11. Jednostavna linearna regresijska analiza

TEME

1.	Teorijski uvod o analizi
2.	Provedba jednostavne linearne regresijske analize u JASP-u
3.	Provjere pretpostavki
4.	Interval pouzdanosti za regresijski koeficijent

11.1. Teorijski uvod o analizi

Jednostavna linearna regresijska analiza dijelom je nalik korelacijskoj analizi utoliko što podrazumijeva analizu međuodnosa kvantitativnih varijabli. No, za razliku od korelacijske analize koja nam govori o smjeru i jačini povezanosti (obično) dviju varijabli, ako postoji linearna povezanost između dvije varijable, linearna regresija omogućuje predikciju vrijednosti zavisne varijable pomoću vrijednosti nezavisne varijable (što korelacijskom analizom nije moguće).

Također, za razliku od korelacijske analize koja ne razlikuje zavisnu i nezavisnu varijablu, prije provedbe linearne regresijske analize istraživač mora sam odrediti (obično vođen teorijskim pretpostavkama) koju će varijablu tretirati kao zavisnu, a koju kao nezavisnu. U toj analizi zavisna varijabla mora biti kvantitativna (npr. bodovi na testu, visina, mjesečna primanja itd.), dok nezavisna varijabla može biti kvantitativna ili dihotomna nominalna.

Terminologija (hrv.)	Nezavisna varijabla	Zavisna varijabla
	Prediktor	Kriterij

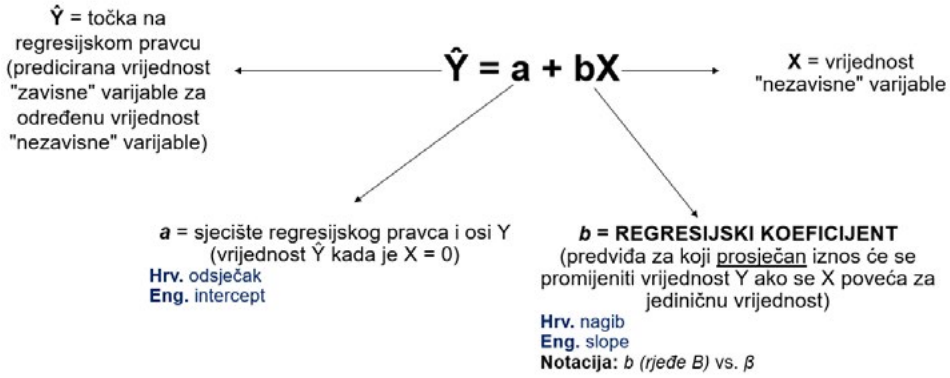
Terminologija (engl.)	Independent variable (IV)	Dependent variable (DV)
	Predictor	Criterion
	Explanatory variable	Outcome
	Regressor variable	Target variable

Pitanja na koja linearna regresijska analiza može odgovori:
- Koliko (posto) varijance zavisne varijable jest protumačeno korelacijom s nezavisnom varijablom?
- Koji se rezultat očekuje na zavisnoj varijabli za određeni rezultat na nezavisnoj varijabli?
- Koliko dobro nezavisna varijabla predviđa vrijednosti zavisne varijable?

Odgovor na pitanje ''Koliko (posto) varijance zavisne varijable jest protumačeno korelacijom s nezavisnom varijablom?'' dobivamo putem **koeficijenta determinacije** (R^2). Za razliku od koeficijenata korelacije koji nam govore o jačini i smjeru povezanosti varijabli, koeficijent determinacije u linearnoj regresijskoj analizi govori o količini varijance zavisne varijable koju objašnjava jedna (u višestrukoj regresiji i više) nezavisna varijabla. Npr. ako je $R^2 = 0{,}40$, kažemo da je 40% varijance zavisne varijable protumačeno variranjem nezavisne varijable. Koeficijent determinacije ujedno je i **pokazatelj veličine učinka** (engl. *effect size*) modela koji dobivamo analizom. Ne postoje univerzalne smjernice za njegovu interpretaciju, već ona ovisi o predmetu analizu. Na primjer, možemo se pitati je li 40% puno ili malo protumačene varijance? To ovisi što smo s čime tumačili. U sociologiji se rijetko viđaju jednostavni linearni regresijski modeli s velikom količinom protumačene varijance pa tako primjerice ako utvrdimo da stupanj religioznosti tumači 40% varijacija varijable politička orijentacija, možemo to smatrati iznimno važnim istraživačkim nalazom.

