Osnove statistike u demografiji

Predavanje 10

Analiza odnosa varijabli

- Univarijatna analiza: jedna varijabla (obilježje)
  - Mjere deskriptivne statističke analize
    - Mjere centralne tendencije
    - Mjere disperzije
    - Mjere asimetrije
  - Metode inferencijalne statistike:
    - Testiranje hipoteza i procjena parametara jedne populacije
    - Usporedba parametara **iste** varijable za različite populacije (grupe)
- Jedna varijabla: prosječna dob (jedne populacija ili usporedba više populacija
- Pored univarijatne analize u demografskim istraživanjima se često analizira i odnos varijabli

Analiza odnosa varijabli

- Odnos varijabli:
    - Plaća i dob
    - Plaća i spol
- Odnos dvije varijable: bivarijatna analiza
- Odnos više od dvije varijable: multivarijatna analiza
- Metode analize ovise o metričkim svojstvima varijabli
- Numeričke varijable

Analiza odnosa varijabli

- Najzastupljenije metode u analizi statističke povezanosti varijabli su:
  - korelacijska i
  - regresijska analiza.
- Obje metode analiziraju linearnu povezanost varijabli.
- Pritom, u korelacijskoj se analizi utvrđuje smjer i jakost povezanosti dviju slučajnih varijabli x i y koje se tretiraju simetrično.
- S druge strane, u regresijskoj analizi se pretpostavlja odnos između varijabli, tj. varijable se dijele na zavisnu (varijabla y) i nezavisnu varijablu (varijabla x).

Korelacijska analiza

- Pretpostavka o linearnoj povezanosti pojava česta je u empirijskim istraživanjima.
- Linearnu povezanost dviju pojava moguće je analizirati:

1. izračunavanjem odgovarajuće mjere povezanosti ili
2. grafičkim putem (dijagramom rasipanja).

- Najčešće korišteni pokazatelji statističke povezanosti među pojavama su:
  - Pearsonov koeficijent korelacije, kojim se analizira linearna povezanost dviju numeričkih varijabli, te
  - Spearmanov i Kendallov koeficijent korelacije ranga kojima se analizira stupanj povezanosti rangova dviju varijabli

Dijagram rasipanja

- Dijagram rasipanja (engl. *scatter plot*) je grafički prikaz točaka u pravokutnom koordinatnom sustavu na temelju kojeg se analizira povezanost dviju varijabli
- Točke (parovi vrijednosti varijabli x i y) se crtaju u pravokutnom koordinatnom sustavu s aritmetičkim mjerilom za vrijednosti x_i na osi apscisa i aritmetičkim mjerilom za vrijednosti y_i na osi ordinata
- Aritmetička mjerila ne moraju imati nulu kao početnu vrijednost
- Analizom oblika “raspršenosti” točaka utvrđuje se oblik, smjer i intenzitet povezanosti dviju pojava

Pearsonov koeficijent korelacije

- Najčešće korištena mjera jakosti i smjera linearne statističke povezanosti dviju varijabli.
- Računa se kao omjer kovarijance dviju varijabli i umnoška njihovih standardnih devijacija.
- Kovarijanca je mjera zajedničke varijabilnosti među varijablama *x* i *y* $Cov(x,y) = E[(x - E(x))(y - E(y))] = E[xy] - E(x)E(y)$
- Kovarijanca je apsolutna mjera povezanosti čije vrijednosti ovise o mjernim jedinicama promatranih varijabli (pojava).
  - Ako ne postoji linearna povezanost varijabli, kovarijanca je jednaka nuli.
  - Ako se (u prosjeku) povećanjem vrijednosti x_i povećavaju i vrijednosti y_i , kovarijanca je pozitivna.
  - Povezanost velikih vrijednosti x_i s malim vrijednostima y_i , rezultirat će negativnom vrijednošću kovarijance

Pearsonov koeficijent korelacije

- Procjena kovarijance varijabli x i y na temelju uzorka parova vrijednosti varijabli, definira se kao prvi mješoviti moment

- Standardizirana mjera jakosti i smjera linearne statističke povezanosti dviju varijabli je Pearsonov koeficijent korelacije

- Koeficijent linearne korelacije poprima vrijednosti iz intervala [−1, 1]

Regresijska analiza

- Regresijska analiza je najčešće korištena metoda u empirijskim istraživanjima.
- Njome se želi kvantificirati odnos dviju (ili više) pojava.
- Osnova regresijske analize je regresijski model.
- Regresijski model je hipotetički model (formula) kojim se izražava statistička (stohastička) povezanost između pojava.
- Na temelju uzorka vrijednosti odabranih varijabli procjenjuju se parametri pretpostavljenog modela i testiraju pretpostavke kako bi se odredila adekvatnost procijenjenog modela.
- Ako je procijenjeni model adekvatan, koristi se za testiranje ekonomske teorije i u prognostičke svrhe.

