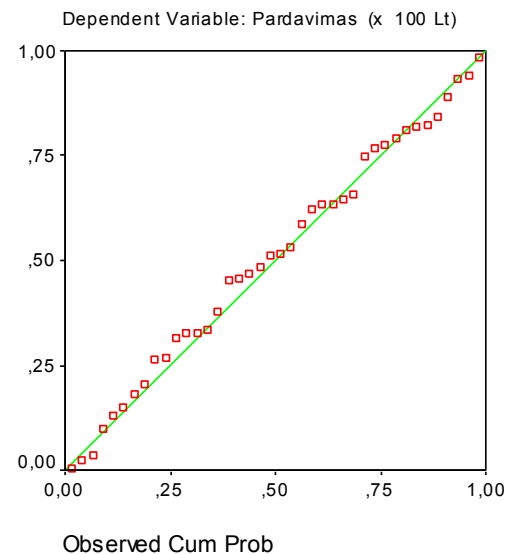
Liekanų normalumo tikrinimas

Išvada. Vizualiai palyginę, galime teigti, kad standartizuotųjų liekanų histograma yra suderinta su standartinio normaliojo skirstinio kreive (t.y. standartizuotųjų liekanų skirstinys yra suderintas su standartiniu normaliuoju skirstiniu).

Histogram



Normal P-P Plot of Regression Standardized Residual



Itakos matas $DfFit_i$ parodo i -tojo stebėjimo pašalinimo įtaką

prognozuojamai reikšmei \hat{Y}_i , $DfFit_i = \hat{Y}_i - \hat{Y}_{i(i)}$, čia $\hat{Y}_{i(i)}$ yra prognozė pagal regresijos lygtį gautą pašalinus i -tąjį stebėjimą. Nustatant išskirtis

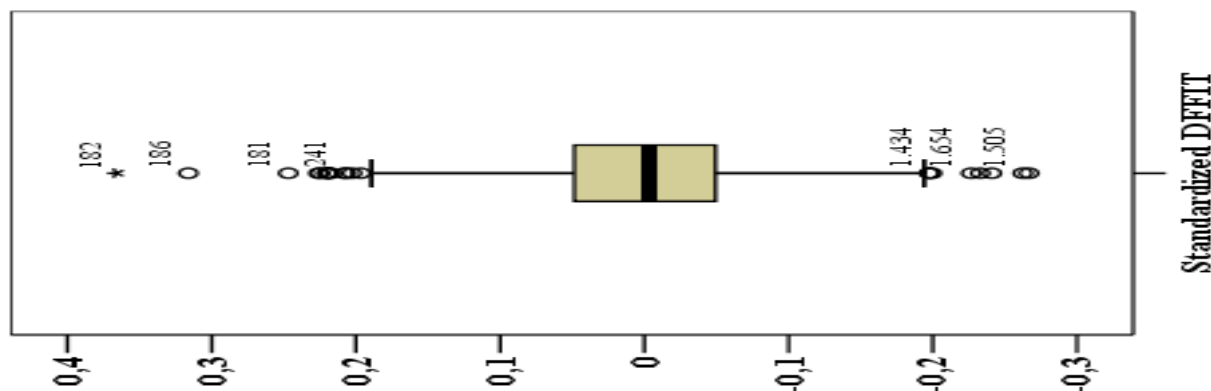
naudojami standartizuotoji $DfFit_i$ reikšmė $Std.DfFit_i$. Jeigu,

$|Std.DfFit_i| > 2 \cdot \sqrt{(K+1)/n}$, tai i -tasis stebėjimas laikomas išskirtimi, jo pašalinimas

įtakoja prognozę \hat{Y}_i . Mūsų nagrinėjamame pavyzdyje įtakos mato $|Std.DfFit_i|$

