5. Linearna regresija s jednom i više regresionih varijabli

Odluke menadžmenta često se zasnivaju na odnosima između dve ili više varijabli. Na primer, menadžer marketinga može pokušati da predvidi prodaju za određeni nivo troškova oglašavanja nakon ispitivanja odnosa između tih izdataka i prodaje.

U drugom slučaju, javno poduzeće može koristiti odnos između maksimalne dnevne temperature i potrebe za električnom energijom, za predviđanje potrošnje električne energije. Menadžer se ponekad oslanja na intuiciju i intuitivno procenjuje kako su dve varijable povezane. Međutim, ako je moguće dobiti podatke, ima smisla koristiti statistički postupak koji se zove regresiona analiza kako bi se pokazalo kako su te dve varijable međusobno povezane.

U terminologiji regresije, varijabla čija se vrednost predviđa naziva se zavisna varijabla.

Varijabla ili varijable koje se koriste za predviđanje vrednosti zavisne varijable nazivaju se nezavisne varijable.

U analizi uticaja izdataka za oglašavanje na prodaju, prodaja bi bila zavisna varijabla. Izdaci za oglašavanje bili bi nezavisna varijabla. U statističkom zapisu y označava zavisnu varijablu, a x označava nezavisnu varijablu.

U ovom odjeljku će se objasniti najjednostavnija vrsta regresione analize koja uključuje jednu nezavisnu varijablu i jednu zavisnu varijablu. Odnos između dve varijable se aproksimira pravom linijom. Naziva se jednostavnom linearnom regresijom. Regresiona analiza koja uključuje dve ili više nezavisnih varijabli naziva se višestruka regresiona analiza.

5.1 Jednostavni linearni regresioni model

Best Burger je lanac restorana brze hrane koji se nalazi u području s više država. Najbolje lokacije Burgera nalaze se u blizini univerzitetskih kampusa. Menadžeri veruju da je tromesečna prodaja u ovm restoranima (označeno s y) u pozitivnoj

korelaciji s veličinom studentske populacije (označeno s x). Restorani u blizini kampusa s velikim brojem studenata obično generišu veću prodaju od onih u blizini kampusa s malim brojem studenata. Pomoću regresione analize možemo razviti jednačinu koja pokazuje kako je y zavisna varijabla povezana s nezavisnom varijablom x.

5.2 Regresioni model i regresiona jednačina

U slučaju Best Burgera, populaciju čine svi restorani Best Burger. Za svaki restoran u populaciji postoji vrednost x (studentska populacija) i odgovarajuća vrednost y (tromesečna prodaja). Jednačina koja opisuje kako je y povezana s x zove se regresioni model.

$$y = \beta_0 + \beta_1 x + \epsilon$$

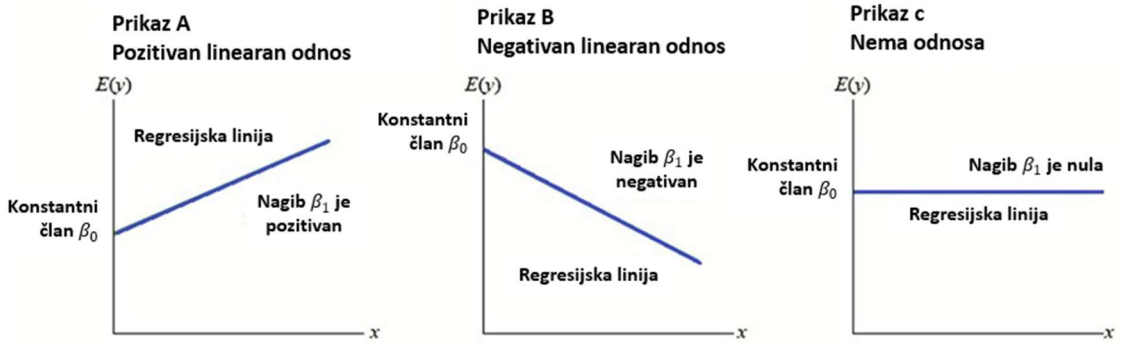
β_0 i β_1 nazivaju se parametri modela, ϵ (grčko slovo epsilon) je slučajna varijabla koja se naziva greška modela. Greška predstavlja varijabilnost y što se ne može objasniti linearnim odnosom između x i y.