Regresijska analiza

- Model jednostavne linearne regresije pretpostavlja linearnu povezanost između zavisne varijable y i jedne eksplanatorne (nezavisne) varijable x.
- Model višestruke regresije: varijacije zavisne varijable opisuju se većim brojem eksplanatornih (nezavisnih) varijabli

Model jednostavne linearne regresije

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

- β_0 i β_1 regresijski parametri (parametri populacije) koji se procjenjuju na osnovi opažanja (mjerenja) varijabli
- Varijabla ε naziva se greška relacije i komponenta je koja modelu daje statistički (stohastički) karakter.
  - Slučajna varijabla s nepoznatom funkcijom gustoće vjerojatnosti.
  - Nemjerljiva i uključuje sve “ostale” faktore (osim varijable x) koji utječu na zavisnu varijablu y.
- Uključivanje varijable ε u model posljedica:
  - utjecaja varijabli koje nisu uključene u model,
  - utjecaja pogrešaka pri mjerenju vrijednosti zavisne varijable,
  - nepredvidivih ekonomskih ili prirodnih utjecaja na zavisnu varijablu kao što je npr. donošenje novih mjera ekonomske politike, prirodne katastrofe i slično.

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

- Najjednostavniji oblik modela u kojem se pretpostavlja aditivnost determinističke i slučajne komponente
- Linearan regresijski model
- Linearnost modela u pravilu se odnosi na način na koji se parametri i varijable pojavljuju u modelu
- U statističkim analizama, pojam linearan regresijski model podrazumijeva linearnost u parametrima (parametri se ne potenciraju)

Polazne pretpostavke u analizi modela jednostavne linearne regresije

1. Model populacije (osnovnog skupa) je linearan.

- Cilj regresijske analize je da se na temelju *n* opažanja varijabli procijene nepoznati regresijski parametri (parametri populacije). U tu svrhu potrebno je uvesti teorijske pretpostavke o greškama relacije (slučajnoj varijabli) ε

Procjene parametara u modelu jednostavne linearne regresije

- Uz pretpostavku da se povezanost varijabli opisuje linearnom funkcijom, tj. da je model populacije

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

- zadatak regresijske analize je da se pronađu procjene nepoznatih parametara (parametara populacije) β_0 i β_1, te procjena nepoznate varijance σ^2 slučajnih varijabli ε_i
- U tu svrhu potrebno je odabrati slučajni uzorak od n parova vrijednosti varijabli.

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\varepsilon}_i, \quad i = 1, 2, \ldots, n.$$

- Procijenjen model na temelju uzorka je

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \ldots, n,$$

$$y_i = \hat{y}_i + \hat{\varepsilon}_i, \quad i = 1, 2, \ldots, n.$$

Procjene parametara u modelu jednostavne linearne regresije

- Razlika između stvarne vrijednosti zavisne varijable i njene procijenjene vrijednosti,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \ldots, n,$$

- je rezidualno odstupanje i procjena je vrijednosti varijable εi

Metoda najmanjih kvadrata

- Odabire one procjene parametara za koje će zbroj kvadrata rezidualnih odstupanja biti najmanji.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \rightarrow \textbf{minimum}.$$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \rightarrow \textbf{minimum}.$$

- dobiva se sustav normalnih jednadžbi čijim se rješavanjem dolazi do izraza za izračunavanje regresijskih koeficijenata (procjena nepoznatih parametara) koristeći empirijske vrijednosti iz uzorka

Procijenjeni model (model uzorka)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- $\hat{\beta}_1$ Regresijski koeficijent
- $\hat{\beta}_0$. Konstantni član
- Ako su ispunjene polazne pretpostavke o modelu jednostavne linearne regresije, procjenitelji imaju dobra statistička svojstva:
    - nepristrani su i
    - imaju najmanju varijancu u klasi svih linearnih procjenitelja (efikasni su).
- Za navedena svojstva procjenitelja pretpostavka o normalnoj distribuiranosti grešaka relacije nije nužna.
- Ta je pretpostavka nužna u kasnijim koracima analize kako bi se odredile intervalne procjene parametara i testirale hipoteze.

