REGRESIJSKA ANALIZA

Vježba 11

Regresijska analiza

Regresijska analiza je matematičko-statistički postupak kojim se utvrđuje odgovarajuća funkcionalna veza (relacija) između jedne **zavisne** ili **kriterijske varijable** i jedne ili više **nezavisnih** ili **prediktorskih varijabli**.

Zavisna (kriterijska) varijabla je varijabla čiji se varijabilitet objašnjava putem nezavisnih varijabli.

Nezavisne (prediktorske) varijable su varijable na temelju kojih se objašnjava varijabilitet zavisne varijable.

Regresijska analiza se u kineziologiji najčešće koristi u svrhu:

- utvrđivanja utjecaja jedne varijable ili skupa varijabli na neku kriterijsku varijablu (npr. utvrđivanje utjecaja građe tijela na rezultat u bacanju kugle) i

- utvrđivanje trenda razvoja rezultata u nekom sportu (npr. utvrđivanje trenda razvoja najboljih rezultata u bacanju kugle na svjetskim prvenstvima)

Regresijska analiza

Funkcionalna veza između prediktorskih varijabli i kriterijske varijable definira se utvrđivanjem odgovarajuće **regresijske jednadžbe**. Opći oblik regresijske jednadžbe izgleda ovako:

$$y = f(x_1, x_2, \ldots, x_m) + \mathrm{e}$$

- y - zavisna (kriterijska) varijabla
- f - odgovarajuća funkcija
- $x_1, x_2, \ldots, x_m$ - nezavisne (prediktorske) varijable
- e - greška prognoze.

Regresijska analiza

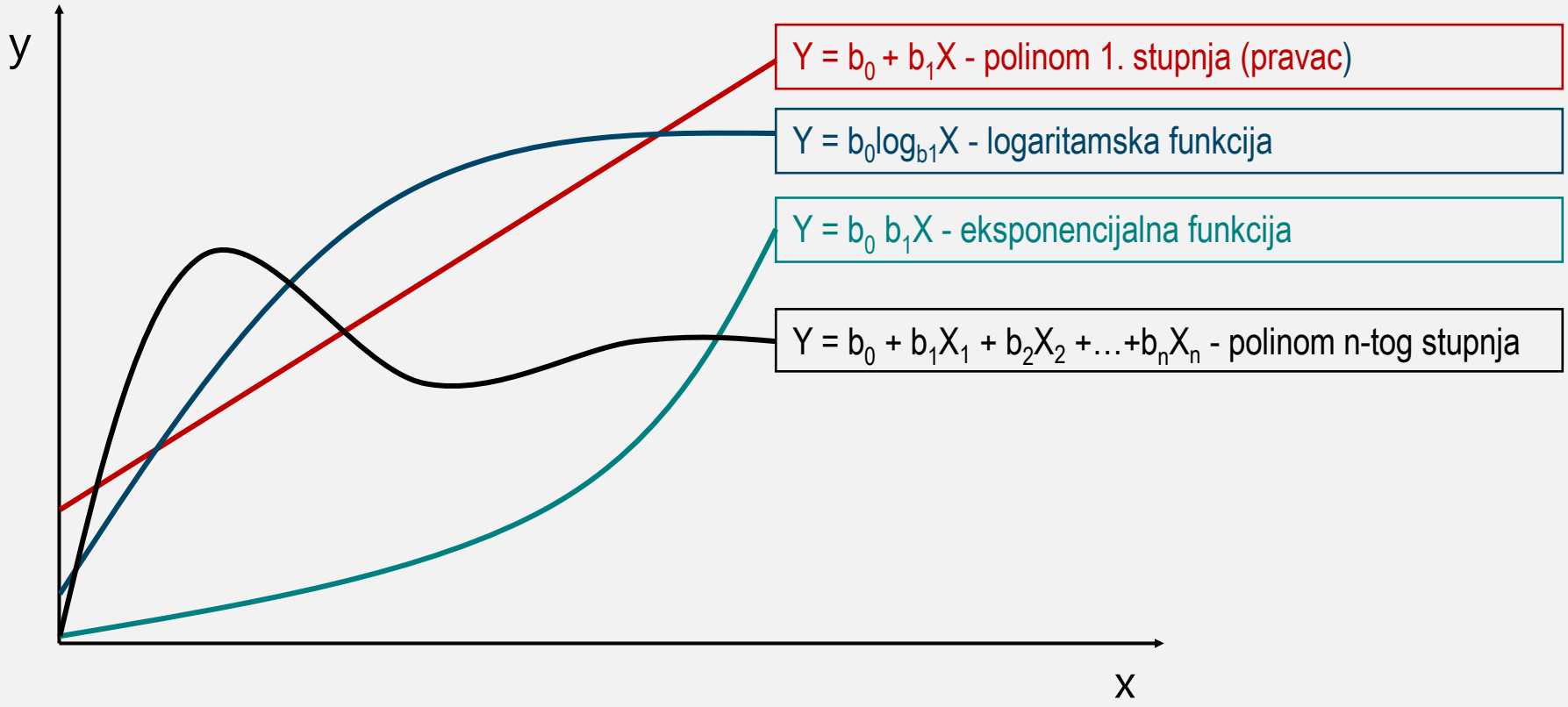
Regresijske modele moguće je generalno podijeliti na temelju dvaju kriterija i to:

- prema broju nezavisnih varijabli na:
  - ✓ **jednostavne (simple) regresijske modele** i
  - ✓ **višestruke (multiple) regresijske modele**, te

- prema odnosu između zavisne i nezavisnih varijabli na:
  - ✓ **linearne regresijske modele** i
  - ✓ **nelinearne regresijske modele**.

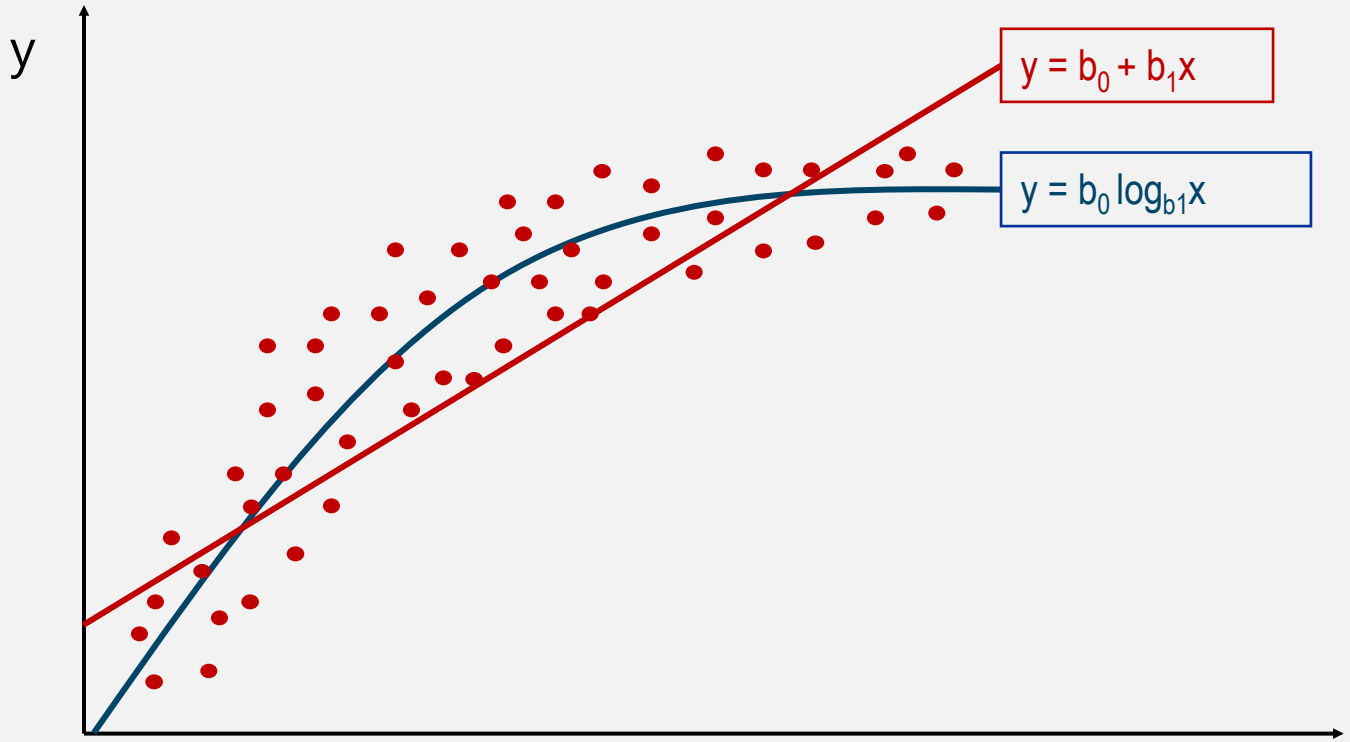
Regresijska analiza

Linearni i nelinearni modeli jednostavne regresijske analize:



Regresijska analiza

Odabir modela jednostavne regresijske analize vrši se pomoću korelacijskog dijagrama.



