UNIVERZITET U BEOGRADU

MATEMATIČKI FAKULTET

STUDIJSKI PROGRAM
STATISTIKA, FINANSIJSKA I AKTUARSKA MATEMATIKA

LINEARNI STATISTIČKI MODELI I NJIHOVA PRIMENA KROZ PROGRAMSKI PAKET SPSS

-MASTER RAD-

Mentor:
prof. dr Vesna Jevremović

Student:
Milica Obradović 1079/2012

Beograd, 2014.

SADRŽAJ

Uvod

U ovom radu biće prikazana primena linearnih modela kroz programski paket SPSS[1]. Prvi deo rada predstavlja teorijski uvod u linearne statističke modele, dok je drugi deo posvećen programu SPSS, unosu i obradi podataka, i svakako konstrukciji linearnih regresionih modela u njemu. Akcenat će biti na upoznavanju osobina i mogućnosti SPSS-a. To je program koji podržava istraživačku praksu i za razliku od nekih drugih programa (na primer R) na jednostavan način se dolazi do rezultata, odnosno nije potrebno ni osnovno znanje iz programiranja. Konstrukcija linearnih modela će biti prikazana na osnovu istraživanja, koje je sprovedeno u jednoj hladnjači koja otkupljuje malinu u Zapadnoj Srbiji. Naime, proizvođači predaju svežu malinu u hladnjaču, gde se zamrzava i dalje prodaje. Proizvođači imaju različite površine zasada i svi u zavisnosti od svojih mogućnosti ili volje ulože novac u svoj voćnjak. Hladnjača pruža svojim proizvođačima mogućnost uzimanja preparata i druge robe avansno, pre sezone branja. U toku istraživanja na slučajan način je uzet uzorak od 50 proizvođača, i za njih su uzeti sledeći podaci:

1) Količina predate maline u kilogramima u hladnjaču tokom poslednje sezone branja koja traje oko 50 dana

2) Površina malinjaka u arima

3) Starost malinjaka

4) Vrednost robe koju je svako od njih posebno avansno uzeo

Cilj istraživanja je da ispita da li obeležja pod 2), 3) i 4) utiču na obeležje 1). Odnosno, ako važi da površina malinjaka utiče na količinu proizvedene maline (ako je površina malinjaka veća, veća je i proizvedena količina), onda proizvođač na pravi način brine o svom zasadu. Po nekoj logici, ako se proizvođač bavi dugo malinom, trebalo bi da je ima više, zato što je o njoj sve naučio. Pitanje je, da li ljudi čiji je malinjak stariji imaju veću količinu? Takođe, ispitujemo zavisnost uzete robe pre sezone, i količine proizvedene maline. Da li oni proizvođači koji su više uložili u malinjak imaju više proizvedene maline, bez obzira na površinu malinjaka? Na ova pitanja će biti dati odgovori kroz primenu linearnih modela u

1 Statistical Program for Social Science

programu SPSS, i ispitivanja jačine eventualne linearne veze. Obeležje pod 1) ćemo, radi lakšeg snalaženja, pri prikazivanju jednačina imenovati zavisnom promenljivom Y, a obeležja pod 2),3) i 4) redom nezavisnim promenljivima X_1,X_2,X_3.

Linearni statistički modeli

Osnovna svrha primene regresione analize je da se na osnovu jedne poznate promenljive može predvideti vrednost druge, nepoznate promenljive i to iz relacije koja pokazuje njihovu zavisnost. Veze između pojava mogu biti funkcionalne (determinističke) i statističke (stohastičke). Glavni zadatak regersione analize jeste otkrivanje zakonitosti i pravilnosti koje

vladaju u odnosima među masovnim statističkim pojavama i kreiranje matematičkih modela koji pomoću simbola opisuju ponašanje pojava u stvarnim uslovima funkcionisanja. Kada se u analizi međuzavisnosti definiše koja je promenljiva zavisna a koja nezavisna, onda se koriste metode regresione analize. Zavisnost pojava se utvrđuje prema prethodnim teorijskim i empirijskim saznanjima o prirodi pojava i njihovim odnosima. Model koji pokazuje kako na vrednost zavisne promenljive utiče vrednost više ili jedne nezavisne promenljive naziva se regresioni model. Ukoliko je zavisnost linearna dobijeni model je linearni regresioni model.

Istorija linearnih statističkih modela

U istraživanjima često se interes istraživača usmerava prema problemu povezanosti među promenljivim (obeležjima). Pri tom je od posebnog interesa mogućnost prognoziranja ili predviđanja vrednosti jedne zavisne promenljive na osnovu drugih nezavisnih promenljivih. Prvi tako formulisani problem potiče od engleskog antropologa Francisa Galtona. On je studirajući zajedno sa Pirsonom nasleđivanje u biologiji, mereći visine očeva i sinova ustanovio neku vrstu paradoksa, odnosno, da visoki očevi imaju visoke sinove ali u proseku ne tako visoke kao što su oni sami, i slično, da niski očevi imaju niske sinove ali opet u proseku ne tako niske kao što su oni. Ovu tendenciju proseka neke karakteristike (u ovom slučaju visine) odabrane grupe da u sledećoj generaciji sinova teži ka proseku populacije a ne proseku njihovih očeva, Galton je nazvao regresijom, tačnije, regresijom prema proseku. Da bi dobio informaciju zavisnosti visine sinova od visine njihovih očeva, Pirson je pretpostavio da se ta zavisnost može izraziti kao funkcija, $Y = f(X)$, pri čemu je Y zavisna promenljiva, odnosno promenljiva koju želimo da objasnimo ili predvidimo (u Galtonovom primeru visina sinova), a X nezavisna promenljiva koju koristimo da objasnimo zavisnu promenljivu (visina očeva).

Osnovni pojmovi

Opšti oblik modela regresije je :

$$Y = f(X_1, X_2, \ldots\ldots, X_n) + \varepsilon .$$

Iz navedene relacije vidimo da se model sastoji iz determinističkog dela, koji predstavlja funkciju kojom se izražava zavisnost zavisne promenljive od određenog broja nezavisnih promenljivih, i stohastičkog dela koji predstavlja eventualno odstupanje od te navedene funkcionalne zavisnosti. Modele regresije mozemo podeliti u odnosu na broj nezavisnih promenljivih uključenih u model i u odnosu na oblik funkcije determinističkog dela regresinog modela. U odnosu na broj nezavisnih promenljivih, modeli regresije se dele na

modele jednostruke regresije i na modele višestruke regresije. Model jednostruke regresije ima jednu zavisnu i jednu nezavisnu promenljivu, a model višestruke regresije ima jednu zavisnu i više nezavisnih promenljivih. Prema obliku funkcije determinističkog dela, modele regresije delimo na linearne i nelinearne regresione modele. Veza između promenljivih linearnog modela predstavljena je linearnom funkcijom čiji je grafik prava, a veza između promenljivih nelinearnog modela ima oblik neke druge matematičke funkcije čiji je grafik neka kriva linija. Iz tog razloga se nelinearni model naziva još i krivolinijski regresioni model. Pomoću regresione i korelacione analize određujemo jačinu, smer i oblik veze između posmatranih pojava. Oblik veze, kao što je već navedeno, predstavlja oblik matematičke funkcije iz determinističkog dela modela. Smer veze može biti pozitivan i negativan. Jačina veze se određuje analizom slučajne promenljive regresionog modela.

Jednostruki linearni modeli

Jednostruki linearni modeli imaju sledeći oblik:

$$Y = a X + b + \varepsilon \qquad (*)$$

gde su,

$Y=(y_1, y_2, \ldots, y_n)^\intercal$ - zavisna promenljiva

$X=(x_1, x_2, \ldots, x_n)^\intercal$ -nezavisna promenljiva

a, b - koeficijenti

ε-slučajna promenljiva, odnosno slučajna greška koja se javlja pri merenju, koja ima normalnu raspodelu $N(0, \sigma^2)$, pri čemu σ^2 nije poznato.

Koeficijenti a i b su nepoznati, i njih treba oceniti. Konstanta b predstavlja vrednost y kada je

x =0 i naziva se odsečak na y osi. Koeficijent regresije, parametar a , pokazuje kako se linearno menja vrednost zavisne promenljive Y ako se nezavisna promenljiva X promeni za jedinicu mere. Koeficijent a ima pozitivan predznak kada se sa povećanjem vrednosti promenljive X povećava vrednost promenljive Y , a negativan kada se sa povećanjem vrednosti promenljive X smanjuje vrednost promenljive Y . Ako su obe promenljive izražene u istim dimenzijama, a predstavlja tangens ugla koji prava linija zaklapa sa x osom.

Metoda najmanjih kvadrata

Metoda koja se koristi da se izračunaju parametri linearne jednačine iz datih tačaka naziva se metoda najmanjih kvadrata. Ovom metodom se na osnovu uzorka $(x_i, y_i), i=1,\ldots,n$ određuje zavisnost Y od X, pri pretpostavci da je ona oblika $Y=aX+b+\varepsilon$. Koeficijente a

i b određujemo iz uslova da suma kvadrata odstupanja $S(a,b)=\sum_{i=1}^{n}(y_i-(a x_i+b))^2$

bude minimalna, odnosno:

$$\min_{a,b\in R} S(a,b)=\min_{a,b\in R}\sum_{i=1}^{n}(y_i-(a x_i+b))^2$$

Iz sistema jednačina

$$\frac{\partial S(a,b)}{\partial a}=0,\frac{\partial S(a,b)}{\partial b}=0$$

dobijamo:

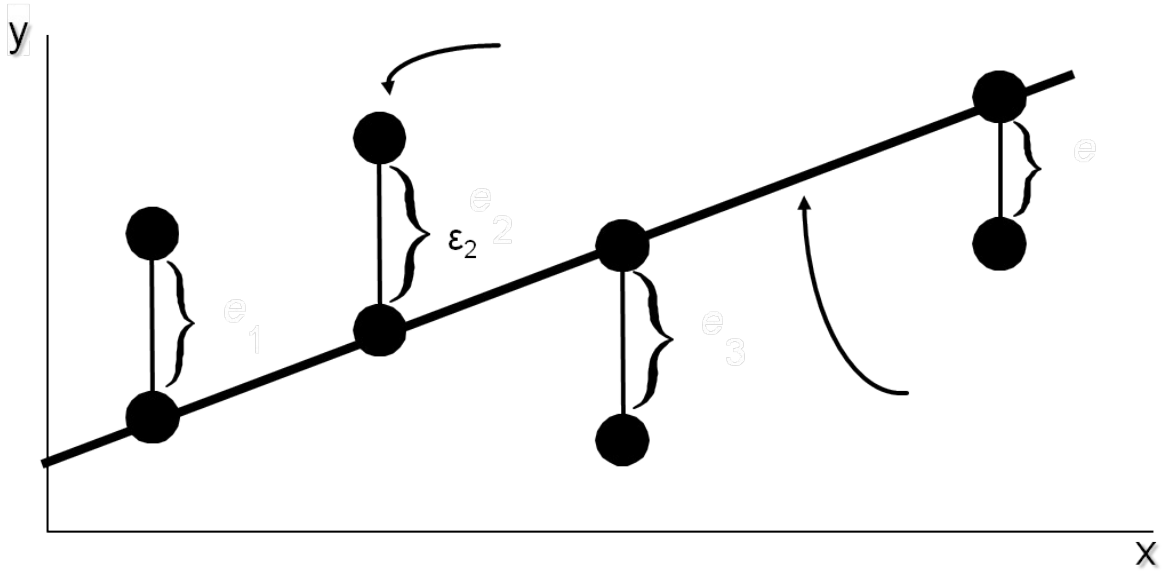
$$a=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\cdot\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

(**)

$$b = \acute{y}_n - a\acute{x}_n$$

Jasno je da postoji više pravih linija relativno bliskih svim tačkama iz uzorka, a najbolja je ona kod koje je zbir kvadrata odstupanja tačaka od regresione prave najmanji mogući. Prava $y = ax + b$ se naziva regresiona prava. Grafička interpretacija je data na sledećoj slici:

$y = ax + b$ $y_2 = ax_2 + b + \varepsilon_2$



Slika 1

Neka su $\hat{a}$ i $\hat{b}$ rešenja sistema (**). Ako posmatramo kao slučajne promenljive, tada su $\hat{a}$ i $\hat{b}$ nepristrasne[2] ocene parametara a i b iz modela (*).

2 Nepristrasna ocena parametra a je $\hat{a}$ ako važi $E(\hat{a}) = a$

Prava koja se prema kriterijumu najmanjih kvadrata najbolje uklapa u grupu tačaka naziva se regresiona prava, a model koji je definiše naziva se regresioni model. Osobine koje regresioni model ima su sledeće:

- Razlika između stvarne vrednosti i izračunate vrednosti za Y je najmanja moguća
- Iz srednje vrednosti za promenljivu X možemo da izračunamo i srednju vrednost za promenljivu Y
- Kada vrednost promenljive X odstupa od srednje vrednosti možemo da očekujemo i da vrednost promenljive Y odstupa od svoje srednje vrednosti

Analiza disperzije u modelu jednostruke linearne regresije

Prvo je potrebno reći nešto osnovno o analizi disperzije. ANOVA[3] je analitički model koji ukupnu disperziju deli na nekoliko delova, pri čemu se svaki od njih vezuje posebnim sistemom variranja tako da je moguće odrediti ne samo koji su izvori variranja u pitanju, nego i koliki je doprinos svakog dela u ukupnoj disperziji. Prednost ove metode se ogleda u tome što u model ulaze u obzir svi varijabiliteti, kao i njihov međusobni uticaj, što je nemoguće proceniti na drugi način.

