PRIMJENA REGRESIONE I KORELACIONE ANALIZE U MS EXCELU

APPLICATION OF REGRESSION AND CORRELATION ANALYSIS IN MS EXCEL

Lejla Dacić
Haris Dacić

SAŽETAK

Mnoge pojave u poslovanju, privredi i drugim područjima djelatnosti u međusobnoj su vezi. Prilikom istraživanja međusobnih veza dviju promjenjivih primjenjuju se modeli proste (linearne i krivolinijske) regresione i korelacione analize, a u slučaju posmatranja više promjenjivih metoda višestruke regresione i korelacione analize. U ovom radu je pokazano na koji način se istražuju veze među pojavam. U tu svrhu obrađeni su jednostavniji modeli regresione i korelacione analize. Činjenica je da se svakodnevno pojavljuju mnogobrojni računarski programi koji se intenzivno koriste i olakšavaju edukaciju kadrova, što omogućava njihovu dostupnost velikom broju korisnika. Oni su do te mjere usavršeni da je njihova primjena moguća i u slučajevima kada čitalac nije profesionalno obučen za rad sa statističkim metodama. Korištenje statističkih programa u MS Excel-u predstavlja veliki napredak u statističkoj obradi podataka uz značajnu uštedu vremena. Iz tog razloga rad je koncipiran tako da ukratko pruži uputstva za korištenje statističkih metoda u Microsoft Excelu. Postepeno je objašnjeno na koji način se može doći do rješenja problema, matematički i primjenom statističkog programa u Excel-u. Prostim primjerima zadatka je pokazana lakoća kojom se dolazi do konačnih rezultata što predstavlja značajnu uštedu u vremenu u odnosu na klasični način određivanja parametara.

Ključne riječi: regresiona analiza, korelaciona analiza, stadardna greška regresije , koeficijent varijacije, koeficijent determinacije.

SUMMARY

Many phenomena in business, economy and other areas of activity are interrelated. When investigating the mutual connections of two variables, models of simple (linear and curvilinear) regression and correlation analysis are applied, and in the case of observing several variable methods, multiple regression and correlation analysis. This paper shows how the connections between phenomena are evaluated. For this purpose, simple models of regression and correlation analysis were processed. The fact is that numerous computer programs appear every day, which are intensively used and facilitate the education of staff, which enables their availability to a large number of users. They are so advanced that their application is possible even in cases when the reader is not professionally trained to work with statistical methods. The use of statistical programs in MS Excel represents a major advance in statistical data processing with significant time savings. For this reason, the paper is designed to briefly provide instructions for using statistical methods in Microsoft Excel. It is gradually explained how to solve a problem, mathematically and by applying a statistical program in Excel. Simple examples of the task show the ease with which the final results are obtained, which represents a significant saving in time compared to the classical method of determining parameters.

Keywords: regression analysis, correlation analysis, standard error, coefficient of variation, coefficient of determination.

UVOD

Statistika je naučna disciplina koja na organizovan način pristupa prikupljanju, selekciji, grupisanju, prezentaciji i analizi informacija ili podataka, te interpretiranju rezultata provedene analize, a u svrhu realizacije postavljenih istraživačkih ciljeva. Da bi se ostvarili navedeni zadaci i ciljevi, statistika zahtijeva sopstveni instrumentarij: statističke metode i tehnike. Jedna od osnovnih analiza kojom se koristi statistika je regresiono-korelaciona analiza.

Danas većina stanovnika ima mogućnost korištenja kompjutera, metoda papir-olovka-kalkulator neizbježno zastarjeva i ustupa mjesto različitim računarskim programima. Iako postoje «jači» računarski statistički programi (SPSS, SAS, STATISTICA, MINITAB, STATGRAPHICS, SYSTAT,MYSTAT, itd.) odabran je Ms Excel iz praktičnog razloga. Ovaj program je sadržan u svakom paketu office programa te je dostupan svakom korisniku novijih Windowsa, a i veliki dio korisnika računara ima neka osnovna znanja iz Excela. Korištenje ovih metoda podrazumijeva poznavanje statistike a djelimično i Excela-a. Za osobe koje nedovoljno poznaju statističke metode preporučljivo je uporedo praćenje ovih programa i nekog od udžbenika statistike. Excel može zadovoljiti većinu potreba vezanih uz statističu analizu iako u odnosu na neke programe nedostaju mu neki neparametarski testovi, a i neki oblici statističke analize zahtijevaju dodatne korake u svrhu donošenja konačnih zaključaka. Osim toga, odličan je za tablički i grafički prikaz podataka. Kao drugo, većina stručnih i znanstvenih, diplomskih, magistarskih i doktorskih radova sadrži i statističku analizu prikupljenih podataka, te prezentaciju i interpretaciju dobivenih statističkih vrijednosti. Iz tog razloga, se može smatrati da se znanje koje se stekne prilikom izrade ovog rada može korisiti u mnogim segmentima za dalju edukaciju. Dalje, stručna literature pretrpana je raznovrsnim statističkim metodama i testovima, što statistiku čini kompliciranijom nego što ona stvarno jeste. Zbog toga je u radu ukazano i na prednosti koji nude računarski programi, odnosno Ms Excel.

Mnoge pojave u poslovanju, privredi i dr. područjima djelatnosti u međusobnoj su vezi. Prilikom istraživanja međusobnih veza dviju promjenjivih primjenjuju se modeli proste (linearne i krivolinijske) regresione i korelacione analize, a u slučaju posmatranja više promjenjivih metoda višestruke regresione i korelacione analize.U ovom radu osvrt je bio na prostoj regresionoj i korelacionoj analizi. Svrha korelacione analize je da se ispita da li između varijacija posmatranih pojava postoji slaganje i ako postoji u kom stepenu. Cilj regresije je da se utvrdi priroda veze, odnosno oblik zavisnosti između posmatranih pojava, što se postiže pomoću regresionog modela. On pokazuje prosječno slaganje varijacija posmatranih pojava i sredstvo je pomoću koga možemo da ocjenimo i predvidimo ponašanje zavisne promjenljive za željene vrijednosti nezavisne promjenjive.

U drugom poglavlju rada biće riječi o analizi veza među pojavama, odnosno o funkcionalnoj i stohastičnoj vezi među pojavama, direktnoj i indirektnoj te općenito o regresionoj i korelacionoj analizi i njihovim ciljevima.

U trećem dijelu rada govorit će se o linearnom modelu regresije, te kako matematičkim putem doći do rješenja problema. Zatim, na primjeru je pokazan grafički prikaz regresione analize, te objašnjeno na koji način se može doći do rješenja problema primjenom Excela odnosno izračunavanje parametara a i b te regresione prave.

U četvrtom dijelu će za razliku od linearnog modela regresije biti riječi o krivolinijskom modelu regresije. Ravnomjerne promjene nezavisne varijable ne uzrokuju ravnomjerne promjene zavisne varijable, te je u u tom slučaju potrebno je pronaći neku drugu funkciju (nelinearnu) koja najbolje pokazuje vezu između zavisne i nezavisne varijable. Najjednostavniji oblici krivolinijske regresije su model jednostavne eksponencijalne

regresije i dvostruko logaritamski model regresije (POWER) i bit će prikazani preko primjera.
Peti dio govori o mjerama reprezentativnosti. Da bismo odredili reprezentativnost i pouzdanost ocijenjenog modela potrebno je analizirati pokazatelje koji nam to omogućavaju. Kao pokazatelji reprezentativnosti analirani su varijansa regresije, standardna greška regresije, koeficijent varijacije i koeficijent determinacije.
I u zadnjem poglavlju obradit će se korelacija. Ova analiza bavi istraživanjem veza među posmatranim pojavama u smislu utvrđivanja stepena i smjera povezanosti. Na primjeru će se pokazati njena primjena u Excelu.

ANALIZA VEZA MEĐU POJAVAMA

Prilikom statističke analize neke vremenske serije utvrđuju se bitna svojstva pojave u prošlom i sadašnjem vremenu uz mogućnost predviđanja za budući period. Osim toga, statističkom analizom se utvrđuje odnos posmatrane serije podataka (skupa) prema nekoj drugoj seriji podataka.
Prema tome, ovom analizom se ne ispituju uzroci i posljedice pojava, već utvrđuje da li promjena svojstva jedinice neke serije podataka zavisi od promjene svojstva jedinice nekog drugog skupa. Dalje to znači, da se utvrđuje saglasnost u kretanju među pojavama pomoću saglasnosti u kretanju njihovih varijacija.
Postojanje veza između jedinica dva ili više skupa može se utvrditi grafički na osnovu paralelnosti kretanja promjena svojstva jedinica posmatranih skupova. Ako postoji paralelnost kretanja promjena svojstava jedinica posmatranih skupova, onda se može reći da postoji izvjesna zavisnost između ta dva skupa. Ako ne postoji paralelnost kretanja promjena svojstava jedinica posmatranih skupova, tada se može reći da ne postoji zavisnosti između ta dva skupa ili pojave.
Zavisnost između dva posmatrana skupa može biti funkcionalna (deterministička) i stohastička (Dacić, 2004).
Funkcionalna zavisnost postoji onda ako svakoj vrijednost jedne promjenjive (pojave) odgovara jedna vrijednost neke druge promjenjive (pojave). Kada se zna odnos promjenjivih među posmatranim pojavama, odnosno njihovim varijacijama, tada se na osnovu jedne pojave može predvidjeti druga, na primjer: Y = f(x). U ovom slučaju y je zavisno promjenjiva a x nezavisno promjenjiva. Primjera funkcionalnih veza u prirodnim naukama ima mnogo: utjecaj temperature na različite predmete, utjecaj ishrane na zdravlje čovjeka itd. Pored toga, postoji znatna međuzavisnost i u društvenim naukama: ekonomiji, sportu, sociologiji, psihologiji itd.
Pod stohastičkim vezama među pojavama podrazumijeva se slučaj kada jednoj vrijednosti nezavisne promjenljive y odgovara čitav niz mogućih vrijednosti zavisne promjenjive x. Zato se uglavnom kaže, da jednoj vrijednosti nezavisne promjenjive najvjerovatnije odgovara određena vrijednost zavisne promjenjive. U vezi s ostvarenjem što objektivnije analize za zavisnu promjenjivu y koriste se prosječne vrijednosti. Na primjer, na potrošnju specijalne vrste hrane, osim prosječne plate, utiče i veliki broj drugih faktora (cijene, starosna struktura, bavljenje sportom, socijalni faktori, itd).
Stohastički oblik linearnog odnosa između varijable x i y formalno se izražava na osnovu jednačine:

$$y_i = a + bx + \varepsilon_i$$

u kojoj je " ε i" tzv. slučajno odstupanje.