(*Primjer:* logaritamska funkcija je prikladan model)

Regresijska analiza

Jednostavna linearna regresijska analiza

Jednostavnom linearnom regresijskom analizom utvrđuje se linearna povezanost između jedne nezavisne (prediktorske) i jedne zavisne (kriterijske) varijable pri čemu regresijska jednadžba ima sljedeći oblik:

$$y_i = b_0 + b_1 x_i + e_i$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- b_0 i b_1 - regresijski koeficijenti
- x_i - rezultat entiteta i u prediktorskoj varijabli
- e_i - rezidualna vrijednost entiteta i
- $i = 1,\ldots,n$
- n – *broj entiteta*

Regresijska analiza

Jednostavna linearna regresijska analiza

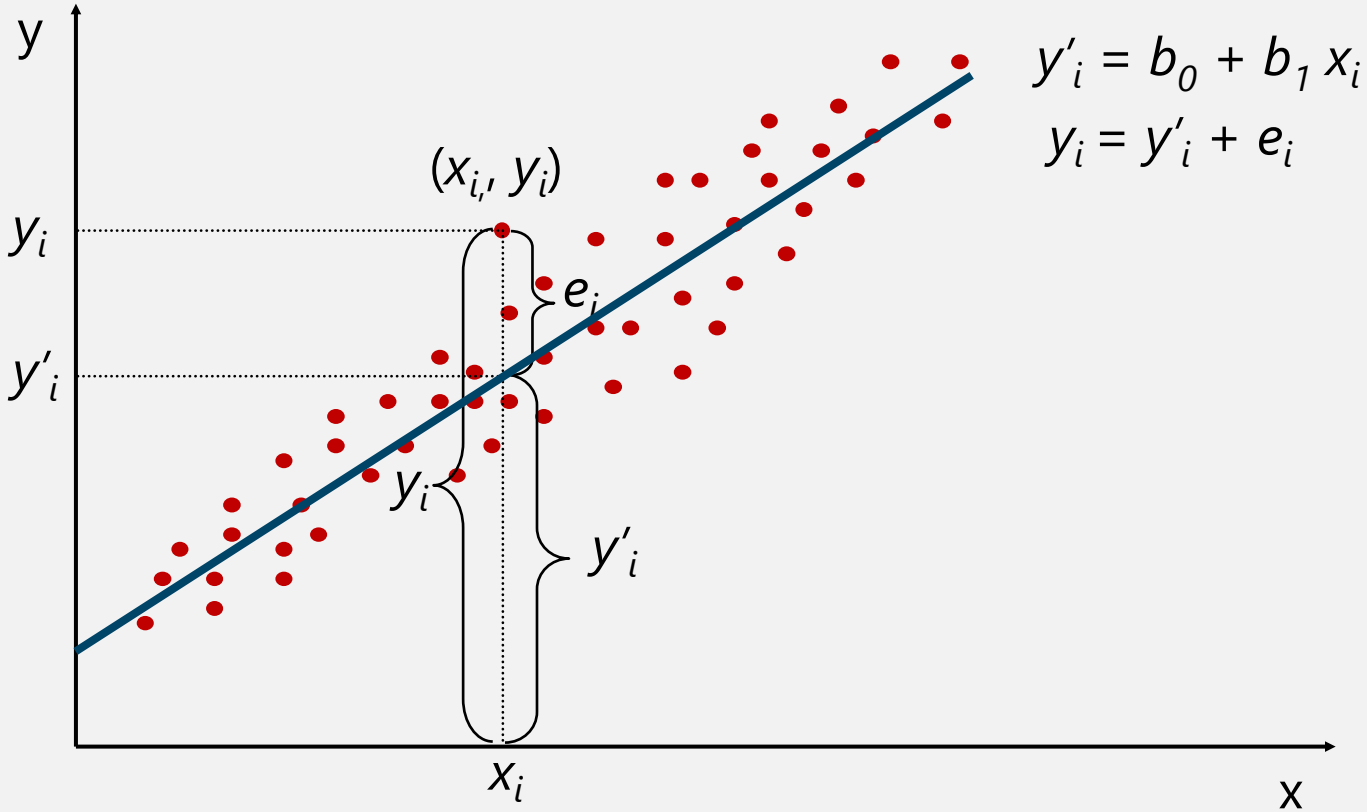
Regresijski koeficijenti omogućavaju prognoziranje rezultata entiteta u kriterijskoj varijabli na temelju rezultata u prediktorskoj varijabli putem sljedeće formule:

$$y'_i = b_0 + b_1 x_i$$

gdje je

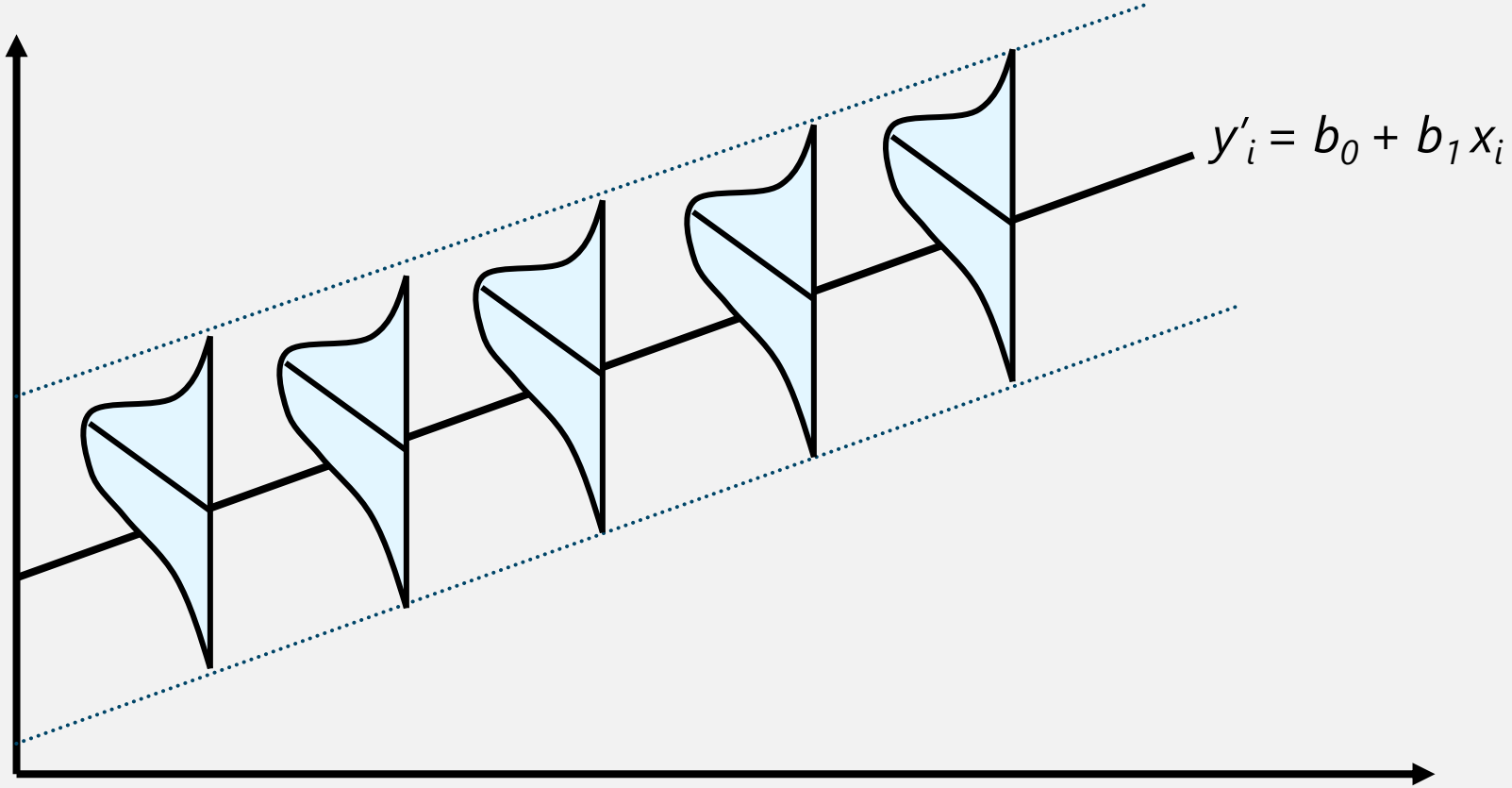
- y'_i - prognozirani rezultat entiteta i u kriterijskoj varijabli
- b_0 i b_1 - regresijski koeficijenti
- x_i - rezultat entiteta i u prediktorskoj varijabli.

Jednostavna linearna regresijska analiza



(Prikaz regresijskog pravca, originalnih i prognoziranih rezultata u kriterijskoj varijabli i rezidualnih vrijednosti)

Jednostavna linearna regresijska analiza



(Distribucija rezidualnih vrijednosti oko regresijskog pravca)

Regresijska analiza

Jednostavna linearna regresijska analiza

Koeficijenti regresijskog pravca utvrđuju se **metodom najmanjih kvadrata**.

