Uporaba metode regresijske analize u rješavanju problema vezanih za inženjersku praksu

Lulić, Ivan

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

ZAVRŠNI RAD

Ivan Lulić

Zagreb, 2014.

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE

ZAVRŠNI RAD

Mentori:

Prof. dr. sc. Nedeljko Štefanić, dipl.ing.
Dr. sc. Hrvoje Cajner, dipl.ing.

Student:

Ivan Lulić

Zagreb, 2014.

Izjavljujem da sam ovaj rad izradio samostalno koristeći stečena znanja tijekom studija i navedenu literaturu.

Zahvaljujem se svome mentoru, prof.dr.sc. Nedeljku Štefaniću i asistentu dr.sc. Hrvoju Cajneru na stručnoj pomoći i savjetima tijekom izrade ovog rada.

Najviše se zahvaljujem Saneli na razumijevanju i pomoći te svojoj obitelji na podršci tijekom školovanja.

Ivan Lulić

SVEUČILIŠTE U ZAGREBU
FAKULTET STROJARSTVA I BRODOGRADNJE
Središnje povjerenstvo za završne i diplomske ispite
Povjerenstvo za završne ispite studija strojarstva za smjerove:
proizvodno inženjerstvo, računalno inženjerstvo, industrijsko inženjerstvo i menadžment, inženjerstvo materijala i mehatronika i robotika

Sveučilište u Zagrebu Fakultet strojarstva i brodogradnje	
Datum 18 -09- 2014	Prilog
Klasa: 602-04/14-6/2	
Ur.broj: 15-1703-14-346	

ZAVRŠNI ZADATAK

Student: **IVAN LULIĆ** Mat. br.: 0035183690

Naslov rada na hrvatskom jeziku: **UPORABA METODE REGRESIJSKE ANALIZE U RJEŠAVANJU PROBLEMA VEZANIH ZA INŽENJERSKU PRAKSU**

Naslov rada na engleskom jeziku: **THE USE OF REGRESSION ANALYSIS METHOD IN SOLVING PROBLEMS FROM ENGINEERING PRACTICE**

Opis zadatka:

Zadatak regresijske analize, kao neizostavnog dijela inženjerske statistike, jest definirati oblik veze među varijablama. Varijable predstavljaju kvantificirane uzroke i njihove posljedice (odzive) u nekom promatranom procesu. U proizvodnim procesima te procesima vezanim za strojarstvo i tehniku općenito, ponekad je teško isplanirati ispitivanje u obliku plana pokusa stoga se primjenjuje klasični postupak regresijske analize.

U radu je potrebno:

1. Detaljno objasniti pojam te teorijsku podlogu regresijske analize.
2. Napraviti presjek postojećih vrsta regresijske analize kao što su jednostavna, višestruka, linearna i nelinearna.
3. Izraditi algoritam u nekom od dostupnih programskih alata koji rješava probleme primjenom jedne od spomenutih vrsta regresijske analize.
4. Primijeniti metodu regresijske analize na proizvoljnim literaturnim primjerima te primjerima iz strojarske prakse uz uporabu razrađenog algoritma te gotovih programskih paketa.

Zadatak zadan: 17. travnja 2014.

Rok predaje rada: **2. rok:** 12. rujna 2014.

Predviđeni datumi obrane: **2. rok:** 22., 23. i 24. rujna 2014.

Zadatak zadao: Štefanić N.

Prof.dr.sc. Nedeljko Štefanić

Predsjednik Povjerenstva:

Prof. dr. sc. Zoran Kunica

Sadržaj

Popis slika

Popis tablica

Popis oznaka

Oznaka	Jedinica	Opis
Mo		Mod
Me		Medijan
$\bar{x}$		Aritmetička sredina
G		Geometrijska sredina
H		Harmonijska sredina
σ^2		Varijanca
σ		Standardna devijacija
z_i		Standardizirano obilježje
V		Koeficijent varijacije
e		Nezavisna slučajna varijabla
H_0		Nulta hipoteza
H_1		Alternativna hipoteza
F_0		Parametar F razdiobe
S_x^2		Srednje kvadratno odstupanje varijable x od $\bar{x}$
S_y^2		Srednje kvadratno odstupanje varijable y od $\bar{y}$
S_{xy}		Uzročna kovarijanca
ê		Rezidual
t		Oznaka Studentove t-razdiobe
SSE		Suma kvadrata reziduala
SST		Suma kvadrata odstupanja varijable x od $\bar{x}$
SSE		Suma kvadrata odstupanja empirijskih vrijednosti varijable y od $\hat{y}$
SSR		Suma kvadrata odstupanja regresijskih vrijednosti varijable y od $\bar{y}$
r^2		Koeficijent determinacije
$\bar{r}^2$		Korigirani koeficijent determinacije
E		Očekivana vrijednost
$\hat{\sigma}_y^2$		Nepristrana procjena varijance varijable y

Sažetak

U završnom radu predstavljena je regresijska analiza s naglaskom na jednostavnu i višestruku linearnu regresiju te logaritamsku i polinomnu nelinearnu regresiju.

Prvo poglavlje predstavlja pojam i definiciju statistike te opisuje osnovne pojmove statistike koji su važni za razumijevanje rada. Nakon toga predstavljen je regresijski model. Središnji dio rada sastoji se od dva poglavlja: jednostavne linearne regresije i višestruke linearne regresije. U poglavlju jednostavne linearne regresije opisana je metoda najmanjih kvadrata koja je važna za dobivanje procijenjenih regresijskih parametara. Nadalje, predstavljen je model jednostavne linearne regresije nakon kojeg slijedi statističko zaključivanje i verificiranje osnovnih statističkih podataka (standardna pogreška modela, tablica analize varijance, koeficijent determinacije, te F i T testovi koji su važni za testiranje nulte hipoteze). U poglavlju višestruke linearne regresije opisana je metoda najmanjih kvadrata za dobivanje procijenjenih regresijskih parametara te je predstavljen matrični pristup višestrukoj linearnoj regresiji. Kao i kod jednostavne linearne regresije poglavlje je zaključeno statističkim zaključivanjem i verificiranjem osnovnih statističkih podataka. Nakon linearne regresije predstavljena je nelinearna regresija, odnosno polinomna i eksponencijalna regresija. Na kraju završnog rada dani su primjeri jednostavne i višestruke linearne regresije te eksponencijalne regresije koji uključuju usporedbu ručnog načina računanja parametara regresije sa vlastitim algoritmom u programskom alatu Excel za olakšavanje računanja parametara regresije.

1. Uvod

1.1. Pojam i definicija statistike

Statistika je posebna znanstvena disciplina koja u svrhu realizacije postavljenih ciljeva istraživanja na organiziran način prikuplja, odabire, grupira, prezentira i vrši analizu informacija ili podataka, te interpretira rezultate provedene analize [1]. U raznim segmentima društva u ekonomiji statističke se metode i tehnike koriste na razini poduzeća i na makroekonomskoj razini. Statistika se može podijeliti na deskriptivnu i inferencijalnu statistiku.

Deskriptivna ili **opisna statistika** temelji se na potpunom obuhvatu statističkog skupa, čiju masu podataka organizirano prikuplja, odabire, grupira, prezentira i interpretira dobivene rezultate analize. Na taj način se, izračunavanjem različitih karakteristika statističkog skupa, sirova statistička građa svodi na lakše razumljivu i jednostavniju formu. Ako se statističke metode i tehnike primjenjuju na čitav statistički skup, dakle ako su istraživanjem obuhvaćeni svi elementi skupa oni tvore *populaciju*.

Inferencijalna statistika temelji se na dijelu (uzorku) jedinica izabranih iz cjelovitog statističkog skupa, pomoću kojeg se uz primjenu odgovarajućih statističkih metoda i tehnika donose zaključci o čitavom statističkom skupu. Uvijek je prisutan odgovarajući stupanj rizika kada se koriste rezultati iz uzorka, za kojeg je poželjno da bude izabran na slučajan način i da bude reprezentativan.

1.2. Predmet proučavanja statistike - statistički skup, statističke jedinice

Predmet proučavanja statistike su određene zakonitosti koje se javljaju u masovnim pojavama [1]. Zadaća statistike je da uoči zakonitosti u masovnim i slučajnim pojavama, te da ih iskaže brojčano.

Pri definiranju statističkog skupa potrebna je velika preciznost da bi se na temelju takve definicije moglo jednoznačno utvrditi da li neki element pripada ili ne pripada tom skupu. **Statistički skup** je skup elemenata kojima proučavamo jedno ili više obilježja čije se vrijednosti mijenjaju od elementa do elementa.

Statistički skup potrebno je **definirati pojmovno, prostorno i vremenski.**

- **Pojmovno** - odrediti pojam ili svojstvo svakog elementa promatranog skupa.
- **Prostorno** - odrediti prostor na koji se odnosi ili kojemu pripadaju elementi statističkog skupa.
- **Vremenski** - odrediti vremenski trenutak ili razdoblje kojim će se obuhvatiti svi elementi koji ulaze u statistički skup.

1.3. Osnovne faze statističkog istraživanja

Osnovne faze statističkog istraživanja su [1]:

- statističko promatranje (mjerenje, brojanje, opažanje, evidencija, anketiranje)
- grupiranje (tabelarno i grafičko prikazivanje statističkih podataka)
- statistička analiza i interpretacija rezultata provedene analize.

1.3.1. Sekundarni i primarni podaci

U ovisnosti o karakteru izvora podataka, statistički podaci se dijele na:

- sekundarne podatke
- primarne podatke.

Sekundarni podaci su oni koji se pribavljaju iz već postojećih baza podataka različitih državnih ustanova, npr. u Hrvatskoj su to: Državni zavod za statistiku, Hrvatska gospodarska komora i slične institucije. Sekundarni podaci su uglavnom brojčani. Predočeni su tablicama, a vrlo često i grafičkim prikazima.
Jedan od najčešće korištenih sekundarnih izvora podataka o gospodarskom i društvenom životu, te o demografskom razvitku i stanju okoliša u Republici Hrvatskoj je Statistički ljetopis u izdanju Državnog zavoda za statistiku.

Primarni podaci prikupljaju se neposrednim promatranjem svojstava elemenata statističkog skupa u skladu s unaprijed definiranim ciljevima statističkog istraživanja.

1.4. Uređivanje i prikazivanje podataka

Prikupljeni statistički podaci u svom izvornom obliku često nisu pregledni, pa ih je potrebno na odgovarajući način urediti [1]. Ako se urede prikupljeni statistički podaci prema nekom obilježju ili karakteristici, dobiva se statistički niz.

Grupiranje statističkih podataka je postupak diobe statističkog skupa na određeni broj pod skupova prema prethodno utvrđenim modalitetima promatranog obilježja i uz poštivanje načela isključivosti i iscrpnosti.

- **Načelo isključivosti** podrazumijeva da svaki element statističkog skupa istovremeno može pripadati samo jednoj grupi, tj. podskupu.
- **Načelo iscrpnosti** podrazumijeva da postupkom grupiranja trebaju biti obuhvaćeni svi elementi statističkog skupa.

Tabeliranje je postupak svrstavanja grupiranih prikupljenih statističkih podataka u tablice. Tablica nastaje crtanjem okomitih i vodoravnih linija prema određenim pravilima. Svaka statistička tablica mora imati:

- **Naslov tablice** mora biti jasan i kratak, a istovremeno mora u sebi sadržavati pojmovnu, prostornu i vremensku definiciju statističkog skupa.
- **Tekstualni dio** statističke tablice sastoji se od dva dijela: zaglavlja i pred stupca. U zaglavlju ili tumaču stupaca opisuje se i objašnjava sadržaj redaka. To su najčešće oblici statističkog obilježja po kojemu je promatran statistički niz.
- **Brojčani ili numerički dio** tablice sastoji se od polja u koja se unose frekvencije, odnosno rezultati grupiranja statističkih podataka. Zbirni ili marginalni stupac sadrži zbrojeve pojedinih redaka.
- **Izvor podataka** se navodi ispod tablice. On omogućuje provjeru ispravnosti prikupljenih podataka u tablici, kao i eventualnu dopunu podataka.

Statističke tablice mogu se podijeliti na:

- opće ili izvještajne statističke tablice
- analitičke ili sumarne statističke tablice.

Opće statističke tablice prikazuju ogroman broj statističkih podataka o nekom promatranom statističkom skupu.

Analitičke statističke tablice se konstruiraju za neku posebne analize i one su u pravilu preglednije.

Grafikonima se na jednostavan i pregledan način uz pomoć različitih geometrijskih likova prezentiraju osnovne karakteristike statističkih nizova. Grafički prikazi statističkih podataka su pregledniji i razumljiviji u odnosu na njihovo prikazivanje statističkom tablicom. Grafikoni omogućuju jednostavnije uočavanje glavnih karakteristika promatranih pojava, ali vrlo često ta preglednost ide na štetu preciznosti statističkih informacija. Stoga je poželjno uz grafički prikaz prezentirati i tablicu s originalnim vrijednostima statističkog niza. Oznake na grafikonu moraju biti takve da onaj tko čita sliku može jasno raspoznati koje su jedinice i koja je pojava prikazana.

Grafikon mora imati **naslov, jedinice mjere promatranog obilježja, oznake modaliteta obilježja, izvor podataka i po potrebi kazalo ili tumač oznaka.**

Postoje 3 skupine grafičkih prikaza:

1. **površinski grafikoni**
2. **linijski grafikoni i**
3. **kartogrami.**

Površinski grafikoni su: jednostavni stupci, dvostruki stupci, razdijeljeni stupci, strukturni krugovi ili polukrugovi, kvadrati.

Jednostavni stupci imaju jednake baze i jednako su udaljeni jedan od drugog i razlikuju se samo po visini koja odgovara veličini apsolutnih frekvencija.

Dvostruki ili razdijeljeni stupci upotrebljavaju se za grafičko prikazivanje dvaju ili više statističkih skupova koji su grupirani prema modalitetima istog obilježja.

Strukturni stupci također služe za prikazivanje više statističkih skupova podijeljenih na iste grupe prema jednom obilježju. Oni su jednake veličine, a razlikuju se po strukturi.

Proporcionalni strukturni krugovi i polukrugovi također prikazuju dva ili više skupova podijeljena na jednake grupe. Krugovi, odnosno polukrugovi, mogu biti iste veličine i razlikovati se samo po strukturi, a mogu biti i različitih polumjera, tj. proporcionalni veličinama promatranih statističkih skupova

Kartogrami su geografske karte na kojima se na različite načine pokazuje prostornu rasprostranjenost statističkog skupa. Razlikuju se tri vrste kartograma:

- **Dijagramske karte** - crtaju se spajanjem zemljovida i površinskih grafikona, na primjer, kvadrata, trokuta, krugova, i slično. Površinski grafikoni, odnosno likovi, moraju biti proporcionalni apsolutnim frekvencijama skupa, čime se izražava intenzitet promatranog obilježja ili pojave. Likovi se ucrtavaju unutar granica površine na zemljovidu koja predočava odgovarajući modalitet prostorno statističkog obilježja, čime se izražava prostorna rasprostranjenost elemenata statističkog skupa.

