Regresija

Jedan od ciljeva u velikom broju istraživanja je da se opišu veze među pojavama koje nas okružuju. To se može postići pronalaženjem formule ili jednačine koja povezuje veličine koje posmatramo. Na primer, može nas interesovati veza između temperature i pritiska u hemijskom procesu; ili veza između broja jabuka na drveću u voćnjaku i količine đubriva koje je korišćeno u voćnjaku; ili kako nova vakcina utiče na bolest. Veza između količine padavina, temperature i vlažnosti ili veza između prinosa i sorte žita.

U statistici pronalaženjem statističkih veza između pojava bavi se regresiona analiza, regresija. Regresija je od velikog značaja, kako u ekonomiji i privredi, tako i u drugim prirodnim naukama, kao što su: hemija, fizika, biologija, farmakologija, toksikologija, biohemija i sudska medicina....

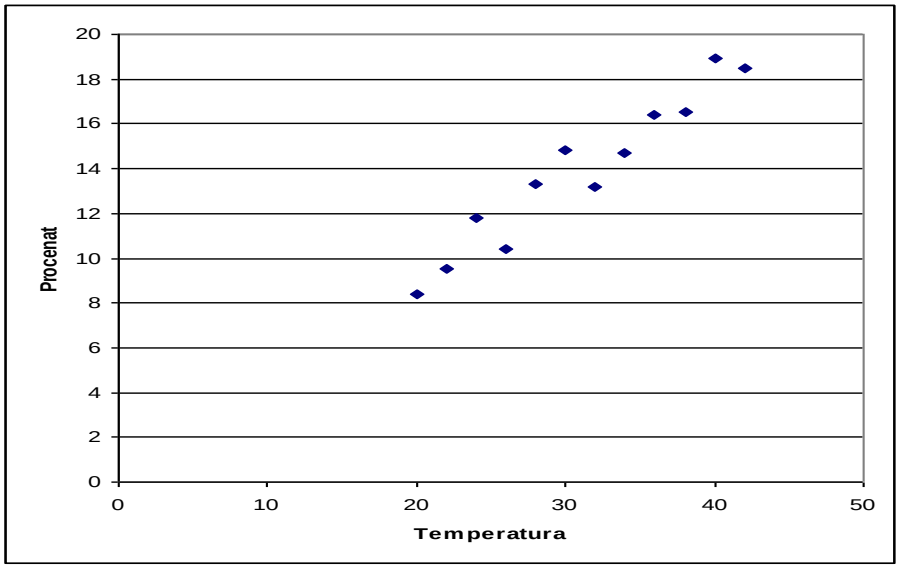
Primer 1.

U tabeli 1. dati su podaci o procentu izdvojene supstance za različite visine temperature pri nekom hemijskom procesu.

Tabela 1.

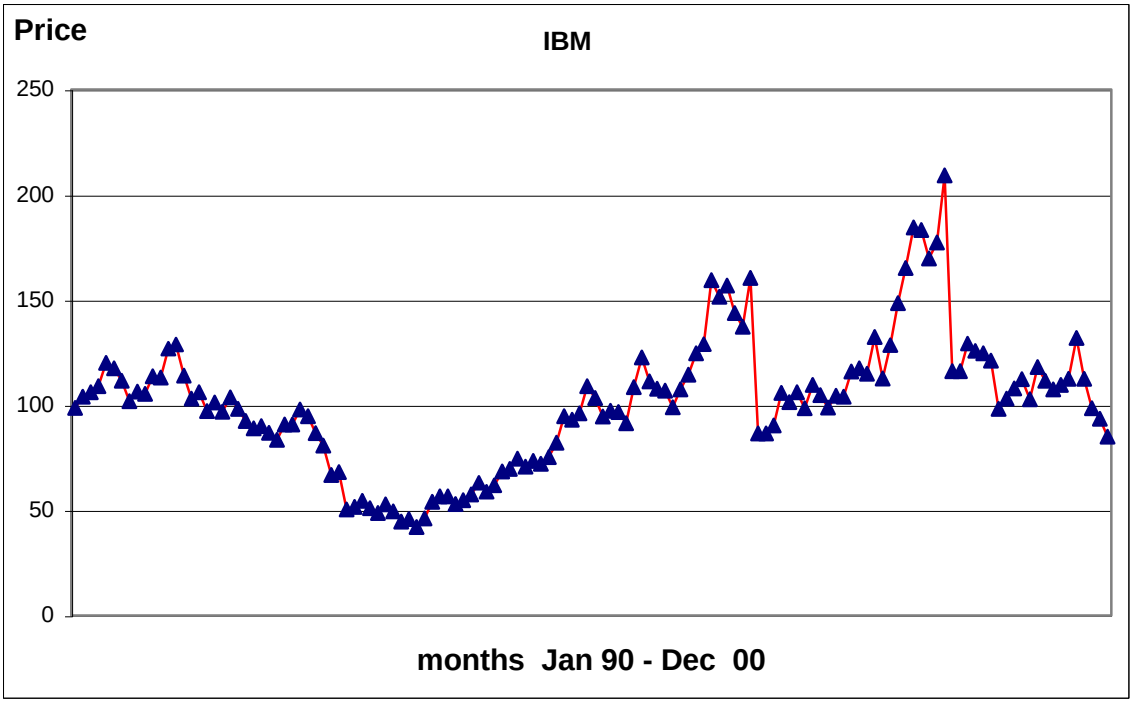
temperatura	20	22	24	26	28	30	32	34	36	38	40	42
procenti	8.4	9.5	11.8	10.4	13.3	14.8	13.2	14.7	16.4	16.5	18.9	18.5

Na slici 1. ovi podaci su prikazani grafički u obliku **dijagrama rasipanja**



Slika 1.

Takođe, često se traže veze između ekonomskih pojava, kao na primer veza između potražnje, ponude i cene nekog proizvoda, kretanje cene nekog proizvoda ili akcije tokom vremena. Na slici 2. prikazane su promene cene akcije IBM-a od januara 1990. do decembra 2000. godine.



Slika 2.

Regresija je jedna od metoda u okviru metoda koje čine statistical learning - veliki skup metoda, tehnika, alata statistike za modeliranje i razumevanje kompleksnih skupova podataka. Regresija je jedan od načina da se sastavi model za predviđanje i ocenjivanje jedne ili više zavisnih promenljivih na osnovu jedne ili više nezavisnih promenljivih. U regresiji postoji izlaz, za razliku od drugih statističkih tehnika koje se bave problemima u kojima ne postoji zavisna promenljiva.

Ulaz / Izlaz (Input / output);
Prediktor / Odgovor (Predictor / response);
Nezavisna promenljiva / Zavisna promenljiva (Independent variable / Dependent variable);

Primeri. Predvideti :

- Visinu plate na osnovu godina školovanja i godina starosti.
- Prinos žita na osnovu količine đubriva, količine padavina, vrste žita.
- VaR na osnovu istorijskih podataka.
- Da li će indekx S&P sutra rasti ili padati.
- Vrednost indeksa S&P na osnovu istorijskih podataka.
- Da li je poruka spam ili ham.
- Količinu prodatih proizvoda na osnovu količine novca uloženog u reklamu preko TV, radia i novina.

Problem opisivanja ovakvih veza svodi se na pronalaženje modela koji povezuje jednu ili više zavisnih promenljivh $Y_1, Y_2, ..., Y_p$ sa jednom ili više nezavisnih promenljivih $x_1, x_2, ..., x_k$ pomoću neke funkcionalne zavisnosti. Oblik ove funkcionalne zavisnosti je najčešće nepoznat, pa ostaje na istraživaču da izabere onu koja je po nekom krierijumu najbolja. Veoma često se koriste polinomne funkcije, ali isto tako i eksponencijalne ili neke druge funkcije.

Opšti problem nalaženja funkcije koja dobro aproksimira dobijeni skup podataka, često se naziva "***fitovanje***" krive, ili određivanje ***regresione linije***.

Model

$$Y = f(X) + \varepsilon$$

f sistematska informacija koju X daje o Y. U opštem slučaju nepoznata, ocenjuje se na osnovu registrovanih podataka, uzorka.

ε slučajna greška: sadrži neizmerene promenljive koje utiču na Y, greške pri merenju.

Postoje dva pristupa ovom problemu, prema prirodi podataka koji se aproksimiraju.

1. Posmatra se uticaj nezavisne promenljive X, koja je slučajna promenljiva na zavisnu promenljivu Y koja je isto slučajna veličina. Kako svaka realizovana vrednost x slučajne promenljive X proizvodi realizovanu vrednost y slučajne promenljive Y, potrebno je pronaći funkciju μ koja dobro aproksimira vrednosti slučajne promenljive Y pri svakoj realizovanoj vrednosti slučajne promenljive X. Oblik funkcije μ je nepoznat, a jedan od mogućih načina određivanja ove funkcije je da se kao kriterijum "dobrog" aproksimiranja uzme srednje kvardatna greška

$$E(Y - \mu(X))^2$$

Ovako odabrana funkcija se naziva ***funkcija regresije*** Y na X. Ova tip regresije se naziva ***regresija prve vrste***.

