

STATISTIKA LINEARNA REGRESIJA

Testiranje zavisnosti - Parametarski i neparametarski

Ne pruža mogućnost ocenjivanja ili predviđanja vrednosti jedne pojave na osnovu poznavanja vrednosti neke druge.

Tu uvodimo linearni regresioni model.

Posmatramo dve promenljive X i Y .

Moramo definisati (za razliku od zavisnosti) zavisnu i nezavisnu slučajnu promenljivu.

Na zavisnu promenljivu Y utiče nezavisna X .

Primer – od primanja u domaćinstvu će zavisiati potrošnja

Dijagram raspršenosti (tabela i slika)!!

Cilj je da nadujemo linearnu zavisnost između X i Y .

Tražimo pravu koja najbolje reprezentuje (aproksimira) izmerene vrednosti – ta prava je linija regresije.

Neka je obeležje Y ***zavisna promenljiva***, a X ***kontrolisana (nezavisna) promenljiva***.

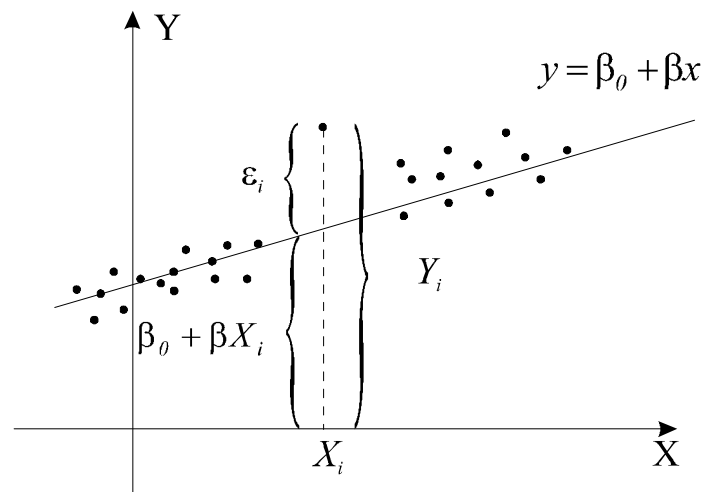
Linearni statistički model će imati oblik (na populaciji)

$$Y_i = \alpha X_i + \beta + \varepsilon_i$$

Promenljive ε_i su slučajne promenljive (***slučajan šum - greška***), koje odgovaraju elementima uzorka. (poremećaj, stohastički član)

Deterministička (tačno jedna vrednost - površina) i stohastička (vrednosti se uzimaju sa verovatnoćama) veza. Kod stohastičke veze može doći do znatnih odstupanja Y od proseka..

Izmerene vrednosti u uzorku su X i Y , i predstavljaju roj tačaka u ravni XOY ,



Na uzorku će linija regresije – regresiona prava biti oblika:

$$\hat{Y}_i = \hat{\alpha} X_i + \hat{\beta}$$

Pošto ne znamo koliko su α i β cilj nam je da ih ocenimo na uzorku.
To radimo metodom najmanjih kvadrata.

Metoda najmanjih kvadrata

Definišemo odstupanja empirijskih tačaka (dobijenih vrednosti) od regresione prave. Ta odstupanja su vertikalna odstupanja između stvarne i ocenjene vrednosti y i nazivaju se reziduali.

Rezidual: $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} X_i + \hat{\beta})$

Koeficijente α i β ćemo oceniti tako da minimiziramo sumu kvadrata odstupanja empirijskih tačaka (dobijenih vrednosti) od regresione prave.

$$\sum_{i=1}^n \left[Y_i - (\hat{\alpha} X_i + \hat{\beta}) \right]^2 = F(\alpha, \beta),$$

Odredićemo vrednosti α i β tako da ova funkcija ima najmanju moguću vrednost, tj. ocenićemo ih tako da ukupno kvadratno odstupanje bude minimalno.

Izjednačavanjem parcijalnih izvoda funkcije F sa nulom

$$\frac{\partial F(\alpha, \beta)}{\partial \alpha} = 0 \qquad \frac{\partial F(\alpha, \beta)}{\partial \beta} = 0 \qquad \dots$$

Dobijamo parametre:

$$\hat{\alpha} = \frac{\frac{1}{n} \sum X_i Y_i - \bar{x}\bar{y}}{S_x^2}$$

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

α – pokazuje prosečnu promenu zavisne promenljive Y kada se nezavisna poveća za jednu svoju jedinicu.