Populacija svih restorana Best Burger takođe se može posmatrati kao zbirka podpopulacija, jedna za svaku zasebnu vrednost x. Na primer, jednu podpopulaciju čine svi restorani Best Burger u blizini univerzitetskih kampusa s 8000 studenata. Drugu podpopulaciju čine svi restorani Best Burger koji se nalaze u blizini univerzitetskih kampusa s 9000 studenata i tako dalje. Svaka subpopulacija ima odgovarajuću raspodelu vrednosti y. Svaka raspodela vrednosti y ima svoju srednju ili očekivanu vrednost. Jednačina koja opisuje šta je očekivana vrednost y, označena s $E(y)$, čime je povezana s x naziva se regresiona jednačina. Regresiona jednačina za jednostavnu linearnu regresiju je sledeća:

$$E(y) = \beta_0 + \beta_1 x$$

Grafikon jednostavne jednačine linearne regresije je prava linija. β_0 predstavlja početnu vrednost regresione linije, β_1 je koeficijent smera linije i $E(y)$ srednju vrednost ili očekivanu vrednost y za datu vrednost x.

Primeri mogućih regresionih linija prikazani su na slici 5.1 u nastavku. Regresiona linija y u slučaju A pokazuje da je vrednost y u pozitivnoj korelaciji s x. Kako se vrednosti x povećavaju, vrednosti $E(y)$ se takođe povećavaju. Tamo gde su manje vrednosti $E(y)$ povezane su s višim vrednostima x. Regresiona linija na prikazu C prikazuje slučaj kada vrednost y nije povezana s x. To znači da je vrednost y ista za svaku vrednost x.



Slika 4.7Primeri grafikona linearnog odnosa.

5.3 Procenjena regresiona jednačina

Kad bi bile poznate vrednosti parametara populacije β_0 i β_1, mogli bismo koristiti gornju jednačinu za izračunavanje vrednosti y za zadatu vrednost x. U praksi je tim parametrima teško pristupiti, pa se jednostavno procenjuju korišćenjem podataka uzorka. Statistika uzorka (označena s b_0 i b_1) utvrđena je kao procena parametara populacije β_0 i β_1. Zamenom vrednosti statistike uzorka b_0 i b_1 umesto β_0 i β_1 u regresionoj jednačini dobija se nova, procenjena regresiona jednačina. Procenjena regresiona jednačina za jednostavnu linearnu regresiju je sledeća:

$$\hat{y} = b_0 + b_1 x$$

Grafikon procenjene jednostavne linearne regresije naziva se procenjena regresiona linija. b_0 predstavlja početnu vrednost regresione linije, b_1 je koeficijent smera linije.

U nastavku ćemo pokazati kako koristiti metodu najmanjih kvadrata za izračunavanje vrednosti b_0 i b_1 u procenjenoj regresionoj jednačini.

Generalno je $\hat{y}$ (rezultat za $E(y)$) prosečna vrednost y za datu vrednost x. Ako sada želimo proceniti očekivanu vrednost tromesečne prodaje za sve restorane Best Burger koji se nalaze u blizini kampusa s 10 000 studenata, vrednost x bi bila zamijenjena vednošću 10 000 u posljednjoj jednačini. U nekim slučajevima, međutim, možda ćemo biti više zainteresovani za predviđanje prodaje samo za jedan određeni restoran, na primer, pretpostavimo da želite predvideti kvartalnu prodaju za restoran koji planirate izgraditi u blizini fakulteta s 10 000 studenata, pokazalo se da je čak i u ovom slučaju najbolji prediktor vrednosti y za datu x vrednost $\hat{y}$.

5.4 Metoda najmanjih kvadrata

Metoda najmanjih kvadrata je postupak u kojem se pomoću uzoraka podataka nalazi jednačina procenjene regresione linije. Kako bismo ilustrovali metodu najmanjih kvadrata, pretpostavimo da su podaci prikupljeni iz uzorka od 10 restorana s najboljim hamburgerima u blizini univerzitetskih kampusa. Sa x_i će se označiti veličina studentske populacije (u hiljadama) i veličina y_i tromesečna prodaja (u hiljadama EUR). Vrednosti za x_i *i* y_i za 10 uzoraka restorana sažeti su u tabeli u nastavku. Vidimo da je restoran 1, za x_1= 2 i y_1= 58, blizu kampusa s 2000 studenata i ima kvartalnu prodaju od 58 000 €. Restoran 2, s x_2= 6 i y_2= 105, blizu je kampusa sa 6000 studenata i ima kvartalnu prodaju od 105.000 €. Restoran s najvećom prodajnom vrednošću je restoran 10, koji se nalazi u blizini kampusa s 26.000 studenata i ima kvartalnu prodaju od 202.000 €.

Sledi dijagram rasipanja podataka na slici 5.2 u nastavku. Studentska populacija prikazana je na horizontalnoj osi, a kvartalna prodaja na vertikalnoj osi. Dijagrami rasipanja za regresionu analizu konstruisani su s nezavisnom varijablom x na horizontalnoj osi i zavisnom varijablom y na vertikalnoj osi. Stoga, dijagram rasipanja omogućava izvođenje preliminarnih zaključaka o mogućem odnosu između varijabli.