2. U pristupu traženja funkcionalne zavisnosti između dve promenljive se posmatra uticaj jedne nezavisne promenljive x koja je deterministička na zavisnu promenljivu Y koja je slučajna veličina. Neslučajna veličina ili veličine koje se posmatraju nazivaju se kontrolisani faktori. Ishod posmatranja, obeležje Y, je slučajan, jer na njega sem kontrolisanih utiču i, po pravilu, slučajni faktori koji se ne mogu kontrolisati (na primer greške pri merenju), ali i neki drugi neslučajni faktori koji su obično prisutni ali se njihov uticaj ne može meriti. Funkcionalna zavisnost koja se određuje naziva se ***regresija druge vrste***, a često se takvi modeli nazivaju i linearni modeli, zbog određenih pretpostavki o linearnosti. Mi ćemo ovde izložiti osnovne pojmove o regresiji druge vrste jer se one primenjuje u analizi vremenskih serija.

Problemi povezani sa regresijom su se prvi put posmatrali u 18. veku, i bili su povezani sa navigacijom pomoću astronomije Legendre (Ležandr) je razvio metod najmanjih kvadrata 1805. Godine. Gauss (Gaus) je tvrdio da je ovaj metod razvio nekoliko godina ranije i pokazao je da je metoda najmanjih kvardrata optimalno rešenje ako greške imaju normalnu raspodelu.

Sama reč ***regresija*** znači vraćanje unatrag. Naziv ***regresija*** za ovakve modele potiče od statitičara Sir Fransis Galtona, (1822 - 1911) koji je prvi upotebio date modele za ispitivanje bioloških i psiholoških pojava, posmatrajući regresionu jednačinu oblika:

$$\frac{y-\bar{y}}{SD_y}=\frac{x-\bar{x}}{SD_x}.$$

Galton je ispitivao vezu između visina roditelja i dece. Ustanovio je da roditelji sa natprosečnom visinom teže da imaju decu koja su takođe viša od proseka, ali ne toliko visoka kao roditelji. Ta činjenica važi i za decu izuzetno niskih roditelja: deca su taođe bila niska ali ne toliko niska kao roditelji. Visine su se "vraćale natrag" , regresirale prema proseku. Galton je ovu činjenicu nazvao "regresija prema proseku", a naziv je ostao i za metodu.

Osnovni razlozi zašto se ocenjuje f : predviđanje i zaključivanje.
Predviđanje.

Česte su situacije, kada su vrednosti za X dostupne, a za Y ne. Kako su, u proseku greške nula, Y se predviđa kao $\hat{Y}=\hat{f}(X)$. U principu, nije bitan oblik funkcije f, ako se dobija dobro predviđanje. Tačnost predviđanja $\hat{Y}$ zavisi od dve veličine: greška koja se može smanjiti (reducible error) i greška koja se ne može smanjiti (ireducible error). $\hat{f}$ neće savršeno oceniti f, i ta netačnost izaziva grešku. Ova greška se može smanjiti poboljšanjem aproksimacije, uvođenjem nove funkcije, koja je bolja ocena. Međutim čak i kada bismo mogli da nađemo pravu vezu između X i Y, predviđanje bi u sebi sadržalo grešku. To je zbog toga što je Y funkcija i od ε , greška koja se ne može predvideti pomoću X. Ova greška se ne može smanjiti.

Zaključivanje.
Često nas interesuje na koji način X utiče na Y. Interesuje nas da razumemo vezu koja postoji između X i Y, odnosno kako se Y menja kao funkcija od X. Nije nam osnovni cilj da predvidimo vrednosti za Y.

- Koji prediktori su povezani sa odgovorom?
- Kakva je povezanost odgovora sa svakom promenljivom?
- Kolika je jačina veze?
- Da li se povezanost odgovora i prediktora moze aproksimirati linearnom jednačinom, ili je potebna komplikovanija veza?

U primeru sa prodajom i reklamom: Dati su podaci o prodaji proizvoda na 200 različitih tržišta, zajedno sa sumom uloženom u reklamiranje proizvoda preko tri medija – TV, radio, novine. Podaci o količino prodatih proizoda su dati u hiljadama jedinica, a novac uložen u reklame je dat u hiljadama dolara. Interesuje nas:

- Da li postoji veza između ulaganja u reklamu i količine prodatih proizvoda?
- Koji medij doprinosi većoj prodaji?
- Koji medij najviše povećava prodaju?
- Koliki porast prodaje je rezultat određenog povećanja ulaganja u TV reklame?
- Koliko je precizna ocena dejstva svakog medija na prodaju?
- Da li je relacija linearna?

- Da li postoji udruženo delovanje medija na prodaju? (Interakcija).

Kako se ocenjuje f?
Na osnovu registrovanog uzorka, (x_1, y_1), (x_2, y_2),..., (x_n, y_n). Training data. Pomoću ovih merenja treniramo, učimo naš model kako da oceni f.

Dva pristupa ocenjivanju funkcije f:

Parametarski metod: Pretpostavimo kako funkcija f izgleda, na primer

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Ocenimo parametre nekom od metoda, na primer, metodom najmanjih kvadrata.
Ovaj metod je pogodan jer se parametri ocenjuju, a model je linearna funkcija parametara. Problem je što tako odabrana fukcija najčešće ne odgovara stvarnoj funkciji f.

Neparametarski metod: Ne pretpostavljamo kako f izgleda. Tražimo ocenu za f dok ne priđemo registrovanim podacima što je moguće bliže. U ovom slučaji potrebni su veliki skupovi podataka.

Linearni model - linearna regresija druge vrste

Linearan model, koji se koristi da se objasni slučajna promenljiva Y pomoću neslučajne promenljive x, obično se daje u obliku

$$Y = \mu(x) + \varepsilon, \qquad (1)$$

gde su Y i ε slučajne promenljive, x je neslučajna promenljiva, a μ je funkcija od x. Funkcija μ se naziva deterministički deo modela, a ε je slučajni deo modela. Nezavisna promenljiva x je kontrolisana, vrednosti zavisno promenljive Y se mogu meriti, dok se vrednosti promenljive ε, koja se naziva i greška, ne mogu meriti. Promenljiva ε sadrži u sebi sve ostale promenljive koje utiču na vrednosti promenljive Y, ali nisu obuhvaćene posmatranjem. Na primer, prinos jabuka u voćnjaku zavisi, pored količine đubriva i od sastava zemljišta, količine padavina temperature… Sve ove promenljive dovodiće do slučajnih pomeranja i odstupanja od predviđanog modela. Iako se vrednosti promenljive ε ne mogu meriti, često se kao deo modela daju pretpostavke o njenoj raspodeli.

Oblik funkcije μ je poznat, ali često zavisi od nekih nepoznatih parametara. Naziv linearan znači da je funkcija μ linearna funkcija nepoznatih parametara. Naopštiji oblik ove funkcije zavisi od $p+1$ nepoznatih parametara $\beta_0, \beta_1, ..., \beta_p$ i može se zapisati kao

$$\mu(x) = \beta_0 + \beta_1 q_1(x) + \beta_2 q_2(x) + \cdots \beta_p q_p(x),$$

gde su $q_1, q_2, ..., q_p$ poznate funkcije.

Primer 1. Neki mogući oblici zavisnosti.

$Y=\beta_0+\beta_1 x+\varepsilon$	$E(\varepsilon)=0$	$D(\varepsilon)=\sigma^2$ (nepoznata)
$Y=\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\varepsilon$	$E(\varepsilon)=0$	$D(\varepsilon)=\sigma^2$ (nepoznata)
$Y=\beta_1 x+\beta_2 e^x+\varepsilon$	ε : N $(0,\sigma^2)$	σ^2 nepoznata

Kako je $E(\varepsilon)=0$ imamo da je $E(Y)=\mu(x)$, pa se model (1) može zapisati i kao

$$Y_x=\mu(x)+\varepsilon_x,$$

što je model za populaciju, odnosno daju vezu između Y i bilo koje vrednosti x-a. Za svako fiksirano x, slučajna promenljiva Y_x ima svoju (uslovnu) raspodelu, F_{Y_x}, koja u opštem slučaju može biti različita za svako x.

Da bi se ocenili nepoznati paratmetri, mora se koristiti uzorak. Za fiksiranih n vrednosti $x_1,x_2,...,x_n$ nezavisno promenljive x, određuju se vrednosti promenljive Y. Na taj način se dobija n parova $(x_1,Y_1),(x_2,Y_2),\ldots,(x_n,Y_n)$ koji čine *uzorački model*:

$$\begin{aligned} &Y_i=\mu(x_i)+\varepsilon_i,\\ &E(\varepsilon_i)=0 \end{aligned} \qquad i=1,2,...,n, \tag{2}$$

gde su:

- $Y_1,Y_2,...,Y_n$ slučajne promenljive koje se mogu meriti,
- $x_1,x_2,...,x_n$ su ne slučajne promenljive,
- μ je funkcija koja linearno zavisi od p parametara $\beta_0,\beta_1,...,\beta_p$,
- $\varepsilon_1,\varepsilon_2,...,\varepsilon_n$ su slučajne promenljive, greške, *reziduali*, koje se ne mogu registrovati takve da je $\operatorname{cov}(\varepsilon_i,\varepsilon_j)=\sigma_{ij}$, $i,j=1,2,...,n$.