3 ANalysis Of VAriances

Formula analize disperzije glasi:

$$(\hat{y}_i - \acute{y})^2 + ¿ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$(y_i - \acute{y})^2 = ¿ \sum_{i=1}^{n} ¿$$

$$\sum_{i=1}^{n} ¿$$

odnosno,

$$ST = SP + SR \quad .$$

ST je oznaka za ukupnu sumu kvadrata, odnosno za sumu kvadrata odstupanja vrednosti promenljive Y od srednje vrednosti:

$$ST = \sum_{i=1}^{n} (y_i - \acute{y})^2 \quad .$$

SP je oznaka za sumu kvadrata objašnjenu modelom, odnosno sumu kvadrata odstupanja regresionih vrednosti od srednje vrednosti:

$$SP=\sum_{i=1}^{n}(\hat{y}_i-\acute{y})^2 \quad .$$

SR je oznaka za rezidualnu ili neobjašnjenu modelom sumu kvadrata. To je suma odstupanja opaženih od regresionih vrednosti:

$$SR=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \quad .$$

Dakle, ukupna suma kvadrata ST se rastavlja na sumu kvadrata odstupanja regresionih vrednosti od srednje vrednosti SP i sumu kvadrata odstupanja regresionih od opaženih vrednosti, tj. rezidualnu sumu kvadrata SR. Ako se sume kvadrata podele sa odgovarajućim stepenima slobode dolazi se do sredina kvadrata koje su nezavisne procene komponenti disperzije. Sume kvadrata, stepeni slobode, sredine kvadrata i druge informacije predstavljaju se u tabeli analize disperzije (ANOVA), za model jednostruke linearne regresije:

Izvor disperzije	Stepeni slobode	Sume kvadrata	Sredina kvadrata	F-statistika	p>F

Objašnjen modelom Neobjašnjem modelom	1 n-2	SP SR	SP/1 SR/(n-2)	$\frac{SP/1}{SR/(n-2)}$	
Ukupno	n-1	ST			

Tabela 1

Standardna greška u modelu jednostruke linearne regresije

Vrednost odstupanja tačaka od prave izražava standardna greška regresione prave, koja se računa prema formuli:

$$S_{y,x}=\sqrt{\frac{\sum (y_i-\hat{y}_i)^2}{n-2}}$$

U ovoj formuli suma $\sum (y_i-\hat{y}_i)^2$ predstavlja sumu kvadrata odstupanja dobijenih vrednosti od izračunatih vrednosti, a broj n-2 predstavlja broj stepeni slobode (n se umanjuje za dva zato što su a i b izračunati iz istih podataka). Naime, ocena diperzije regresije je rezidualna suma kvadrata podeljena sa n-2 (stepeni slobode):

$$\hat{\sigma}^2=\frac{SR}{n-2}$$

Pozitivni koren iz ocenjene disperzije regresije je ocena standardnog odstupanja regresije:

$$\hat{\sigma}=\sqrt{\frac{SR}{n-2}}$$

Ukoliko postoji jaka linearna veza između X i Y znači da su tačke veoma blizu prave , pa je stoga suma kvadrata odstupanja dobijenih vrednosti od izračunatih vrednosti mala. Samim tim mala je i standardna greška $S_{y,x}$. Važi i obrnuto, kada je linearna zavisnost između promenljivih X i Y slaba, tačke su rasute oko prave i suma kvadrata odstupanja tačaka od prave je velika, pa je velika i standardna greška. Razlika između dobijenih i izračunatih vrednosti naziva se rezidual. Iz tog razloga, standardna greška se drugačije naziva rezidualno standardno odstupanje.

Koeficijent determinacije u modelu jednostruke linearne regresije

Uz pomoć regresione analize dobija se relacija koja objašnjava prirodu zavisnosti između dve promenljive, a ukoliko hoćemo da odredimo koliki je stepen te zavisnosti koristimo korelacionu analizu. Naime, kao što je već rečeno ukupna suma kvadrata ST sastoji se iz sume kvadrata odstupanja regresionih vrednosti od srednje vrednosti SP i rezidualne sume kvadrata SR.

Odnos objašnjene sume kvadrata SP i ukupne sume kvadrata ST je koeficijent determinacije. Koeficijent determinacije se obeležava sa R^2.

$$R^2=\frac{SP}{ST}=\frac{\sum(\hat{y}_i-\acute{y})^2}{\sum(y_i-\acute{y})^2}$$

Ako tačke leže na pravoj, rezidualna suma kvadrata SR je jednaka nuli, pa je ukupna disperzija jednaka sumi kvadrata odstupanja regresionih vrednosti od srednje vrednosti SP i koeficijent determinacije je jednak 1. Analogno, ako su tačke raštrkane oko prave, onda je koeficijent determinacije manji od 1. To znači da što je veća linearna zavisnost između X i Y, koeficijent determinacije je bliži 1, i obrnuto. Pored R^2 posmatra se korigovani koeficijent determinacije definisan relacijom:

$$\acute{R}^2=1-\frac{n-1}{n-2}(1-R^2)$$

Koeficijent korelacije

Koeficijenti korelacije predstavljaju meru povezanosti između dve promenljive. Korelacija proučava povezanost i uzajamni odnos među pojavama. Postoji više koeficijenata korelacije koji se koriste u različitim slučajevima. U praksi se prilikom rada sa linearnim modelima najčešće koristi Pirsonov koeficijent korelacije. Pirsonov koeficijent korelacije koristi se u slučajevima kada su promenljive posmatranog modela linearno povezane i normalno raspodeljene promenljive (prema (9)) . Vrednost Pirsonovog koeficijenta korelacije kreće se od +1 (savršena pozitivna korelacija) do –1 (savršena negativna korelacija)i obeležava se sa *r*. Kada je vrednost r blizu -1 ili 1 znači da između X i Y postoji jaka linearna veza i da regresiona jednačina može da se koristi za predviđanje. Ako je vrednost koeficijenta korelacije blizu nule, to znači da između promenljivih postoji slaba linearna zavisnost. Predznak koeficijenta nas upućuje na smer korelacije, to jest da li je ona pozitivna ili negativna, ali nas ne upućuje na snagu korelacije. Pirsonov koeficijent korelacije bazira se na poređenju stvarnog uticaja posmatranih promenljivih jedne na drugu u odnosu na maksimalni mogući uticaj dve promenljive. Međutim, ne mora visoka apsolutna vrednost za r nužno da znači da su X i Y u jakoj vezi. Moguće je da se desi da su ove dve promenljive u jakoj vezi sa nekom trećom promenljivom, pa su zbog toga i one povezane. Za računanje koeficijenta korelacije potrebne su tri različite sume kvadrata : suma kvadrata promenljive *X*, suma kvadrata promenljive *Y* i suma umnožaka promenljivih *X* i *Y*, to jest:

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

Veliki značaj u regresionoj analizi ima korelacijska matrica, koja pruža uvid u linearnu povezanost promenljivih. To je matrica koja sadrži koeficijente linearne korelacije nultog reda, odnosno koeficijente jednostruke linearne korelacije između svih parova promenljivih uključenih u regresioni model. Redovi i kolone matrice predstavljaju posmatrane promenljive, a član matrice na preseku određenog reda i kolone predstavlja koeficijent korelacije između promenljivih u odgovarajućem redu i koloni. Matrica na dijagonali ima član 1 (svaka promenljiva je sama sa sobom u potpunoj korelaciji). Dobijena matrica je simetrična, to jest podaci iznad i ispod dijagonale za isti par promenljivih su identični. Zbog tih svojstava dovoljno je posmatrati jedan njen deo, iznad dijagonale ili ispod dijagonale. Vizuelno možemo utvrditi u kojoj meri su dve pojedinačne promenljive u korelaciji, koje promenljive u međusobnom odnosu imaju najveći ili najmanji koeficijent korelacije, i koji skupovi promenljivih se ističu sličnim koeficijentima. Na taj način ne možemo utvrditi na koji način i u kolikoj meri više promenljivih zajednički utiče na drugu pojedinačnu promenljivu.

$$K=\begin{bmatrix} 1 & r_{y1} & r_{y2} & \dots & r_{yk} \\ r_{1y} & 1 & r_{12} & \dots & r_{1k} \\ r_{2y} & r_{21} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \vdots & 1 & r_{3k} \\ r_{ky} & r_{k1} & r_{k2} & r_{k3} & 1 \end{bmatrix}$$

Testiranje hipoteza u modelu jednostruke linearne regresije

Test hipoteze o pretpostavljenoj vrednosti regresionog parametra a u modelu jednostruke linearne regresije, moguće je sprovesti pomoću t-testa ili F-testa.

t-test

Cilj testiranja jeste da se utvrdi da li odabrana nezavisna promenljiva treba biti uključena u model, ili je suvišna u modelu. Model se uvek formira sa namerom da bude prihvaćen, pa se iz tog razloga alternativna hipoteza uvek formira tako da bude u skladu sa pretpostavkom istraživača. U regresionoj analizi se t-test najčešće formuliše kao jednostrani test.

Postavljene hipoteze (nulta H_0 i alternativna H_1) su:

- H_0: $a=0$
- H_1: $a>0$

Test statistika je:

$$t=\frac{\hat{a}}{SE(\hat{a})}$$

Ako izračunata vrednost t-statistike pripada kritičnoj oblasti, nulta hipoteza se odbacuje, i obrnuto ako izračunata vrednost t-statistike ne pripada kritičnoj oblasti nulta hipoteza se ne odbacuje. Kritična oblast je definisana intervalom $D=(t_{\alpha,(n-2)},+\infty)$, gde je $t_{\alpha,(n-2)}$ vrednost iz tablica za Studentovu raspodelu sa datim nivoom značajnosti α (verovatnoća greške prve vrste[4]) i sa (n-2) stepena slobode. Odnosno, ako je izračunata vrednost za t manja od tablične za izabrani nivo značajnosti α i broj stepeni slobode n-2, onda ne postoji značajna linearna zavisnost između X i Y , i ne odbacujemo nultu hipotezu.

4 Greška prve vrste se javlja kada se odbaci tačna nulta hipoteza.

F-test

Polazeći od elemenata tabele ANOVA moguće je sprovesti test o značajnosti nezavisne promenljive u modelu jednostruke linearne regresije ekvivalentnom t-testu. Cilj testiranja jeste da se utvrdi da li je disperzija odabrane nezavisne promenljive značajna za objašnjenje disperzije zavisne promenljive, pa treba biti uključena u model ili je suvišna u modelu. Kao i prethodni test, i F-test je formulisan kao jednostrani. Postavljene hipoteze (nulta H_0 i alternativna H_1) su:

- H_0: $a=0$
- H_1: $a>0$

Uz polazne pretpostavke o modelu važi:

$$\frac{SR}{\sigma^2} \sim \chi^2(n-2)$$

(***)

$$\frac{SP}{\sigma^2} \sim \chi^2(1).$$

Na osnovu (***) zaključujemo da sledeća F-statistika ima Fišerovu raspodelu $F(1,(n-2))$

$$F=\frac{\frac{SP}{1}}{\frac{SR}{n-2}} \quad F(1,(n-2)) \quad .$$

Postupak testiranja se sprovodi na sledeći način. Kritična oblast je definisana intervalom $D=(F^{\alpha}_{1,(n-2)},\infty)$, (gde je α zadat nivo značajnosti, a $F^{\alpha}_{1,(n-2)}$ vrednost iz tablice za Fišerovu raspodelu sa datim nivoom značajnosti i stepenima slobode (1,n-2). Ukoliko izračunata vrednost test statistika F pripada kritičnoj oblasti D, nulta hipoteza se odbacuje. Odnosno, za zadati nivo značajnosti α odbacuje se H_0 ako je $F^{\alpha}_{1,(n-2)}$ manja od vrednosti F-statistike, a u suprotnom se H_0 ne odbacuje (prema (1)).

Takođe, zaključke možemo doneti i na osnovu *p*-vrednosti testa, koja predstavlja verovatnoću, uz pretpostavku da je nulta hipoteza tačna, da vrednost test statistike F_1 bude veća ili jednaka apsolutnoj vrednosti test statistike dobijene na osnovu uzorka *F*, odnosno $P\left(F_1 \geq \left|F\right|\right)=p$. Ako je *p*-vrednost testa mala (u odnosu na odabrani nivo značajnosti α), nulta hipoteza se odbacuje.

Višestruki linearni modeli

Opšti oblik modela višestruke regresije je:

$$Y = f(X_1, X_2, \ldots\ldots, X_k) + \varepsilon$$

Kao i kod prostog linearnog modela Y je zavisna promenljiva. To je promenljiva čija se disperzija izražava u funkciji disperzija nezavisnih promenljivih $X_1, X_2, \ldots\ldots, X_k$. Promenljiva ε izražava nepoznata odstupanja od funkcionalnog odnosa. Ako pretpostavimo da je veza između Y i ($X, X_2, \ldots\ldots, X_k$) linearna, model višestruke linearne regresije ima sledeći oblik:

$$Y=\beta_0+\beta_1 X_1+\ldots+\beta_k X_k+\varepsilon \qquad (\#)$$

U ovoj relaciji $X_1, X_2, \ldots, X_k$ su nezavisne promenljive, β_0 ,…., β_k su parametri, a ε su slučajne promenljive. Ako se pretpostavi da se linearna regresiona veza između promenljive Y i odabranog skupa nezavisnih promenljivih utvrđuje na osnovu uzorka veličine n, tada se formula (#) može napisati u vidu sistema od n jednačina:

$$y_1=\beta_0+\beta_1 x_{11}+\ldots+\beta_k x_{1k}+\varepsilon_1$$

$$y_2=\beta_0+\beta_1 x_{21}+\ldots+\beta_k x_{2k}+\varepsilon_2$$

.

.

.

.

$$y_n = \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \varepsilon_n$$

Navedeni sistem jednačina može da se predstavi i u matričnom obliku:

$$Y = X\beta + \varepsilon \quad ,$$

gde su

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{bmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} \quad .$$

U analizi modela polazi se od određenih pretpostavki koje se odnose na prirodu uključenih promenljivih. To su sledeće pretpostavke:

- Veza između zavisne promenljive i odabranog skupa nezavisnih promenljivih je linearna
- Promenljive $X_i, i=1,\ldots,k$ su međusobno nezavisne i rang matrice X je k+1
- Slučajne promenljive ε_i imaju normalnu raspodelu sa očekivanjem nula i sa konstantnom disperzijom σ^2 I međusobno su nekorelisane, odnosno važi:

$$E(\varepsilon_i)=0, Var(\varepsilon_i)=0, \varepsilon_i=N(0,\sigma^2)$$

$$Cov(\varepsilon_i,\varepsilon_j)=E(\varepsilon_i,\varepsilon_j)=0, i\neq j, i,j=1,\ldots,n \quad .$$

Kao i kod jednostrukog linearnog modela, parametre β_0 ,…., β_k određujemo metodom najmanjih kvadrata. Naime, ocenjeni parametri $\hat{\beta}_0$,…., $\hat{\beta}_k$ omogućavaju minimum funkcije $S(\beta_0,\ldots,\beta_k)$.

$$S(\beta_0,\ldots,\beta_k)=\sum_{i=1}^{n}(y_j-(a_0+\sum_{j=1}^{k}a_j x_{ji}))^2 \quad .$$

Regresione vrednosti

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} \quad , \quad i=1,\ldots,n$$

su vrednosti zavisne promenljive za zadate vrednosti nezavisnih promenljivih $x_{i1}, x_{i2}, \ldots\ldots, x_{ik}$, i=1,….,n. Rezidualna odstupanja su razlike između empirijskih i regresionih vrednosti zavisne varijable, to jest:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Analiza disperzije u modelu višestruke linearne regresije

Kao što je već navedeno, formula analize disperzije je:

$$(\hat{y}_i - \acute{y})^2 + ¿ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$(y_i - \acute{y})^2 = ¿ \sum_{i=1}^{n} ¿$$

$$\sum_{i=1}^{n} ¿$$

odnosno,

$$ST = SP + SR \quad .$$

Objašnjenje za ovu formulu je u delu o jednostrukoj linearnoj regresiji.

Sume kvadrata, stepeni slobode i sredine kvadrata predstavljaju se u tabeli analize disperzije (ANOVA), o čemu je u ovom radu već bilo reči. Tabela analize disperzije za model višestruke regresije je Tabela2:

Izvor disperzije	Stepeni slobode	Sume kvadrata	Sredina kvadrata	F-statistika	p>F
Objašnjen modelom Neobjašnjem modelom	k n-(k+1)	SP SR	SP/k SR/(n-(k+1))	$\frac{SP/k}{SR/(n-(k+1))}$	
Ukupno	n-1	ST			

Tabela 2

Standardno odstupanje u modelu višestruke linearne regresije

Rezidualna suma kvadrata podeljena sa (n-(k+1)) (stepeni slobode) je ocenjena disperzija regresije:

$$\hat{\sigma}^2 = \frac{SR}{n-(k+1)}$$

Pozitivni koren iz ocenjene disperzije regresije je ocena standardnog odtupanja regresije i može se interpretirati kao prosečno odstupanje empirijskih od regresionih vrednosti.