Restoran	Studentska populacija (u 1000-ama)	Kvartalna prodaja (u 1000ama eura)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Slika 4.8 Dijagram rasipanja podataka.

Koji se preliminarni zaključci mogu izvući sa slike 5.3? Veća tromesečna prodaja događa se u kampusima s većom populacijom studenata. Osim toga, postoji konstantan odnos između veličine studentske populacije i tromesečne prodaje, koji se može opisati pravom linijom. Između x *i* y zaista postoji pozitivan linearni

odnos. Stoga smo odabrali jednostavan linearni regresioni model za prikaz odnosa između tromesečne prodaje i studentske populacije. S obzirom na ovaj izbor, naš sledeći zadatak je koristiti tabelu podataka uzorka za određivanje vrednosti b_0 i b_1, koji su važni parametri u proceni jednostavne jednačine linearne regresije. Za i-ti restoran procenjena regresiona jednačina je:

$$\hat{y}_i = b_0 + b_1 x_i$$

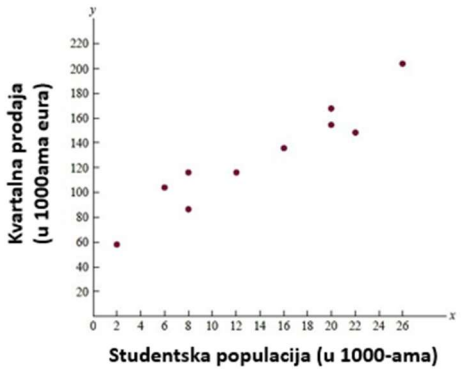
gde je

$\hat{y}_i$_ procenjena vrednost tromesečne prodaje (1000 €) za i-ti restoran

b_0_ početna vrednost procenjene regresione linije

b_1_ koeficijent smera procenjene regresione linije

x_i _veličina studentske populacije (1000) za i-ti restoran



Slika 4.9 Grafikon rasipanja.

y_i označava opaženu (stvarnu) prodaju za restoran i i $\hat{y}_i$, predstavljajući procenjenu vrednost prodaje za restoran i, svaki restoran u uzorku će imati opaženu prodajnu vrednost od y_i i predviđenu prodajnu vrednost $\hat{y}_i$. Kako bi procenjena regresiona linija osigurala dobro uklapanje u podatke, želimo da razlike između opaženih prodajnih vrednosti i predviđenih prodajnih vrednosti budu što manje.

Metoda najmanjih kvadrata koristi uzorke podataka za dobijanje vrednosti b_0 i b_1.

Minimizirajte zbir kvadrata odstupanja između posmatranih vrednosti zavisne varijable y_i i predviđene vrednosti zavisne varijable $\hat{y}_i$. Polazna tačka za izračunavanje minimalnog zbira metodom najmanjih kvadrata data je izrazom

Kriterijum minimalnog iznosa: $\min \sum(y_i - \hat{y}_i)^2$

gde je

y_i=posmatrana vrednost zavisne varijable za i-to opažanje

$\hat{y}_i$ =predviđena vrednost zavisne varijable za i-to opažanje

Koeficijent smera regresione linije i početna vrednost:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

x_i_ vrednost nezavisne varijable za i-to opažanje

y_i_ vrednost zavisne varijable za i-to opažanje

$\bar{x}$ _ prosečna vrednost za nezavisnu varijablu

$\bar{y}$ _ prosečna vrednost za zavisnu varijablu

n_ukupan broj opažanja

Neki od proračuna potrebnih za izradu procenjene linije regresije najmanjih kvadrata prikazani su u nastavku. Na uzorku od 10 restorana imamo n=10 opažanja. Gornje jednačine prvo zahtevaju utvrđivanje srednje vrednosti x i prosečne vrednosti y.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Alternativna jednačina izračunava b_1:

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

Koristeći poslednje jednačine i informacije na slici 5.4, možemo izračunati usmereni koeficijent regresione linije za primer restorana Best Burger. Izračunavanje nagiba (b_1) je kako slijedi.

Slika 5.5 prikazuje dijagram ove jednačine na dijagramu rasipanja.

Nagib procenjene regresione jednačine ili koeficijent smera jednačine ($b_1 = 5$) je pozitivan.

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Slika 4.10 Prikaz jednačine na dijagramu rasipanja.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

Nakon toga sledi utvrđivanje početne vrednosti (b_0).

