# REGRESIJA I KORELACIJA

Šošić I.(2006). Primijenjena statistika, pp.379-546

**U velikom se broju istraživanja analizira dvije ili više statističkih varijabli s ciljem da se utvrdi.**

- postoji li povezanost među varijablama
- jakost veze
- može li se varijabla koja je predmet statističke analize prognozirati pomoću opaženih vrijednosti druge varijable (drugih varijabli)

**Regresijska se analiza bavi ispitivanjem ovisnosti jedne varijable o jednoj ili više nezavisnih varijabli s ciljem da se utvrdi analitički izraz takve povezanosti, odnosno model koji služi u analitičke i prediktivne svrhe.**

Model može povezivati dvije varijable (bivariatna veza) ili više varijabli (multivariatna veza).

Model može biti deterministički (funkcionalan) ili statistički (stohastički, probabilistički).

Deterministčkim se modelom pretpostavlja egzaktna veza među varijablama (za svaku vrijednost nezavisne varijable jednoznačno je određena vrijednost zavisne varijable).

$$Y=f(X)$$

Statistički model izražava labaviju vezu među varijablama. Vrijednost zavisne varijable (Y) nije jednoznačno određena za zadanu vrijednost nezavisne varijable. Postoje neobjašnjene varijacije Y-a zbog neuključivanja varijabli koje utječu na ponašanje zavisne varijable ili zbog slučajnih utjecaja.

Y= deterministička komponenta+slučajna pogreška

# JEDNOSTAVNA REGRESIJA I KORELACIJA

**Jednostavna se regresija bavi pronalaženjem analitičkog izraza kojim se opisuje povezanost zavisne ili regresand varijable s jednom nezavisnom ili regresorskom varijablom.**

**Model jednostavne linearne regresije je probabilistički model. Pretpostavlja se da je zavisna varijabla (Y) (varijabla koju se modelira) slučajna varijabla povezana s nezavisnom varijablom (X) slijedećim izrazom:**

$$Y_i = \alpha + \beta X_i + e_i \quad i = 1,2,...,n$$

**Pretpostavlja se da za svaku vrijednost varijable X postoji distribucija vrijednosti varijable Y.**
**U gornjem su izrazu:** $\alpha, \beta$ **nepoznati parametri**

$e_i$ , $i = 1,2,...,n$ **su slučajne varijable ( greške relacije). To su nepoznate slučajne varijable za koje se pretpostavlja da su međusobno nezavisne i normalno distribuirane slučajne varijable sa sredinom nula i varijancom**

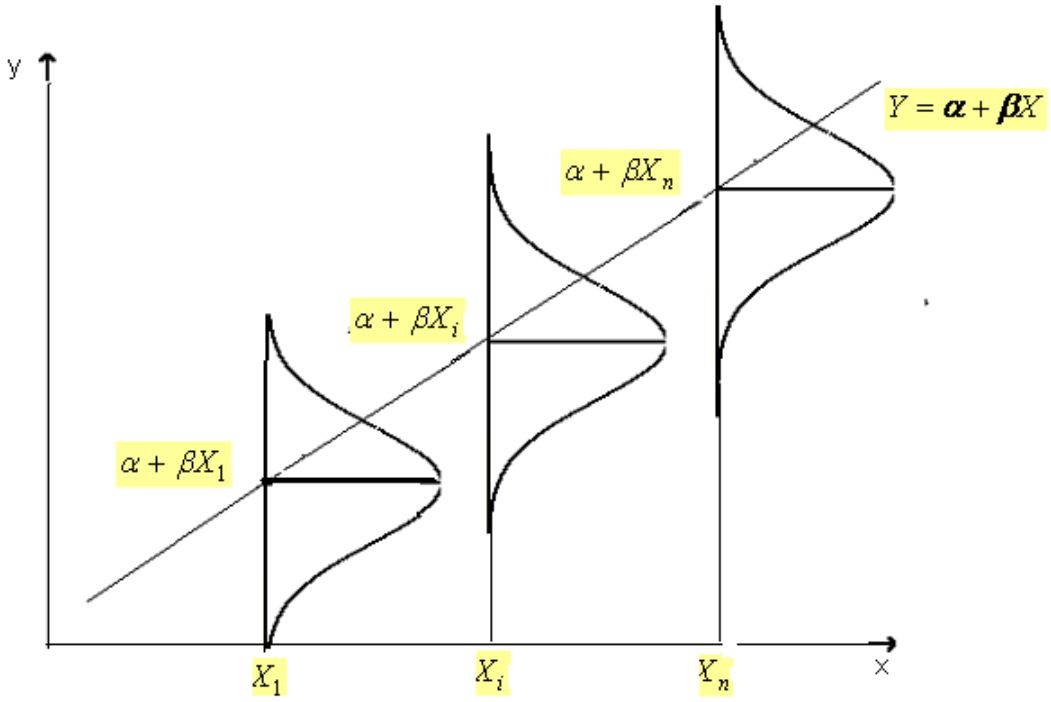
$$\sigma^2, \text{tj. } e_i \sim N(0,\sigma^2) \quad E(e_i e_j) = 0 \quad \forall i \neq j$$