Definicija modela i procjene parametara

- Pretpostavlja se da je zavisna varijabla *y* linearna funkcija *k* nezavisnih varijabli, odnosno da je model populacije

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_k x_k + \varepsilon$$

- Slučajna varijabla *ε* je greška relacije koja nije mjerljiva i njenim uključivanjem u model u obliku aditivnog člana, model postaje stohastički (statistički).
- Ako se jednadžba procjenjuje na osnovi *n* opažanja (mjerenja) varijabli, jednadžba se može zapisati kao sustav od n jednadžbi

Pretpostavke

odnosno, za svako i, $i = 1, 2, \dots, n$ vrijedi da je

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

- Pretpostavke modela višestruke linearne regresije su jednake pretpostavkama za model jednostavne linearne regresije uz dodatnu pretpostavku da su regresorske varijable nekorelirane, tj. da nisu linearno povezane.

1. Veza između zavisne varijable i odabranog skupa nezavisnih varijabli je linearna (u parametrima)

Definicija modela i procjene parametara

- Parametri regresijskog modela procjenjuju se metodom najmanjih kvadrata.
- U tu svrhu potrebno je odabrati slučajni uzorak
- U procjeni se polazi od osnovnog modela (modela populacije)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_k x_{ik} + \varepsilon_i. \qquad \text{za svaki } i = 1, 2, \ldots, n,$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_j x_{ij} + \cdots + \hat{\beta}_k x_{ik}, \quad i = 1, 2, \ldots, n$$

$$y_i = \hat{y}_i + \hat{\varepsilon}_i \quad i = 1, 2, \ldots, n.$$

- Razlika između stvarne vrijednosti zavisne varijable i njene procijenjene vrijednosti je rezidualno odstupanje

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_j x_{ij} + \cdots + \hat{\beta}_k x_{ik}), \quad i = 1, 2, \ldots, n,$$

Metoda najmanjih kvadrata

- Odabire procjene parametara za koje će zbroj kvadrata rezidualnih odstupanja biti najmanji

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \rightarrow \textbf{minimum},$$

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j x_{ij} + \cdots + \hat{\beta}_k x_{ik}) \right)^2 \rightarrow \textbf{minimum}.$$

- Minimiziranjem rezidualnog zbroja kvadrata, iz nužnog uvjeta za minimum dobiva se sustav (k + 1) normalnih jednadžbi nepoznatih parametara

Primjer

- Analizira se odnos broja stanovnika (u milijunima, oznaka pop) Danske u odnosu na skup odabranih nezavisnih varijabli u razdoblju od 1990 do 2017 godine (izvor: EUROSTAT)
- Odabrane nezavisne varijable su:
- Broj iseljenika (em) u tis.
- Broj useljenika (im) u tis.
- Pokazatelj fertiliteta (fert)
- Prosječna dob majki pri porodu (dob)
- Broj sklopljenih brakova (brak) u tis.
- Broj razvedenih brakova (raz) u tis.
- Bruto domaći proizvod (bdp) u mlrd EUR

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,996383
R Square	0,992778
Adjusted R Square	0,990251
Standard Error	0,017551
Observations	28

ANOVA

	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95,0%	Upper 95,0%
Intercept	4,168317	0,971706	4,289691	0,000357	2,141374	6,19526	2,141374	6,19526
em	0,004531	0,001389	3,263417	0,003889	0,001635	0,007428	0,001635	0,007428
im	0,000753	0,00059	1,276592	0,216371	-0,00048	0,001983	-0,00048	0,001983
fert	-0,03183	0,103469	-0,3076	0,761564	-0,24766	0,184006	-0,24766	0,184006
dob	0,03085	0,035853	0,860458	0,399735	-0,04394	0,105638	-0,04394	0,105638
brak	-0,00764	0,001841	-4,15181	0,000493	-0,01149	-0,0038	-0,01149	-0,0038
raz	-0,00168	0,003681	-0,45714	0,652497	-0,00936	0,005995	-0,00936	0,005995
bdp	0,002102	0,000543	3,869618	0,000954	0,000969	0,003235	0,000969	0,003235

