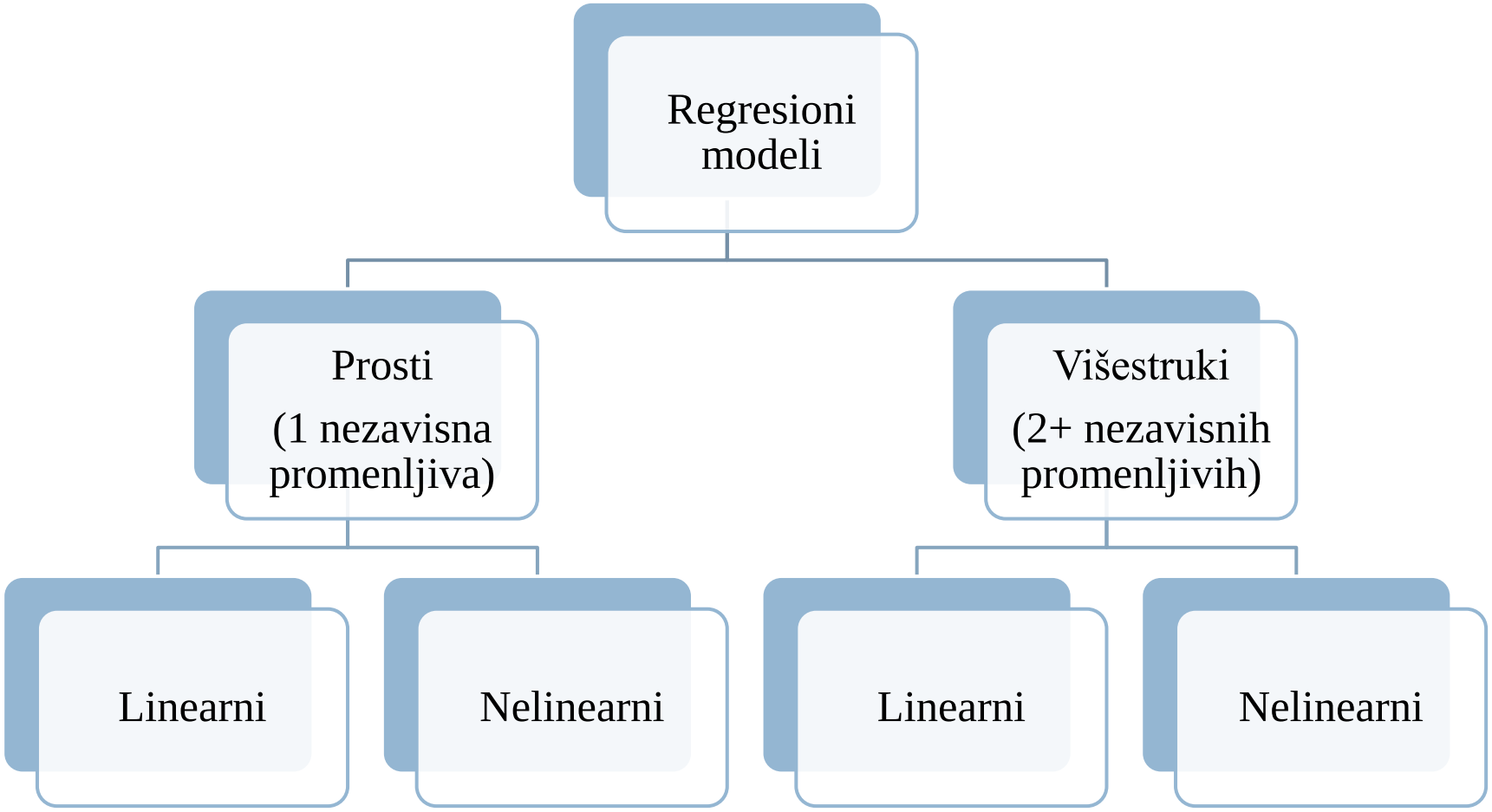
VIŠESTRUKA LINEARNA REGRESIJA

Malo podsećanje



- Naziv *višestruka linearna regresija* znači:

- Višestruka - ima više nezavisnih promenljivih X
- Linearna - regresiona funkcija je linearna po koeficijentima β
- Regresija - koristi se regresiona funkcija kao najbolje predviđanje za Y na osnovu Xi, i=1,..,n

Regresiona analiza

~ Ukoliko se problem koji posmatramo može tretirati kao problem jedne zavisne i više nezavisnih promenljivih, radi se o pogodnoj situaciji za analizu podataka metodom višestruke regresije. Ako je veza između njih linearna, slučaj se svodi na višestruki linearni model.

~ Neka su:

Y zavisna promenljiva

$X_1, X_2, ..., Xp$ nezavisne promenljive

~ Tada je linearni model:

$Y = \beta_0 + \beta_1 X_1 + ... + \beta p Xp + \varepsilon$

Pri čemu su $\beta_0, \beta_1, ..., \beta p$ nepoznati parametri koje treba oceniti, a ε greška merenja, tj. reziduali.

Promenljivu Y zovemo i promenljiva odgovora, tj. output promenljiva, dok su X-promenljive zvane input, tj. objašnjavajuće promenljive.

Ukoliko imamo n eksperimenata model možemo zapisati matrično na sledeći način:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad , i = 1,\ldots,n$$

Uvodimo oznaku:

$$\beta = \begin{bmatrix} \beta_0 \\ \ldots \\ \beta_p \end{bmatrix}$$

Smatra se da je $E\varepsilon_i = 0$, $i = 1,\ldots,n$, jer tu vrednost uvek možemo sračunati sa β_0.
Konačno dobijamo:

$$\begin{bmatrix} y_1 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \ldots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

Kraći zapis je:

$Y = X\beta + \varepsilon$

gde se $X\beta$ zove sistemska komponenta modela,

ε slučajna komponenta modela.