- **Piktogrami** - prostornu rasprostranjenost i intenzitet elemenata statističkog skupa prikazuju gušće ili rjeđe raspoređenim točkama (ili nekim drugim znakovima) na odgovarajućem zemljovidu..

- **Statističke karte** - crtaju se tako da se na zemljovidu različitim bojama ili sjenčanjem po pojedinim dijelovima nekog područja pokazuje intenzitet neke pojave, koji je najčešće izražen relativnim brojevima.

1.5. Srednje vrijednosti i mjere raspršenosti

1.5.1. Srednje vrijednosti - mod, medijan, aritmetička sredina, geometrijska i harmonijska sredina

Prikupljeni statistički podaci u svom izvornom obliku često nemaju razumljivu formu [1]. Zbog toga se vrši njihovo grupiranje, odnosno formiranje statističkih nizova. Na taj način se dobiva detaljniji uvid u svojstva promatranog statističkog niza.

Računanjem srednjih vrijednosti dolazi se do informacija o vrijednostima statističkog obilježja oko kojih se raspoređuju elementi statističkog niza.

Srednja vrijednost je vrijednost statističkog obilježja oko koje se grupiraju podaci statističkog niza. Drugi naziv: mjera centralne tendencije.

Srednje vrijednosti mogu se podijeliti na:

1. **Položajne srednje vrijednosti** – određuju se položajem podataka u nizu.

   Najvažnije položajne srednje vrijednosti jesu:

   a) mod
   b) medijan.

2. **Potpune srednje vrijednosti** – računaju se upotrebom svih podataka u statističkom nizu.

   Potpune srednje vrijednost jesu:

   a) aritmetička sredina,
   b) geometrijska sredina i
   c) harmonijska sredina.

Mod je vrijednost statističkog obilježja koja se najčešće javlja u nekom nizu, tj. vrijednost obilježja kojoj pripada najveća frekvencija.

Izraz za izračunavanje moda:

$$Mo = L_1 + \frac{(b-a)}{(b-a)+(b-c)}\, x\, i \tag{1.1}$$

gdje je:

*L*1– donja prava ili precizna granica modalnog razreda,
b – najveća korigirana frekvencija (tj. frekvencija modalnog razreda),
a – korigirana frekvencija ispred frekvencije modalnog razreda
c – korigirana frekvencija iza frekvencije modalnog razreda,
i – veličina modalnog razreda.

Medijan je vrijednost statističkog obilježja koja statistički niz dijeli na dva jednaka dijela.

Izraz za izračunavanje medijana:

$$Me = L_1 + \frac{\frac{N}{2} - \sum_{i=1}^{m} f_i}{f_{med}} \; x \; i \tag{1.2}$$

gdje je:

*L*1 – donja prava ili precizna granica medijalnog razreda,
N/2 – polovina elemenata statističkog niza,
$\sum_{i=1}^{m} f_i$– zbroj svih apsolutnih frekvencija do medijalnog razreda, ne uključujući medijalni razred, tj. kumulativna frekvencija ispred kumulativne frekvencije medijalnog razreda,
f_{med} – apsolutna frekvencija medijalnog razreda,
i – veličina medijalnog razreda.

Aritmetička sredina je omjer zbroja svih vrijednosti numeričkog obilježja jednog niza i broja elemenata tog niza.

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N} \tag{1.3}$$

Geometrijska sredina je *N*-ti korijen umnoška svih vrijednosti negrupiranog numeričkog obilježja jednog niza.

$$G = \sqrt[N]{\prod_{i=1}^{N} x_i} = \sqrt[N]{x_1 * x_2 * \ldots * x_N} \tag{1.4}$$

Harmonijska sredina je recipročna vrijednost aritmetičke sredine recipročnih vrijednosti numeričkog obilježja u jednom nizu.

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}} \tag{1.5}$$

1.5.2. Varijanca, standardna devijacija i koeficijent varijacije

Varijanca je prosječno kvadratno odstupanje vrijednosti numeričkog obilježja od aritmetičke sredine [1]. Varijanca spada u potpune mjere raspršenosti jer obuhvaća sve elemente odabranog numeričkog statističkog niza. Ovaj pokazatelj mjeri odstupanja, tj. raspršenost elemenata skupa od aritmetičke sredine.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N} = \frac{\sum_{i=1}^{N} {X_i}^2}{N} - \bar{X}^2 \tag{1.6}$$

Standardna devijacija je pozitivan korijen iz varijance i izražena je u originalnim jedinicama mjere. Stoga je kao potpuna i apsolutna mjera disperzije vrlo često u upotrebi. Može se definirati kao prosječno odstupanje vrijednosti numeričkog obilježja od aritmetičke sredine. Pomoću standardne devijacije u originalnim mjernim jedinicama obilježja može se uspoređivati raspršenost oko aritmetičke sredine nizova koji su grupirani po jednakom obilježju.

$$\sigma = +\sqrt{\sigma^2} = +\sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N}} = +\sqrt{\frac{\sum_{i=1}^{N} {X_i}^2}{N} - \bar{X}^2} \tag{1.7}$$

Standardizirano obilježje (z_i) je linearna transformacija originalnih vrijednosti numeričkog obilježja x_i, a pokazuje odstupanje vrijednosti obilježja od aritmetičke sredine u standardnim devijacijama.

$$z_i = \frac{X_i - \bar{X}}{\sigma}, \text{ i=1,2,3, …, n.} \tag{1.8}$$

Koeficijent varijacije je postotak standardne devijacije od aritmetičke sredine. Koeficijent varijacije spada u potpune relativne mjere raspršenosti jer obuhvaća sve elemente odabranog numeričkog statističkog niza, a izražava se u postocima (%).

$$V = \frac{\sigma}{\bar{X}} * 100. \tag{1.9}$$

2. Regresijska analiza

Regresijska analiza je statistički postupak za procjenu odnosa među varijablama. Cilj istraživanja odnosa među varijablama je utvrditi statističku ovisnost i pokazatelje jakosti takve ovisnosti [3]. Odnosi među pojavama mogu biti funkcionalni (deterministički) i statistički (stohastički):

- Funkcionalni ili deterministički odnosi su postojani, izražavaju zakonitosti koje se iskazuju analitički (formulom, jednadžbom). Svakoj vrijednosti jedne pojave odgovara točno određena vrijednost druge pojave.

$$Y = f(X) \tag{2.1}$$

- Statistički ili stohastički odnosi su slabiji od funkcionalnih. Jednoj vrijednosti jedne pojave odgovara više različitih vrijednosti druge pojave. Takva odstupanja su u praksi češća.

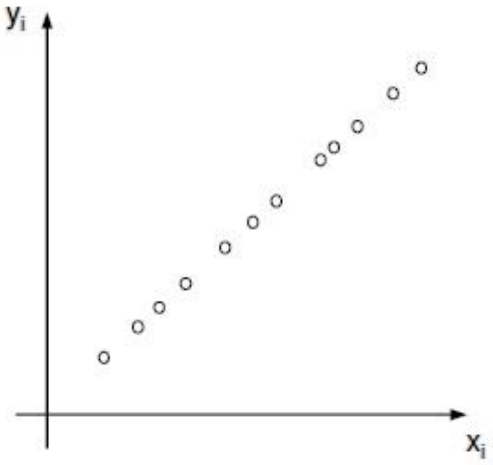
$$Y = f(X) + e \tag{2.2}$$

gdje je:

f(X) - funkcionalna (deterministička) komponenta
e - stohastička varijabla koja predočuje nesistematske utjecaje na zavisnu varijablu.

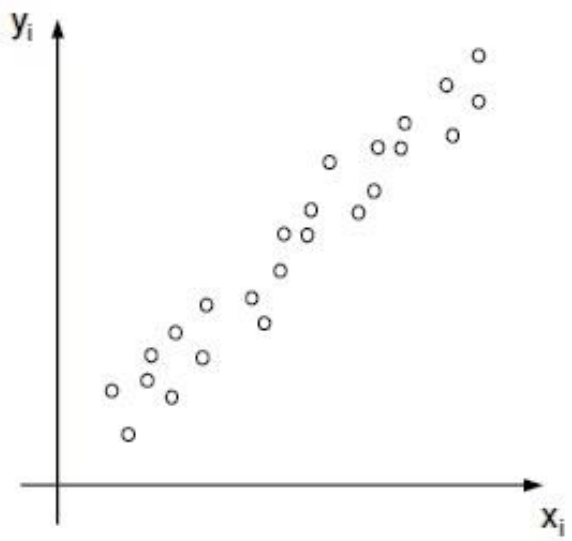
Dijagram rasipanja u pravokutnom koordinatnom sustavu točkama prikazuje parove vrijednosti dviju promatranih numeričkih varijabli. U sljedećim slikama prikazane su različite veze dviju promatranih numeričkih varijabli.

a) pozitivna funkcionalna veza



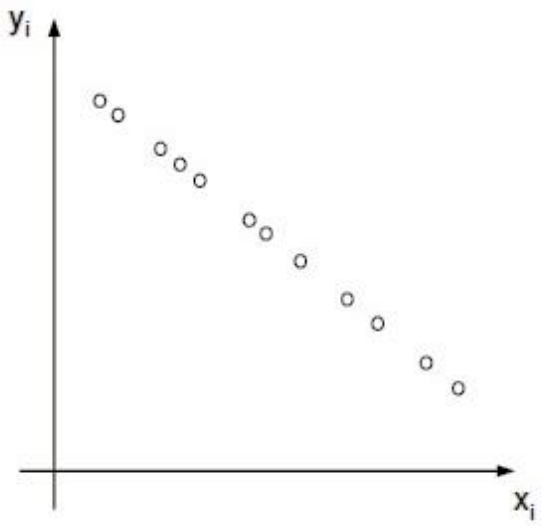
Slika 1: Pozitivna funkcionalna veza [3]

b) pozitivna statistička veza



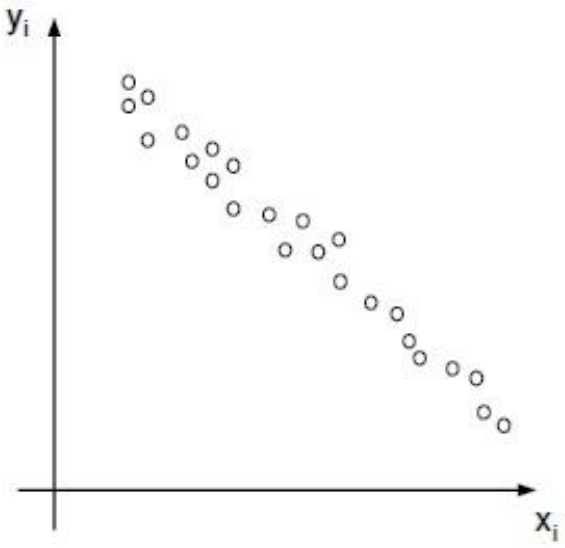
Slika 2: Pozitivna statistička veza [3]

c) negativna funkcionalna veza



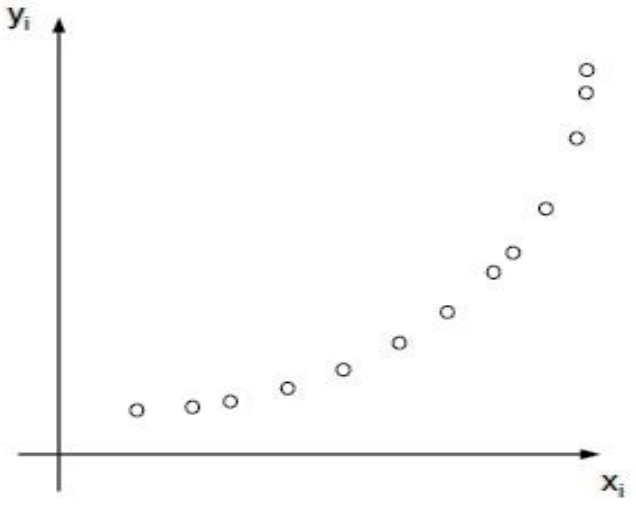
Slika 3: Negativna funkcionalna veza [3]

d) negativna statistička veza



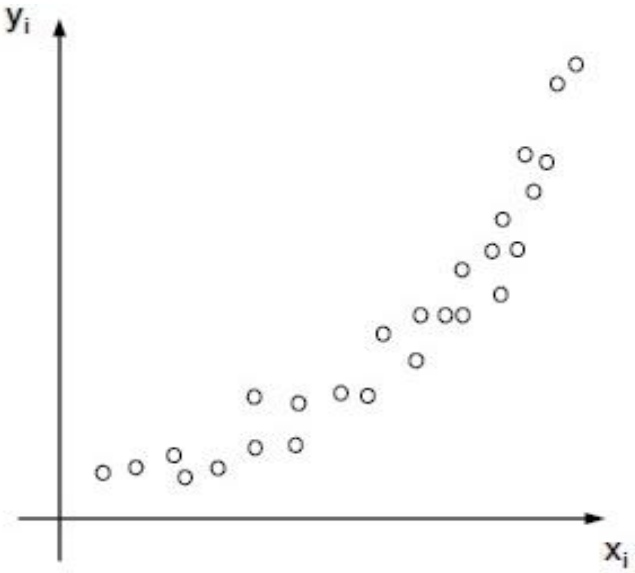
Slika 4: Negativna statistička veza [3]

e) pozitivna funkcionalna krivolinijska veza



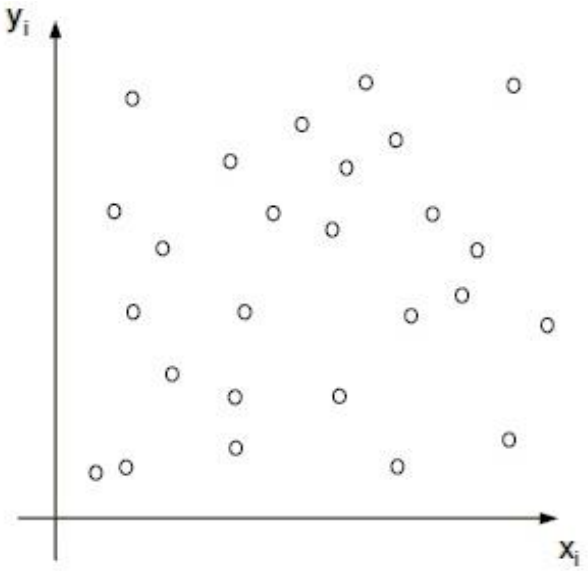
Slika 5: Pozitivna funkcionalna krivolinijska veza [3]

f) pozitivna statistička krivolinijska veza



Slika 6: Pozitivna statistička krivolinijska veza [3]

g) nema veze među pojavama



Slika 7: Nema veze među pojavama [3]

Regresijska analiza uključuje mnoge tehnike za modeliranje i analizu varijabli, gdje se fokus stavlja na odnosu između zavisne varijable i jedne ili više nezavisnih varijabli [8]. Konkretnije, regresijska analiza pomaže razumjeti kako se mijenja vrijednost zavisne varijable kada bilo koja nezavisna varijabla varira, dok su ostale nezavisne varijable fiksne. Najčešće, regresijska analiza procjenjuje uvjetno očekivanje zavisne varijable s obzirom na nezavisne varijable - to jest, prosječnu vrijednost zavisne varijable kada su nezavisne varijable fiksne. Ciljana procjena je funkcija nezavisnih varijabli odnosno regresijska funkcija. U regresijskoj analizi važno je karakterizirati varijacije zavisne varijable oko regresijske funkcije, a to se može opisati pomoću distribucije vjerojatnosti.