Mogu se staviti i dodatne pretpostavke na promenljive $\varepsilon_1,\varepsilon_2,...,\varepsilon_n$, kao na primer da su međusobno nezavisne, ili da imaju normalnu raspodelu. Uslov da su im disperzije jednake, odnosno

$$D(\varepsilon_i)=\sigma^2,\quad i=1,2,...,n,$$

naziva se *homoskedastičnost*. Slučaj kada uslov homoskedastičnosti nije ispunjen, naziva se *heteroskedastičnost*. Uslov $E(\varepsilon_i\varepsilon_j)=0$ je uslov nekoreliranosti grešaka. Ako on nije ispunjen kažemo da su greške autokotelirane.

Iz jednačina u (2) dobijaju se ocene za nepoznate parametre $\beta_0, \beta_1, ..., \beta_p$ funkcije μ. Za konkretno izmerene vrednosti, odnosno za realizovani uzorak dobijamo parove $(x_1, y_1), (x_2, y_2)$ $,\ldots, (x_n, y_n)$, i dobijaju se konkretne, realizovane vrednosti nepoznatih parametara.

Ocene parametara u linearnom modelu

Određivanje parametara u funkciji μ u modelu vrši se tako da se pomoću nje najbolje

opiše veza između promenljivih Y i x, odnosno da funkcija μ što manje odstupa od registrovanih vrednosti promenljive Y. Jedna od mogućih mera odtupanja funkcije μ od vrednosti promenljive Y je suma kvadrata odstupanja

$$F = \sum_{i=1}^{n} (Y_i - \mu(x_i))^2 = \sum_{i=1}^{n} \varepsilon_i^2 .$$

Ocene parametrara se dobijaju tako da se traži minimum funkcije F. Ova metoda traženja ocena nepoznatih parametara u funkciji μ naziva se *metoda najmanjih kvadrata.* Nalaženje minimuma funkcije

$$F = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 q_1(x) + \beta_2 q_2(x) + \cdots \beta_p q_p(x)))^2 \to \min$$

svodi se na rešavanje sistema **normalnih jednačina**:

$$\frac{\partial F}{\partial \beta_i} = 0, \qquad i = 0, 1, 2, ..., p .$$

Iz sistema normalnih jednačina dobijaju se ocene za nepoznate parametre $\beta_0, \beta_1, ..., \beta_p$ funkcije μ. Za konkretno izmerene vrednosti, odnosno za realizovani uzorak dobijamo parove $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ pomoću kojih se dobijaju konkretne, realizovane vrednosti ocena nepoznatih parametara.

Dakle, ocene $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ parametara $\beta_0, \beta_1, ..., \beta_p$ zadovoljavaju sistem jednačina:

$$\frac{\partial F}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 q_1(x_i) + \beta_2 q_2(x_i) + \cdots \beta_p q_p(x_i))) = 0$$

$$\frac{\partial F}{\partial \beta_1} - 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 q_1(x_i) + \beta_2 q_2(x_i) + \cdots \beta_p q_p(x_i)))q_1(x_i) = 0$$

$\cdots$

$$\frac{\partial F}{\partial \beta_p} - 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 q_1(x_i) + \beta_2 q_2(x_i) + \cdots \beta_p q_p(x_i)))q_p(x_i) = 0$$

Ovaj sistem se naziva sistem normalnih jednačina i svodi se na:

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} q_1(x_i) + \beta_2 \sum_{i=1}^{n} q_2(x_i) + \cdots \beta_p \sum_{i=1}^{n} q_p(x_i) = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} q_1(x_i) + \beta_1 \sum_{i=1}^{n} q_1^2(x_i) + \beta_2 \sum_{i=1}^{n} q_2(x_i) q_1(x_i) + \cdots \beta_p \sum_{i=1}^{n} q_p(x_i) q_1(x_i) = \sum_{i=1}^{n} Y_i q_1(x_i)$$

…

$$\beta_0 \sum_{i=1}^{n} q_p(x_i) + \beta_1 \sum_{i=1}^{n} q_1(x_i) q_p(x_i) + \beta_2 \sum_{i=1}^{n} q_2(x_i) q_p(x_i) + \cdots \beta_p \sum_{i=1}^{n} q_p^2(x_i) = \sum_{i=1}^{n} Y_i q_p(x_i)$$

Rešavanjem ovog sistema po $\beta_0, \beta_1, ..., \beta_p$ dobijamo ocene $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$ nepoznatih parametara, dobijene metodom najmanjih kvadrata.

Primer 2. Ako je funkcija μ polinom stepena p, odnosno ako je

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots \beta_p x^p$$

sistem normalnih jednačina je

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 + \cdots \beta_p \sum_{i=1}^{n} x_i^p = \sum_{i=1}^{n} Y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_2 \sum_{i=1}^{n} x_i^3 + \cdots \beta_p \sum_{i=1}^{n} x_i^{p+1} = \sum_{i=1}^{n} Y_i x_i$$

…

$$\beta_0 \sum_{i=1}^{n} x_i^p + \beta_1 \sum_{i=1}^{n} x_i^{p+1} + \beta_2 \sum_{i=1}^{n} x_i^{p+2} + \cdots \beta_p \sum_{i=1}^{n} x_i^{2p} = \sum_{i=1}^{n} Y_i x_i^p .$$

Opšte rešenje ovog sistema se lakše predstavlja u matričnom zapisu datog modela, ali to prevazilazi okvire ovog teksta. Zainteresovanog čitaoca upućujemo na [Graybill].

Ocene nepoznatih parametara $\beta_0, \beta_1, ..., \beta_p$ dobijene metodom najmanjih kvadrata imaju, pod nekim uslovima, dobre osobine. O tome govori teorema Gaus Markova, koju dajemo bez dokaza.

Teorema (Gaus Markova) Neka je u modelu

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots \beta_p x_i^p + \varepsilon_i \qquad i = 1, 2, ..., n$$
$$E(\varepsilon_i) = 0 \qquad , \tag{3}$$

ispunjeno

1. $E(\varepsilon_i)=0$, $D(\varepsilon_i)=\sigma^2$, $i=1,2,...,n$,

2. $\text{cov}(\varepsilon_i,\varepsilon_j)=0$, za $i\neq j$.

Tada su ocene $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_p$ nepoznatih parametara $\beta_0,\beta_1,...,\beta_p$ dobijene metodom najmanjih kvadrata najefikasnije ocene u klasi svih linearnih centriranih ocena. Dakle, ocene $\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_p$ su centrirane i imaju najmanju disperziju.

Ocene parametara u linearnom modelu mogu se dobiti i na drugi način, na primer metodom maksimalne verodostojnosti.

Neki specijalni slučajevi linearnog modela

Prosta linearna regresija

Funkcija μ je linearna po nezavisnoj promenljivoj.

Sada je uzoračkи model

$$\begin{aligned} &Y_i=\mu(x_i)+\varepsilon_i=\beta_0+\beta_1x_i+\varepsilon_i \\ &E(\varepsilon_i)=0,\ D(\varepsilon_i)=\sigma^2, && i=1,2,...,n \\ &\text{cov}(\varepsilon_i\varepsilon_j)=0 && i\neq j \end{aligned} \qquad (4)$$

.

Ocene paramerata

Metodom najmanjih kvadrata traži se minimum funkcije:

$$F=\sum_{i=1}^{n}(Y_i-(\beta_0+\beta_1x))^2=\sum_{i=1}^{n}\varepsilon_i^2,$$

a sistem normalnih jednačina postaje

$$\beta_0 n+\beta_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}Y_i$$

$$\beta_0\sum_{i=1}^{n}x_i+\beta_1\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}Y_ix_i.$$

Ako označimo

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}_n, \quad \tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n,$$

$$S_{xY} = \sum_{i=1}^{n} x_i Y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i (x_i - \bar{x})$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2,$$

rešenja ovog sistema su

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}.$$

Ocenjeni model (fitovani model) je

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, 2, ..., n,$$

koji se može napisati i u obliku

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x},) = \hat{\beta}_0^{'} + \hat{\beta}_1 (x_i - \bar{x},) \qquad i = 1, 2, ..., n, \tag{5}$$

gde je označeno $\bar{Y} = \beta_0^{'}$.

Ocene grešaka $\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$, $i = 1, 2, ..., n$ date su sa $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Uvedimo oznake: $S_{YY} = \sum (Y_i - \bar{Y})^2$, $R_{xY} = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}$.