**U klasičnoj regresijskoj analizi pretpostavlja se da je varijabla X nestohasička (tj da u ponovljenim uzorcima ima fiksne vrijednosti .**

S obzirom da su slučajne varijable $Y_i$ linearne funkcije normalno distribuiranih varijabli $e_i$, one su također normalno distribuirane s parametrima:

$$E(Y_i) = E(\alpha + \beta X_i + e_i) = \alpha + \beta X_i + E(e_i) = \alpha + \beta X_i$$

$$Var(Y_i) = Var(\alpha + \beta X_i + e_i) = Var(e_i) = \sigma^2$$




**Koraci u analizi modela jednostavne linearne regresije:**

- Pretpostavlja se linearna regresijska veza među varijablama Y i X, pri čemu je model populacije:

$$y_i = \alpha + \beta x_i + e_i$$

- Polazeći od ***n*** empirijskih (opaženih) vrijednosti varijabli x i y (koje se smatraju uzorkom iz hipotetičke populacije) crta se dijagram rasipanja.

- Nepoznati se parametri procjenjuju metodom najmanjih kvadrata. Računaju se procjene pokazatelja reprezentativnosti modela, kao što su procjena varijance, standardne devijacije i koeficijenta varijacije regresije, koeficijent determinacije, koeficijent korelacije i dr.
- Ispituje se kakvoća i upotrebljivost dobivenih rezultata. Računaju se elementi analize varijance, testiraju se hipoteze o parametrima u

regresijskom modelu, te se ispituje jesu li ispunjene polazne pretpostavke o modelu.

- Ako je model zadovoljavajući, koristi se za predviđanje, procjenjivanje i dr.

## PROCJENJIVANJE PARAMETARA: METODA NAJMANJIH KVADRATA

**Metoda najmanjih kvadrata sastoji se u određivanju regresijskog pravca koji minimizira sumu kvadrata rezidualnih odstupanja.**

Model uzorka s procijenjenim parametrima glasi:

$$y_i = \hat{\alpha} + \hat{\beta}x_i + \hat{e}_i$$

pri čemu su $\hat{\alpha}$ i $\hat{\beta}$ procjene parametara, a $\hat{e}_i$ su rezidualna odstupanja ili procjene slučajnih varijabli. Gornja se jednadžba može napisati u obliku:

$$\boldsymbol{y_i = \hat{y}_i + \hat{e}_i}$$

gdje je s $\hat{y}_i$ označena i-ta procijenjena ili regresijska vrijednost zavisne varijable. Iz tog izraza slijedi:

$$\hat{e}_i = y_i - \hat{y}_i, \quad \hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

**Suma kvadrata rezidualnih odstupanja glasi:**

$$SR = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

**Iz zahtjeva da ta suma bude minimalna dolazi se do normalnih jednadžbi za procjenitelje metodom najmanjih kvadrata:**

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

**Procijenjeni regresijski pravac je:**

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

**Regresijske vrijednosti su:**

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i \quad i = 1,2,...,n$$

**Procjenitelj varijance regresije:**

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

**Procjenitelj standardne devijacije regresije** (procijenjena standardna pogreška regresijskog modela je:

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

**a procjenitelj koeficijenta varijacije ($C.V \equiv \hat{V}$):**
**:**

$$\hat{V} = \frac{\hat{\sigma}}{\bar{y}}100$$

## INTERPRETACIJA PROCJENA

Konstantni član (*intercept*) $\hat{\alpha}$ je vrijednost regresije ako je vrijednost nezavisne varijable x=0.

