Статистички софтвер 4

Шести час

Марија Радичевић

Математички факултет, Београд

2015.

Садржај

Процедуре за тестирање везе између променљивих

Фокус истраживања односа између променљивих је усмерен на откривање конзистентне и системске везе између нивоа, односно ознака, две или више променљивих.

- питања везана за истраживање односа:
  1. Да ли веза између променљивих постоји?
  2. Ако веза постоји, који је смер те везе?
  3. Која је јачина везе?
  4. Који је тип везе?
- одговори на питања:
  1. закључује се на основу статистичке значајности
  2. може да буде позитивна и негативна
  3. може да буде: не постоји веза, слаба веза, умерене веза и јака веза
  4. може бити линеарна и нелинеарна

Процедуре за тестирање везе између променљивих

1. корелациона анализа
2. регресиона анализа
3. хи-квадрат анализа

Напомена

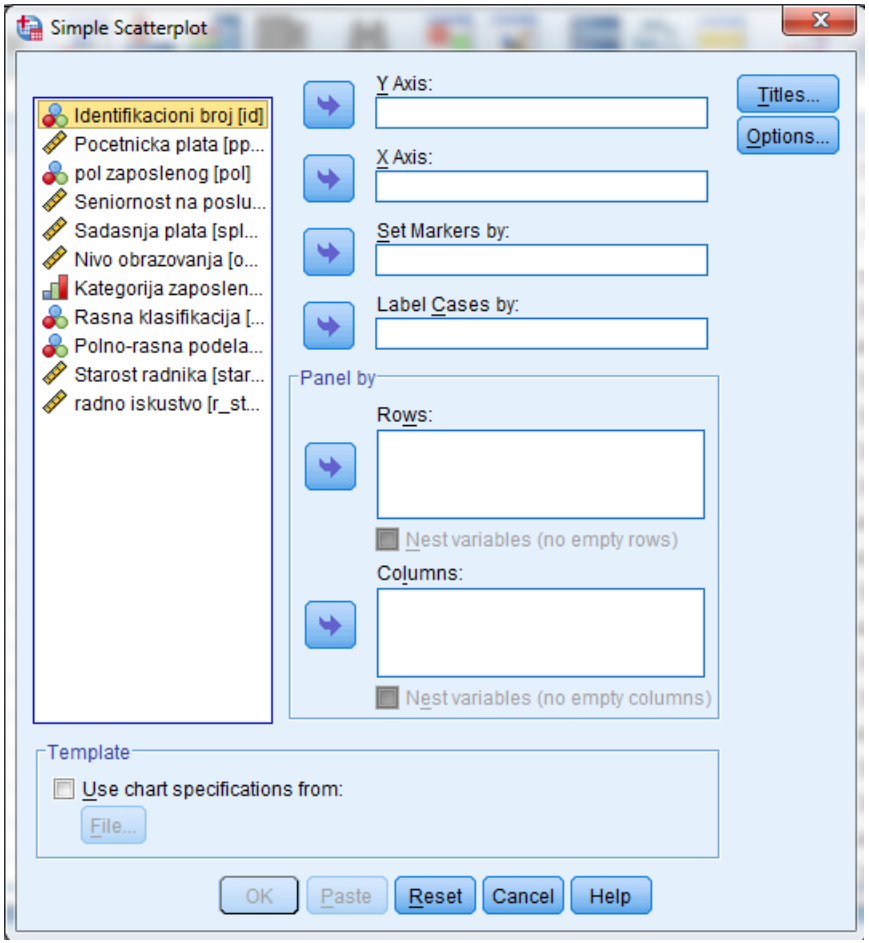
1. и 2. се користе за везе између метричких променљивих, а 3. за везу између категоријских променљивих.

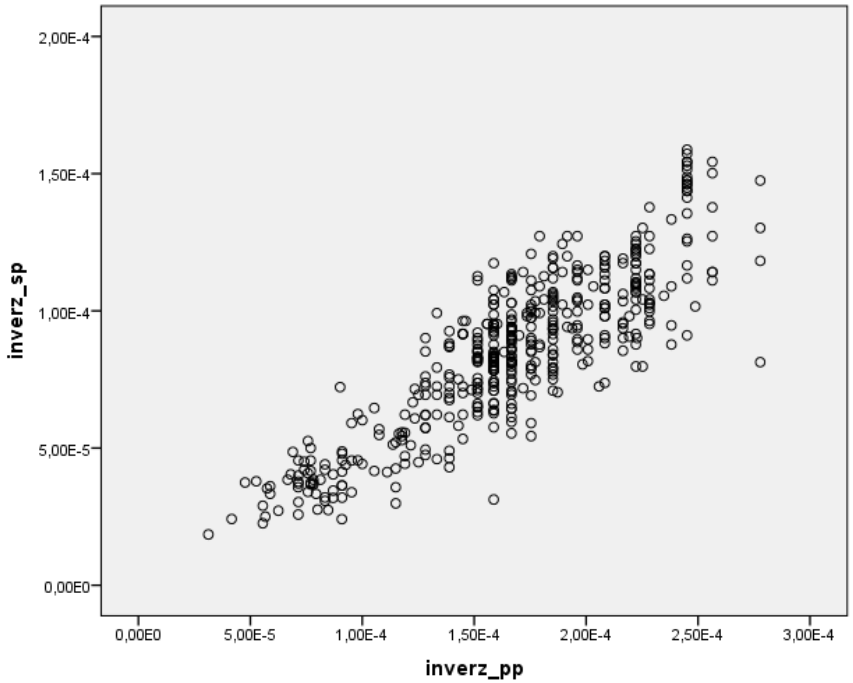
Процедуре за корелациону анализу

- степен везе између променљивих које се посматрају се испитује корелационом анализом
- мера корелације се изражава коефицијентом корелације који узима вредности у интервалу $[-1, 1]$
- корелација се може поделити према:
  1. смеру (позитивна или негативна)
  2. броју посматраних променљивих (проста или вишеструка)
  3. типу везе (линеарна или нелинеарна)

Диаграм распршености

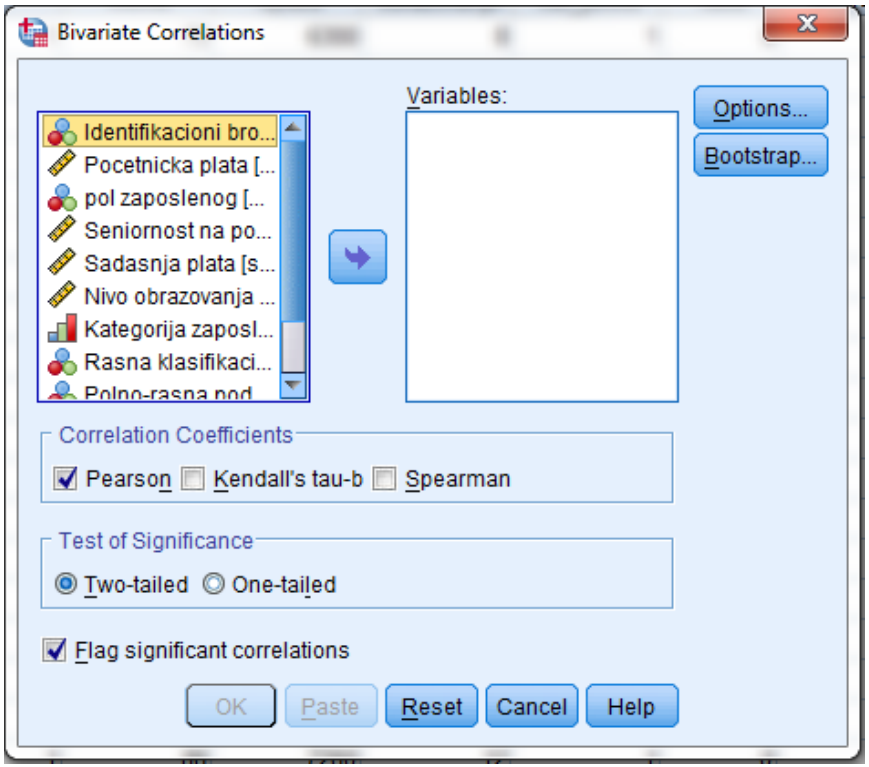
- график који приказује све тачке узорка (x_i, y_i) две нумеричке променљиве
- на основу њега може се одредити облик, смер и приближна јачина везе између посматраних обележја
- ако су тачке распршене без икаквог правила, можемо закључити да нема никакве везе између променљивих
- *SPSS* : *Graphs* $\Rightarrow$ *Legacy Dialogs* $\Rightarrow$ *Scatter/Dot...*, за испитивање просте корелације изабрати *Simple Scatterplot*