Metoda najmanjih kvadrata temelji se na uvjetu da je suma kvadrata rezidualnih vrijednosti minimalna

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y'_i)^2 = \min$$

gdje je

- e_i - rezidualna vrijednost entiteta i
- y_i - rezultat entiteta i u kriterijskoj varijabli
- y'_i - prognozirani rezultat entiteta i u kriterijskoj varijabli

Regresijska analiza

Jednostavna linearna regresijska analiza

Regresijski koeficijent b_0 predstavlja odsječak na osi zavisne varijable y, odnosno, vrijednost zavisne varijable y ukoliko je vrijednost nezavisne varijable x = 0 .

$$b_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x^2{}_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2}$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- x_i - rezultat entiteta i u prediktorskoj varijabli

Regresijska analiza

Jednostavna linearna regresijska analiza

Regresijski koeficijent b_1 određuje nagib pravca, odnosno, pokazuje koliko se u prosjeku linearno mijenja vrijednost zavisne varijable *y* za jedinični porast vrijednosti nezavisne varijable *x*.

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x^2{}_i - (\sum_{i=1}^{n} x_i)^2}$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- x_i - rezultat entiteta i u prediktorskoj varijabli

Regresijska analiza

Jednostavna linearna regresijska analiza

Regresijski koeficijenti se također mogu izračunati i rješavanjem regresijske jednadžbe u matričnom obliku:

$$\underbrace{\boldsymbol{y}}_{\begin{vmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{vmatrix}} = \underbrace{\boldsymbol{X}}_{\begin{vmatrix} 1 & x_1 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{vmatrix}} \cdot \underbrace{\boldsymbol{b}}_{\begin{vmatrix} b_0 \\ b_1 \end{vmatrix}} + \underbrace{\boldsymbol{e}}_{\begin{vmatrix} e_1 \\ . \\ . \\ . \\ e_n \end{vmatrix}}$$

gdje je

- $\boldsymbol{y}$ - vektor n rezultata entiteta u kriteriju
- $\boldsymbol{X}$ - matrica reda $n \cdot 2$ rezultata entiteta u prediktoru
- $\boldsymbol{b}$ - vektor regresijskih koeficijenata
- $\boldsymbol{e}$ - vektor n rezidualnih vrijednosti

Regresijska analiza

Jednostavna linearna regresijska analiza

$$y = X\,b \qquad / X^T$$

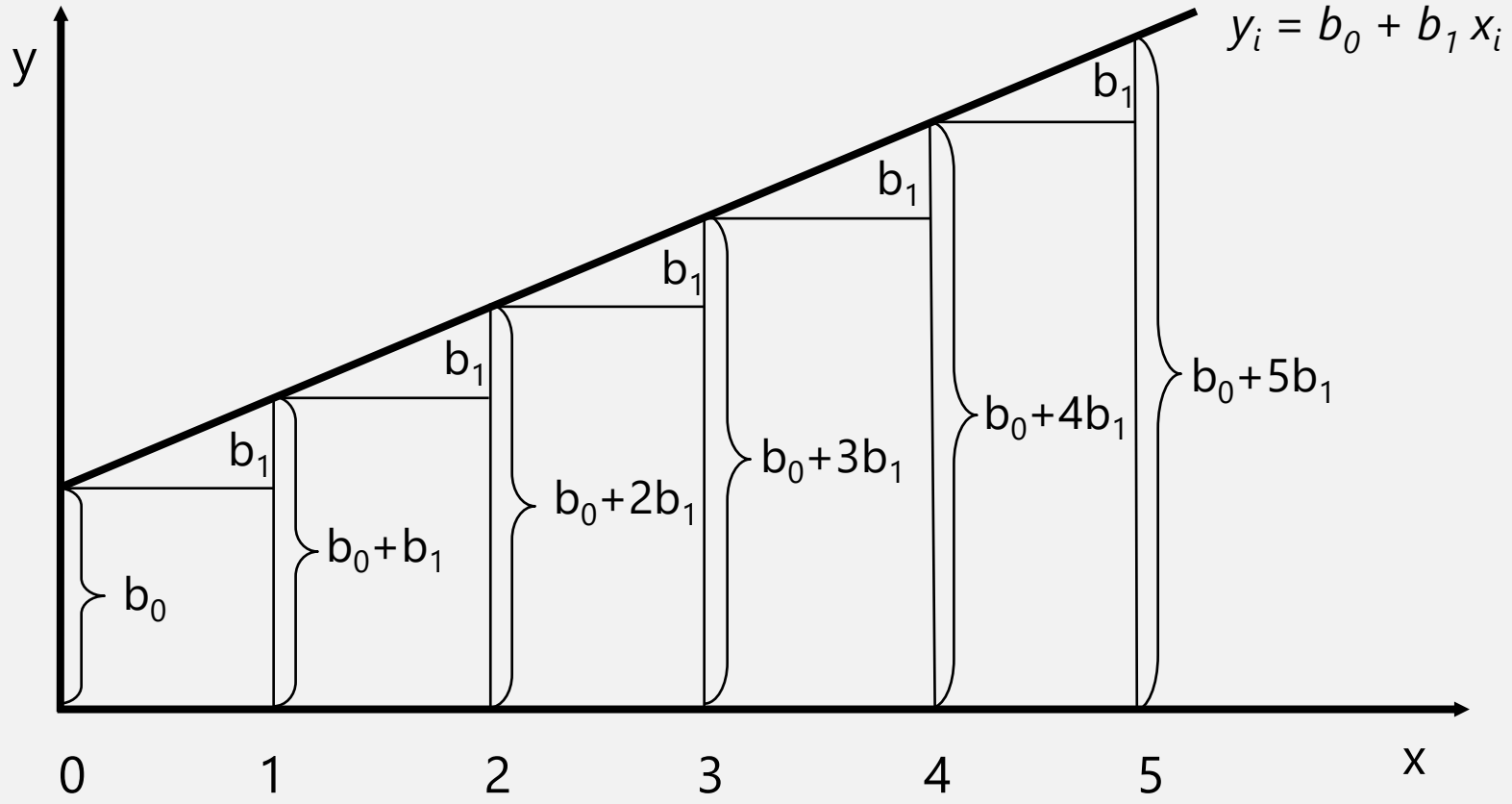
$$X^T y = X^T X\,b \qquad / (X^T X)^{-1}$$