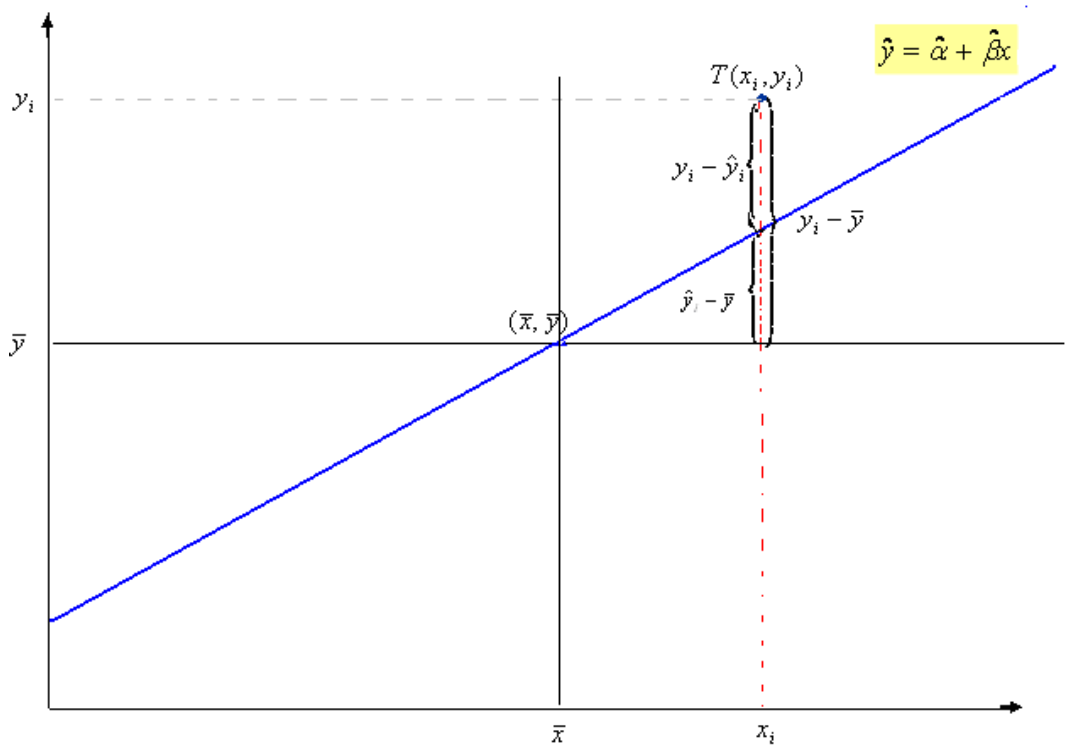
Regresijski koeficijent $\hat{\beta}$ je promjena regresijske vrijednosti zavisne varijable za jedinično povećanje varijable x. Ili:
Regresijski koeficijent $\hat{\beta}$ je prosječna promjena zavisne varijable y za jedinično povećanje varijable x.

Regresijske vrijednosti $\hat{y}_i,\ i = 1,2,\cdots,n$ su procijenjene vrijednosti zavisne varijable za zadane vrijednosti nezavisne varijable $x_i,\ i = 1,2,\cdots,n$. (To su ordinate na regresijskom pravcu).

Rezidualna odstupanja $\hat{e}_i$, $i = 1,2,\cdots,n$ su procjene slučajnih varijabli $e_i$, $i = 1,2,\cdots,n$ na osnovi uzorka. To su razlike empirijskih i regresijskih vrijednosti.

Procjena varijance $\hat{\sigma}^2$, procjena standardne devijacije $\hat{\sigma}$ i procjena koeficijenta varijacije C.V su mjere disperzije regresijskog modela. Procjena standardne devijacije regresije se interpretira kao prosječno odstupanje empirijskih od regresijskih vrijednosti. Model je dobar ako su procjene varijance i standardne devijacije male. Procjena standardne devijacije regresije izražena je u istim mjernim jedinicama kao i vrijednosti zavisne varijable. Procjena koeficijenta varijacije je relativna mjera disperzije oko regresijskog pravca.