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

Ovako se procenjuje regresiona jednačina:

$$\hat{y} = 60 + 5x$$

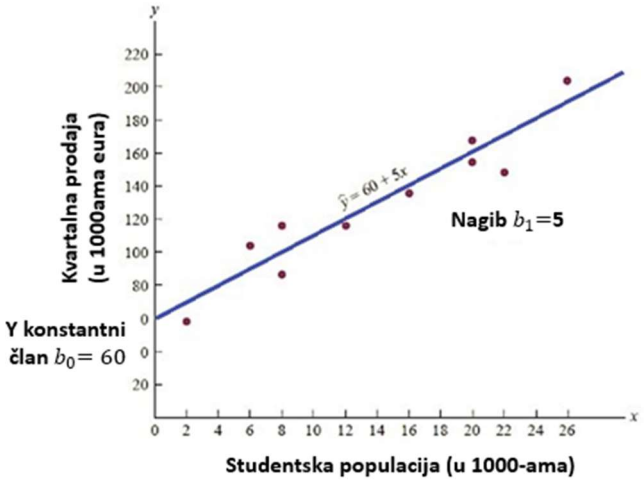
Slika prikazuje dijagram ove jednačine na dijagramu rasturanja.

Nagib procenjene regresione jednačine ($b_1 = 5$) je pozitivan, što znači da kako se broj studenata povećava, prodaja se povećava. Zapravo, možemo zaključiti (na osnovu izmerene prodaje u 1000-ama i studentske populacije u 1000-ama), što znači da je porast u studentskoj populaciji od 1000 povezan s povećanjem očekivane prodaje od 5000; tj. očekuje se povećanje tromesečne prodaje za 5 € po studentu.

Ako verujemo da regresiona jednačina, procenjena najmanjim kvadratima, adekvatno opisuje odnos između x i y, čini se razumnim koristiti procenjenu regresionu jednačinu za predviđanje vrednosti y za datu vrednost x. Na primer, ako želite predvideti tromesečnu prodaju za restoran koji se nalazi u blizini kampusa od 16.000 studenata, izračunali biste

$$\hat{y} = 60 + 5(16) = 140$$

Stoga bismo pretpostavili kvartalnu prodaju od 140.000 za ovaj restoran. U sledećim odjeljcima raspravljamo o metodama za procenu prikladnosti korišćenja procenjene regresione jednačine za procenu i predviđanje.



Slika 4.11Dijagram rasipanja studentske populacije i tromesečne prodaje.

5.5 Koeficijent determinacije

Za primer restorana Best Burger razvili smo procenjenu regresionu jednačinu: $y = 60 + 5x$ za približno linearni odnos između veličine studentske populacije x i tromesečne prodaje y. Sada je pitanje: koliko dobro procenjena regresiona jednačina odgovara podacima? U ovom odjeljku pokazujemo da koeficijent determinacije daje meru dobrog uklapanja za procenjenu regresionu jednačinu. Za i-to opažanje, razlika između opažene vrednosti zavisne varijable y_i i predviđene vrednosti zavisne varijable naziva se i-to rezidualno odstupanje.

Zbir kvadrata ovih rezidualnih odstupanja ili grešaka je vrednost koja je minimizirana metodom najmanjih kvadrata. Ova vrednost, takođe poznata kao zbir kvadrata reziduala, označava se sa SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

SSE vrednost je mera greške u korišćenju procenjene regresione jednačine za predviđanje vrednosti zavisne varijable u uzorku. Slika 5.6 prikazuje proračune potrebne za izračunavanje zbira kvadrata zbog greške za slučaj Best Burger.

Restoran i	x_i = Studentska populacija (u 1000-ama)	y_i = Kvartalna prodaja (u 1000ama eura)	Predviđena prodaja $\hat{y}_i = 60 + 5x_i$	Pogreška $y_i - \hat{y}_i$	Standardna pogreška $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

Slika 4.12 Kvadrati grešaka u slučaju Best Burger.

Pretpostavimo da se od nas traži da uradimo procenu tromesečne prodaje pri čemu ne znamo veličinu studentske populacije. Bez poznavanja bilo koje povezane varijable, koristili bismo prosek uzorka kao procenu tromesečne prodaje u bilo kom restoranu. Tabela na slici 5.6 pokazala je da je podatke o prodaji y_i=1300. Stoga je prosečna tromesečna vrednost prodaje za uzorak od 10 najboljih restorana s hamburgerima y_i/n = 1300/10 = 130. Na slici 5.7 prikazujemo zbir kvadrata odstupanja dobijenih korišćenjem srednje vrednosti uzorka od 130 za predviđanje vrednosti kvartalne prodaje za svaki restoran u uzorku. Za i-ti restoran u uzorku razlika y_i daje meru greške koja je uključena u aplikaciju za predviđanje prodaje. Odgovarajući zbir kvadrata, koji se naziva ukupan zbir kvadrata, označava se sa SST.

$$SST = \sum (y_i - \bar{y})^2$$

Restoran i	x_i = Studentska populacija (u 1000-ama)	y_i = Kvartalna prodaja (u 1000ama eura)	Devijacija $y_i - y$	Standardna devijacija $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