Vrijednosti varijabli x i y spadaju u red primjetnih a vrijednosti ε u red neprimjetnih varijabli. Ovakav oblik regresionog modela podrazumijeva da se vrijednost za y ne može tačno predvidjeti. Ova neizvjesnost proizilazi iz prisutnosti slučajnog odstupanja ε koje, pridaje slučajnost varijabli y.
Osnovne pretpostavke regresionog modela, zasnovane na distribuciji vjerovatnoće odstupanja i određivanja vrijednosti

varijable, su slijedeće:

1. ε i je normalno distribuirano,
2. sredina je jednaka nuli: E(ε)= σ^2,
3. odsutnost autokorelacije: Cov (ε i, ε j) = 0, (i$\neq$ j),
4. nestohastičnost varijable x.

Veze među pojavama u regresionom modelu mogu biti direktne i inverzne. O direktnim vezama govorimo u slučaju kada povećanju jedne pojave direktno odgovara povećanje druge pojave. Inverzne veze postoje ako porastu jedne promjenjive odgovara opadanje druge ili obrnuto. Na kraju, ako vrijednostima podataka koji rastu ili opadaju ne odgovara bilo kakav porast ili pad vrijednosti druge pojave, tada se kaže da nema međusobnih veza među njima.

U ispitivanju kvantitativnih veza varijacija dvije ili više pojava ili njihovih karakteristika koriste se regresiona i korelaciona analiza. Pojam se pominje još u radovima Engleskog naučnika Francis Galtona (1855), a pretpostavlja se da potiče od latinske riječi (regressio), koja znači uzvrat ili uzvraćanje.

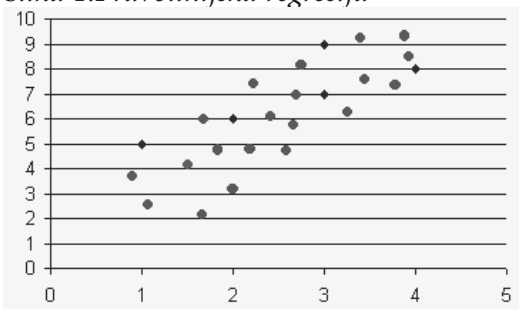
Regresiona analiza se bavi istraživanjem varijabiliteta i otkrivanjem funkcionalnog oblika, kojem se najviše približava kvantitativno slaganje varijacija posmatranih pojava. Ona treba da pokaže kako se zavisno promjenjive mijenjaju u odnosu na nezavisno promjenjive i da na osnovu stepena slaganja njihovih varijacija omogući ocjenu i predviđanje ponašanja zavisne promjenjive.

Za razliku od regresione analize korelaciona analiza treba da utvrdi mjeru slaganja između pojava ne ulazeći u kvantitativno određivanje te veze. To znači da se ova analiza bavi istraživanjem veza među posmatranim pojavama u smislu utvrđivanja stepena i smjera povezanosti.

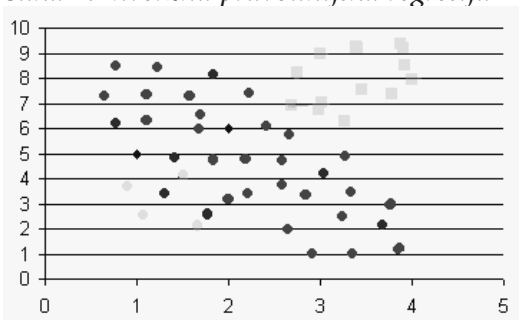
Zavisno od toga da li utvrđujemo povezanost između dvije pojave zanemarujući utjecaj ostalih pojava, kažemo da se radi o prostoj regresionoj analizi ili prostoj korelaciji. Analiza gdje se ispituje međusobni utjecaj više pojava naziva se višestrukom(multiplom) analizom regresije i korelacije.

Regresiona analiza je jedna od najčešće korištenih statističkih metoda sa velikom primjenom u ekonomiji, sociologiji, sportu, psihologiji, medicini i drugim oblastima nauke.

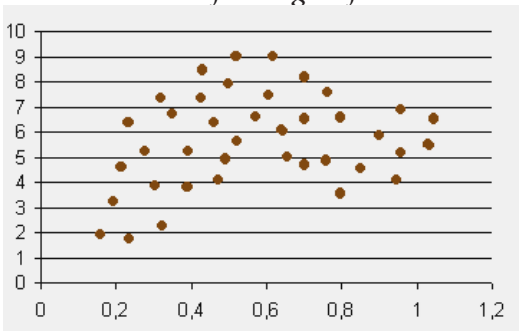
Slika 1.Pravolinijska regresija



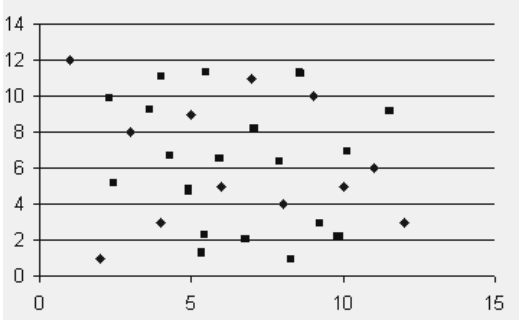
Slika 2. Inverzna pravolinijska regresija



Slika 3. Krivolinijska regresija.



Slika 4. Odsustvo regresije



Na grafikonima su prikazana četiri oblika mogućih regresionih veza: a) linearna regresija, b) inverzna linearna regresija, c) odsustvo regresije, d) krivolinijska regresija.
Kod konstrukcije pravouglog koordinatnog sistema, na apscisu se nanose jedinice pojave koju smo odredili kao nezavisnu promjenjivu, a na ordinatu jedinice zavisne promjenjive. Ovakav grafički prikaz zove se dijagramom raspršenosti ili oblakom raspršenosti. Pod raspršenošću se podrazumijeva udaljenost tačaka od linije regresije na grafiku. Uočavamo da postoje tačke koje su bliže ili dalje od kretanja linije regresije. Dijagram raspršenosti pruža polaznu informaciju o obliku zavisnosti između dvije varijable (Somun-Kapetanović, 2014).
Da bismo stupili modelizaciji veza između dvije ili više varijabli polazimo od sljedećih pretpostavki:

1. Modeliziranje možemo vršiti ukoliko postoji zavisnost između varijabli.
2. Mogu se modelizirati jedino kvantitativne varijable, jer je u tom slučaju moguće kompletirati oblak (dijagram) raspršenosti, računati mjere centralne tendencije i disperzije (Somun-Kapetanović, 2014).

Uz zavisnu varijablu možemo imati samo jednu nezavisnu varijablu (jednostavni regresijski model) ili veći broj nezavisnih varijabli (model višestruke regresije). U daljem tekstu opisat ću tri u praksi najčešće primjenjivana jednostavna regresijska modela, dok o višestrukoj regresiji neće biti riječi u ovom radu.
Dva su osnovna cilja koja želimo ostvariti pri konstrukciji regresijskog modela:

1. Pronaći funkciju koja najbolje opisuje vezu između posmatranih varijabli.
2. Parametre te funkcije ocijeniti tako da neobjašnejni dio varijanse zavisne varijable bude što manji (Papić, 2008).

LINEARNI MODEL REGRESIJE

Jednostavni linearni model se koristi u situaciji kada empirijski podaci u nekoj seriji pokazuju tendenciju linearnog povećanja ili smanjenja. Matematički postupak je u suštini isti kao i kod linearnog modela trenda. Jedina je razlika što zavisna i nezavisna promjenjiva mogu mijenjati mjesta u funkciji dok je kod trenda zavisna promjenljiva uvijek ista (vrijeme). To znači da se u funkciji linearnog modela regresije mogu javiti bilo koje dvije pojave za koje se ispituje moguća povezanost. Zato je i linearna funkcija ove regresije ista kao kod trenda, to jest:

$$Y = a + bx$$

Parametri a i b su konstate, a odnos između x i y predstavaljaju sve moguće vrijednosti koje zadovoljavaju jednačinu. Normalno je da oblik karakteristične jednačine daje oblik odgovarajućem odnosu: linearna jednačina opisuje linearni odnos, eksponencijalne jednačine opisuju eksponencijalni odnos i tako dalje. Sistem jednačina linearne funkcije je slijedećeg oblika:

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum\ x + b\sum x^2$$

Na osnovu ovog sistema jednačina, a prema već poznatom matematičom postupku, dobijaju se vrijednosti za y uvrštavanjem različitih vrijednosti nezavisne promjenjive x.
Sistem jednačina se također, može riješiti primjenom Kramerove metode za rješavanje sistema linearnih jednačina, na osnovu:

$$a = \frac{\Delta a}{\Delta}$$

$$b = \frac{\Delta b}{\Delta}$$

gdje je: Δ - oznaka za determinantu koju formiraju koeficijenti uz nepoznate veličine, to jest:

$$\Delta = \begin{vmatrix} N,,, \sum x \\ \sum x, \sum x^2 \end{vmatrix} = \mathrm{N}\sum x^2 - (\sum x)^2 \ .$$

Δa - oznaka determinante nepoznate veličine a.

$$\Delta a = \begin{vmatrix} \sum y \cdots \sum x \\ \sum xy \ldots \sum x^2 \end{vmatrix} = \sum y \sum x^2 - \sum x \sum xy$$