## JEDNADŽBA ANALIZE VARIJANCE. TABELA ANOVA




Odstupanje empirijske vrijednosti $y_i$ od prosjeka može se raščlaniti na protumačeno odstupanje (odstupanje odgovarajuće regresijske vrijednosti od prosjeka) i neprotumačeno ili rezidualno odstupanje:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad \mathbf{i = 1,2,\cdots,n}$$

S obzirom da je suma odstupanja pojedinačnih vrijednosti varijable od prosjeka uvijek jednaka nuli:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0,$$

računa se suma kvadrata odstupanja:

$$\boxed{\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Dobivena se jednadžba zove jednadžba analize varijance. Njene su komponente:

**Ukupna suma kvadrata ST** (*The total sum of squares, corrected sum of squares* $SS_{yy}$):

$$ST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

**Protumačena suma kvadrata SP** (*the sum of regression due to the linear regresion, model, explained sum of squares SSR*) je suma kvadrata odstupanja regresijskih vrijednosti od prosjeka:

$$SP = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

**Suma kvadrata neprotumačenih ili rezidualnih odstupanja SR** (*residual, unexplained sum of squares, sum of squared errors, SSE*):

$$SR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Jednadžba analize varijance se simbolički može zapisati:

ST=SP+SR

Elementi analize varijance (jednostavna regresija) predočeni su u tabeli analize varijance (tabeli ANOVA):

| Izvor varijacije | Stupnjevi slobode DF | Sume kvadrata SS | Sredine kvadrata MS | F-omjer | PROB>F |
|---|---|---|---|---|---|
| Protumačen modelom | 1 | SP | SP/1 | $\frac{SP/1}{SR/(n-2)}$ | |
| Neprotumačena odstupanja | n-2 | SR | SR/(n-2) | | |
| Ukupno | n-1 | ST | | | |

$\hat{\sigma}^2 = \frac{SR}{n-2}$ je nepristran procjenitelj varijance regresije; $\hat{\sigma} = \sqrt{\frac{SR}{n-2}}$ je nepristrani procjenitelj standardne devijacije regresije - Root MSE,

$r^2 = \frac{SP}{ST}$ je koeficijent determinacije, R-Square;

$\bar{r}^2$ je korigirani koeficijent determinacije – Adj R-Sq;

## KOEFICIJENT DETERMINACIJE, KORIGIRANI KOEFICIJENT DETERMINACIJE, KOEFICIJENT KORELACIJE

Koeficijent determinacije (*Coefficient of determination*) je proporcija varijacija iz uzorka protumačena linearnom regresijskom vezom:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad 0 \le r^2 \le 1$$

Interpretacija: 100($r^2$)% varijacija iz uzorka (mjerenih sumom kvadrata ukupnih odstupanja empirijskih vrijednosti od prosjeka) može se protumačiti uporabom x-a za procjenu (predviđanje) y-a u modelu jednostavne linearne regresije.

Korigirani koeficijent determinacije (*corrected coefficient of determination*) je mjera reprezentativnosti modela koja se izračunava korigiranjem koeficijenta determinacije faktorom koji ovisi o broju stupnjeva slobode:

$$\bar{r}^2 = 1 - \frac{n-1}{n-2}(1 - r^2)$$

Koeficijent linearne korelacije (coefficient of linear corelation) je mjera jakosti i smjera linearne veze između varijabli x i y. Definiran je izrazom:

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \qquad r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}, \qquad -1 \le r \le 1$$

Koeficijent jednostavne linearne korelacije može se odrediti i kao drugi korijen iz koeficijenta determinacije, s tim da se predznak od r određuje u skladu s predznakom regresijskog koeficijenta:

$$r = \sqrt{r^2}\ ; \quad \mathrm{r}^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}; \qquad \mathrm{sign(r)} = \mathrm{sign}(\hat{\beta})$$

## INTERVALNA PROCJENA PARAMETRA $\beta$

Ako su ispunjene pretpostavke o modelu jednostavne linearne regresije sampling distribucija procjenitelja parametra $\beta$ je normalna s očekivanom vrijednosti jednakom parametru $\beta$ i standardnom devijacijom jednakoj standardnoj pogreški regresijskog koeficijenta.