Simple Scatterplot
Identifikacioni broj [id]
Pocetnicka plata [pp...
pol zaposlenog [pol]
Seniornost na poslu...
Sadasnja plata [spl...
Nivo obrazovanja [o...
Kategorija zaposlen...
Rasna klasifikacija [...
Polno-rasna podela...
Starost radnika [star...
radno iskustvo [r_st...
Y Axis:
X Axis:
Set Markers by:
Label Cases by:
Titles...
Options...
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help


2,00E-4
1,50E-4
1,00E-4
5,00E-5
0,00E0
inverz_sp
0,00E0
5,00E-5
1,00E-4
1,50E-4
2,00E-4
2,50E-4
3,00E-4
inverz_pp

Коефицијент корелације

1. Пирсонов коефицијент корелације
2. Спирманов коефицијент корелације
3. Кендалов тау-б коефицијент корелације



Пирсонов коефицијент корелације

- користи се за испитивање степена линеарне везе између две нумеричке променљиве са нормалном расподелом

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

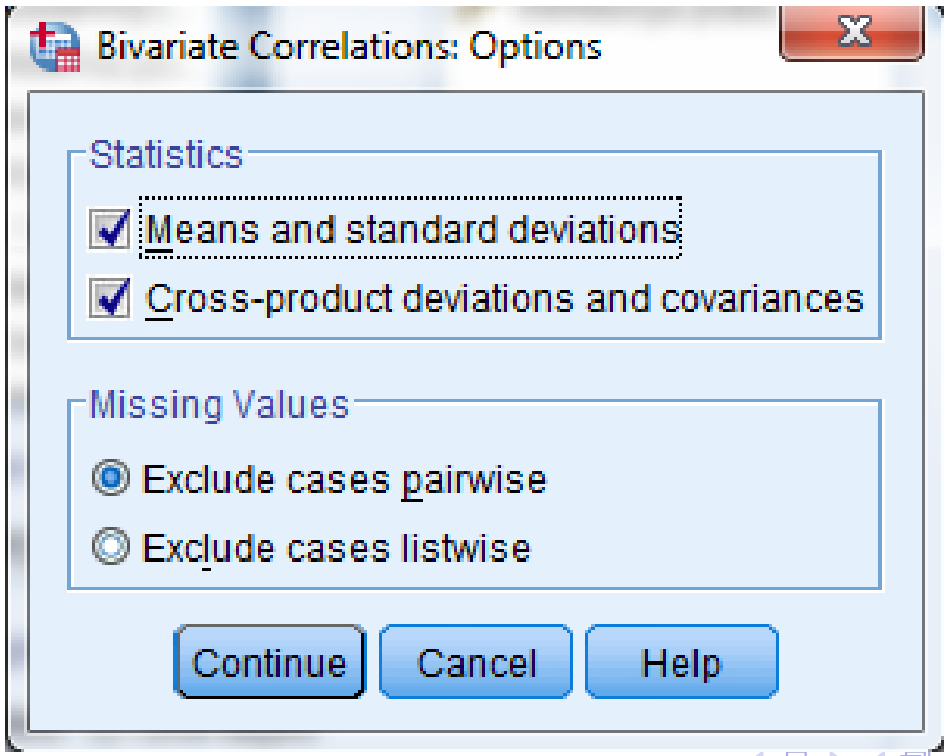
- коефицијент детерминације: показује који је део варијансе једне променљиве објашњен другом променљивом

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Табела корелација

1. вредности Пирсоновог коефицијента
   - коефицијент корелације сваке променљиве са самом собом је увек 1
   - корелација између променљивих X и Y је иста као између Y и X
2. $p-$вредност теста:
   - $H_0 : r = 0$
   - $H_1 : r \neq 0$ за двострану критичну област $(Two - tailed)$, односно $H_1 : r > 0$ или $H_1 : r < 0$ за једнострану критичну област $(One - tailed)$
   - тест статистика: $t = r\sqrt{\frac{n-2}{1-r^2}} : t_{n-2}$ при H_0
3. звездице су дефинисане легендом испод табеле (једна звездица означава да је корелација статистички значајна уз могуност грешке од 0.05, а две уз могућност грешке од 0.01)

- Sum of Cross-products: $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- Sum of Squares: $\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Covariance: $\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$



Descriptive Statistics

	Mean	Std. Deviation	N
inverz_pp	,0002	,00005	474
inverz_sp	,0001	,00003	474

Correlations

		inverz_pp	inverz_sp
inverz_pp	Pearson Correlation	1	,861**
	Sig. (2-tailed)		,000
	Sum of Squares and Cross-products	,000	,000
	Covariance	,000	,000
	N	474	474
inverz_sp	Pearson Correlation	,861**	1
	Sig. (2-tailed)	,000	
	Sum of Squares and Cross-products	,000	,000
	Covariance	,000	,000
	N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

⇒висока позитивна корелација

Sig.(2 – *tailed*) показује са колико поверења треба прихватити добијене резултате (не говори о јачини везе)

Спирманов коефицијент корелације

- користи се када није испуњени бар један услови за примену Пирсоновог коефицијента корелације
  - X или Y нема нормалну расподелу
  - X или Y су мерене ординалном скалом
  - узорак је мали
  - веза између X и Y није линеарна
- његова вредност не зависи од нумеричких вредности променљивих већ од њихових релативних односа (рангова)
- непараметарски је јер не претпоставља расподелу променљивих X и Y

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

где је d_i разлика између рангова два елемента

Correlations

			Pocetnicka plata	Sadasnja plata
Spearman's rho	Pocetnicka plata	Correlation Coefficient	1,000	,826**
		Sig. (2-tailed)	.	,000
		N	474	474
	Sadasnja plata	Correlation Coefficient	,826**	1,000
		Sig. (2-tailed)	,000	.
		N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

$\Rightarrow$висока позитивна корелација

Тестира се хипотеза $H_0 : \rho = 0$ на исти начин као код Пирсоновог коефицијента

Кендалов тау-б коефицијент корелације

- ако се у узорку јавља пуно једнаких вредности (једнаки рангови) Спирманов коефицијент постаје велики
- у том случају пожељно је користити Кендалов тау-б коефицијент који такође упоређује рангове
- непараметарска статистика
- најчешће има мању вредност него Спирманов коефицијент

$$\tau = \frac{S}{\frac{1}{2}n(n-1)}$$

где је $S = broj\ saglasnih\ parova - broj\ nesaglasnih\ parova$

- Пар (x_i, y_i) и (x_j, y_j) је сагласан ако важи $x_i > x_j$ и $y_i > y_j$ или $x_i < x_j$ и $y_i < y_j$
- Пар (x_i, y_i) и (x_j, y_j) је несагласан ако важи $x_i > x_j$ и $y_i < y_j$ или $x_i < x_j$ и $y_i > y_j$
- Пар (x_i, y_i) и (x_j, y_j) је везан ако важи $x_i = x_j$ или $y_i = y_j$