Interpretacija parametara regresijskog modela

- Procijenjeni modela (model uzorka)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_k x_k.$$

- $\hat{\beta}_0$ konstantni član.
  - Označava procijenjenu (prosječnu) vrijednost zavisne varijable ako su vrijednosti svih regresorskih varijabli jednake nuli.
  - Premda konstantni član najčešće nema suvislu interpretaciju, on se uvijek uključuje u model
  - Izuzetak je analiza modela na temelju transformiranih vrijednosti varijabli (standardiziranje, prve diferencije, logaritamske transformacije) ili ako se to eksplicitno pretpostavlja na temelju ekonomske teorije

Interpretacija procjena parametara regresijskog modela

- Uključivanje konstante osigurava da je prosječna vrijednost reziduala jednaka nuli što je nužan uvjet metode najmanjih kvadrata.
- Nadalje, uloga konstantnog člana u modelu povezana je i s jednadžbom regresijskog pravca.
- Naime, kada bi regresijski pravac bio definiran bez konstantnog člana, geometrijski bi to značilo da pravac prolazi ishodištem

Interpretacija parametara regresijskog modela

- Procijenjeni model (model uzorka)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_k x_k.$$

$$\hat{\beta}_j = \frac{\partial \hat{y}}{\partial x_j}$$

- $\hat{\beta}_1$ regresijski koeficijent .
  - procjena regresijskog parametra uz *j*-tu regresorsku varijablu
  - vrijednost parcijalnog utjecaja varijable x_j na $\hat{y}$
  - promjena prosječne (regresijske) vrijednosti zavisne varijable, za jedinično povećanje vrijednosti varijable x_j , uz uvjet da su vrijednosti ostalih regresorskih varijabli nepromijenjene

	Coefficients	Standard Error	t Stat	P-value
Intercept	4,168317	0,971706	4,289691	0,000357
em	0,004531	0,001389	3,263417	0,003889
im	0,000753	0,00059	1,276592	0,216371
fert	-0,03183	0,103469	-0,3076	0,761564
dob	0,03085	0,035853	0,860458	0,399735
brak	-0,00764	0,001841	-4,15181	0,000493
raz	-0,00168	0,003681	-0,45714	0,652497
bdp	0,002102	0,000543	3,869618	0,000954

$$\widehat{pop} = 4{,}17 + 0{,}004 \cdot em + 0{,}0007 \cdot im - 0{,}03 \cdot fert + 0{,}03 \cdot dob - 0{,}0076 \cdot brak - 0{,}0017 \cdot raz + 0{,}002 \cdot bdp$$

Standardne pogreške procjena parametara u modelu višestruke linearne regresije

- Kako bi se odredila preciznost procjena regresijskih parametara, potrebno je definirati mjeru preciznosti procjene, tj. standardnu pogrešku procjene
- Ako su ispunjene sve pretpostavke o regresijskom modelu sampling-distribucija procjenitelja regresijskog parametra je $\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j})$
- Varijanca procjenitelja ovisi o:
  - veličini uzorka n,
  - disperzijama (varijancama) nezavisnih varijabli,
  - korelaciji između nezavisnih varijabli i
  - varijanci regresije.
- U slučaju modela s dvjema nezavisnim varijablama

Standardne pogreške procjena parametara u modelu višestruke linearne regresije

- Ako je varijanca regresije nepoznata, zamjenjuje se (nepristranom) procjenom na bazi uzorka

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

- a standardizirana varijabla pridružena procjenitelju

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim t\left(n-(k+1)\right)$$

- Ako je uzorak *n* dovoljno velik, tada prema centralnom graničnom teoremu aproksimativno vrijedi

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0,1).$$

Intervalne procjene parametara u modelu višestruke linearne regresije

- Intervalna procjena regresijskog parametra je interval koji uz zadanu pouzdanost uključuje stvarnu vrijednost parametra.
- Granice intervala procjene ovise o obliku sampling-distribucije procjenitelja
- Uz razinu pouzdanosti (1−α) , intervalna procjena regresijskog parametra je

$$P(\hat{\beta}_j - t_{\alpha/2}\sigma_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + t_{\alpha/2}\sigma_{\hat{\beta}_j}) = 1 - \alpha$$

- Ako je:
  - uzorak dovoljno velik ili
  - ako je varijanca populacije poznata,
- umjesto koeficijenta pouzdanosti $t_{\alpha/2}$ uzima se odgovarajući percentil jedinične normalne distribucije $z_{\alpha/2}$.