Regresijska analiza često se koristi za predviđanje i prognoziranje. Također koristi se za razumijevanje odnosa nezavisnih o zavisnim varijablama i istraživanje oblika tih odnosa. U određenim okolnostima, regresijska analiza se može koristiti za zaključivanje uzročnih odnosa između nezavisnih i zavisnih varijabli. Međutim to može dovesti do pogrešnih ili lažnih odnosa iz razloga što korelacija ne podrazumijeva uzročnost tako da je poželjan oprez.

Razvijene su mnoge tehnike regresijske analize kao što su jednostavna, višestruka, linearna i nelinearna. Najpoznatije metode su linearna regresija i metoda najmanjih kvadrata gdje se regresijska funkcija definira preko konačnog broja nepoznatih parametara koji se procjenjuju na temelju podataka.

3. Povijest regresije

Najraniji oblik regresije je metoda najmanjih kvadrata koju su objavili Legendre 1805. i Gauss 1809. godine [8]. Na temelju astronomskih promatranja Legendre i Gauss su primijenili metodu na problem utvrđivanje orbite nebeskih tijela oko Sunca (uglavnom kometa, a kasnije i tadašnjih novootkrivenih manjih planeta). Gauss je objavio daljnji razvoj metode najmanjih kvadrata u 1821. godini u Gauss-Markov teoremu.

Izraz "regresija" prvi put je uveo Francis Galton u devetnaestom stoljeću kako bi opisao biološki fenomen spuštanja visine potomaka visokih predaka prema normalnom prosjeku, odnosno rasta visine potomaka niskih predaka. Taj fenomen nazvao je regresija prema prosjeku. Njegov rad kasnije su proširili Udny Yule i Karl Pearson. U radu Yule i Pearson-a pretpostavlja se da je zajednička distribucija zavisnih i nezavisnih varijabli Gaussova. Pretpostavlja se da visina sinova ovisi o visini njihovih očeva te se može izraziti kao funkcija y = f(x), pri čemu je y zavisna ili kriterijska varijabla (visina sinova), a x nezavisna ili prediktorska varijabla (visina očeva). Tu pretpostavku nadopunio je R.A. Fisher u svojim djelima 1922 i 1925. godine u kojima pretpostavlja da zajednička distribucija nužno ne treba biti Gaussova.

1950-ih i 1960-ih godina ekonomisti su koristili elektromehaničke stolne kalkulatore za izračun regresije. Prije 1970. godine na rezultate jedne regresije čekalo se preko 24 sata.

4. Regresijski modeli

Neka je (X, Y) vektor slučajnih varijabli X i Y [9]. Zanima nas kako je distribuirana varijabla Y uz uvjet da varijabla X poprima vrijednost x (Pišemo Y I X = x). Funkcija f (x) = E (Y I X = x) zove se regresija. U slučaju kada imamo slučajni vektor (X, Y) normalno distribuiran N(α, β, σ, τ, φ), regresijska funkcija je oblika $E(Y \mid X = x) = \beta + \varphi \frac{\tau}{\sigma}(x - \alpha)$ što naravno predstavlja pravac $y = \alpha + \beta x$ koji nazivamo pravac regresije. Parametar β se zove regresijski koeficijent, a α je regresijska konstanta.

Regresijski model ima tri osnovna elementa: jednadžbe, hipoteze i uzročne pretpostavke.

1. Jednadžba regresijskog modela

Neka nezavisna varijabla X ima vrijednost $x_1, \ldots, x_n$, a zavisna varijabla Y ima vrijednosti $y_1, \ldots, y_n$. Promatramo vrijednost zavisne varijable Y.

Odnos između dvije varijable X i Y može se prikazati preko modela

$$y_i = f(x_i) + e_i \, (i = 1, \ldots, n) \tag{4.1}$$

gdje je f(x) regresijska funkcija, a $e_{1, \ldots}, e_n$ nezavisne slučajne varijable. Kada imamo samo jednu nezavisnu varijabla tada govorimo o jednodimenzionalnom ili jednostavnom regresijskom modelu, a kada imamo više od jedne nezavisne varijable tada se radi o višestrukom regresijskom modelu.

2. Hipoteza

Hipoteze u regresijskoj analizi su definirane na temelju regresijskih koeficijenata u regresijskoj funkciji.

3. Uzročne pretpostavke

Mjerenja provedena na manjem dijelu populacije nazivamo uzorak. S obzirom da je uzorak reprezentativan mora biti izabran nepristrano i mora biti dovoljno velik.}

5. Jednostavna linearna regresija

Regresijska analiza koristi se za donošenje zaključaka o nizu slučajnih varijabli $Y_1, \ldots, Y_n$ koje ovise o nezavisnoj varijabli x [2]. Zaključci se donose na temelju niza sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$, gdje su $x_1, \ldots, x_n$ vrijednosti nezavisne varijable x, a $y_1, \ldots, y_n$ odgovarajuće vrijednosti slučajnih varijabli $y_1, \ldots, y_n$.

U konkretnim primjerima nezavisnu varijablu x često zovemo **kontroliranom** ili **prediktornom varijablom.**

Cilj je na temelju sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$ ustanoviti ovisnost slučajnih varijabli y_i o nezavisnoj varijabli x.

Matematički model definirali smo relacijom

$$y_i = f(x_i) + e_i \quad (i = 1, \ldots, n) \tag{5.1}$$

gdje je:

$x \rightarrow f(x)$ realna funkcija jedne realne varijable,
$e_1, \ldots, e_n$ nezavisne slučajne varijable tako da je $E[e_i] = 0$ i $Var(e_i) = \sigma^2$.

Prvi korak u uspostavljanju veza između varijabli y i x je prikaz podataka u dijagramu raspršenosti iz kojeg se lako vidi grupiraju li se točke (sparena mjerenja) oko pravca (linearna zavisnost) ili neke krivulje.

Da bismo postavili što realniju pretpostavku o regresijskoj funkciji, parove podataka $(x_1, y_1), \ldots, (x_n, y_n)$ prikazujemo točkama u koordinatnom sustavu (dijagram raspršenosti ili scatterplot).

Ako pretpostavimo da je graf funkcije f(x) pravac, tj. da je zakonitost koja povezuje nezavisnu varijablu x i vrijednosti slučajnih varijabli y_i linearnog tipa, regresijsku funkciju u algebarskom obliku zapisujemo na sljedeći način:

$$f(x) = \alpha + \beta x \tag{5.2}$$

U tom se slučaju parametar β (koeficijent smjera) naziva **koeficijent regresije**, a pravac $y = \alpha + \beta x$ **regresijski pravac**.

5.1. Statistički model jednostavne linearne regresije

Pretpostavljamo da su vrijednosti zavisne varijable y_i povezane s vrijednostima nezavisne varijable na sljedeći način [2]:

$$y_i = \alpha + \beta x_i + e_i \quad (i = 1, \ldots, n) \tag{5.3}$$

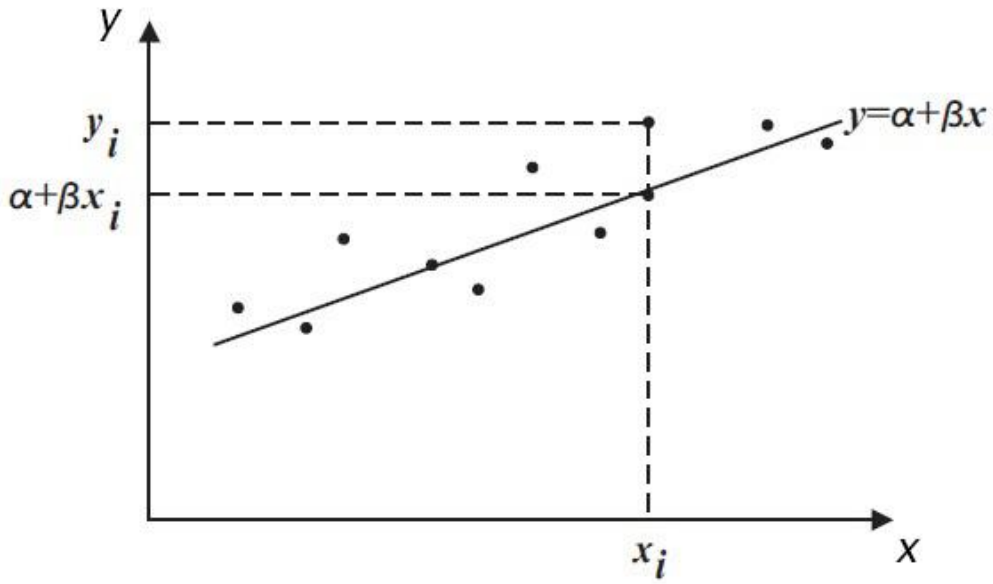
gdje je:

$x_1, x_2, \ldots, x_n$ - vrijednosti nezavisne (prediktorne) varijable x
$e_1, e_2, \ldots, e_n$ - nepoznate komponente greške koja je dodana na linearnu vezu. Ovo su slučajne varijable za koje pretpostavljamo da su međusobno nezavisne i da sve imaju normalnu distribuciju s očekivanjem 0 i istom varijancom σ^2,
α *i* β - nepoznati parametri pretpostavljene veze koje treba procijeniti

5.2. Metoda najmanjih kvadrata

Problem procjene nepoznatih parametara α i β identificira sa procjenom nepoznatog regresijskog pravca [2].

U sklopu dijagrama raspršenja nacrtan je proizvoljan pravac $y = \alpha + \beta x$. Iz slike je vidljivo da za vrijednost x_i nezavisne varijable x, zavisna varijabla y_i poprima vrijednost (predicted value) $\alpha + \beta x_i$.



Slika 8: Dijagram raspršenja

Nas zanima razlika $d_i = y_i - (\alpha + \beta x_i)$. Pretpostavljamo da su podaci $(x_1, y_1), \ldots, (x_n, y_n)$ zadani eksperimentalno. Tada regresijske parametre α i β procjenjujemo metodom najmanjih kvadrata.

Ideja metode je da se minimizira suma kvadratnih odstupanja teoretskih od eksperimentalnih vrijednosti, tj. procjene $\hat{\alpha}\ i\ \hat{\beta}$ regresijskih parametara α i β trebamo odrediti tako da vrijedi:

$$D = (\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2 = \min_{(\alpha,\beta)\in R^2} \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 = \min_{(\alpha,\beta)\in R^2} D(\alpha, \beta) \tag{5.4}$$

Procjene $\hat{\alpha}\ i\ \hat{\beta}$ regresijskih parametara α i β nazivamo procjenama u smislu metode najmanjih kvadrata ili least square estimates. Najbolja procjena nepoznatog regresijskog pravca y = α + βx je pravac $\hat{y} = \hat{\alpha} + \hat{\beta} x$.

Za analitičko rješenje procjena $\hat{\alpha}\ i\ \hat{\beta}$ regresijskih parametara α i β potrebno je:

Aritmetička sredina varijable x_i

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.5}$$

Aritmetička sredina varijable y_i

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{5.6}$$

Srednje kvadratno odstupanje varijable x od $\bar{x}$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5.7}$$

Srednje kvadratno odstupanje varijable y od $\bar{y}$

$$S_y^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.8}$$

Uzročna kovarijanca

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{5.9}$$

Dobre procjena $\hat{\alpha}\ i\ \hat{\beta}$ regresijskih parametara α i β u smislu metode najmanjih kvadrata su:

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} \tag{5.10}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{5.11}$$

Konačni izrazi za koeficijente regresijskog pravca oblika y = α + βx:

$$\beta = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} \tag{5.12}$$

$$\alpha = \bar{y} - \beta\bar{x} \tag{5.13}$$

Koristeći formulu regresijskog pravca, za svaku pojedinu eksperimentalnu vrijednost x_i možemo izračunati pripadnu teorijsku vrijednost y_i, pa možemo točno izračunati koliko iznosi svako odstupanje teorijske od eksperimentalne vrijednosti:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x) \tag{5.14}$$

Ovako dobiven niz vrijednosti $\hat{e}_1, \ldots, \hat{e}_n$ zovemo **rezidualima**. Suma kvadrata svih reziduala (sum of squares of errors = SSE) je minimalna postignuta vrijednost za $D(\alpha, \beta)$ i predstavlja mjeru kvalitete modela koju označavamo sa **SSE**:

$$SSE = \sum_{i=1}^{n}\left[y_i - \left(\hat{\alpha} + \hat{\beta}x_i\right)\right]^2 = \sum_{i=1}^{n}\hat{e}_i^2 \tag{5.15}$$

5.3. Statističko zaključivanje

Nakon ocijene parametara regresijskog modela postavlja se pitanje reprezentativnosti, odnosno sposobnosti modela da objasni kretanje ovisne varijable y uz pomoć odabrane neovisne varijable x [1]. U tu svrhu koriste se neki apsolutni i relativni pokazatelji. Ovi pokazatelji temelje se na raspodjeli odstupanja vrijednosti ovisne varijable y_i u regresijskom modelu od njene aritmetičke sredine $\bar{y}$ i njenih očekivanih vrijednosti $\hat{y}_i$.

5.3.1. Procjena varijance σ^2 (standardne greške regresije)

To je objektivna procjena pogreške regresijskog modela i označava se s $\hat{\sigma}^2$, a zapisujemo ju kako slijedi [1]:

$$\hat{\sigma}^2 = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2} = \sqrt{\frac{SSE}{n-2}} \tag{5.16}$$

Standardna greška regresije ili standardna devijacija regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti [1]. Prikazani izraz odnosi se na standardna grešku (devijaciju) regresije jednostrukog modela. Ovaj pokazatelj izražen je u originalnim jedinicama mjere ovisne varijable y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira relativni pokazatelj - **koeficijent varijacije regresije**, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$V = \frac{\hat{\sigma}^2}{\bar{y}} \times 100 \tag{5.17}$$

Najmanja vrijednost koeficijenta varijacije je 0%, a najveća nije definirana. Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji. Često se uzima dogovorena granica reprezentativnosti od 30%. Dakle, ako je koeficijent varijacije manji od 30%, kaže se da je model dobar.

