5. Linearna regresija s jednom i više regresorskih varijabli

Odluke menadžmenta često se temelje na odnosu između dvije ili više varijabli. Na primjer, voditelj marketinga može pokušati predvidjeti prodaju na određenoj razini troškova oglašavanja nakon ispitivanja odnosa između tih izdataka i prodaje.

U drugom slučaju, javno poduzeće može koristiti omjer između maksimalne dnevne vrijednosti temperature i potrebe za električnom energijom za predviđanje potrošnje električne energije. Ponekad se menadžer oslanja na intuiciju. Intuitivno prosuđuje kako su dvije varijable povezane. Međutim, ako je moguće dobiti podatke, ima smisla koristiti statistički postupak koji se zove regresijska analiza kako bi se pokazalo kako su te dvije varijable povezane jedna s drugom.

U terminologiji regresije, predviđena varijabla se naziva zavisna varijabla.

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

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

U ovom ćemo odjeljku pogledati najjednostavniju vrstu regresijske analize koja uključuje jednu nezavisnu varijablu i jednu zavisnu varijablu. Odnos između dviju varijabli bit će aproksimiran ravnom linijom. Naziva se jednostavnom linearnom regresijom. Regresijska analiza koja uključuje dvije ili više nezavisne varijable naziva se višestruka regresijska analiza.

5.1 Jednostavni linearni regresijski model

Best Burger je lanac restorana brze hrane koji se nalazi u području s više država. Najbolje lokacije Burgera nalaze se u blizini sveučilišnih kampusa. Menadžeri vjeruju da je tromjesečna prodaja ovih restorana (označeno s y) u pozitivnoj korelaciji s veličinom studentske populacije (označeno s x). Restorani u blizini kampusa s velikim brojem studenata obično generiraju veću prodaju od onih u blizini kampusa s malim brojem studenata. Pomoću regresijske analize možemo razviti jednadžbu koja pokazuje kako je y zavisna varijabla povezana s nezavisnom varijablom x.

5.2 Regresijski model i regresijska jednadžba

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

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

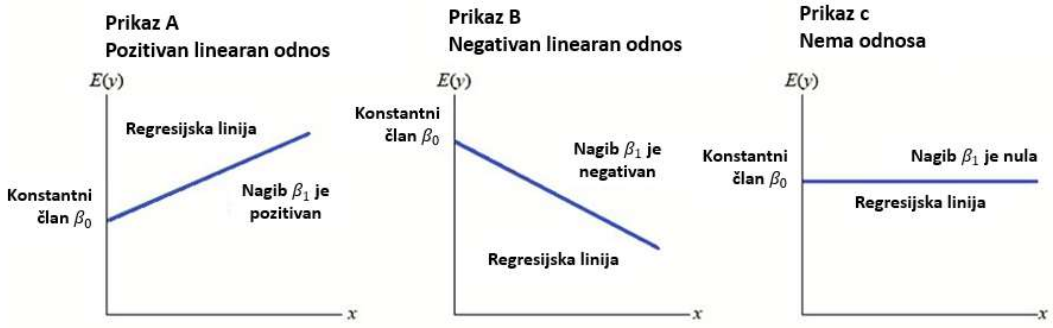
β_0 i β_1 nazivaju se parametri modela, ϵ (grčko slovo epsilon) je slučajna varijabla koja se naziva pogreška modela. Greška predstavlja varijabilnost y što se ne može objasniti linearnim odnosom između x i y.