$$b = (X^T X)^{-1} X^T y$$

$$y' = Xb$$

$$e = y - y'$$

Jednostavna linearna regresijska analiza



(Prikaz regresijskih koeficijenata b_0 i b_1)

Regresijska analiza

Jednostavna linearna regresijska analiza

Standardna pogreška prognoze (σ_e) je drugi korijen iz varijance rezidualnih vrijednosti, a predstavlja mjeru reprezentativnosti regresijskog modela.

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{n-2}}$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- y'_i - prognozirani rezultat entiteta i u kriterijskoj varijabli

Regresijska analiza

Jednostavna linearna regresijska analiza

Koeficijent korelacije između kriterijske i prediktorske varijable izražava veličinu njihove linearne povezanosti.

Kada je $r_{x,y} = 0$ to znači da nezavisna varijabla x nema nikakav utjecaj na varijabilitet kriterijske varijable y .

Ako koeficijent korelacije ima maksimalnu vrijednost $r_{x,y} = 1$, to znači da je cjelokupan varijabilitet varijable y moguće pripisati utjecaju varijable x.

Kvadrat koeficijenta korelacije (r^2) naziva se **koeficijent determinacije**, a predstavlja proporciju varijance kriterijske varijable koju je moguće objasniti putem prediktorske varijable.

Regresijska analiza

Višestruka linearna regresijska analiza

Višestrukom linearnom regresijskom analizom utvrđuje se linearna povezanost između dviju ili više nezavisnih (prediktorskih) i jedne zavisne (kriterijske) varijable pri čemu regresijska jednadžba ima sljedeći oblik:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_m x_{im} + e_i$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- $b_0,\ldots,b_m$ - regresijski koeficijenti
- $x_{i1},\ldots,x_{im}$ - rezultati entiteta i u m prediktorskih varijabli
- e_i - rezidualna vrijednost entiteta i
- $i = 1,\ldots,n$ (n - broj entiteta), a m - broj prediktora

Višestruka linearna regresijska analiza

Regresijski koeficijenti mogu se izračunati rješavanjem regresijske jednadžbe u matričnom obliku:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{b} + \boldsymbol{e}$$

$$\begin{vmatrix} y_1 \\ . \\ . \\ y_n \end{vmatrix} = \begin{vmatrix} 1 & x_{11} & . & . & x_{1m} \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & . & . & x_{nm} \end{vmatrix} . \begin{vmatrix} b_0 \\ b_1 \\ . \\ b_m \end{vmatrix} + \begin{vmatrix} e_1 \\ . \\ . \\ e_n \end{vmatrix}$$

Regresijska analiza

Višestruka linearna regresijska analiza

$$y = X\,b \qquad / X^T$$

$$X^T y = X^T X b \qquad / (X^T X)^{-1}$$

$$b = (X^T X)^{-1} X^T y$$

$$y' = Xb$$

$$e = y - y'$$

Regresijska analiza

Višestruka linearna regresijska analiza

Regresijski koeficijent b_0 predstavlja vrijednost zavisne varijable *y* ukoliko je vrijednost svih nezavisnih varijabli jednaka *0* .