To je nagib prave. Pokazuje zavisnost Y od X. Može biti direktna i inverzna veza.

β - odsečak na Y osi

$$E(\hat{\alpha}) = \alpha$$

$$E(\hat{\beta}) = \beta$$

- ocene su nepristrasne, najbolje i najefikasnije
- Imaju Normalnu ili približno Normalnu raspodelu

$$\hat{Y}_i = \hat{\alpha}X_i + \hat{\beta}$$

$$E(\hat{Y}_i) = \alpha X_i + \beta$$

ε_i - slučajan šum

$$E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2,$$

$$\text{Var}(Y_i) = \sigma^2 - \text{konstanta}$$

Očekivane vrednosti slučajnih promenljivih ε_i jednake nuli, a varijanse su sve međusobno jednake i iznose σ^2 (homoskedastičnost), a $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ su nezavisne slučajne promenljive.

Varijansa od epsilon je isto što i varijansa od Y (rezidualna varijansa)

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum Y_i X_i - \hat{\beta} \sum Y_i \right)$$

Standardna greška regresije – koren iz varijanse promenljive Y

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Testiranje Hipoteza

$$H_0(\alpha = 0)$$

$$H_1(\alpha > 0)$$

desnostrana

$$H_1(\alpha < 0)$$

levostrana

$$H_1(\alpha \neq 0)$$

obostrana

$$\tau = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2(n-2)} : t_{n-2}$$

$$H_0(\beta = \beta_0)$$

$$H_1(\beta > \beta_0)$$

desnostrana

$$H_1(\beta < \beta_0)$$

levostrana

$$H_1(\beta \neq \beta_0)$$

obostrana

$$\tau = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \sqrt{S_x^2 + \bar{X}^2}} \sqrt{S_x^2(n-2)} : t_{n-2}$$

VAŽNO: X i Y su uvek kontinualnog tipa
Prvo proveriti iz teksta zadatka šta je X a šta Y pa onda raditi zadatak

ZADACI:

- Komadi određene vrste rude izdvojeni su na osnovu njihove specifične gustine (X) i u njima je merena sadržina gvožđa (Y). Dobijeni su sledeći rezultati

X	2.8	2.9	3.0	3.1	3.2	3.2	3.2	3.3	3.4
Y	27	23	30	28	30	32	34	33	30

Oceniti jednačinu linearne regresije sadržine gvožđa u odnosu na specifičnu gustinu.

Rešenje:

	x	y	x ²	y ²	xy
	2.80	27.00	7.84	729	75.6
	2.90	23.00	8.41	529	66.7
	3.00	30.00	9	900	90
	3.10	28.00	9.61	784	86.8
	3.20	30.00	10.24	900	96
	3.20	32.00	10.24	1024	102.4
	3.20	34.00	10.24	1156	108.8
	3.30	33.00	10.89	1089	108.9
	3.40	30.00	11.56	900	102
suma	28.10	267.00	88.03	8011	837.2
prosek	3.12	29.67	9.781111	890.1111	93.02222

$$\bar{X} = 3.12 \quad \bar{Y} = 29.67$$

$$\hat{\alpha} = \frac{\frac{1}{n} \sum X_i \cdot Y_i - \bar{X} \cdot \bar{Y}}{S_x^2} = 12,067$$

$$\hat{\beta} = \bar{Y} - \hat{\alpha} \bar{X} = 29.67 - 12,067 \cdot 3.12 = -8,011$$

$$\hat{Y} = 12,067X - 8,011$$

2. Pretpostavlja se da između broja kupaca (Y) i broja ponuđenih artikala (X) u prodavnicama postoji linearna veza. U cilju provere te pretpostavke, izvršena su merenja u 20 prodavnica i metodom najmanjih kvadrata dobijeno je sledeće:

$$\hat{a} = 1.2 \quad \hat{b} = 0.8 \quad \hat{\sigma} = 19.98 \quad \text{и} \quad S_x^2 = 9.86$$

Da li se, na osnovu dobijenih podataka, a sa rizikom od 5%, može smatrati da postoji linearna veza takva da, kad se povećava broj ponuđenih artikala, raste i broj kupaca? Formulirati hipoteze.