Δb - oznaka za determinantu nepoznate veličine b.

$$\Delta b = \begin{vmatrix} N \ldots \sum y \\ \sum x \sum xy \end{vmatrix} = N \sum xy - \sum x \sum y$$

Odakle je (Dacić, 2004):

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$$

$$b = \frac{N \sum xy - \sum x \sum y}{N \sum x^2 - (\sum x)^2}$$

ili

$$a = \bar{y} - b\bar{x}$$

Objašnjenje značenja parametara u regresijskoj jednačini su kako slijedi (Papić, 2008):

- $a \Rightarrow$ Konstantni član, tj. očekivana vrijednost zavisne varijable (y) kada je vrijednost nezavisne varijable (x) nula. (Napomena: vrijednost parametra a ponekad je " nelogična", tj ne odgovara realnoj situaciji.)
- $b \Rightarrow$ Regresijski koeficijent koji pokazuje prosječnu primjenu zavisne varijable (y) kada se nezavisna varijabla (x) poveća za jedan (jednu jedinicu mjerenja). Intenzitet tih promjena apsolutno je određen parametrom b, tako su mogući slučajevi:

U prvom slučaju b>0, znači da povećanjem pojave (x) povećava se i pojava (y), ako je b < 0, znači da će povećanjem pojave (x) doći do smanjenja pojave (y) i treće, b=0, značit će da promjena pojave (x) neće utjecati na promjenu pojave (y).

Geometrijski gledano on predstavlja koeficijent smjera pravca, pa će biti pozitivan ako pravac raste (varijable su upravno proporcionalne), a negativan ako pravac pa (varijable su obrnuto proporcionalne).

Pojednostavljenje modela se, kao kod linearnog trenda, postiže tako što se središnji vremenski period kod serije sa neparnim brojem podataka označi sa nulom (0), a periodi prije i poslije nje sa +1, +2, +3, +4, odnosno, -1, -2, -3,-4. Ili kod parne serije, središnji sa –0,5, i 0,5 a prethodni i naredni sa $\pm 1, \pm 2, \pm 3, \ldots, \pm n$. Ovaj način rada kao što znamo, zasniva se na relaciji:

$$\sum x = 0$$

i specifičnog oblika normalnih jednačina:

$$\sum y = na$$

$$\sum yx = b \sum x^2$$

iz kojih je:

$$a = \frac{\sum y}{n}$$

$$b = \frac{\sum xy}{\sum x^2}$$

Već je kazano da varijable mogu mijenjati mjesta u jednačini. Na primjer: utjecaj prihoda preduzeća na plaće radnika, i obrnuto, utjecaj promjena plaća radnika na prihod preduzeća.

Ako pođemo od jednačine:

$$y_i = a + bx + \varepsilon_i$$

i utvrdimo odstupanje ε_i:

$$\varepsilon = Y_i - a - bx$$

veličine ε_i u ovom izrazu nazivamo rezidualnim odstupanjima. Njima se iskazuju ocjene greške relacije posmatranog modela. Ovaj model linearne regresije se ocjenjuje, kako je pokazano, metodom najmanjih kvadrata.

Primjer 1: U sljedećoj tabeli prikazan je ukupni prihod (kolona *x*) i ukupni troškovi (kolona *y*) u turističkoj agenciji Fibula u Sarajevu za period od 5 mjeseci. [1] Podaci su dati u hiljadama (000 KM).

1 Podaci nisu preuzeti iz poslovne prakse.

Tabela 1. Podaci o ukupnim prihodima i troškovima

Ukupan prihod (000KM)	9	11	15	20	25
Troškovi (000 KM)	7	9	11	13	15

Treba izračunati: a) međuzavisnot između ukupnog prihoda i troškova u navedenom, b) Izvršiti analizu dobijenih rezultata, c) Grafički prikazati na dijagramu date pojave i regresionu funkciju.

Rješenje:

Na osnovu sistema normalnih jednačina formiramo tabelu iz koje je:

$$\bar{x} = \frac{80}{5} = 16$$

što predstavlja prosječno ostvaren prihod agencije.

Tabela 2. Radna tablica za primjer 1.

x	y	xy	x^2	y_c	y^2
9	7	63	81	7,66	49
11	9	99	121	8,62	81
15	11	165	225	10,52	121
20	13	260	400	12,91	169
25	15	375	625	15,29	225
80	55	962	1452	55,00	645

Na isti način je:

$$\bar{y} = \frac{55}{5} = 11$$

što predstavlja prosječno ostvarene troškove agencije.

Ocjena parametara sistemom normalnih jednačina bit će:

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{N\sum x^2 - (\sum x)^2} = \frac{55 \cdot 1452 - 80 \cdot 962}{5 \cdot 1452 - 80^2} = 3,372$$

$$b = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} = \frac{5 \cdot 962 - 80 \cdot 55}{5 \cdot 1452 - 80^2} = 0,4767$$

što znači da je jednačina regresionog modela:

$$y = 3,372 + 0,4767x$$

Na osnovu dobijenih regresijskih koeficijenata, može se zaključiti sljedeće:

– Koeficijent b=0,47 nam kazuje da u konkretnom slučaju porast ukupnog prihoda za jednu jedinicu utječe na porast ukupnih troškova za 0,47, odnosno 470 KM. Regresija je pozitivna što znači da povećanje jedne pojave utječe na povećanje druge u skladu sa regeresijskim parametrom b. Ocijenjena vrijednost a u praksi nema neku posebnu važnost.

Veličina i znak ocijenjenih vrijednosti često mogu biti u suprotnosti sa očekivanim rezultatima ili logikom naučnog zaključivanja. Takve vrijednosti se smatraju nezadovoljavajućim a rezultat su neadekvatne veličine uzorka (ispod 30 ispitanika). Pored toga, nezadovoljavajuće vrijednosti mogu nastati i zbog narušenih pretpostavki o kojima je bilo riječi kao i zbog nereprezentativnosti uzorka. U našem primjeru vrijednost parametra ima pozitivan predznak u jednom i drugom slučaju što je u skladu sa logikom zaključivanja.

Ako postoji suprotna zavisnost među pojavama i uslov x =f(y), što je kod mnogih pojava normalno, tada funkcija glasi:

$$x = a' + b'y$$

a ocijenjeni parametri

$$a' = \frac{\sum x \sum y^2 - \sum xy \sum y}{N\sum y^2 - (\sum y)^2} = \frac{80 \cdot 645 - 962 \cdot 55}{5 \cdot 645 - 55^2} = -6,55$$

$$b' = \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - (\sum y)^2} = \frac{5 \cdot 962 - 80 \cdot 55}{5 \cdot 645 - 55^2} = 2,05$$

iz čega je najmanja regresiona prava :

$$x = -6,55 + 2,05y$$

Do istog rezultata se dolazi ako se stavi da je $\sum y = 0$, uz koeficijente:

$$a' = \frac{\sum x}{N},$$

$$b' = \frac{\sum xy}{\sum y^2}.$$

uz napomenu da parametar a ima drugu vrijednost za drugi metod računanja.

Regresione prave: x=-6,55+2,05x i y=3,372+0,4767x se ne poklapaju. U općem obliku poklapat će se samo u slučaju kada je:

$$(\frac{a'}{b'} = a) \quad i \quad (\frac{1}{b'} = b).$$

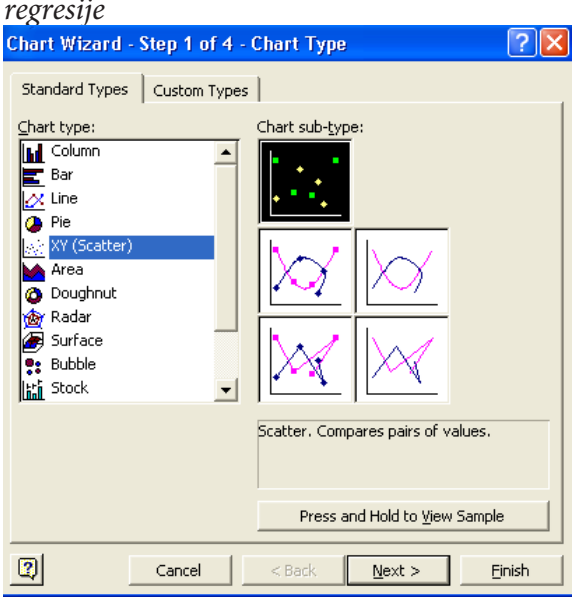
Ovakvi slučajevi podrazumijevaju potpunu podudarnost varijabli x i y što je kod primjene na empirijskim podacima pojava gotovo nemoguće. Ovo važi iz razloga što potpuna podudarnost varijabli znači punu funkcionalnu vezu među pojavama što praktično isključuje utjecaj drugih pojava na posmatrani odnos.

GRAFIČKI PRIKAZ REGRESIONE ANALIZE

Unosimo prvo podatke iz tablice. U slučaju regresijske analize moramo prvo tačno odrediti koja je varijabla nezavisna (x) , ako ja zavisna (y). U ovom primjeru nam i logika i tekst zadatka sugeriraju da je nezavisna varijabla "ukupan prihod", a zavisna "troškovi". Tim se redoslijedom podaci trebaju unositi u radni list MS Excela. Dakle, u stupcu A su redni brojevi, u B ukupni prihod i u C su troškovi.