5.3.2. Analiza varijance (ANOVA)

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ Suma kvadrata odstupanja varijable x od $\bar{x}$ [1] (5.18)

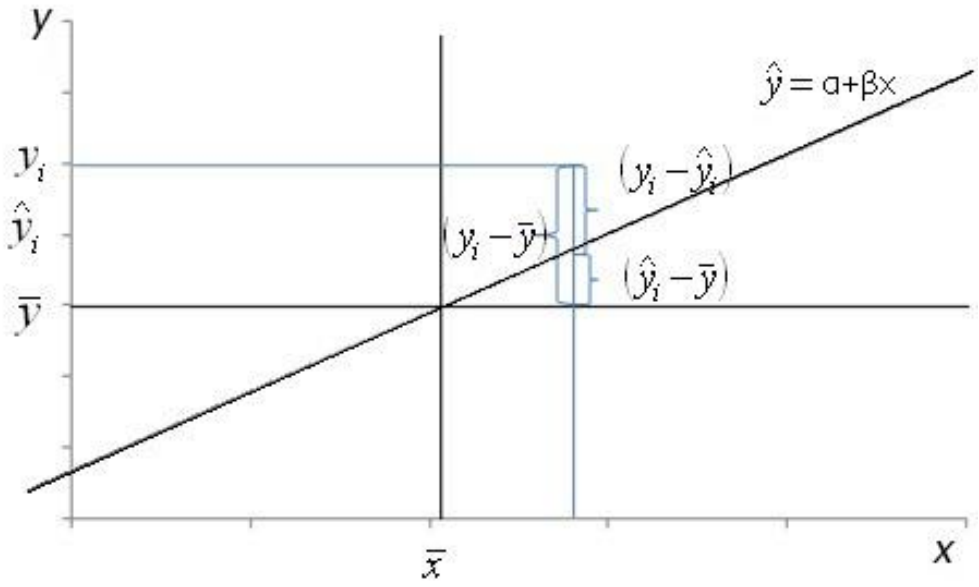
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ Suma kvadrata odstupanja regresijskih vrijednosti varijable y od $\bar{y}$ (odstupanja protumačena modelom) (5.19)

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Suma kvadrata odstupanja empirijskih vrijednosti varijable y od $\hat{y}$ (odstupanja ne protumačena modelom) (5.20)

$$SST = SSR + SSE$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5.21)$$

Ovaj izraz koji je u skraćenom obliku dan formulom SST = SSR + SSE zove se jednadžba analize varijance i predstavlja temelj analize reprezentativnosti regresijskog modela.



Slika 9: Dekompozicija sume kvadrata odstupanja [10]

Tablica 1: ANOVA - Analiza varijance jednostavne linearne regresije [10]

Izvor varijacije	***ss***	***Suma kvadrata***	***Sredina kvadrata***	***F-omjer***
Protumačen modelom	1	SSR	$SSR = s_P^2$	$\frac{s_P^2}{s_R^2}$
Ne protumačen modelom	n-2	SSE	$\frac{SSE}{n-2} = s_R^2$	
Ukupno	n-1	SST		

5.3.3. Test jakosti modela

Koeficijent determinacije r^2:

Koeficijent determinacije r^2 je pokazatelj reprezentativnosti regresijskog modela, koji se temelji na analizi varijance [1,2]. Daje nam informaciju o tome koliko rasipanja izlaznih podataka potječe od funkcijske ovisnosti $x \rightarrow \alpha + \beta x$, a koliko otpada na tzv. rezidualno ili neobjašnjeno rasipanje. Također nam daje informaciju o tome koliko je jaka funkcijska veza između x i y. Što je vrijednost koeficijenta r^2 bliža 1, zavisnost je jača. Koeficijent determinacije r^2 zapisujemo na sljedeći način:

$$r^2 = \frac{SSR}{SST} \tag{5.22}$$

odnosno:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.23}$$

Vrijednost koeficijenta determinacije kreće se u intervalu $0 \leq r^2 \leq 1$. Regresijski model reprezentativniji je ako je ovaj pokazatelj bliži 1. Teorijska granica reprezentativnosti modela je 0,9. U praksi je ponekad vrlo teško pronaći varijablu koja dobro objašnjava ovisnu pojavu pa se ta granica reprezentativnosti spušta i do 0,6.

Korigirani koeficijent determinacije $\bar{r}^2$:

$$\bar{r}^2 = 1 - \frac{n-1}{n-2}(1 - r^2) \tag{5.24}$$

- može biti manji ili jednak od koeficijenta determinacije
- ovisi i o broju vrijednosti za koje se računa
- nepovoljno obilježje je to što može biti manji od nule
- kada se govori o jačini veze ne smije se govoriti na nivou r, već treba uzeti u obzir i koeficijent determinacije

5.3.4. Testovi hipoteza jednostavne linearne regresije

Važan dio procjene adekvatnosti linearnog regresijskog modela je testiranje statističkih hipoteza o parametrima modela [1,2]. Za testiranje hipoteza moraju biti zadovoljeni uvjeti da su greške relacije e_i međusobno nezavisne, normalno distribuirane slučajne varijable s očekivanom vrijednosti nula i varijancom σ^2.

Osnova ovog testa hipoteza je utvrditi je li model $y_i = \alpha + \beta x_i + e_i$ bolji od nul-modela $y_i = \alpha + e_i$, tj. modela u kojemu je koeficijent regresije $\beta = 0$.
Ukoliko je $\beta = 0$ tada ne postoji linearna ovisnost između y i x. Taj problem svodi se na testiranje nulte hipoteze koju formuliramo na jedan od sljedeća dva načina:

H_0: Linearna veza između y i x ne postoji.
H_0: $\beta = 0$.

Alternativna hipoteza postavljena je na sljedeći način:

H_1: Linearna veza između y i x postoji.
H_1: $\beta \neq 0$.

Za testiranje ovih hipoteza koristi se T-test, pri čemu je vrijednost parametra „t“ dana izrazom:

$$t = \frac{S_x \hat{\beta}}{\hat{\sigma}^2}\sqrt{n-1} \sim T(n-2) \tag{5.25}$$

gdje je:

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.26}$$

$$\hat{\sigma}^2 = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}\hat{e}_i^2}{n-2}} \tag{5.27}$$

odnosno:

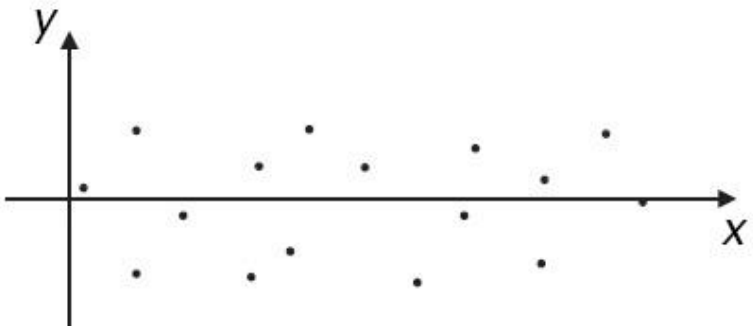
$$t = |r|\sqrt{\frac{n-2}{1-r^2}} \tag{5.28}$$

5.3.5. Analiza reziduala

Potrebno je utvrditi da li su ispunjene sve početne pretpostavke koje reziduali trebaju ispunjavati, a to su [2]:

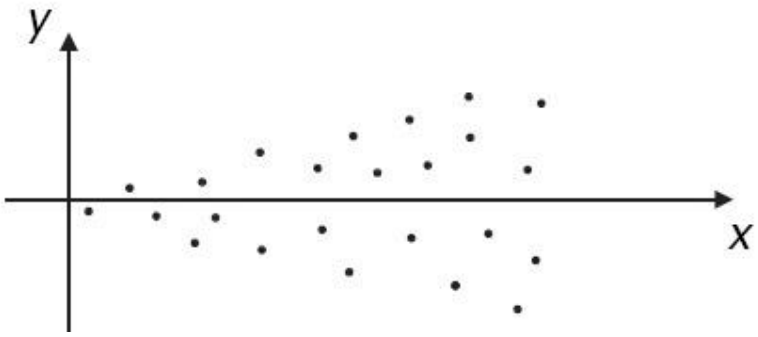
1. Varijance grešaka su jednake. Homogenost varijanci reziduala provjeravamo analizom grafičkog prikaza ovisnosti reziduala e_i o procijenjenim vrijednostima $\hat{y} = \hat{\alpha} + \hat{\beta}x$. Jednostavno donošenje zaključaka o varijanci dano je pomoću sljedećih slika:

a) Horizontalno raspoređene točke sugeriraju homogenost varijanci.



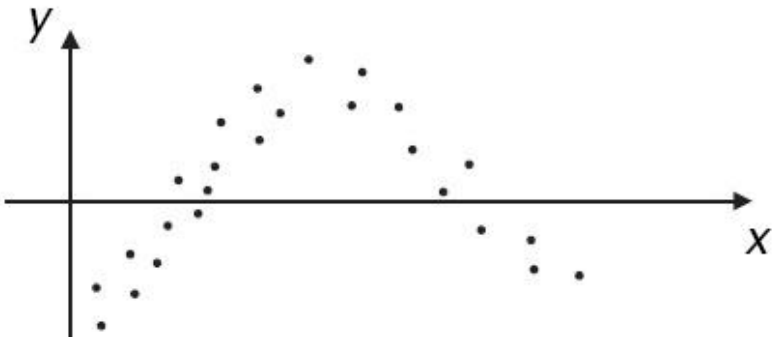
Slika 10: Horizontalno raspoređene točke koje sugeriraju na homogenost varijance [2]

b) Ovakav raspored točaka sugerira stalan rast varijance, dakle varijance nisu homogene.



Slika 11: Raspored točaka koji sugerira na nehomogenost varijance [2]

c) Ovakav raspored točaka sugerira neadekvatnost linearnog modela.



Slika 12: Raspored točaka koji sugerira na neadekvatan linearni model [2]

2. Reziduali su normalno distribuirani. Normalnost reziduala provjeravamo analizom histograma reziduala i p-plota reziduala.

3. Reziduali moraju biti međusobno nezavisni, tj. vrijednost reziduala koji se odnosi na realizaciju y_i slučajne varijable y nema nikakvog utjecaja na vrijednost reziduala koji se odnosi na realizaciju y_j iste slučajne varijable. Ovu pretpostavku provjeravamo analizom grafičkog prikaza reziduala za svaki pojedini slučaj i crtanjem dijagrama raspršenja uređenih parova reziduala $(\hat{e}_i, \hat{e}_{i-1})$, i = 2, . . . , n.

4. Ako reziduali $\hat{e}_i$ zadovoljavaju prethodno navedene pretpostavke smatramo ih dobrim procjenama stvarnih normalnih grešaka e.

.

6. Višestruka linearna regresija

Mnoge primjene regresijske analize uključuju situacije u kojima postoji više od jedne nezavisne ili regresorske varijable [6]. Regresijski model koji sadrži više od jedne regresorske varijable naziva model višestruke regresije.

Opći oblik modela višestruke regresije je:

$$y = f(x_1, x_2, x_3, \dots, x_i, \dots, x_K) + e \tag{6.1}$$

ili

$$y = f(x_1, x_2, x_3, \dots, x_i, \dots, x_K)\ x\ e \tag{6.2}$$

U navedenom modelu y je zavisna varijabla [5]. To je pojava čije se varijacije izražavaju pomoću nezavisnih (regresorskih) varijabli $x_1, x_2, \dots, x_K$. Varijabla e izražava nepoznata odstupanja od funkcionalnog odnosa. Prva jednadžba je aditivni, a druga multiplikativni oblik modela. Pretpostavi li se da je veza između y i $x_1, x_2, \dots, x_K$ linearna, model iz gornje jednadžbe je model višestruke linearne regresije:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_k x_k + e \tag{6.3}$$

U navedenom modelu y je zavisna ili regresand varijabla, $x_1, x_2, x_3, \dots, x_K$ su nezavisne, regresorske ili eksplanatorne varijable, $\beta_1, \beta_2, \beta_3, \dots, \beta_K$ su nepoznati parametri ili regresijski koeficijenti, a e je slučajna varijabla. Pretpostavi li se da se linearna regresijska veza između varijable y i odabranog skupa regresorskih varijabli utvrđuje na osnovi uzorka veličine n (n opažanja ili mjerenja odabranih varijabli), tada se vektorska jednadžba može napisati u vidu sustava od n jednadžbi:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_j x_{1j} + \cdots + \beta_k x_{1k} + e_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_j x_{2j} + \cdots + \beta_k x_{2k} + e_2 \\
&\vdots \\
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_k x_{ik} + e_i \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_j x_{nj} + \cdots + \beta_k x_{nk} + e_n
\end{aligned} \tag{6.4}$$

Polazne pretpostavke u analizi modela višestruke linearne regresije su jednake polaznim pretpostavkama za model jednostavne linearne regresije, proširenim za pretpostavku kojom se izražava priroda odnosa regresorskih varijabli. Te se pretpostavke mogu izraziti na slijedeći način:

- Veza između zavisne varijable i odabranog skupa nezavisnih varijabli je linearna.
- Greške relacije su međusobno nezavisne, identično normalno distribuirane slučajne varijable s očekivanom vrijednosti nula i varijancom σ^2:

$$E[e_i] = 0 \qquad Var(e_i) = \sigma^2 \; e_i \sim N(0, \sigma^2)$$
$$Cov(e_i, e_j) = E(e_i, e_j) = 0, \quad i \neq j \quad i,j = 1,\ldots, n$$

- Nadalje se pretpostavlja da su varijable x_i međusobno nezavisni vektori, te da je matrica X punoga ranga. Ta se pretpostavka uvodi iz numeričkih razloga.

6.1. Procjena parametara metodom najmanjih kvadrata

Uvrste li se u $Y = X\beta + e$ umjesto vektora parametara β vektor njihovih procjena, tada je [5]:

$$Y = X\hat{\beta} + \hat{e} \tag{6.5}$$

$$\hat{e} = Y - X\hat{\beta} \tag{6.6}$$

$\hat{e}$ je procjena slučajne varijable na bazi uzorka e i zove se rezidualna odstupanja. Procjena koeficijenta β metodom najmanjih kvadrata dobit će se iz zahtjeva da se minimizira suma kvadrata rezidualnih odstupanja:

$$L = \sum_{i=1}^{n} \hat{e}_i{}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij})^2 \rightarrow min \tag{6.7}$$

S obzirom da su u točki u kojoj funkcija dostiže minimum njene prve parcijalne derivacije jednake nuli, to se svodi na rješavanje sustava jednadžbi [6]:

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0 \tag{6.8}$$

i

$$\frac{\partial L}{\partial \hat{\beta}_j} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) x_{ij} = 0 \tag{6.9}$$

j = 1, 2, .. , k

Pojednostavljivanjem gornjeg izraza dobivamo normalne jednadžbe metode najmanjih kvadrata:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} {x_{i1}}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} x_{ik}y_i \tag{6.10}$$

Postoje p = k + 1 normalne jednadžbe, po jedna za svaki od nepoznatih regresijskih koeficijenata. Rješenja normalnih jednadžbi biti će **procjenitelji najmanjih kvadrata** regresijskih koeficijenata. Normalne jednadžbe mogu se riješiti rješavanjem sustava linearnih jednadžbi.