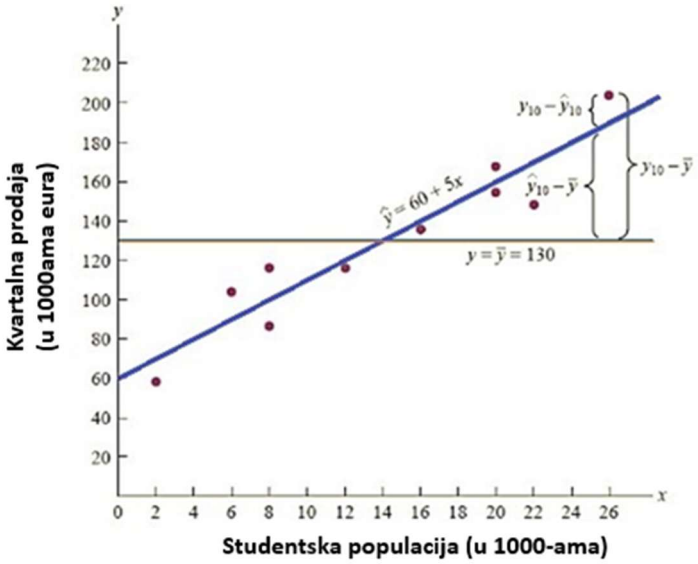
Slika 4.13 Zbir kvadrata odstupanja.

Zbir na dnu poslednje kolone na slici 5.7 je ukupni zbir kvadrata za BestBurgerove restorane SST = 15,730. Na slici 5.8 prikazujemo procenjenu regresionu liniju $y = 60 + 5x$ i liniju koja odgovara y = 130. Imajte na umu da se tačke grupišu bliže oko procenjene regresione linije

nego oko linije y = 130. Na primer, za 10. restoran u uzorku, vidi se da je greška puno veća kada se 130 koristi za predviđanje y = 10 nego kada se 130 koristi $y = 60 + 5x$ i iznosi 190. Možemo se setiti SST kao mere koliko dobro se opažanja grupišu oko linije i SSE kao mere koliko dobro se opažanja grupišu oko linije.

Da bi se izmerilo koliko vrednosti na procenjenoj regresionoj liniji odstupaju od sledećeg, izračunava se još jedan zbir kvadrata. Ovaj zbir kvadrata, koji se naziva regresionim zbirom kvadrata, označava se kao SSR.

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$



Slika 4.14 Regresiona linija u slučaju Best Burgera.

Iz prethodne rasprave, trebali bismo očekivati da su SST, SSR i SSE povezani. Zapravo, odnos između ta tri zbira kvadrata je jedan od najvažnijih rezultata u statistici.

5.6 Odnos između SST, SSR i SSE

$$SST = SSR + SSE$$

Gde je:

SST = ukupan zbir kvadrata

SSR = regresioni zbir kvadrata

SSE = rezidualni zbir kvadrata

Jednačina ($SST = SSR + SSE$) pokazuje da se ukupni zbir kvadrata može podeliti na dve komponente, regresioni zbir kvadrata i rezidualni zbir kvadrata. Dakle, ako su poznate vrednosti bilo koja od ova dva zbira kvadrata, treći zbir kvadrata može se lako izračunati. Na primer, u slučaju Best Burger restorana, već znamo da je SSE = 1530 i SST = 15,730; stoga, rešavanjem za SSR u gornjoj jednačini, nalazimo da je regresioni zbir kvadrata

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Sada ćemo videti kako možemo koristiti tri zbira kvadrata, SST, SSR i SSE, da bismo dobili kriterijum prilagođavanja za procenjenu regresionu jednačinu. Procenjena regresiona jednačina bi savršeno odgovarala ako bi svaka vrednost zavisne varijable $yi\_$ ležala slučajno na procenjenoj regresionoj liniji. U ovom slučaju to bi bilo nula za svako opažanje, što bi rezultiralo SSE =0. Budući da je SST = SSR + SSE, vidimo da za savršeno poklapanje SSR mora biti jednak SST i odnos (SSR/SST) mora biti jednak jedinici. Lošije prilagođavanje rezultiraće većim vrednostima za SSE. Rešavajući SSE u jednačini, vidimo da je SSE = SST - SSR. Stoga se najveća vrednost za SSE (a time i najlošije uklapanje) javlja kada je SSR = 0 i SSE = SST.

Za procenu se koristi odnos SSR/SST, koji ima vrednosti između nula i jedan koji odgovara procenjenoj regresionoj jednačini.

Taj odnos se naziva koeficijent determinacije i označava se s r^2.