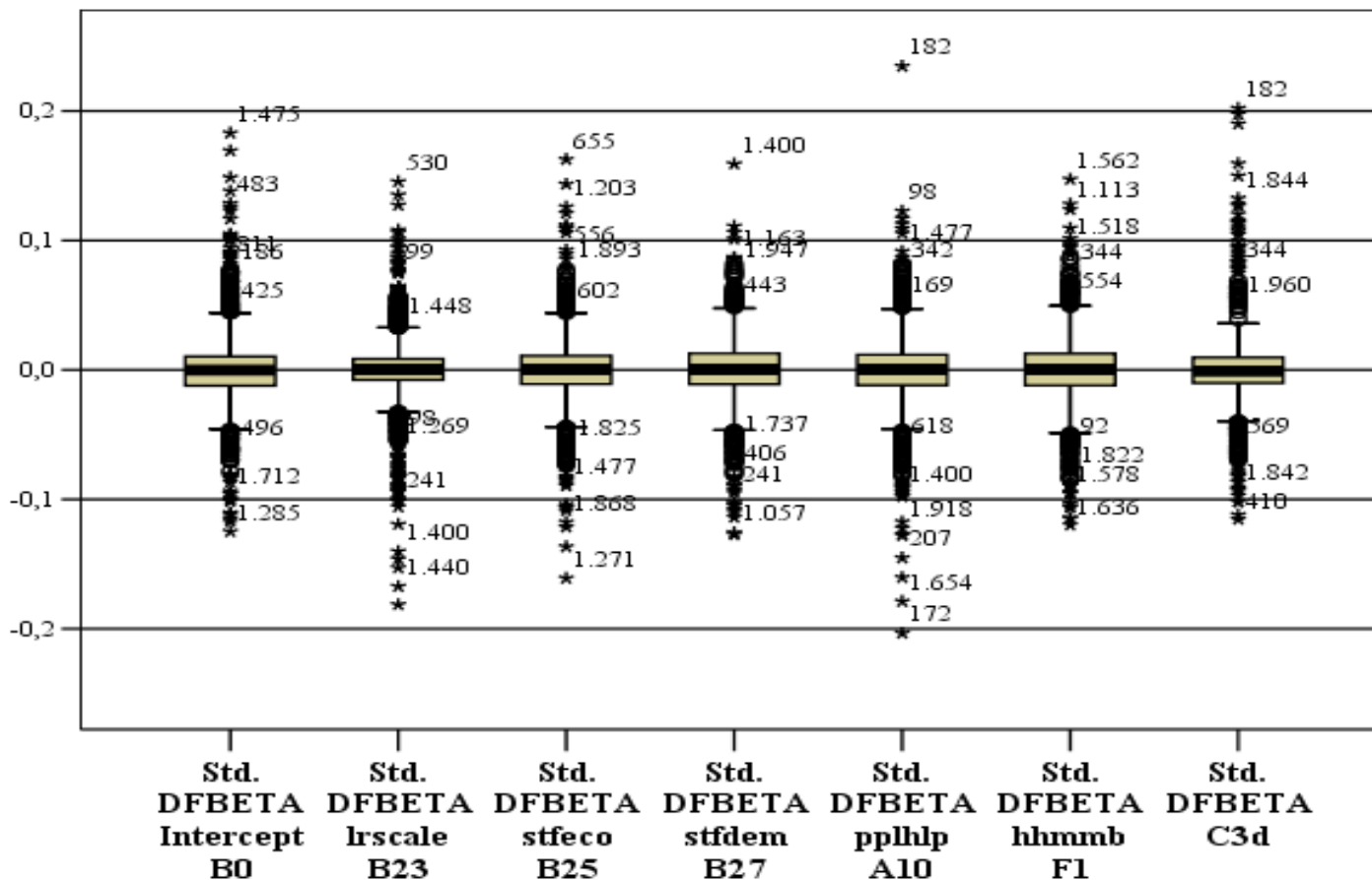
ribinė reišmė yra $2 \cdot \sqrt{(6+1)/1221} = 0,1514$. Gavome, kad imtyje yra daug

išskirčių, nes daug $|Std.DfFit_i|$ reikšmių viršija ribinę reikšmę 0,1514.