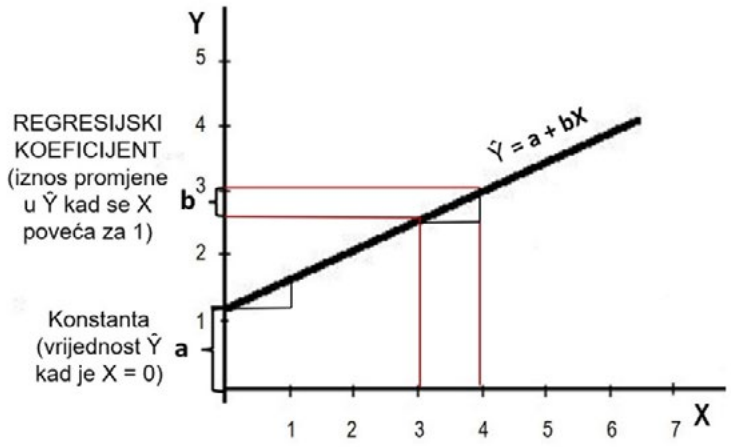
Linearna regresijska analiza temelji se na pronalasku odgovarajućeg **regresijskog pravca** kreiranog pomoću postojećih podataka. Linearna regresijska analiza primjerena je samo ako dijagram raspršenja ukazuje na linearnu povezanost varijabli, tj. samo se tada njihova zajednička distribucija može adekvatno sumirati regresijskim pravcem koji omogućuje najtočniju moguću predikciju rezultata zavisne varijable na temelju rezultata nezavisne varijable. Takav pravac koji najbolje opisuje neki set podataka utvrđuje se tzv. metodom najmanjih kvadrata, odnosno metodom koja pronalazi onaj pravac za koji je suma kvadriranih odstupanja (reziduala) pojedinih rezultata od pravca najmanja moguća.

Predikcija rezultata na zavisnoj varijabli u linearnoj je regresiji utemeljena na **regresijskoj jednadžbi,** tj. na jednadžbi pravca regresije Y na X:



Slika 1. Jednadžba regresijskog pravca

Upravo nam jednadžba regresijskog pravca daje odgovor na pitanje: koji se rezultat očekuje na zavisnoj varijabli za određeni rezultat na nezavisnoj varijabli?



Slika 2. Regresijski pravac u koordinatnom sustavu

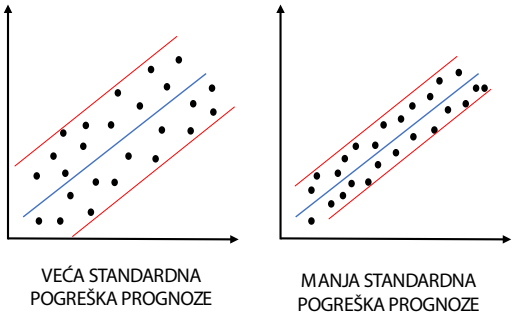
Iako se ne navodi u ovom obliku jednadžbe regresijskog pravca, za interpretaciju rezultata same analize biti će nam korisno razlikovati *b* i beta koeficijent:

- ***b koeficijent =*** nestandardizirani regresijski koeficijent (izražen u izvornoj metrici varijable)
- ***β (beta) koeficijent =*** standardizirani regresijski koeficijent (sve vrijednosti pretvorene u z-vrijednosti zbog usporedivosti, raspon od -1 do +1)

Kada bismo vrijednosti nekih dviju varijabli prvo standardizirali (pretvorili u z-vrijednosti) pa potom na njima proveli jednostavnu linearnu regresijsku analizu, vrijednosti *b* i beta koeficijenata bile bi identične.