6.1. Matrični pristup kod višestruke linearne regresije

U postavljanju modela višestruke regresije praktičnije je izraziti matematičke operacije koristeći matrični zapis [6]. Kao što smo ranije pretpostavili da postoji n regresorskih varijabli i k promatranja $(x_{i1}, x_{i2}, x_{i3}, \dots, x_{iK}, y_i), i = 1,2, \dots, n$, model koji povezuje regresore zapisali smo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i \qquad i = 1,2, \dots, n \tag{6.11}$$

Ovaj model je sustav od n jednadžbi koje se mogu izraziti u matričnom zapisu kao

$$y = X\beta + e \tag{6.12}$$

gdje je:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nk} \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \\ \vdots \\ \beta_k \end{bmatrix}$$

U jednadžbi $y = X\beta + e$ y je vektor opaženih vrijednosti zavisne varijable, X je matrica čiji prvi stupac sadrži jedinice, a ostali stupci promatrane vrijednosti nezavisnih varijabli $x_1, x_2, \dots, x_K$, β je vektor nepoznatih parametara, a e je vektor slučajnih varijabli e_i, tj n-dimenzionalna slučajna varijabla. X matrica se također naziva matrica modela.

Želimo pronaći vektor procjene najmanjih kvadrata $\hat{\beta}$. Procjenitelj najmanjih kvadrata $\hat{\beta}$ dobiva se iz jednadžbe:

$$\frac{\partial L}{\partial \hat{\beta}} = 0 \tag{6.13}$$

gdje je:

$$L = \sum_{i=1}^{n} \hat{e}_i{}^2 = \hat{e}'\hat{e} = (y - X\hat{\beta})'(y - X\hat{\beta}) \tag{6.14}$$

Rješenje gornje derivacije daje nam normalne jednadžbe metode najmanjih kvadrata, a one iznose:

$$X'X\hat{\beta} = X'y \tag{6.15}$$

Ove jednadžbe predstavljaju normalne jednadžbe metode najmanjih kvadrata zapisane u matričnom obliku. One su identična ranije prikazanom skalarnom zapisu normalnih jednadžbi metode najmanjih kvadrata. Iz tih jednadžbi dijeljenjem sa X'X dobivamo procjenitelje najmanjih kvadrata varijable β:

$$\hat{\beta} = (X'X)^{-1}X'y \tag{6.16}$$

Ako je ispunjena polazna pretpostavka o nezavisnosti regresorskih varijabli matrica X'X je inverzibilna matrica.

Iz jednadžbi $X'X\hat{\beta} = X'y$ slijedi da je:

$$X'\left(y - X\hat{\beta}\right) = X'\hat{e} = 0 \tag{6.17}$$

Odnosno da je:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = hy \tag{6.18}$$

pri čemu je matrica $X(X'X)^{-1}X'$ (tzv. „hut“ matrica) projektor (odnosno simetrična i idempotentna matrica).
Kao što je već dokazano procjenitelj vektora parametara metodom najmanjih kvadrata je uz pretpostavku da su ispunjene sve polazne pretpostavke o modelu, najbolji linearni nepristrani procjenitelj.

6.2. Procjena varijance σ^2

Kao i kod jednostavne linearne regresije važna je procjena varijance $\hat{\sigma}^2$ [6]. Prisjetimo se kako se kod jednostavne linearne regresije procjena standardne greške regresije $\hat{\sigma}^2$ dobila dijeljenjem zbroja kvadrata reziduala sa n-2. S obzirom da kod jednostavne linearne regresije postoje dva parametra, kod višestruke linearne regresije javlja se p parametara, stoga je logična procjena varijance $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{SSE}{n - p} \tag{6.19}$$

Formula za zbroj kvadrata reziduala SSE glasi:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} {e_i}^2 = e'e \tag{6.20}$$

Uvrštavanjem $e = y - \hat{y} = y - X\hat{\beta}$ u gornju formulu dobivamo formulu za zbroj kvadrata reziduala SSE matričnog zapisa:

$$SSE = y'y - \hat{\beta}'X'y \tag{6.21}$$

6.3. Testovi hipoteza kod višestruke linearne regresije

Kod višestruke linearne regresije određeni testovi hipoteza korisni su za mjerenje adekvatnosti modela [6]. U ovom poglavlju opisat ćemo nekoliko važnih procedura za testiranje hipoteza. Kao i kod jednostavne linearne regresije, za testiranje hipoteza moraju biti zadovoljeni uvjeti da su greške relacije e_i međusobno nezavisne, identično normalno distribuirane slučajne varijable s očekivanom vrijednosti nula i varijancom σ^2.

6.3.1. Testiranje hipoteze o značajnosti regresije

Testiranje hipoteze o značajnosti regresije je test kako bi se utvrdilo da li postoji linearni odnos između zavisne varijable y i nezavisnih regresorskih varijabli $x_1, x_2, \dots, x_K$ [6]. Postupci testiranja hipoteza o značajnosti regresorskih varijabli u modelu višestruke linearne regresije mogu se svrstati u tri grupe [5]:

- Test o značajnosti jedne regresorske varijable (pojedinačni test)
- Test o značajnosti svih regresorskih varijabli (skupni test; test o značajnosti regresije)
- Test o značajnosti podskupa regresorskih varijabli

Test o značajnosti jedne regresorske varijable (pojedinačni test)

U regresijskoj analizi najčešće se provode jednosmjerni testovi o značajnosti pojedine varijable x_j, jer se na osnovi kvalitativne ekonomske analize zna smjer veze između varijabli y i x_j. Pri tom se tvrdnja istraživača (radna hipoteza) formulira kao alternativna hipoteza.

Hipoteze jednosmjernih testova o značajnosti regresijskog parametra:

$$H_0: \beta_j = 0 \qquad H_0: \beta_j = 0$$

$$H_1: \beta_j > 0 \qquad H_1: \beta_j < 0$$

Uz razinu signifikantnosti α (tj. uz zadanu vjerojatnost pogreške da se nul-hipoteza odbaci premda je istinita) hipoteza H_0 se odbacuje ako je $t_j > t_\alpha$ u testu na gornju granicu, odnosno ako je $t_j < -t_\alpha$ u testu na donju granicu. Ako testovi pokazuju da podaci nisu u skladu s odabranim modelom, model je neprihvatljiv i treba ga zamijeniti novim modelom.

Test o značajnosti svih regresorskih varijabli (skupni test; test o značajnosti regresije)

Nultom se hipotezom pretpostavlja da niti jedna od k regresorskih varijabli nema utjecaja na varijacije zavisne varijable. Alternativna hipoteza je po svom sadržaju suprotna nultoj hipotezi, pa glasi da je barem jedna od regresorskih varijabli značajna u modelu. Formalno, hipoteze se formuliraju:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \qquad j = 1,2,\dots,k$$

Odbacivanje $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ znači da barem jedan od regresorskih varijabli $x_1, x_2, \dots, x_K$ značajno doprinosi modelu [6].
Test o značajnosti regresije je generalizacija postupka koji se koristi kod jednostavne linearne regresije. Ukupnu suma kvadrata SST čini zbroj sume kvadrata odstupanja regresijskih vrijednosti varijable y od $\bar{y}$ i sume kvadrata odstupanja empirijskih vrijednosti varijable y od regresijskih vrijednosti odnosno:

$$\text{SST} = \text{SSR} + \text{SSE} \tag{6.22}$$

Ako je hipoteza $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ istinita, SSR/σ^2 je hi-kvadrat slučajna varijabla s k stupnjeva slobode. Broj stupnjeva slobode hi-kvadrat slučajne varijable jednak je broju regresorskih varijabli u modelu. Također SSE/σ^2 je hi-kvadrat slučajna varijabla s n – p stupnjeva slobode. Statistika ispitivanja za $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ glasi:

$$F_0 = \frac{SSR/k}{SSE/(n-p)} \tag{6.23}$$

Ukoliko ni jedna regresorska varijabla nije značajna u modelu procjene varijanci vrijednosti u brojniku i nazivniku trebale bi biti približno jednake. Hipotezu H_0 odbacujemo ako su izračunate vrijednosti F_0 veće od $f_{\alpha,k,n-p}$. Postupak se najčešće prikazuje preko tablice analize varijance.

Tablica 2: ANOVA - Analiza varijance višestruke linearne regresije

Izvor varijacije	***ss***	***Suma kvadrata***	***Sredina kvadrata***	***F-omjer***
Protumačen modelom	k	SSR	$\frac{SSR}{k}$	$\frac{SSR/k}{SSE/(n-p)}$
Ne protumačena odstupanja	n-p	SSE	$\frac{SSE}{n-p}$	
Ukupno	n-1	SST		

6.4. Koeficijent determinacije r^2 i korigirani koeficijent determinacije $\bar{r}^2$

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{(n-p)\,\hat{\sigma}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.24}$$

Taj pokazatelj poprima vrijednosti u intervalu $0 \leq r^2 \leq 1$, a promatrani model je reprezentativniji što je koeficijent determinacije bliži jedinici [5]. Međutim, kao što se vidi iz gornjeg izraza taj pokazatelj ima nedostatak da nije nepristran. Ako je populacijski koeficijent determinacije jednak nuli, očekivana vrijednost r^2 je različita od nule, jer je koeficijent determinacije veći što je veći broj regresorskih varijabli uključenih u model, bez obzira na to jesu li one značajne za objašnjavanje varijacija regresand varijable ili nisu. Nepristrana procjena varijance varijable y iznosi:

$$\hat{\sigma}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{6.25}$$

Ona uvrštena u izraz za koeficijent determinacije daje izraz za korigirani koeficijent determinacije $\bar{r}^2$:

$$\bar{r}^2 = 1 - \frac{n-1}{n-p}\,(1 - r^2) = 1 - \frac{n-1}{n-p}\,\frac{SSE}{SST} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2} \tag{6.26}$$

Korigirani koeficijent determinacije koristi se kao jedan od mogućih kriterija za izbor modela višestruke linearne regresije. Prema tom kriteriju najbolji je model s najvećim $\bar{r}^2$, što je ekvivalentno izboru modela s najmanjom procijenjenom varijancom.

7. Nelinearni regresijski modeli

7.1. Polinomni regresijski model

U situacijama u kojima funkcionalan odnos između zavisne varijable y i nezavisne varijable x ne može biti adekvatno aproksimiran linearnim odnosom, odnosno kada je odgovor (nezavisna varijabla) nelinearan, koristi se polinomna regresija [7].

Model polinomne regresije drugog stupnja sa jednom nezavisnom varijablom glasi:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + e \tag{7.1}$$

U navedenom modelu y je zavisna ili regresand varijabla, x je u nezavisna ili regresorska varijabla, $\beta_0, \beta_1, \dots, \beta_K$ su nepoznati parametri ili regresijski koeficijenti koje trebamo procijeniti, a e je slučajna varijabla. Procjene parametara $\beta_1, \beta_2, \dots, \beta_K$ dobit ćemo metodom najmanjih kvadrata tako da minimiziramo

$$L = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2 - \cdots - \hat{\beta}_k x_i^k)^2 \rightarrow min \tag{7.2}$$

gdje su $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ procjene parametara $\beta_0, \beta_1, \dots, \beta_k$. S obzirom da su u točki u kojoj funkcija dostiže minimum njene prve parcijalne derivacije jednake nuli, to se svodi na rješavanje jednadžbe:

$$\frac{\partial L}{\partial \hat{\beta}} = 0 \tag{7.3}$$

Pojednostavljivanjem gornjeg izraza dobivamo normalne jednadžbe metode najmanjih kvadrata:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i{}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_i^3 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} {x_i}^2 + \hat{\beta}_1 \sum_{i=1}^{n} x_i^3 + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+2} = \sum_{i=1}^{n} x_i^2 y_i$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} {x_i}^k + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} x_i^k y_i \tag{7.4}$$

Postoje k + 1 normalne jednadžbe, po jedna za svaki od nepoznatih regresijskih koeficijenata. Rješavanjem normalnih jednadžbi dobiti ćemo **procjenitelje najmanjih kvadrata** regresijskih koeficijenata.

7.2. Jednostavni eksponencijalni regresijski model

Jednostavni eksponencijalni regresijski model je nelinearan u parametrima i ima oblik [11]:

$$y = \alpha\beta^x \tag{7.5}$$

Da bi pronašli parametre "metodom najmanjih kvadrata" potrebno je model transformirati da bude linearan u parametrima pomoću logaritamskih transformacija:

$$logy = log\alpha + x\, log\beta \tag{7.6}$$

Pri tome logaritmirane vrijednosti parametara dobijemo minimizirajući:

$$L = \sum_{i=1}^{n} log^2 e_i = \sum_{i=1}^{n} (logy_i - log\hat{y}_i)^2 \rightarrow min \tag{7.7}$$

odnosno vrijedi:

$$\beta = \frac{\sum_{i=1}^{n} x_i logy_i - n\overline{xlogy}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} \tag{7.8}$$

$$\alpha = \overline{logy} - log\beta\bar{x} \tag{7.9}$$

Logaritamskom transformacijom model postaje linearan u parametrima (iako nelinearan u varijablama), a nezavisna varijabla x ostaje u nepromijenjenom obliku. Konačno logaritmirane vrijednosti parametara anti logaritmiramo po bazi 10 da bismo dobili parametre u početnoj jednadžbi eksponencijalne regresije.

Mjere reprezentativnosti, kao i jednadžba analize varijance se izvode isto kao kod linearne regresije, koristeći logaritmirane vrijednosti log y i originalne vrijednosti varijable x.