Prvo se grafički prikazuju zadani podaci. To se preporučuje kao prvi korak regresijske analize, zbog toga što već na temelju dijagrama rasipanja možemo otprilike procijeniti kolika je povezanost između varijabli , kao i koji bi model regresije najbolje mogao odražavati vezu između njih. Dakle, označimo ćelije od B2 do C6, zatim kliknemo na ikonicu Chart Wizard i u dobijenom dijaloškom okviru odaberemo XY(Scater) pod Chart Type i gornja podopcija pod Chart sub-type:

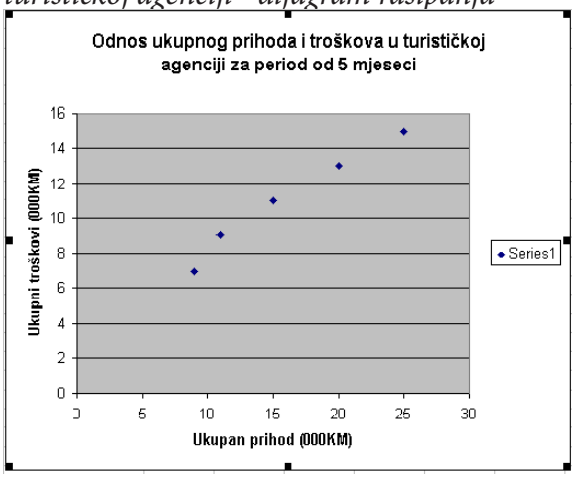
Slika 5. Prvi korak za određivanje modela regresije



Izvor: Autori

Traženi graf izgleda:

Slika 6. Odnos ukupnog prihoda i troškova u turističkoj agenciji - dijagram rasipanja



Izvor:Autori

Iz gornjeg dijagrama rasipanja vidljivo je:

- korelacija između posmatranih varijabli pozitivna je i visoka (ako se zamisli pravac koji prolazi kroz "oblak" tačaka na grafu očito je da taj pravac raste (r>0) i da dosta aproksimira tačke na grafu (r bliže jedinici nego nuli).
- ravnomjeran rast varijable x prati ravnomjeran rast varijable y, tj.opravdano je korisiti linearni model regresije,
- budući da onaj zamišljeni pravac raste, regresijski je koeficijent b pozitivan.

Postoje tri načina za dobivanje jednačine linearnog regresijskog modela, od kojih svaki ima svoje prednosti i nedostatke.[2]

1. način

Za izračunavanje parametara a i b u regresijskoj jednačini $y = a + bx$ u Ms Excelu postoje gotove funkcije:

- - za izračunavanje koeficijenta regresije b: =SLOPE(raspon varijable y; raspon varijable x)
- za izračunavanje konstantnog člana a: =INTERCEPT(raspon varijable y; raspon varijable x)

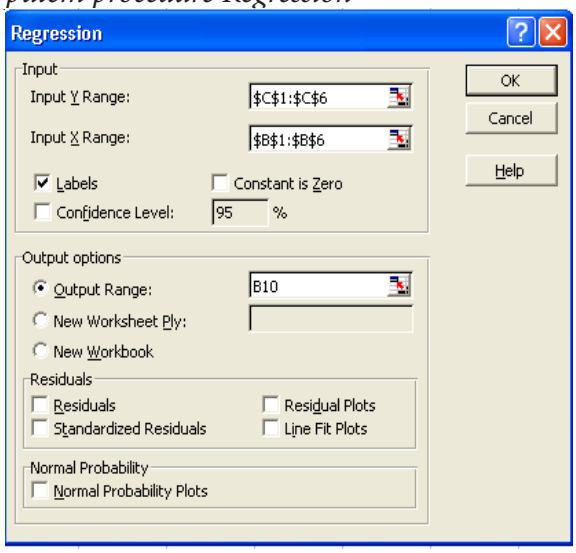
U primjeru 1. izračunate su vrijednosti b i a .
$b \Rightarrow$ =SLOPE(C2:C6;B2:B6). Cap. b=0,48
$a \Rightarrow$ =INTERCEPT(C2:C6;B2:B6).Cap a=3,37

Dakle, jednačina modela linearne regresije glasi: $y = 3,37 + 0,48x$.

2. način

Drugi način određivanja jednačine linearne regresije je korištenje procedure Regression, a pruža puno više informacija o parametrima jednačine i odnosu između posmatranih varijabli. Na padajućem izborniku Tools treba izabrati opciju Data Analysis i u dobivenom dijaloškom okviru označiti proceduru Regression.

Slika 7. Dobivanje jednačine linearne regresije putem procedure Regression



Izvor: Autori

Koristeći se gornjom slikom, upisuju se potrebni podaci u dijaloški okvir. Važno je u okvirić Labels staviti kvačicu jer su upisane i naslovne ćelije. Output-tablica koja se dobije izgleda ovako:

Tabela 3. Output-tablica procedure Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,988598
R Square	0,977326
Adjusted R Square	0,969767
Standard Error	0,549841
Observations	5

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	39,09302	39,09302	129,3077	0,001459
Residual	3	0,906977	0,302326		
Total	4	40			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3,372093	0,714449	4,719848	0,018014	1,098394	5,645792
Ukupni prihod (000KM)	0,476744	0,041925	11,37135	0,001459	0,34332	0,610168

2 Izvor: M.Papić, Primijenjena statistika u MS Excelu, 2.izdanje, ZORO d.o.o., Zagreb-Sarajevo, januar 2008.god., str. 146.

Za naš konkretni problem (određivanje jednačine linearne regresije) bitni su samo neki podaci iz dobivene tablice, i to Coefficients: Intercept (3,37) vrijednost je konstantnog člana a, a koeficijent uz varijablu "Ukupni prihod (000 KM)" (0,48) vrijednost regresijskog koeficijenta b. Dakle, jednačina linearnog modela regresije glasi:

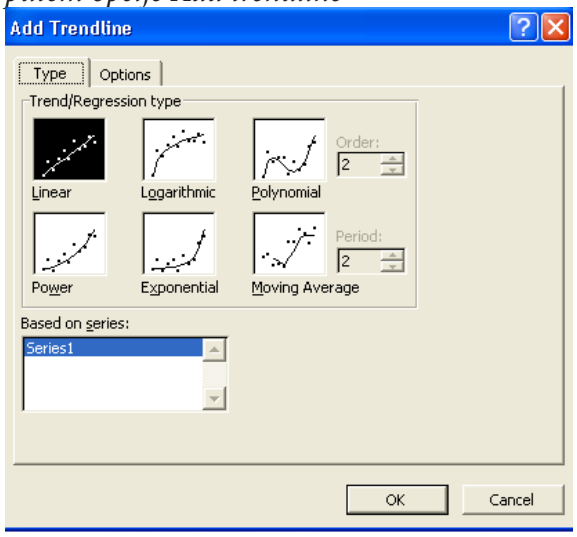
$$y = 3,372093 + 0,476744x$$

Na ovaj način je vidljivo da se dobije ista jednačina kao kod funkcija SLOPE i INTERCEPT.

3. način

Jednačinu modela linearne regresije možemo dobiti direktno iz dijagrama raspianja (XY Scattera) koji prikazuje odnos zavisne i nezavisne varijable. Kliknemo na pozadinu grafika (označimo ga). Na vrpci izbornika, umjesto izbornika Data, pojavit će se izbornik Chart. Kliknemo na Chart i izaberemo opciju Add Trendline:

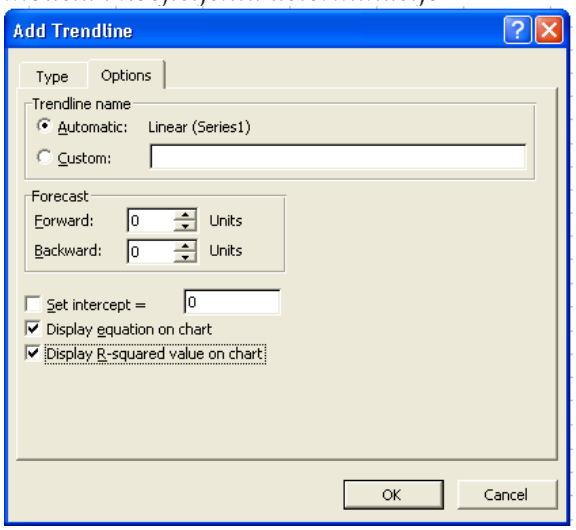
Slika 8. Dobivanje jednačine linearne regresije putem opcije Add trendline



Izvor: Autori

Na kartici Type dijaloškog okvira Add Trendline bira se odgovarajući regresijski model. U ovom primjeru treba kliknuti na kvadratić ispod kojeg piše Linear budući da je u pitanju linearni regresijski model. Nakon toga, potrebno je otvoriti drugu karticu-Options:

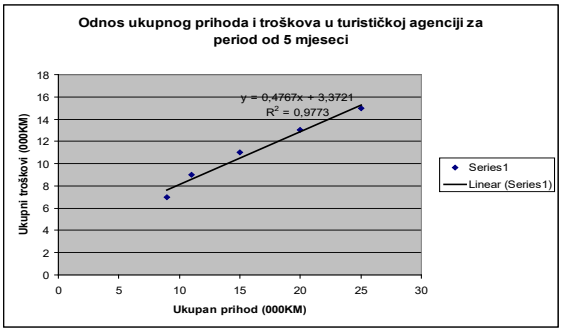
Slika 9. Način dobivanja jednačine regresijskog modela i koeficijenta determinacije



Izvor: Autori

Na kartici Options označavaju se dvije opcije: Display equation on chart, radi prikaza jednačine regresijskog modela, i Display R-squared value on chart, radi izračunavanja koeficijenta deterinacije o čemu će biti riječ kasnije.

Slika 10. Odnosu ukupnog prihoda i troškova u turističkoj agenciji - linearni regresijski model



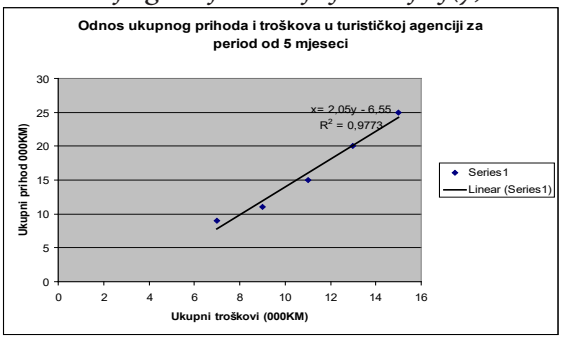
Izvor: Autori

Sa grafika čitamo jednačinu linearnog regresijskog modela: y=0,4767x+3,3721
Na osnovu koeficijenta deteminacije $R^2 = 0,9733$ zaključujemo da je 97,73% veze između ukupnog prihoda i ukupnih troškova u turističkoj agenciji objašnjeno linearnim

regresijskim modelom.