Tada se ocenjeni parametri mogu napisati na sledeći način:

$$\hat{\beta}_1 = R_{xY} \sqrt{\frac{S_{YY}}{S_{xx}}} \quad \text{i} \quad \hat{\beta}_0 = \bar{Y} - \bar{x} R_{xY} \sqrt{\frac{S_{YY}}{S_{xx}}}.$$

a jednačina prave u linearnom modelu (5) može zapisati u sledeća tri oblika:

$$\hat{Y}_i = \sqrt{\frac{S_{YY}}{S_{xx}}} R_{xY} x_i + \bar{Y} - \bar{x} \sqrt{\frac{S_{YY}}{S_{xx}}} R_{xY}, \qquad \hat{Y}_i - \bar{Y} = \sqrt{\frac{S_{YY}}{S_{xx}}} R_{xY} (x_i - \bar{x}),$$

$$\frac{\hat{Y}_i - \bar{Y}}{\sqrt{S_{YY}}} = R_{xY} \frac{x_i - \bar{x}}{\sqrt{S_{xx}}} \quad i = 1, 2, ..., n.$$

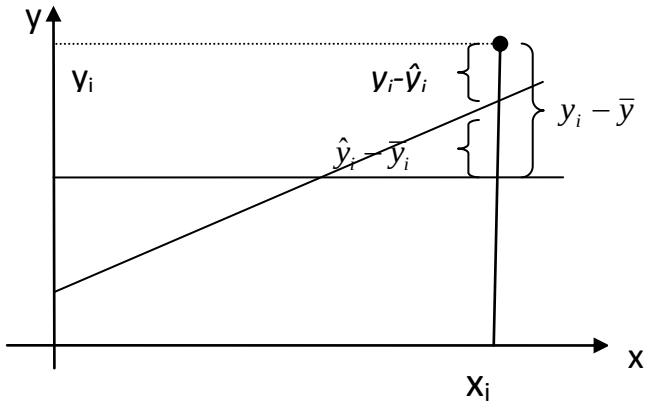
Kada se na osnovu uzorka dobiju realizovane vrednosti zavisno promenljive Y, $y_1, y_2, ..., y_n$, odnosno parovi $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, izračunavaju se realizovane vrednosti

ocena parametara b_0 i b_1, kao i ocenjene vrednosti zavisno promenljive na osnovu modela, $\hat{y}_i$. Takođe, dobijamo i realizovane vrednosti ocena grešaka, $\hat{\varepsilon}$,

$$e_i = y_i - \hat{y}_i .$$

Zamenom registrovanih vrednosti, u uzorku dobijamo realizovane vrednosti ovih ocena, b_0 i b_1 s^2 , ocene za vrednosti za zavisnu promenljivu, $\hat{y}_i = b_0 + b_1 x_i , i = 1,2, \dots , n,$ ocene grešaka $\hat{\varepsilon}_i$, $e_i = y_i - \hat{y}_i$.

U slučaju kada je kada je funkcija μ linearna metod najmanjih kvadrata se sastoji u tome da se od svih pravih koje se mogu povući u ravni, odredi ona za koju će suma kvadrata odstupanja od realizovaih vrednosti zavisno promenljive biti minimalna. Grafički prikaz je dat na slici 3.



Slika 3.

Primer 2 U primeru 1, imamo

$$n = 12, \quad \sum_{i=1}^{12} x_i = 372, \quad \sum_{i=1}^{12} y_i = 166.4, \quad \sum_{i=1}^{12} x_i^2 = 12104, \quad \sum_{i=1}^{12} y_i^2 = 2435.14,$$

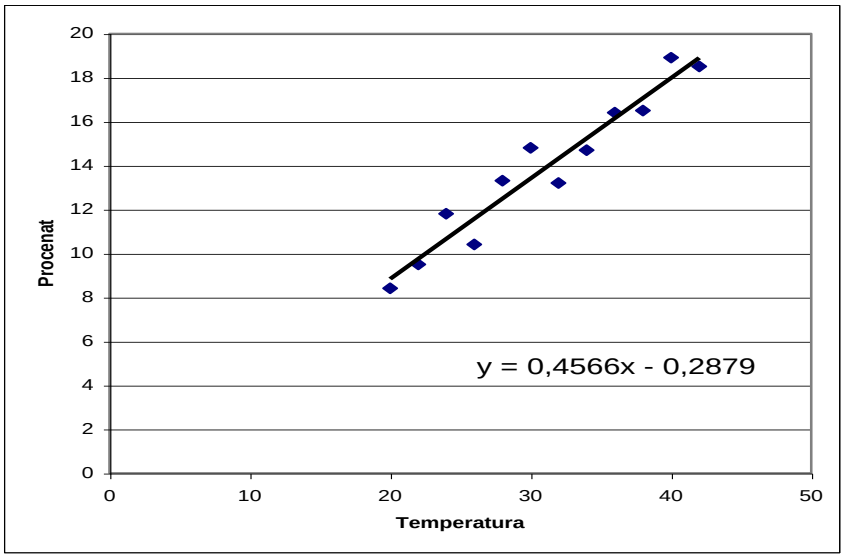
$$\bar{x} = 31, \quad \bar{y} = 13.87, \quad \sum_{i=1}^{12} x_i y_i = 5419.6$$

$$S_{xx} = 572.0, \quad S_{xy} = 261.2, \quad b_0 = -0.2879, \quad b_1 = 0.4566$$

Model je

$$\hat{Y}_i = -0.2879 + 0.4566 x_i , \quad i = 1, 2, ..., n .$$

Grafik ove prave je ucrtan u dijagram rasipanja i prikazan na slici 4.



Slika 4.

Vrednosti nezavisne promenljive, zavisne promenljive, ocenjene vrednosti zavisno promenljive, odstupanja između izmerenih vrednosti zavisne promenljive i pocenjenih vrednosti zavisno promenljive, reziduali, kao i njihovi kvadrati dati su u sledećoj tabeli.

x_i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
20	8,4	8,84	-0,44	0,20
22	9,5	9,76	-0,26	0,07
24	11,8	10,67	1,13	1,28
26	10,4	11,58	-1,18	1,40
28	13,3	12,50	0,80	0,64
30	14,8	13,41	1,39	1,93
32	13,2	14,32	-1,12	1,26
34	14,7	15,24	-0,54	0,29
36	16,4	16,15	0,25	0,06
38	16,5	17,06	-0,56	0,32
40	18,9	17,98	0,92	0,85
42	18,5	18,89	-0,39	0,15

Osobine ocena $\hat{\beta}_0$ i $\hat{\beta}_1$

Posmatrajmo prvo ocenu $\hat{\beta}_1$. Primetimo da je

$$E(\bar{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 \bar{x},$$

$$D(\bar{Y}) = \frac{\sigma^2}{n},$$

jer je $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$, $i = 1, 2, ..., n$. Kako je

$$Y_i - \bar{Y} = \beta_0 + \beta_1 x_i + \varepsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\varepsilon} = \beta_1 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon} \qquad (6)$$

imamo da je

$$E(Y_i - \bar{Y}) = \beta_1 (x_i - \bar{x}).$$

Očekivana vrednost od $\hat{\beta}_1$ je:

$$E(\hat{\beta}_1) = E\left(\frac{S_{xY}}{S_{xx}}\right) = E\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}}\right) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(Y_i - \bar{Y})}{S_{xx}} = \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})^2}{S_{xx}} = \beta_1,$$

odnosno ocena $\hat{\beta}_1$ je nepristrasna. Odmah sledi da je i ocena $\hat{\beta}_0$ nepristrasna.

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1 \bar{x}) = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0.$$

Za određivanje disperzije ovih ocena pođimo od

$$S_{xY} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i (x_i - \bar{x}).$$

Kako su slučajne promenljive ε_i, $i = 1, 2, ..., n$ po pretpostavci modela nekorelirane, sledi da su i slučajne promenljive Y_i, $i = 1, 2, ..., n$ nekorelirane, pa je

$$D(\hat{\beta}_1) = D(\frac{S_{xY}}{S_{xx}}) = \frac{1}{S_{xx}^2} D(S_{xY})$$

$$= \frac{1}{S_{xx}^2} D(\sum_{i=1}^{n} Y_i (x_i - \bar{x})) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 D(Y_i)$$

$$= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$D(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \qquad (7)$$

Uvodeći oznaku

$$h_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{x_i - \overline{x}}{S_{xx}}, \qquad \sum_{i=1}^{n} h_i = 0,$$

imamo

$$\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{S_{xx}} = \sum_{i=1}^{n} h_i (Y_i - \overline{Y})$$

Iz $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$ sledi

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x} = \overline{Y} - \overline{x}\sum_{i=1}^{n} h_i (Y_i - \overline{Y}) = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}h_i\right) Y_i ,$$

$$D(\hat{\beta}_0) = D(\sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}h_i\right) Y_i) = \sigma^2 \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}h_i\right)^2$$

$$D(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}) = \sigma^2(\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}) \qquad (8)$$

.

Kvadratni koreni iz ovih disperzija, standardne devijacije ocean $\hat{\beta}_0$ i $\hat{\beta}_1$, su njihove standardne greške. Primetimo da je $D(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$ manja što je S_{xx} veća. Što se x-sevi više rasipaju, manja je varijacija nagiba.

Na sličan način se može pokazati da je

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{S_{xx}}.$$

Ocena disperzije reziduala σ^2

Ocena za disperziju reziduala σ^2 se može dobiti pomoću ocene $\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ za zbir kvarata reziduala $\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \mu(x_i))^2$.