Primjer: Intervalne procjene parametara

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	4,168317	0,971706	4,289691	0,000357	2,141374	6,19526
em	0,004531	0,001389	3,263417	0,003889	0,001635	0,007428
im	0,000753	0,00059	1,276592	0,216371	-0,00048	0,001983
fert	-0,03183	0,103469	-0,3076	0,761564	-0,24766	0,184006
dob	0,03085	0,035853	0,860458	0,399735	-0,04394	0,105638
brakovi	-0,00764	0,001841	-4,15181	0,000493	-0,01149	-0,0038
raz	-0,00168	0,003681	-0,45714	0,652497	-0,00936	0,005995
bdp	0,002102	0,000543	3,869618	0,000954	0,000969	0,003235

$$P(\hat{\beta}_j - t_{\alpha/2}\sigma_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + t_{\alpha/2}\sigma_{\hat{\beta}_j}) = 1 - \alpha$$

$$P(0{,}000969 < \beta_7 < 0{,}003235) = 0{,}95$$

Analiza varijance u modelu jednostavne linearne regresije

- Je li procijenjeni regresijski model reprezentativan:
- Koliko dobro varijabla x "objašnjava" zavisnu varijablu y , tj. koliki je dio varijabilnosti varijable y objašnjen pretpostavljenom linearnom vezom s varijablom x
- Kako bi se odredilo koliko dobro varijabla x objašnjava varijaciju zavisne varijable y , tj. koliko je procijenjeni regresijski model dobar, polazi se od rastava varijance zavisne varijable procijenjene na bazi uzorka na:
  - dio varijance protumačen modelom i
  - rezidualni dio, tj. dio varijance neprotumačen modelom.

$$(y_i - \overline{y}) = (\hat{y}_i - \overline{y}) + (y_i - \hat{y}_i)$$

Analiza varijance u modelu višestruke linearne regresije

- U modelu višestruke linearne regresije ukupna varijacija zavisne varijable y oko prosječne vrijednosti y nastoji se što bolje objasniti skupom nezavisnih varijabli
- U tu svrhu polazi se od rastava varijance zavisne varijable procijenjene na bazi uzorka na dvije komponente:
  - varijacije koje se mogu objasniti linearnom funkcijom nezavisnih varijabli i
  - varijacije koje ostaju neprotumačene
- Osnova analize je jednadžba analize varijance:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{ST}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SP}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SR}.$$

$$\boldsymbol{ST=SP+SR}$$

Tablica analize varijance (ANOVA)

Tablica 12.6. Analiza varijance za model višestruke linearne regresije – tablica ANOVA

Izvor varijacije	Stupnjevi slobode	Zbroj kvadrata	Sredina kvadrata	F-omjer
Protumačen modelom	k	SP	$\frac{SP}{k}$	$\dfrac{\frac{SP}{k}}{\frac{SR}{n-(k+1)}}$
Neprotumačen modelom	$n-(k+1)$	SR	$\frac{SR}{n-(k+1)}$	
Ukupno	$n-1$	ST		

- Procjena varijance regresije

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

- Procjena standardne devijacije regresije

$$\hat{\sigma} = \sqrt{\frac{SR}{n-(k+1)}}$$

- Pridružena relativna mjera je procjena koeficijenta varijacije regresije

$$\hat{V} = \frac{\hat{\sigma}}{\overline{y}} \cdot 100\ \%$$

Regresijske vrijednosti i rezidualna odstupanja

- Regresijske vrijednosti dobivaju se uvrštavanjem odgovarajućih empirijskih vrijednosti nezavisnih varijabli u procijenjenu regresijsku jednadžbu.
- Regresijske vrijednosti su procjene dobivene na temelju uzorka i razlikuju se od stvarnih vrijednosti zavisne varijable.
- Razlika je rezidualno odstupanje i procjena je greške relacije

$$\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

- Rezidualna odstupanja izražena u mjernim jedinicama zavisne varijable y (apsolutno rezidualno odstupanje).
- Pripadna relativna rezidualna odstupanja su

$$\hat{\varepsilon}_{i,rel} = \frac{\hat{\varepsilon}_i}{y_i} \cdot 100\ \% = \frac{y_i - \hat{y}_i}{y_i} \cdot 100\ \%.$$