Na isti način se dobije funkcija $f(y)$ koju smo prethodno analitički dobili:

Slika 11. Odnos ukupnog prihoda i troškova u turističkoj agenciji kada je funkcija f(y)



Izvor: Autori

Na isti način sa grafika čitamo jednačinu linearnog regresijsko modela:

$$x = -6,55 + 2,05y.$$

Na grafikonima linearne regresije prikazane su jednačine f(x) i f(y), na osnovu kojih se donose zaključci o međuzavisnosti pojava.

MODELI KRIVOLINIJSKE REGRESIJE

Ravnomjerne promjene nezavisne varijable ne uzrokuju ravnomjerne promjene zavisne varijable). U tom slučaju potrebno je pronaći neku drugu funkciju (nelinearnu) koja najbolje pokazuje vezu između zavisne i nezavisne varijable. Jedan od nelinearnih modela koji se često koristi je model jednostavne eksponencijalne regresije.

MODEL JEDNOSTAVNE EKSPONENCIJALNE REGRESIJE

Standardni oblik regresijsje jednačine u ovom modelu je:

$Y = a \times b^x$, pri čemu vrijedi : a, b>0.

Značenje regresijskih parametara u gornjoj regresijskoj jednačini je kako slijedi:

– $a \Rightarrow$ očekivana vrijednost zavisne varijable (y) kada je vrijednost nezavisne varijable (x) jednaka nuli.

– $b \Rightarrow$ prosječna relativna promjena zavisne varijable kada se nezavisna varijala povećava za jednu jedinicu; kada se x poveća za 1 (jednu jedinicu), očekuje se povećanje (smanjenje) varijable y b puta (ako je b>1, s povećanjem varijable x povećava se i varijabla y; ako je b<1, povećanje varijable x uzrokuje smanjenje varijable y). Parametar b jasnije se interpretira preko prosječne stope promjene: $s = (b-1) \times 100$ koja pokazuje povećanje (smanjenje ako je s negativan) varijable y izraženo u postotcima ako se varijabla x poveća za jedan (jednu jedinicu). Npr, ako je $b = 1,25 \Rightarrow s = (1,25-1) \times 100 = 25\% \Rightarrow$ povećanjem varijable x za jedan varijabla y se povećava za 25%. Za $b = 0,7 \Rightarrow s = (0,7-1) \times 100 = -30\% \Rightarrow$ povećanje varijable x za jedan uzrokuje smanjenje varijable y za 30%.

U programu MS Excel jednačina eksponencijalnog modela regresije se dobiva posredno preko XY Scattera na isti način kao i jednačina linearnog modela regresije. To će biti ilustrovano na sljedećem primjeru.

Primjer 3:

Zadani su podaci o godišnjim primanjima i izdvajanju za ljetovanje i zimovanje za 15 slučajno odabranih porodica. Treba odrediti jednačinu eksponencijalnog modela regresije koji pokazuje zavisnost izdvajanja za ljetovanje i zimovanje po godišnjim primanjima, zatim objasniti značaj dobivenih parametara i procijeniti reprezentativnost regresijskog modela. Na temelju dobivene jednačine procijeniti koliko bi za ljetovanje I zimovanje izdvajala obitelj čija godišnja primanja iznose 55 000 KM.

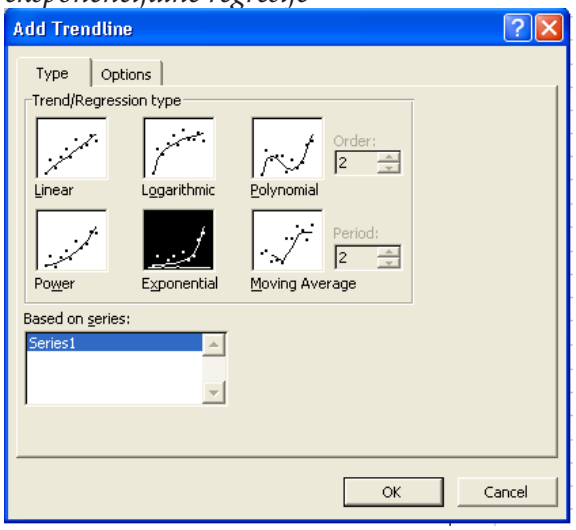
Tabela 4. Podaci o godišnjim primanjima i izdvajanju za ljetovanje i zimovanje 15 slučajno odabranih prodica

Redni broj porodice	Ukupna godišnja primanja (000KM)	Izdvajanja za ljetovanje i zimovanje (000KM)
1	7,50	0,50
2	11,25	0,75
3	12,50	1
4	15	1
5	17,5	1,75
6	18,75	2
7	20	2,25
8	22,50	2,75
9	25	3,50
10	27,50	4,50
11	30	5,75
12	35	7
13	37,50	8,75
14	45	11,25
15	50	16,25

Rješenje:

U stupac A unosimo redne brojeve porodice, u B ukupna godišnja primanja i u C podatke o izdvajanje za ljetovanje i zimovanje. Zatim se označavaju ćelije B2:C16 ⇒ Chart Wizard ⇒ XY(Scatter) ⇒ označiti gornju opciju među ponuđenim grafovima ⇒ Next ⇒ opcija Columns ⇒ Next ⇒ upisati odgovarajući naslov i oznake na koordinatnim osama (slika 14.) ⇒ Next ⇒ Finish. Zatim je potrebno kliknuti na pozadinu grafikona(označiti ga). Na vrpci izbornika, umjesto izbornika Data, pojavit će se izbornik Chart na koji je potrebno kliknuti i izabrati opciju Add trendline. (slika 13.)
Među ponuđenim regresijskim modelima označimo Exponential (slika 13), zatim se otvori kartica Options i označavaju opcije Display equation on chart i Display R-squared value on chart:

Slika 12.. Opcija trendline u slučaju eksponencijalne regresije



Dobivena jednačina eksponencijalnog regresijskog modela je:

$$y = 0,3785\ e^{0,0813x}$$

Parametar b se izračunava na sljedeći način: =EXP(0,0813) ⇒ 1,084696

Dakle, jednačina eksponencijalnog modela regresije glasi:

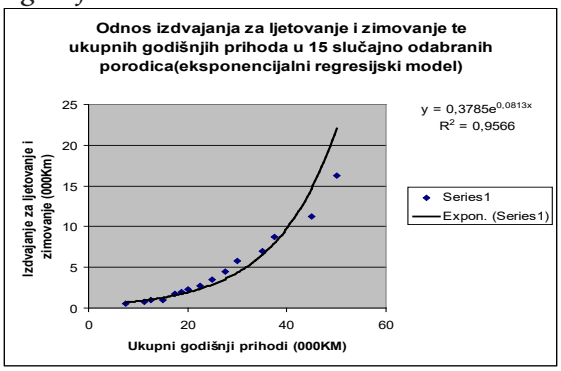
$$y = 0,3785 \times 1,084696^{x}$$

Konkretno značenje dobivenih parametara je sljedeće:

- a ⇒ 0,3785 ⇒ porodica čija su godišnja primanja jednaka nuli za ljetovanje i zimovanje trošila bi 378,5 KM (vrijednost parametra je a je ponekad malo čudna, samim tim što je čudna i sama pretpostavka da porodica nema nikavih primanja).
- b ⇒ 1,084696 ⇒ računa se prvo prosječna stopa promjene: s=(1,084696-1)× 100=8,46% ⇒ ako se ukupna godišnja primanja porodice povećaju za 1000 KM, očekuje se da će se iznos koji izdvajaju za ljetovanje i zimovanje povećati za 8,46%.

Traženi grafik izgleda kao na slici 14.

Slika 14. Odnos izdvajanja za ljetovanje i zimovanje te ukupnih godišnjih prihoda u 15 slučajno odabranih porodica – eksponencijalni regresijski model



Reprezentativnost dobivenog eksponencijalnog regresijskog modela procjenjuje se na temelju koeficijenta determinacije: $R^2 = 0,9566 \Rightarrow 95,66\%$ veze između izdvajanja za ljetovanje i zimovanje i ukupnih godišnjih primanja obitelji objašnjeno je eksponencijalnim regresijskim modelom.

U primjeru je potrebno utvrditi koliko bi za ljetovanje i zimovanje izdvajala porodica čija su ukupna godišnja primanja 55 000 KM. Jednostavno u dobivenu jednačinu umjesto x uvrstimo 55 (jedinica x je 1 000 KM):

$$y = 0,3785 * 1,084696^{55} \Rightarrow 33,1137$$

Dakle, na temelju eksponencijalnog regresijskog modela procjenjujemo da bi porodica čija su ukupna godišnja primanja 55 000 KM za ljetovanje i zomovanje izdvajala 33 113,7 KM. Očito je da je dobivena vrijednost prevelika. Iz grafika (slika 14.) vidljivo da kod većih vrijednosti varijable x eksponencijalna funkcija poprima veće vrijednosti (očekivane ili regresijske) od empirijskih. Prema tome, prognoza na temelju ovog modela nije baš pouzdana za veće vrijednosti varijable x. Radi veće reprezentativnosti ovih rezultata u nastavku će se prikazati drugi model na istom primjeru.

DVOSTRUKO LOGARITAMSKI MODEL REGRESIJE (POWER)

U ekonomskim istraživanjima se ovaj model regresije prilično često koristi. Standardni je oblik jednačine dvostruko logaritamskog regresijskog modela (poznatiji kao POWER):

$$Y = a \times X^b$$

Značenje parametara je sljedeće:

- $a \Rightarrow$ očekivana vrijednost zavisne varijable kada je vrijednost nezavisne varijable jedan (očekivana jedinica)
- $b \Rightarrow$ Očekivana promjena zavisne varijable (izražena u postotku) kada se nezavisna varijabla poveća za 1%.