Procjena varijance varijable y iznosi:

$$\hat{\sigma}_{logy} = \sqrt{\frac{SSE}{n-p}} = \sqrt{\frac{\sum_{i=1}^{n} log^2 e_i}{n-p}} \tag{7.10}$$

Koeficijent varijacije regresije iznosi:

$$V = \frac{\hat{\sigma}_{logy}}{\overline{logy}} \times 100 \tag{7.11}$$

8. Primjeri

8.1. Jednostavna linearna regresija

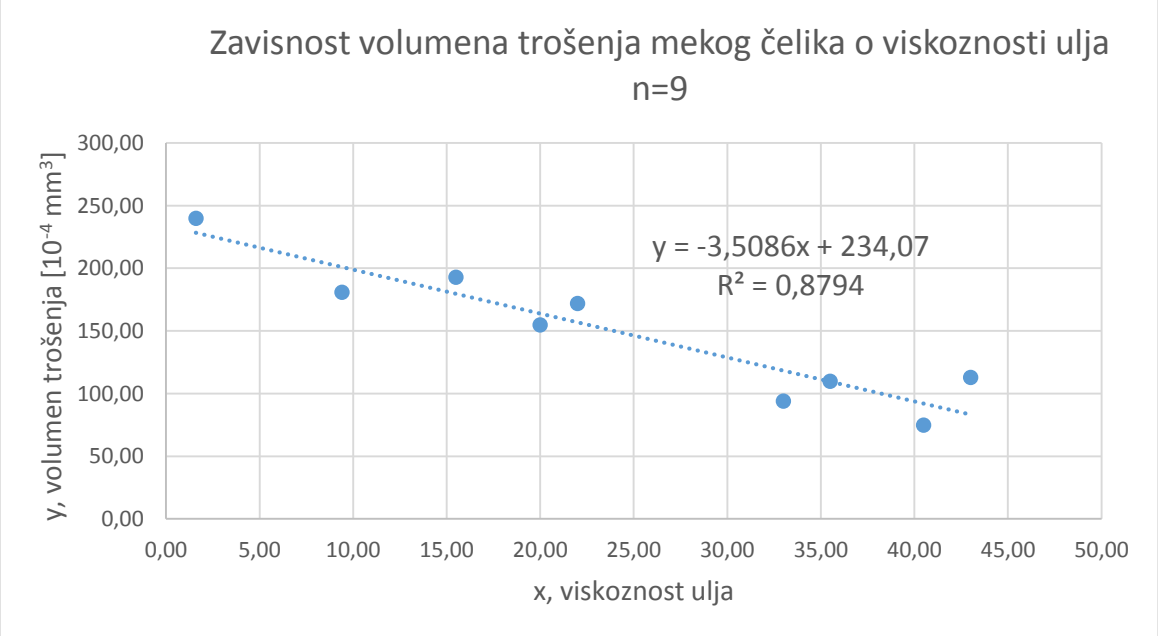
Sljedeći primjer preuzet je iz literature Montgomery D.C., Runger G.C.: Applied Statistics and Probability for Engineers, str. 412. Članak u časopisu Wear (Vol. 152, 1992, str. 171–181) prikazuje podatke o trošenju mekog čelika i viskoznosti ulja. Varijabla x prikazuje viskoznost ulja, a varijabla y volumen trošenje (10^{-4} mm^3)

Tablica 3: Zavisnost trošenja mekog čelika o viskoznosti ulja 1

y	240	181	193	155	172	110	113	75	94
x	1,6	9,4	15,5	20,0	22,0	35,5	43,0	40,5	33,0

Prvi korak u regresijskoj analizi je crtanje dijagrama rasipanja, grafičkog prikaza točaka T(x,y). Na horizontalnoj osi ističe se dio aritmetičkog mjerila koji obuhvaća opažene vrijednosti varijable x, a na vertikalnoj dio aritmetičkog mjerila koji obuhvaća opažene vrijednosti varijable y. Dijagram rasipanja omogućuje da se uoči:

- Oblik veze među odabranim varijablama
- Smjer povezanosti
- Jakost povezanosti



Slika 13: Dijagram rasipanja

Na temelju dijagrama rasipanja zaključuje se da je veza između x i y linearna (jer su točke raspoređene blizu nekog zamišljenog pravca) i negativna. Realno je za pretpostaviti da se trošenje mekog čelika i viskoznost ulja može opisati modelom:

$$y = \alpha + \beta x + e$$

Kako bi se odredio procijenjen model:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x$$

Potrebno je odrediti vrijednosti regresijskih koeficijenata $\hat{\alpha}$ *i* $\hat{\beta}$

Tablica 4: Zavisnost trošenja mekog čelika o viskoznosti ulja 2

i	x	y	xy	x^2	y^2
1	1,60	240,00	384,00	2,56	57600,00
2	9,40	181,00	1701,40	88,36	32761,00
3	15,50	193,00	2991,50	240,25	37249,00
4	20,00	155,00	3100,00	400,00	24025,00
5	22,00	172,00	3784,00	484,00	29584,00
6	35,50	110,00	3905,00	1260,25	12100,00
7	43,00	113,00	4859,00	1849,00	12769,00
8	40,50	75,00	3037,50	1640,25	5625,00
9	33,00	94,00	3102,00	1089,00	8836,00
Suma	***220,50***	***1333,00***	***26864,40***	***7053,67***	***220549,00***

Iz gornje tablice dobiveno je:

$$\sum_{i=1}^{9} x = 220{,}50 \qquad \sum_{i=1}^{9} y = 1333{,}00$$

$$\sum_{i=1}^{9} xy = 26864{,}40 \qquad \sum_{i=1}^{9} x^2 = 7053{,}67 \qquad \sum_{i=1}^{9} y^2 = 220549{,}00$$

Uvrštavanjem konkretnih vrijednosti u jednadžbe za dobivanje koeficijenata regresije dobiva se da je:

$$\beta = \frac{n\sum(xy) - \sum x \sum y}{n\sum(x^2) - (\sum x)^2} = \frac{9x26864{,}4 - 220{,}5x1333}{9x7053{,}67 - 220{,}5^2} = -3{,}5086$$

$$\alpha = \frac{\sum y \sum(x^2) - \sum x \sum(xy)}{n\sum(x^2) - (\sum x)^2} = \frac{\sum y - \beta \sum x}{n} = \frac{1333 - (-3{,}5086 x 220{,}5)}{9} = 234{,}0707$$

U konkretnom slučaju, procijenjena regresijska jednadžba glasi:

$$\hat{y} = 234{,}0707 - 3{,}5086x$$

Regresijski koeficijent $\beta = -3{,}5086$ pokazuje da će se na temelju procijenjenog modela, za smanjenje volumena trošenja mekog čelika za 10^{-4} milimetara kubnih viskoznost ulja u prosjeku smanjiti za 3,5086.

U konkretnom slučaju vrijednost $\alpha = 234{,}0707$ označava očekivanu vrijednost volumena trošenja (10^{-4} milimetara kubnih) ako viskoznost ulja iznosi 0.

Ako se za svaki i = 1, 2 ,..., n u procijenjenu regresijsku jednadžbu $\hat{y} = 234{,}0707 - 3{,}5086x$ uvrste stvarne vrijednosti nezavisne varijable x, dobivaju se regresijske vrijednosti $\hat{y}$ zavisne varijable y. Prva regresijska vrijednost $\hat{y}$ dobiva se uvrštavanjem prve vrijednosti varijable x koja iznosi $x_1 = 1{,}60$ pa je

$$\hat{y} = 234{,}0707 - 3{,}5086x1{,}6 = 228{,}46$$

Analogno se dobivaju i ostale regresijske vrijednosti. Regresijske vrijednosti $\hat{y}$ procjene su vrijednosti stvarnih vrijednosti zavisne varijable y . U konkretnom se primjeru, $\hat{y}$ se interpretira na slijedeći način: za viskoznost ulja od 1,6 očekivana vrijednost volumena trošenja mekog čelika iznosi 228,46 milimetara kubnih. Stvaran volumen trošenja mekog čelika y za vrijednost viskoznosti ulja x = 1,6 je 240. Razliku čini rezidualno odstupanje $\hat{e}$.

Rezidualna odstupanja razlike su stvarnih vrijednosti zavisne varijable od procijenjenih.

$$\hat{e}_i = y_i - \hat{y}_i \qquad i = 1{,}2, \ldots, n$$

U konkretnom primjeru, prvo rezidualno odstupanje je:

$$\hat{e}_1 = y_1 - \hat{y}_1 = 240 - 228{,}46 = 11{,}54$$

U sljedećoj tablici prikazana su sve ostale regresijske vrijednosti i rezidualna odstupanja:

Tablica 5: Zavisnost trošenja mekog čelika o viskoznosti ulja 3

i	x	y	$\hat{y}_i$	e	e^2
1	1,60	240,00	228,46	11,54	133,24
2	9,40	181,00	201,09	-20,09	403,62
3	15,50	193,00	179,69	13,31	177,21
4	20,00	155,00	163,90	-8,90	79,20
5	22,00	172,00	156,88	15,12	228,54
6	35,50	110,00	109,52	0,48	0,23
7	43,00	113,00	83,20	29,80	887,87
8	40,50	75,00	91,97	-16,97	288,12
9	33,00	94,00	118,29	-24,29	589,93
Suma	**220,50**	**1333,00**	**1333,00**	**0**	**2787,96**

$$SSE = \sum_{i=1}^{9} \hat{e}_i^2 = 2787{,}96$$

Pearsonov koeficijent regresije r iznosi:

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{[n\sum(x^2) - (\sum x)^2][n\sum(y^2) - (\sum y)^2]}}$$
$$= \frac{9x26864{,}4 - 220{,}5x1333}{\sqrt{[9x7053{,}67 - 220{,}5^2][9x220549 - 1333^2]}} = -0{,}9378$$

Pripadna p vrijednost iznosi p = 0,0002. Budući da je p < 0,05, dobiveni koeficijent je statistički značajan, tj. na uzorku od 9 promatranja postoji statistički značajna linearna veza između volumena trošenja mekog čelika i viskoznosti ulja.

Procjena pogreške regresijskog modela $\hat{\sigma}$ iznosi:

$$\hat{\sigma} = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (\alpha + \beta x_i)]^2 = \frac{SSE}{n-2} = \frac{2787{,}96}{9-2} = 19{,}96$$

Standardna greška regresije ili standardna devijacija regresije je apsolutni pokazatelj reprezentativnosti regresijskog modela, a pokazuje prosječni stupanj varijacije stvarnih vrijednosti ovisne varijable u odnosu na očekivane regresijske vrijednosti. Prikazani izraz odnosi se na standardna grešku (devijaciju) regresije jednostrukog modela. Ovaj pokazatelj izražen je u originalnim jedinicama mjere ovisne varijable y. Stoga je na temelju standardne greške regresije teško uspoređivati reprezentativnost modela s različitim mjernim jedinicama.

Taj problem eliminira relativni pokazatelj - **koeficijent varijacije regresije**, koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$V = \frac{\hat{\sigma}}{\bar{y}} \times 100 = \frac{19{,}96}{1333/9} \times 100 = 13{,}29\ \%$$

Najmanja vrijednost koeficijenta varijacije je 0 %, a najveća nije definirana. Što je koeficijent varijacije regresijskog modela bliži nuli, to je model reprezentativniji. Često se uzima dogovorena granica reprezentativnosti od 30 %. U konkretnom primjeru koeficijent varijacije je manji od 30 % stoga zaključujemo da je model dobar.

Kako bi olakšali izračun regresijskih koeficijenata te pokazatelja reprezentativnosti modela izrađena je Excel tablica. Narančastom bojom označene su varijable x i y koje samostalno unosimo dok su žutom bojom označene varijable koje su automatski izračunate pomoću Excel funkcija. Na sljedećoj slici prikazana je Excel tablica sa svim potrebnim podacima za jednostavnu linearnu regresijsku analizu.

	A	B	C	D	E	F	G	H	I
1	i	x	y	xy	X^2	y^2	$\hat{y}$	e	e^2
2	1	1,60	240,00	384,00	2,56	57600,00	228,46	11,54	133,24
3	2	9,40	181,00	1701,40	88,36	32761,00	201,09	-20,09	403,62
4	3	15,50	193,00	2991,50	240,25	37249,00	179,69	13,31	177,21
5	4	20,00	155,00	3100,00	400,00	24025,00	163,90	-8,90	79,20
6	5	22,00	172,00	3784,00	484,00	29584,00	156,88	15,12	228,54
7	6	35,50	110,00	3905,00	1260,25	12100,00	109,52	0,48	0,23
8	7	43,00	113,00	4859,00	1849,00	12769,00	83,20	29,80	887,87
9	8	40,50	75,00	3037,50	1640,25	5625,00	91,97	-16,97	288,12
10	9	33,00	94,00	3102,00	1089,00	8836,00	118,29	-24,29	589,93
11	10			0,00	0,00	0,00			0,00
12	Suma	220,50	1333,00	26864,40	7053,67	220549,00	1333,00	0,00	2787,96
13									
14									
15									
16		n					9		
17		Koeficijent β (Slope)					-3,5086		
18		Koeficijent α (Intercept)					234,0707		
19		Pearsonov koeficijent regresije, r (Correlation coefficient)					-0,9378		
20		Koeficijent determinacije, r2 (R-squared)					0,8794		
21		Standardna pogreška, σ (Standard error)					19,9570		
22		Test linearne regresije, t					-7,1444		
23		Pripadna p-vrijednost, p					0,0002		

Slika 14: Excel tablica jednostavne regresijske analize

Funkcije korištene pri izradi tablice:

Tablica 6: Excel funkcije za jednostavnu linearnu regresiju

n	**COUNT**
Koeficijent β (Slope)	**SLOPE**
Koeficijent α (Intercept)	**INTERCEPT**
Pearsonov koeficijent regresije, r (Correlation coefficient)	**CORREL**
Koeficijent determinacije, r2 (R-squared)	**RSQ**
Standardna pogreška, σ (Standard error)	**STEYX**
Pripadna p-vrijednost, p	**TDIST**

Dobivene rezultate možemo usporediti sa rezultatima dobivenim pomoću programske naredbe „Dana Analysis“ u programskom alatu Excel. Na sljedećoj slici prikazani su rezultati dobiveni pomoću programske naredbe „Dana Analysis“ u programskom alatu Excel:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,9378
R Square	0,8794
Adjusted R Square	0,8622
Standard Error	19,9570
Observations	9,0000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1,0000	20328,9259	20328,9259	51,0417	0,0002
Residual	7,0000	2787,9630	398,2804		
Total	8,0000	23116,8889			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	234,0707	13,7484	17,0253	0,0000	201,5609	266,5805	201,5609	266,5805
x	-3,5086	0,4911	-7,1444	0,0002	-4,6698	-2,3473	-4,6698	-2,3473

RESIDUAL OUTPUT

Observation	*Predicted y*	*Residuals*	*Standard Residuals*
1	228,4570	11,5430	0,6183
2	201,0903	-20,0903	-1,0762
3	179,6881	13,3119	0,7131
4	163,8996	-8,8996	-0,4767
5	156,8825	15,1175	0,8098
6	109,5170	0,4830	0,0259
7	83,2028	29,7972	1,5962
8	91,9742	-16,9742	-0,9093
9	118,2884	-24,2884	-1,3011

PROBABILITY OUTPUT

Percentile	*y*
5,5556	75,0000
16,6667	94,0000
27,7778	110,0000
38,8889	113,0000
50,0000	155,0000
61,1111	172,0000
72,2222	181,0000
83,3333	193,0000
94,4444	240,0000

Slika 15: Rezultati jednostavne linearne regresije dobiveni pomoću naredbe Data Analysis u programskom alatu Excel

Kao što je vidljivo na slici svi rezultati podudaraju se s rezultatima dobivenim ručnim postupkom ili pomoću izrađenog algoritma u Excel tablici. U praksi koeficijenti regresijske analize kao i svi ostali statistički pokazatelji ne računaju se ručno nego se koriste gotovi programski alati kao npr. naredba „Dana Analysis“ u programskom alatu Excel ili programski alat Statistica.