Koeficijent determinacije u modelu višestruke linearne regresije

Parametar

$$R^2 = \frac{SP}{ST} = 1 - \frac{(n-(k+1))\hat{\sigma}^2}{\sum_{i=1}^{n}(y_i - \acute{y})^2}$$

naziva se koeficijent determinacije. Kao što je već rečeno, u slučaju jednostruke regresije, on uzima vrednosti iz intervala [0,1] i posmatrani model je bolji kako je koeficijent bliži jedinici. Međutim, ovaj pokazatelj ima nedostatak što nije nepristrasan. Korigovani koeficijent determinacije[5] definisan je sledećim izrazom:

5 Adjusted R Square

$$\acute{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2)$$

Korigovani koeficijent je uvek manji od pravog koeficijenta determinacije, utoliko više što je više nezavisnih promenljivih za isti uzorak (osim za vrednost $\acute{R}^2 = R^2$).

Testiranje hipoteza u modelu višestruke linearne regresije

Testovi značajnosti regresionih promenljivih mogu se podeliti u dve grupe:

1) Test o značajnosti jedne regresione promenljive

t-test

Za model višestruke regresije, t-test se sprovodi analogno kao i za model jednostruke linearne regresije. Cilj testiranja jeste da se utvrdi da li odabrana nezavisna promenljiva treba biti uključena u model, ili je suvišna u modelu. Postavljaju se nulta i alternativna hipoteza:

- H_0: $\beta_j=0$
- H_1: $\beta_j\geq 0$, j=1,…,k

Test statistika je:

$$t=\frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}}$$

Definiše se kritična oblast $D=\left(t_{\alpha,n-|k+1|},+\infty\right)$, sa datim nivoom značajnosti α, gde je $t_{\alpha,n-|k+1|}$ vrednost koja se određuje iz tablica za Studentovu raspodelu sa (n-(k+1)) stepeni slobode. Ukoliko vrednost test statistike pripada kritičnoj oblasti nulta hipoteza se odbacuje, i obrnuto, ukoliko vrednost test statistike ne pripada kritičnoj oblasti *D* nulta hipoteza se ne odbacuje.

2) Test o značajnosti svih regresionih promenljivih

Globalni F-test

Ovaj test se naziva globalni test iz razloga što se nultom hipotezom pretpostavlja da disperzija nijedne od k nezavisnih promenljivih nema uticaj na disperziju zavisne promenljive. Alternativnom hipotezom se pretpostavlja ono što istraživaču odgovara, da je bar jedna od nezavisnih promenljivih značajna u modelu. Postavljaju se nulta i alternativna hipoteza:

- H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
- H_1: $\exists\ \beta_j \neq 0,\ j=1,\ldots,k$

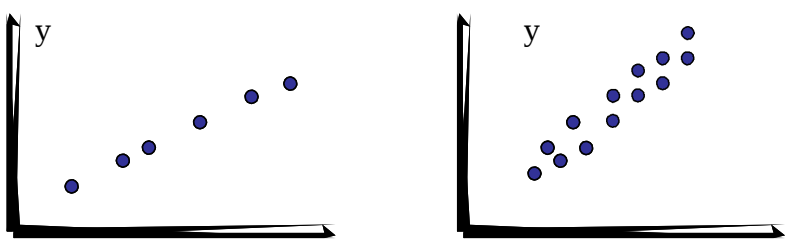
Test statistika je:

$$F = \frac{\frac{SP}{k}}{\frac{SR}{n-(k+1)}}$$

Definiše se kritična oblast $D=\left(F_{k,n-(k+1)}^{\alpha},+\infty\right)$, sa datim nivoom značajnosti α, gde je $F_{k,n-(k+1)}^{\alpha}$ vrednost koja se određuje iz tablica za Fišerovu raspodelu sa (k,n-(k+1)) stepeni slobode. Ukoliko test statistika pripada kritičnoj oblasti, onda se odbacuje nulta hipoteza.

Grafički prikaz

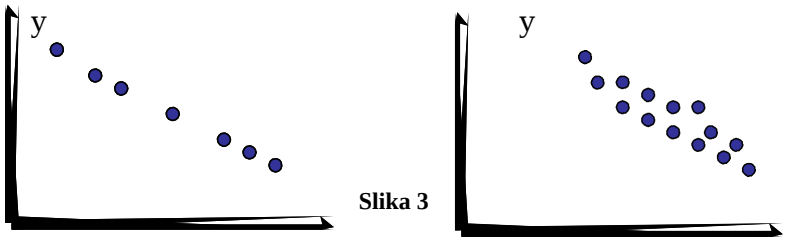
Prilikom istraživanja zavisnosti između promenljivih prvo što treba uraditi jeste formiranje dijagrama rasipanja, koji predstavlja grafički prikaz zavisnosti i međuzavisnosti izmedju promenljivih. Dijagram rasipanja ili scatter dijagram, konstruiše se na osnovu dobijenog eksperimentalnog skupa podataka i prikazuje parove vrednosti promenljivih *X* i *Y*. Dijagram rasipanja omogućava nam da steknemo predstavu o tome da li ima smisla ili ne tražiti zavisnost između promenljivih *X* i *Y*. Na osnovu dijagrama rasipanja može da se uoči oblik aproksimativne linije: prava, kriva, rastuća ili opadajuća, tačke minimum ili maksimuma ili prevojne tačke. Pomoću ovog dijagrama se takođe otkrivaju autlajeri i na njemu se lako mogu uočiti grube greške, a često i sistematske i slučajne greške ravnomernim rasipanjem eksperimentalnih podataka oko aproksimativne krive. Kao što je već navedeno u radu, postoje dva oblika zavisnosti, funkcionalna i statistička zavisnost. Na sledećim slikama možemo uočiti razliku između njih.



Slika 2

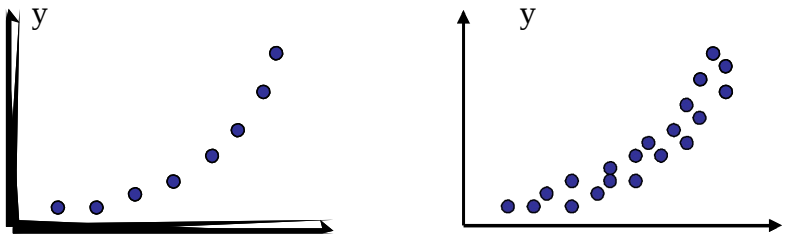
Sa dijagrama levo na Slici2 vidimo da u funkcionalnoj vezi zamišljena linija koja povezuje sve tačke je prava. Od te prave nema nikakvog odstupanja. Vidimo takođe da porast vrednosti jedne promenljive prati porast vrednosti druge posmatrane promenljive, pa je iz tog razloga ova veza pozitivna. Sa grafika desno na Slici2 vidimo da su ovde prisutna pozitivna i negativna odstupanja od prave linije, što se tumači raznim uticajima drugih promenljivih. Za ovu vezu se kaže da je statistička. Ovde je veza takođe pozitivna, zato što u proseku porast jedne promenljive prati porast druge promenljive.

Takođe, postoje negativna funkcionalna i negativna statistička veza. Odgovarajući dijagrami rasipanja su predstavljeni na Slici3.



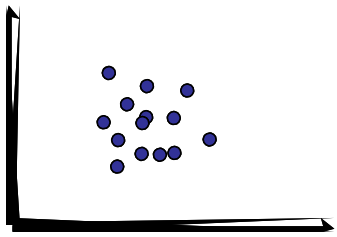
Slika 3

Međutim, kao što je već navedeno međusobne veze između dve promenljive ne moraju da budu pravolinijske, već mogu imati i drugi oblik. Slika4 predstavlja krivolinijske pozitivne veze između promenljivih X i Y.



Slika 4

Takođe, sa dijagrama rasipanja možemo i da utvrdimo da nema povezanosti između promenljivih. Odnosno, zamišljena linija koja prolazi između tačaka ne postoji i ne može da se oceni da li porast jedne promenljive prati rast ili pad druge promenljive, jer se pri jednoj vrednosti promenljive X javlja više različitih vrednosti promenljive Y. Na Slici5 vidimo kako izgleda jedan takav dijagram rasipanja.



Slika 5

Kolmogorov-Smirnov test za ispitivanje normalnosti

Ovaj test određuje da li je opravdana pretpostavka da je uzorak iz populacije sa nekom teorijskom raspodelom (u ovom slučaju normalnom). Primenjuje se za obeležja koja imaju neprekidne raspodele. Postavljene hipoteze su:

- H_0: Obeležje na populaciji iz koje je uzet uzorak ima normalnu raspodelu
- H_1: Obeležje na populaciji iz koje je uzet uzorak nema normalnu raspodelu

Neka je $F_0(x)$ potpuno određena funkcija raspodele, kad je hipoteza H_0 tačna. $F_0(x)$ predstavlja verovatnoću da obeležje u slučaju da podleže pretpostavljenoj raspodeli, nema vrednost veću od realnog broja x. Označimo sa $F_n^{¿}(x)$ uzoračku funkciju raspodele iz slučajnog uzorka sa n opservacija. Ona predstavlja vrednosti relativnih kumulativnih frekvencija i često se naziva empirijska funkcija raspodele. Ako označimo sa k broj opservacija iz uzorka obima n koje su manje ili jednake realnom broju x, biće:

$$F_n^{¿}(x)=\frac{k}{n}.$$

Kada je H_0 tačna , očekuje se da za svaku vrednost x , $F_n^{¿}(x)$ treba da bude vrlo bliska $F_0(x)$. Najveća vrednost $|F_0(x)-F_n^{¿}(x)|$ je maksimalno odstupanje D, i predstavlja statistiiku testa.

Neka je d realizovana vrednost statistike Kolmogorov-Smirnova:

$$d=^{¿}-\infty<x<+\infty|F_0(x)-F_n^{¿}(x)| \quad .$$

Kritična oblast je definisana sa $M=(d_{n,\alpha},+\infty)$ sa datim nivoom značajnosti α, gde je $d_{n,\alpha}$ vrednost koja se određuje u tablicama kritičnih vrednosti za test Kolmogorov-Smirnov za jedan uzorak. Hipotezu H_0 odbacujemo (za dati nivo značajnosti i za dati uzorak), ako važi $d>d_{n,\alpha}$.

Statistički softver SPSS

Program SPSS, predstavlja složen računski paket koji koriste istraživači iz oblasti društvenih nauka, kao i profesionalni analitičari koji se bave statističkom analizom podataka. Njegova verzija IBM SPSS Statistic 20 sadrži niz procedura koje se odnose na proces statističke analize, počev od planiranja istraživanja i prikupljanja podataka, unošenja podataka u program, analiziranja podataka, pa sve do pravljenja izveštaja i vizuelnog predstavljanja izlaznih rezultata analize. Ovaj softver je baš zbog svoje jednostavnosti i širokog spektra mogućnosti našao primenu i u držanju nastave statistike, kao u izradi mnogih naučnih projekata. Još jedna praktična stavar kod ovog softvera je to što je omogućeno učitavanje podataka iz standardnih baza podataka (Excel, Access…) ili iz bilo kog standardnog editora ukoliko su sačuvani u ASCII[6] format.

6 American Standard Code for Information Interchange

Linearni regresioni modeli u SPSS-u

Programom SPSS definisane su tri metode konstrukcije linearnih regresionih modela: standardni (standard), hijerarhijski i postepeni (stepwise). Navedene metode se razlikuju po dva osnova: po tretmanu dela disperzije koji se preklapa između nezavisnih promenljivih, jer je obično prisutna korelacija između njih, i po redosledu uključenja nezavisnih promenljivih u regresioni model.

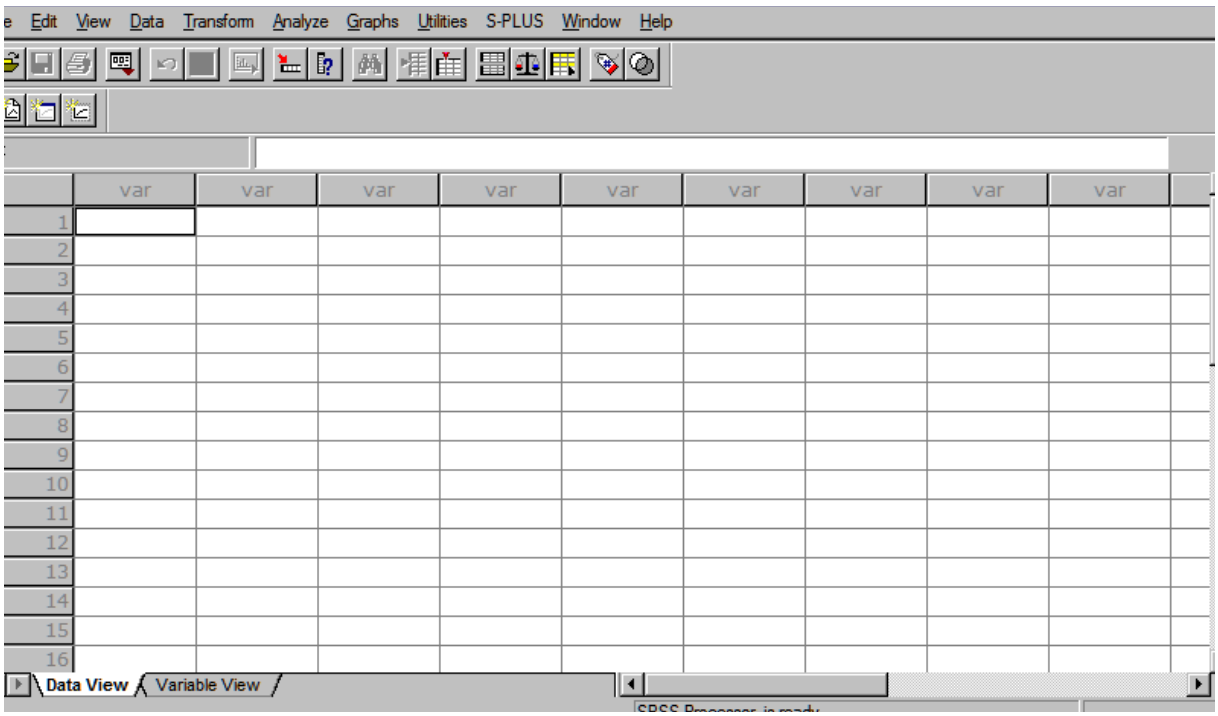
Kod **standard** metode sve nezavisne promenljive se se uključuju u regresioni model zajedno, jer je cilj analize da se ispita veza između skupa nezavisnih promenljivih i zavisne promenljive. Takođe, pomoću ove metode najlakše ispitujemo i konstruišemo model jednostruke linearne regresije.

U slučaju primene hijerarhijske višestruke regresije, sam istraživač određuje, na osnovu prethodnog znanja iz statistike i dobijenih vrednosti određenih parametara, redosled uključenja nezavisnih promenljivih u model.