Ocena parametara

Jedan od kriterijuma ocene je metoda najmanjih kvadrata. Metod se sastoji u tome da se za ocenu parametra $\hat{\beta}$ uzima ona vrednost za koju je zbir kvadrata reziduala $\sum_{i=1}^{n} \varepsilon_i^2$ minimalan.

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T * \varepsilon = (Y - X\beta)^T (Y - X\beta)$$

$$\frac{\partial S}{\partial \beta} = 0$$

S obzirom da tražimo minimum sume, izvršili smo diferenciranje. Međutim, to je upravo sledeće diferenciranje po vektoru:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

Gde je f realna funkcija od X.

$$\frac{\partial f}{\partial X} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

Osobine

1. $\frac{\partial(A^T X)}{\partial X} = A$ Gde je $A = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix}$ $f(X) = A^T X = \sum_{i=1}^{n} a_i x_i$

2. $\frac{\partial(X^T A)}{\partial X} = A$

3. $\frac{\partial(X^T X)}{\partial X} = 2X$ $f(X) = X^T X = \sum_{i=1}^{n} x_i^2$

4. Ako je A simetrična matrica, onda: $\frac{\partial(X^T AX)}{\partial X} = 2AX$

   Ako A nije simetrična matrica, onda: $\frac{\partial(X^T AX)}{\partial X} = (A + A^T)X$

Nakon što smo primenili prethodne osobine, dobijamo da je:

$$\frac{\partial S}{\partial \beta} = -X^T Y - Y^T X + 2X^T X\beta = 0$$

$$2X^T X\beta = X^T Y + Y^T X = 2X^T Y$$

$$X^T X\beta = X^T Y$$

Ako postoji: $(X^T X)^{-1}$

Onda je konačno ocena za β upravo: $\hat{\beta} = (X^T X)^{-1} X^T Y$

~ Ocena sistematske komponente Xβ: $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY$

~ Ocena greške: $\hat{\varepsilon} = Y - X\hat{\beta} = (E - H)Y$

Ciljevi regresione analize

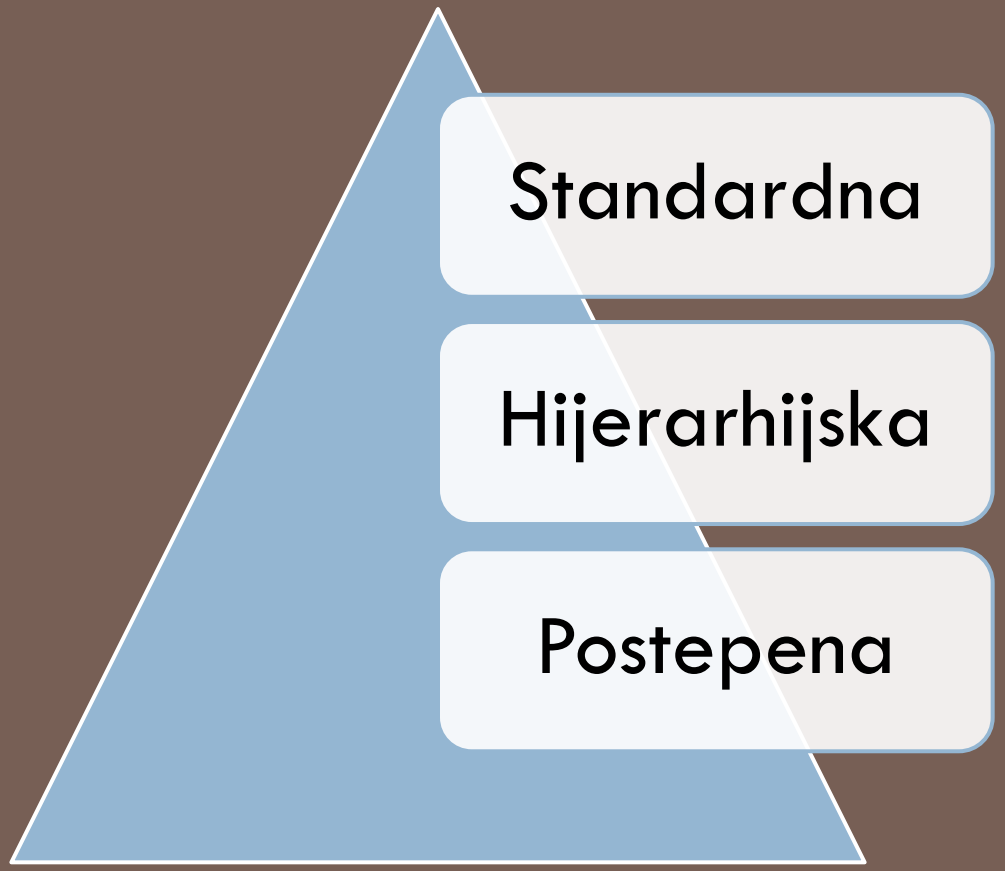
~ Ispitivanje da li nezavisna promenljiva (nezavisne promenljive) objašnjavaju značajni deo varijabiliteta zavisne promenljive, tj. da li postoji veza.

~ Odrediti koji deo varijabiliteta zavisne promenljive može biti objašnjen nezavisnim promenljivim, tj. jačina veze.

~ Odrediti strukturu veze.

~ Predvideti vrednosti zavisne promenljive.

Glavne vrste višestruke regresije



~ Standardna – Kod njega se sve nezavisne promenljive unose istovremeno u model.

~ Hijerarhijski linearni modeli (ili višestepena regresija) organizuje podatke u hijerarhiji regresija. Na primer , gde je A regresovano od B , a B je u zavisnosti od C. Često se koristi kada podaci imaju prirodnu hijerarhijsku strukturu , kao što je u slučajevima statistika vezanih za obrazovanje, gde su učenici grupisani u učionicama , učionice po školama , a škole su uklopljene u nekoj upravnoj jedinici, kao što je školski okrug ili grad. Promenljiva odgovora može biti merilo postignuća učenika za rezultat testova, pa različite "podpromenljive" će prikupljati na nivoima učionica, škola i na posmatranom području. Kod ovog modela istraživač sam zadaje kojim redosledom se nezavisne promenljive uključuju u model.