Iznos beta koeficijenta nam zapravo odgovara na pitanje: Koliko dobro nezavisna varijabla predviđa vrijednosti zavisne varijable?

Za interpretaciju regresijskog modela važno je poznavati i pojam standardne pogreške prognoze. Standardna pogreška prognoze jest standardna devijacija distribucije odstupanja opaženih rezultata od rezultata predviđenih, odnosno prediciranih regresijskim pravcem. Kao mjera raspršenja rezultata oko pravca regresije, ona je zapravo mjera pogreške u predviđanju, odnosno mjera razlike između predviđenih i stvarnih vrijednosti.



Slika 3. Standardna pogreška prognoze

U jednostavnoj linearnoj regresiji testiramo nekoliko hipoteza.
Putem F-testa testiramo nultu hipotezu da varijanca zavisne varijable nije protumačena korelacijom s nezavisnom varijablom (**H_0: $R^2 = 0$**).
Putem t-testa testiramo nultu hipotezu da nema povezanosti između prediktora i kriterija, odnosno da nezavisna varijabla X u populaciji nema učinak na zavisnu varijablu Y (**H_0: $b = 0$**).
U višestrukoj (multiploj) linearnoj regresiji testiramo po jednu hipotezu za svaki regresijski koeficijent (b_1, b_2, ... b_k, gdje je k = broj prediktora u modelu).

Pretpostavke koje moraju biti zadovoljene da bi se smjela provoditi linearna regresijska analiza:

— kvantitativne normalno distribuirane varijable u linearnoj vezi
— bez ekstremnih vrijednosti
— homoskedastičnost (podaci trebaju biti jednako distribuirani oko regresijskog pravca na svim dijelovima pravca)
— nezavisnost opservacija (bez autokorelacije reziduala)
— nezavisnost prediktora (bez multikolinearnosti, tj. visokih korelacija među prediktorima)

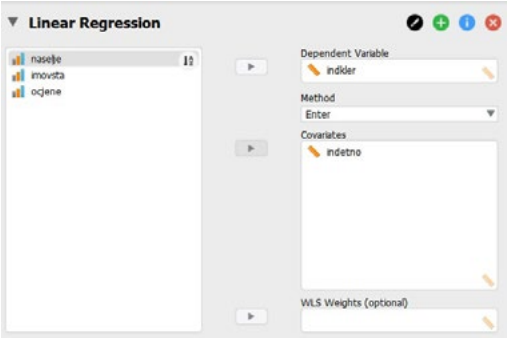
11.2. Provedba jednostavne linearne regresijske analize u JASP-u

Radimo na primjerima u datoteci pod nazivom:
11_jednostavna_linearna_regresija.sav

Želimo uz 1% rizika odrediti predviđaju li rezultati na varijabli INDETNO (indeks etnocentrizma na kojemu viši rezultat označava viši stupanj etnocentrizma) rezultate na varijabli INDKLER (indeks klerikalizma na kojemu viši rezultat označava viši stupanj klerikalizma) i ako da, koliko dobro.