$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ se naziva i rezidualna suma kvadrata (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i))^2 = S_E = SSE.$$

$$\begin{aligned}
\sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n} (Y_i - \overline{Y} + \hat{\beta}_1 \overline{x} - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y} + \hat{\beta}_1 (\overline{x} - x_i))^2 \\
&= \sum_{i=1}^{n} (\beta_1 (x_i - \overline{x}) + \varepsilon_i - \overline{\varepsilon} + \hat{\beta}_1 (\overline{x} - x_i))^2 \qquad \text{vidi (6)} \\
&= \sum_{i=1}^{n} ((\beta_1 - \hat{\beta}_1)(x_i - \overline{x}) + \varepsilon_i - \overline{\varepsilon})^2 \\
&= \sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \overline{x})^2 + 2(\beta_1 - \hat{\beta}_1) \sum_{i=1}^{n} (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) + \sum_{i=1}^{n} (\varepsilon_i - \overline{\varepsilon})^2 \qquad (9)
\end{aligned}$$

Tražimo $E(\sum_{i=1}^{n} \hat{\varepsilon}_i^2)$. Odredićemo očekivanje svakog od sabiraka u (9)

$$E(\sum_{i=1}^{n} (\beta_1 - \hat{\beta}_1)^2 (x_i - \overline{x})^2) = \sum_{i=1}^{n} (x_i - \overline{x})^2 D(\hat{\beta}_1) = \sum_{i=1}^{n} (x_i - \overline{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \sigma^2 .$$

Iz $\hat{\beta}_1 = \sum_{i=1}^{n} h_i (Y_i - \overline{Y})$ sledi

$$\hat{\beta}_1 = \sum_{i=1}^{n} h_i (Y_i - \overline{Y}) = \sum_{i=1}^{n} h_i (\beta_1 (x_i - \overline{x}) + \varepsilon_i - \overline{\varepsilon}) = \beta_1 + \sum_{i=1}^{n} h_i \varepsilon_i$$

i dobijamo

$$\begin{aligned}
E(2(\beta_1 - \hat{\beta}_1) \sum_{i=1}^{n} (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})) &= -2E(\sum_{i=1}^{n} h_i \varepsilon_i \sum_{j=1}^{n} (x_j - \overline{x})(\varepsilon_j - \overline{\varepsilon})) \\
&= -2E(\sum_{i=1}^{n} \sum_{j=1}^{n} h_i (x_j - \overline{x}) \varepsilon_i \varepsilon_j - \sum_{i=1}^{n} h_i \varepsilon_i \overline{\varepsilon} \sum_{i=1}^{n} (x_i - \overline{x}) \\
&= -2 \sum_{i=1}^{n} h_i (x_i - \overline{x}) \sigma^2 = -2\sigma^2.
\end{aligned}$$

Konačno,

$$E(\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2)=E(\sum_{i=1}^{n}\varepsilon_i^2-2\bar{\varepsilon}\sum_{i=1}^{n}\varepsilon_i+n\bar{\varepsilon}^2)$$
$$=n\sigma^2-2n\frac{1}{n}\sigma^2+n\frac{1}{n}\sigma^2=(n-1)\sigma^2.$$

Sledi $E(\sum_{i=1}^{n}\hat{\varepsilon}_i^2)=\sigma^2-2\sigma^2+(n-1)\sigma^2=(n-2)\sigma^2.$

odnosno da je

$$\widehat{\sigma^2}=s^2=\frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i))^2=\frac{1}{n-2}RSS=\frac{1}{n-2}S_E.$$

centrirana ocena disperzije gešaka σ^2.

Formule (7) i (8) daju disperziju ocena $\hat{\beta}_0$ i $\hat{\beta}_1$, koeficijenata β_0 i β_1 u linearnom modelu, ali se u njima pretpostavlja da je σ^2 poznato, što je retko kad slučaj. Pomoću ocene za σ^2, mogu se dobiti ocene za ove disperzije:

$$\widehat{D(\hat{\beta}_1)}=\frac{s^2}{S_{xx}},\qquad \widehat{D(\hat{\beta}_0)}=s^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}).$$

Ocenjene standardne greške koeficijenata

$$SE(\hat{\beta}_0)=s_{\hat{\beta}_0}=\sqrt{\widehat{D(\hat{\beta}_0)}}=\sqrt{s^2(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}})},$$
$$SE(\hat{\beta}_1)=s_{\hat{\beta}_1}=\sqrt{\widehat{D(\hat{\beta}_1)}}=\sqrt{\frac{s^2}{S_{xx}}},$$

gde je s^2 gornja ocena za σ^2,

$$s^2=\widehat{\sigma^2}=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i))^2=\frac{1}{n-2}RSS.$$

U testiranju hipoteza o koeficijentu linearnom modelu i određivanju intervala poverenja koriste se standardne devijacije ocena $s_{\hat{\beta}_0}$ i $s_{\hat{\beta}_1}$.

Sa slike 3. vidi se da se ukupno odstupanje jedne registrovane vrednosti promenljive y_i od srednje vrednosti $\bar{y}$ može podeliti na: objašnjeno odstupanje $\hat{y}_i-\bar{y}$, odstupanje objašnjeno modelom i neobjašnjeno odstupanje $y_i-\hat{y}_i$ registrovanih vrednosti od vrednosti određenih modelom. Slično razlaganje važi i za kvadrate ovih odstupanja, odnosno važi

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2.$$

Imamo da je

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+2\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2;$$

a kako je

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)x_i=\frac{\partial F}{\partial \beta_1}=0 \quad i \quad \sum_{i=1}^{n}(Y_i-\hat{Y}_i)=\frac{\partial F}{\partial \beta_0}=0\,,$$

imamo da je

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y}) &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{\beta}_1 x_i+\hat{\beta}_0-\overline{Y})\\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{\beta}_1 x_i+\overline{Y}-\hat{\beta}_1\overline{x}-\overline{Y})=\\ &= \hat{\beta}_1\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(x_i-\overline{x})\\ &= \hat{\beta}_1[\sum_{i=1}^{n}(Y_i-\hat{Y}_i)x_i-\overline{x}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)]=0.\end{aligned}$$

Sledi

$$S_{YY}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\,.$$

Ako uvedemo oznake

$$S_R=\sum_{i=1}^{n}(\hat{Y}-\overline{Y})^2,\quad S_E=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

tada je

$$S_{YY}=S_R+S_E\,.$$

Zbir kvadrata $S_E=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ se može izraziti i na drugi način.

$$\begin{aligned}S_E &= \sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{i=1}^{n}(\hat{\beta}_1 x_i+\hat{\beta}_0-\overline{Y})^2\\ &= \sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{i=1}^{n}(\hat{\beta}_1 x_i+\overline{Y}-\hat{\beta}_1\overline{x}-\overline{Y})^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2-\sum_{i=1}^{n}(\hat{\beta}_1 x_i-\hat{\beta}_1\overline{x})^2\\ &= \sum_{i=1}^{n}(Y_i-\overline{Y})^2-\hat{\beta}_1{}^2\sum_{i=1}^{n}(x_i-\overline{x})^2=\\ &= S_{YY}-\hat{\beta}_1{}^2 S_{xx}=S_{YY}-R_{xY}^2\,\frac{S_{YY}}{S_{xx}}S_{xx}\\ &= S_{YY}(1-R_{xY}^2),\end{aligned}$$

Pokazalii smo da je $s^2 = \frac{1}{n-2} S_E = \frac{1}{n-2} RSS$ centrirana ocena za σ^2. Može se pokazati da S_E ima χ^2 raspodelu sa $n-2$ stepeni slobode. Takođe, može se pokazati da S_R ima necentriranu χ^2 raspodelu sa jednim stepenom slobode i da je , kako i da su S_E i S_R nezavisne. Odatle sledi da ako je hipoteza $H_0(\beta_1 = 0)$ tačna, statistika

$$F_0 = \frac{S_R / 1}{S_E / (n-2)}$$

ima Fišerovu $F_{1,n-2}$ raspodelu. Pomoću ove statistike može se testirati hipoteza $H_0(\beta_1 = 0)$. Ako je alternavivna $H_1(\beta_1 > 0)$, H_0 hipoteza se odbacuje ako je $F_0 > F_{1,n-2,\alpha}$. rezultati se prikazuju u tabeli.

Izvor varijacije	Zbir kvadrata	Stepeni slobode	Sredina zbira kvadrata	Realizovana vrednost	Tablična vrednost
Regresija	S_R	1	S_R	$F_0 = S_R / S_E (n-2)$	$F_{1,n-2,\alpha}$
Reziuali	S_E	$n-2$	$S_E / n-2$		
Total	S_{YY}	$n-1$			

Testiranje hipoteza i intervali poverenja u za linearnu regresiju

Ako se u modelu (4) uved i pretpostavka o raspodeli reziduala, moguće je testirati hipoteze o vrednosti oba parametra, kao i odrediti odgovarajuće intervale poverenja.

Pretpostavimo da reziduali imaju normalnu raspodelu odnosno da je $\varepsilon_i : \mathrm{N}(0, \sigma^2)$. Ova pretpostavka, zajedno sa $\mathrm{cov}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$ znači da su reziduali nezavisne slučajne promenljive. Iz petpostavke da reziduali imaju normalnu raspodelu, sledi da i promenljive Y_i imaju normalnu raspodelu. Dalje, kako su ocene $\hat{\beta}_0$ i $\hat{\beta}_1$ linearne funkcije slučajnih promenljivih Y_i, sledi da i one imaju normalnu raspodelu:

$$\hat{\beta}_0 : \mathrm{N}(\beta_0, \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n S_{xx}}) \quad \hat{\beta}_1 : \mathrm{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$$

Takođe, može se pokazati da ocena disperzije grešaka ima hi-kvadrat raspodelu, odnosno da

$$\frac{n-2}{\sigma^2}s^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \ : \ \chi^2_{n-2},$$

kao i da je s^2 nezavisna od $\hat{\beta}_0$ i $\hat{\beta}_1$, vidi [Magnus]. Sada se može lako pokazati da

$$(\hat{\beta}_1 - \beta_1)/\sqrt{\frac{\sigma^2}{S_{xx}}} : \mathrm{N}\ (0,1)$$

i

$$t = \frac{\frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{\sigma^2}{S_{xx}}}}}{\sqrt{\frac{\frac{n-2}{\sigma^2}s^2}{n-2}}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\frac{s^2}{S_{xx}}}} = \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\hat{D}\left(\hat{\beta}_1\right)}} \ : \ t_{n-2}$$

odnosno da statistika $\frac{(\hat{\beta}_1 - \beta_1)}{s_{\hat{\beta}_1}}$ ima Studentovu raspodelu sa $n-2$ stepena slobode.