Correlations

			Pocetnicka plata	Sadasnja plata
Kendall's tau_b	Pocetnicka plata	Correlation Coefficient	1,000	,656**
		Sig. (2-tailed)	.	,000
		N	474	474
	Sadasnja plata	Correlation Coefficient	,656**	1,000
		Sig. (2-tailed)	,000	.
		N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

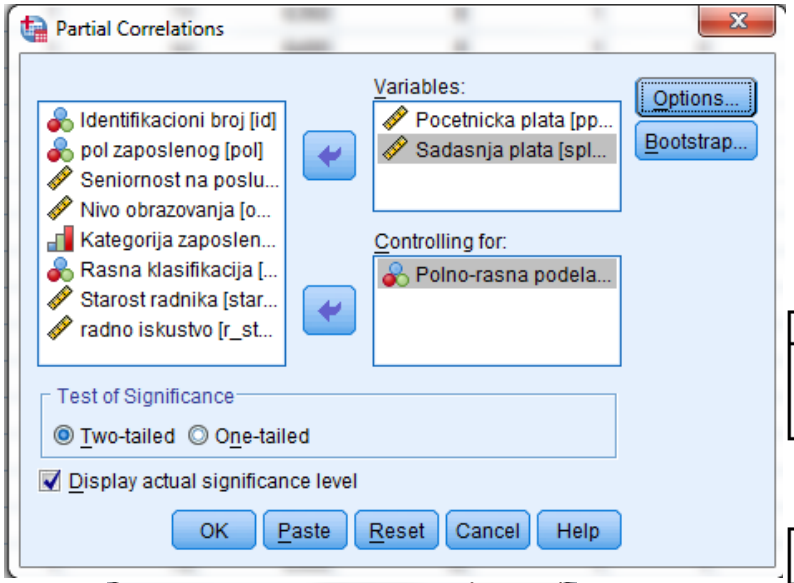
$\Rightarrow$умерена до јака позитивна корелација

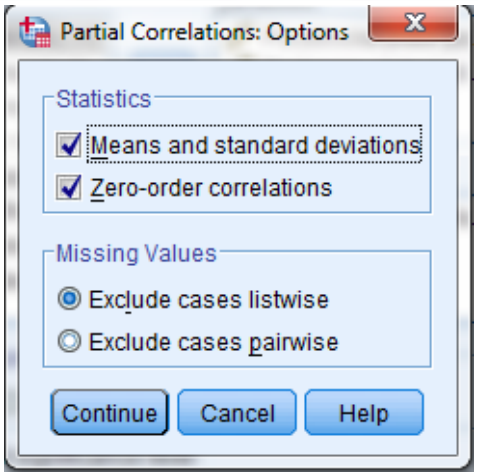
Тестира се хипотеза $H_0 : \tau = 0$ на исти начин као код Пирсоновог коефицијента

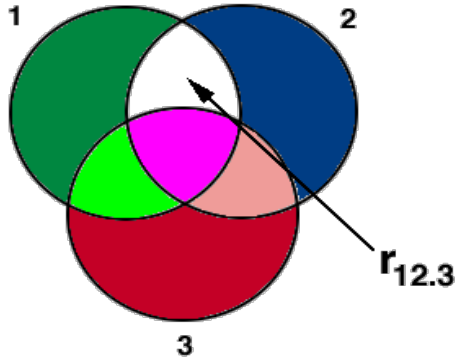
Парцијална корелација

- представља корелацију између две променљиве након уклањања утицаја треће променљиве или више других променљивих
- омогућава да се открије прави однос између посматраних појава
- најчешће се примењује на мале моделе од три до пет променљивих
- ако су $\rho_{12}, \rho_{13}, \rho_{23}$ коефицијенти корелација између променљивих X и Y, X и Z и између Y и Z, парцијални коефицијент корелације измеу X и Y без утицаја Z рачуна се на следећи начин:

$$\frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}$$







Descriptive Statistics

	Mean	Std. Deviation	N
Pocetnicka plata	6806,43	3148,255	474
Sadasnja plata	13767,83	6830,265	474
Polno-rasna podela	2,13	1,050	474

Correlations

Control Variables			Pocetnicka plata	Sadasnja plata	Polno-rasna podela
-none-[a]	Pocetnicka plata	Correlation	1,000	,880	-,496
		Significance (2-tailed)	.	,000	,000
		df	0	472	472
	Sadasnja plata	Correlation	,880	1,000	-,497
		Significance (2-tailed)	,000	.	,000
		df	472	0	472
	Polno-rasna podela	Correlation	-,496	-,497	1,000
		Significance (2-tailed)	,000	,000	.
		df	472	472	0
Polno-rasna podela	Pocetnicka plata	Correlation	1,000	,841	
		Significance (2-tailed)	.	,000	
		df	0	471	
	Sadasnja plata	Correlation	,841	1,000	
		Significance (2-tailed)	,000	.	
		df	471	0	

Процедуре за регресиону анализу

- једна од најкоришћенијих статистичких техника
- процедура за анализу везе између нумеричке зависне променљиве и једне или више независних променљивих, које су по правилу такође нумеричке
- подела регресије:
  1. проста (једна независна променљива)
     - линеарна
     - нелинеарна
  2. вишеструка (више независних променљивих)
     - линеарна
     - нелинеарна

Циљеви регресионе анализе

- испитивање да ли независна променљива (независне променљиве) објашњава знашајни део варијабилитета зависне променљиве, тј. да ли постоји веза
- одређивање који део варијабилитета зависне променљиве може бити објашњен независним променљивим, тј. јачина везе.
- одређивање структуре везе
- предвиђање вредности зависне променљиве

Вишеструки линеарни регресиони модел

- вишеструки - има више независних променљивих X
- линеарни - регресиона функција је линеарна по коефицијентима β
- регресиони - користи се регресиона функција као најбоље предвиђање за Y на основу $X_i, i = 1, ..., n$

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

- Y зависна променљива
- X_i $i-$та независна променљива
- $\beta_0, \beta_1, ..., \beta_p$ непознати параметри које треба оценити
- ϵ грешка мерења, тј. резидуал

Претпоставке за примену вишеструки линеарни регресиони модел

- опсервације су независне
- грешка има нормалну расподелу са очекивањем 0 и константном дисперзијом (хомоскедастичност)
- грешке су међусобно некорелисане
- независне променљиве не смеју бити савршено корелисане
- број података у узорку је значајно већи од броја параметара који се оцењују
- мора постојати линеарна зависност између зависне и било које независне променљиве, или групе истих

- већина претпоставки се може испитати примером одговарајућих опција у процедури за регресиону анализу
- вишеструки регресиони модел је осетљив на присуство нетипичних тачака
- оне се могу открити у прелиминарној анализи или на дијаграму стандардизованих резуидуала

Врсте вишеструког регресионог модела

стандардна све независне променљиве се уносе истовремено у модел

хијерархијска истраживач сам задаје којим редоследом се независне променљиве укључују у модел; оцењује се допринос сваке променљиве предикцији зависне променљиве тако што се уклонио утицај свих зависних променљивих које су претходно ушле у модел и оцењује се способност целог модела да предвиди зависну променљиву, као и релативни допринос сваког блока променљивих

постепена на основу статистичких критеријума програм одлучује које променљиве и којим редоследом се укључују у модел; прва променљива која се уноси у модел је она која има највећи коефицијент корелације са зависном променљивом

Вишеструки регресиони модел у SPSS-u

Enter стандардна линеарна регресија

Remove искључивање једне или више независних променљивих из стандардног модела (обрнуто од *Enter*)