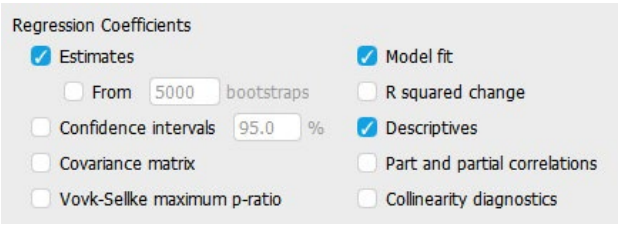
Regresijsku jednadžbu regresije varijable INDKLER (''zavisna'' varijabla ili **kriterij**) na varijablu INDETNO (''nezavisna'' varijabla ili **prediktor**) možemo odrediti pozivanjem procedure (*Classical*) *Linear Regression* u modulu *Regression.*

U okvir *Dependent Variable* prebaciti ćemo varijablu INDKLER, a u okvir *Covariates* varijablu INDETNO. Ako u okvir *Covariates* uvedemo više od jedne varijable, tada ćemo provesti višestruku (ili multiplu) linearnu regresiju (više o tome u sljedećem poglavlju).



Metodu izgradnje regresijskog modela (zadana: Enter) nema smisla mijenjati u jednostavnoj linearnoj regresiji, no ima u višestrukoj.

Prije interpretacije dobivenih rezultata, pod *Statistics* uključit ćemo opciju *Descriptives* jer će nam interpretacija deskripcije varijabli trebati prije interpretacije same regresijske analize.



Dobivene tablice sadržavaju sve ključne pokazatelje potrebne za interpretaciju provedene analize.

U dijelu ispisa naslovljenom *Model Summary* dobivamo pokazatelje za dva modela: prvi (u prvom retku) nulti je model bez prediktora (ne interpretiramo!), dok je drugi (u drugom retku) model s uključenim prediktorom/ima (njega interpretiramo!)

Dobivamo koeficijent korelacije (*R*), koeficijent determinacije *(R Square*), korigirani koeficijent determinacije (*Adjusted R Square*) te standardnu pogrešku prognoze (*RMSE*).

Korigirani koeficijent determinacije konzervativnija je procjena objašnjene varijance koja uzima u obzir veličinu uzorka i broj prediktora.

Pokazatelj RMSE (engl. *Root Mean Square Error*) standardna je pogreška prognoze, odnosno standardna devijacija reziduala (odstupanja opaženih rezultata od predviđenih, prediciranih regresijskim pravcem).

U našem primjeru vidimo da je (na podacima s uzorka) 30,4% varijance kriterija objašnjeno prediktorom u modelu.

Model Summary - indkler

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	5.226
H_1	0.553	0.306	0.304	4.360

ANOVA-tablica prezentira test statističke značajnosti koeficijenta determinacije, tj. testira se (nulta) hipoteza da koeficijent determinacije u populaciji iznosi 0, odnosno
H_0: $R^2 = 0$.

U našem primjeru odbacujemo nultu hipotezu ($F_{(1, 347)} = 4556{,}823$; $p < 0{,}001$) te uz 1% rizika zaključujemo da je u populaciji koju naš uzorak reprezentira varijanca zavisne varijable protumačena korelacijom s nezavisnom varijablom, tj. da koeficijent determinacije u populaciji ne iznosi 0. To zapravo znači da je model s prediktorom statistički značajan uz 1% rizika ili manje.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	2901.063	1	2901.063	152.641	< .001
	Residual	6575.994	346	19.006		
	Total	9477.057	347			

Note. The intercept model is omitted, as no meaningful information can be shown.

U tablici *Coefficients* gledamo H_1 model (model s prediktorom) i njegove parametre. Ovdje se nalaze svi elementi za rješavanje regresijske jednadžbe, čiji je opći oblik:

$\hat{Y} = a + bX$

U našem slučaju: INDKLER = a + b * INDETNO

U tablici koeficijenata nalazimo da **a** (engl. *Constant; Intercept*) iznosi 2,107, a **b** (nestandardizirani regresijski koeficijent; engl. *Unstandardized Coefficient; Slope*) 0,319. Jednadžba linearne regresije varijable INDKLER na varijablu INDETNO glasi, dakle:

INDKLER = 2,107 + 0,319 * INDETNO
U stupcima t i p prezentirana je statistička značajnost regresijskog koeficijenata (*b*) i odsječka (*a*). Preko t-distribucije testiraju se hipoteze: a = 0, odnosno b = 0.