Tim modelom omogućava se direktno određivanje elastičnosti zavisne varijable u odnosu na nezavisnu varijablu. Ako je $|b|$ >1, promatrana je funkcija elastična (zavisna varijabla «brže» se mijenja od nezavisne, npr. slučaj u ovom primjeru), a ako je $|b|$ <1, funkcija je neelastična (zavisna varijabla «sporije» se mijenja od nezavisne varijable, npr. izdaci za hranu rastu sporije odnosu na rast mjesečnih primanja).

Primjer 4:

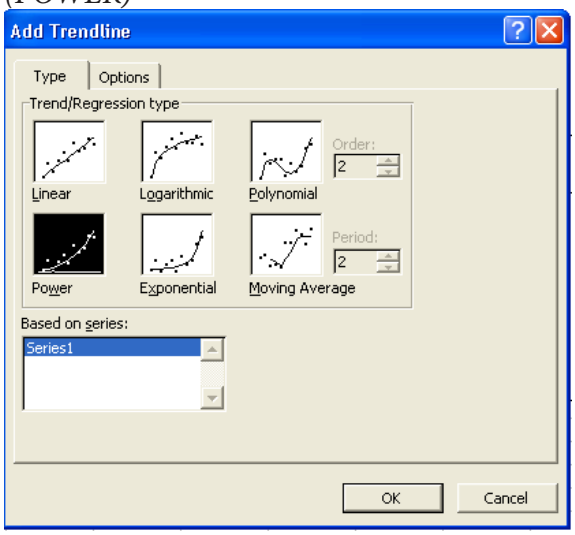
Za podatke iz primjera odrediti jednačinu dvostrukog logaritamskog modela regresije koji pokazuje zavisnost izdvajanja za ljetovanje i zimovanje o godišnjim primanjima. Objasniti značenje dobivenih parametara i procijeniti reprezentativnost regresijskog modela. Na temelju dobivene jednačine procijeniti koliko bi za ljetovanje i zimovanje izdvajala porodica čija ukupna godišnja primanja iznose 55 000 KM.

Rješenje:

Za rješavanje ovog primjera koristit će se već dobiveni grafički prikaz (slika 14.) iz prethodnog primjera. Najprije treba izbrisati sve podatke vezane uz eksponencijalni regresijski model (desni klik na krivulji

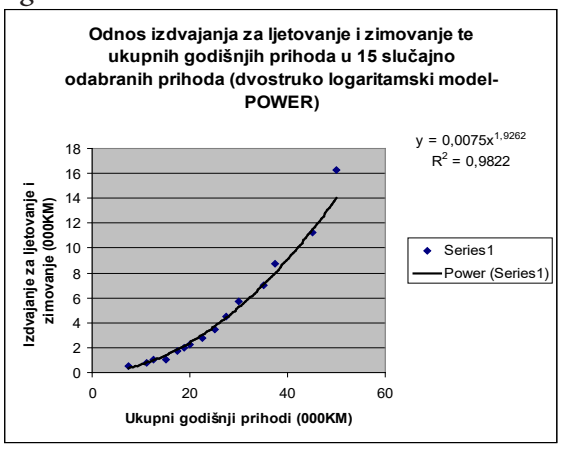
eksponencijalne funkcije, zatim pritisnuti Clear) a zatim označiti grafik i na izborniku Chart izabrati opciju Add trendline:
U dijaloškom okviru Add trendline označiti opciju Power (slika 15.), zatim otvoriti karticu Options i označiti Display equation on chart i Display R-squared value on chart.

Slika 13. Opcija Add trendline u slučaju dvostruko logaritamskg modela regresije (POWER)



Traženi grafik izgleda ovako:

Slika 14. Odnos izdvajanja za ljetovanje i zimovanje te ukupnih godišnjih prihoda u 15 slučajno odabranih porodica-dvostruko logaritamski model - POWER



Jednačina dvostruko logaritamskog modela regresije glasi:

$$y = 0,0075x^{1,9262}$$

Značenje parametara je sljedeće:

- $a \Rightarrow$ porodica čija su godišnja primanja 1 000 KM za ljetovanje i zimovanje trošila bi 7,50 KM.
- $b \Rightarrow$ ako se ukupna godišnja primanja porodice povećaju za 1% očekuje se povećanje izdvajanja za ljetovanje i zimovanje za 1,93%.

Na temelju parametra b može se zaključiti da je funkcija koja odražava zavisnost izdvajanja za ljetovanje i zimovanje o ukupnim godišnjim primanjima porodice elastična.
Reprezentativnost dobivenog regresijskog modela procjenjuje se na osnovu koeficijenta determinacije. $R^2 = 0,9822 \Rightarrow 98,22\%$ veze između izdvajanja za ljetovanje i zimovanje te ukupnih godišnjih primanja obitelji objašnjeno je dvostruko logaritamskim modelom.
S obzirom na to da treba odgovoriti na pitanje koliko bi za ljetovanje i zimovanje izdvajala porodica čija ukupna godišnja primanja iznose 55 000 KM, u dobivenu jednačinu umjesto x treba uvrstiti 55 (jedinica za X je 1 000 KM):

$$y = 0,0075 * 55^{1,9262} \Rightarrow 16,8790$$

Na temelju dvostrukog logaritamskog regresijskog modela procjenjuje se da bi porodica čija su ukupna godišnja primanja 55 000 KM za ljetovanje i zimovanje izdvajala 16 879 KM.
Uspoređujući prognozirane vrijednosti na temelju dva posmatrana nelinearna regresijska modela jasno je da je razlika velika. Logično je da se u ovom primjeru, birajući između ova dva modela, treba odlučiti za dvostruko logaritamski modela (R^2 za eksponencijalni je 0,9566, a za dvostruko logaritamski 0,9822).
Već iz samog dijaloškog okvira Add trendline jasno je da pored opisana tri regresijska modela program MS Excel omogućavan određivanje i nekih drugih regresijski modela kao pokazatelja veze između posmatranih varijabli. Tri opisana modela najčešśća su

u praktičnoj primjeni statističke analize, međutim ukoliko se njihovom primjenom ne dobiju dovoljno reprezentativni podaci može se primijeniti neki drugi ponuđeni model. Način dobivanja jednačine, mjere reprezentativnost i grafičkog prikaza odgovarajuće funkcije je isti kao i kod opisanih modela. Interpretacija parametara regresijske jednačine složenija je i zahtijeva malo bolje matematičko poznavanje dobivenih funkcija. Ali, ako regresijski model služi prije svega u svrhu procjene zavisne varijable za određene vrijednosti nezavisne, treba korisiti onaj model sa najvećim koeficijentom determinacije. U sljedećem primjeru prikazan je parabolični oblik krive regresijskog modela.

MJERE REPREZENTATIVNOSTI REGRESIJE

Nakon što se odredi jednačina modela jednostavne linearne regresije, potrebno je utvrditi valjanost tog modela. Postavlja se pitanje u kojoj mjeri vrijednost zavisne varijable prognozirane modelom odgovara stvarnim (empirijskim) vrijednostima, tj. koliko tačno se može predvidjeti vrijednost zavisne varijable za određenu vrijednost nezavisne varijable.

Procjena reprezentativnosti modela temelji se na analizi rezidualnih odstupanja (empirijskih, originalnih) vrijednosti (y_i) od očekivanih (regresijskih, teorijskih) vrijednosti ($\hat{y}_i$) zavisne varijable.

Da bismo odredili reprezentativnost i pouzdanost ocijenjenog modela potrebno je analizirati pokazatelje koji nam to omogućavaju. Kao pokazatelje reprezentativnosti analirani su varijansa regresije, standardna greška regresije, koeficijent varijacije i koeficijent determinacije (Somun-Kapetanović, 2014).

Komponente jednačine analize varijanse su:

- $\sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2 \Rightarrow$ ukupna suma kvadrata odstupanja (ST)
- $\sum_{i=1}^{N} \left(\hat{y}_i - \overline{Y}\right)^2 \Rightarrow$ protumačeni dio ukupne sume kvadrata odstupanja (SP)
- $\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2 \Rightarrow$ rezidualni (neprotumačeni) dio ukupne sume kvadrata odstupanja (SR)

VARIJANSA REGRESIJE

Varijansa regresije je apslutna mjera reprezentativnosti modela i predstavlja prosječno kvadratno odstupanje empirijskih od regresijskih vrijednosti.

$$\hat{\sigma}_{\hat{y}}^{\,2} = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2}{n}$$

ili

$$\hat{\sigma}_{\hat{y}}^{\,2} = \frac{SR}{n}$$

STANDARDNA GREŠKA REGRESIJE

Odstupanja empirijskih (originalnih) podataka uopće, mogu se posmatrati trojako: kao odstupanja empirijskih podataka od aritmetičke sredine, odstupanja empirijskih podataka od teorijskih podataka koji predstavljaju regresiju, i odstupanje aritmetičke sredine od serije koju čini regresija.