~ Postepena regresija (sekvencijalno ispitivanje) je statistički metod u kom veličina baze nije fiksna unapred. Umesto toga podaci se obrađuju kako se prikupljaju i dalje ispitivanje se zaustavlja u skladu s unapred definisanim pravilom, čim se dosegne značajni rezultat. Dakle, zaključak može ponekad biti postignut u mnogo ranijoj fazi nego što bi to bilo moguće ostalim tipovima regresije. Kod ovog modela se na osnovu statističkih kriterijuma odlučuje koje promenljive i kojim redosledom se uključuju u model.

ANALYZE > REGRESSION > LINEAR...

Primer višestruke linearne regresije - SPSS

Opis korišćene baze:

~ Baza se sastoji iz 6 promenljivih

1. VO2max – Maksimalni kapacitet izdržljivosti tokom vežbanja, indikator za fitnes
2. age - Godine ispitanika
3. weight - Težina ispitanika
4. heart_rate - Otkucaji srca
5. gender – Pol ispitanika
6. caseno – Broj ispitanika

~ Promenljiva caseno nam služi za brisanje autlajera, ako na njih naiđemo tokom ispitivanja. Način na koji vršimo brisanje se vrši postavljanjem početnih pretpostavki.

Cilj ispitivanja:

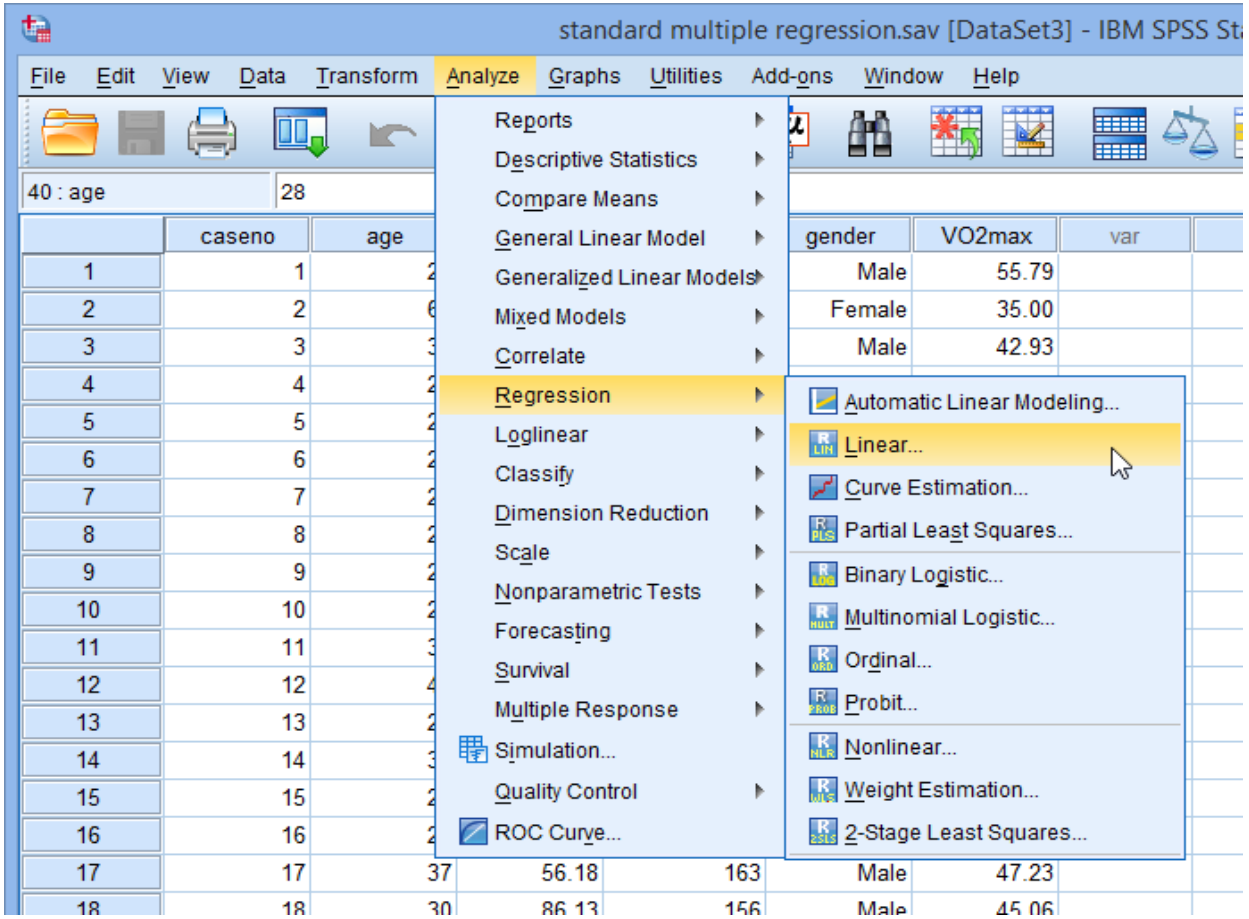
~ Želja je predvideti promenljivu VO2max, indikator zdravlja i fitnes sposobnosti. S obzirom da bi za ovo istraživanje bilo potrebno mnogo novca i laboratorijske opreme gde bi ispitanici vežbali do granice iscrpljenja, koja čak može biti opasna po zdravlje, ovu promenljivu ćemo predvideti na jeftiniji način. U istraživanju je učestvovalo 100 učesnika, kojima su zabeležene vrednosti date u bazi.