Standardizirani regresijski koeficijent (beta ponder; β) u bivarijatnoj regresiji uvijek je jednak koeficijentu korelacije prediktora s kriterijem. Ovdje iznosi β = 0,553. Značajnost *b* koeficijenta ujedno je i značajnost β koeficijenta.

Coefficients

Model		Unstandardized		Standard Error	Standardized	t	p
H_0	(Intercept)	15.575		0.280		55.595	< .001
H_1	(Intercept)	2.107	a	1.115		1.890	0.060
	indetno	0.319	b	0.026	0.553	12.355	< .001

Vrijednost **odsječka (a)** interpretiramo na sljedeći način:
Ako je rezultat na indeksu etnocentrizma jednak nuli, vrijednost na indeksu klerikalizma u prosjeku će iznositi 2,107.

Vrijednost **nestandardiziranog regresijskog koeficijenta (b)** interpretiramo na sljedeći način:
Ako se rezultat na indeksu etnocentrizma poveća za 1, vrijednost na indeksu klerikalizma povećat će se u prosjeku za 0,319.

Vrijednost **standardiziranog regresijskog koeficijenta (β)** interpretiramo na sljedeći način:
Ako se rezultat na indeksu etnocentrizma poveća za 1 standardnu devijaciju te varijable, vrijednost na indeksu klerikalizma povećat će se u prosjeku za 0,553 standardnih devijacija te varijable.

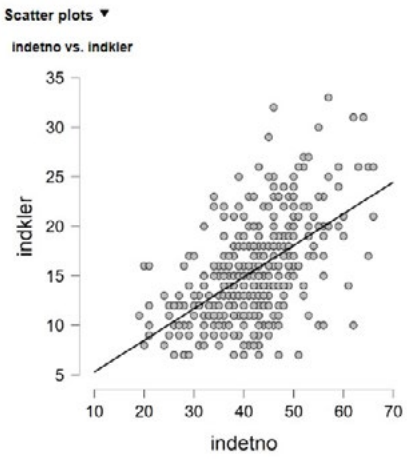
Uz ovaj zadatak možemo napisati rješenje:

Korištena je jednostavna linearna regresijska analiza za predviđanje vrijednosti na indeksu klerikalizma na temelju vrijednosti na indeksu etnocentrizma. Dobiven je statistički značajni regresijski model [$F_{(1, 346)} = 152{,}641$; $p < 0{,}001$] uz 1% rizika u kojem je $R^2 = 0{,}304$ (korigirani koeficijent determinacije). Drugim riječima, etnocentrizam objašnjava 30,4% varijance klerikalizma.

Regresijska jednadžba ($\hat{Y} = a + b*X$) glasi: *INDKLER* = 2,107 + 0,319 * *INDETNO*.

Predviđena vrijednost rezultata na indeksu klerikalizma jednaka je 2,107 + 0,319 * INDETNO, što znači da ako se rezultat na indeksu etnocentrizma poveća za 1, vrijednost na indeksu klerikalizma povećat će se u prosjeku za 0,319.
Smjer i jačina veze između rezultata na indeksima klerikalizma i etnocentrizma mogu se iščitati iz *b* koeficijenta: b = 0,553, što znači da je veza klerikalizma i etnocentrizma umjerena i pozitivna (no to smo mogli saznati i iz korelacijske analize).

Dijagram raspršenja podataka na varijablama *INDKLER i INDETNO*:



Izradu dijagrama raspršenja zatražili smo u modulu *Regression* odabirom procedure (*Classical) Corelation* te pod *Plots*, odabirom *'Scater plots'*.