Kod primene modela **stepwise** jedan broj nezavisnih promenljivih se uključuje u model, a deo se odmah odbacuje. Redosled njihovog uključenja zavisi isključivo od statističkih kriterijuma, koji su ugrađeni u Stepwise procedure u SPSS programu. Metoda uključenja nezavisnih promenljivih u regresioni model može biti unapred (forward), unazad (backward) ili kombinacija ove dve metode-stepwise. U slučaju primene **forward** metode , uključuju se jedna po jedna nezavisna promenljiva u model. Poredak uključenja i njihov opstanak u modelu određeni su pomoću statističkih kriterijuma. U statističke kriterijume pomoću kojih odlučujemo o opstanku promenljivih u modelu koristimo vrednost F-statistike, koja treba da bude veća od određene kritične vrednosti. I obrnuto, kod **backward** metode polazi se od toga da su sve nezavisne promenljive uključene u regresioni model, a zatim se jedna po jedna promenljiva isključuju iz modela. Metoda stepwise je kombinacija prethodne dve metode, kao što je već navedeno. Karakteristično za nju jeste to da se naizmenično proverava opravdanost i uključenja i eventualnog isključenja promenljive iz modela, tako da se može desiti da se iz modela isključi nezavisna promenljiva koja je u nekom prethodnom koraku zadovoljila kriterijum za uključenje u model.

Učitavanje podataka u SPSS i definisanje promenljivih

Prozor koji se prikazuje odmah po pokretanju SPSS radne sesije je veoma prilagodljiv i dat je u formi radnog lista (spreadsheet). On se koristi za unošenje, korigovanje i prikazivanje podataka. U njemu može da se kreira nova ili da se menja neka od postojećih datoteka podataka. Prozor editor podataka je dat u formi radnog lista u kome su opservacije predstavljene u redovima, a obeležja (promenljive) u kolonama (Slika6).



Slika 6

Za svaku novu promenljivu koja se unosi u SPSS program podrazumeva se da je numerička promenljiva i određuje joj širinu od 8 mesta sa dva decimalna mesta, dodeljuje joj naziv var00001 i daje joj se desno poravnanje. Međutim, svaku od ovih karakteristika možemo promeniti prema svojim potrebama i namerama. Naziv promenljive možemo promeniti kada mišem kliknemo na ćeliju sa već zadatim imenom. U principu, proces definisanja promenljivih sadrži sedam opcionih koraka. Primarni korak je dodeljivanje naziva promenljivoj, a ostali koraci uključuju određivanje tipa promenljive (**Variable type**), opisa promenljive (**Variable labels**), vrednosti promenljive (**Variable values**), nedostajućih vrednosti (**Missing values**),formata kolone u editoru podataka (**Column format**), pripadnosti određenoj skali merenja (Measurement level) i uloge promenljive (**Role**).

Radi lakšeg snalaženja moramo dati i odgovarajuće nazive promenljivima. Nazivi promenljivih moraju da ispunjavaju sledeće osobine:

- Naziv mora početi slovom, a znaci posle njega mogu biti slova, cifre tačke ili simboli @,#,_, $, i naziv se ne može završiti tačkom ili znakom za podvučeno.
- Svaki naziv mora biti jedinstven, odnosno duplikati nisu dozvoljeni
- Naziv ne može biti duzi od 64 bajta
- Rezervisane reči se ne mogu koristiti kao nazivi. U rezervisane reči spadaju na primer: ALL,AND,BY,NOT,OR,TO,WITH...

Nazivi promenljivih dodeljeni za ovaj slučaj su sledeći:

- Promenljiva **Količina_u_kg** predstavlja količinu proizvedene maline za sezonu i količina je predstavljena u kilogramima
- Promenljiva **Površina_zasada** predstavlja vrednost robe koju su proizvođači uzeli iz poljoprivredne apoteke na ime avansa
- Promenljiva **Broj_godina** predstavlja površinu malinjaka u arima
- Promenljiva **Avans_roba** predstavlja starost malinjaka

Posle definisanja promenljivih klikom na prvu ćeliju možemo pristupiti unosu podataka. Na druge ćelije prelazimo ili strelicama za pomeranje kursora ili klikom mišana njih. Uneti podaci se trajno zapisuju naredbom SAVE iz FILE menija. Pri prvom zapisivanju pokreće se opcija SAVE AS i tom prilikom je potrebno dati ime fajlu sa podacima. U ovom slučaju je to Master_rad.sav. U prilogu je data tabela sa podacima iz uzorka kao **Prilog1**.

Osnovni parametri za sve promenljive:

Postupak za generisanje tabele frekvencija , mere centralne tendencije i disperzije je sledeći:

1) Iz linije glavnog menija izabrati opciju **Analyze**

2) Odabrati stavku **Descriptive Statistics**, a zatim stavku **Frequencies**

3) Izabrati potrebnu promenljivu (jedna ili više), a zatim pritisnuti dugme sa strelicom udesno da bi promeenljive bile prenete u polje **Variables**

4) Pritiskom na dugme **Statistics** se otvara podokvir za dijalog **Frequencies:Statistics**

5) U delu okvira **Percentile Values** potvrditi polje **Quartiles**

6) U delu okvira **Central Tendency** potvrditi polja **Mean**, **Median** i **Mode**

7) U delu okvira **Dispersion** potvrditi polja **Std**, **Variance**, **Minimum** i **Maximum**.

8) Pritiskom na dugme **Continue** se prelazi na sledeći korak

9) Pritiskom na dugme **Continue,** a zatim na dugme **OK** se realizuje operacija

Statistics

		Količina_u_kg	Površina_zasada	Broj_godina	Avans_roba
N	Valid	50	50	50	50
	Missing	3	3	3	3
Mean		2377,0200	16,8200	13,5000	36079,6200
Std. Error of Mean		179,10995	1,11527	,85583	2663,74537

Median		2358,5000	16,2500	14,0000	36495,0000
Mode		486,00[a]	8,50	16,00	4582,00[a]
Std. Deviation		1266,49862	7,88615	6,05165	18835,52412
Variance		1604018,755	62,191	36,622	354776968,730
Minimum		486,00	4,00	1,00	4582,00
Maximum		5531,00	35,50	27,00	78320,00
Percentiles	25	1189,7500	10,7500	9,0000	22777,5000
	50	2358,5000	16,2500	14,0000	36495,0000
	75	3224,5000	21,7500	18,0000	44802,0000

a. Multiple modes exist. The smallest value is shown

Tabela 3

Iz Tabele3 da je prosečna proizvedena količina maline 2377,02kg,a da je standardna greška 179,10995. Prosečna površina zasada pod malinom je 16,82ari, a standardna greška je 1,11527. Takođe, prosečne vrednosti za broj godina malinjaka i za avans dat u robi su redom 13.50 i 36079,62din , a odgovarajuće vrednosti standardnog odstupanja su redom 0,85583 i 2633,74537.

Da bi uzorak bio reprezentativan i da bi bili ispunjeni uslovi za analizu disperzije, potrebno je da podaci za svaku promenljivu budu normalno raspodeljeni. Da li je taj uslov ispunjen, ispitaćemo pomoću Kolmogorov-Smirnovljevog testa za testiranje normalnosti raspodele. Uz

njega se formira i **normal probability plot**. Ovaj test se u programu SPSS sprovodi na sledeći način:

1) Iz glavnog menija izabrati komandu **Analyze**
2) Pritiskom na stavku **Descriptive Statistics**, a zatim na **Explore** otvara se okvir za dijalog **Explore**
3) Izabrati promenljivu koja vam je potrebna i pritisnuti dugme sa strelicom sa desne strane da biste prebacili ovu promenljivu u polje **Dependent List**
4) U delu **Display** proveriti da li je potvrđeno dugme **Both**
5) Izabrati komandno dugme **Plots** da biste dobili **Explore:Plots** podokvir za dijalog
6) Potvrditi polje **Normality plots with tests**
7) Pritisnuti dugme **Continue**, a zatim **OK**

Tests of Normality

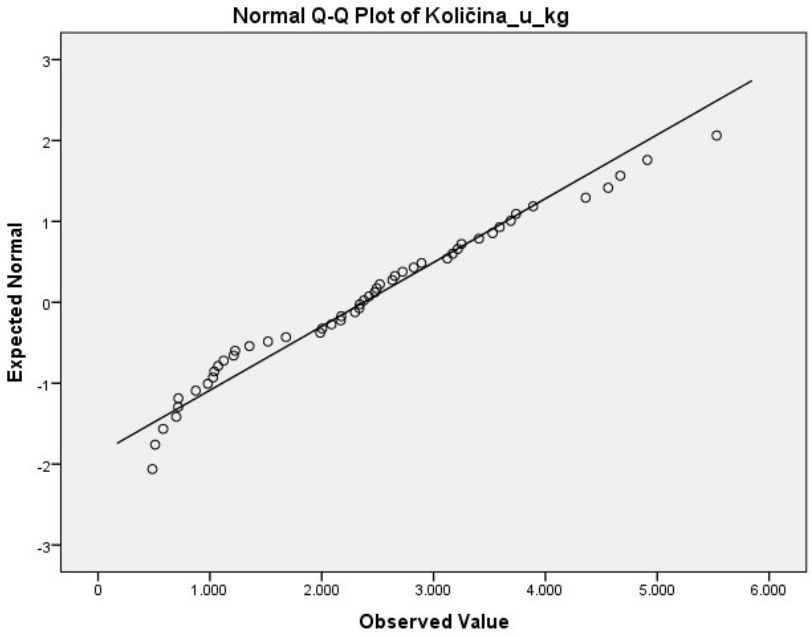
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Količina_u_kg	,098	50	,200*	,962	50	,107
Površina_zasada	,081	50	,200*	,964	50	,127
Broj_godina	,100	50	,200*	,984	50	,715
Avans_roba	,095	50	,200*	,953	50	,046

*. This is a lower bound of the true significance.

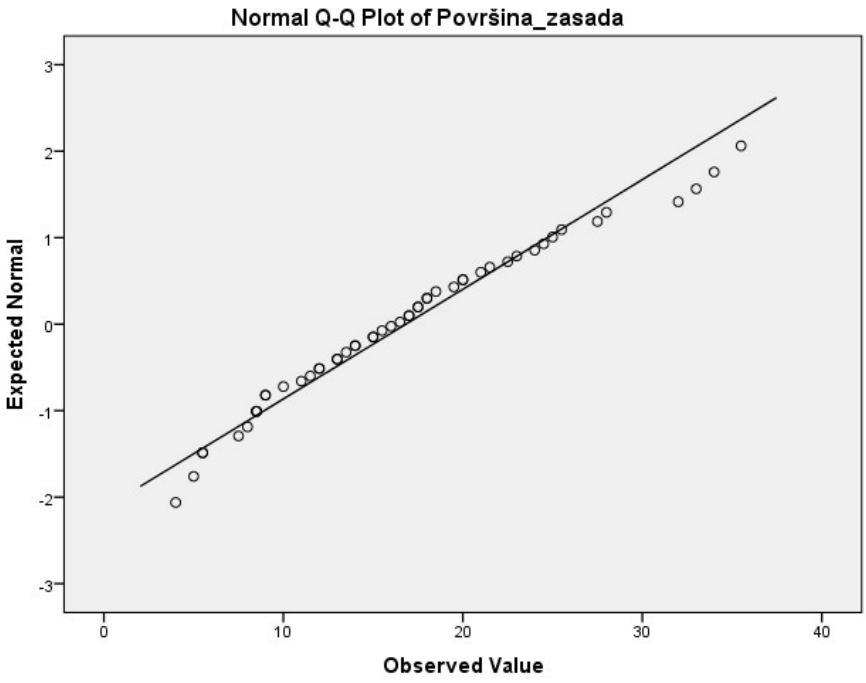
a. Lilliefors Significance Correction

Tabela 4

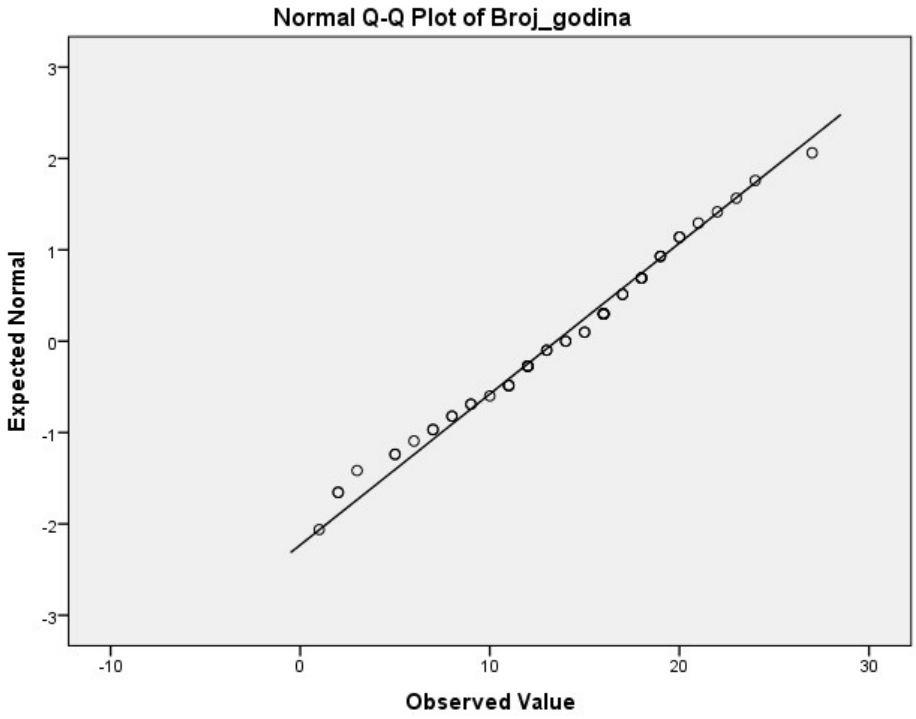
Tablično $d_{50;0,05}=1,92$ pa zaključujemo da za svaku vrednost Kolmogorov-Smirnov statistike važi $d<d_{50;0,05}$. Na osnovu teorijskog dela rada znamo da ne treba da odbacimo hipotezu H_o, odnosno zaključujemo da promenljive imaju normalnu raspodelu. Pošto je uzorak manji od 100 program je automatski odradio i Šapiro-Vilk test. Ako je vrednost za svaku promenljivu u koloni Sig za Šapiro-Vilk test veća od 0,05, onda prihvaatamo nultu hipotezu H_0 da su raspodelee promenljivih normalne. Za promenljivu Avans_roba Sig=0,046, međutim s obzirom da je dokazana normalnost po Kolmogorov-Smirnov testu i da je vrednost jako blizu 0,05, prihvatićemo da Avans_roba ima relativno normalnu raspodelu. Sve ove zaključci su i grafički prikazani na sledećim dijagramima.