Slično se pokazuje da i $\frac{(\hat{\beta}_0 - \beta_0)}{s_{\hat{\beta}_0}}$ ima Studentovu raspodelu sa $n-2$ stepena slobode.

Interval poverenja kod linearne regresije

- Interval poverenja za **srednju vrednost** μ_0 promenljive Y za datu vrednost x_0 promenljive x na nivou γ je

$$\hat{Y}(x_0) - t \cdot S_{\hat{Y}} < \mu_0 < \hat{Y}(x_0) - t \cdot S_{\hat{Y}},$$

gde je $\hat{Y}(x_0)$ vrednost linearne regresije u tački x_0, t je kvantil reda $(1+\gamma)/2$ Studentove raspodele sa $n-2$ stepena slobode, a $S_{\hat{Y}}$ je standardna greška za $\hat{Y}(x_0)$.

$$S_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- Interval poverenja (interval predikcije) za pojedinačnu vrednost $Y(x_0)$ promenljive Y za datu vrednost x_0 promenljive x na nivou γ je

$$\hat{Y}(x_0) - t \cdot S_Y < Y(x_0) < \hat{Y}(x_0) - t \cdot S_Y,$$

gde je $\hat{Y}(x_0)$ vrednost linearne regresije u tački x_0, t je kvantil reda $(1+\gamma)/2$ Studentove raspodele sa $n-2$ stepena slobode, a S_Y je standardna greška za $Y(x_0)$.

$$S_{\hat{Y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Testiranje značajnosti parametara modela

Značajnost koeficijenata regresionog modela $Y = \beta_0 + \beta_1 x + \varepsilon,$ vrši se pomoću odgovarajućih statistika.

Testiranje hipoteze o vrednosti parametra β_0, $H_0(\beta_0 = 0)$ protiv alternativne $H_1(\beta_0 \neq 0)$ vrši se pomoću statistike

$$T = \frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}},$$

koja ima Studentovu raspodelu na $n-2$ stepena slobode. Ovde je $s_{\widehat{\beta}_0}$ ocena standardne greške parametra β_0:

$$s_{\widehat{\beta}_0} = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}, \quad .$$

Testiranje hipoteze

H_0 (ne postoji veza između X i Y) , H_1 (postoji veza između X i Y).

za prostu linearnu regresiju ovo se svodi na

H_0 ($\beta_1 = 0$) , H_1 ($\beta_1 \neq 0$).

Test statistika

$$T = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}},$$

koja ima Studentovu raspodelu na $n-2$ stepena slobode. Ovde je $s_{\widehat{\beta}_1}$ ocena standardne greške parametra β_1:

$$s_{\widehat{\beta}_1} = \sqrt{\frac{s^2}{S_{xx}}}.$$

Primer 4 U primeru 1 je

i	x_i	y_i	$(x_i-31)^2$	$\hat{y}_i$	$y_i-\hat{y}_i$	$(y_i-\hat{y}_i)^2$
1	20	8,4	121	8,84	-0,44	0,20
2	22	9,5	81	9,76	-0,26	0,07
3	24	11,8	49	10,67	1,13	1,28
4	26	10,4	25	11,58	-1,18	1,40
5	28	13,3	9	12,50	0,80	0,64
6	30	14,8	1	13,41	1,39	1,93
7	32	13,2	1	14,32	-1,12	1,26
8	34	14,7	9	15,24	-0,54	0,29
9	36	16,4	25	16,15	0,25	0,06
10	38	16,5	49	17,06	-0,56	0,32
11	40	18,9	81	17,98	0,92	0,85
12	42	18,5	121	18,89	-0,39	0,15
Σ	372		572		0	8,45

Tabela 8.6

Imamo da je $\overline{x} = 372/12 = 31$, $\sum_{i=1}^{n}(x_i-\overline{x})^2$ standardna greška linearne regresije $s = 0,919$, pa je

a) Interval poverenja za **srednju vrednost** μ_0 promenljive Y za datu vrednost $x_0 = 29$ promenljive x na nivou $\gamma = 0,95$ je

$$\hat{y}(x_0) - t \cdot s_{\hat{Y}} < \mu_0 < \hat{y}(x_0) - t \cdot s_{\hat{Y}},$$

gde je $\hat{y}(29) = -0,2879 + 0,4566 \cdot 29 = 12,9535$ vrednost linearne regresije u tački $x_0 = 29$, $t = 2,28$ je kvantil reda $0,975$ Studentove raspodele sa $n-2 = 10$ stepeni slobode, a $S_{\hat{Y}}$ je standardna greška za $\hat{Y}(x_0)$.

$$s_{\hat{Y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 0,919 \cdot \sqrt{\frac{1}{12} + \frac{(29-31)^2}{572}} = 0,276.$$

Sledi $12,3242 < \mu_0 < 13,5827$.

b) Interval poverenja za pojedinačnu vrednost $Y(x_0)$ promenljive Y za datu vrednost $x_0 = 29$ promenljive x na nivou $\alpha = 0,95$ je

$$\hat{y}(x_0) - t \cdot s_Y < Y(x_0) < \hat{y}(x_0) - t \cdot s_Y,$$

gde je $\hat{y}(29) = 12,9535$ vrednost linearne regresije u tački $x_0 = 29$, $t = 2,28$ je kvantil reda $0,975$ Studentove raspodele sa $n-2 = 10$ stepena slobode, a s_Y je standardna greška za $\hat{Y}(x_0)$,

$$s_Y = s \cdot \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} = 0{,}919 \cdot \sqrt{1+\frac{1}{12}+\frac{(29-31)^2}{572}} = 0{,}9596\,.$$

Sledi $10{,}7656 < Y(x_0) < 15{,}1414\,.$

c) $s_{\hat{\beta}_0} = 0{,}919 \cdot \sqrt{\frac{12104}{12 \cdot 572}} = 1{,}22$, $t = \frac{-0{,}2879}{1{,}22} = -0{,}0236$. Ovu vrednost upoređujemo sa kvantilom reda $1-\alpha/2 = 0{,}975$ Studentove raspodele sa $n-2=10$ stepeni slobode, $c = 2{,}228$. Kako je $-2{,}228 < -0{,}0236 < 2{,}228$, ne odbacujemo hipotezu $H_0(\beta_0 = 0)$.

d) $s_{\hat{\beta}_1} = 0{,}919 \cdot \sqrt{\frac{1}{572}} = 0{,}038$, $t = \frac{0{,}4566}{0{,}038} = 12{,}016$. Ovu vrednost upoređujemo sa kvantilom reda $1-\alpha/2 = 0{,}975$ Studentove raspodele sa $n-2=10$ stepeni slobode, $c = 2{,}228$. Kako je $12{,}016 > 2{,}228$, odbacujemo hipotezu $H_0(\beta_1 = 0)$.

Primer sa prodajom i reklamom:

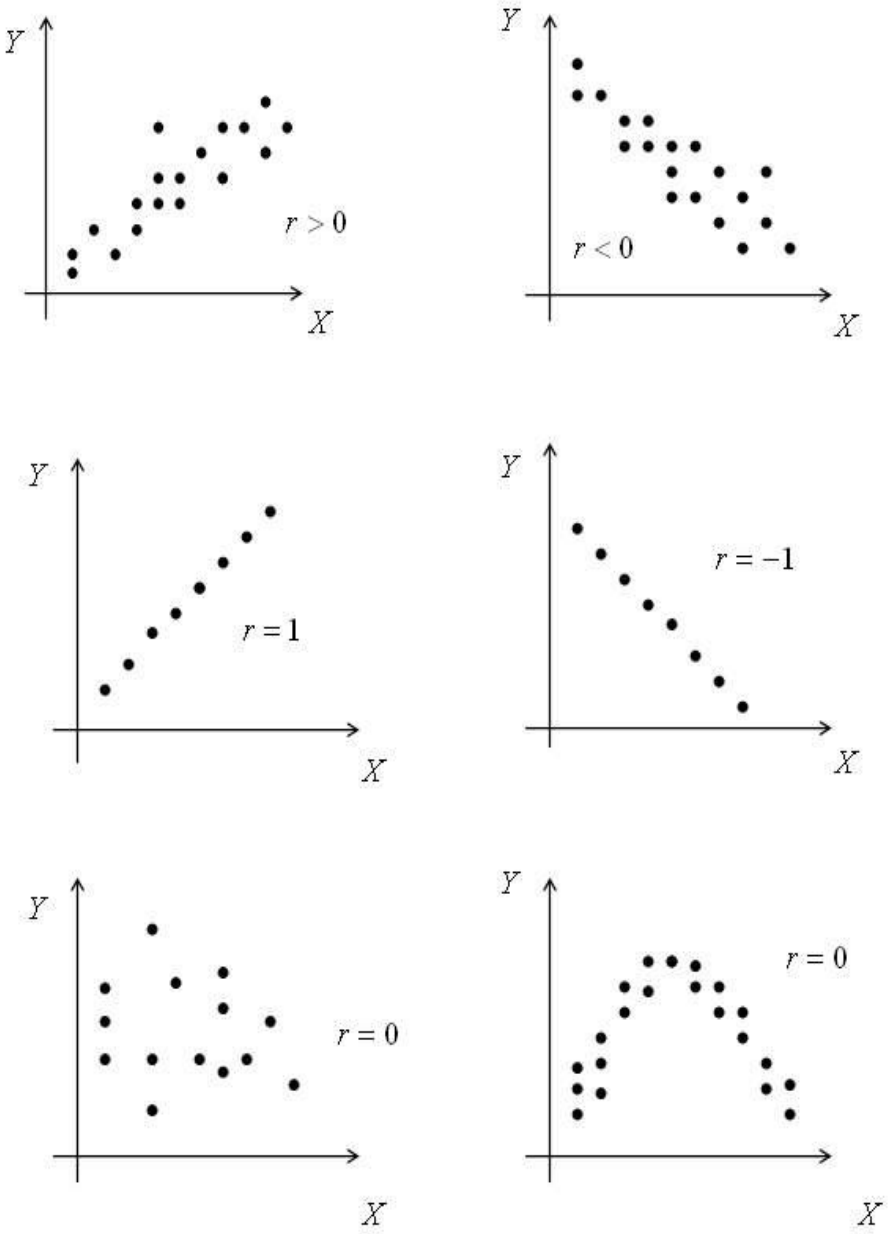
N=200	Regression Summary for Dependent Variable: Sales (Advertising) R= .78222442 R²= .61187505 Adjusted R²= .60991482 F(1,198)=312.14 p<0.0000 Std.Error of estimate: 3.2587					
	b*	Std.Err. of b*	b	Std.Err. of b	t(198)	p-value
Intercept			7.032594	0.457843	15.36028	0.000000
TV	0.782224	0.044274	0.047537	0.002691	17.66763	0.000000