Įtakos matas $DfBetas_{ji}$ parodo i -tojo stebėjimo įtaką regresijos koeficientui β_j . $DfBetas_{ji} = b_j - b_{j(i)}$, čia $b_{j(i)}$ yra koeficiento β_j taškinis įvertis, kai pašalintas i -tasis stebėjimas, $j=0,1,\dots,K$. Nustatant išskirtis naudojami standartizuotieji $DfBetas_{ji}$. Jeigu $|\text{Std.DfBetas}_{ji}| > 2/\sqrt{n}$, tai i -tasis stebėjimas laikomas išskirtimi, jo pašalinimas įtakoja regresijos koeficiento β_j taškinį įvertį b_j , o tuo pačiu ir išvadas, kurios formuluojamos pagal šį koeficientą. Mūsų nagrinėjamame pavyzdyje įtakos mato $|\text{Std.DfBetas}_{ji}|$ ribinė reikšmė yra $2/\sqrt{n} = 0,0572$. Gavome, kad imtyje yra daug išskirčių, nes daug $|\text{Std.DfBetas}_{ji}|$ reikšmių viršija ribinę reikšmę 0,0572.

Įtakos matų Std. DfBetas; stačiakampės diagramos



Įtakos matas	Ribinė reikšmė
Standartizuota liekana (standardized residual)	3
Studentizuota liekana (studentized residual)	2 arba 3
Centruotas įtakos indeksas (centered leverage) ch_i	$2(K+1)/n$ arba $3(K+1)$
Kuko įtakos matas (CooksD) $_i$	$4/n$
$ \text{Std.DfFit}_i $	$2 \cdot \sqrt{(K+1)/n}$
CovRatio_i	$1-3K/n$ ir $1+3K/n$
$ \text{Std.DfBetas}_{ji} $	$2/\sqrt{n}$

K – nepriklausomų kintamųjų skaičius modelyje,
 n – stebėjimų skaičius (imties didumas).

Daugialypės regresinės analizės išvados

Regresinė analizė yra galingas statistikos įrankis, deja dažnai ši analizė yra neteisingai taikoma arba traktuojama, dėl to daromos neteisingos prognozės arba priimami neteisingi sprendimai. Pavyzdžiui, tiesinė regresinė analizė vis tiek taikoma, nors netenkinamos jos taikymo prielaidos (dažnai tyrėjai nepatikrina regresijos liekanų normalumo, homoskedastiškumo, išskirčių įtakos regresijos koeficientams, multikolinearumo ir kitų prielaidų). Taigi, korektiškos išvados regresinėje analizėje galimos tik tuo atveju, jeigu tenkinamos modelio tinkamumo prielaidos.

- Įtraukti į regresijos lygtį naujus nepriklausomus kintamuosius, kurie geriau paaiškintų priklausomą kintamąjį.
- Pašalinti dalį nepriklausomų kintamųjų iš modelio, nes didelis nepriklausomų kintamųjų skaičius gali pabloginti prognozavimą dėl nepriklausomų kintamųjų multikolinearumo. Kintamųjų pašalinimui galima naudoti SPSS tiesinės regresinės analizės procedūroje realizuotą kintamųjų išbraukimo metodą (Backward).
- Transformuoti kintamuosius, taip, kad tiesinė regresija jiems tiktų. Dažnai ši problema sprendžiama logaritmuojant tiek priklausomą kintamąjį, tiek dalį nepriklausomų kintamųjų. Kartais į regresijos modelį įtraukiant kintamieji pakelti atitinkamais laipsniais, arba įtraukiamos dviejų ir daugiau kintamųjų tarpusavio sąveikos (sandaugos). Tokiu atveju visada reikia patikrint ar neatsirado multikolinearumo problema.
- Rinktis netiesinį modelį. Kartais kintamųjų negalima transformuoti taip, kad tiesinė regresija tiktų. Tokiu atveju reikia taikyti netiesinę regresinę analizę.

Imties daugialypės tiesinės regresijos lygtis:

$$\begin{aligned} \text{„B24|Pasitenkinimas dabartiniu gyvenimu apskritai“} = & \\ & 2,44 + 0,085 * \text{„B23|Pozicija kairės-dešinės skalėje“} + \\ & 0,295 * \text{„B25|Pasitenkinimas dabartine Lietuvos ekonomine situacija“} + \\ & 0,245 * \text{„B27|Pasitenkinimas tuo, kaip demokratija veikia Lietuvoje“} + \\ & 0,087 * \text{„A10|Žmonės dažniausiai stengiasi padėti kitiems, ar rūpinasi} \\ & \text{tik savimi“} + 0,133 * \text{„F1|Kiek žmonių nuolatos gyvena namų ūkyje,} \\ & \text{įskaitant respondentą ir vaikus“} - 0,875 * \text{„C3d|Ar turi su kuo aptarti} \\ & \text{asmeninius ir intymius reikalus“} \end{aligned}$$

$$\hat{Y} = 2,44 + 0,085 \cdot X_1 + 0,295 \cdot X_2 + 0,245 \cdot X_3 + 0,087 \cdot X_4 + 0,133 \cdot X_5 - 0,875 \cdot D.$$