Regresijski koeficijenti $b_1,\ldots,b_m$ pokazuju koliko se u prosjeku linearno mijenja vrijednost zavisne varijable *y* za jedinični porast vrijednosti odgovarajuće nezavisne varijable ($x_1,\ldots,x_m$) uz uvjet da su vrijednosti ostalih nezavisnih varijabli konstantne.

Regresijska analiza

Višestruka linearna regresijska analiza

Ako se kriterijska i prediktorske varijable prethodno standardiziraju regresijska jednadžba poprima sljedeći oblik:

$$k_i = \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_m z_{im} + \varepsilon_i$$

gdje je

- k_i - standardizirani rezultat entiteta i u kriterijskoj varijabli
- $\beta_1,\ldots,\beta_m$ - standardizirani regresijski koeficijenti
- $z_{i1},\ldots,z_{im}$ - standardizirani rezultati entiteta i u m prediktorskih varijabli
- ε_i - standardizirana rezidualna vrijednost entiteta i
- $i = 1,\ldots,n$ (n - broj entiteta), a m - broj prediktora

Višestruka linearna regresijska analiza

Standardizirani regresijski koeficijenti mogu se izračunati rješavanjem sljedeće jednadžbe u matričnom obliku:

$$k = Z \beta + \varepsilon$$

$$\begin{vmatrix} k_1 \\ . \\ . \\ k_n \end{vmatrix} = \begin{vmatrix} z_{11} & . & . & z_{1m} \\ . & . & . & . \\ . & . & . & . \\ z_{n1} & . & . & z_{nm} \end{vmatrix} \cdot \begin{vmatrix} \beta_1 \\ . \\ . \\ \beta_m \end{vmatrix} + \begin{vmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{vmatrix}$$

Regresijska analiza

Višestruka linearna regresijska analiza

$$\boldsymbol{k} = \boldsymbol{Z}\,\beta \qquad / \boldsymbol{Z}^{\boldsymbol{T}} n^{-1}$$

$$\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{k}n^{-1} = \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z}n^{-1}\beta$$

$$\boldsymbol{r} = \boldsymbol{R}\,\beta \qquad / \boldsymbol{R}^{-1}$$

$$\beta = \boldsymbol{R}^{-1}\,\boldsymbol{r}$$

gdje je

- $\boldsymbol{k}$ - vektor n standardiziranih rezultata entiteta u kriteriju
- $\boldsymbol{Z}$ - matrica reda $n \cdot m$ standardiziranih rezultata entiteta u m prediktora
- $\boldsymbol{\beta}$ - vektor standardiziranih regresijskih koeficijenata
- $\boldsymbol{r}$ - vektor korelacija m prediktora s kriterijem
- $\boldsymbol{R}$ - matrica međusobnih korelacija m prediktora

Regresijska analiza

Višestruka linearna regresijska analiza

Standardizirani regresijski koeficijenti $\beta_1,\ldots,\beta_m$ su relativni koeficijenti utjecaja, a predstavljaju veličinu promjene zavisne varijable izraženu u dijelovima standardne devijacije za jedinični porast standardizirane vrijednosti odgovarajuće nezavisne varijable ($z_1,\ldots,z_m$) uz uvjet da su vrijednosti preostalih nezavisnih varijabli konstantne.

Statistička značajnost svakog pojedinog regresijskog koeficijenta se testira putem Studentove t-distribucije.

Pri tome je za svaki regresijski koeficijent moguće postaviti sljedeću alternativnu (H1), odnosno nultu (H0) hipotezu:

- H0: Uz pogrešku *p* ne možemo tvrditi da je utjecaj prediktora *j* na kriterijsku varijablu statistički značajan.

- H1: Utjecaj prediktora *j* na kriterijsku varijablu je statistički značajan uz pogrešku *p*.

Regresijska analiza

Višestruka linearna regresijska analiza

Standardna pogreška prognoze (σ_e) je drugi korijen iz varijance rezidualnih vrijednosti, a predstavlja mjeru reprezentativnosti regresijskog modela.