Povećanje od 1000$ za reklamu preko TVa povezano je sa povećanjem prodaje od oko 50 jedinica (0,0475 x 1000).

Ocenjivanje koeficijenta korelacije.

Za ispitivanje lineane povezanosti dva obeležja X i Y, koristi se Pirsonov (Karl Pearson) koeficijent linearne korelacije $\rho(X,Y)$:

$$\rho(X,Y) = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{D(X)D(Y)}} = \frac{E(XY)-E(X)E(Y)}{\sqrt{D(X)D(Y)}}.$$



Slika Različiti slučajevi korelacije izmedju promenljivih

Karl Pirson (Karl Pearson, 1857-1936) engleski matematičar i statističar, koji se bavio isključivo matematičko – statističkom obradom bioloških problema, primenio je metode statističke analize da bi dokazao Galtonovu teoriju. U radovima Karla Pirsona prvi put nailazimo na termin korelacija. Korelacija potiče od latinske reči "correlatia" i znači medjusobni odnos i zavisnost, stvari koje stoje u uzajamnom odnosu. Od tada su regresiona i korelaciona analiza opšte prihvaćeni statistički metodi putem kojih se ispituju zavisnosti izmedju pojava, odnosno predvidja i ocenjuje jedna pojava na osnovu vrednosti neke druge pojave (ili grupa pojava).

Iako su korelacija i regresija matematički vrlo usko povezani, ipak postoji značajna razlika. Naime, kod regresione analize se ispituju zavisnosti izmedju dve ili više promenljivih tj.

sagledava se uticaj promene jedne ili više promenljivih na promenu drugih promenljivih i potrebno je unapred odrediti koja pojava će imati ulogu zavisne promenljive a koja nezavisne promenljive. Ovo se utvrdjuje na osnovu prethodnih empirijskih saznanja. Kod korelacione analize se ispituje samo medjuzavisnost ili veza izmedju dve promenljive i prilikom ispitivanja dve pojave svejedno je koja se klasifikuje kao nezavisna a koja kao zavisna promenljiva

Ocena koeficijenta linearne korelacije dobija se osnovu uzorka obima n:

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n).$$

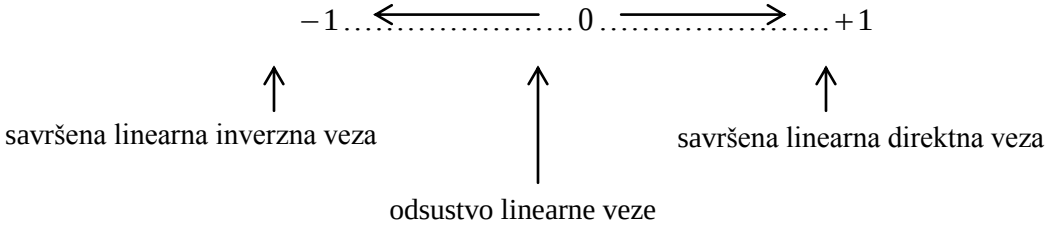
Uzorački koeficijent korelacije je

$$R_{XY} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \bar{X}_n \bar{Y}_n}{\bar{S}_X \bar{S}_Y}$$

Na osnovu registrovanih parova vrednosti $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ ocenjuje se vrednost koeficijenta korelacije $\rho(X, Y)$.

$$r = \frac{\sum_{i=1}^{n} x_i y - n\bar{x} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}.$$

Vrednost koeficijenta linearne korelacije se kreće u intervalu od -1 do $+1$, i to kod pozitivne korelacije se kreće izmedju 0 i 1 a kod negativne korelacije izmedju 0 i –1. Ako koeficijent korelacije uzima pozitivne vrednosti, tada je korelacija direktna ili pozitivna, vrednost 1 ukazuje na savršenu pozitivnu korelaciju i tada pojave istovremeno rastu tj. postoji funkcionalna veza, gde svakoj promeni jedne promenljive odgovara tačno odredjena promena druge promenljive. U slučaju da je koeficijent korelacije negativan, korelacija je inverzna ili negativna, tada jedna pojava raste a druga opada ili obrnuto a vrednost –1 ukazuje na savršenu negativnu korelaciju. Medjutim, u nekim ekstremnim slučajevima postoji i odsustvo korelacije a to se dešava kada je vrednost koeficijenta korelacije jednak nuli. Ipak, kada je $r = 0$, to pre treba protumačiti kao odsustvo linearne korelacije nego korelacije uopšte. To je iz tog razloga što u takvom slučaju izmedju pojava možda postoji neki oblik krivolinijskog slaganja.



Sledeća tabela nam ukazuje na empirijsko pravilo tumačenja vrednosti koeficijenta korelacije:

Tabela

Vrednost r	Tumačenje
1,00	Funkcionalna pozitivna veza
0,99 – 0,95	Vrlo jaka pozitivna veza
0,94 – 0,75	Jaka pozitivna veza
0,74 – 0,50	Srednje jaka pozitivna veza
0,49 – 0,25	Slaba pozitivna veza
0,24 – 0,01	Neznatna pozitivna veza
0,00	Odsustvo linearne veze
$-0,01$ – $-0,25$	Neznatna negativna veza
$-0,26$ – $-0,50$	Slaba negativna veza
$-0,51$ – $-0,75$	Srednje jaka negativna veza
$-0,76$ – $-0,95$	Jaka negativna veza
$-0,96$ – $-0,99$	Vrlo jaka negativna veza
$-1,00$	Funkcionalna negativna veza

Medjutim, treba naznačiti da postoji mogućnost da u tumačenju datih vrednosti razlike ipak variraju u zavisnosti od oblasti istraživanja posmatranih pojava.

Primer Odrediti koeficijent korelacije za promenljive čije su registrovane vrednosti date u sledećoj tabeli

Redni broj	Dnevna zarada X	Godine školovanja Y
1	8,5	12
2	12,0	14
3	9,0	10
4	10,5	12
5	11,0	16
6	15,0	16
7	25,0	18
8	12,0	18
9	6,5	12
10	8,25	10

Tabela 8.1

Potrebna izračunavanja su data u sledećoj tabeli

Redni broj	X	Y	XY	X^2	Y^2
1	8,50	12	102	72,25	144
2	12,00	14	168	144,00	196
3	9,00	10	90	81,00	100
4	10,00	12	120	100,00	144
5	11,00	16	176	121,00	256
6	15,00	16	240	225,00	256
7	25,00	18	450	625,00	324
8	12,00	18	216	144,00	324
9	6,50	12	78	42,25	144
10	8,25	10	82,5	68,06	100
Σ	117,25	138,00	1722,50	1622,56	1988,00

Tabela 8.2

Imamo da je $\overline{x} = 11,75,\ \overline{y} = 13,8 \quad r = 0,726$.

Testiranje hipoteze o koeficijentu korelacije.

Testiranje hipoteze o koeficijentu korelacije $H_0(\rho = 0)$ (među promenljivima ne postoji linearna korelacija) protiv alternativne $H_1(\rho \neq 0)$ (među promenljivima postoji linearna korelacija značajno različita od nule) vrši se pomoću statistike

$$T = R\sqrt{\frac{n-2}{1-R^2}},$$

koja ima Studentovu raspodelu na $n-2$ stepena slobode.

Primer Za podatke u primeru 8.1 testirati hipotezu o koeficijentu korelacije na nivou značajnosti $\alpha = 0,05$.