Rešenje:

$$\hat{Y} = 1.2X + 0.8$$

$$H_0(\alpha = 0)$$

$$H_1(\alpha > 0)$$

$$\tau = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2(n-2)} : t_{n-2}$$

$$\tau = \frac{1,2}{19,98} \sqrt{9,86 \cdot 18} = 0,8$$

$$P\{\tau > t_0\} = \alpha = 0.05$$

tab

$$P\{\tau < t_0\} = 0.95 \Rightarrow t_0 = 1.8 \Rightarrow C = (1.8; +\infty) \quad \text{Ne prihvatamo } H_0 \text{ hipotezu}$$

t_{18}

3. U 6 preduzeća su mereni dohodak i lična primanja. Izraženi su pokazateljima X i Y:

X	0.767	0.112	2.677	0.865	0.832	5.503
Y	0.313	0.047	1.101	0.380	0.365	2.277

Izvršiti statističku analizu zavisnosti ova dva pokazatelja

Rešenje:

$$\hat{a} = 0.411$$

$$\hat{Y} = 0.411X + 0.009$$

$$\hat{b} = 0.009$$

$$H_0(\alpha = 0)$$

$$H_1(\alpha \neq 0)$$

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = 3.368$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum Y_i X_i - \hat{\beta} \sum Y_i \right) = 0.000119$$

$$\tau = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2 (n-2)} = 138,388$$

$$P\{|\tau| < t_0\} = 0.95 \quad 2F(t_0) - 1 = 0.95 \quad F(t_0) = 0.975 \quad \xRightarrow[t_4]{\text{tab}} t_0 = 2.8$$

$$C = (-\infty; -2.8) \cup (2.8; +\infty) \quad \tau \in C \Rightarrow \text{Ne prihvatamo } H_0 \text{ hipotezu}$$

4. Metodom najmanjih kvadrata ocenjen je linearni model:

$$\hat{Y} = 2,1X + 0,01$$

ca odgovarajućim parametrima:

$$\bar{X} = 15 \quad S_x^2 = 25 \quad \hat{\sigma}^2 = 90 \quad n = 100$$

Da li se, sa rizikom od 3%, može smatrati da je presek regresione prave sa Y-osom pozitivan.

Rešenje:

$$H_0(\beta = 0)$$

$$H_1(\beta > 0)$$

$$\tau = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \sqrt{S_x^2 + \bar{X}^2}} \sqrt{(n-2) S_x^2} = 0,0033$$

$$P\{t_{98} > t_0\} = 0.03 \quad P\{t_{98} < t_0\} = 0.97 \quad \Rightarrow t_0 \stackrel{\text{tab}}{=} 1.89$$

$$C = (1.89; +\infty) \quad \tau \notin C \Rightarrow H_0$$

5. Na osnovu izjava 10 studenata FON-a, dobijeni su sledeći podaci:

Student	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Broj sati na vežbama	26	16	34	23	47	52	18	22	50	42
Broj poena na ispitu	64	61	74	60	88	92	52	64	82	78

a) Oceniti odgovarajući regresioni model;

b) Testirati hipotezu da se, povećanjem broja sati na vežbama, povećava i broj poena na ispitu.

c) Predvideti broj poena na ispitu za studenta koji na vežbama provede 40 sati

Rešenje:

a) Oceniti odgovarajući regresioni model;

$$\bar{X} = \frac{\sum X}{N} = \frac{330}{10} = 33 \quad \bar{Y} = \frac{\sum Y}{N} = \frac{715}{10} = 71.5$$

$$S_x^2 = \frac{\sum X^2}{N} - \bar{X}^2 = \frac{12602}{10} - 1089 = 171.2$$

$$\hat{\alpha} = \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{S_x} = \frac{\frac{25176}{10} - 2359.5}{171.2} = 0.9235 \quad \hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} = 41.0251$$

$$\hat{Y} = 0.9235X + 41.0251$$

b) Testirati hipotezu da se, povećanjem broja sati na vežbama, povećava i broj poena na ispitu.