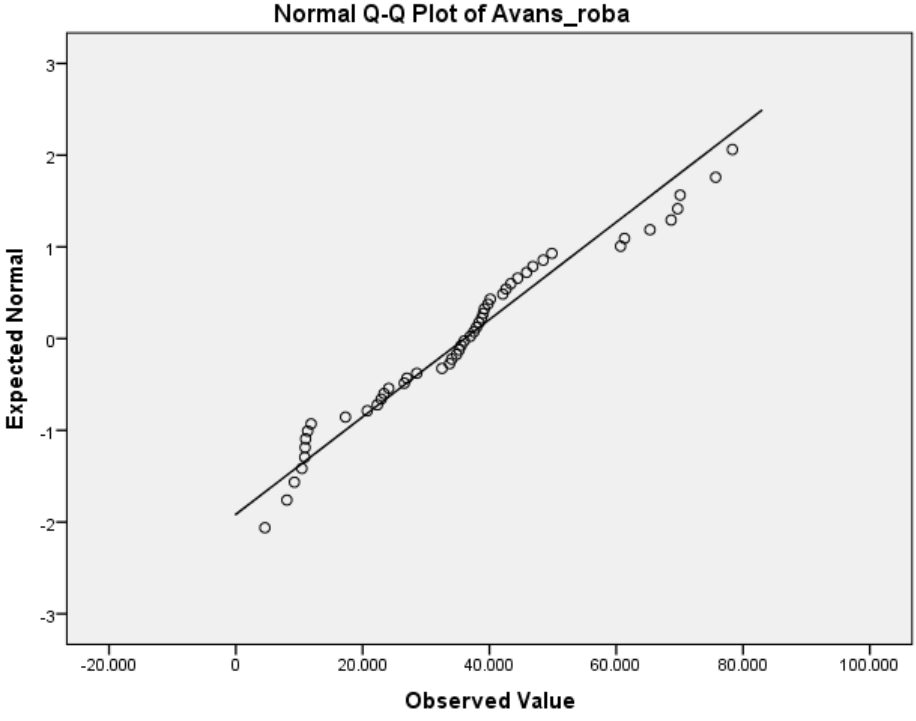
Dijagram 1



Dijagram 2



Dijagram 3



Dijagram 4

Korelacija između promenljivih

Prisutnost linearne veze možemo donekle zaključiti na osnovu dijagrama rasipanja. Pirsonov koeficijent korelacije meri korelaciju između dve neprekidne promenljive, a u SPSS programu se dobija primenom **Analyze** i **Correlate** menija. Korelaciona analiza počiva na pretpostavkama da su podaci prikupljeni za uparene parove (ako su podaci uzeti od jednog ispitanika za promenljivu X, onda od njega mora da se ima i podatak za promenljivu Y) i da podaci moraju biti mereni na intervalnoj ili relacionoj skali. Takođe, podaci za svaku promenljivu moraju biti normalno raspodeljeni i da veza između njih mora biti linearna. Primećujemo da dati uzorak zadovoljava sve ove zahteve, i sad ostaje da ispitamo još linearnost i stepen korelacije.

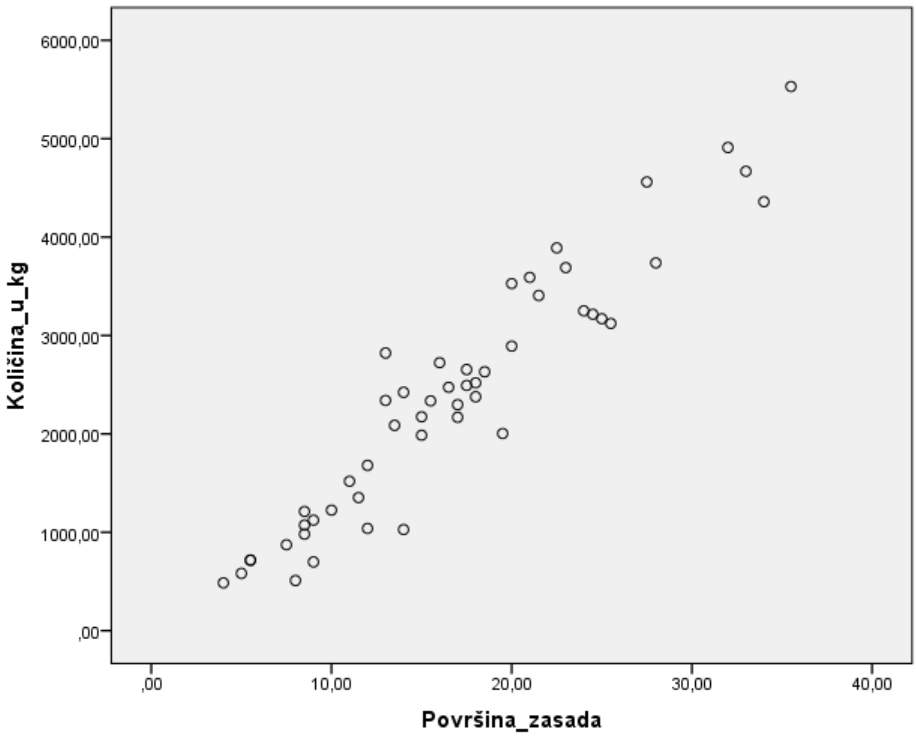
Dijagram rasipanja

Kada smo učitali podatke i definisali promenljive potrebno je da izradimo dijagram rasipanja na osnovu koga ćemo da zaključimo da li uopšte ima smisla tražiti linearnu vezu između promenljivih.

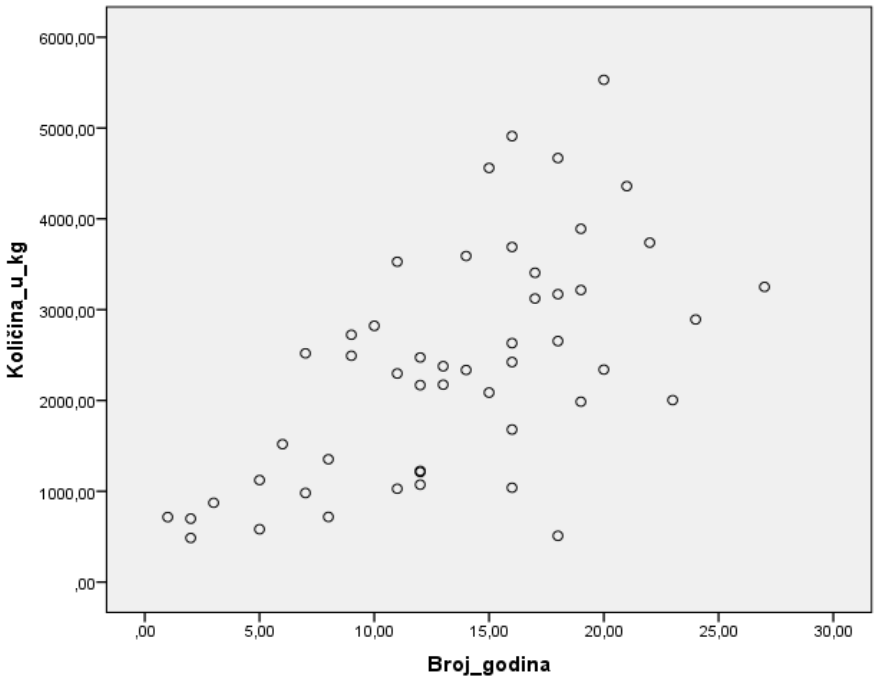
Dijagram rasipanja se konstruiše na sledeći način:

1) Iz glavnog menija izabrati komandu **Graphs**

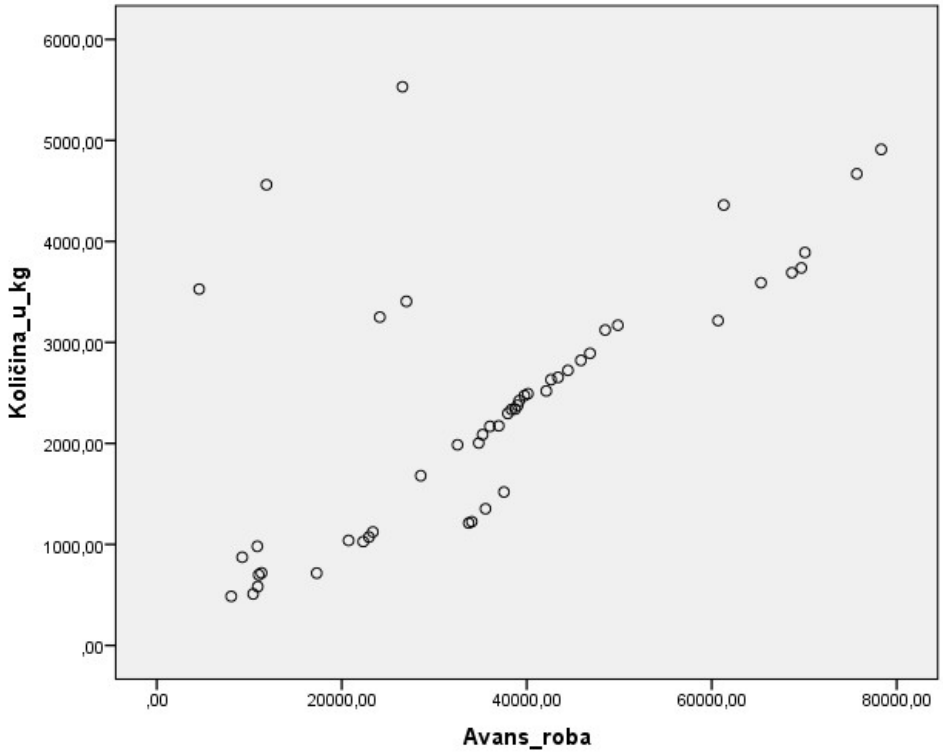
2) Pritiskom na stavku **Legacy Dialogs** otvara se galerija dijagrama iz koje se bira **Scatterplot** i njegovi osnovni elementi (dve promenljive čiju zavisnost ispitujemo)



Dijagram 5



Dijagram 6



Dijagram 7

Sa Dijagrama5 vidimo da je količina proizvedene maline u tesnoj vezi sa površinom pod zasadom maline, pa stoga tu možemo pretpostaviti da postoji tesna linearna veza. Takođe, možemo zaključiti da je u pitanju pozitivna linearna veza. Sa Dijagrama6 vidimo da tu najverovatnije ne postoji linearna veza, ali još ćemo ispitati preko drugih parametara. Dijagram7 ukazuje na relativno moguću linearnu vezu.

Pirsonov koeficijent korelacije

Pored ovih dijagrama ćemo još izračunati i Pirsonov koeficijent korelacije, koji je objašnjen u teorijskom delu rada. Postupak za generisanje Pirsonovog koeficijenta korelacije je sledeći:

1) Iz linije glavnog menija izabrati opciju **Analyze**

2) Izabrati stavku **Correlate**, a zatim **Biavariate** da bi se otvorio okvir za dijalog **Bivariate Correlations**.

3) Odabirom željenih promenljivih i pritiskom na dugme sa strelicom udesno se te promenljive prenose u polje **Variables**

4) Potrebno je da bude potvrđeno polje **Pearson correlation coefficient**

5) U delu **Test of Significance** odabrati dugme **One-taled**

6) Pritiskom na dugme **OK** realizuje se zahtev za računanje Pirsonovog koeficijenta korelacije

Correlations

	Količina_u_kg	Površina_zasada

Količina_u_kg	Pearson Correlation	1	,950**
	Sig. (1-tailed)		,000
	N	50	50
Površina_zasada	Pearson Correlation	,950**	1
	Sig. (1-tailed)	,000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 5

Rezultati Tabele5 potvrđuju zaključak koji je izveden iz tumačenja dijagrama rasipanja-da između promenljivih Količina_u_kg i Površina_zasada postoji značajna pozitivna veza (r=0,950).

Correlations

		Količina_u_kg	Broj_godina
Količina_u_kg	Pearson Correlation	1	,608**
	Sig. (1-tailed)		,000
	N	50	50
Broj_godina	Pearson Correlation	,608**	1
	Sig. (1-tailed)	,000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 6

Tumačenjem Tabele6 vidimo da postoji linearna veza i između promenljivih Količina_u_kg i Broj_godina (r=0,608), što bi značilo da oni proizvođači koji imaju veću količinu maline rade sa njom veći broj godina. Međutim, već sada možemo da primetimo da je ovo najslabija veza.

Correlations

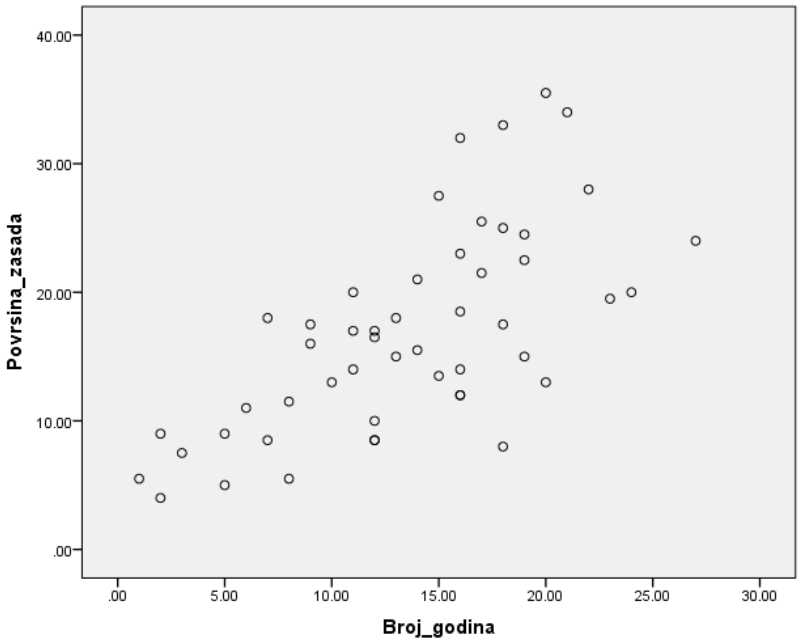
		Količina_u_kg	Avans_roba
Količina_u_kg	Pearson Correlation	1	,653**
	Sig. (1-tailed)		,000
	N	50	50
Avans_roba	Pearson Correlation	,653**	1
	Sig. (1-tailed)	,000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 7

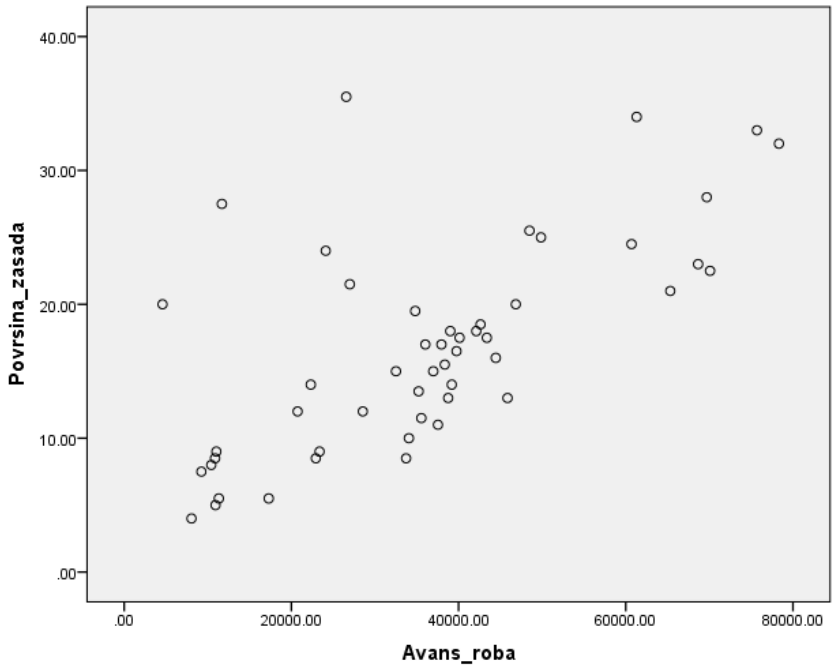
Na osnovu Tabele7 veza postoji i između promenljivih Količina_u_kg i Avans_roba (r=0,653). To znači da oni koji više ulože u svoj malinjak imaju i bolji prinos. Ova veza je neznatno jača od prethodne.