Procjena varijance i standardne devijacije regresije

- Nepristrana procjena varijance regresije

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

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

- Procjena standardne devijacije regresije
  - interpretira se kao prosječno odstupanje empirijskih vrijednosti zavisne varijable od regresijskih (procijenjenih) vrijednosti
  - interpretira se kao prosječno odstupanje empirijskih vrijednosti zavisne varijable (yi) od regresijskih (procijenjenih) vrijednosti
- Odgovarajuća relativna mjera disperzije je procjena koeficijenta varijacije regresije

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

Koeficijent determinacije R^2

$$ST = SP + SR$$

$$R^2 = \frac{SP}{ST} = 1 - \frac{SR}{ST}$$

- Koeficijent determinacije:
- proporcija varijacije varijable y protumačena modelom
- Najčešće korištena mjera reprezentativnosti regresijskog modela.
- poprima vrijednosti iz intervala [0, 1]
- mogu se koristiti za usporedbu regresijskih modela samo ako su modeli procijenjeni na temelju istog uzorka empirijskih vrijednosti

Koeficijent determinacije

- Ako je veliki dio varijabilnosti zavisne varijable protumačen modelom, $SP \approx ST$, te je $R^2 \approx 1$.
- S druge strane, ako je zanemariv dio varijabilnosti zavisne varijable protumačen modelom, tada je $SP \approx 0$ pa je $R^2 \approx 0$
- oprez pri interpretaciji!
- Empirijske analize pokazuju da je u analizi vremenskih nizova vrijednost R^2 obično vrlo visoka (0,8 i veća).
- Ako se regresijska analiza provodi na prostornim podatcima vrijednost je obično između 0,4 i 0,6.
- Modeli koji uključuju varijable koje se odnose na pojedine ljudske karakteristike obično imaju vrlo niske vrijednosti (između 0,1 i 0,2).
- Monotono neopadajuća funkcija broja nezavisnih varijabli k , te se povećanjem broja regresorskih varijabli u modelu njegova vrijednost povećava

Korigirani koeficijent determinacije

- Jedan od nedostataka koeficijenta determinacije je da se njegova vrijednost povećava s brojem nezavisnih varijabli u modelu, bez obzira na proporciju varijacija zavisne varijable koje objašnjavaju.
- Budući da je osnovna ideja regresijske analize parsimonija, odnosno da se uz što manje nezavisnih varijabli objasni što više varijacija zavisne varijable y, promatra se korigirani koeficijent determinacije

$$\overline{R}^2 = 1 - \frac{n-1}{n-(k+1)}(1-R^2), \quad \overline{R}^2 \leqslant R^2$$

- "Kažnjava" uključivanje u model regresorskih varijabli koje zanemarivo (ili nedovoljno) smanjuju rezidualni zbroj kvadrata

Primjer: ANOVA

- Analizirajte tablicu analize varijance (ANOVA)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,996383
R Square	0,992778
Adjusted R Square	0,990251
Standard Error	0,017551
Observations	28

ANOVA

	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			

ANOVA					
	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			

$$\frac{SP}{k} = \frac{0,846896}{7} = 0,120985$$

$$k = 7$$

$$n - (k+1) = 28 - 8 = 20$$

$$\frac{SR}{n-(k+1)} = \frac{0,00616}{20} = 0,000308$$

$$ST = SP + SR$$
$$0,853057 = 0,846896 + 0,00616$$

$$F = \frac{\frac{SP}{k}}{\frac{SR}{n-(k+1)}} = \frac{\frac{0,846896}{7}}{\frac{0,00616}{20}} = \frac{0,120985}{0,000308} = 392,7823$$

Primjer: pokazatelji reprezentativnosti

- Analizirajte pokazatelje reprezentativnosti modela

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,996383
R Square	0,992778
Adjusted R Square	0,990251
Standard Error	0,017551
Observations	28

$$R^2 = \frac{SP}{ST} = \frac{0,846896}{0,853057} = 0,992778$$

ANOVA

	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			

Testiranje hipoteza u modelu višestruke linearne regresije

1. test značajnosti (jedne) regresorske varijable (pojedinačni t -test)
2. test značajnosti regresije (skupni test značajnosti svih regresorskih varijabli, F -test)
3. test značajnosti podskupa regresorskih varijabli (parcijalni F -test).