Populacija svih restorana Best Burger također se može promatrati kao zbirka podpopulacija, jedna za svaku zasebnu vrijednost x. Na primjer, jednu podpopulaciju čine svi restorani Best Burger u blizini sveučilišnih kampusa s 8000 studenata. Drugu podpopulaciju čine svi restorani Best Burger koji se nalaze u blizini sveučilišnih kampusa s 9000 studenata i tako dalje. Svaka subpopulacija ima odgovarajuću raspodjelu vrijednosti y. Svaka raspodjela vrijednosti y ima svoju srednju ili očekivanu vrijednost. Jednadžba koja opisuje što je očekivana vrijednost y, označena s $E(y)$, čime je povezana x naziva se regresijska jednadžba. Regresijska jednadžba za jednostavnu linearnu regresiju je sljedeća

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

Grafikon jednostavne jednadžbe linearne regresije je ravna linija. β_0 predstavlja početnu vrijednost regresijske linije, β_1 je koeficijent smjera linije i $E(y)$ srednju vrijednost ili očekivanu vrijednost y za danu vrijednost x.

Primjeri mogućih regresijskih linija prikazani su na slici 5.1 u nastavku. Regresijska y linija u slučaju A pokazuje da je vrijednost y u pozitivnoj korelaciji s x. Kako se vrijednosti povećavaju x, vrijednosti se također $E(y)$ povećavaju. Tamo gdje su manje vrijednosti $E(y)$ povezane su s višim vrijednostima x. Regresijska linija na prikazu C prikazuje slučaj u kojem vrijednost y nije povezana s x. To znači da je vrijednost y ista za svaku vrijednost x.



Slika 4.7Primjeri grafikona linearnog odnosa.

5.3 Procijenjena regresijska jednadžba

Kad bi bile poznate vrijednosti parametara populacije β_0 i β_1, mogli bismo koristiti gornju jednadžbu za izračunavanje vrijednosti y za zadanu vrijednost x. U praksi je tim parametrima teško pristupiti, pa se jednostavno procjenjuju korištenjem podataka uzorka. Statistika uzorka (označena s b_0 i b_1) izračunata je kao procjene parametara populacije β_0 i β_1. Zamjenom vrijednosti statistike uzorka b_0 i b_1 umjesto β_0 i β_1 u regresijskoj jednadžbi dobivamo novu, procijenjenu regresijsku jednadžbu. Procijenjena regresijska jednadžba za jednostavnu linearnu regresiju je sljedeća

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

Grafikon procijenjene jednostavne linearne regresije naziva se procijenjena regresijska linija. b_0 predstavlja početnu vrijednost regresijske linije, b_1 je koeficijent smjera linije.

U nastavku ćemo pokazati kako koristiti metodu najmanjih kvadrata za izračunavanje vrijednosti b_0 i b_1 u procijenjenoj regresijskoj jednadžbi.

Općenito je $\hat{y}$ (rezultat za $E(y)$) prosječna vrijednost y za danu vrijednost x. Ako sada želimo procijeniti očekivanu vrijednost tromjesečne prodaje za sve restorane Best Burger koji se nalaze u blizini kampusa s 10 000 studenata, vrijednost bi x bila zamijenjena vrijednošću 10

000 u posljednjoj jednadžbi. U nekim slučajevima, međutim, možda ćemo biti više zainteresirani za predviđanje prodaje samo za jedan određeni restoran, na primjer, pretpostavimo da želite predvidjeti 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 vrijednosti y za danu x vrijednost $\hat{y}$.

5.4 Metoda najmanjih kvadrata

Metoda najmanjih kvadrata je postupak u kojem se pomoću uzoraka podataka nalazi jednadžba procijenjene regresijske linije. Kako bismo ilustrirali metodu najmanjih kvadrata, pretpostavimo da su podaci prikupljeni iz uzorka od 10 restorana s najboljim hamburgerima u blizini sveučilišnih kampusa. Sa x_i će označavati veličinu studentske populacije (u tisućama) i veličinu y_i tromjesečne prodaje (u tisućama EUR). x_i *i* y_i za 10 uzoraka restorana sažeti su u tablici u nastavku. Vidimo da je restoran 1, z 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 vrijednošću je restoran 10, koji se nalazi u blizini kampusa s 26.000 studenata i ima kvartalnu prodaju od 202.000 €.