Potrebne pretpostavke:

1. Vrednosti promenljivih treba da se nalaze u nekom kontinualnom opsegu.
2. Potrebne su 2 ili više promenljivih.
3. Opservacije moraju biti nezavisne.
4. Mora postojati linearna zavisnost između zavisne i bilo koje nezavisne promenljive, ili grupe istih.
5. Podaci moraju biti slični kako se krećemo kroz bazu – homogenost.
6. Baza ne sme imati više međusobno zavisnih promenljivih.
7. Baza ne treba da ima previše autlajera, vrednosti koje mnogo iskaču ili tačke koje čak i previše utiču na istraživanje.
8. Reziduali (greške) treba da budu približno normalno raspodeljeni.

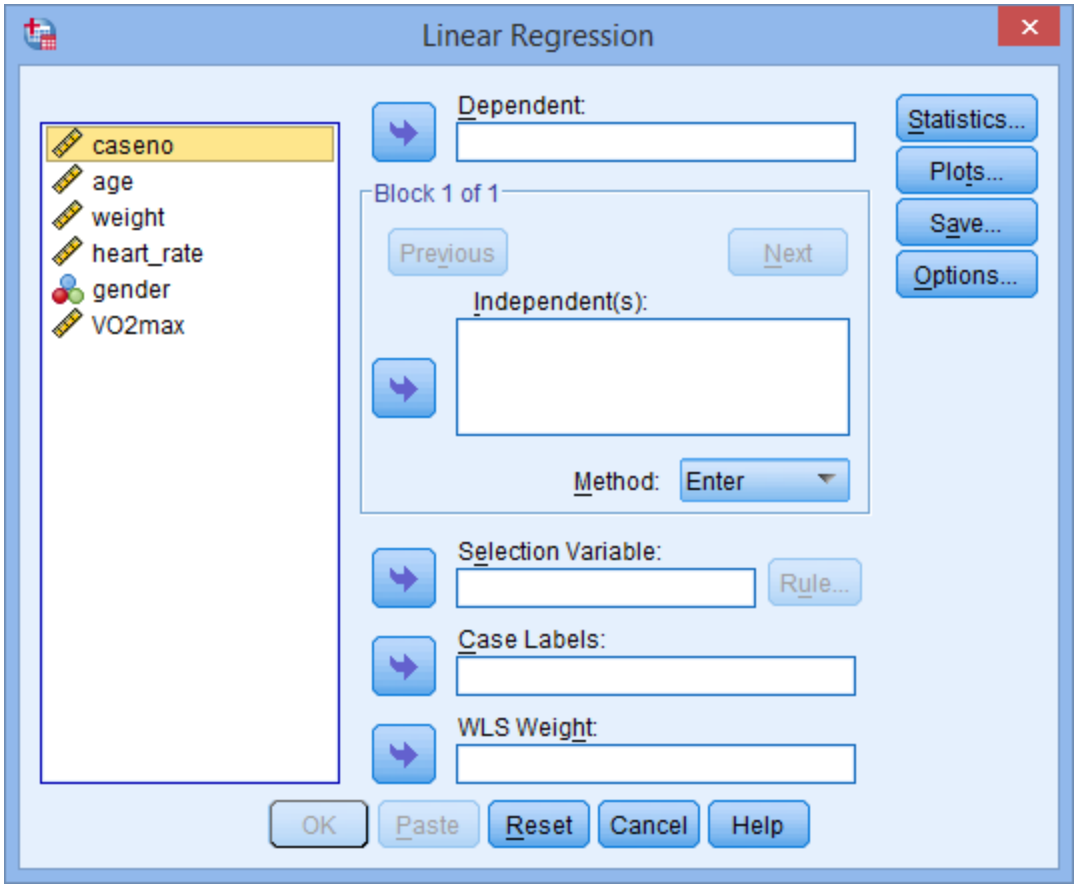
Postupak regresije:

1.



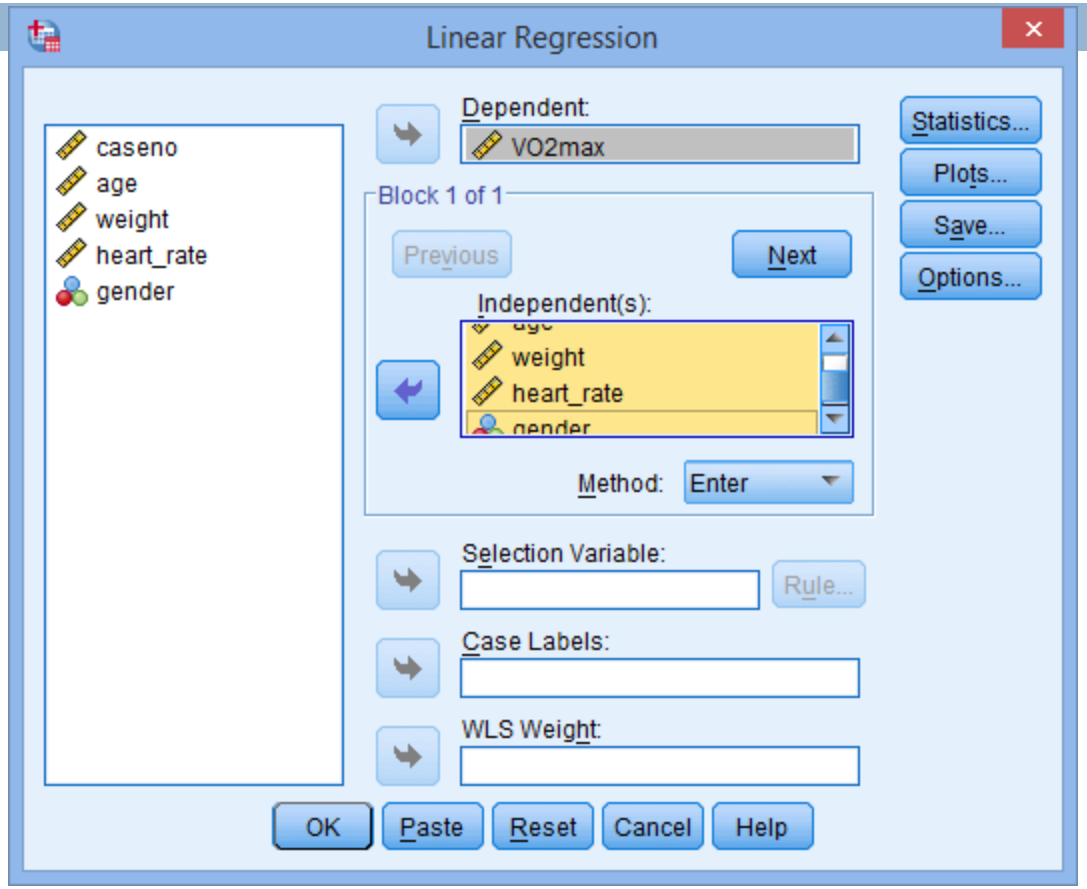
U ovom delu nam se otvara prozor gde sa leve strane vidimo prikaz promenljivih, a sa desne izbor za nezavisne i zavisnu promenljivu.