- Postupak testiranja bazira se na obliku sampling-distribucije standardiziranog procjenitelja
- Za danu razinu značajnosti α , testna veličina se uspoređuje s teorijskom (kritičnom) vrijednosti odgovarajuće sampling-distribucije.
- U postupku testiranja polazi se od modela (populacije),

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_j + \cdots + \beta_k x_k + \varepsilon,$$

- i procijenjenog modela na bazi uzorka
-
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_k x_k.$$

t -test značajnosti regresorske varijable

- Ako su ispunjene polazne pretpostavke o modelu i ako je H_0 istinita, testna veličina

$$t_j = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_j}} \sim t(n-(k+1)).$$

p-vrijednost ili empirijska razina značajnosti

- Odluka o ishodu testa može se donijeti i usporedbom empirijske i teorijske razine značajnosti.
- p-vrijednost je vjerojatnost da test veličina koja, uz pretpostavku da je H_0 istinita, ima t -distribuciju s $df = n - 2$ stupnjeva slobode, poprimi vrijednost jednaku ili veću od apsolutne vrijednosti testne veličine izračunane na osnovi podataka iz uzorka $p\text{-vrijednost} < \alpha \implies H_1$

Primjer: Pojedinačni testovi

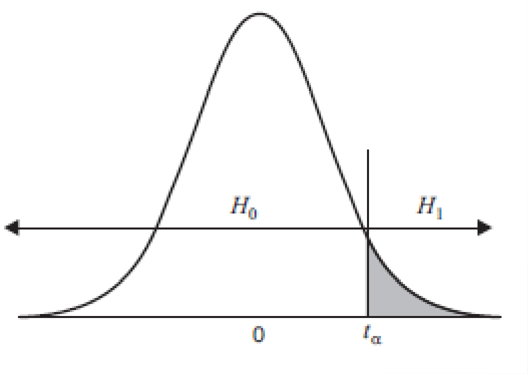
- Hipoteze jednosmjernih testova?
- Za pojedine varijable, naznačite pripadne p-vrijednosti jednosmjernih testova
- Na razini značajnosti 5%, koje varijable su statistički značajne?

	Coefficients	Standard Error	t Stat	P-value
Intercept	4,168317	0,971706	4,289691	0,000357
em	0,004531	0,001389	3,263417	0,003889
im	0,000753	0,00059	1,276592	0,216371
fert	-0,03183	0,103469	-0,3076	0,761564
dob	0,03085	0,035853	0,860458	0,399735
brak	-0,00764	0,001841	-4,15181	0,000493
raz	-0,00168	0,003681	-0,45714	0,652497
bdp	0,002102	0,000543	3,869618	0,000954

$$\widehat{pop} = 4{,}17 + 0{,}004 \cdot em + 0{,}0007 \cdot im - 0{,}03 \cdot fert + 0{,}03 \cdot dob - 0{,}0076 \cdot brak - 0{,}0017 \cdot raz + 0{,}002 \cdot bdp$$

$$H_0: \beta_7 = 0$$

$$H_1: \beta_7 > 0$$



$$p - vrijednost = \frac{0{,}000954}{2} = 0{,}000477$$

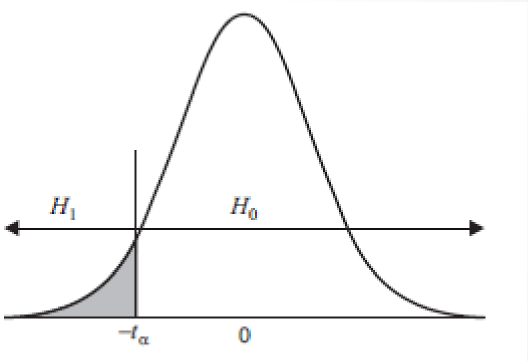
Za $\alpha = 0{,}05$, p-vrijednost $< \alpha \implies H_1$.