Slijedi dijagram raspršenosti podataka na slici 5.2 u nastavku. Studentska populacija prikazana je na vodoravnoj osi, a kvartalna prodaja na okomitoj osi. Dijagrami raspršenosti za regresijsku analizu konstruirani su s nezavisnom varijablom x na vodoravnoj osi i ovisnom varijablom y na okomitoj osi. Dijagram raspršenosti nam stoga omogućuje 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 raspršenosti podataka.

Koji se preliminarni zaključci mogu izvući iz slike ispod 5.3? Veća tromjesečna prodaja događa se u kampusima s većom populacijom studenata. Osim toga, postoji konstantan odnos između veličine studentske populacije i tromjesečne prodaje, koji se može opisati pravocrtno. Između x i y se doista podrazumijeva pozitivan linearni odnos. Stoga smo odabrali jednostavan linearni regresijski model za prikaz odnosa između tromjesečne prodaje i studentske populacije. S obzirom na ovaj izbor, naš sljedeći zadatak je koristiti tablicu podataka uzorka za određivanje vrijednosti b_0 i b_1, koji su važni parametri u procjeni jednostavne jednadžbe linearne regresije. Za i-ti restoran procijenjena regresijska jednadžba je

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

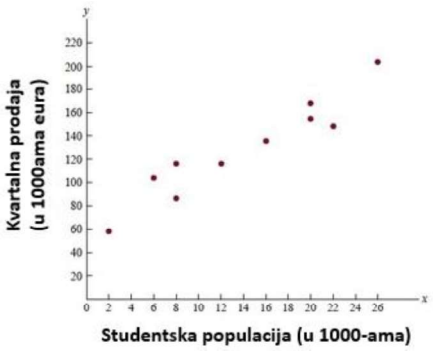
gdje je

$\hat{y}_i$_ procijenjena vrijednost tromjesečne prodaje (1000 €) za i-ti restoran

b_0_ početna vrijednost procijenjene regresijske linije

b_1_ koeficijent smjera procijenjene regresijske linije

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



Slika 4.9 Raspršeni grafikon.

y_i označava opaženu (stvarnu) prodaju za restoran i i $\hat{y}_i$, predstavljajući procijenjenu vrijednost prodaje za restoran i, svaki će restoran u uzorku imati opaženu prodajnu vrijednost od y_i i predviđenu prodajnu vrijednost $\hat{y}_i$. Kako bi procijenjena regresijska linija osigurala dobro

uklapanje u podatke, želimo da razlike između opaženih prodajnih vrijednosti i predviđenih prodajnih vrijednosti budu što manje.

Metoda najmanjih kvadrata koristi uzorke podataka za pružanje vrijednosti b_0 i b_1.

Minimizirajte zbroj kvadrata odstupanja između promatranih vrijednosti zavisne varijable y_i i predviđene vrijednosti zavisne varijable $\hat{y}_i$. Polazna točka za izračunavanje minimalnog zbroja metodom najmanjih kvadrata dana je izrazom

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

gdje je

y_i=promatrana vrijednost zavisne varijable za i-to opažanje

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

Koeficijent smjera regresijske linije i početna vrijednost:

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

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

x_i_ vrijednost nezavisne varijable za i-to opažanje

y_i_ vrijednost zavisne varijable za i-to opažanje

$\bar{x}$ _ prosječna vrijednost za nezavisnu varijablu

$\bar{y}$ _ prosječna vrijednost za zavisnu varijablu

n_ukupan broj opažanja

Neki od izračuna potrebnih za izradu procijenjene linije regresije najmanjih kvadrata prikazani su u nastavku. Na uzorku od 10 restorana imamo n=10 opažanja. Gornje jednadžbe prvo zahtijevaju izračun srednje vrijednosti x i prosječne vrijednosti 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 jednadžba izračuna 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 posljednje jednadžbe i informacije na slici 5.4, možemo izračunati usmjereni koeficijent regresijske linije za primjer restorana Best Burger. Izračunavanje nagiba (b_1) je kako slijedi.