Potrebno je proveriti i da li između nezavisnih promenljivih postoji korelacija.



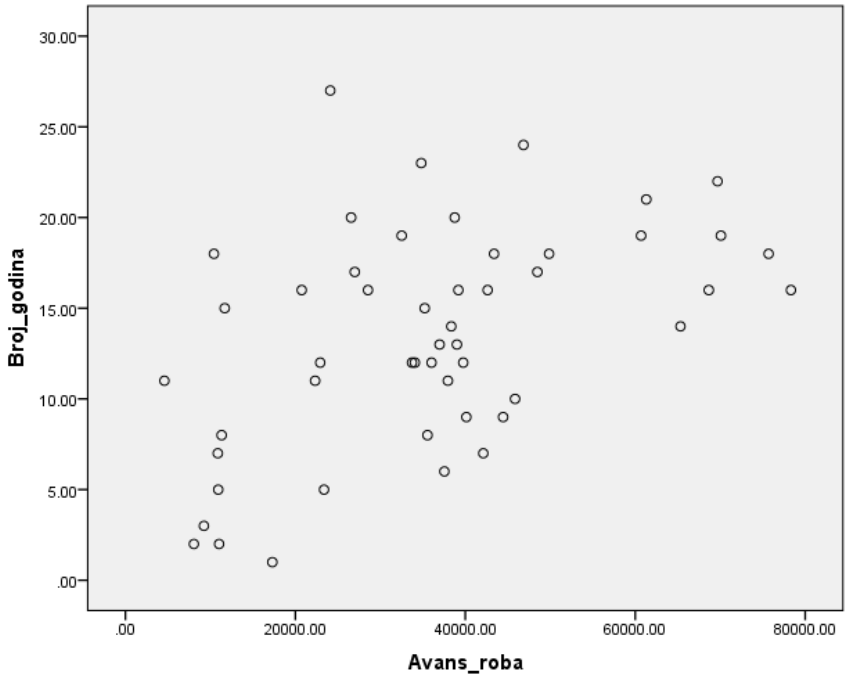
Dijagram 8

Sa Dijagrama8 vidimo da su promenljive Površina_zasada i Broj_godina u mogućoj korelaciji, pa možemo pretpostaviti da će program iz modela izbaciti promenljivu Broj_godina.



Dijagram 9

Sa Dijagrama9 takođe možemo zaključiti da postoji korelacija između promenljivih Površina_zasada i Avans_roba.



Dijagram 10

Dijagram10 pokazuje da ne bi trebalo da postoji zavisnost između promenljivih Broj_godina I Avans_roba, ali to još treba ispitati.

Ove tri tvrdnje ćemo proveriti pomoću Pirsonovog testa korelacije.

Correlations

		Povrsina_zasada	Broj_godina
Povrsina_zasada	Pearson Correlation	1	.672**

	Sig. (1-tailed)		.000
	N	50	50
	Pearson Correlation	.672**	1
Broj_godina	Sig. (1-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 8

Correlations

		Povrsina_zasada	Avans_roba
	Pearson Correlation	1	.655**
Povrsina_zasada	Sig. (1-tailed)		.000
	N	50	50
	Pearson Correlation	.655**	1
Avans_roba	Sig. (1-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 9

Correlations

		Avans_roba	Broj_godina
Avans_roba	Pearson Correlation	1	.486**
	Sig. (1-tailed)		.000
	N	50	50
Broj_godina	Pearson Correlation	.486**	1
	Sig. (1-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (1-tailed).

Tabela 10

Na osnovu prethodnih tabela u kojima je izračunat Pirsonov koeficijent korelacije vidimo da su ove tri promenljive korelisane između sebe (čak su i promenljive Avans_roba i Broj_godina korelisane), pa na osnovu prethodnih rezultata, možemo pretpostaviti da će najbolji model biti jednostruki linearni model sa nezavisnom promenljivom Površina_zasada i zavisnom promenljivom Količina_u_kg. Promenljive Avans_roba i Broj_godina ćemo koristiti radi dalje ilustracije mogućnosti programa SPSS.

Konstrukcija linearnog regresionog modela

U ovom delu će biti prikazana konstrukcija jednostrukih i višestrukih linearnih regresionih modela metodama **standard**, **hijerarhijski** i **stepwise**, i biće izabran najbolji od njih.

Standard metoda

Model konstruišemo na sledeći način:

1) Izabrati opciju **Analyze** iz glavnog menija

2) Izabrati stavku **Regression**, a zatim **Linear**

3) U okviru za dijalog **Linear Regression** se odabrati zavisnu promenljivu i premestiti u polje **Dependent**

4) Pritisnuti dugme sa oznakom strelice udesno da biste ih preneli u polje **Independent**

5) U padajućoj listi **Method** potvrditi opciju **Enter**

6) Pritiskom na dugme Statistics se otvara podokvir za dijalog **Linear Regression** i potrebno je da budu potvrđena polja **Estimates** i **Model fit** u delu **Regression Coefficients**. U delu okvira **Residuals** potrebno je potvrditi polje **Casewise** diagnostics i polje **Outliers outside**.

7) Pritisnuti dugme **Continue**

8) Izabrati ***ZRESID** i pritisne dugme sa oznakom strelice udesno da bi se izabrana stavka prenela u polje **Y**

9) Izabrati ***ZPRED** i pritisnuti dugme sa oznakom strelice udesno da bi se izabrana stavka prenela u polje **X**

10) U delu okvira **Standardized Residual Plots** izabrati polje **Normal probability plot**

11) Pritisnuti dugme **Continue**

12) Pritisnuti dugme **Save** da bi se otvorio podokvir za dijalog **Linear Regression: Save**

13) Pritiskom na dugme **Continue**, a potom **OK** dobijamo linearni model sa svim parametrima

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Avans_roba, Broj_godina, Površina_zasada[b]	.	Enter

a. Dependent Variable: Količina_u_kg

b. All requested variables entered.

Tabela 11

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,952[a]	,907	,901	399,48668

a. Predictors: (Constant), Avans_roba, Broj_godina, Površina_zasada

b. Dependent Variable: Količina_u_kg

Tabela 12

Iz Tabele12 zaključujemo da sve tri nezavisne promenljive koje su uključene u model objašnjavaju 90,7% disperzije zavisne promenljive, odnosno proizvedene količine maline.Ovaj procenat disperzije je izuzetno značajan, što pokazuje i visoka vrednost F-statistike (148,831) u Tabeli13 :

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	71255797,047	3	23751932,349	148,831	,000[b]
	Residual	7341121,933	46	159589,607		
	Total	78596918,980	49			

a. Dependent Variable: Količina_u_kg

b. Predictors: (Constant), Avans_roba, Broj_godina, Površina_zasada

Tabela 13

Na osnovu F-testa i vrednosti $F^{\alpha}_{3,(n-4)}=8{,}59$ i na osnovu jednakosti $F^{\alpha}_{3,(n-4)}<F$ odbacujemo nultu hipotezu H_0 i vidimo da je procenat objašnjene disperzije izuzetno značajan (od strane bar jedne zavisne promenljive).

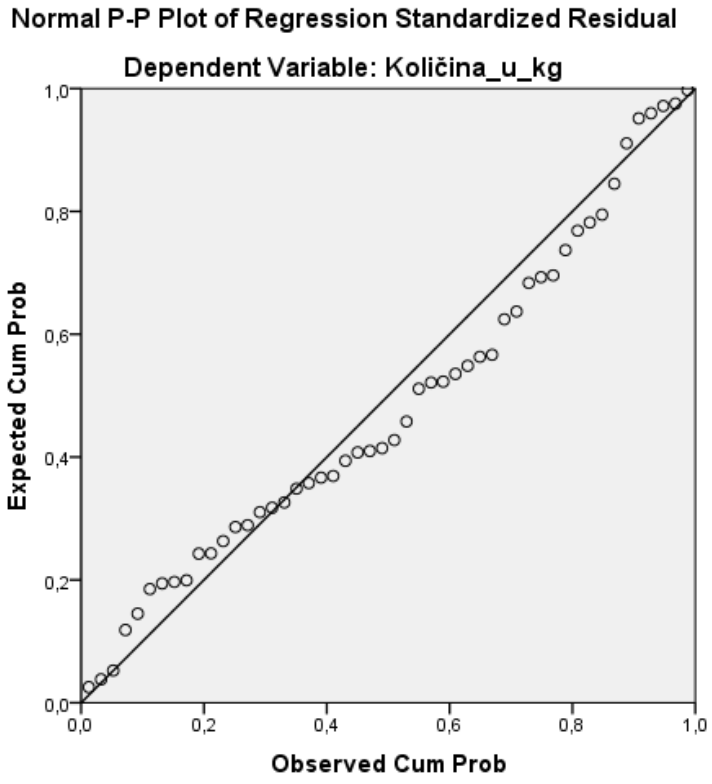
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-166,140	151,514		-1,097	,279
	Površina_zasada	153,157	11,350	,954	13,494	,000
	Broj_godina	-12,680	12,780	-,061	-,992	,326
	Avans_roba	,004	,004	,057	,952	,346

a. Dependent Variable: Količina_u_kg

Tabela 14

Posmatranjem t-vrednosti iz Tabele14 i tablične vrednosti $t_{0,5;46}=1,684$ možemo izvesti zaključak da na količinu proizvedene maline najviše utiče površina pod zasadom. Vrednost t-statistike za promenljivu Broj_godina je negativna i ona nema uticaja. Vrednost t-statistike za promenljivu Avans_roba je 0,952, što je manje od tablične $t_{0,5;46}$. Zaključak je da u model trebamo uključiti samo nezavisnu promenljivu Površina_zasada.



Dijagram 11

Dijagram normalnosti raspodele za standardizovane reziduale (Dijagram11) ukazuje na postojanje relativno normalne raspodele reziduala, što dokazuju i sledeći testovi normalnosti ($d < d_{50;0,05}$ i svaka vrednost u koloni Sig za Šapiro-Vilk test je veća od 0,05, Tabela15).

Tests of Normality

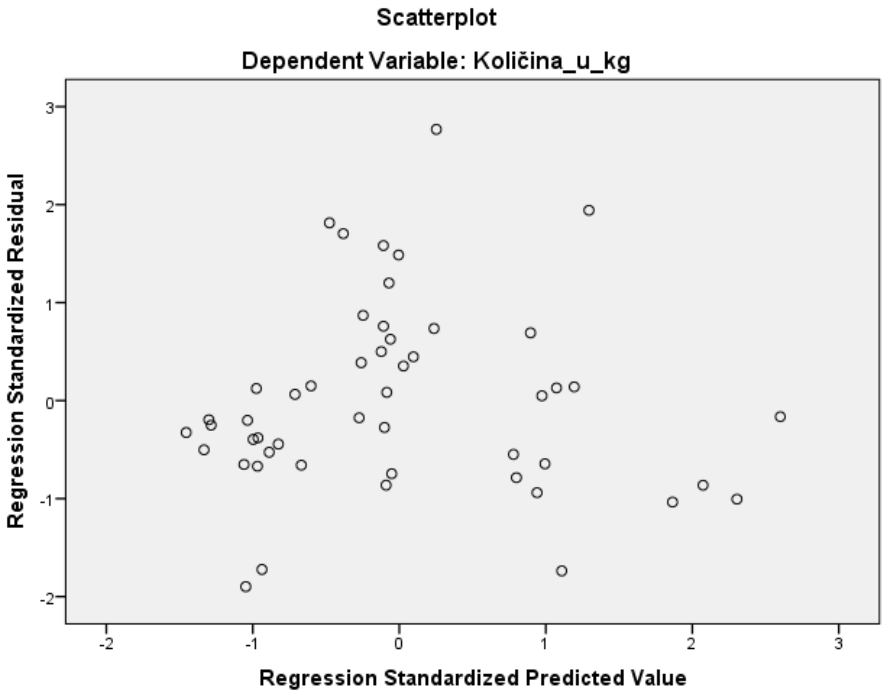
	Kolmogorov-Smirnov[a]	Shapiro-Wilk

	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.098	50	.200*	.989	50	.923

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabela 15



Dijagram 12

Na dijagramu rasipanja reziduala u odnosu na modelom predviđene vrednosti (Dijagram 12) može se zapaziti da ne postoji jasna zavisnost između reziduala i predviđenih vrednosti, što je u skladu sa pretpostavkom o linearnosti.

Na osnovu svih ovih prethodnih rezultata , možemo zaključiti da uticaj na disperziju količine proizvedene maline ima samo površina pod zasadom. Uticaj ostale dve promenljive nije značajan.Model, za koji već sada znamo da nije dobar, a koji je formiran, jeste:

$$Y = -166{,}40 + 153{,}157 * X_1 - 12{,}680 * X_2 + 0{,}004 * X_3$$

Standard metodom će, u skladu sa prethodnim zaključcima, biti postavljeni sledeći jednostruki linearni regresioni modeli, u kojima je zavisna promenljiva Količina_u_kg:

1) $Y = a X_1 + b$

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Povrsina_zasada[b]	.	Enter

a. Dependent Variable: Kolicina_u_kg

b. All requested variables entered.