$$H_0 (\alpha = 0)$$

$$H_1 (\alpha > 0)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum X_i \cdot Y_i - \hat{\beta} \sum Y_i \right) = 12.602$$

$$\tau = \frac{\alpha - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2 (n-2)} = \frac{0.9235}{\sqrt{12.602}} \sqrt{171.2 \cdot 8} = 9.627$$

$$P\{\tau > t_0\} = \alpha \quad P\{\tau < t_0\} = 1 - \alpha \stackrel{mao}{\Rightarrow}_{t_8} t_0 = 2.9 \quad c = (2.9; \infty)$$

$$\tau \in c \Rightarrow H_0 \text{ odbacujemo}$$

c) Predvideti broj poena na ispitu za studenta koji na vežbama provede 40 sati

$$\hat{Y} = 0.9235 \cdot 40 + 41.0251 = 77,9651$$

6. Merenjem snage automobila i njegove potrošnje dobijeni su sledeći podaci:

Snaga	50	60	45	80	92	110
Potrošnja	7.6	7.7	6.3	8.1	7.5	9.2

Pretpostavljajući da posmatrana obeležja imaju Normalnu raspodelu:

Odrediti regresionu pravu za potrošnju automobila.

Da li se sa rizikom od 3% može smatrati da je presek regresione prave sa y osom pozitivan?

Rešenje:

$$\hat{Y} = \hat{\alpha}X + \hat{\beta}$$

$$\bar{X} = 72,83 \quad \bar{Y} = 7,73 \quad S_x^2 = 543,472$$

$$\hat{\alpha} = 0.029 \quad \hat{\beta} = 7.73 - 0.03 \cdot 72.83 = 5,58$$

$$\alpha = 0.03$$

$$H_0 (\beta = 0)$$

$$H_1 (\beta > 0)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum X_i Y_i - \hat{\beta} \sum Y_i \right) = 0,264 \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2} = 0.514$$

$$\tau = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \sqrt{S_x^2 + \bar{X}^2}} \sqrt{(n-2)S_x^2} : t_{n-2} \quad \tau = 6,629$$

$$p\{\tau > t_0\} = 0.03 \quad p\{\tau < t_0\} = 0.97 \xrightarrow{ma\delta} t_0 = 2.6$$

$$c = (2.6, +\infty) \quad \tau \in c$$

7. Prosečna starost mladoženja i nevesta u 10 opština bila je sledeća

Mladoženja	19	22	26	32	37	42	47	52	57	62
Nevesta	19	21	24	29	35	38	43	41	49	51

- a) Odrediti parametre linearne veze između ova dva pokazatelja
b) Da li se sa rizikom od 0.01 može smatrati da postoji takva linearna veza da prosečna starost neveste raste kada raste prosečna starost mladoženje?

Rešenje:

- a) Odrediti parametre linearne veze između ova dva pokazatelja

$$\bar{X} = 39,6 \quad \bar{Y} = \frac{\sum Y}{N} = \frac{715}{10} = 71,5$$

$$S_x^2 = \frac{\sum X^2}{N} - \bar{X}^2 = 200,24$$

$$\hat{\alpha} = \frac{\frac{1}{n} \sum XY - \bar{X}\bar{Y}}{S_x} = 0,756 \quad \hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} = 5,062$$

$$\hat{Y} = 0,756X + 5,062$$

- b) Da li se sa rizikom od 0.01 može smatrati da postoji takva linearna veza da prosečna starost neveste raste kada raste prosečna starost mladoženje?

$$H_0(\alpha = 0)$$

$$H_1(\alpha > 0)$$

$$\hat{\sigma}^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum X_i \cdot Y_i - \hat{\beta} \sum Y_i \right) = 2,55$$

$$\tau = \frac{\alpha - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2 (n-2)} = \frac{0,9235}{\sqrt{12,602}} \sqrt{171,2 \cdot 8} = 19,03$$

$$P\{\tau > t_0\} = 0,01 \quad P\{\tau < t_0\} = 0,99 \xrightarrow[t_0]{ma\delta} t_0 = 2.9 \quad c = (2.9; \infty)$$

$$\tau \in c \Rightarrow H_0 \text{ odbacujemo}$$

8. Izdržljivost sintetičkog materijala merena je pre bojenja (X) i posle bojenja (Y). Dobijeni su sledeći podaci:

Broj merenja	5	5	15	10	40	5	20	30
X	4.0	3.5	4.1	5.5	4.6	6.0	5.1	4.3
Y	3.0	3.0	3.8	2.1	4.9	5.3	3.1	2.7

- a) Odrediti linearnu vezu između izdržljivosti posle bojenja i pre bojenja.
b) Testirati hipotezu o nezavisnosti ovih promenljivih.