	Coefficients	Standard Error	t Stat	P-value
Intercept	4,168317	0,971706	4,289691	0,000357
em	0,004531	0,001389	3,263417	0,003889
im	0,000753	0,00059	1,276592	0,216371
fert	-0,03183	0,103469	-0,3076	0,761564
dob	0,03085	0,035853	0,860458	0,399735
brak	-0,00764	0,001841	-4,15181	0,000493
raz	-0,00168	0,003681	-0,45714	0,652497
bdp	0,002102	0,000543	3,869618	0,000954

$$\widehat{pop} = 4{,}17 + 0{,}004 \cdot em + 0{,}0007 \cdot im - 0{,}03 \cdot fert + 0{,}03 \cdot dob - 0{,}0076 \cdot brak - 0{,}0017 \cdot raz + 0{,}002 \cdot bdp$$

$H_0: \beta_6 = 0$

$H_1: \beta_6 < 0$



$$p - vrijednost = \frac{0{,}652497}{2} = 0{,}326249$$

$$p - vrijednost > \alpha \rightarrow H_1$$

	Coefficients	Standard Error	t Stat	P-value
Intercept	4,168317	0,971706	4,289691	0,000357
em	0,004531	0,001389	3,263417	0,003889
im	0,000753	0,00059	1,276592	0,216371
fert	-0,03183	0,103469	-0,3076	0,761564
dob	0,03085	0,035853	0,860458	0,399735
brak	-0,00764	0,001841	-4,15181	0,000493
raz	-0,00168	0,003681	-0,45714	0,652497
bdp	0,002102	0,000543	3,869618	0,000954

F-test značajnosti regresije

- Test značajnosti regresije je skupni test o značajnosti svih regresorskih varijabli.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \exists \beta_j \neq 0, \quad j = 1, 2, \ldots, k$$

- Ako su ispunjene polazne pretpostavke o modelu i ako je H_0 istinita, testna veličina

$$F = \frac{\frac{SP}{k}}{\frac{SR}{n-(k+1)}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-(k+1)}}$$

- pripada F-distribuciji s $df_1 = k$ stupnjeva slobode u brojniku i $df_2 = n - (k + 1)$ stupnjeva slobode u nazivniku
- Uz razinu značajnosti a, odluka o ishodu testa donosi se usporedbom testne veličine F i teorijske vrijednosti F distribucije s (k, n − (k + 1)) stupnjeva slobode.
- H_0 se odbacuje ako je $F > F^{\alpha}_{(k,n-(k+1))}$

Primjer *F*-test

ANOVA					
	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			

$H_0: \beta_1 = \beta_2 = \cdots = \beta_7 = 0$

$H_1: \exists \beta_j \neq 0, j = 1,2, \ldots, 7$

$p - vrijednost = 5{,}38 \cdot 10^{-20}$ Za $\alpha = 0{,}05$, p-vrijednost $< \alpha \implies H_1$.

Korelacijska matrica

- U modelu višestruke linearne regresije, koji uključuje zavisnu varijablu y i k nezavisnih varijabli, za svaki par varijabli može se izračunati koeficijent linearne korelacije, tj. pokazatelj smjera i jakosti linearne statističke povezanosti dviju varijabli.

Korelacijska matrica

- Dobivene vrijednosti zapisuju se u korelacijskoj matrici R oblika

Primjer: korelacija

- Interpretirajte vrijednosti koeficijenata korelacije

	pop	em	im	fert	dob	brakovi	raz	bdp
pop	1							
em	0,822645	1						
im	0,774462	0,670238	1					
fert	0,111644	0,088996	0,061183	1				
dob	0,978242	0,817136	0,744308	0,219267	1			
brakovi	-0,43081	-0,03103	-0,24026	0,451796	-0,31491	1		
raz	0,760095	0,610075	0,587805	-0,19606	0,752883	-0,41562	1	
bdp	0,988107	0,81481	0,765067	0,20836	0,99154	-0,34744	0,747643	1

Koeficijent višestruke linearne korelacije

- Standardizirana mjera jakosti, linearne statističke povezanosti zavisne varijable y i skupa nezavisnih varijabli

$$R = \sqrt{R^2}$$

- Vrijednost koeficijenta je uvijek nenegativna i ne pridružuje mu se predznak.
- Naime, povezanost varijable y i pojedine nezavisne varijable iz skupa od k nezavisnih varijabli može biti različitog smjera.
- Vrijednost R ≈ 0 označava zanemarivu, a R ≈ 1 jaku povezanost varijabli

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,996383
R Square	0,992778
Adjusted R Square	0,990251
Standard Error	0,017551
Observations	28

$R = 0{,}996383$

ANOVA

	df	SS	MS	F	Significance F
Regression	7	0,846896	0,120985	392,7823	5,38E-20
Residual	20	0,00616	0,000308		
Total	27	0,853057			