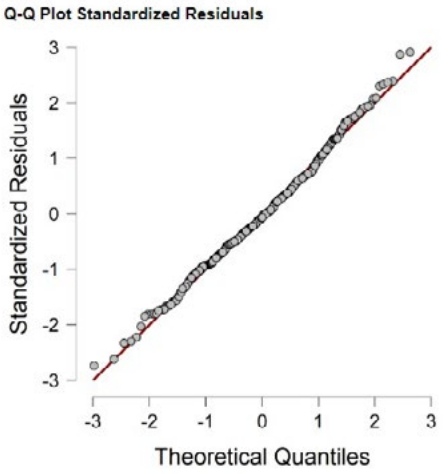
Napomena: U linearnoj je regresiji uobičajeno kriterijsku (zavisnu varijablu) prikazati na osi Y. To ćemo učiniti tako da u popis varijabli u okviru *Variables* prvo postavimo nezavisnu varijablu iz regresijskog modela (koje će biti na osi X), potom kriterijsku (zavisnu) koja će biti na osi Y.

11.3. Provjere pretpostavki

11.3.1. Linearnost veze i normalnost distribucije varijabli

Radimo na primjerima u datoteci pod nazivom:
11_jednostavna_linearna_regresija.sav

Pod *Plots* označiti *Q-Q plot standardized residuals*.



Q-Q plot pokazuje da standardizirani reziduali leže duž dijagonalne linije, što ukazuje na to da ni jedna od ove dvije pretpostavke, o normalnosti raspodjele i linearnosti odnosa, nije prekršena.

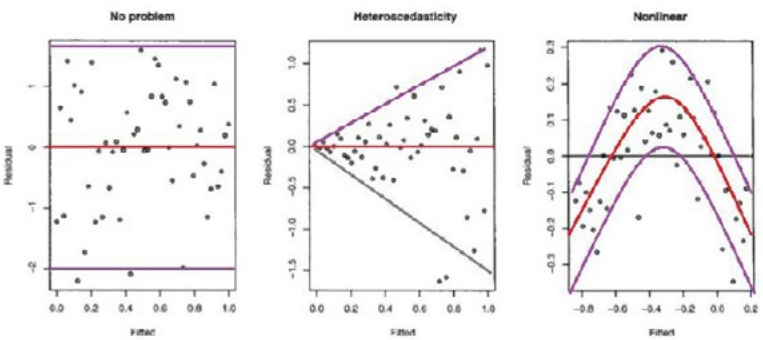
11.3.2. homoskedastičnost

Radimo na primjerima u datoteci pod nazivom:
11_jednostavna_linearna_regresija.sav

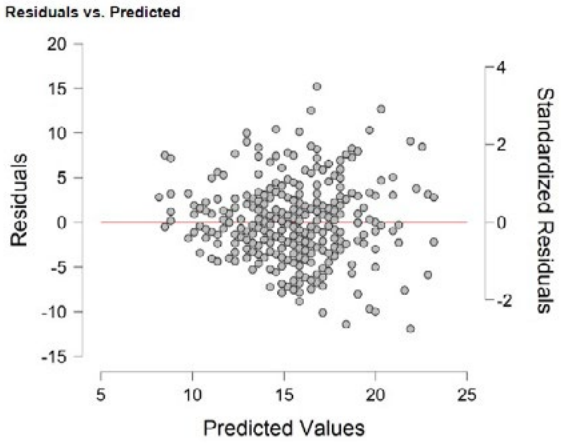
Provjera pretpostavke o **homoskedastičnosti** (hipoteza o homogenosti varijanci reziduala, tj. pretpostavka da je variranje podataka oko regresijskog pravca jednako za sve podatke prediktora).

Pod *Plots* označiti *Residuals vs. predicted*.