2.



S obzirom da želimo da predvidimo promenljivu VO2max, nju biramo za zavisnu, a u opciji za zavisne stavljamo sve ostale promenljive, kao što je prikazano na slici.

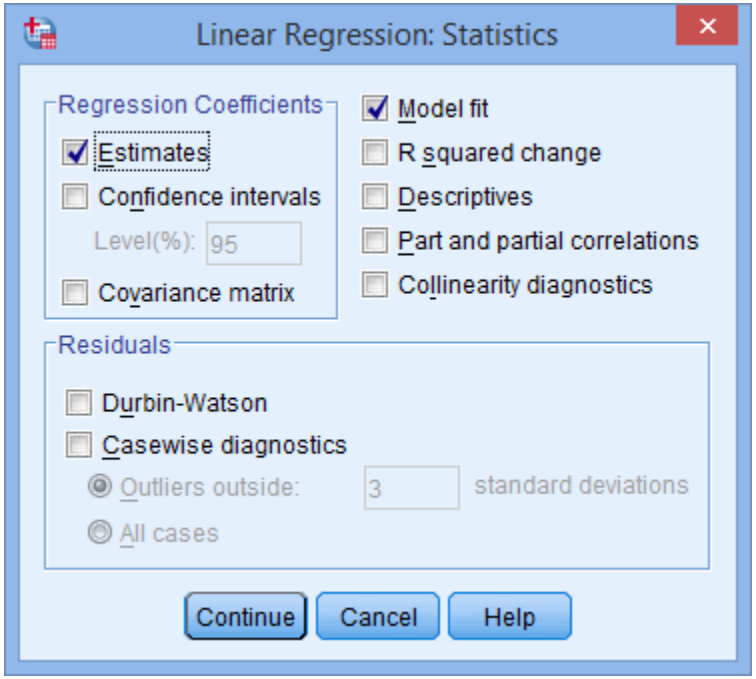
3.



~ Napomena1: Primer ilustruje standardnu višestruku regresiju, tako da dugmad "Previous" i "Next" ignorišemo, jer oni služe za postepenu i hijerarhijsku regresiju.
~ Napomena2: Ako iz bilo kog razloga, metod "Enter" nije selektovan, potrebno je vratiti se na njega, jer je to SPSS ugrađen metod upravo za standardnu regresiju.

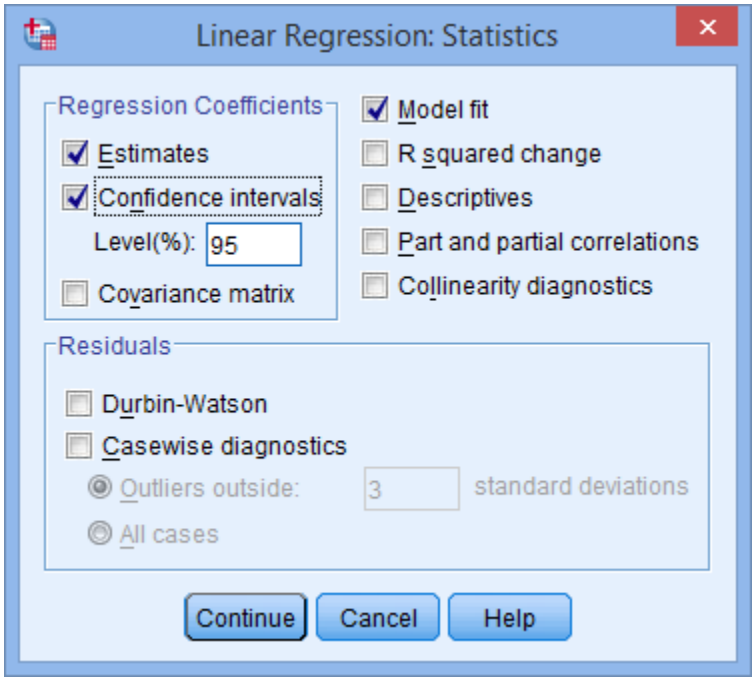
Pritiskom na dugme “Statistics”, otvara nam se prozor sa slike ispod. Određeni check-box- ovi su već selektovani, obzirom da želimo da procenimo vrednost, tj. da fitujemo model, ostavićemo tako kako već i jeste.

4.



U ovom delu ćemo još selektovati i check-box za intervale poverenje i ostaviti 95%-ni interval. Sa desne strane prozora možemo uključiti i opciju parcijalne korelacije, kao i dijagnostiku za međusobnu zavisnost promenljvih, koju smo pomenuli u pretpostavkama.

5.



Takođe smo u pretpostavkama spomenuli i reziduale, i u ovom delu možemo primeniti Durbin-Watson- ov test, koji nam govori o nezavisnosti istih. I ostalu dijagnostiku, npr. za autlajere, možemo ovde selektovati.

6.

- Klikom na dugme “Continue” vraćamo se na prozor Linear Regression.

7.

- Nakon toga klikom na dugme “Ok” dobijamo output vrednost.

Analiza rezultata i izveštaj

- SPSS će generisati na izlazu nekoliko tabela. Od njih ćemo prokomentarisati 3 koje su nam potrebne za interpretiranje rezultata podrazumevajući da su pretpostavke zadovoljene.