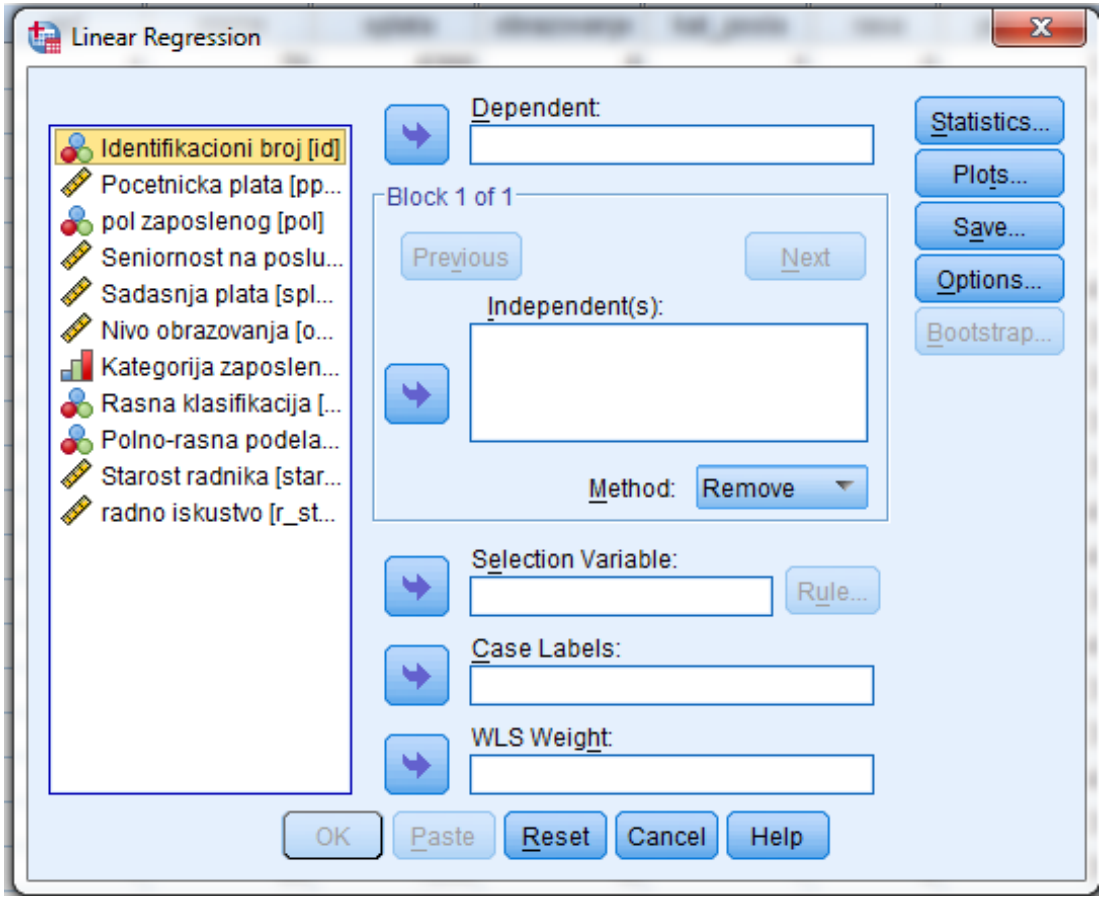
Forward постепено укључивање променљивих у модел на основу $p-$вредности $F-$теста

Backward из модела са свим независним променљивим се избацују једна по једна променљива на основу $p-$вредности $F-$теста

Stepwise постепено се додају променљиве у модел, при чему се у сваком кораку проверава $p-$вредност $F-$теста

Хијерархијска регресија се добија преко стандардне регресије груписањем независних променљивих у више слојева (*next*).

Analyze⇒Regression⇒Linear...



http://www.ats.ucla.edu/stat/spss/webbooks/reg/chapter2/spssreg2.htm

коефицијент корелације (R) показује линеарну корелацију између вредности зависне променљиве и регресиом предвиђене вредности

коефицијент детерминације (R^2) мери јачину везе између зависне и независних променљивих и представља пропорцију укупног варијабилитета зависне променљиве која се објашњава варијацијама независних променљивих: $R^2 = \frac{objasnjeni\ varijavilitet}{ukupan\ varijabilitet}$

кориговани коефицијент детерминације коефицијент детерминације коригован према броју независних променљивих и величини узорка: $\hat{R}^2 = 1 - \frac{N-1}{N-k-1}(1 - R^2)$, N величина узорка, k број независних променљивих

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Starost radnika, Pocetnicka plata, pol zaposlenog, Nivo obrazovanja, Kategorija zaposlenih[b]	.	Enter

a. Dependent Variable: Sadasnja plata

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,904[a]	,817	,815	2935,552

a. Predictors: (Constant), Starost radnika, Pocetnicka plata, pol zaposlenog, Nivo obrazovanja, Kategorija zaposlenih

F-test користи се за тестирање статистичке занчајности целокупног регресионог модела

- $H_0 : R^2 = 0 \Leftrightarrow H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0$
- $H_1 : R^2 \neq 0 \Leftrightarrow H_1 :$ нису сви $\beta_i = 0$
- тест статистика:

$$F = \frac{R^2/k}{(1-R^2)/(N-k-1)} : F_{k,N-k-1}$$

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18033665423	5	3606733085	418,538	,000[b]
	Residual	4032973847	468	8617465,485		
	Total	22066639270	473			

a. Dependent Variable: Sadasnja plata

b. Predictors: (Constant), Starost radnika, Pocetnicka plata, pol zaposlenog, Nivo obrazovanja, Kategorija zaposlenih

регресиони коефицијенти B оцењене вредности регресионих параметара

стандардизовани регресиони коефицијенти β коефицијенти добијени на основу стандардизованих података $\beta_i = B_i \frac{s_i}{s_y}$; приказују стварни утицај независних променљивих

t-test тестира хипотезу да су регресиони параметри (сваки појединачно) различити од 0:

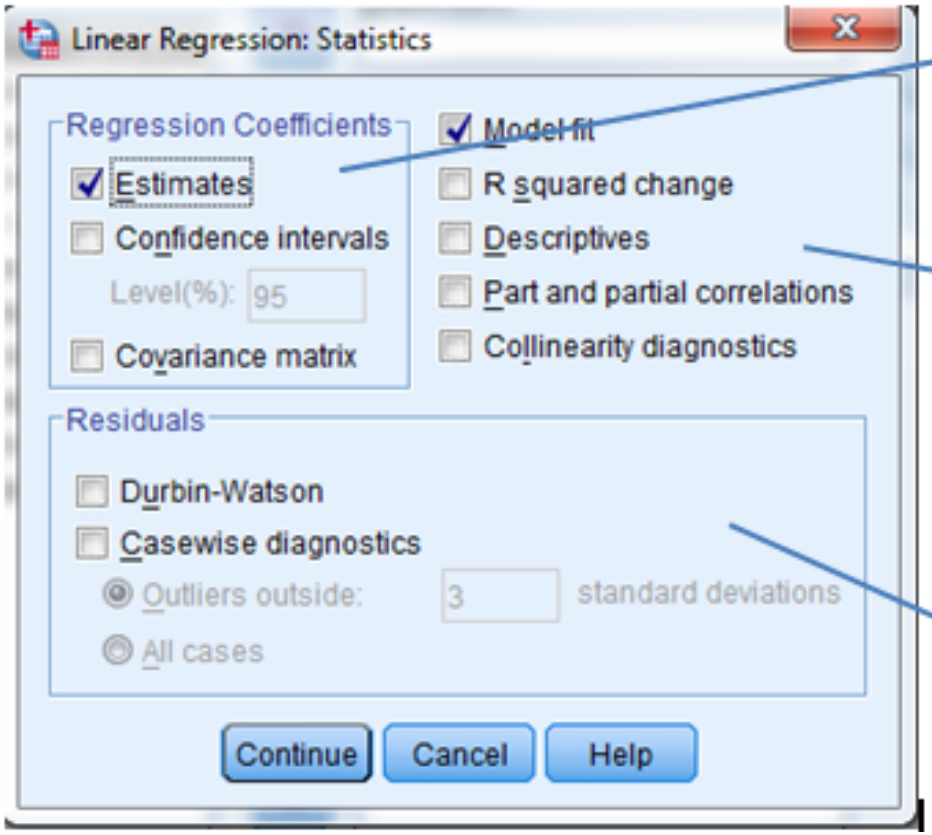
- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$
- тест статитика: $t = \frac{B}{s_B} : t_{N-k-1}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	749,075	1002,203		,747	,455
	Pocetnicka plata	1,365	,081	,629	16,815	,000
	pol zaposlenog	-728,663	306,457	-,053	-2,378	,018
	Nivo obrazovanja	295,521	64,869	,125	4,556	,000
	Kategorija zaposlenih	921,997	152,692	,190	6,038	,000
	Starost radnika	-49,172	12,331	-,085	-3,988	,000