**Procjena jednim brojem** parametra je $\hat{\beta}$.

**Intervalna procjena od** $\beta_1$ uz pouzdanost (1-γ) definirana je izrazom:

$$P\left\{\hat{\beta} - t_{\gamma/2}\sigma_{\hat{\beta}} < \beta < \hat{\beta} + t_{\gamma/2}\sigma_{\hat{\beta}}\right\} = 1-\gamma$$

**Standardna pogreška** $s_{\hat{\beta}_1}$ definirana je s:

$t_{\gamma/2}$ je koeficijent pouzdanosti koji pripada t-distribuciji s n-2 stupnja slobode.

$$\sigma_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

# PREDVIĐANJE POJEDINAČNIH VRIJEDNOSTI ZAVISNE VARIJABLE ZA DANU VRIJEDNOST NEZAVISNE VARIJABLE

Procjenitelj jednim brojem zavisne varijable za zadanu vrijednost $x_f$ nezavisne varijable je:

$$\hat{y}_f = \hat{\alpha} + \hat{\beta} x_f$$

Prognostički interval uz pouzdanost (1-$\gamma$) je:

$$P(\hat{y}_f - t_{\gamma/2}\sigma_{\hat{y}_f} < y_f < \hat{y}_f + t_{\gamma/2}\sigma_{\hat{y}_f}) = 1-\gamma$$

pri čemu je $\hat{y}_f$ procjena jednim brojem zavisne varijable za $x = x_f$, $t_{\gamma/2}$ je koeficijent pouzdanosti, a $\sigma_{\hat{y}_f}$ je standardna pogreška prognostičke vrijednosti definirana izrazom:

$$\sigma_{\hat{y}_f} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

# TESTIRANJE HIPOTEZA O PARAMETRU $\beta$

Testiranje hipoteza o parametru $\beta$ moguće je provesti pomoću dvosmjernog testa ili pomoću jednosmjernih testova.

| Dvosmjerni test | Jednosmjerni test na gornju granicu | Jednosmjerni test na donju granicu |
|---|---|---|
| $H_0 : \beta = 0$<br>$H_1 : \beta \neq 0$ | $H_0 : \beta \leq 0$<br>$H_1 : \beta > 0$ | $H_0 : \beta \geq 0$<br>$H_1 : \beta < 0$ |
| Područje odbacivanja nulte hipoteze | Područje odbacivanja nulte hipoteze | Područje odbacivanja nulte hipoteze |
| (1) Testovna veličina:<br>$t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ | (1) Testovna veličina:<br>$t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ | (1) Testovna veličina:<br>$t = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}$ |
| $\lvert t \rvert > t_{\alpha/2}(n-2) \Rightarrow H_1$ | $t > t_{\alpha}(n-2) \Rightarrow H_1$ | $t < -t_{\alpha/2}(n-2) \Rightarrow H_1$ |
| (2)<br>$\left.\begin{array}{c} \hat{\beta} < t_{\alpha/2} \cdot \sigma_{\hat{\beta}} \\ ili \\ \hat{\beta} > t_{\alpha/2} \cdot \sigma_{\hat{\beta}} \end{array}\right\} \Rightarrow H_1$ | (2)<br>$\hat{\beta} > t_{\alpha} \cdot \sigma_{\hat{\beta}} \Rightarrow H_1$ | (2)<br>$\hat{\beta}_1 < -t_{\alpha} \cdot \sigma_{\hat{\beta}} \Rightarrow H_1$ |
| (3)<br>$\text{p - vrijednost} < \alpha \Rightarrow H_1$<br>$p - vrijednost = 2P(t > t_{emp})$ | (3)<br>$\text{p - vrijednost} < \alpha \Rightarrow H_1$<br>$p - vrijednost = P(t > t_{emp})$ | (3)<br>$\text{p - vrijednost} < \alpha \Rightarrow H_1$<br>$p - vrijednost = P(t > \lvert t_{emp} \rvert)$ |