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

Za primer restorana Best Burger, vrednost koeficijenta determinacije je

$$r^2 = \frac{SSR}{SST} = \frac{14200}{15730} = 0.9027$$

Kada se koeficijent determinacije izrazi kao procenat, r^2 se može tumačiti kao procenat ukupnog zbira kvadrata koji se može objasniti pomoću procenjene regresione jednačine. Za najbolje restorane s hamburgerima možemo zaključiti da se 90,27% ukupnog zbira kvadrata može objasniti korišćenjem procenjene regresione jednačine $y = 60 + 5x$ za predviđanje

kvartalne prodaje. Drugim riječima, 90,27% varijabilnosti u prodaji može se objasniti linearnim odnosom između veličine studentske populacije i prodaje. Trebalo bismo biti zadovoljni što vidimo da se tako dobro uklapa u procenjenu regresionu jednačinu.

5.7 Koeficijent korelacije

Koeficijent korelacije može se smatrati opisnom merom snage linearnog odnosa između dve varijable, x i y. Vrednosti koeficijenta korelacije su uvek između -1 i +1. Vrednost +1 znači da su $x\ i\ y$ dve varijable u savršenoj korelaciji u pozitivnom linearnom smislu. To znači da su svi podaci na liniji a da je prava s pozitivnim nagibom. Vrednost -1 znači da su $x\ i\ y$ savršeno povezane u negativnom linearnom smislu, sa svim podacima na pravoj liniji s negativnim nagibom. Vrednosti koeficijenta korelacije blizu nule znače da x i y nisu linearno povezani.

Ako je regresiona analiza već sprovedena i koeficijent determinacije r^2 je izračunat, koeficijent korelacije uzorka može se izračunati na sledeći način.

$$r_{xy} = (predznak\ b_1)\sqrt{koeficijent\ determinacije}$$

$$r_{xy} = (predznak\ b_1)\sqrt{r^2}$$

PEARSONOV KOEFICIJENT KORELACIJE: UZORCI PODATAKA

$$rxy = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

$$rxy = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

gde su:

$s_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$, $s_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, $s_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$

Predznak koeficijenta korelacije uzorka je pozitivan ako procenjena regresiona jednačina ima pozitivan nagib ($b_1 > 0$) i negativan ako procenjena regresiona jednačina ima negativan nagib ($b_1 < 0$).

Za slučaj Best Burger, vrednost koeficijenta determinacije koji odgovara procenjenoj regresionoj jednačini $y = 60 + 5x$ je 0,9027. Budući da je nagib procenjene regresione jednačine pozitivan, jednačina pokazuje da je koeficijent korelacije uzorka prema koeficijentu korelacije uzorka Rxy = 0,9501, zaključili bismo da postoji jaka pozitivna linearna veza između x i y.

U slučaju linearnog odnosa između dve varijable, oba koeficijenta determinacije i koeficijent korelacije uzorka daju meru jačine odnosa.

Koeficijent determinacije daje meru između 0 i 1, dok koeficijent korelacije uzorka daje meru između -1 i +1. Iako je koeficijent korelacije uzorka ograničen na linearni odnos između dve varijable, koeficijent determinacije može se primeniti na nelinearne odnose i na odnose koji imaju dve ili više nezavisnih varijabli. Dakle, koeficijent determinacije pruža širi raspon primenjivosti.

5.8 Model višestruke regresije

U sledećim odjeljcima nastavljamo naše proučavanje regresione analize razmatrajući situacije koje uključuju dve ili više nezavisnih varijabli. Ovo predmetno područje, koje se naziva višestruka regresiona analiza, omogućava nam da uzmemo u obzir više faktora i tako dobijemo bolja predviđanja nego što je to moguće s jednostavnom linearnom regresijom.

Višestruka regresiona analiza je proučavanje o tome kako je zavisna varijabla y povezana s dve ili više nezavisnih varijabli. U opštem slučaju, s p ćemo označiti broj nezavisnih varijabli.

5.9 Regresioni model i regresiona jednačina

Koncepti regresionog modela i regresione jednačine uvedeni u prethodnom odjeljku primenjuju se u slučaju višestruke regresije. Jednačina koja opisuje kako je zavisna varijabla y povezana s nezavisnim varijablama $x_1, x_2, \ldots, x_p$ i greškom naziva se modelom višestruke regresije. Počinjemo s pretpostavkom da model višestruke regresije ima sledeći oblik:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

U modelu višestruke regresije $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ su parametri, a greška (ϵ) je slučajna varijabla. Pažljivo ispitivanje ovog modela otkriva da je y linearna funkcija varijabli $x_1, x_2, \ldots, x_p$ plus

greška ϵ. Izraz greške uzima u obzir varijabilnost y koja se ne može objasniti linearnim učinkom p nezavisnih varijabli.