Slika 5.5 prikazuje dijagram ove jednadžbe na raspršenom dijagramu.

Nagib procijenjene regresijske jednadžbe ili koeficijent smjera jednadžbe ($b_1 = 5$) 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 jednadžbe na dijagramu raspršenosti.

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

Nakon toga slijedi izračun početne vrijednosti (b_0).

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

Ovako se procjenjuje regresijska jednadžba:

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

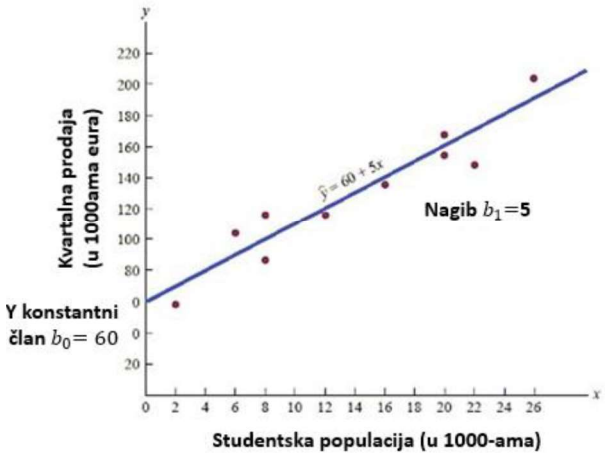
Slika prikazuje dijagram ove jednadžbe na dijagramu raspršenosti.

Nagib procijenjene regresijske jednadžbe ($b_1 = 5$) je pozitivan, što znači da kako se broj studenata povećava, prodaja se povećava. Zapravo, možemo zaključiti (na temelju izmjerene 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 tromjesečne prodaje za 5 € po studentu.

Ako vjerujemo da regresijska jednadžba, procijenjena najmanjim kvadratima, adekvatno opisuje odnos između x i y, čini se razumnim koristiti procijenjenu regresijsku jednadžbu za predviđanje vrijednosti y za danu vrijednost x. Na primjer, ako želite predvidjeti tromjesečnu 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 sljedećim odjeljcima raspravljamo o metodama za procjenu prikladnosti korištenja procijenjene regresijske jednadžbe za procjenu i predviđanje.



Slika 4.11Dijagram raspršenosti studentske populacije i tromjesečne prodaje.

5.5 Koeficijent determinacije

Za primjer restorana Best Burger razvili smo procijenjenu regresijsku jednadžbu $y = 60 + 5x$ za približno linearni odnos između veličine studentske populacije x i tromjesečne prodaje y. Sada je pitanje: koliko dobro procijenjena regresijska jednadžba odgovara podacima? U ovom odjeljku pokazujemo da koeficijent determinacije daje mjeru dobrog uklapanja za procijenjenu regresijsku jednadžbu. Za i-to opažanje, razlika između opažene vrijednosti zavisne varijable y_i i predviđene vrijednosti zavisne varijable naziva se i-ti rezidualno odstupanje.

Zbroj kvadrata ovih rezidualnih odstupanja ili pogrešaka je količina koja je minimizirana metodom najmanjih kvadrata. Ova količina, također poznata kao rezidualni zbroj kvadrata, označava se sa SSE.

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

SSE vrijednost je mjera pogreške u korištenju procijenjene regresijske jednadžbe za predviđanje vrijednosti zavisne varijable u uzorku. Slika 5.6 prikazuje izračune potrebne za izračunavanje zbroja kvadrata zbog pogreške 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 pogrešaka u slučaju Best Burger.