Realizovana vrednost test statistike je $t = 2,986$. Kako je reč o dvostranom testu, ovu vrednost upoređujemo sa kvantilom reda $1-\alpha/2 = 0,975$ Studentove raspodele sa $n-2 = 8$ stepeni slobode, $c = 2,306$. Kako je $t = 2,986 > 2,306 = c$, odbacujemo hipotezu $H_0(\rho = 0)$.

Procena tačnosti modela.

Mere kvaliteta fitovane krive - koliko se predviđene vrednosti slažu sa registrovanim podacima.

Srednja kvadratna greška (Mean Square Error, MSE)

$$training \quad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

MSE će biti mala ako su pedviđeni odgovori bliski registrovanim odgovorima. Ovo je training MSE. *MSE = (1/n)RSS.*

Međutim, nas ne interesuje koliko je model dobar sa poznatim, registrovanim podacima, interesije nas koliko će biti dobar kada ga primenimo na još neviđene, podatke – test data,

$$(\check{x}_1, \check{y}_1), (\check{x}_2, \check{y}_2),\ldots, (\check{x}_k, \check{y}_k).$$

Ako napravimo model za predviđanje cene akcije pomoću istorijskih podataka, ne intresuje nas koliko je on dobar za stare podatke, nego kolio dobro predviđa buduće cene akcije. Zato nas interesuje test MSE

$$test \quad MSE = \frac{1}{k}\sum_{i=1}^{k}(\check{y}_i - \hat{f}(\check{x}_i))^2.$$

Kvalitet modela se obično opisuje pomoću rezidualne standardne greške RSE i pomoću koeficijenta determinacije R^2.

Rezidualna standardna greška je

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

Ovo je ocena standardne devijacije od ε. Govori koliko predviđanja odstupaju od registrovanih podataka. RSE se posmatra kao mera nedostatka fita (lack of fit) modela prema podacima. Ako je $y_i \approx \hat{y}_i$ RSE je mala. Ona je Apsolutna mera nedostatka fita

U primeru: Stvarna prodaja na tržištima u proseku odstupa od stvarne regresione linije za 3,26 jedinica. Ili na drugi način, kad bi bile poznate vrednosti parametara β_0, β_1, svako predviđanje prodaje na osnovu ulaganja u TV reklamu bi ipak u proseku odstupalo za 3,26 jedinica od stvarne vrednosti, zbog slučajne greške ε.

Koeficijent determinacije R^2

$$R^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{S_{YY}-S_E}{S_{YY}}$$

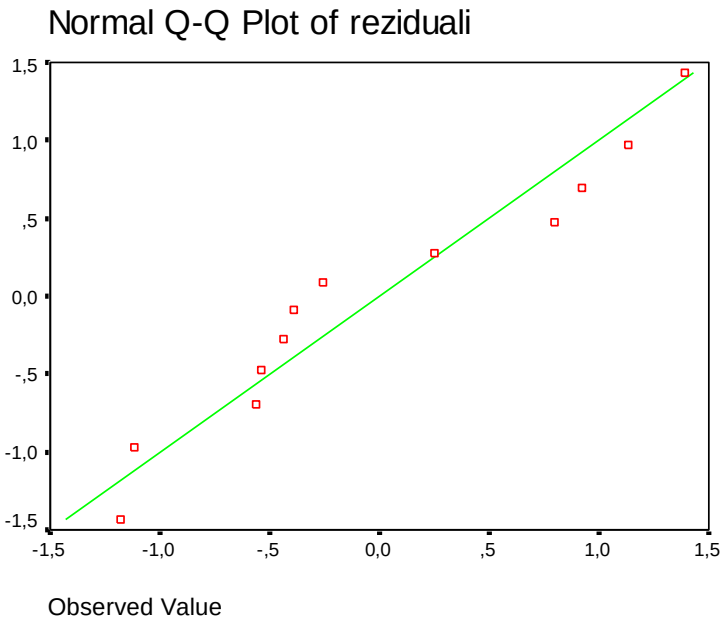
Relativna mera fitovanja. Uvek između 0 i 1. Nezavisna od jedinica mere za *Y*.
To je proporcija objašnjene varijanse. Mera ukupne varijanse za Y je S_{YY} i može se posmatrati kao količina varjanse koja postoji u Y pre nego sto je uvedena regresija. RSS meri varijansu koja je ostala neobjašnjena posle uvođenja regresije. Zbog toga $S_{YY} - S_E = S_{YY} - RSS$, meri varijansu koja je objašnjena regresijom. Zato R^2 meri proporciju, udeo varijanse od Y koja se može objasniti regresijom, objasniti korišćenjem promenljive *X*.

U primeru: R^2 je 0,6, pa se dve trećina varijanse količine prodatih proizvoda može objasniti ulaganjem u reklamu preko TV.

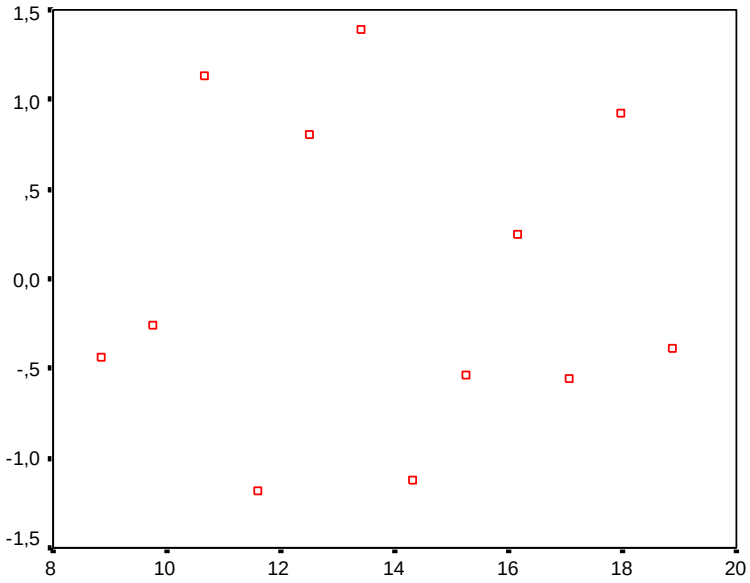
Koliki R^2 je dobar R^2 zavisi od primene U nekim oblastima kao što je hemija ili fizika R^2 je blizu 1, a u medicini R^2 od 0,6 se smatra odličnim.
U prostoj linearnoj regresiji je R^2 jednak kvadratu koeficijenta korelacije.

Analiza reziduala.

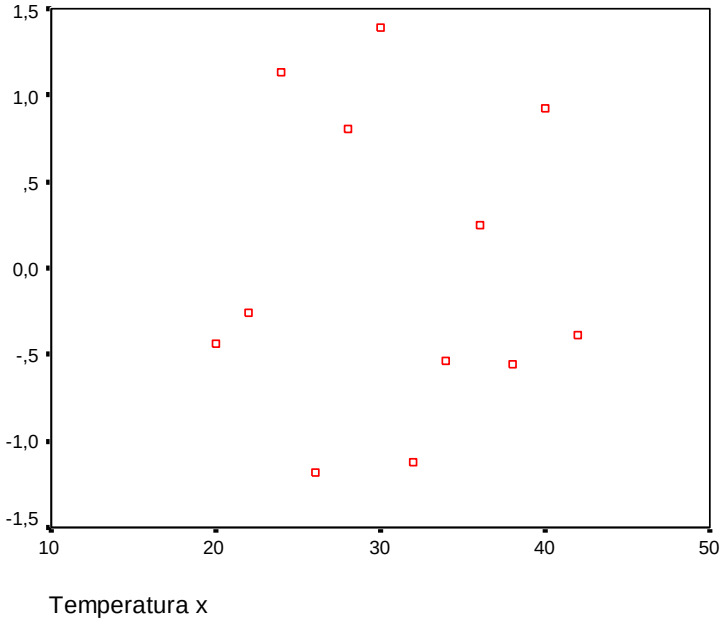
Prilikom formiranja linearnog modela, potrebno je izvršiti analizu reziduala, da bi se uvrdilo koliko je model odgovarajući, koliko dobro je prilagođen podacima. Obično se ispituje da li reziduali imaju normalnu raspodelu, upoređuju se reziduali sa fitovanim vrednostima, kao i sa svakom od nezavisnih promenljivih u modelu. Takođe, ako postoje promenljive koje nisu uključene u model, a mogu biti interesantne, i njih treba uporediti sa rezidualima. Upoređivanje se vrši tako što se crtaju dijagrami rasturanja, i ako se na njima može uočiti neka pravilnost, model je moguće još poboljšati.



Slika 5



Slika 6



Slika 7

Literatura

1. Z. Lozanov Crvenković, Statistika PMF, 2012.
2. Julian J. Faraway, *Practical Regression and Anova using R*, 2002, http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf
3. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning, with Applications in R*, Springer, 2013, http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
4. Douglas C. Montgomery, *Design and analysis of experiments*, John Wiley & Sons Inc. 2001.
5. Andy Field, *Discovering Statistics using IBM SPSS Statistics*, SAGE, 2012, http://www.uk.sagepub.com/field4e/main.htm
6. N. R. Draper, H. Smith,1998: *Applied regression analysis.* Wiley-Interscience, New-York, 736 pp.
7. J. Kmenta,1997: *Počela ekonometrije.* Mate, Zagreb, 787 str.
8. S. Hadživuković, 1979*: Statistika.* Rad, Beograd, 244 str.