8.2. Višestruka linearna regresija

Sljedeći primjer preuzet je iz literature Ross S., Probability and Statistics for Engineers and Scientists, str. 434. Vrijeme potrebno za otkaz komponente stroja povezano je s radnim naponom x_1, brojem okretaja motora u okretajima po minuti x_2 i radnom temperaturom x_3. Na temelju eksperimenata provedenih u laboratoriju za istraživanje i razvoj dobiveni su sljedeći podaci gdje y predstavlja vrijeme potrebno do pojave kvara komponente stroja u minutama:

Tablica 7: Zavisnost vremena otkazivanja komponente stroja o radnoj temperaturi, radnom naponu i broju okretaja motora 1

i	***y, min***	***x_1, V***	***x_2, min^{-1}***	***x_3, °C***
1	2145	110	750	140
2	2155	110	850	180
3	2220	110	1000	140
4	2225	110	1100	180
5	2260	120	750	140
6	2266	120	850	180
7	2334	120	1000	140
8	2340	130	1000	180
9	2212	115	840	150
10	2180	115	880	150
Suma	***22337***	***1160***	***9020***	***1580***

Kako bi se odredio procijenjen model:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

Potrebno je odrediti vrijednosti regresijskih koeficijenata $\hat{\alpha}$ *i* $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

Iz tablice 6 dobiveni su sljedeći podaci:

$$\sum_{i=1}^{10} x_1 = 1160 \qquad \sum_{i=1}^{10} x_2 = 9020 \qquad \sum_{i=1}^{10} x_3 = 1580 \qquad \sum_{i=1}^{10} y = 22337$$

$$\sum_{i=1}^{10} x_1^2 = 134950 \qquad \sum_{i=1}^{10} x_2^2 = 8260000 \qquad \sum_{i=1}^{10} x_3^2 = 253000 \qquad \sum_{i=1}^{10} y^2 = 49934931$$

$$\sum_{i=1}^{10} x_1 x_2 = 1046800 \qquad \sum_{i=1}^{10} x_1 x_3 = 183500 \qquad \sum_{i=1}^{10} x_2 x_3 = 1432000$$

$$\sum_{i=1}^{10} x_1 y = 2594430 \qquad \sum_{i=1}^{10} x_2 y = 20179580 \qquad \sum_{i=1}^{10} x_3 y = 3530540$$

Normalne jednadžbe za višestruku linearnu regresiju dobivene metodom najmanjih kvadrata u poglavlju 6.1. glase:

$$n\hat{\alpha} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\alpha} \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} {x_{i1}}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\hat{\alpha} \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} x_{ik} y_i$$

Uvrštavanjem konkretnih vrijednosti u normalne jednadžbe dobiva se:

$$10\hat{\alpha} + 1160\hat{\beta}_1 + 9020\hat{\beta}_2 + 1580\hat{\beta}_3 = 22337$$

$$1160\hat{\alpha} + 134950\hat{\beta}_1 + 1046800\hat{\beta}_2 + 183500\hat{\beta}_3 = 2594430$$

$$9020\hat{\alpha} + 1046800\hat{\beta}_1 + 8260000\hat{\beta}_2 + 1432000\,\hat{\beta}_3 = 20179580$$

$$1580\hat{\alpha} + 183500\hat{\beta}_1 + 1432000\hat{\beta}_2 + 253000\hat{\beta}_3 = 3530540$$

Rješavanjem sustava jednadžbi dobivaju se koeficijenti višestruke linearne regresije:

$$\hat{\alpha} = 1108{,}72 \qquad \hat{\beta}_1 = 8{,}6393$$

$$\hat{\beta}_2 = 0{,}2608 \qquad \hat{\beta}_3 = -0{,}7114$$

Za konkretni slučaj, procijenjena regresijska jednadžba glasi:

$$\hat{y} = 1108{,}72 + 8{,}6393x_1 + 0{,}2608x_2 - 0{,}7114x_3$$

Ako se za svaki i = 1, 2 ,..., n u procijenjenu regresijsku jednadžbu $\hat{y} = 1108{,}72 + 8{,}6393x_1 + 0{,}2608x_2 - 0{,}7114x_3$ uvrste stvarne vrijednosti nezavisnih varijabli x, dobivaju se regresijske vrijednosti $\hat{y}$ zavisne varijable y. Prva regresijska vrijednost $\hat{y}$ dobiva se uvrštavanjem varijabli x_1, x_2, x_3 pa je

$$\hat{y} = 1108{,}72 + 8{,}6393x110 + 0{,}2608x750 - 0{,}7114x140 = 2155{,}03$$

Analogno se dobivaju i ostale regresijske vrijednosti. Regresijske vrijednosti $\hat{y}$ procjene su vrijednosti stvarnih vrijednosti zavisne varijable y . U konkretnom se primjeru, $\hat{y}$ se interpretira na slijedeći način: za radni napon od 110 V, broj okretaja motora 750 min^{-1} i radnu temperaturu 140 °C očekivano vrijeme pojave kvara komponente stroja je nakon 2155,03 minute. Stvaran vrijeme pojave kvara komponente stroja y za radni napon od 110 V, broj okretaja motora 750 min^{-1} i radnu temperaturu 140 °C iznosi 2145 minuta. Razliku čini rezidualno odstupanje $\hat{e}$.

Rezidualna odstupanja razlike su stvarnih vrijednosti zavisne varijable od procijenjenih.

$$\hat{e}_i = y_i - \hat{y}_i \qquad i = 1{,}2, \dots, n$$

U konkretnom primjeru, prvo rezidualno odstupanje je:

$$\hat{e}_1 = y_1 - \hat{y}_1 = 2145 - 2155{,}03 = -10{,}03$$

U sljedećoj tablici prikazana su sve ostale regresijske vrijednosti i rezidualna odstupanja:

Tablica 8: Zavisnost vremena otkazivanja komponente stroja o radnoj temperaturi, radnom naponu i broju okretaja motora 2

i	***y, min***	***x_1, V***	***x_2, min^{-1}***	***x_3, °C***	***$\hat{y}_i$***	***e_i***	***e^2***
1	2145	110	750	140	2155,03	-10,03	100,64
2	2155	110	850	180	2152,65	2,35	5,51
3	2220	110	1000	140	2220,22	-0,22	0,05
4	2225	110	1100	180	2217,85	7,15	51,18
5	2260	120	750	140	2241,43	18,57	345,01
6	2266	120	850	180	2239,05	26,95	726,50
7	2334	120	1000	140	2306,62	27,38	749,76
8	2340	130	1000	180	2364,56	-24,56	602,96
9	2212	115	840	150	2214,58	-2,58	6,68
10	2180	115	880	150	2225,01	-45,01	2026,35
Suma	***22337***	***1160***	***9020***	***1580***		***0***	***4615***

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = 4615$$

Kao i kod primjera jednostavne linearne regresije važna je procjena varijance $\hat{\sigma}$. Kod jednostavne linearne regresije procjena standardne greške regresije $\hat{\sigma}$ dobila dijeljenjem zbroja kvadrata reziduala sa n-2. S obzirom da kod jednostavne linearne regresije postoje dva parametra, kod višestruke linearne regresije javlja se p parametara, stoga je logična procjena varijance $\hat{\sigma}$:

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-p}} = \sqrt{\frac{4615}{10-4}} = 27{,}73$$

Koeficijent varijacije regresije koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$V = \frac{\hat{\sigma}}{\bar{y}} \times 100 = \frac{27{,}73}{22337/10} \times 100 = 1{,}24\ \%$$

U konkretnom primjeru koeficijent varijacije iznosi 1,24 % što je manje od 30 % stoga zaključujemo da je model dobar.

Kako bi olakšali izračun regresijskih koeficijenata kao i kod primjera jednostavne linearne regresije izrađena je Excel tablica. Narančastom bojom označene su varijable x_1, x_2, x_3, i y koje samostalno unosimo dok su žutom bojom označene varijable koje su automatski izračunate pomoću Excel funkcija. Na sljedećoj slici prikazana je Excel tablica sa svim potrebnim podacima za višestruku linearnu regresijsku analizu:

	A	B	C	D	E	F	G	H	I
1	i	x1	x2	x3	y	(x1)^2	(x2)^2	(x3)^2	y^2
2	1	110	750	140	2145	12100	562500	19600	4601025
3	2	110	850	180	2155	12100	722500	32400	4644025
4	3	110	1000	140	2220	12100	1000000	19600	4928400
5	4	110	1100	180	2225	12100	1210000	32400	4950625
6	5	120	750	140	2260	14400	562500	19600	5107600
7	6	120	850	180	2266	14400	722500	32400	5134756
8	7	120	1000	140	2334	14400	1000000	19600	5447556
9	8	130	1000	180	2340	16900	1000000	32400	5475600
10	9	115	840	150	2212	13225	705600	22500	4892944
11	10	115	880	150	2180	13225	774400	22500	4752400
12	Suma	1160	9020	1580	22337	134950	8260000	253000	49934931
13									
14									
15									
16									
17									
18									
19	x1x2	x1x3	x1y	x2x3	x2y	x3y	$\hat{y}$	e	e^2
20	82500	15400	235950	105000	1608750	300300	2155,03	-10,03	100,64
21	93500	19800	237050	153000	1831750	387900	2152,65	2,35	5,51
22	110000	15400	244200	140000	2220000	310800	2220,22	-0,22	0,05
23	121000	19800	244750	198000	2447500	400500	2217,85	7,15	51,18
24	90000	16800	271200	105000	1695000	316400	2241,43	18,57	345,01
25	102000	21600	271920	153000	1926100	407880	2239,05	26,95	726,50
26	120000	16800	280080	140000	2334000	326760	2306,62	27,38	749,76
27	130000	23400	304200	180000	2340000	421200	2364,56	-24,56	602,96
28	96600	17250	254380	126000	1858080	331800	2214,58	-2,58	6,68
29	101200	17250	250700	132000	1918400	327000	2225,01	-45,01	2026,35
30	1046800	183500	2594430	1432000	20179580	3530540	22337,00	0	4615
31									
32									
33		Rješenja koeficijenata jednadžbe pravca:							
34		β3	β2	β1	α				
35		-0,7114	0,2608	8,6393	1108,7245				

Slika 16: Excel tablica višestruke regresijske analize

Funkcija korištena pri izradi tablice:

Tablica 9: Excel funkcija za višestruku linearnu regresiju

Koeficijenti α i β	**LINEST**

Kao i u primjeru jednostavne linearne regresije dobivene rezultate možemo usporediti sa rezultatima dobivenim pomoću programske naredbe „Dana Analysis“ u programskom alatu Excel. Na sljedećoj slici prikazani su rezultati dobiveni pomoću programske naredbe „Dana Analysis“ u programskom alatu Excel:

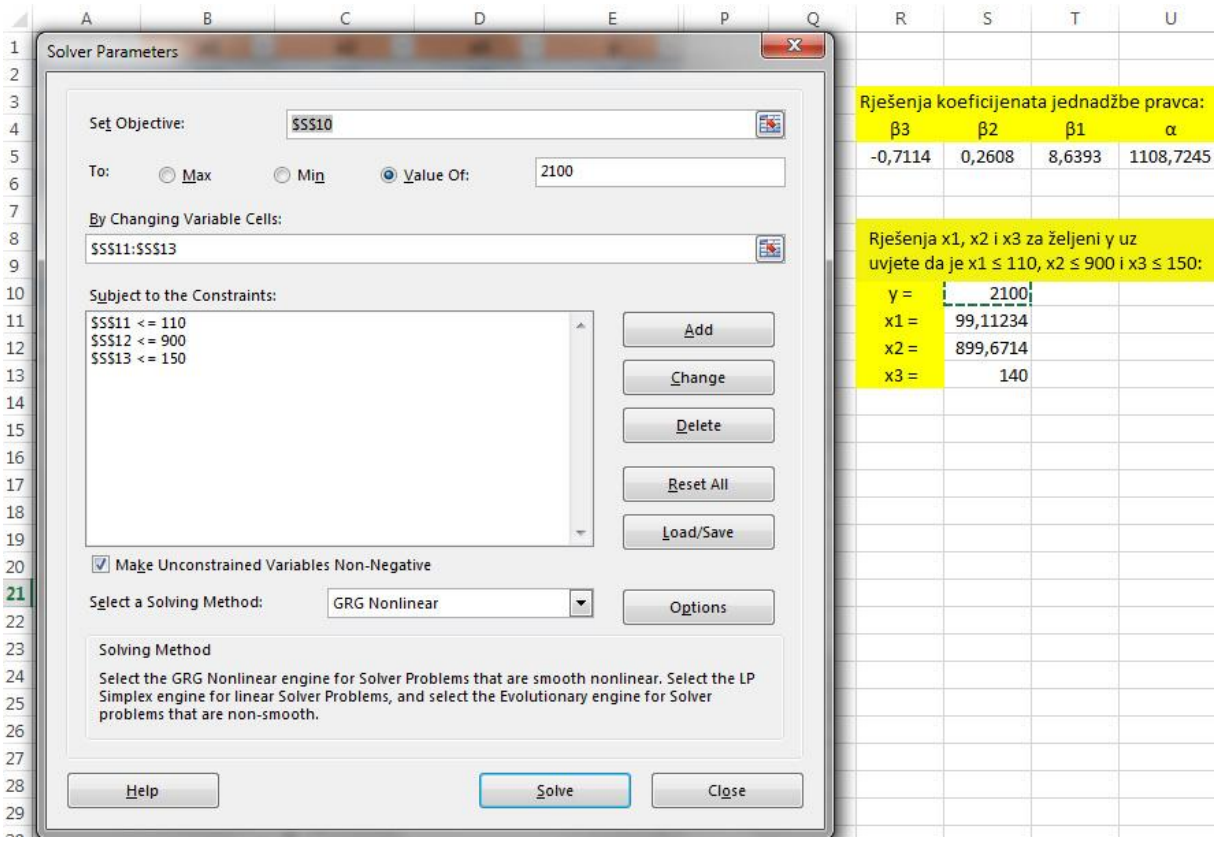
Regression Statistics								
Multiple R	0,9417							
R Square	0,8868							
Adjusted R Square	0,8302							
Standard Error	27,7328							
Observations	10,0000							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3,0000	36159,4495	12053,1498	15,6716	0,0030			
Residual	6,0000	4614,6505	769,1084					
Total	9,0000	40774,1000						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	1108,7245	175,8723	6,3041	0,0007	678,3804	1539,0686	678,3804	1539,0686
x1	8,6393	1,4310	6,0373	0,0009	5,1378	12,1409	5,1378	12,1409
x2	0,2608	0,0836	3,1191	0,0206	0,0562	0,4653	0,0562	0,4653
x3	-0,7114	0,5162	-1,3781	0,2174	-1,9746	0,5517	-1,9746	0,5517
RESIDUAL OUTPUT					PROBABILITY OUTPUT			
Observation	*Predicted y*	*Residuals*	*Standard Residuals*		*Percentile*	*y*		
1	2155,0322	-10,0322	-0,4430		5,0000	2145,0000		
2	2152,6530	2,3470	0,1036		15,0000	2155,0000		
3	2220,2249	-0,2249	-0,0099		25,0000	2180,0000		
4	2217,8457	7,1543	0,3160		35,0000	2212,0000		
5	2241,4255	18,5745	0,8203		45,0000	2220,0000		
6	2239,0463	26,9537	1,1903		55,0000	2225,0000		
7	2306,6182	27,3818	1,2092		65,0000	2260,0000		
8	2364,5552	-24,5552	-1,0844		75,0000	2266,0000		
9	2214,5841	-2,5841	-0,1141		85,0000	2334,0000		
10	2225,0150	-45,0150	-1,9880		95,0000	2340,0000		

Slika 17: Rezultati višestruke linearne regresije dobiveni pomoću naredbe Data Analysis u programskom alatu Excel

Kao i u primjeru jednostavne linearne regresije svi rezultati se podudaraju sa ručno dobivenim rezultatima te rezultatima dobivenih pomoću izrađenog algoritma u Excel tablici.