Pretpostavimo da se od nas traži da napravimo procjenu tromjesečne prodaje bez da znamo veličinu studentske populacije. Bez poznavanja bilo koje povezane varijable, koristili bismo prosjek uzorka kao procjenu tromjesečne prodaje u bilo kojem restoranu. Tablica na slici 5.6 pokazala je da je podatke o prodaji y_i=1300. Stoga je prosječna tromjesečna vrijednost prodaje za uzorak od 10 najboljih restorana s hamburgerima y_i/n = 1300/10 = 130. Na slici 5.7 prikazujemo zbroj kvadrata odstupanja dobivenih korištenjem srednje vrijednosti uzorka od 130 za predviđanje vrijednosti kvartalne prodaje za svaki restoran u uzorku. Za i-ti restoran u uzorku razlika y_i daje mjeru pogreške koja je uključena u aplikaciju za predviđanje prodaje. Odgovarajući zbroj kvadrata, koji se naziva ukupan zbroj kvadrata, označava se sa SST.

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

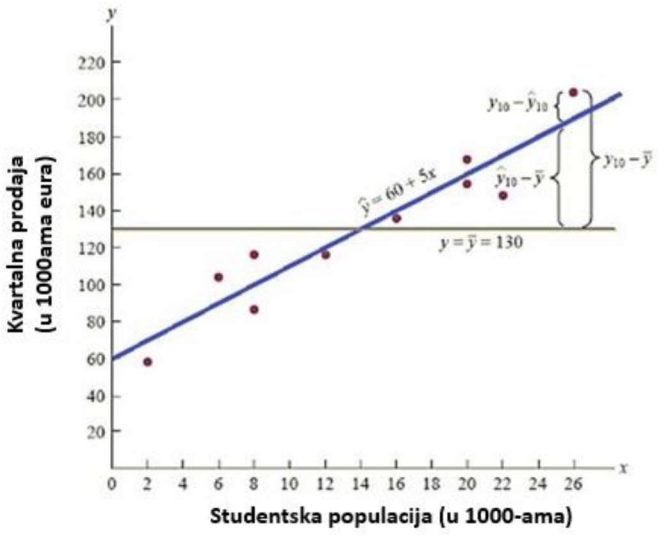
Restoran i	x_i = Studentska populacija (u 1000-ama)	y_i = Kvartalna prodaja (u 1000ama eura)	Devijacija $y_i - y$	Standardna devijacija $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10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

Slika 4.13 Zbroj kvadrata odstupanja.

Zbroj na dnu zadnjeg stupca na slici 5.7 je ukupni zbroj kvadrata za BestBurgerove restorane SST = 15,730. Na slici 5.8 prikazujemo procijenjenu regresijsku liniju $y = 60 + 5x$ i liniju koja odgovara y = 130. Imajte na umu da se točke grupiraju bliže oko procijenjene regresijske linije nego oko linije y = 130. Na primjer, za 10. restoran u uzorku, vidi se da je pogreška puno veća kada se 130 koristi za predviđanje y = 10 nego kada se 130 koristi $y = 60 + 5x$ i iznosi 190. Možemo se sjetiti SST kao mjere koliko dobro se opažanja grupiraju oko linije i SSE kao mjere koliko dobro se opažanja grupiraju oko linije.

Da bi se izmjerilo koliko vrijednosti na procijenjenoj regresijskoj liniji odstupaju od sljedećeg, izračunava se još jedan zbroj kvadrata. Ovaj zbroj kvadrata, koji se naziva regresijskim zbrojem kvadrata, označava se kao SSR.

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



Slika 4.14 Regresijska linija u slučaju Best Burgera.

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

5.6 Odnos između SST, SSR i SSE

$$SST = SSR + SSE$$

Gdje je:

SST = ukupan zbroj kvadrata

SSR = regresijski zbroj kvadrata

SSE = rezidualni zbroj kvadrata

Jednadžba ($SST = SSR + SSE$) pokazuje da se ukupni zbroj kvadrata može podijeliti na dvije komponente, regresijski zbroj kvadrata i rezidualni zbroj kvadrata. Dakle, ako su poznate vrijednosti bilo koja dva od ovih zbroja kvadrata, treći zbroj kvadrata može se lako saznati izračunom. Na primjer, u slučaju Best Burger restorana, već znamo da je SSE = 1530 i SST = 15,730; stoga, rješavanjem za SSR u gornjoj jednadžbi, nalazimo da je regresijski zbroj kvadrata

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

Sada ćemo vidjeti kako možemo koristiti tri zbroja kvadrata, SST, SSR i SSE, da bismo dobili kriterij prilagodbe za procijenjenu regresijsku jednadžbu. Procijenjena regresijska jednadžba bi savršeno odgovarala ako bi svaka vrijednost zavisne varijable $yi\_$ ležala nasumično na procijenjenoj regresijskoj 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 pristajanje SSR mora biti jednak SST i omjer (SSR/SST) mora biti jednak jedan. Lošija prilagodba rezultirat će većim vrijednostima za SSE. Rješavajući SSE u jednadžbi, vidimo da je SSE = SST - SSR. Stoga se najveća vrijednost za SSE (a time i najlošije uklapanje) javlja kada je SSR = 0 i SSE = SST.

Za procjenu se koristi omjer SSR/SST, koji ima vrijednosti između nula i jedan koji odgovara procijenjenoj regresijskoj jednadžbi.

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

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

Za primjer restorana Best Burger, vrijednost koeficijenta determinacije je

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

Kada se koeficijent determinacije izrazi kao postotak, r^2 se može tumačiti kao postotak ukupnog zbroja kvadrata koji se može objasniti pomoću procijenjene regresijske jednadžbe. Za najbolje restorane s hamburgerima možemo zaključiti da se 90,27% ukupnog zbroja kvadrata može objasniti korištenjem procijenjene regresijske jednadžbe $y = 60 + 5x$ 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 procijenjenu regresijsku jednadžbu.

5.7 Koeficijent korelacije

Koeficijent korelacije može se smatrati opisnom mjerom snage linearnog odnosa između dviju varijabli, x i y. Vrijednosti koeficijenta korelacije su uvijek između -1 i +1. Vrijednost +1 znači da su $x\ i\ y$ dvije varijable u savršenoj korelaciji u pozitivnom linearnom smislu. To znači da su sve podatkovne točke na a ravna linija s pozitivnim nagibom. Vrijednost -1 znači da su $x\ i\ y$ savršeno povezane u negativnom linearnom smislu, sa svim podatkovnim točkama na ravnoj liniji s negativnim nagibom. Vrijednosti koeficijenta korelacije blizu nule znače da x i y nisu linearno povezani.

Ako je regresijska analiza već provedena i koeficijent determinacije r^2 je izračunat, koeficijent korelacije uzorka može se izračunati na sljedeći način.

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

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

PEARSONOV KOEFICIJENT KORELACIJE: UZORCI PODATAKA

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

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

gdje 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 procijenjena regresijska jednadžba ima pozitivan nagib ($b_1 > 0$) i negativan ako procijenjena regresijska jednadžba ima negativan nagib ($b_1 < 0$).

Za slučaj Best Burger, vrijednost koeficijenta determinacije koji odgovara procijenjenoj regresijskoj jednadžbi $y = 60 + 5x$ je 0,9027. Budući da je nagib procijenjene regresijske jednadžbe pozitivan, jednadžba 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 dviju varijabli, oba koeficijenta determinacije i koeficijent korelacije uzorka daju mjeru snage odnosa.

Koeficijent determinacije daje mjeru između nula i jedan, dok koeficijent korelacije uzorka daje mjeru između -1 i +1. Iako je koeficijent korelacije uzorka ograničen na linearni odnos između dviju varijabli, koeficijent determinacije može se primijeniti na nelinearne odnose i na odnose koji imaju dvije ili više nezavisnih varijabli. Dakle, koeficijent determinacije pruža širi raspon primjenjivosti.

5.8 Model višestruke regresije

U sljedećim odjeljcima nastavljamo naše proučavanje regresijske analize razmatrajući situacije koje uključuju dvije ili više neovisnih varijabli. Ovo predmetno područje, koje se naziva višestruka regresijska analiza, omogućuje nam da uzmemo u obzir više faktora i tako dobijemo bolja predviđanja nego što je to moguće s jednostavnom linearnom regresijom.

Višestruka regresijska analiza je studija o tome kako je zavisna varijabla y povezana s dvije ili više nezavisnih varijabli. U općem slučaju, s p ćemo označiti broj nezavisnih varijabli.

5.9 Regresijski model i regresijska jednadžba

Koncepti regresijskog modela i regresijske jednadžbe uvedeni u prethodnom odjeljku primjenjuju se u slučaju višestruke regresije. Jednadžba koja opisuje kako je zavisna varijabla

y povezana s nezavisnim varijablama $x_1, x_2, \ldots, x_p$ i pogreškom naziva se modelom višestruke regresije. Počinjemo s pretpostavkom da model višestruke regresije ima sljedeći 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 pogreška (ϵ) je slučajna varijabla. Pomno ispitivanje ovog modela otkriva da je y linearna funkcija varijabli $x_1, x_2, \ldots, x_p$ plus pogreška ϵ. Izraz pogreške 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 vrijednost (ϵ) nula. Implikacija ove pretpostavke je da je srednja ili očekivana vrijednost y, označena s $E(y)$, jednaka $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. Jednadžba koja opisuje kako je srednja vrijednost y povezana s $x_1, x_2, \ldots, x_p$ naziva se jednadžba višestruke regresije.

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

5.10 Procijenjena jednadžba višestruke regresije

Ako su vrijednosti $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ poznate, jednadžba iz 5.9 se može koristiti za izračunavanje prosječne vrijednosti y pri danim vrijednostima $x_1, x_2, \ldots, x_p$. Nažalost, ove vrijednosti parametara općenito neće biti poznate i moraju se procijeniti iz podataka uzorka. Jednostavan slučajni uzorak koristi se za izračun statistike uzorka $b_0, b_1, b_2, \ldots, b_p$ koja se koristi kao točkasti procjenitelj parametara $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Ovi uzorci statistike daju sljedeću procjenu jednadžbe višestruke regresije:

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

gdje su

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

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

Primjer: Frigo transportna tvrtka

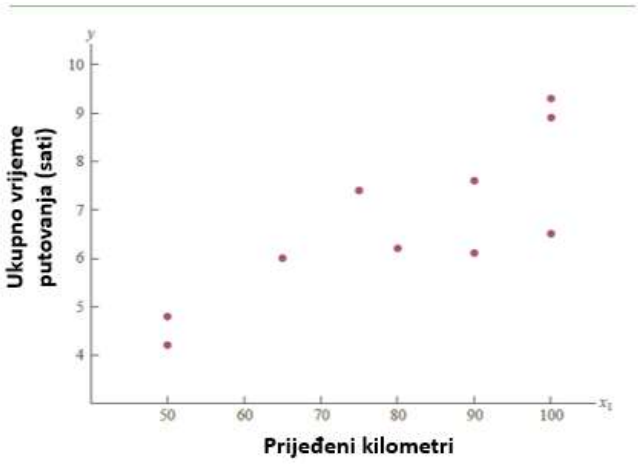
Kao ilustraciju višestruke regresijske analize, razmotrit ćemo problem s kojim se susreće Frigo transportna tvrtka, neovisna autoprijevoznička tvrtka u južnoj Italiji. Najveći dio poslovanja

tvrtke Frigo odnosi se na dostavu na cijelom lokalnom području. Za bolje rasporede rada, menadžeri žele planirati zajedničko dnevno vrijeme putovanja za svoje vozače.