Suformuluosime išvadą apie imties regresijos lygties koeficientą b_3 prie kintamojo: „Padidėjus **respondentų** pasitenkinimo tuo, kaip demokratija veikia Lietuvoje, vertinimui vienu balu, **vidutinis** pasitenkinimas dabartiniu gyvenimu apskritai **padidėja** 0,245 balo, kai likusieji kintamieji yra fiksuoti“. Analogiškai formuluojamos išvados ir apie kitus koeficientus.

Suformuluosime išvadą apie populiacijos regresijos lygties koeficiento β_3 pasiklovimo intervalą $PI_{0,95}(\beta_3) = (0,185 ; 0,304)$: „Su 95% garantija galime prognozuoti, kad padidėjus **Lietuvos gyventojų** pasitenkinimui tuo, kaip demokratija veikia Lietuvoje vertinimui vienu balu, **vidutinis** pasitenkinimo dabartiniu gyvenimu apskritai **padidėjimas** yra intervale nuo 0,185 iki 0,304 balo, kai kitų kintamųjų reikšmės yra fiksuotos“.

Tarkime **prognozuojame** Lietuvos žmonių su fiksuotomis charakteristikomis pasitenkinimą dabartiniu gyvenimu apskritai.

B23 | Pozicija kairės-dešinės skalėje - **4**,

B25 | Pasitenkinimas dabartine Lietuvos ekonomine situacija - **3**,

B27 | Pasitenkinimas tuo, kaip demokratija veikia Lietuvoje - **2**,

A10 | Žmonės dažniausiai stengiasi padėti kitiems, ar rūpinasi tik savimi - **4**,

F1 | Kiek žmonių nuolatos gyvena namų ūkyje, įskaitant respondentą - **3**,

C3d | Ar turi su kuo aptarti asmeninius ir intymius reikalus - **0 – taip**.

Su **95% garantija** galime prognozuoti, kad Lietuvos gyventojų su tokiomis fiksuotomis nepriklausomų kintamųjų reikšmėmis, **pasitenkinimo** dabartiniu gyvenimu apskritai **vidurkis** yra intervale nuo **4,718** iki **5,097** balo

PABAIGAI

Statistikos mokslui jo taikymo sritys bei informacinės technologijos yra „**gyvybės šaltinis**“, kuris stimuliuoja naujų teorijų ir metodų kūrimą.

Vienas iš pagrindinių **ateities statistikos** mokslo ir jos taikymų krypčių - **didelių ir sudėtingos struktūros duomenų masių bei didelio dažnio duomenų statistinė analizė realiam laike virtualioje erdvėje.**

Ateities analitikai ekonomistai, informatikai, inžinieriai, sociologai, biologai, psichologai, medikai, edukologai ir t. t. turi būti gerai įvaldę statistikos metodus, kad galėtų siekti karjeros aukštumų. Lietuvos universitetų studijų programose, kurios rengia šiuos specialistus, šiandien statistikai skiriama per mažai laiko.

Herbertas Džordžas Velsas
(1866-1946 m.)– anglų rašytojas,
kartu su Žiuliu Vernu, laikomas
„Mokslinės fantastikos tėvu“.
Žymiausi kūriniai: „Laiko mašina“,
„Pasaulių karas“, „Nematomas
žmogus“, „Daktaro Moro sala“ ir kt.,
kurie ne kartą ekranizuoti.

Herbert George Wells



1903 m. H.G. Velsas pasakė:

**„Statistinis mąstymas kada nors taps toks pat
reikalingas žmonėms, kaip gebėjimas rašyti ir
skaityti“.**

AČIŪ UŽ DĖMESĮ!



“I keep saying the sexy job in the next ten years will be statisticians. People think I’m joking.....” *

*** Hal Varian, *Google’s chief economist*, The McKinsey Quarterly, January, 2009**