1. Provera da li je model dobar:

~ Sledeća tabela, Model Summary, se sastoji iz 4 vrednosti:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.559	5.69097

a. Predictors: (Constant), gender, age, heart_rate, weight

Kolona "R" predstavlja vrednost koeficijenta višestruke korelacije. On služi da bi se odredio kvalitet predviđanja zavisne promenljive, u ovom slučaju VO2max. Vrednost 0,76 predstavlja dobar nivo predviđanja.

Kolona "R Square" predstavlja koeficijent odlučivanja, tj. proporciju disperzije zavisne promenljive koja se može objasniti nezavisnom. Naša vrednost 0,577 predstavlja 57,7% varijabiliteta zavisne promenljive koji može biti objašnjen nezavisnim promenljivima, tako da je jačina veze jaka.

2. Značajnost testa:

F-vrednost u ANOVA tabeli, koja je prikazana ispod, testira da li je regresioni model dobar za ove vrednosti. Tabela pokazuje da nezavisne promenljive dobro statistički predviđaju zavisnu promenljivu. Drugim rečima, regresioni model je dobar.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4196.483	**4**	1049.121	**32.393**	**.000[b]**
	Residual	3076.778	**95**	32.387		
	Total	7273.261	99			

a. Dependent Variable: VO2max

b. Predictors: (Constant), gender, age, heart_rate, weight

$F(4,95) = 32,392$, $p<0.05$.

3. Ocene koeficijenata modela:

$VO_2max = 87.83 – (0.165 \times age) – (0.385 \times weight) – (0.118 \times heart\_rate) + (13.208 \times gender)$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**87.830**	6.385		13.756	.000	75.155	100.506
	age	**-.165**	.063	-.176	-2.633	.010	-.290	-.041
	weight	**-.385**	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	**-.118**	.032	-.252	-3.667	.000	-.182	-.054
	gender	**13.208**	1.344	.748	9.824	.000	10.539	15.877

a. Dependent Variable: VO2max

Iz tabele vidimo da je koeficijent za godine negativan, što nam govori da je regresija obrnuta, odnosno sa većim brojem godina, slabija joj je fitnes sposobnost. Isto važi i za telesnu težinu i otkucaje srca.

4. Značaj nezavisnih promenljivih u testu:

Testom možemo proveriti koliko je značajna svaka nezavisna promenljiva u našem modelu. Ako je p <0,05, zaključujemo da je koeficijent statistički značajno različit od 0, tj. da je odgovarajuća promenljiva potrebna u istraživanju.

Vrednost statistike i odgovarajuća p-vrednost se nalaze u kolonama “t” i “”Sig.”, respektivno, kao što je obojeno u sledećoj tabeli.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	87.830	6.385		**13.756**	**.000**	75.155	100.506
	age	-.165	.063	-.176	**-2.633**	**.010**	-.290	-.041
	weight	-.385	.043	-.677	**-8.877**	**.000**	-.471	-.299
	heart_rate	-.118	.032	-.252	**-3.667**	**.000**	-.182	-.054
	gender	13.208	1.344	.748	**9.824**	**.000**	10.539	15.877

a. Dependent Variable: VO2max

Zaključak: Sve promenljive su bitne!

Primer vrednosti predviđene modelom

~ Na osnovu jednačine linearne regresije koju nam je model obezbedio možemo za bilo koje proizvoljne vrednosti nezavisnih promenljivih predvideti, odnosno izračunati vrednost zavisne promenljive.
Npr. Osoba A, ženskog pola, ima 23 godine, telesnu masu 140lbs i 57 otkucaja srca po minuti.

VO_2max =
= 87.83 – (0.165 x age) – (0.385 x weight) – (0.118 x heart_rate) + (13.208 x gender) =
= 87.83 – (0.165 x 23) – (0.385 x 140) – (0.118 x 57) + (13.208 x 1) = 36,617

Možemo zaključiti da osoba A ima niži koeficijent zdravlja i fitnesa.

Ostale metode linearne regresije

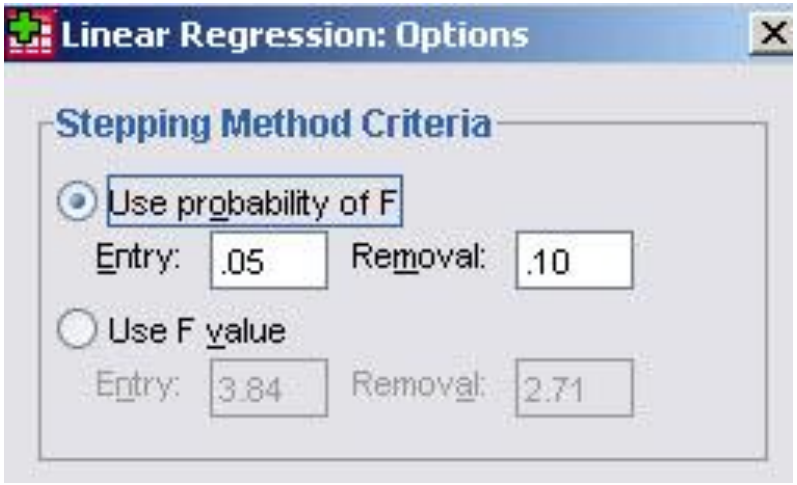
Metode zapravo kontrolišu način na koji se promenljive uključuju u proces regresije. Veoma često znamo koje od promenljivih želimo da uključimo u regresiju i tada ćemo koristiti prethodno opisan model Enter, koji je ujedno i osnovni.

Osim ovog metoda, nekoliko drugih metoda je dostupno za izgradnju modela i oni kontrolišu koje će promenljive i na koji način biti ubačene. Postoji mogućnost i kombinovanja metoda. Glavni cilj je odrediti najbolji podskup promenljivih koje objašnjavaju zavisnu promenljivu.