Varijabilitet funkcije y vezan za odstupanje originalnih podataka te funkcije prema seriji podataka iz regresije naziva se standardnom greškom regresije. Na osnovu toga se donosi sud o reprezentativnosti funkcije. Što je veća varijacija oko funkcije regresije to je reprezentativnost manja, tj. funkcija slabo predstavlja prosječan odnos između pojava i obrnuto. To znači, da se smanjenjem varijacija povećava funkcionalna veza između pojava, tako da se, u idealnom

slučaju ako su varijacije jednake nuli, radi o potpunoj vezi među pojavama.
Dakle, standardna devijacija ili standardna greška regresije je apsolutna mjera reprezentativnosti modela i predstavlja prosječno odstupanje empirijskih od regresijskih vrijednosti .

$$\hat{\sigma}_{\hat{y}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

ili

$$\hat{\sigma}_{\hat{y}} = \sqrt{\frac{SR}{n}}$$

KOEFICIJENT VARIJACIJE

Koeficijent varijacije regresije je relativna mjera reprezentativnosti modela I predstavlja postotni udio standardne greške regresije u odnosu na aritmetičku sredinu zavisne varijable:

$$\hat{V}_{\hat{y}} = \frac{\hat{\sigma}_{\hat{y}}}{\overline{Y}} \cdot 100$$

KOEFICIJENT DETERMINACIJE

Koeficijent determinacije definiše se kao omjer sume kvadrata odstupanja protumačene regresijom i sume kvadrata ukupnih odstupanja:

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

ili

$$R^2 = \frac{SP}{ST}$$

U skladu s prethodnim relacijama, R^2 se može izračunati i iz rezidualne sume kvadrata:

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

ili

$$R^2 = 1 - \frac{SR}{ST}$$

gdje se omjer sume kvadrata rezidualnih i ukupnih odstupanja naziva koeficijent alineacije ($1 - R^2$). U okviru regresijske analize koeficijent determinacije interpretira se kao proporcija veze između posmatranih varijabli objašnjena primijenjenim regresijskim modelom.
Koeficijent determinacije i koeficijent linearne korelacije povezani su relacijom: $R^2 = r^2$ tj. $r = \pm\sqrt{R^2}$
pri čemu je predznak koeficijenta korelacije jednak predznaku regresijskog koeficijenta b (vidljivo iz oblaka rasipanja).
Koeficijent determinacije može se objasniti kao proporcija varijanse zavisne varijable objašnjena nezavisnom varijablom. Npr, koeficijent korelacije između koeficijenta inteligencije i školskog uspjeha je oko 0,45 $\Rightarrow R^2$ =0,2025, što znači da je oko 20% školskog uspjeha objašnjeno inteligencijom učenika, odnosno studenata.
Koeficijent determinacije je u praktičnoj primjeni jednostavan za interpretaciju, a pomoću MS Excela se lako dolazi do njegove vrijednosti.
U cilju pojašnjenja koeficijenta determinacije koristit će se podaci iz primjera 1.
Ukoliko se koristi 1. način određivanja parametara jednačine linearnog regresijskog modela (funkcije SLOPE i INTERCEPT), tada je najprimjerenije izračunati koeficijent determinacije pomoću funkcije:
=RSQ(raspon varijable y; raspon varijable x)
U ovom primjeru koeficijent determinacije je : =RSQ(C2:C6;B2:B6) $\Rightarrow R^2$ = 0,977326
Dobivena vrijednost se može interpretirati na dva načina. Interpretirajući koeficijent determinacije kao mjeru povezanu sa Pearsonovim koeficijentom linearne korelacije, može se reći da je 97,73% varijanse ukupnih troškova u turističkoj agenciji objašnjeno je ostvarenim prihodom. S druge strane, interpretirajući koeficijent deteminacije kao mjeru reprezentativnosti

regresijskog modela. Kaže se da je 97,73% veze između posmatranih varijabli objašnjeno linearnim regresijskim modelom.
Procedura Regression (2. način) izračunava veličinu navedenih mjera reprezentativnosti modela. Output-tablica procedure Regression o kojo je bilo riječi u primjeru 1. izgleda ovako:

Tabela 5. Output-tabela procedure Regession:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,988598
R Square	0,977326
Adjusted R Square	0,969767
Standard Error	0,549841
Observations	5

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	39,09302	39,09302	129,3077	0,001459
Residual	3	0,906977	0,302326		
Total	4	40			

	Coefficients	Standard Error	t Stat	P-value	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	3,372093	0,714449	4,719848	0,018014	5,645792	1,098394	5,645792
Ukupni prihod (000KM)	0,476744	0,041925	11,37135	0,001459	0,610168	0,34332	0,610168

Unutar tablice ANOVA (analiza varijanse) izračunate su ove vrijednosti:

$$\sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2$$

- ukupna suma kvadrata odstupanja (ST)=40 (SS/Total)

$$\sum_{i=1}^{N} \left(\hat{y}_i - \overline{Y}\right)^2$$

- protumačeni dio ukupne sume kvadrata odstupanja (SP)=39,09302 (SS/Regression)

$$\sum_{i=1}^{N} \left(y_i - \hat{y}_i\right)^2$$

- rezidualni (neprotumačeni) dio ukupne sume kvadrata odstupanja(SR)=0,906977 (SS/Residual)
Standardna greška regresije iznosi 0,549841 (Standard error). Koeficijent determinacije je 0,977326 (R Square). Korigirani koeficijent determinacije (primjenjuje se u slučaju malog uzorka $\Rightarrow N \leq 30$) iznosi 0,969767 (Adjust R Square).
3. način dobivanja jednačine linearnog modela regresije (posredno preko XY Scattera) kao jednu od opcija nudi izračunavanje koeficijenta determinacije (Display R-squared value on chart), kao što se vidi na grafičkom prikazu.

KORELACIJA

Kod istraživanja međusobne povezanosti dvije ili više pojava, težište analize usmjereno je na iznalaženje i iskazivanje stepena i smjera njihove međusobne veze. Ovakav vid regresione analize naziva se korelacijom. Njome se ukazuje na postojanje veza između posmatranih pojava sa naznakom na uzročnu povezanost. Ovaj koeficijent ne zavisi od jedinica mjere. To je, dakle, neimenovan broj.
Imajući u vidu postojanje brojnih raznolikih veza među pojavama, razlikuju se i brojne vrste i oblici korelacije. Korelacija se može razlikovati u odnosu na oblik slaganja varijacija posmatranih pojava, na: linearnu i nelinearnu korelaciju. S obzirom na brojnost pojava koje istražuje, korelacija može biti: prosta i složena. Prosta pravolinijska

korelacija (r) je statistička metoda koja se koristi, ako se traži korelacija pravolinijskog oblika samo između dvije pojave (x) i (y), čije se promjenjive vrijednosti vežu u parove, uz zanemarivanje mogućnosti zavisnosti tih pojava od nekih drugih pojava. Pod složenom (multiplom) korelacijom podrazumijevat ćemo odnos povezanosti između više od dvije pojave.

Problem se svodi na kvantitativno obuhvatanje: stepena, jačine i smjera međusobnih odnosa i veza među pojavama. U statističkoj analizi taj kvantitativni oblik obuhvatanja veza naziva se koeficijentom korelacije.

KOEFICIJENT PRAVOLINIJSKE KORELACIJE

Koeficijent korelacije je relativna mjera stepena korelacije serije podataka posmatrane pojave. Poznat je i pod imenom svoga autora, kao Pearsonov koeficijent. (Dacić, 2004). Spada u red najjednostavnijih oblika ispoljavanja korelacionih odnosa između obilježja dvije pojave, računa se osnovu formule:

$$r = \frac{N\Sigma xy - \Sigma x\Sigma y}{\sqrt{N\sum x^2 - (\sum x)^2}\sqrt{N\sum y\ - (\sum y)^2}}$$

gdje su: x i y oznake za varijablu 1 i varijablu 2.

Vrijednost koeficijenta korelacije kreće se u granicama ± 1, odnosno: -1 < r<1. Ako je 1>r>0 u pitanju je povezanost kod koje su promjene obje promjenjive istog smjera, a ako je 0>r>-1, promjene pojava su suprotnog smjera. U praksi bi to bio slučaj da se sa povećanjem vrijednosti jedne pojave istovremeno smanjuju vrijednosti druge pojave. U slučaju da se r približava broju: $\pm$ 1, povezanost između pojava "jača", a ako se kreće prema nuli (0), povezanost između pojava "slabi". U ekstremnoj situaciji, kada koeficijent korelacije uzme vrijednost 1, linearna veza je potpuna i direktna. U drugom slučaju, kada koeficijent korelacije uzme vrijednost nula (0), ne postoji nikakva povezanost između pojavama. I treće, ako koeficijent korelacije uzme vrijednost minus (–1), korelacija je potpuna i inverzna.

Tabela 6. Jačina povezanosti između varijabli ovisno o apslutnoj vrijednosti koeficijenta korelacije

APSOLUTNA VRIJEDNOST KOEFICIJENTA KORELACIJE	JAČINA POVEZANOSTI IZMEĐU VARIJABLI
$\lvert r = 1 \rvert$	Potpuna korelacija
$0,8 \le \lvert r \rvert < 1$	Jaka korelacija
$0,5 \le \lvert r \rvert < 0,8$	Srednje jaka korelacija
$0,2 \le \lvert r \rvert < 0,5$	Relativno slaba korelacija
$0 < \lvert r \rvert < 0,2$	Neznatna korelacija
$\lvert\ r\ \rvert = 0$	Potpuna odsutnost korelacije

PRIMJENA KORELACIJE U EXCELU

U programu Ms Excel Pirsonov koeficijent linearne korelacije možemo izračunati na više načina. Radi lakšeg razumijevanja objašnjeni su na istom primjeru kao i linearna regresija, tj. na primjeru 1. Prisjetimo se podataka iz primjera 1:

Tabela 7. Podaci o ukupnim prihodima i troškovima

Ukupan prihod (000KM)	9	11	15	20	25
Troškovi (000 KM)	7	9	11	13	15

Treba odrediti stupanj korelacije između ove dvije varijable i objasnit njihovo značenje.