a. Dependent Variable: Sadasnja plata

Додатне могућности



регресиони коефицијенти, интервали поверења за коефицијенте, ковариациона матрица коефицијената

дескриптивне мере, коефицијент детерминације, парцијални и семи-парцијални коефицијент корелације, статистике мултиколинеарности

Дурбан-Вотсонова статистика, нетипични резидуали, нетипичне тачке

Correlations

		Sadasnja plata	Pocetnicka plata	pol zaposlenog	Nivo obrazovanja	Kategorija zaposlenih	Starost radnika
Pearson Correlation	Sadasnja plata	1,000	,880	-,450	,661	,762	-,145
	Pocetnicka plata	,880	1,000	-,457	,633	,772	-,011
	pol zaposlenog	-,450	-,457	1,000	-,356	-,319	,051
	Nivo obrazovanja	,661	,633	-,356	1,000	,498	-,280
	Kategorija zaposlenih	,762	,772	-,319	,498	1,000	-,085
	Starost radnika	-,145	-,011	,051	-,280	-,085	1,000
Sig. (1-tailed)	Sadasnja plata	.	,000	,000	,000	,000	,001
	Pocetnicka plata	,000	.	,000	,000	,000	,409
	pol zaposlenog	,000	,000	.	,000	,000	,133
	Nivo obrazovanja	,000	,000	,000	.	,000	,000
	Kategorija zaposlenih	,000	,000	,000	,000	.	,032
	Starost radnika	,001	,409	,133	,000	,032	.
N	Sadasnja plata	474	474	474	474	474	474
	Pocetnicka plata	474	474	474	474	474	474
	pol zaposlenog	474	474	474	474	474	474
	Nivo obrazovanja	474	474	474	474	474	474
	Kategorija zaposlenih	474	474	474	474	474	474
	Starost radnika	474	474	474	474	474	474

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Pocetnicka plata	pol zaposlenog	Nivo obrazovanja	Kategorija zaposlenih	Starost radnika
1	1	4,991	1,000	,00	,00	,01	,00	,00	,00
	2	,692	2,685	,00	,01	,45	,00	,03	,00
	3	,195	5,063	,01	,00	,34	,00	,27	,12
	4	,074	8,238	,01	,02	,00	,10	,21	,49
	5	,038	11,422	,06	,84	,16	,03	,48	,01
	6	,010	21,913	,91	,13	,04	,86	,01	,38

a. Dependent Variable: Sadasnja plata

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	749,075	1002,203		,747	,455					
	Pocetnicka plata	1,365	,081	,629	16,815	,000	,880	,614	,332	,279	3,587
	pol zaposlenog	-728,663	306,457	-,053	-2,378	,018	-,450	-,109	-,047	,780	1,281
	Nivo obrazovanja	295,521	64,869	,125	4,556	,000	,661	,206	,090	,520	1,922
	Kategorija zaposlenih	921,997	152,692	,190	6,038	,000	,762	,269	,119	,396	2,528
	Starost radnika	-49,172	12,331	-,085	-3,988	,000	-,145	-,181	-,079	,861	1,161

a. Dependent Variable: Sadasnja plata

Висока корелисаних независних променљивих узрокује беспотребно укључивање неких променљивих и отежава израчунавање коефицијената.

$Tolerance_i = 1 - R_i^2$ варијанса променљиве која није везана за друге независне променљиве; шо је ближа 1 то је слабија корелисаност са другим променљивим

$VIF_i = \frac{1}{1-R_i^2}$ фактор пораста варијансе; што је ближе 1 то је мања мултиколинеарност

Condition Index ако је велика вредност (> 15) онда је висока међузависност

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,904[a]	,817	,815	2935,552	,817	418,538	5	468	,000	1,102

a. Predictors: (Constant), Starost radnika, Pocetnicka plata, pol zaposlenog, Nivo obrazovanja, Kategorija zaposlenih

b. Dependent Variable: Sadasnja plata

Дурбан-Воцонов тест је тест серијске корелације (независност грешака предвиђања).

$0 <= d < 1.3$ позитивна аутокорелација резидуала

$1.3 <= d < 1.7$ позитивна аутокорелација која може да буде проблем

$1.7 <= d < 2.3$ нема аутокорелације

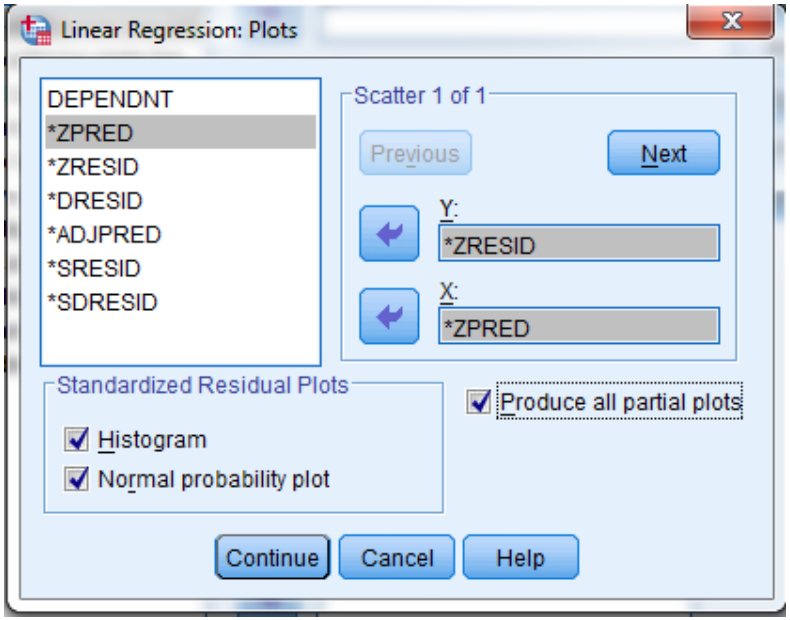
$2.3 <= d < 2.7$ негативна аутокорелација која може да буде проблем

$2.7 <= d <= 4$ негативна аутокорелација резидуала

Ако постоји аутокорелација, оцене коефицијената нису ефикасне и јављају се грешке.

http://www.math.nsysu.edu.tw/ lomn/homepage/class/92/DurbinWatsonTest.pdf

Plot



дијаграм распршености
$y = ZRESID, x = ZPRED$
приказује испуњеност претпоставки о нормалности, линеарности и хомоскедастичности резидуала

парцијални регресиони дијаграми и дијаграми стандардизованих резидуала

DEPENDNT зависна променљива

**ZPRED* стандардизована предвиђена вредност зависне променљиве

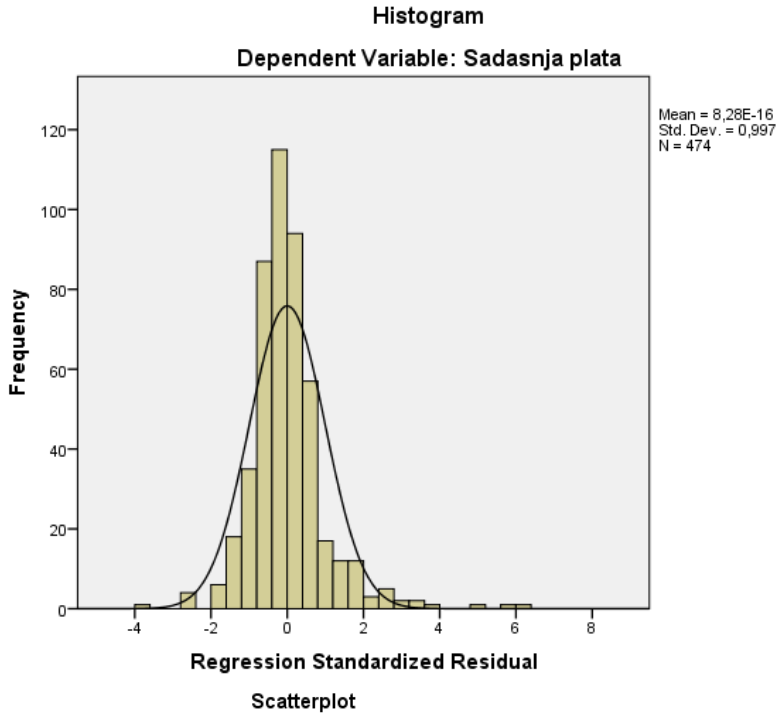
**ZRESID* стандардизовани резидуали

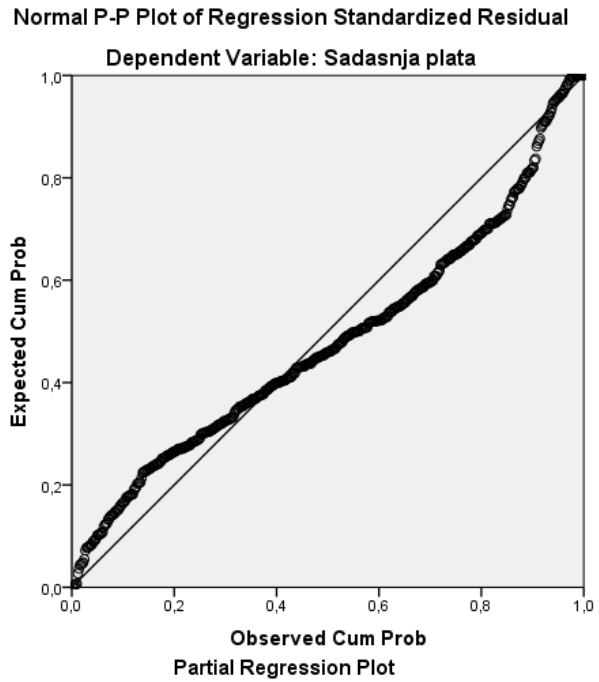
**DRESID* "избрисани" резидуали, за случај када су искључени из рачуна за регресију

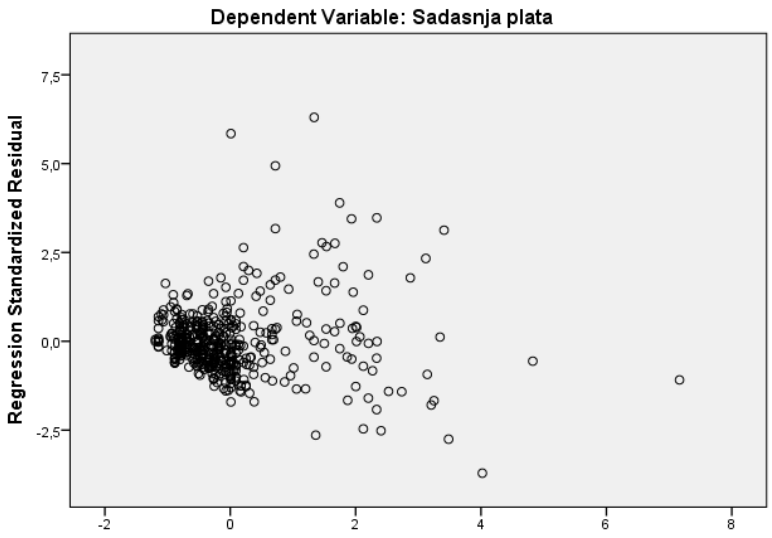
**ADJPRED* кориговане предвиђене вредности, предвиђене вредности за случај када су резидуали искључени из рачуна за регресију

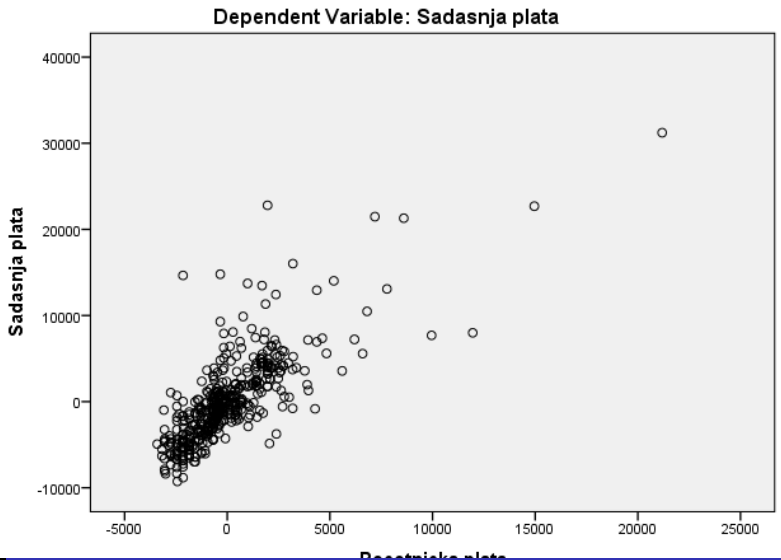
**SRESID* резидуали на основу студентове расподеле

**SDRESID* "избрисани" резидуали на основу студентове расподеле

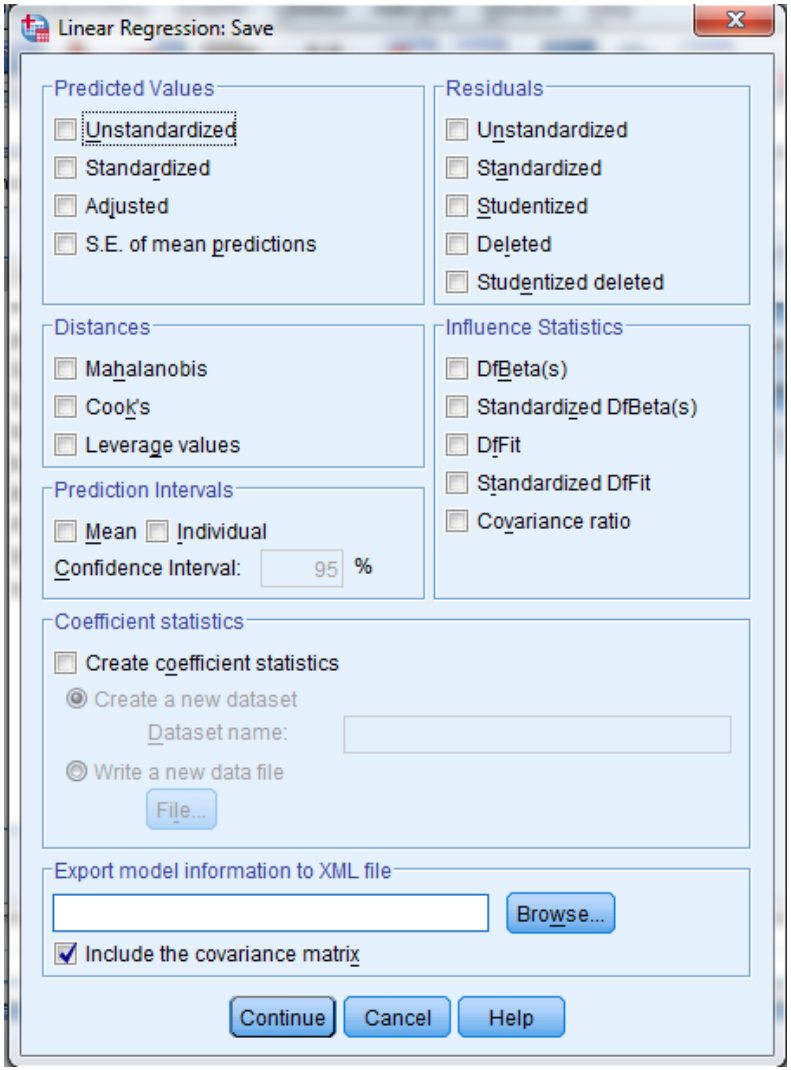

Histogram
Dependent Variable: Sadasnja plata
Mean = 8,28E-16
Std. Dev. = 0,997
N = 474
Frequency
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Sadasnja plata
Expected Cum Prob
Observed Cum Prob


Scatterplot
Dependent Variable: Sadasnja plata
Regression Standardized Residual


Partial Regression Plot
Dependent Variable: Sadasnja plata
Sadasnja plata

Save



опције за контролу облика у ком ће бити приказане предвиђене вредности, резидуали, растојања, битне статистике и интервали предвиђања

- предвиђене вредности
- резидуали:
  - *Unstandardized* разлика вредност зависне променљиве и његове предвиђена вредности
  - *Standardized* резидуали подељени проценом њихове стандардне грешке
  - *Studentized* резидуали подељени проценом њихове стандардне грешке која варира од случаја до случаја, на основу растојања посматране вредности независне променљиве од њене средње вредности
  - *Deleted* резидуали, код којих су вредности неког случаја искључене из рачуна коефицијената регресије
  - *Studentized deleted* "избрисани" резидуали подељени проценом њихове стандардне грешке
- мерење удаљености:
  - *Mahalanobis* мера разлике посматране вредности од просечне вредности читаве зависне променљиве
  - *Cook* мера колико ће се резидуали свих вредности променити ако се посматрана вредност искључи из рачуна
  - *Leverage Values* мера колико много посматрана вредност утиче на фитовање регресионог модел
- интервали поверања за предвиђене вредности
- *Influence Statistics*

Категоријске променљиве у регресији

- ако је независна променљива категоријска треба формирати вештачке променљиве
- ако променљиве има k категорија, онда се формира $k-1$ променљива

На пример:

категорије	1	2	3	4
променљива *I*	0	1	0	0
променљива *II*	0	0	1	0
променљива *III*	0	0	0	1

Нарушеност претпоставки регресије

Ако су претпоставке нарушене, може се урадити нешто од следећег:

- трансформације променљивих
- промена појединачних вредности
- избацивање појединачних вредности

⋮

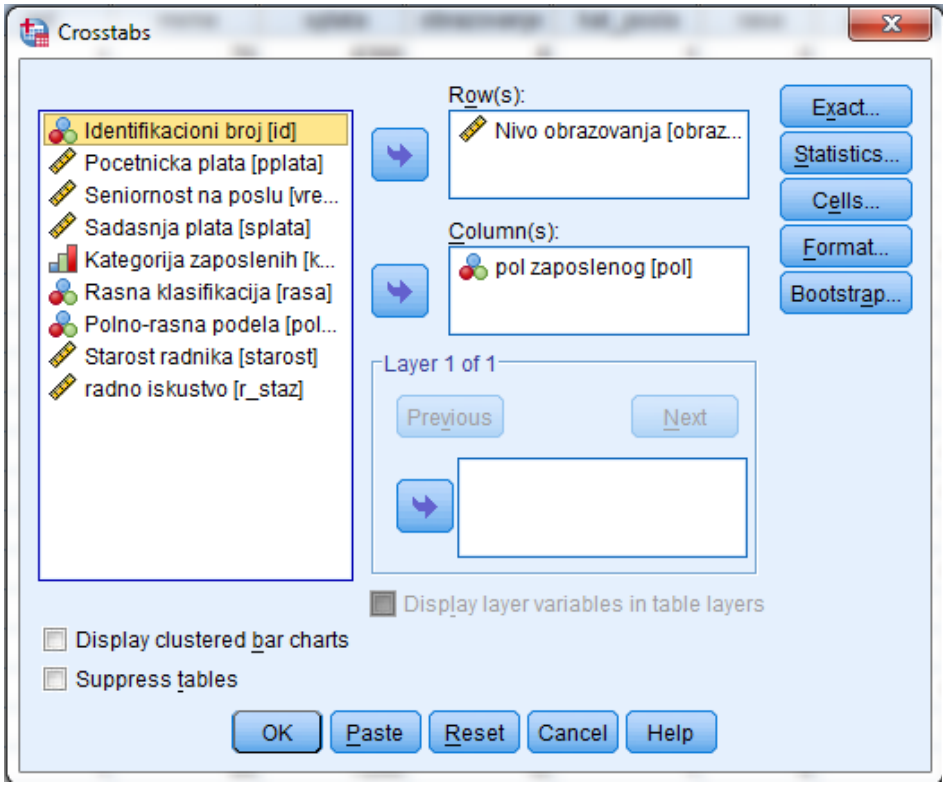
Процедуре за хи-квадрат анализу

- анализа категоријских обележја
- Хи-квадрат тест:
  - непараметарски тест
  1. тест фреквенција (касније ће бити речи о њему)
  2. тест независности
- у хи-квадрат анализи се користи:
  1. табела фреквенција
  2. унакрсно табелирање
  3. хи-квадрат тест независности

Напомена

У Пирсоновом Хи-квадрат тесту (и свим његовим варијантама) обим узорка треба да буде велики, бар толико да се обезбеди да очекиване фреквенције у свим класама буду веће од 5.

Analyze -> Descriptive Statistics -> Crosstabs...



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Nivo obrazovanja * pol zaposlenog	474	100,0%	0	0,0%	474	100,0%

Nivo obrazovanja * pol zaposlenog Crosstabulation

Count

		pol zaposlenog		
		muskarci	zene	Total
Nivo obrazovanja	8	23	30	53
	12	62	128	190
	14	6	0	6
	15	83	33	116
	16	35	24	59
	17	10	1	11
	18	9	0	9
	19	27	0	27
	20	2	0	2
	21	1	0	1
Total		258	216	474

Хи-квадрат тест независности

- испитивање да ли су два обележја X и Y независна
- $H_0 : F_{X,Y} = F_X F_Y$
- $H_1 : F_{X,Y} \neq F_X F_Y$
- одговор треба дати на основу дводимензионалног простог узорка обима $n : (x_i, y_i), i = 1, 2, , n$
- подаци се дају у табели контигенције са r категорија обележја X и s категорија по вредностима обележја Y
- n_{ij} је број појављивања пара (x_i, y_i) у узорку
- тест статистике:
$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(nn_{ij} - n(x_i)n(y_j))^2}{nn(x_i)n(y_j)} : \chi^2_{(r-1)(s-1)} \text{ при } H_0$$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	96,856[a]	9	,000
Likelihood Ratio	115,880	9	,000
Linear-by-Linear Association	59,941	1	,000
N of Valid Cases	474		

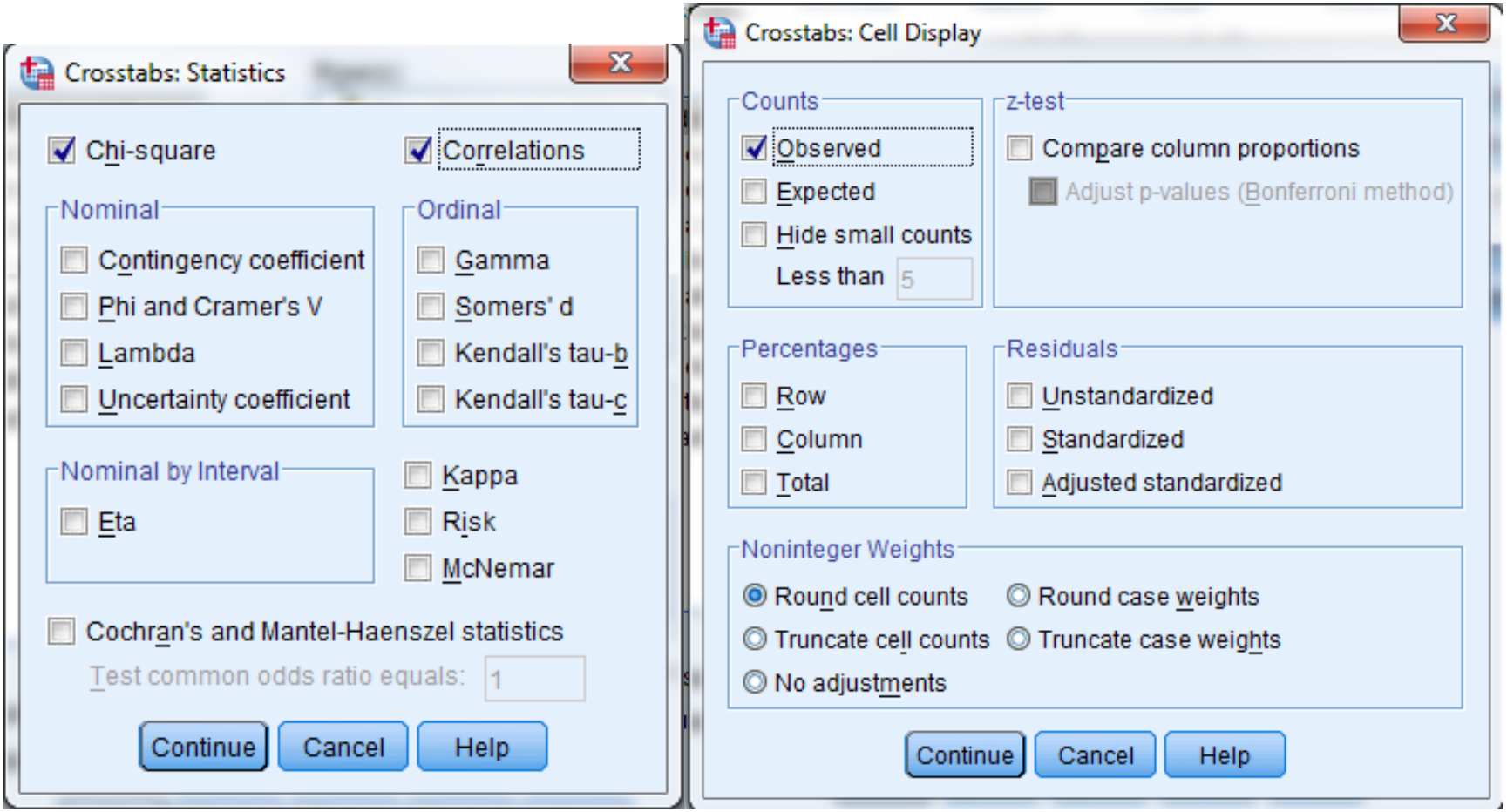
a. 8 cells (40,0%) have expected count less than 5. The minimum expected count is ,46.

Након прекодирања:

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	94,709[a]	4	,000
Likelihood Ratio	108,879	4	,000
Linear-by-Linear Association	64,187	1	,000
N of Valid Cases	474		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 22,78.


Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1
Continue
Cancel
Help
Crosstabs: Cell Display
Counts
Observed
Expected
Hide small counts
Less than 5
z-test
Compare column proportions
Adjust p-values (Bonferroni method)
Percentages
Row
Column
Total
Residuals
Unstandardized
Standardized
Adjusted standardized
Noninteger Weights
Round cell counts
Round case weights
Truncate cell counts
Truncate case weights
No adjustments
Continue
Cancel
Help

Јачина везе

- хи-квадрат тестом се утврђује да ли постоји или не веза између променљивих
- ако се закључи да постоји веза (одбаци се H_0) како одредити јачину везе?
  1. коефицијент контигенције
  2. Крамеров V коефицијент
  3. Фи-коефицијент

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,447	,000
	Cramer's V	,447	,000
	Contingency Coefficient	,408	,000
N of Valid Cases		474	

$\Rightarrow$умерена веза

Садржај

Пример 1

Да ли постоји веза (корелација) између променљивих *ppslata* и *splata*? Да ли се јачина везе промени ако се искључи утицај променљиве пол?

Пример 1

Да ли постоји веза (корелација) између променљивих *ppslata* и *splata*? Да ли се јачина везе промени ако се искључи утицај променљиве пол?

- Analyze⇒ Correlate⇒ Bivariate...
- Analyze⇒ Correlate⇒ Partial...

Correlations

			Pocetnicka plata	Sadasnja plata
Kendall's tau_b	Pocetnicka plata	Correlation Coefficient	1,000	,656**
		Sig. (2-tailed)	.	,000
		N	474	474
	Sadasnja plata	Correlation Coefficient	,656**	1,000
		Sig. (2-tailed)	,000	.
		N	474	474
Spearman's rho	Pocetnicka plata	Correlation Coefficient	1,000	,826**
		Sig. (2-tailed)	.	,000
		N	474	474
	Sadasnja plata	Correlation Coefficient	,826**	1,000
		Sig. (2-tailed)	,000	.
		N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

Control Variables			Pocetnicka plata	Sadasnja plata
pol zaposlenog	Pocetnicka plata	Correlation	1,000	,849
		Significance (2-tailed)	.	,000
		df	0	471
	Sadasnja plata	Correlation	,849	1,000
		Significance (2-tailed)	,000	.
		df	471	0

Пример 2

Да ли су пол и категорија посла повезани? Ако је одговор потврдан, одредити јачину те везе?

Пример 2

Да ли су пол и категорија посла повезани? Ако је одговор потврдан, одредити јачину те везе?

- Analyze⇒ Descriptive Statistics⇒ Crosstabs...

pol zaposlenog * Kategorija zaposlenih Crosstabulation

Count

		Kategorija zaposlenih							Total
		sluzbenik	obuceni sluzbenik	sluzbenik obezbedjenja	sluzbenik sa fakult. diplomom	posebna grupa sluzbenika	MBA diploma	tehnicka sluzba	
pol zaposlenog	muskarci	110	47	27	34	30	4	6	258
	zene	117	89	0	7	2	1	0	216
Total		227	136	27	41	32	5	6	474

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	87,230[a]	6	,000
Likelihood Ratio	106,109	6	,000
Linear-by-Linear Association	48,161	1	,000
N of Valid Cases	474		

a. 4 cells (28,6%) have expected count less than 5. The minimum expected count is 2,28.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,429	,000
	Cramer's V	,429	,000
N of Valid Cases		474	

Пример 3

Наћи најбољи модел за одређивање вредности почетне плате преко осталих променљивих.

Пример 3

Наћи најбољи модел за одређивање вредности почетне плате преко осталих променљивих.

- Analyze⇒ Regression⇒ Linear...

Хвала на пажњи!