8.2.1. Optimizacija nezavisnih varijabli uz ciljanu izlaznu vrijednost (nezavisnu varijablu)

Nakon dobivanja procijenjenih koeficijenata višestruke linearne regresije te procijenjene regresijske jednadžbe $\hat{y} = 1108{,}72 + 8{,}6393x_1 + 0{,}2608x_2 - 0{,}7114x_3$ pretpostavimo da želimo dobiti ciljanu izlaznu vrijednost odnosno vrijeme potrebno za otkaz komponente stroja y = 2100 min uz uvjete da nezavisne varijable iznose: radni napon stroja $x_1 \leq 110$ V, broj okretaja motora $x_2 \leq 900$ min^{-1}, radna temperatura stroja $x_3 \leq 150$ °C. S obzirom da je vrijeme potrebno za otkaz komponente stroja povezano s radnim naponom x_1, brojem okretaja motora x_2 i radnom temperaturom x_3 imamo jednu jednadžbu s tri nepoznanice. Za optimizaciju parametara x_1, x_2 i x_3 koristi se naredba „Solver“ u programskom alatu Excel. Na sljedećoj slici prikazana je naredba „Solver“ te dobiveni rezultati uz zadane uvjete:



Slika 18: Naredba "Solver" u programskom alatu Excel

Za ciljanu vrijednost zavisne varijable y = 2100 optimizirane nezavisne varijable iznose x_1 = 99,11, x_2 = 899,67, x_3 = 140 što znači da će se otkaz komponente stroja pojavit nakon 2100 minuta rada ako je radni napon stroja 110 V, brzina vrtnje stroja 899,67 min^{-1} te radna temperature stroja 140 °C.

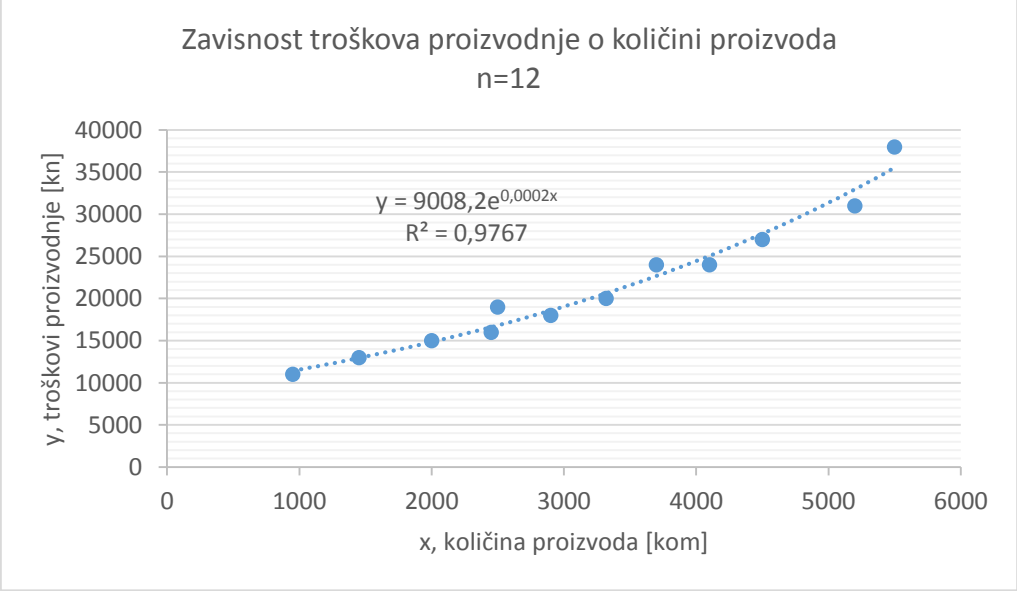
8.3. Jednostavna eksponencijalna regresija

U slijedećem primjeru podaci su simulirani. Troškovi proizvodnje tvrtke ovise o količini proizvedenih proizvoda. Promatrani su troškovi i proizvodnja svaki mjesec tijekom jedne godine i dobiveni su sljedeći podaci:

Tablica 10: Zavisnost troškova proizvodnje o količini proizvoda 1

i	*x, kom*	*y, kn*
1	950	11000
2	2500	19000
3	1450	13000
4	2000	15000
5	4500	27000
6	2450	16000
7	3700	24000
8	5500	38000
9	4100	24000
10	2900	18000
11	3320	20000
12	5200	31000
Suma	*38570*	*256000*

Prvi korak u regresijskoj analizi je crtanje dijagrama rasipanja, grafičkog prikaza točaka T(x,y). Na horizontalnoj osi ističe se dio aritmetičkog mjerila koji obuhvaća opažene vrijednosti varijable x, a na vertikalnoj dio aritmetičkog mjerila koji obuhvaća opažene vrijednosti varijable y.



Slika 19: Dijagram rasipanja

Na temelju dijagrama rasipanja zaključujemo se da je veza između x i y nelinearna odnosno logaritamska i pozitivna. Realno je za pretpostaviti da se troškovi proizvodnje i količina proizvedenih proizvoda može opisati modelom:

$$y = \alpha\beta^{x}$$

Kako bi se odredio procijenjen model:

$$\hat{y} = \hat{\alpha}\hat{\beta}^{x}$$

Potrebno je odrediti vrijednosti regresijskih koeficijenata $\hat{\alpha}\ i\ \hat{\beta}$. Da bi pronašli parametre metodom najmanjih kvadrata potrebno je model transformirati da bude linearan u parametrima pomoću logaritamskih transformacija:

$$log\hat{y} = log\hat{\alpha} + x\ log\hat{\beta}$$

Metodom najmanjih kvadrata dolazimo do izraza za regresijske koeficijente $\hat{\alpha}\ i\ \hat{\beta}$:

$$log\hat{\beta} = \frac{\sum_{i=1}^{n} x_i log y_i - n\overline{xlogy}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2}$$

$$log\hat{\alpha} = \overline{logy} - log\beta\bar{x}$$

Tablica 11: : Zavisnost troškova proizvodnje o količini proizvoda 2

i	***x***	***y***	***log y***	***logy x***	***x^2***	***$logy^2$***
1	950	11000	4,04	3839,32	902500	16
2	2500	19000	4,28	10696,88	6250000	18
3	1450	13000	4,11	5965,22	2102500	17
4	2000	15000	4,18	8352,18	4000000	17
5	4500	27000	4,43	19941,14	20250000	20
6	2450	16000	4,20	10300,09	6002500	18
7	3700	24000	4,38	16206,78	13690000	19
8	5500	38000	4,58	25188,81	30250000	21
9	4100	24000	4,38	17958,87	16810000	19
10	2900	18000	4,26	12340,29	8410000	18
11	3320	20000	4,30	14279,42	11022400	18
12	5200	31000	4,49	23355,08	27040000	20
Suma	***38570***	***256000***	***51,63***	***168424,09***	***146729900***	***222,44***

Iz gornje tablice dobiveno je:

$$\sum_{i=1}^{12} x = 38570 \qquad \sum_{i=1}^{12} y = 256000 \qquad \sum_{i=1}^{12} logy = 51{,}6335$$
$$\sum_{i=1}^{12} logyx = 168424{,}09 \qquad \sum_{i=1}^{12} x^2 = 146729900 \qquad \sum_{i=1}^{12} logy^2 = 222{,}44$$

Uvrštavanjem konkretnih vrijednosti u jednadžbe za dobivanje koeficijenata regresije dobiva se da je:

$$log\hat{\beta} = \frac{\sum_{i=1}^{n} x_i logy_i - n\overline{xlogy}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = \frac{168424{,}09 - 12x\frac{38570}{12}x\frac{51{,}63}{12}}{146729900 - 12x(\frac{38570}{12})^2}$$
$$= 0{,}0001008325$$

$$log\hat{\alpha} = \overline{logy} - log\hat{\beta}\bar{x} = \frac{51{,}63}{12} - 0{,}0001008325x\frac{38570}{12} = 3{,}9546$$

U konkretnom slučaju, logaritamska regresijska jednadžba glasi:

$$log\hat{y} = 3{,}9546 + 0{,}000108325x$$

Nakon antilogaritmiranja logaritamska regresijska jednadžba glasi:

$$\hat{y} = 9007{,}8 * 1{,}00023^x$$

Parametar $\beta = 1{,}00023$ pokazuje kako možemo očekivati porast troškova u prosjeku za 0,023 % ako proizvodnja poraste za 1 komad.
Parametar $\alpha = 9007{,}8$ pokazuje kako možemo očekivati troškove od 9007,8 kuna, kada proizvodnja iznosi 0 komada (fiksni troškovi).

Ako se za svaki i = 1, 2 ,..., n u procijenjenu regresijsku jednadžbu $\hat{y} = 9007{,}8 * 1{,}00023^x$ uvrste stvarne vrijednosti nezavisne varijable x, dobivaju se regresijske vrijednosti $\hat{y}$ zavisne varijable y. Prva regresijska vrijednost $\hat{y}$ dobiva se uvrštavanjem prve vrijednosti varijable x koja iznosi $x_1 = 950$ pa je

$$\hat{y} = 9007{,}8 * 1{,}00023^x = 9007{,}8 * 1{,}00023^{950} = 11207$$

Analogno se dobivaju i ostale regresijske vrijednosti. Regresijske vrijednosti $\hat{y}$ procjene su vrijednosti stvarnih vrijednosti zavisne varijable y . U konkretnom se primjeru, $\hat{y}$ se interpretira na slijedeći način: za 950 komada gotovih proizvoda očekivana vrijednost troškova iznosi 11207 kuna. Stvarni troškovi y za 950 komada proizvoda je 11000 kuna. Razliku čini rezidualno odstupanje $\hat{e}$.

U sljedećoj tablici prikazana su sve ostale regresijske vrijednosti i rezidualna odstupanja:

Tablica 12: : Zavisnost troškova proizvodnje o količini proizvoda 3

i	x	y	$\log y$	$\hat{y}_i$	$\log\hat{y}_i$	$\log e$	$\log^2 e$
1	950	11000	4,04	11207,3	4,0495	-0,0081	0,0001
2	2500	19000	4,28	16006,98	4,2043	0,0744	0,0055
3	1450	13000	4,11	12573,01	4,0994	0,0145	0,0002
4	2000	15000	4,18	14268,27	4,1544	0,0217	0,0005
5	4500	27000	4,43	25354,9	4,4041	0,0273	0,0007
6	2450	16000	4,20	15823,97	4,1993	0,0048	0,0000
7	3700	24000	4,38	21094,09	4,3242	0,0561	0,0031
8	5500	38000	4,58	31910,83	4,5039	0,0758	0,0058
9	4100	24000	4,38	23126,58	4,3641	0,0161	0,0003
10	2900	18000	4,26	17549,3	4,2443	0,0110	0,0001
11	3320	20000	4,30	19328,93	4,2862	0,0148	0,0002
12	5200	31000	4,49	29783,46	4,4740	0,0174	0,0003
Suma	**22337**	**1160**	**9020**	**238027,6**	**51,3077**	**0,3259**	**0,0169**

$$SSE = \sum_{i=1}^{12} (logy_i - log\hat{y}_i)^2 = \sum_{i=1}^{12} log^2 e_i = 0{,}0169$$

Kao i kod primjera linearne regresije važna je procjena varijance. **Procjena varijance** varijable y iznosi:

$$\hat{\sigma}_{logy} = \sqrt{\frac{SSE}{n-p}} = \sqrt{\frac{0{,}0169}{12-2}} = 0{,}0411$$

Koeficijent varijacije regresije koji predstavlja postotak standardne greške regresije od aritmetičke sredine varijable y:

$$V = \frac{\hat{\sigma}_{logy}}{\overline{logy}} \times 100 = \frac{0{,}0411}{51{,}6335/12} \times 100 = 0{,}96\,\%$$

U konkretnom primjeru koeficijent varijacije iznosi 0,96 % što je manje od 30 % stoga zaključujemo da je model dobar.

9. Literatura

[1] Pivac S., Šego B.: Statistika, Alka script, Zagreb, 2014
[2] Šuvak N., Benšić M.: Primijenjena statistika, Osijek, 2013
[3] Štambuk A., Biljan-August M.: Regresijska i korelacijska analiza, Rijeka, 2013
[4] Cajner H.: Korelacija i regresija, skripte s predavanja, 2012
[5] Bahovec V.: Jednostavna i višestruka linearna regresija, skripte s predavanja
[6] Montgomery D.C., Runger G.C.: Applied Statistics and Probability for Engineers, John Wiley & Sons, 2010
[7] Ross S.: Probability and Statistics for Engineers and Scientists, Elsevier, 2009
[8] http://en.wikipedia.org/wiki/Regression_analysis
[9] Korenjak A.: Regresijska analiza, diplomski rad, Maribor, 2010
[10] Čižmešija M.: Regresijska analiza, skripte s predavanja
[11] Jurun E.: Nelinearna regresijska analiza, skripte s predavanja