Rešenje:

- a) Odrediti linearnu vezu između izdržljivosti posle bojenja i pre bojenja.

$$\begin{aligned}\bar{x} &= \frac{599.0}{130} = 4.61 & \bar{y} &= \frac{473.5}{130} = 3.64 \\ S_x^2 &= \frac{2797.2}{130} - 4.61^2 = 0.26 & S_y^2 &= \frac{1862.45}{130} - 3.64^2 = 1.077 \\ \hat{\alpha} &= \frac{\frac{2186.8}{130} - 4.61 \cdot 3.64}{0.26} = 0.136 & \hat{\beta} &= 3.64 - 0.16 \cdot 4.61 = 3.015 \\ \hat{Y} &= 0.136X + 3.015\end{aligned}$$

- b) Testirati hipotezu o nezavisnosti ovih promenljivih.

$$\begin{aligned}H_0 (\alpha = 0) & & \tau &= \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2 (n-2)} = 0.801 \\ H_1 (\alpha \neq 0) & & &\end{aligned}$$

$$\begin{aligned}P\{|\tau| \geq t_0\} &= 0.01 & 2F(t_0) - 1 &= 0.99 & \tau \notin C \Rightarrow H_0 \\ F(t_0) &= 0.995 \xRightarrow[t_{198}]{\text{таб}} \approx \xRightarrow[N]{\text{таб}} t_0 = 2.57 & c &= (-\infty; -2.57) \cup (2.57; +\infty)\end{aligned}$$

$$\begin{aligned}H_0 : \rho &= 0 & r &= \hat{\alpha} \frac{S_y}{S_x} = 0.071 & \tau &= \frac{r}{\sqrt{1-r^2}} \sqrt{n-2} = 0.805 \\ H_1 : \rho &\neq 0 & && &\end{aligned}$$

$$\begin{aligned}P\{|\tau| \geq t_0\} &= 0.01 & 2F(t_0) - 1 &= 0.99 & \tau \notin C \Rightarrow H_0 \\ F(t_0) &= 0.995 \xRightarrow[t_{198}]{\text{таб}} \approx \xRightarrow[N]{\text{таб}} t_0 = 2.57 & c &= (-\infty; -2.57) \cup (2.57; +\infty)\end{aligned}$$

9. Na osnovu 10 obavljenih merenja dobijeni su podaci:

$$\frac{1}{n} \sum XY = 8.6 \quad \bar{x} = 2.0 \quad \bar{y} = 4.0 \quad S_x^2 = 4.0 \quad S_y^2 = 0.25$$

Testirati hipotezu: "Između X i Y postoji linearna veza, takva da, kad raste X, raste i Y". Formulirati nultu i alternativnu hipotezu i sprovesti test sa $\alpha = 0.01$.

Rešenje:

$$H_0 (\alpha = 0)$$

$$H_1 (\alpha > 0)$$

$$\tau = \frac{\hat{\alpha} - \alpha_0}{\hat{\sigma}} \sqrt{S_x^2 (n-2)} : t_{n-2}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (Y_i - \hat{Y}_i)^2 = \frac{1}{n} \left(\sum Y_i^2 - \hat{\alpha} \sum Y_i X_i - \hat{\beta} \sum Y_i \right)$$

$$\hat{\alpha} = \frac{\frac{1}{n} \sum X_i Y_i - \bar{x} \bar{y}}{S_x^2} = 0.15 \quad \hat{\beta} = \bar{y} - \hat{\alpha} \bar{x} = 3.7 \quad \hat{\sigma}^2 = 0.16$$

$$\tau = \frac{0.15}{\sqrt{0.16}} \sqrt{4 \cdot 8} = 2.115$$

$$P\{\tau > t_0\} = \alpha = 0.01$$

tab

$$P\{\tau < t_0\} = 0.99 \Rightarrow \underset{t_8}{t_0} = 2.9 \Rightarrow C = (2.9; +\infty) \quad \text{Ne prihvataamo } H_0 \text{ hipotezu}$$