U odjeljku 5.10 raspravljamo o pretpostavkama za model višestruke regresije i epsilon. Jedna od pretpostavki je da je srednja ili očekivana vrednost (ϵ) nula. Implikacija ove pretpostavke je da je srednja ili očekivana vrednost y, označena s $E(y)$, jednaka $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. Jednačina koja opisuje kako je srednja vrednost y povezana s $x_1, x_2, \ldots, x_p$ naziva se jednačina višestruke regresije:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

5.10 Procenjena jednačina višestruke regresije

Ako su vrednosti $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ poznate, jednačina iz 5.9 se može koristiti za izračunavanje prosečne vrednosti y pri datim vrednostima $x_1, x_2, \ldots, x_p$. Nažalost, ove vrednosti parametara generalno neće biti poznate i moraju se proceniti iz podataka uzorka. Jednostavan slučajni uzorak koristi se za izračunavanje statistike uzorka $b_0, b_1, b_2, \ldots, b_p$ koja se koristi kao tačkasti procenitelj parametara $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Ovi uzorci statistike daju sledeću procenu jednačine višestruke regresije:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

gde su

$b_0, b_1, b_2, \ldots, b_p$ procjene $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

$\hat{y}$ = predviđena vrednost zavisne varijable.

Primer: Frigo transportna kompanija

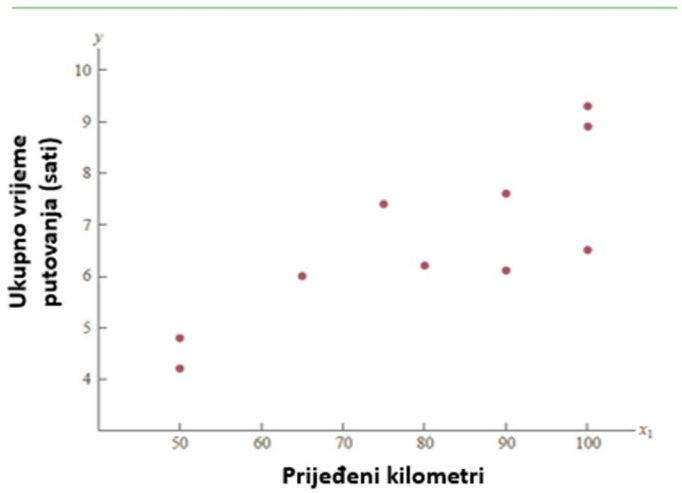
Kao ilustraciju višestruke regresione analize, razmotrićemo problem s kojim se susreće Frigo transportna kompanija, nezavisni autoprevoznik u južnoj Italiji. Najveći deo poslovanja kompanije Frigo odnosi se na dostavu na celom lokalnom području. Za bolji raspored rada, menadžeri žele planirati zajedničko dnevno vreme putovanja za svoje vozače.

U početku su menadžeri verovali da će ukupno dnevno vreme putovanja biti usko povezano s brojem pređenih kilometara u dnevnim isporukama. Jednostavan slučajni uzorak od 10 dodeljenih vozača dao je podatke prikazane na slici 5.9 i dijagramu rasipanja. Nakon

pregleda ovog dijagrama rasipanja, menadžeri su pretpostavili da se jednostavni linearni regresioni model može koristiti za opisivanje odnosa $y = \beta_0 + \beta_1 x_1 + \epsilon$ između ukupnog vremena putovanja (y) i broja pređenih km (x_1).

Vozački zadatak	x_1 = prijeđeni kilometri	y = vrijeme putovanja (sati)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

Slika 4.15 Podaci za primjer Frigo transportne kompanije.



Slika 4.16 Dijagram rasipanja za Frigo transportnu kompaniju.

Za procenu parametara β_0 i β_1 , jednačina najmanjih kvadrata korišćena je za izradu procenjene regresije.

$$\hat{y} = b_0 + b_1 x_1$$

Prethodna slika daje prikaz softverskog programa Minitab-a korišćenjem jednostavne linearne regresije na podatke u prethodnoj tabeli. Procenjena regresiona jednačina je:

$$\hat{y} = 1{,}27 + 0{,}0678x_1$$

Na nivou značajnosti od 0,05, F-vrednost od 15,81 i odgovarajuća p-vrednost od 0,004 pokazuju da je odnos značajan. To znači da možemo odbiti $H_0: \beta_1 = 0$ jer je p-vrednost manja od $\alpha = 0{,}05$. Imajte na umu da isti zaključak proizlazi iz vrednosti $t = 3{,}98$ i pridružene p-vrednosti od 0,004. Stoga možemo zaključiti da je odnos između ukupnog vremena putovanja i broja pređenih kilometara značajan. Duže vreme putovanja povezano je s više pređenih kilometara. S koeficijentom determinacije (izraženim u procentima) $R - Sq = 66{,}4\ \%$, vidimo da se 66,4% varijabilnosti vremena putovanja može objasniti linearnim učinkom broja pređenih kilometara.

Ovo otkriće je prilično dobro, ali menadžeri bi mogli razmotriti dodavanje druge nezavisne varijable kako bi objasnili neke od preostalih varijabli u zavisnoj varijabli.