Spisak metoda:
~Enter
~Stepwise
~Backward
~Forward
~Remove (Sve promenljive u bloku se istovremeno uklanjaju.)

Stepwise methods - Postepena metoda

Stepwise metode uključuju ili uklanjaju jednu nezavisnu promenljivu na svakom koraku, temeljeno (po defaultu) na p-vrednosti (verovatnoća od F). Alternativno, može se koristiti i vrednosti F umesto njegove verovatnoće. Ograničenja za kriterijume koji kontrolišu uključivanje ili uklanjanje promenljive mogu se dodatno precizirati kao F-to-enter/F-to-remove. Ovo možemo promeniti u prozoru Options, kao što se vidi na sledećoj slici.



Na raspolaganju su nam tri metode za postepenu regresiju:

~ Stepwise – Na osnovu p-vrednosti od F, SPSS počinje uključivanjem promenljive sa najmanjom p-vrednosti, u sledećem koraku ubacuje promenljivu s najmanjom p-vrednosti za F iz preostalog skupa promenljivih i tako dalje. Promenljive koje su već u jednačini zavisnosti se uklanjaju ako im p-vrednost postane veća od zadate granice zbog uključivanja druge promenljive. Postupak se završava kada nema više promenljivih koje su podobne za uključivanje ili uklanjanje. Ova metoda temelji se na obe: verovatnoće za unos (PIN) i verovatnoće za uklanjanje (POUT) (ili alternativno FIN i FOUT).

~ Backward - Eliminacija: Prvo sve promenljive ulaze u jednačinu, a zatim se redom uklanjaju. Za svaki korak SPSS nudi statistiku, pod nazivom R^2. Na svakom koraku, najveća verovatnoća za F se uklanja (ako je vrednost veća od POUT). Alternativno, FOUT može biti navedeno kao kriterijum.

~Forward – Odabir unapred: Na svakom koraku promenljive koje još nisu u jednačini, a imaju najmanju p-vrednost za F se dodaju, ali pod uslovom da je ta vrednost manja od PIN. Alternativno, koristi se vrednost F postavljanjem FIN na /CRITERIA. Postupak se zaustavlja kada više nema promenljive koja zadovoljavaja kriterijum za ulaz.

Kartica Plot

Kartica Plot služi za crtanje grafika u SPSS-u. Sa leve strane otvorenog prozora na slici nalaze se opcije za grafik predviđenih vrednosti i reziduala:

~DEPENDNT - zavisna promenljiva.

~*ZPRED - standardizovana predviđena vrednost zavisne promenljive.

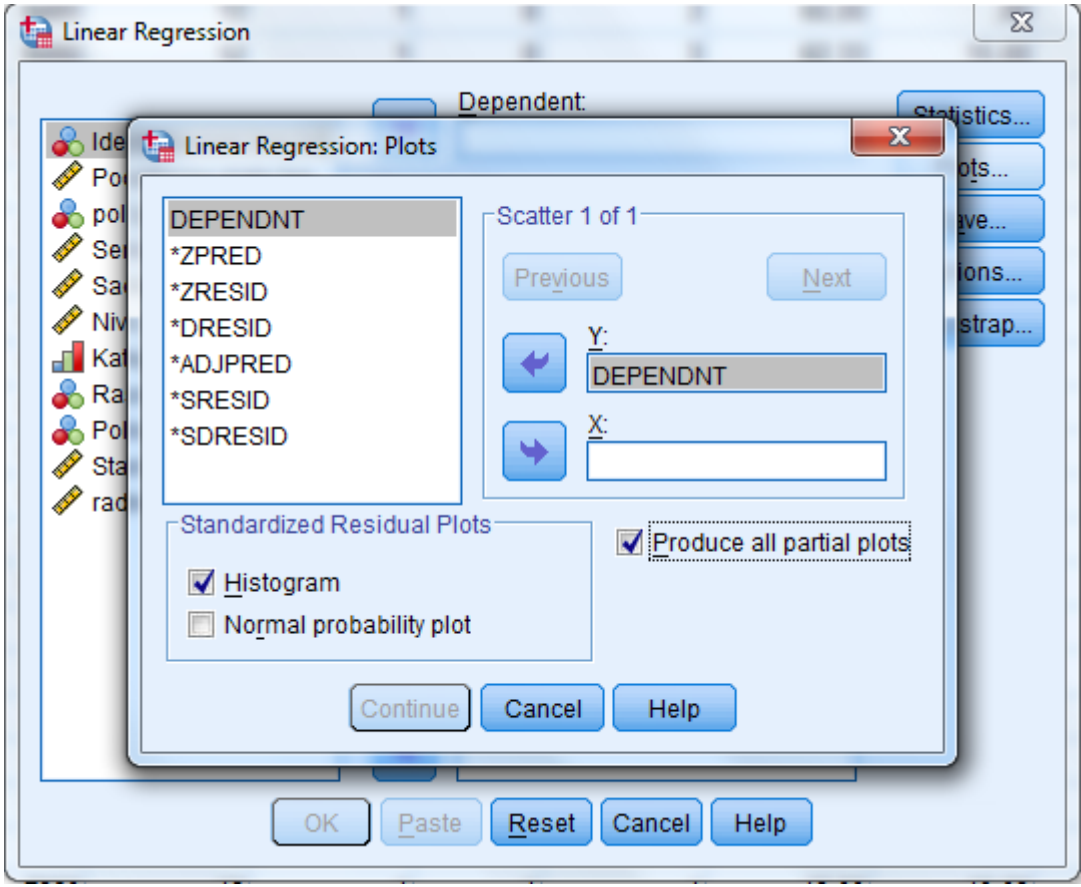
~*ZRESID - standardizovani reziduali.

~*DRESID - "izbrisani" reziduali, za slučaj kada su isključeni iz računa za regresiju.

~*ADJPRED – korigovane predviđene vrednosti, predviđene vrednosti za slučaj kada su reziduali isključeni iz računa za regresiju.