Mogući obrasci podataka (prema Goss-Sampson, 2018:22):

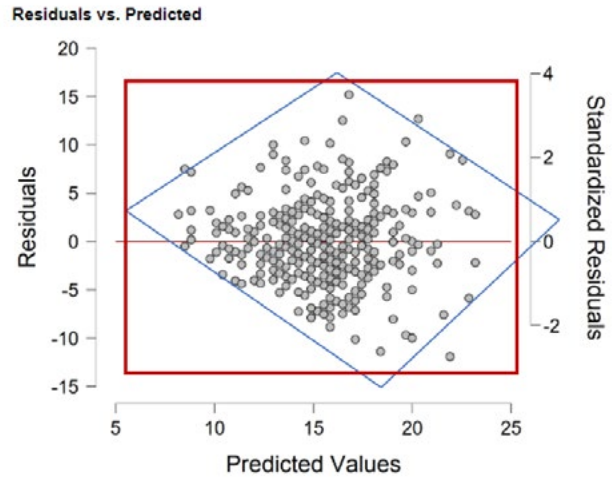


U našem primjeru oblik dijagrama raspršenja reziduala od prediciranih vrijednosti ukazuje na problem heteroskedastičnosti:



Kako smo na ovom grafičkom prikazu uočili problem heteroskedastičnosti?

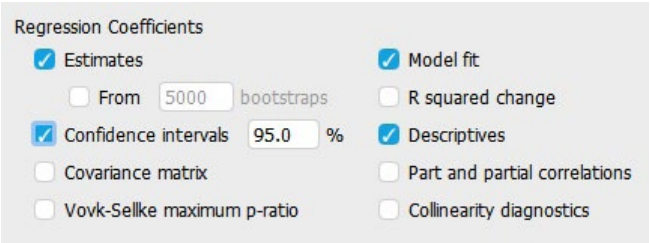
1. Povučemo li zamišljene linije uz rubne točke grafikona, očekujemo kvadratni oblik. U našem slučaju točke (u nekoj mjeri) čine kvadratni oblik.
2. Dobiveni bi kvadratni oblik trebao biti u što većoj mjeri horizontalan, odnosno paralelan s osi *x* (crvene linije). U našem je slučaju taj zamišljeni kvadratni oblik praktički zarotiran za 45°, što ukazuje na problem heteroskedastičnosti.



11.4. Interval pouzdanosti za regresijski koeficijent

Radimo na primjerima u datoteci pod nazivom:
11_jednostavna_linearna_regresija.sav

Koristan dodatak za interpretaciju rezultata provedene linearne regresijske analize jest i interval pouzdanosti za regresijski koeficijent, čije određivanje možemo zatražiti pod *Statistics*, opcijom '*Confidence intervals*'.



Time u tablici s koeficijentima dobivamo:

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	95% CI Lower	95% CI Upper
H_0	(Intercept)	15.575	0.280		55.595	< .001	15.024	16.126
H_1	(Intercept)	2.107	1.115		1.890	0.060	−0.085	4.300
	indetno	0.319	0.026	0.553	12.355	< .001	0.269	0.370

Interpretacija:
Uz 95% pouzdanosti zaključujemo da regresijski koeficijent u regresiji varijable INDKLER na varijablu INDETNO u populaciji iznosi između 0,269 i 0,370.

Literatura

— Navarro, D.J., Foxcroft, D.R., i Faulkenberry, T.J. (2019*). Learning Statistics with JASP: A Tutorialfor Psychology Students and Other Beginners*. Poglavlje 11. Correlation and linear regression. URL: https://tomfaulkenberry.github.io/JASPbook/chapters/chapter11.pdf

— Goss-Sampson, M. A. (2018). Statistical Analysis in JASP: A Guide for Students. Poglavlje Exploring Data Integrity. URL: https://static.jasp-stats.org/Statistical%20Analysis%20in%20JASP%20-%20A%20Students%20Guide%20v2.pdf

— Goss-Sampson, M. A. (2019). *Statistička analiza u JASP programu: vodič za studente*. Poglavlja: Regresija (str. 63-65) i Jednostruka regresija (str. 66-68). URL: http://static.jasp-stats.org/Manuals/Statistic%cc%8cka_analiza_u_JASP_programu_v0.10.2.pdf