MINITAB OUTPUT FOR FRIGO TRUCKING WITH ONE INDEPENDENT VARIABLE

```
The regression equation is
Time = 1.27 + 0.0678 kM

Predictor      Coef  SE Coef     T      p
Constant      1.274    1.401  0.91  0.390
kM          0.06783  0.01706  3.98  0.004

S = 1.00179   R-Sq = 66.4%   R-Sq(adj) = 62.2%

Analysis of Variance

SOURCE           DF      SS      MS      F      p
Regression        1  15.871  15.871  15.81  0.004
Residual Error    8   8.029   1.004
Total             9  23.900
```

Slika 4.17 Rezultati s jednom nezavisnom varijablom.

Kada su pokušali identifikovati drugu nezavisnu varijablu, menadžeri su smatrali da broj dostava takođe može uticati na ukupno vreme putovanja. Podaci Frigo transportne kompanije, s dodanim brojem dostava, prikazani su na slici 5.12. - (x_1) i broj isporuka (x_2), kao nezavisne varijable. Procenjena regresiona jednačina je:

$$\hat{y} = -0.869 + 0.0611x_1 + 0.923x_2$$

PODACI ZA FRIGO TRANSPORTNU TVRTKU S PRIJEĐENIM KILOMETRIMA (x_1) I BROJEM DOSTAVA (x_2) KAO NEZAVISNIM VARIJABLAMA

Vozački zadatak	x_1 = prijeđeni kilometri	x_2 = broj dostava	y = vrijeme putovanja (sati)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Slika 4.18 Podaci Frigo transportnekompanije i nezavisne varijable.

Pogledajmo pažljivije vrednosti b_1= 0,0611 i b_2= 0,923 u poslednjoj jednačini.

Napomena o tumačenju koeficijenata

U ovom trenutku možemo dati jedan komentar o odnosu između procenjene regresione jednačine sa samo pređenim kilometrima kao nezavisnom varijablom i jednačine koja uključuje broj isporuka kao drugu nezavisnu varijablu. Vrednost b_1 nije ista u oba slučaja. U jednostavnoj linearnoj regresiji tumačimo b_1 kao procenu promene y za jednu jediničnu promenu nezavisne varijable. U višestrukoj regresionoj analizi ovo tumačenje treba malo modifikovati. To jest, u višestrukoj regresionoj analizi, svaki regresioni koeficijent se tumači na sledeći način: predstavljao bi procenu promene u y koja odgovara promeni u x_i za jednu jedinicu kada se sve ostale nezavisne varijable konstantne.

U slučaju Frigo transportne kompanije, to uključuje dve nezavisne varijable, b_1 = 0,0611 i b_2 = 0,923.

MINITAB OUTPUT FOR FRIGO TRUCKING WITH TWO INDEPENDENT VARIABLES

The regression equation is
Time = - 0.869 + 0.0611 kM + 0.923 Deliveries

Predictor	Coef	SE Coef	T	p
Constant	-0.8687	0.9515	-0.91	0.392
kM	0.061135	0.009888	6.18	0.000
Deliveries	0.9234	0.2211	4.18	0.004

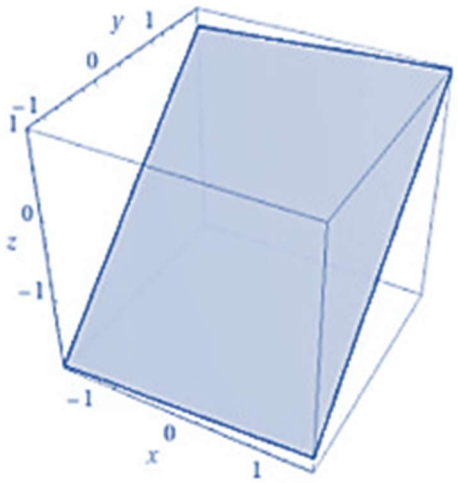
S = 0.573142 R-Sq = 90.4% R-Sq(adj) = 87.6%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	21.601	10.800	32.88	0.000
Residual Error	7	2.299	0.328		
Total	9	23.900			

Slika 4.19 Rezultati za Frigo transportnu kompaniju s dve nezavisne varijable.

Stoga je 0,0611 sati procena očekivanog povećanja vremena putovanja koje odgovara povećanju od jednog kilometra pređene udaljenosti kada je broj dostava konstantan. Slično, budući da je b_2 = 0,923, procena očekivanog povećanja vremena putovanja koje odgovara povećanju jedne isporuke kada je broj pređenih kilometara konstantan iznosi 0,923 sata.



Slika 4.20 Vizualni prikaz rezultata za Frigo transportnu kompaniju.

Literatura 5. poglavlja

- *Introductory Statistics.* Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®
- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved