



5. Linearna regresija z enim in več regresorji

Upravljaljske odločitve pogosto temeljijo na razmerju med dvema ali več spremenljivkami. Vodja trženja lahko na primer poskuša napovedati prodajo pri določeni ravni izdatkov za oglaševanje, potem ko preuči razmerje med temi izdatki in prodajo.

V drugem primeru lahko javno podjetje za napovedovanje porabe električne energije uporabi razmerje med najvišjo dnevno temperaturo in povpraševanjem po električni energiji. Včasih se upravljavec zanaša na intuicijo. Intuitivno oceni, kako sta spremenljivki povezani. Če pa je mogoče pridobiti podatke, je smiselno uporabiti statistični postopek, imenovan regresijska analiza, ki pokaže, kako sta spremenljivki povezani med seboj.

V regresijski terminologiji se napovedana spremenljivka imenuje odvisna spremenljivka.

Spremenljivka ali spremenljivke, ki se uporabljajo za napovedovanje vrednosti odvisne spremenljivke, se imenujejo neodvisne spremenljivke.

Pri analizi učinka izdatkov za oglaševanje na prodajo bi bila torej prodaja odvisna spremenljivka. Neodvisna spremenljivka so izdatki za oglaševanje. V statističnem zapisu y označuje odvisno spremenljivko, in x označuje neodvisno spremenljivko.

V tem razdelku si bomo ogledali najpreprostejšo vrsto regresijske analize, ki vključuje eno neodvisno spremenljivko in eno odvisno spremenljivko. Razmerje med spremenljivkama se približa s premico. To se imenuje preprosta linearna regresija. Regresijska analiza, ki vključuje dve ali več neodvisnih spremenljivk, se imenuje multipla regresijska analiza.

5.1 Enostavni linearni regresijski model

Best Burger je veriga restavracij s hitro prehrano na območju več držav. Lokacije Best Burger so v bližini univerzitetnih kampusov. Poslovodje menijo, da četrtna prodaja teh restavracij (označena z y) je pozitivno povezana z velikostjo študentske populacije (označeno z x). Restavracije v bližini študentskih kampusov z velikim številom študentov običajno ustvarijo večjo prodajo kot restavracije v bližini kampusov z





majhnim številom študentov. Z regresijsko analizo lahko razvijemo enačbo, ki pokaže, kako odvisna spremenljivka y povezana z neodvisno spremenljivko x .

5.2 Regresijski model in regresijska enačba

V primeru restavracije Best Burger so prebivalstvo vse restavracije Best Burger. Za vsako restavracijo v populaciji obstaja vrednost x (študentska populacija) in ustrezna vrednost y (četrtna prodaja). Enačba, ki opisuje, kako y je povezana z x se imenuje regresijski model.

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

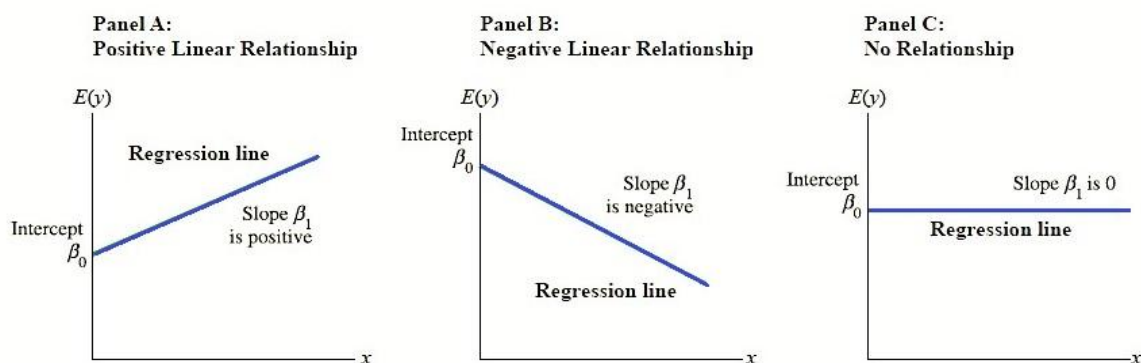
β_0 in β_1 imenujemo parametri modela, ϵ (grška črka epsilon) je naključna spremenljivka, imenovana napaka modela. Napaka predstavlja variabilnost y ki je ni mogoče pojasniti z linearno povezavo med x in y .

Populacijo vseh restavracij Best Burger lahko obravnavamo tudi kot zbirko pod-populacij, po eno za vsako vrednost posebej. x . Ena podpopulacija je na primer sestavljena iz vseh restavracij Best Burger v bližini univerzitetnih kampusov z 8000 študenti. Drugo podpopulacijo sestavljajo vse restavracije Best Burger v bližini univerzitetnih kampusov z 9000 študenti in tako naprej. Vsaka podpopulacija ima ustrezno porazdelitev vrednosti y . Vsaka porazdelitev vrednosti y ima svojo srednjo ali pričakovano vrednost. Enačba, ki opisuje pričakovano vrednost, je y , označena z $E(y)$, ki je povezana z x se imenuje regresijska enačba. Regresijska enačba za preprosto linearno regresijo je naslednja

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

Graf preproste linearne regresijske enačbe je ravna črta. β_0 predstavlja začetno vrednost regresijske premice, β_1 je smerni koeficient premice in $E(y)$ srednja vrednost ali pričakovana vrednost y za dano vrednost x .

Primeri možnih regresijskih linij so prikazani na spodnji sliki 5.1. Regresijska črta v primeru A kaže, da je vrednost y je pozitivno povezana z x . Ko se vrednosti povečujejo x , se vrednosti povečujejo tudi $E(y)$. Regresijska premica v primeru B kaže, da je vrednost y , ki je negativno povezana z x . Kadar so manjše vrednosti $E(y)$ so povezane z višjimi vrednostmi x . Regresijska črta v panelu C prikazuje primer, ko je vrednost y ni povezana z x . To pomeni, da je vrednost y je enaka za vsako vrednost x .



Slika 5.1 Primeri grafov linearnega razmerja.

5.3 Ocenjena regresijska enačba

Če bi bile vrednosti parametrov populacije znane β_0 in β_1 , bi lahko z zgornjo enačbo izračunali vrednosti y za dano vrednost x . V praksi so ti parametri težko dostopni, zato jih preprosto ocenimo z uporabo vzorčnih podatkov. Vzorčna statistika (označena z b_0 in b_1) se izračunajo kot ocene populacijskih parametrov β_0 in β_1 . Zamenjava vrednosti vzorčnih statistik b_0 in b_1 namesto β_0 in β_1 v regresijski enačbi dobimo novo, ocenjeno regresijsko enačbo. Ocenjena regresijska enačba za enostavno linearno regresijo je naslednja

$$\hat{y} = b_0 + b_1x$$



Graf ocenjene enostavne linearne regresije se imenuje ocenjena regresijska črta. b_0 predstavlja začetno vrednost regresijske premice, b_1 je smerni koeficient premice.

V nadaljevanju prikazujemo, kako z metodo najmanjših kvadratov izračunati vrednosti b_0 in b_1 v ocenjeni regresijski enačbi.

Na splošno $\hat{y}$ (ocena za $E(y)$) povprečna vrednost y za dano vrednost x . Če bi zdaj želeli oceniti pričakovano vrednost četrtnetne prodaje za vse restavracije Best Burger, ki se nahajajo v bližini kampusov z 10000 študenti, bi bila vrednost x v zadnji enačbi nadomestili z vrednostjo 10000. V nekaterih primerih pa nas morda bolj zanima napoved prodaje samo za eno določeno restavracijo. Recimo, da želite napovedati četrtnetno prodajo za restavracijo, ki jo nameravate zgraditi v bližini kolidža z 10000 študenti. Kot se izkaže, je tudi v tem primeru najboljši napovedovalec vrednosti y za dano vrednost x vrednost $\hat{y}$.

5.4 Metoda najmanjših kvadratov

Metoda najmanjših kvadratov je postopek, pri katerem na podlagi vzorčnih podatkov poiščemo enačbo ocenjene regresijske premice. Za ponazoritev metode najmanjših kvadratov predpostavimo, da so bili podatki zbrani na vzorcu 10 restavracij Best Burger v bližini univerzitetnih kampusov. S x_i označimo velikost študentske populacije (v tisočih) in z y_i velikost četrtnetne prodaje (v tisočih EUR). x_i in y_i za 10 vzorčnih restavracij so povzete v spodnji preglednici. Vidimo, da je restavracija 1, z $x_1 = 2$ in $y_1 = 58$, blizu kampusa z 2000 študenti in ima četrtnetno prodajo v višini 58 000 EUR. Restavracija 2, z $x_2 = 6$ in $y_2 = 105$, je blizu kampusa s 6000 študenti in ima četrtnetno prodajo 105.000 €. Restavracija z najvišjo vrednostjo prodaje je restavracija 10, ki je blizu kampusa s 26.000 študenti in ima četrtnetno prodajo 202.000 €.



V nadaljevanju je prikazana razpršena slika podatkov na sliki 5.2 spodaj. Na vodoravni osi je prikazana populacija študentov, na navpični osi pa četrtnetna prodaja. Diagrama razpršitve za regresijsko analizo sta sestavljena z neodvisno spremenljivko x na vodoravni osi in odvisno spremenljivko y na navpični osi. Diagram razpršitve nam tako omogoča, da naredimo predhodne sklepe o možnem razmerju med spremenljivkama.

Restaurant i	Student Population (1000s) x_i	Quarterly Sales (€1000s) 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 5.2 Diagram razpršitve podatkov.

Kakšne predhodne ugotovitve lahko potegnemo iz spodnje slike 5.3? V kampusih z večjim številom študentov je četrtnetna prodaja višja. Poleg tega obstaja konstantna povezava med številom študentov in četrtnetno prodajo, ki jo je mogoče opisati z ravno črto. Med x in y pozitivnim linearnim razmerjem je res mogoče sklepati na pozitivno linearno razmerje. Zato smo izbrali preprost linearni regresijski model za predstavitev povezave med četrtnetno prodajo in številom študentov. Glede na to izbiro je naša naslednja naloga, da s pomočjo preglednice z vzorčnimi podatki določimo vrednosti b_0 in b_1 , ki sta





pomembna parametra pri ocenjevanju enačbe preproste linearne regresije. Za i -to restavracijo je ocenjena regresijska enačba naslednja

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

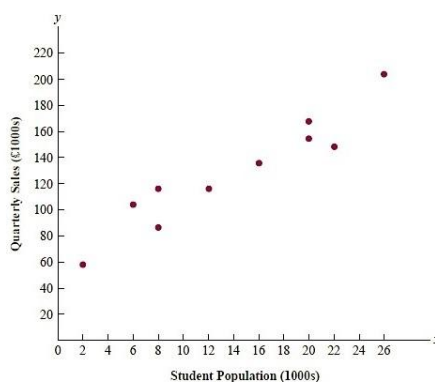
Kje:

$\hat{y}_i$ _ ocenjena vrednost četrtnetne prodaje (1000 EUR) za i -to restavracijo

b_0 _ začetna vrednost ocenjene regresijske premice

b_1 _ smerni koeficient ocenjene regresijske premice

x_i _ velikost populacije učencev (1000) za i -to restavracijo



Slika 5.3 Graf razpršitve.

y_i označuje opazovano (dejansko) prodajo v restavraciji i in $\hat{y}_i$, ki predstavlja ocenjeno vrednost prodaje za restavracijo i . Vsaka restavracija v vzorcu bo imela opazovano vrednost prodaje y_i in predvideno vrednost prodaje $\hat{y}_i$. Da bi ocenjena regresijska premica zagotovila dobro prileganje podatkom, želimo, da so razlike med opazovanimi vrednostmi prodaje in napovedanimi vrednostmi prodaje čim manjše.

Metoda najmanjših kvadratov uporablja vzorčne podatke, ki zagotavljajo vrednosti b_0 in b_1 .

Minimizirati vsoto kvadratov odstopanj med opazovanimi vrednostmi odvisne spremenljivke y_i in napovedano vrednostjo odvisne spremenljivke $\hat{y}_i$. Izhodišče za izračun najmanjše vsote po metodi najmanjših kvadratov je podano z izrazom

Merilo najmanjše vsote: $\min \sum (y_i - \hat{y}_i)^2$

Kje:

y_i = opazovana vrednost odvisne spremenljivke za i -to opazovanje





$\hat{y}_i$ = napovedana vrednost odvisne spremenljivke za i-to opazovanje

smerni koeficient regresijske premice in začetna 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 neodvisne spremenljivke za i-to opazovanje

y_i _ vrednost odvisne spremenljivke za i-to opazovanje

$\bar{x}$ _ povprečna vrednost za neodvisno spremenljivko

$\bar{y}$ _ povprečna vrednost odvisne spremenljivke

n _skupno število opazovanj

V nadaljevanju so prikazani nekateri izračuni, potrebni za izračun ocenjene regresijske premice najmanjših kvadratov. Pri vzorcu 10 restavracij imamo $n=10$ opazovanj. Zgornje enačbe najprej zahtevajo izračun povprečne vrednosti x in povprečno vrednost y .

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

Alternativna enačba za izračun b_1 :

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

Z uporabo zadnjih enačb in informacij na sliki 5.4 lahko izračunamo smerni koeficient regresijske premice za primer restavracije Best Burger. Izračun naklona (b_1) je naslednje.

Slika 5.5 prikazuje graf te enačbe na diagramu razpršitve.

Naklon ocenjene regresijske enačbe ali smerni koeficient enačbe ($b_1 = 5$) je pozitiven.

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

Temu sledi izračun začetne vrednosti (b_0).

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



Restaurant <i>i</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 5.4 Graf enačbe na diagramu razpršitve.

Tako se oceni regresijska enačba:

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

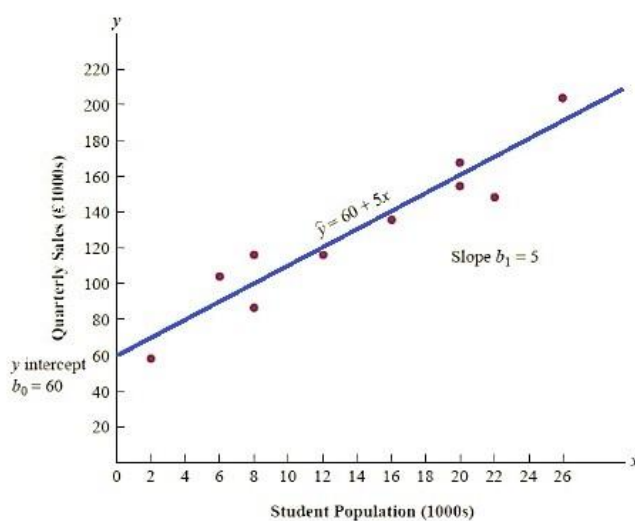
Na sliki je prikazan graf te enačbe na diagramu razpršitve.

Naklon ocenjene regresijske enačbe ($b_1 = 5$) je pozitivna, kar pomeni, da se s povečevanjem števila študentov povečuje tudi prodaja. Dejansko lahko sklepamo (na podlagi izmerjene prodaje v tisočakah in števila študentov v tisočakah), da je povečanje števila študentov za 1000 povezano s povečanjem pričakovane prodaje za 5000; tj. četrtna prodaja naj bi se povečala za 5 EUR na študenta.

Če menimo, da regresijska enačba, ocenjena z metodo najmanjših kvadratov, ustrezno opisuje razmerje med x in y , se zdi smiselno, da z ocenjeno regresijsko enačbo napovemo vrednost y za dano vrednost x . Na primer, če bi želeli napovedati četrtno prodajo za restavracijo, ki je v bližini kampusa s 16.000 študenti, bi izračunali z

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

Zato predvidevamo, da bo četrtna prodaja te restavracije znašala 140.000 evrov. V naslednjih razdelkih obravnavamo metode za ocenjevanje ustreznosti uporabe ocenjene regresijske enačbe za ocenjevanje in napovedovanje.



Slika 5.5 Graf razpršitve števila študentov in četrtnetne prodaje.

5.5 Koeficient determinacije

Za primer restavracij Best Burger smo razvili ocenjeno regresijsko enačbo $y = 60 + 5x$ za približno linearno razmerje med velikostjo študentske populacije x in četrtnetno prodajo y . Vprašanje je, kako dobro se ocenjena regresijska enačba ujema s podatki. V tem razdelku bomo pokazali, da koeficient determinacije zagotavlja merilo dobrega ujemanja ocenjene regresijske enačbe. Za i -to opazovanje je razlika med opazovano vrednostjo odvisne spremenljivke y_i in napovedano vrednostjo odvisne spremenljivke imenujemo i -ti ostanek.

Vsota kvadratov teh ostankov ali napak je količina, ki se minimizira z metodo najmanjših kvadratov. Ta količina, znana tudi kot vsota kvadratov napak, je označena s SSE.



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

Vrednost SSE je merilo napake pri uporabi ocenjene regresijske enačbe za napovedovanje vrednosti odvisne spremenljivke v vzorcu. Slika 5.6 prikazuje izračune, potrebne za izračun vsote kvadratov zaradi napake za primer Best Burger.



Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Predicted Sales $\hat{y}_i = 60 + 5x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
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 5.6 Kvadrati napak v primeru Best Burger.

Recimo, da moramo pripraviti oceno četrtnetne prodaje, ne da bi poznali velikost študentske populacije. Brez poznavanja povezanih spremenljivk bi kot oceno četrtnetne prodaje v kateri koli restavraciji uporabili povprečje vzorca. Tabela na sliki 5.6 je pokazala, da za podatke o prodaji $y_i = 1300$. Zato je povprečna vrednost četrtnetne prodaje za vzorec 10 restavracij Best Burger $y_i/n = 1300/10 = 130$. V preglednici 14.4 prikazujemo vsoto kvadratov odstopanj, ki jih dobimo z uporabo vzorčnega povprečja 130 za napovedovanje vrednosti četrtnetne prodaje za vsako restavracijo v vzorcu. Za i -to restavracijo v vzorcu je razlika y_i predstavlja merilo napake, ki je vključena v aplikacijo za napovedovanje prodaje. Ustrezna vsota kvadratov, imenovana skupna vsota kvadratov, je označena s SST.

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

Restaurant i	x_i = Student Population (1000s)	y_i = Quarterly Sales (€1000s)	Deviation $y_i - \bar{y}$	Squared Deviation $(y_i - \bar{y})^2$
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 5.7 Vsota kvadratov.

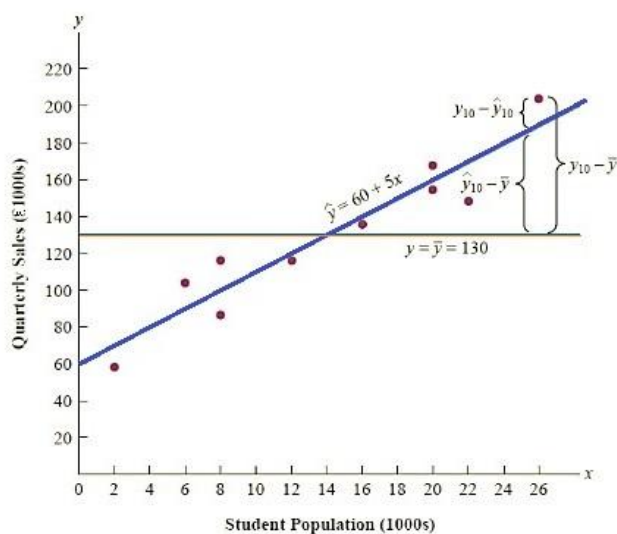
Vsota na dnu zadnjega stolpca na sliki 5.7 je skupna vsota kvadratov za restavracije BestBurger $SST = 15\,730$. Na sliki 5.8 je prikazana ocenjena regresijska črta $y = 60 + 5x$ in črto, ki ustreza $y = 130$. Upoštevajte, da so točke bolj zbrane okoli ocenjene regresijske premice kot okoli premice $y = 130$. Na primer za deseto restavracijo v vzorcu vidimo, da je napaka veliko večja, če za



napovedovanje $y = 10$ uporabimo 130, kot če uporabimo $y = 60 + 5x$ in znaša 190. Na SST lahko gledamo kot na merilo, kako dobro so opazovanja zbrane okoli premice, na SSE pa kot na merilo, kako dobro so opazovanja zbrana okoli premice.

Da bi izmerili, koliko vrednosti na ocenjeni regresijski premici odstopajo od naslednjih, se izračuna še ena vsota kvadratov. Ta vsota kvadratov, imenovana vsota kvadratov zaradi regresije, se označi kot SSR.

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



Slika 5.8 Regresijska črta za primer Best Burger.

Na podlagi prejšnje razprave lahko pričakujemo, da so SST, SSR in SSE povezani. Dejansko je povezava med temi tremi vsotami kvadratov eden najpomembnejših rezultatov v statistiki.

5.6 Razmerje med SST, SSR in SSE:

$$SST = SSR + SSE$$

Kje je:

SST = skupna vsota kvadratov

SSR = vsota kvadratov zaradi regresije

SSE = vsota kvadratov zaradi napake



Enačba ($SST = SSR + SSE$) kaže, da lahko skupno vsoto kvadratov razdelimo na dve komponenti, vsoto kvadratov zaradi regresije in vsoto kvadratov zaradi napake. Če sta torej znani vrednosti



katerih koli dveh od teh vsot kvadratov, lahko tretjo vsoto kvadratov zlahka ugotovimo z izračunom. V primeru restavracij Best Burger na primer že vemo, da je $SSE = 1530$ in $SST = 15,730$; zato z reševanjem za SSR v zgornji enačbi ugotovimo, da je vsota kvadratov zaradi regresije

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

Poglejmo, kako lahko s pomočjo treh vsot kvadratov, SST, SSR in SSE, zagotovimo merilo ustreznosti ocenjene regresijske enačbe. Ocenjena regresijska enačba bi se popolnoma ujemala, če bi vsaka vrednost odvisne spremenljivke y_i ležala naključno na ocenjeni regresijski premici. V tem primeru bi bila za vsako opazovanje enaka nič, kar bi pomenilo $SSE = 0$. Ker je $SST = SSR + SSE$, vidimo, da mora biti za popolno ujemanje SSR enak SST, razmerje (SSR/SST) pa mora biti enako ena. Slabše prileganje bo imelo za posledico večje vrednosti SSE. Če v enačbi (14.11) rešimo SSE, vidimo, da je $SSE = SST - SSR$. Zato je največja vrednost SSE (in s tem najslabše prileganje), ko je $SSR = 0$ in $SSE = SST$.

Za ocenjevanje se uporablja razmerje SSR/SST , ki ima vrednosti med nič in ena, ki ustrezajo ocenjeni regresijski enačbi.

To razmerje se imenuje koeficient determinacije in se označuje z r^2 .

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

Za primer restavracij Best Burger je vrednost koeficienta determinacije

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

Ko je koeficient determinacije izražen v odstotkih, lahko r^2 lahko interpretiramo kot odstotek skupne vsote kvadratov, ki ga lahko pojasnimo z ocenjeno regresijsko enačbo. Za restavracije Best Burger Restaurants lahko sklepamo, da je 90,27 % skupne vsote kvadratov mogoče pojasniti z ocenjeno regresijsko enačbo. $y = 60 + 5x$ za napoved četrletne prodaje. Z drugimi besedami, 90,27 % variabilnosti prodaje je mogoče pojasniti z linearno povezavo med velikostjo študentske populacije in prodajo. Zadovoljni moramo biti, da se tako dobro ujema z ocenjeno regresijsko enačbo.



5.7 Korelacijski koeficient

Korelacijski koeficient je opisno merilo moči linearne povezave med dvema spremenljivkama, x in y , z vrednostmi med -1 in +1. Vrednost +1 pomeni popolno pozitivno linearno povezavo, kjer vse podatkovne točke ležijo na premici a s pozitivnim naklonom, medtem ko vrednost -1 nakazuje popolno negativno linearno povezavo, pri kateri vse točke ležijo na premici a z negativnim naklonom. Vrednosti blizu nič kažejo, da x in y nista linearno povezani.

Če je regresijska analiza že izvedena in je koeficient determinacije r^2 znan, lahko koeficient korelacije vzorca izračunamo na naslednji način

$$r_{xy} = (\text{sign of } b_1) \sqrt{\text{coefficient of determination}}$$

$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$



PEARSONOV KORELACIJSKI KOEFICIENT: VZORČNI PODATKI

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

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

Kje so:

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

Znak za vzorec korelacijskega koeficienta je pozitiven, če ima ocenjena regresijska enačba pozitiven naklon ($b_1 > 0$) in negativen, če ima ocenjena regresijska enačba negativni naklon ($b_1 < 0$).

Za primer Best Burger je vrednost koeficienta determinacije, ki ustreza ocenjeni regresijski enačbi $y = 60 + 5x$ znaša 0,9027. Ker je naklon ocenjene regresijske enačbe pozitiven, enačba (14.13) kaže, da je koeficient korelacije vzorca Z koeficientom korelacije vzorca

$R_{xy}=0,9501$, bi sklepali, da obstaja močna pozitivna linearna povezava med x in y .



V primeru linearnega razmerja med dvema spremenljivkama tako koeficienti determinacije kot koeficient vzorčne korelacije zagotavljajo merilo za moč razmerja.

Koeficient determinacije predstavlja mero med nič in ena, medtem ko koeficient vzorčne korelacije predstavlja mero med -1 in +1. Čeprav je koeficient vzorčne korelacije omejen na linearno razmerje med dvema spremenljivkama, se koeficient determinacije lahko uporablja za nelinearna razmerja in razmerja, ki imajo dve ali več neodvisnih spremenljivk. Tako koeficient determinacije omogoča širši obseg uporabe.

5.8 Večkratni regresijski model

V naslednjih razdelkih nadaljujemo s preučevanjem regresijske analize in obravnavamo situacije, ki vključujejo dve ali več neodvisnih spremenljivk. To tematsko področje, imenovano multipla regresijska analiza, nam omogoča, da upoštevamo več dejavnikov in tako dobimo boljše napovedi, kot so mogoče z enostavno linearno regresijo.



Večkratna regresijska analiza je študija povezanosti odvisne spremenljivke y z dvema ali več neodvisnimi spremenljivkami. V splošnem primeru bomo s p označili število neodvisnih spremenljivk.

5.9 Regresijski model in regresijska enačba

Pojma regresijski model in regresijska enačba, predstavljena v prejšnjem razdelku, veljata tudi v primeru večkratne regresije. Enačba, ki opisuje, kako je odvisna spremenljivka y povezana z neodvisnimi spremenljivkami $x_1, x_2, \dots, x_p$ in izrazom napake, se imenuje model večkratne regresije. Začnemo s predpostavko, da ima model večkratne regresije naslednjo obliko.

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

V modelu multiple regresije so $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ parametri in izraz napake (ϵ) je naključna spremenljivka. Natančen pregled tega modela pokaže, da je y linearna funkcija spremenljivk $x_1, x_2, \dots, x_p$ in izraza napake ϵ epsilon. Člen napake upošteva variabilnost y , ki je ni mogoče pojasniti z linearnim učinkom p neodvisnih spremenljivk.

V razdelku 5.10 obravnavamo predpostavke za model multiple regresije in epsilon. Ena od predpostavk je, da je povprečna ali pričakovana vrednost (ϵ) enaka nič. Posledica te predpostavke je, da je povprečna ali pričakovana vrednost y , označena z $E(y)$, enaka $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots +$



$\beta_p x_p$. Enačba, ki opisuje, kako je srednja vrednost y povezana z $x_1, x_2, \dots, x_p$ se imenuje enačba multiple regresije.

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

5.10 Ocenjena enačba multiple regresije

Če so vrednosti $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ znane, lahko enačbo (5.9) uporabimo za izračun povprečne vrednosti y pri danih vrednostih $x_1, x_2, \dots, x_p$. Žal te vrednosti parametrov običajno niso znane in jih je treba oceniti iz vzorčnih podatkov. Za izračun vzorčne statistike se uporabi enostavni naključni vzorec $b_0, b_1, b_2, \dots, b_p$, ki se uporabijo kot točkovne ocene parametrov $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. Te vzorčne statistike zagotavljajo naslednje ocenjevanje enačb multiple regresije:



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

Kjer so

$b_0, b_1, b_2, \dots, b_p$ ocene $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ in

$\hat{y}$ = napovedana vrednost odvisne spremenljivke

Primer: Prevozno podjetje Frigo

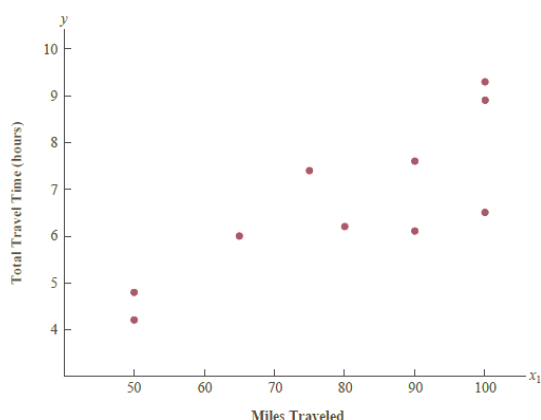
Za ponazoritev večkratne regresijske analize bomo obravnavali problem, s katerim se je soočalo podjetje Frigo Trucking Company, neodvisno prevozniško podjetje v južni Italiji. Največji del poslovanja podjetja Frigo vključuje dostavo na lokalnem območju. Za boljši razvoj delovnih razporedov želijo vodje načrtovati skupni dnevni potovalni čas za svoje voznike.

Sprva so vodje menili, da bo skupni dnevni čas potovanja tesno povezan s številom prevoženih kilometrov pri dnevni dostavi. S preprostim naključnim vzorcem 10 voznških nalog so bili pridobljeni podatki, prikazani na sliki 5.9, in razpršeni diagram. Po pregledu tega razpršenega diagrama so vodje domnevali, da je mogoče za opis povezave uporabiti preprost linearni regresijski model $y = \beta_0 + \beta_1 x_1 + \epsilon$ med skupnim časom vožnje (y) in številom prevoženih kilometrov (x_1).



Driving Assignment	x_1 = KM Traveled	y = Travel Time (hours)
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 5.9 Podatki za primer prevoznega podjetja Frigo.



Slika 5.10 Diagram razpršitve za primer podjetja Frigo Transport

Za oceno parametrov β_0 in β_1 je bila za pripravo ocenjene regresije uporabljena metoda enačbe najmanjših kvadratov.

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

Na zgornji sliki je prikazan rezultat programa Minitab z uporabo preproste linearne regresije za podatke v zgornji tabeli. Ocenjena regresijska enačba je

$$\hat{y} = 1.27 + 0.0678x_1$$

Pri stopnji pomembnosti 0,05 F-vrednost 15,81 in ustrezna p-vrednost 0,004 kažeta, da je razmerje pomembno. To pomeni, da lahko zavrnilo $H_0: \beta_1 = 0$ ker je p-vrednost manjša od $\alpha = 0,05$. Upoštevajte, da enak sklep izhaja iz vrednosti $t = 3,98$ in pripadajoče p-vrednosti 0,004. Tako lahko sklepamo, da je povezanost med skupnim časom potovanja in številom prevoženih kilometrov pomembna. Daljši čas potovanja je povezan z več prevoženimi kilometri. S koeficientom determinacije (izraženim v odstotkih) $R - Sq = 66,4 \%$, vidimo, da lahko 66,4 % variabilnosti potovalnega časa pojasnimo z linearnim učinkom števila prevoženih kilometrov.



Ta ugotovitev je dokaj dobra, vendar bodo vodje morda želeli razmisliti o dodajanju druge neodvisne spremenljivke, da bi pojasnili nekatere preostale spremenljivke v odvisni spremenljivki.

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 5.11 Rezultati z eno neodvisno spremenljivko.

Pri iskanju druge neodvisne spremenljivke so menedžerji menili, da lahko k skupnemu času potovanja prispeva tudi število dostav. Podatki družbe Frigo Trucking z dodanim številom dostav so prikazani na spodnji sliki. (x_1) in število dostav (x_2) kot neodvisni spremenljivki sta prikazani na sliki 5.12. Ocenjena regresijska enačba je

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

DATA FOR FRIGO TRUCKING WITH KM TRAVELED (x_1) AND NUMBER OF DELIVERIES (x_2) AS THE INDEPENDENT VARIABLES

Driving Assignment	x_1 = km Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
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 5.12 Podatki in neodvisne spremenljivke družbe Frigo Trucing.

Podrobneje si oglejmo vrednosti $b_1 = 0,0611$ in $b_2 = 0,923$ v zadnji enačbi.



Opomba o razlagi koeficientov

Na tej točki lahko komentiramo razmerje med ocenjeno regresijsko enačbo z neodvisno spremenljivko samo prevoženi kilometri in enačbo, ki vključuje število dostav kot drugo neodvisno spremenljivko. Vrednost b_1 v obeh primerih ni enaka.



Pri preprosti linearni regresiji razlagamo b_1 kot oceno spremembe y za spremembo neodvisne spremenljivke za eno enoto. Pri večkratni regresijski analizi je treba to razlago nekoliko spremeniti. To pomeni, da se pri večkratni regresijski analizi vsak regresijski koeficient razlaga na naslednji način: predstavlja bi oceno spremembe v y ki ustreza spremembi v x_i za eno enoto, če so vse druge neodvisne spremenljivke nespremenjene.

V primeru družbe Frigo Trucking to vključuje dve neodvisni spremenljivki, $b_1=0,0611$ in $b_2=0,923$.

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

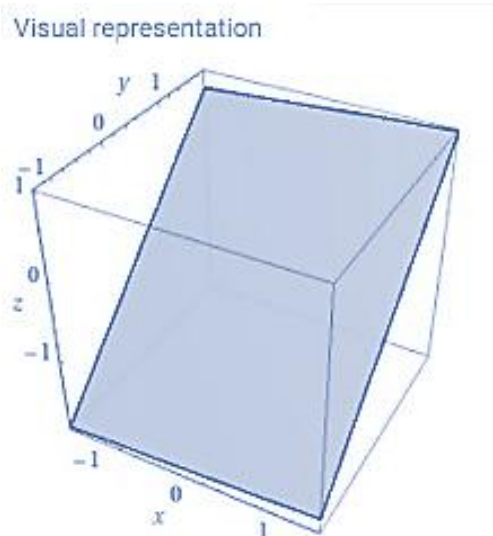
S = 0.573142 R-Sq = 90.4% R-Sq(adj) = 87.6%

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 5.13 Rezultati za Frigo Trucking z dvema neodvisnima spremenljivkama.

Tako je 0,0611 ure ocena pričakovanega podaljšanja potovalnega časa, ki ustreza povečanju za eno miljo na prevoženo razdaljo, če je število dostav konstantno. Podobno, ker je $b_2=0,923$, je ocena pričakovanega povečanja potovalnega časa, ki ustreza povečanju za eno dostavo, ko je število prevoženih milj konstantno, 0,923 ure.



Slika 5.14 Vizualni prikaz rezultatov za primer Frigo Trucking.

Literatura poglavje 5

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/97898151231351230101
- 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