Tabela 16

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.950[a]	.903	.901	398.44341	1.916

a. Predictors: (Constant), Povrsina_zasada

b. Dependent Variable: Kolicina_u_kg

Tabela 17

Iz Tabele17 i Tabele18 vidimo da je ovim modelom objašnjeno 90,3% disperzije zavisne promenljive Količina_u_kg, kao i da je veoma visoka vrednost F-statistike u tabeli za ANOVU. Tablična vrednost $F_{1,48}=251,144$, što je dosta manje od emprijske vrednosti F-statistike. Dakle, ovaj model je dosta dobar.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70976575.840	1	70976575.840	447.076	.000[b]
	Residual	7620343.140	48	158757.149		
	Total	78596918.980	49			

a. Dependent Variable: Kolicina_u_kg

b. Predictors: (Constant), Povrsina_zasada

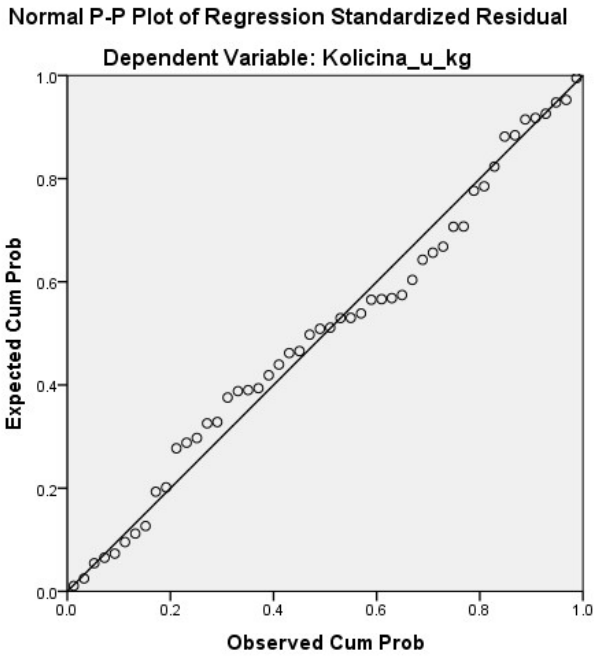
Tabela 18

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-189.947	133.843		-1.419	.162
	Povrsina_zasada	152.614	7.218	.950	21.144	.000

a. Dependent Variable: Kolicina_u_kg

Tabela 19



Dijagram 13

Tests of Normality

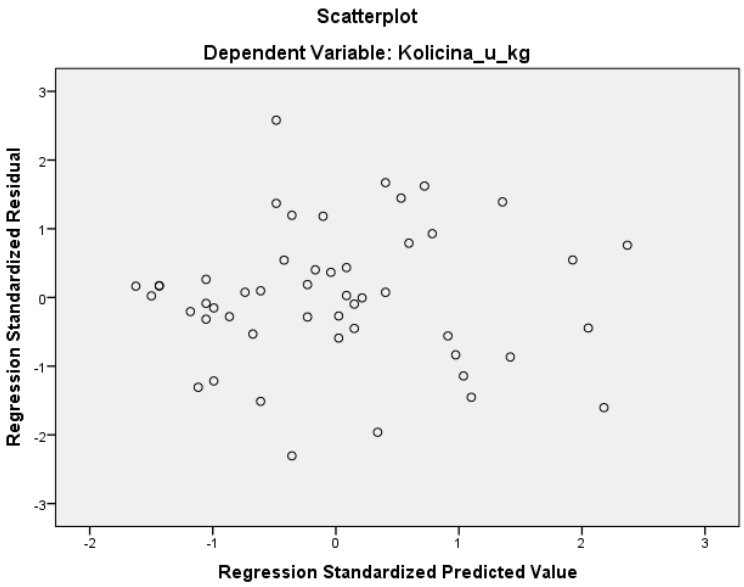
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.085	50	.200*	.986	50	.831

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabela 20

Na osnovu Tabele20 i Dijagrama13, analogno kao i u prethodnim primerima, zaključujemo da su reziduali normalno raspodeljeni.



Dijagram 14

Na dijagramu rasipanja reziduala (Dijagram14) može se videti da ne postoji jasna zavisnost između reziduala I predviđenih vrednosti, pa je pretpostavka o linearnosti korektna.

Dobijeni model je:

$$Y = -189{,}947 + 152{,}614\, X_1$$

2) $Y = a\, X_2 + b$

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Broj_godina[b]	.	Enter

a. Dependent Variable: Kolicina_u_kg

b. All requested variables entered.

Tabela 21

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.608[a]	.370	.357	1015.76921	1.989

a. Predictors: (Constant), Broj_godina

b. Dependent Variable: Kolicina_u_kg

Tabela 22

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29071138.931	1	29071138.931	28.176	.000[b]
	Residual	49525780.049	48	1031787.084		
	Total	78596918.980	49			

a. Dependent Variable: Kolicina_u_kg

b. Predictors: (Constant), Broj_godina

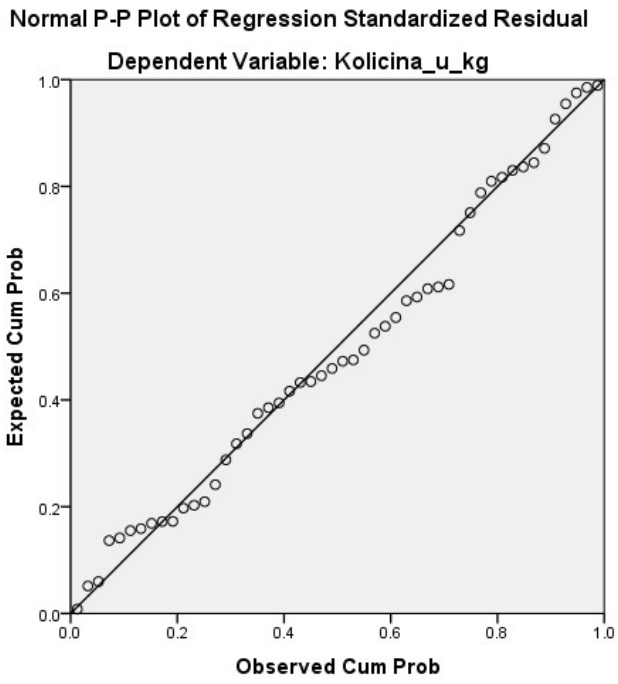
Tabela 23

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	658.743	354.153		1.860	.069
	Broj_godina	127.280	23.979	.608	5.308	.000

a. Dependent Variable: Kolicina_u_kg

Tabela 24



Dijagram 15

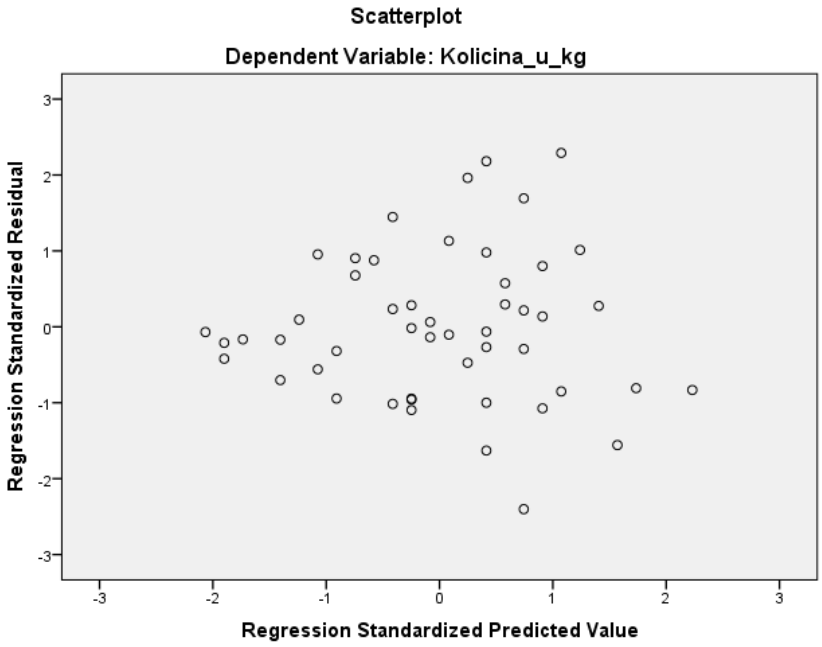
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.103	50	.200*	.980	50	.547

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabela 25



Dijagram 16

Na osnovu svih prethodnih tabela za ovaj model, možemo zaključiti da model nije dobar (R-square je 0,370 ,a F-statistika je jako mala). Reziduali su normalno raspodeljeni, ali ovaj model nije korektan i odbacujemo ga.

3) $Y = a X_3 + b$

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method

1	Avans_roba[b]	.	Enter

a. Dependent Variable: Kolicina_u_kg

b. All requested variables entered.

Tabela 26

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.652[a]	.425	.413	970.12812	2.340

a. Predictors: (Constant), Avans_roba

b. Dependent Variable: Kolicina_u_kg

Tabela 27

ANOVA[a]

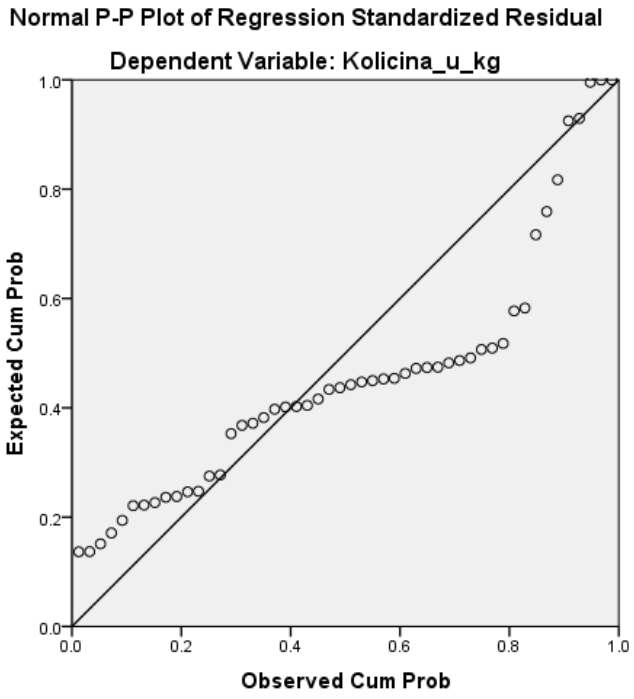
Model	Sum of Squares	df	Mean Square	F	Sig.

	Regression	33421787.561	1	33421787.561	35.512	.000[b]
1	Residual	45175131.419	48	941148.571		
	Total	78596918.980	49			

a. Dependent Variable: Kolicina_u_kg

b. Predictors: (Constant), Avans_roba

Tabela 28



Dijagram 17

Tests of Normality

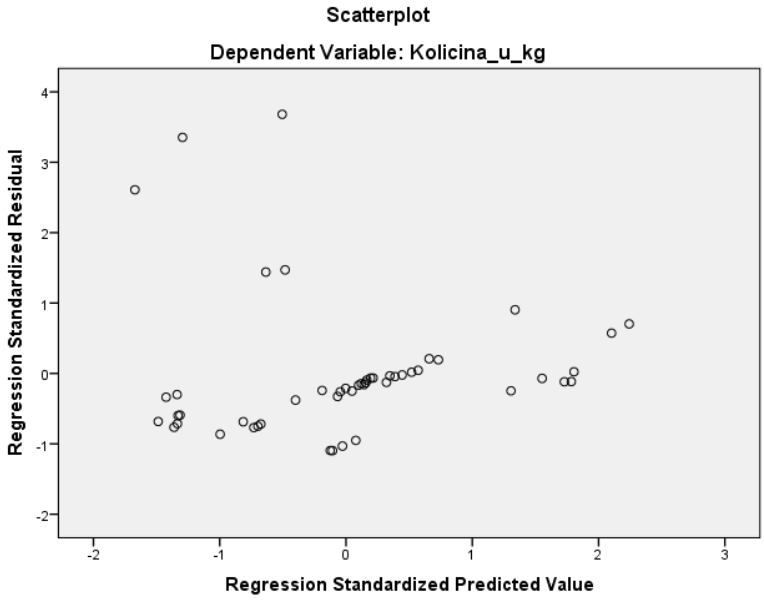
	Kolmogorov-Smirnov[a]	Shapiro-Wilk

	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.282	50	.000	.739	50	.000

a. Lilliefors Significance Correction

Tabela 29

Vidimo da za ovaj model reziduali nisu normalno raspodeljeni ni po Kolmogorov-Smirnov testu, a ni po Šapiro-Vilkovom.



Dijagram 18

Iako su vrednosti za R square (Tabela27) I za F-statistiku (Tabela28) neznatno veći u odnosu na prethodni model, na osnovu svih ostalih parametara ovaj model je lošiji, pa I njega možemo odbaciti.

Hijerarhijska metoda

Postupak za sprovođenje hijerarhijske regresione analize

1) Izabrati opciju **Analyze** iz glavnog menija

2) Odabrati stavku **Regression**, a zatim **Linear** da bi se otvorio okvir za dijalog **Linear Regression**

3) Iz liste promenljivih odabrati zavisnu promenljivu i pritiskom na dugme sa oznakom strelice udesno, promenljiva se prenese u polje **Dependent**

4) Iz liste promenljivih odabrati nezavisnu promenljivu za koju se, na osnovu prethodnih statističkih saznanja, smatra da treba prva da uđe u regresioni model i pritiskom na dugme sa oznakom strelice udesno se prenese u polje **Independent**

5) Iz liste promenljivih odabrati sledeću nezavisnu promenljivu za koju se smatra da treba da uđe u regresioni model i pritiskom na dugme sa oznakom strelice udesno se prenosi u polje **Independent**

6) Iz liste promenljivih odabrati sledeću nezavisnu promenljivu za koju se smatra da treba da uđe u regresioni model i pritiskom na dugme sa oznakom strelice udesno se prenosi u polje **Independent**. U delu okvira iznad treba da piše **Block 3 of 3**

7) Pritiskom na dugme **Statistics** se otvara podokvir za dijalog **Linear Regression:Statistics** i potrebno je da budu potvrđena polja **Estimates**, **Model fit** i **R squared change**.

8) Pritiskom na dugme **Continue** i **OK** dobijamo rezultate ispitivanja

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Površina_zasada[b]	.	Enter
2	Avans_roba[b]	.	Enter
3	Broj_godina[b]	.	Enter

a. Dependent Variable: Količina_u_kg

b. All requested variables entered.

Tabela 30

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,950[a]	,903	,901	398,44341	,903	447,076	1	48	,000
2	,951[b]	,905	,901	399,42096	,002	,765	1	47	,386
3	,952[c]	,907	,901	399,48668	,002	,985	1	46	,326

a. Predictors: (Constant), Površina_zasada

b. Predictors: (Constant), Površina_zasada, Avans_roba

c. Predictors: (Constant), Površina_zasada, Avans_roba, Broj_godina

d. Dependent Variable: Količina_u_kg

Tabela 31

Iz Tabele31 vidimo da nezavisna promenljiva Površina_zasada koja je prva ušla u regresioni model objašnjava 90,3% disperzije zavisne promenljive Količina_u_kg, što znači da predstavlja značajan prediktor zavisne promenljive. U sledećem koraku, kada nezavisna promenljiva Avans_roba ulazi u model, dolazi do promene koeficijenta determinacije-**R Squared Change**.

U tabeli za analizu disperzije (Tabela32), analogno kao i za prethodni model, možemo primetiti da je uticaj disperzije nezavisnih promenljivih na disperziju zavisne promenljive veliki. Međutim, na osnovu zaključaka iz prethodne tabele znamo da je naveći procenat tog uticaja zapravo objašnjava promenljiva Površina_zasada.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70976575,840	1	70976575,840	447,076	,000[b]
	Residual	7620343,140	48	158757,149		
	Total	78596918,980	49			
2	Regression	71098675,157	2	35549337,579	222,828	,000[c]
	Residual	7498243,823	47	159537,103		
	Total	78596918,980	49			

3	Regression	71255797,047	3	23751932,349	148,831	,000[d]
	Residual	7341121,933	46	159589,607		
	Total	78596918,980	49			

a. Dependent Variable: Količina_u_kg

b. Predictors: (Constant), Površina_zasada

c. Predictors: (Constant), Površina_zasada, Avans_roba

d. Predictors: (Constant), Površina_zasada, Avans_roba, Broj_godina

Tabela 32

Na osnovu **t-testa** koji je objašnjen u teorijskom delu rada, vrednosti $t_{\alpha,(n-2)}=1{,}684$, i Tabele15 možemo zaključiti da promenljiva Površina_zasada, koja je prva uključena u model ima najviše značajnosti u objašnjenju disperzije promenljive Količina_u_kg, i njen uticaj se smanjuje kako se koja nova nezavisna promenljiva uključuje u model, ali je i dalje velika. Ostale dve promenljive ne utiču na disperziju zavisne promenljive što možemo primetiti na osnovu jako male t-vrednosti (Tabela33).