U početku su menadžeri vjerovali da će ukupno dnevno vrijeme putovanja biti usko povezano s brojem prijeđenih kilometara u dnevnim isporukama. Jednostavan nasumični uzorak od 10 dodijeljenih vozača dao je podatke prikazane na slici 5.9 i dijagramu raspršenosti. Nakon pregleda ovog dijagrama raspršenosti, upravitelji su pretpostavili da se jednostavni linearni regresijski model može koristiti za opisivanje odnosa $y = \beta_0 + \beta_1 x_1 + \epsilon$ između ukupnog vremena putovanja (y) i broja prijeđenih 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 tvrtke.



Slika 4.16 Dijagram raspršenosti za Frigo transportnu tvrtku.

Za procjenu parametara β_0 i β_1, jednadžba najmanjih kvadrata korištena je za izradu procijenjene regresije.

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

Gornja slika prikazuje prikaz softverskog programa Minitab-a korištenjem jednostavne linearne regresije na podatke u gornjoj tablici. Procijenjena regresijska jednadžba je

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

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

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 identificirati drugu nezavisnu varijablu, menadžeri su smatrali da broj dostava također može pridonijeti ukupnom vremenu putovanja. Podaci Frigo transportne tvrtke, s

dodanim brojem dostava, prikazani su na slici 5.12. - (x_1) i broj isporuka (x_2), kao nezavisne varijable. Procijenjena regresijska jednadžba 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 transportne tvrtke i nezavisne varijable.

Pogledajmo pobliže vrijednosti b_1= 0,0611 i b_2= 0,923 u posljednjoj jednadžbi.

Napomena o tumačenju koeficijenata

U ovom trenutku možemo dati jedan komentar o odnosu između procijenjene regresijske jednadžbe sa samo prijeđenim miljama kao nezavisnom varijablom i jednadžbe koja uključuje broj isporuka kao drugu nezavisnu varijablu. Vrijednost b_1 nije ista u oba slučaja. U jednostavnoj linearnoj regresiji tumačimo b_1 kao procjenu promjene y za jednu jediničnu promjenu nezavisne varijable. U višestrukoj regresijskoj analizi ovo tumačenje treba malo modificirati. To jest, u višestrukoj regresijskoj analizi, svaki regresijski koeficijent se tumači na sljedeći način: predstavljao bi procjenu promjene u koja y odgovara promjeni u x_i za jednu jedinicu kada se sve ostale nezavisne varijable drže konstantnima.

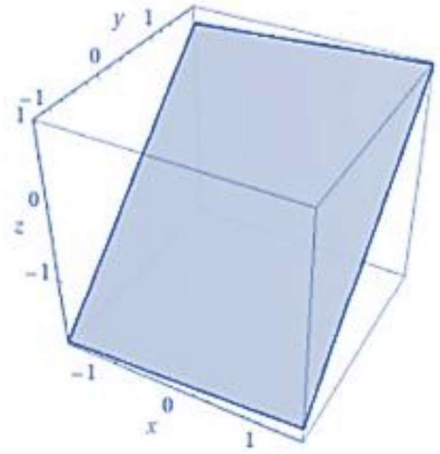
U slučaju Frigo transportne tvrtke, to uključuje dvije nezavisne varijable, b_1 = 0,0611 i b_2 = 0,923.

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 tvrtku s dvije nezavisne varijable.

Stoga je 0,0611 sati procjena očekivanog povećanja vremena putovanja koja odgovara povećanju od jedne milje po prijeđenoj udaljenosti kada je broj dostava konstantan. Slično, budući da je b_2 = 0,923, procjena očekivanog povećanja vremena putovanja koja odgovara povećanju jedne isporuke kada je broj prijeđenih milja konstantan iznosi 0,923 sata.



Slika 4.20 Vizualni prikaz rezultata za Frigo transportnu tvrtku.

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