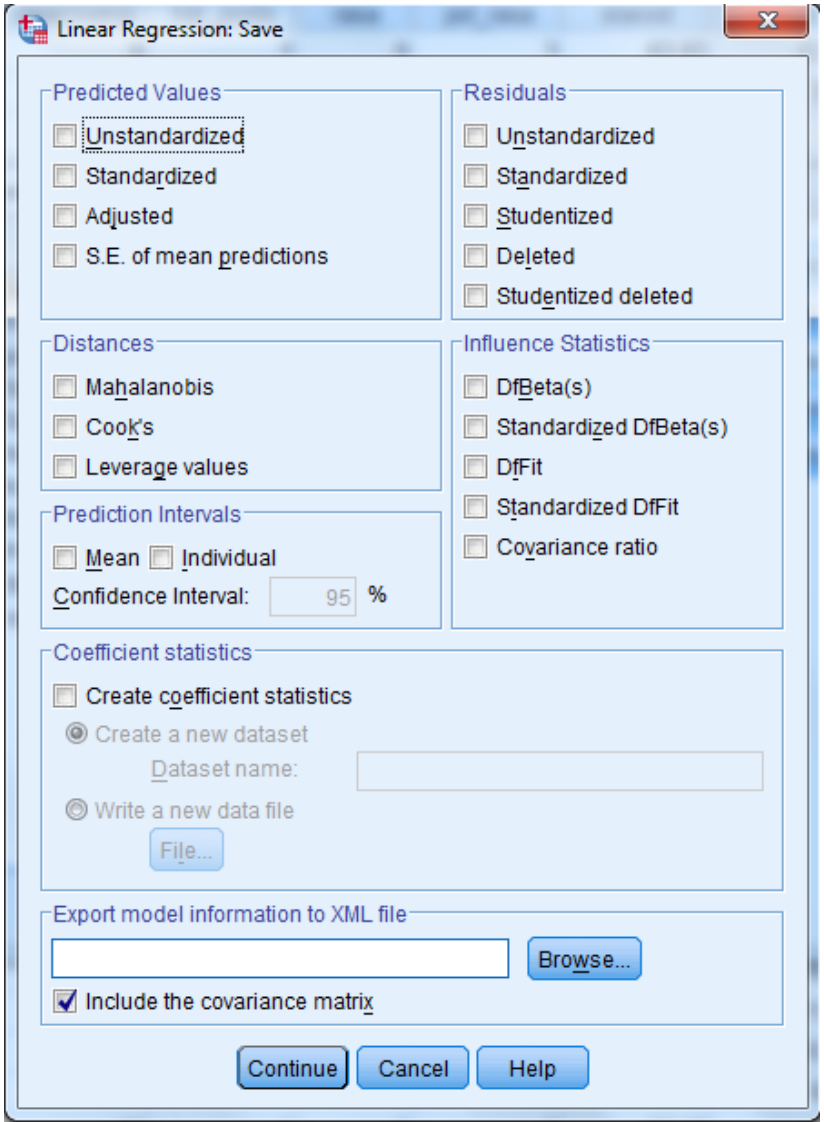
~*SRESID – reziduali na osnovu studentove raspodele.

~*SDRESID – "izbrisani" reziduali na osnovu studentove raspodele.



~ Sa desne strane prethodnog prozora biramo promenljive za X i Y osu željenog grafika.
~ Konačno, pruža nam se i mogućnost izbora vrste grafika: Histogram ili Normal probability plot (pruža nam uvid u bliskost/odsutpanje u odnosu na normalnu raspodelu).
~ Vrlo poželjna mogućnost je čekirati “Produce all partial plots”, što nam daje uvid u više pojedinačnih grafika po željenim uslovima/promenljivima.

Kartica Save



U ovoj kartici nude nam se opcije za kontrolu oblika u kom će nam se prikazati predviđene vrednosti, reziduali, rastojanja, bitne statistike i intervali predviđanja.
U nastavku će biti objašnjeno značenje svake ponuđene opcije.

Neki termini u SPSS-u

Merenje udaljenosti (razlika)

1.Mahalanobis: Mera razlike posmatrane vrednosti od prosečne vrednosti čitave zavisne promenljive.

2.Cook je: Mera koliko će se reziduali svih vrednosti promeniti ako se posmatrana vrednost isključi iz računa.

3. Leverage Values: Mera koliko mnogo posmatrana vrednost utiče na fitovanje regresionog model.

Termini za reziduale:

1. Unstandardized: Vrednost zavisne promenljive minus njegova predviđena vrednosti .

2. Standardized: Reziduali podeljeni procenom njihove standardne greške.

3. Studentized: Reziduali podeljeni procenom njihove standardne greške koja varira od slučaja do slučaja, na osnovu rastojanja posmatrane vrednosti nezavisne promenljive od njene srednje vrednosti.

4. Deleted: Reziduali, kod kojih su vrednosti tog slučaja isključene iz računa koeficijenata regresije.

5. Studentized deleted: “Izbrisani” reziduali podeljeni procenom njihove standardne greške.

Bitne statistike:

1. DfBeta: Nova promenljiva za svaki pojam u regresijskom modelu, uključujući i konstantu, koja sadrži promenu koeficijenta za taj izraz, ako je trenutna vrednost izostavljena iz kalkulacije.

2. Standardized DfBeta: Nova promenljiva za svaki pojam u regresijskom modelu, uključujući i konstanta, koja sadrži vrednost DfBeta podeljenu procenom njegove standardne greške.

3. DfFit: Promena u predviđenoj vrednosti zavisne promenljive ako je trenutna vrednost izostavljena iz računa.

4. Standardized DfFit: DfFit vrednost podeljena procenom njene standardne greške.

5. Covariance Ratio: Determinanta kovarijacione matrice gde je trenutna vrednost isključena iz računa, podeljena determinantom matrice gde je ta vrednost uključena.


Jelena Ljuboja
Jovana Dubljanin
Maša Obradović
Milan Ljuboja