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		

1	(Constant)	-189,947	133,843		-1,419	,162
	Površina_zasada	152,614	7,218	,950	21,144	,000
2	(Constant)	-224,157	139,754		-1,604	,115
	Površina_zasada	147,121	9,580	,916	15,357	,000
	Avans_roba	,004	,004	,052	,875	,386
3	(Constant)	-166,140	151,514		-1,097	,279
	Površina_zasada	153,157	11,350	,954	13,494	,000
	Avans_roba	,004	,004	,057	,952	,346
	Broj_godina	-12,680	12,780	-,061	-,992	,326

a. Dependent Variable: Količina_u_kg

Tabela 33

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Avans_roba	,052[b]	,875	,386	,127	,570

	Broj_godina	-,056[b]	-,919	,363	-,133	,548
2	Broj_godina	-,061[c]	-,992	,326	-,145	,545

a. Dependent Variable: Količina_u_kg

b. Predictors in the Model: (Constant), Površina_zasada

c. Predictors in the Model: (Constant), Površina_zasada, Avans_roba

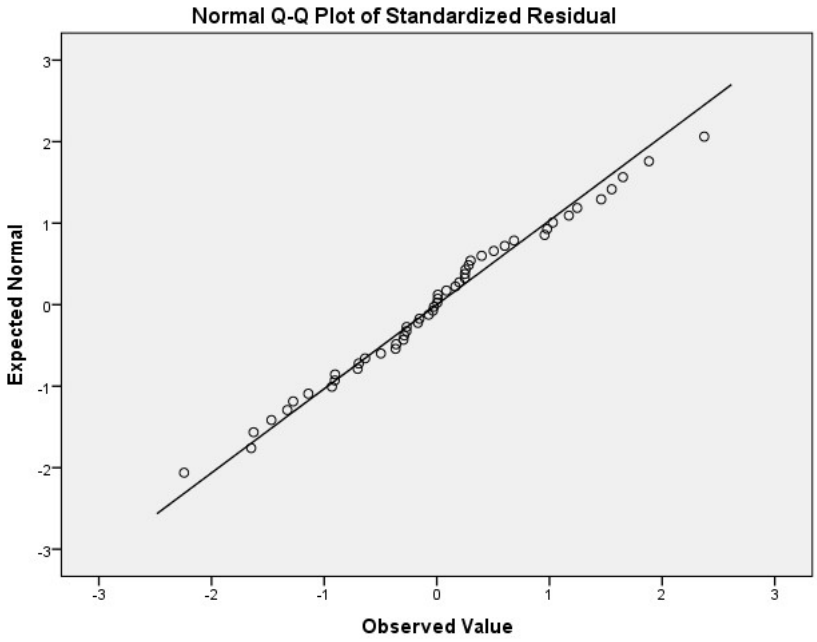
Tabela 34

Očekivano, u tabeli33 vidimo da varijable Broj_godina i Avans_roba ne ispunjavaju statističke kriterijume za uključenje u regresioni model, na šta ukazuje t-vrednost koja nije značajna, pa na osnovu Tabele34 zaključujemo da su modeli koji su konstruisani ovom metodom:

$$Y = -189{,}947 + 152{,}614 * X_1$$

i

$$Y = -166{,}140 + 153{,}157 * X_1 + 0{,}004 * X_3$$



Dijagram 19

Dijagram normalnosti raspodele za standardizovane reziduale (Dijagram19) ukazuje na postojanje normalne raspodele, što dokazuje i test normalnosti (Tabela35).

Tests of Normality

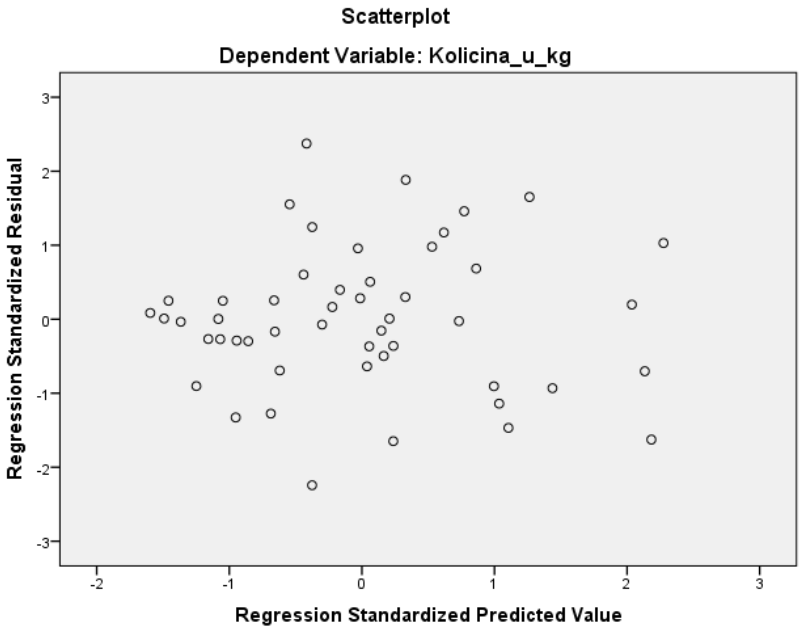
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.

Standardized Residual	.098	50	.200*	.989	50	.923

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabela 35



Dijagram 20

Na dijagramu rasipanja reziduala u odnosu na modelom predviđene vrednosti (Dijagram 20) možemo videti da ne postoji jasna zavisnost između reziduala i predviđenih vrednosti, pa možemo zaključiti da je prisutna linearnost.

Stepwise metoda

Postupak za sprovođenje stepwise regresione analize:

1) Odabrati opciju **Analyze** iz glavnog menija

2) Odabrati stavku **Regression**, a zatim **Linear** da bi se otvorio okvir za dijalog **Linea**r **Regression**

3) Izabrati zavisnu promenljivu i pritiskom na dugme sa oznakom strelice udesno se prenese u polje **Dependent**

4) Analogno, odabrati nezavisne promenljive i preneti u polje **Independent**

5) U padajućoj listi **Method** potvrditi opciju **Stepwise**

6) Pritiskom na dugme **Statistics** se otvara podokvir za dijalog **Linear Regression:Statistics** i potrebno je da budu potvrđena polja **Estimates** i **Model fit** u delu **Regression Coefficients**

7) Pritiskom na dugme **Continue**, a zatim **OK** realizuje se tražena metoda

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Površina_zasada	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= , 100).

a. Dependent Variable: Količina_u_kg

Tabela 36

Na osnovu Tabele36 vidimo da promenljive Avans_roba i Broj_godina nisu ni ušle u regresioni model, već samo promenljiva Površina_zasada.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,950[a]	,903	,901	398,44341

a. Predictors: (Constant), Površina_zasada

b. Dependent Variable: Količina_u_kg

Tabela 37

Na osnovu Tabele37 vidimo da je disperzija zavisne promenljive 90,3% objašnjena kada je u njemu samo nezavisna promenljiva Površina_zasada.Ovaj procenat objašnjene disperzije je veoma značajan, što pokazuje i visoka vrednost F-statistike (F=447,076) u Tabeli 38.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	70976575,840	1	70976575,840	447,076	,000[b]
	Residual	7620343,140	48	158757,149		
	Total	78596918,980	49			

a. Dependent Variable: Količina_u_kg

b. Predictors: (Constant), Površina_zasada

Tabela 38

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-189,947	133,843		-1,419	,162
	Površina_zasada	152,614	7,218	,950	21,144	,000

a. Dependent Variable: Količina_u_kg

Tabela 39

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	Broj_godina	-,056[b]	-,919	,363	-,133	,548
	Avans_roba	,052[b]	,875	,386	,127	,570

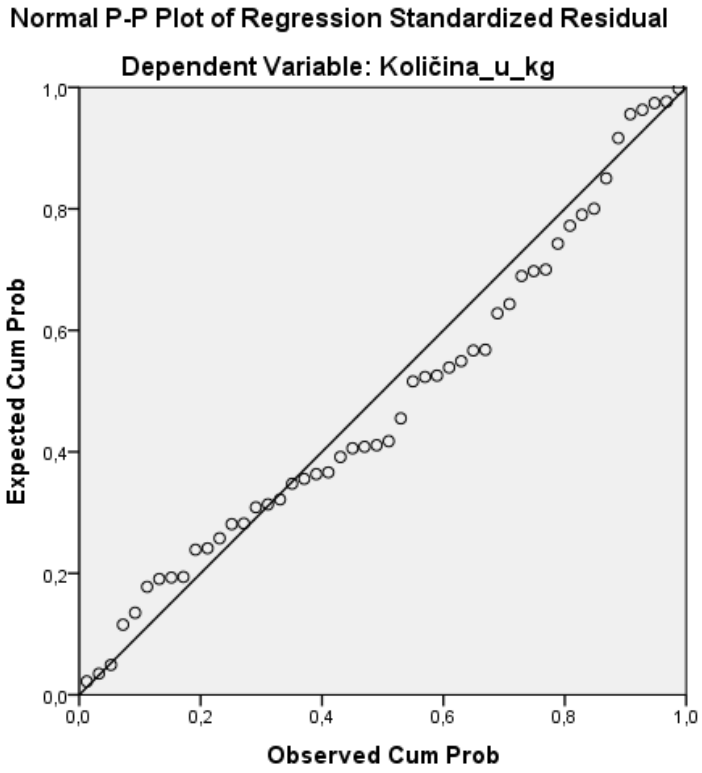
a. Dependent Variable: Količina_u_kg

b. Predictors in the Model: (Constant), Površina_zasada

Tabela 40

I kao što je pretpostavljeno, na osnovu Tabele39 i Tabele40 vidimo zašto su ostale promenljive izbačene iz regresionog modela i zaključujemo da je model koji je konstruisan:

$$Y = -189{,}947 + 152{,}614 * X_1$$



Dijagram 21

Tests of Normality

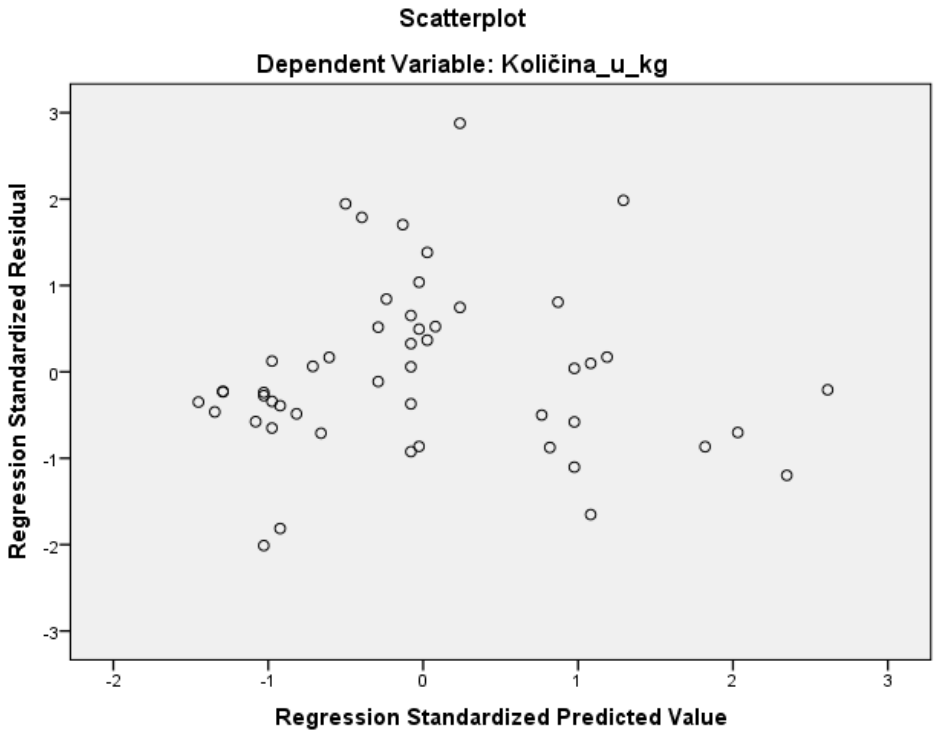
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	.085	50	.200*	.986	50	.831

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tabela 41

Reziduali su normalno raspodeljeni (Tabela41 i Dijagram21). Na dijagramu rasipanja reziduala (Dijagram22) ne postoji jasna zavisnost između reziduala i predviđenih vrednosti, što je u skladu sa pretpostavkom o linearnosti.



Dijagram 22

Zaključak

Na osnovu dobijenih rezultata možemo zaključiti da na količinu proizvedene maline najviše utiče površina pod zasadom, što je i logično. To znači da se malinjaci uglavnom dobro održavaju zato što važi da se sa povećanjem površine pod zasadom povećava i proizvedena količina. Uglavnom, na osnovu istraživanja možemo dati sve odgovore na pitanja iz uvoda. Model koji najbolje opisuje vezu između promenljivih, po svim statističkim kriterijumima, je sledeći:

$$Y = -189{,}947 + 152{,}614 * X_1$$

Najbolji model ima najveću R-square i koeficijenti mu imaju najveću značajnost. Sve ove rezultate smo dobili brzo i jednostavno uz pomoć programa SPSS. Ovaj softver omogućava jasno očitavanje dobijenih rezultata i statističko zaključivanje na osnovu njih. Ono što je obrađeno u ovom radu je samo jedan mali deo širokog spektra mogućnosti koje pruza program SPSS. U njemu je moguće obrađivati i ispitivati na osnovu dobijenih rezultata, istraživačke scenarije različitog stepena složenosti. Statistički softver SPSS daje mogućnost studentima da sprovode različite statističke analize kod kuće ,i to mnogo jednostavnije i brže nego, na primer, uz programski jezik R. On je pogotovo pogodan za istraživače koji nemaju mnogo prakse iz programiranja.Međutim, zbog velike sličnosti programa SPSS i Excela, sasvim je moguće da će novije verzije Excela vremenom sasvim potisnuti SPSS sa informacionog tržišta.

Literatura

(1) „Statističke metode u meterologiji i inženjerstvu“, autori dr Vesna Jevremović i dr Jovan Mališić;
(2) „Statistička kontrola kvaliteta“- autori dr Milan Eremić i dr Radmila Njegić;
(3) „SPSS 20.0 for Windows“-autor Sheridan J. Coakes;
(4) http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis
(5) http://en.wikipedia.org/wiki/Linear_regression
(6) http://www.pmf.ni.ac.rs/pmf/predmeti/5116/predavanja/15%20Regresija.pdf

(7) https://www.google.rs/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&ved=0CFwQFjAJ&url=http%3A%2F%2Fwww.unizd.hr%2Fportals%2F4%2Fnastavni_mat%2F2_godina%2Fstatistika%2Fstatistika_09.ppt&ei=pm0xVLG4Aaf9ywOPx4G4Cw&usg=AFQjCNFAgilWt1z5nL0wP5ysguOPi0xUKA&sig2=RbWWRfeInWYs9j9cc1IlVQ
(8) http://www.ef.uns.ac.rs/Download/multivarijaciona-statisticka-analiza/2013-02-08Visestruka-regresija-i-korelacija.pdf
(9) http://sh.wikipedia.org/wiki/Korelacija
(10) Uvod u programski paket SPSS- Doc. Dr Dragan Bogdanović
(11) Metode regresijske analize-Fran Galetić