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{n-(m+1)}}$$

gdje je

- y_i - rezultat entiteta i u kriterijskoj varijabli
- y'_i - prognozirani rezultat entiteta i u kriterijskoj varijabli
- $i = 1,\ldots,n$
- n - broj entiteta, m - broj prediktorskih varijabli

Regresijska analiza

Višestruka linearna regresijska analiza

Koeficijent multiple korelacije *(ρ)* je korelacija između kriterijske varijable i varijable prognoziranih rezultata, a izražava veličinu linearne povezanosti skupa prediktorskih varijabli s kriterijem.

Koeficijent multiple korelacije se kreće u intervalu od *0* do *1* pri čemu *0* označava nikakavu, a *1* potpunu zavisnost kriterijske varijable o skupu prediktorskih varijabli.

Kvadrat koeficijenta korelacije (ρ^2) naziva se **koeficijent multiple determinacije**, a predstavlja proporciju varijance kriterijske varijable koju je moguće objasniti putem skupa prediktorskih varijabli.

Regresijska analiza

Višestruka linearna regresijska analiza

Statistička značajnost koeficijenta multiple korelacije (ρ) se testira putem Snedecorove F-distribucije.

Pri tome je moguće postaviti sljedeću alternativnu (H1), odnosno nultu (H0) hipotezu:

- H0: $\rho = 0$ - Uz pogrešku p ne možemo tvrditi da je povezanost između skupa prediktora i kriterijske varijable statistički značajna.
- H1: $\rho \neq 0$ - Povezanost između skupa prediktora i kriterijske varijable je statistički značajna uz pogrešku p.

RStudio – Regression Analysis

Zadatak: U datoteci *SKOLA.csv* utvrdite relacije između morfoloških obilježja (VISI, TEZI, OBPO, NABN), testova za procjenu motoričkih sposobnosti (TAPI, POLI, SDAL, POTR, PRRA) i testa *trčanje 6 minuta* (TR6M).

0.83
Multiple Correlation

0.68
Coefficient of Determination

133.48
Standard Error of the Estimate

18.97
F - value

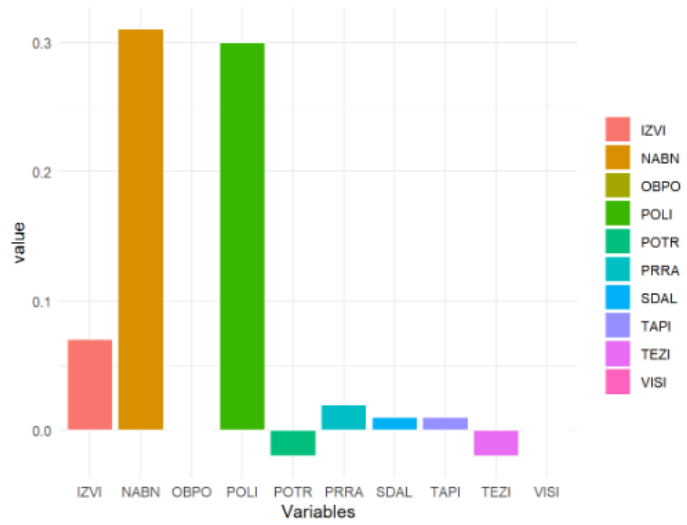
0
p - level

Results of Regression Analysis

Copy

	B	SE(B)	Beta	Part_R	R	P	Tolerance	t-value	p-value
B0	1,665.784	486.731	0.000	0.000	0.000	0.000	0.000	3.422	0.001
VISI	-1.162	2.475	-0.034	-0.050	-0.056	0.002	0.695	0.470	0.640
TEZI	2.960	2.613	0.095	0.119	-0.203	-0.019	0.508	1.133	0.260
OBPO	-3.373	13.259	-0.019	-0.027	-0.260	0.005	0.670	0.254	0.800
NABN	-27.319	4.579	-0.472	-0.535	-0.657	0.310	0.574	5.967	0.000
TAPI	1.695	3.279	0.037	0.055	0.164	0.006	0.687	0.517	0.607
POLI	-16.770	3.191	-0.455	-0.487	-0.648	0.295	0.479	5.256	0.000
SDAL	0.338	0.844	0.031	0.042	0.319	0.010	0.605	0.401	0.689
POTR	-1.615	2.720	-0.046	-0.063	0.400	-0.018	0.595	0.594	0.554
PRRA	0.933	1.559	0.049	0.063	0.493	0.024	0.533	0.598	0.551
IZVI	1.633	0.969	0.130	0.176	0.511	0.067	0.599	1.685	0.096

Graph of Partial Coefficient of Determinations (P)



Graph of Standardized Regression Coefficients, Partial Correlations and Correlations