Rješenje:

Unosimo prvo zadane podatke. U stupac A unosimo redne brojeve, u stupac B ukupan prihod, u stupac C troškove. U ćelije A1, B1 i C1 upisujemo odgovarajuće naslove. Postoje 4 načina dobivanja Pearsonovog koeficijenta linearne korelacije pomoću programa MS

Excel:

1. način

U jednu praznu ćeliju upišemo =CORREL(B2:B6;C2:C6) i dobijemo r=0,9886

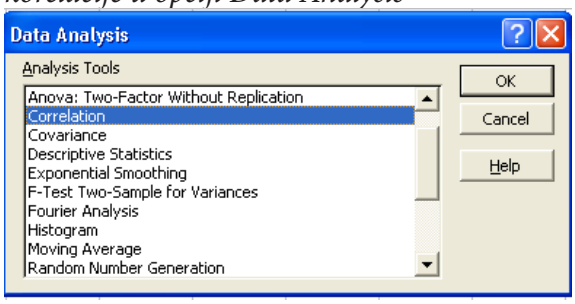
2. način

U neku drugu praznu ćeliju upišemo =PEARSON(B2:B6;C2:C6) i opet dobijemo isto, r=0,9886

3. način

Na padajućem izborniku Tools kliknemo na Data Analysis. U dijaloškom okviru izabremo proceduru Correlation:

Slika 15. Polazni dijalog box za računanje korelacije u opciji Data Analysis



U Input Range upisujemo raspon analiziranih podataka: B1:B6. Upisali smo i naslovne ćelije (jer želimo da nam u output-tablici pišu nazivi varijabli), pa je onda nužno staviti kvačicu na Labels in first row. Budući da su nam podaci za pojedinu varijablu uneseni u stupcima, označavamo opciju Columns u izborniku Grouped by. U drugom dijelu dijaloškog okvira Output options određujemo mjesto nove output tablice.

Slika 16.Tabela za unošenje potrenih podataka za obradu - korelacija

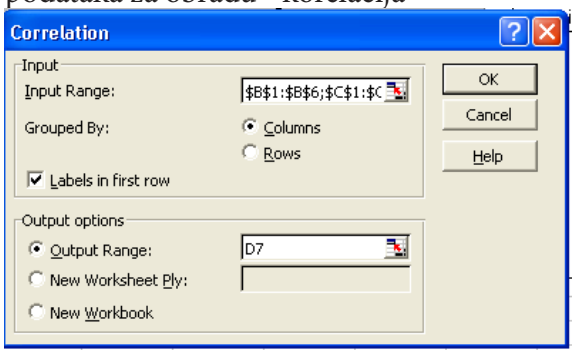


Tabela 8. Output tablica procedure Corerelation:

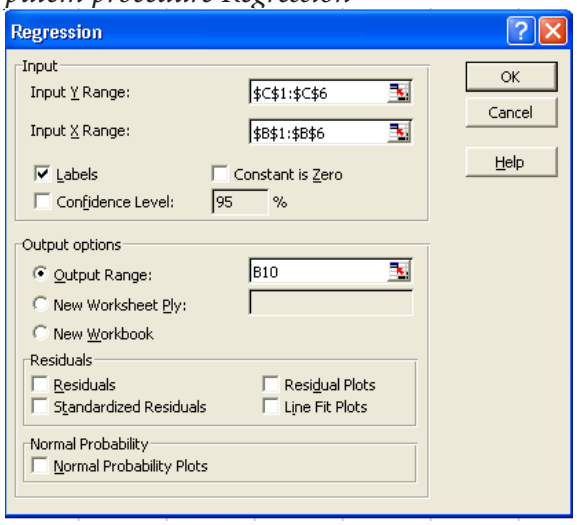
	Ukupan prihod	Troškovi
Ukupan prihod	1	
Troškovi	0,988598	1

U korelacijskoj matrici na dijagonali su uvijek jedinice (jer je korelacija svake varijable sa samom sobom potpuna), a u odgovarajućem polju matrice isčitavamo Pearsonov koeficijent linearne korelacije između posmatranih varijabli. Ovaj postupak je malo duži od prethodna dva, ali nam omogućava da u jednom koraku dobijemo međusobne korelacije između većeg broja varijabli (funkcije CORREL i PEARSON izračunavaju korelaciju isključivo između dvije varijable).

4. način

Oni koji imaju iskustva u primjeni korelacijske anlaizu uočit će da u trima opsianim načinima izračunavanja koeficijenta korelacije nešto nedostaje. Samo na temelju vrijednosti koeficijenta korelacije ne možemo zaključiti je li taj koeficijent statistički značajan, tj. da li je dobivena vrijednost slučajna ili ukazuje na stvarnu povezanost posmatranih varijabli. Zbog toga na ovaj načuin možemo izračunati značajnost Pearsonovog koeficijenta.
Na padajućem izborniku Tools kliknemo na Data Analysis i u dobivenom dijaloškom okviru izaberemo opciju Regression:

Slika 17. Dobivanje koeficijenta korelacije putem procedure Regression



O regresijskoj analaizi je bilo riječi, a sada ćemo iz ponuđenog u gornjem dijaloškom okviru izabrati samo ono što nam treba za rješavanje ovog konkretnog problema. Unosimo podatke u Input Y Range i Input X Range i stavljamo kvačicu na Labels. U Output options odredimo lokaciju output tablice. Dobijemo output tablicu, kao u primjeru 1:

Tabela 9. Output-tablica procedure Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,988598
R Square	0,977326
Adjusted R Square	0,969767
Standard Error	0,549841
Observations	5

ANOVA			
	df	*SS*	*MS*
Regression	1	39,093	39,093
Residual	3	0,906	0,302
Total	4	40	

ANOVA		
	F	*Sig. F*
Regression	129,307	0,001459
Residual		
Total		

	Coefficients	Standard Error
Intercept	3,372093	0,714449
Ukupni prihod (000KM)	0,476744	0,041925

	t Stat	P-value
Intercept	4,719848	0,018014
Ukupni prihod (000KM)	11,37135	0,001459

	Lower 95%	Upper 95%
Intercept	1,098394	5,645792
Ukupni prihod (000KM)	0,34332	0,610168

Iz gornje tablice čitamo potrebne vrijednosti. Vrijednost Pearsonovog koeficijenta linearne korelacije je 0,988598 (vrijednost koja je u tablici nazvana Multiple R). Značajnost koeficijenta korelacije možemo očitati na dva mjesta u gornjoj tablici: Significance F i P-Value za x varijablu. U oba slučaja je jednaka i iznosi 0,001459. Budući da je manja od 0,05 možemo zaključiti da dobiveni Pearsonov koeficijent linearne korelacije je značajan, tj da postoji statistički značajna povezanost između ukupnog prihoda i troškova u turističkoj agenciji.

ZAKLJUČAK

Primjena računara u svakodnevnom životu, ili bar u onom dijelu koji se odnosi na direktnu primjenu softwerskih programa u statistici, u velikoj mjeri utjecala je na sadržinu ovog rada. Činjenica je da se svakodnevno pojavljuju mnogobrojni računarski programi koji se intenzivno koriste u svim oblastima javnog i privatnog života. Oni su do te mjere usavršeni da je njihova primjena moguća i u slučajevima kada čitalac nije profesionalno obučen za rad sa statističkim metodama. Iz tog razloga rad je koncipiran tako da ukratko pruži uputstva za korištenje statističkih metoda u Microsoft Excelu. Upotreba ovog programa važna je i iz razloga što je Excel sadržan u bilo kojem paketu office

programa te je dostupan svakom korisniku novijih Windowsa. Korištenje ovih metoda podrazumijeva poznavanje statistike a djelimično i Excela-a. Statistički programi koji se koriste za multivarijacionu analizu pisani su u posebnim programima SPSS i drugi, koji se uglavnom namjenski prenose na eksternim memorijama. Ovim radom je pokazano na koji način se istražuju veze među pojavama, funkcionalne i stohastičke. U tu svrhu obrađeni su jednostavniji modeli regresione i korelacione analize. Kako bi analiza bila razumljivija i za čitaoce koji nedovoljno poznaju Excel obrađeni su teorijski modeli ove analize kako bi se na kraju svi primjeri ilustrovali tabelarnim i grafičkim prikazima u Excel-u. Naravno, postupno je objašnjeno na koji način se može doći do rješenja problema, matematički i primjenom statističkog programa u Excel-u. Da bi analiza bila uvjerljivija i bliža stvarnosti korišteni su podaci o izdvajanjima stanovništva za ljetovanje i zimovanje i njihovog ukupnog godišnjeg primanja. Primjeri su uglavnom pokazali da postoji međusobna veza između posmatranih pojava.

Obzirom da empirijski podaci, koji govore o kretanju neke pojave, ne uzrokuju uvijek ravnomjerne promjene zavisne varijable, to je u radu pokazano na koji način treba koristiti ostale oblike funkcija. Na primjeru je pokazano kako se određuje veza između varijabli u slučaju da se empirijski podaci kreću u obliku eksponencijalnog i dvostruko logaritamskog modela (POWER) regresije. Ovim primjerima je pokazana lakoća kojom se dolazi do konačnih rezultata što predstavlja značajnu uštedu u vremenu u odnosu na klasični način određivanja parametara.

Iz navedenog izlaganja se može zaključiti, da korištenje statističkih programa u MS Excel-u predstavlja veliki napredak u statističkoj obradi podataka uz značajnu uštedu u vremenu. Uz to upotreba statisičkih programa znatno olakšava edukaciju kadrova što omogućava njihovu dostupnost velikom broju korisnika.

LITERATURA

[1] Berenson, M.L., Levine, D.M., Krehbiel, T:C. (2004). Basic business statistics. 9/e. New Jersey: Pearson Education International.

[2] Dacić, R. (2004). Osnovi statistike, MADŽ D.O.O. - Sarajevo, Sarajevo,

[3] Levine, D.M. i ostali (2005). Statistics for Managers Using Microsoft Excel. New Jersey: Prentice Hall.

[4] Papić, M. (2008). Primijenjena statistika u MS Excelu, 2.izdanje, ZORO d.o.o., Zagreb-Sarajevo

[5] Resić, E. (2006). Zbirka zadataka iz Statistike. Sarajevo: Ekonomski fakultet.

[6] Resić, E., Delalić, A., Balavac, M., Abdić, A. (2010). Statistics in Economics and Management. Sarajevo: Ekonomski fakultet.

[7] Somun-Kapetanović, R. (2014). Statistika u ekonomiji i menadžmentu. Sarajevo: Ekonomski fakultet.

[8] Šošić, I. (2004). Primijenjena statistika. Zagreb: Školska knjiga.