Statistika 2 z računalniško analizo podatkov

Statistično sklepanje

- Multipla regresija
  - Statistično sklepanje o regresijskih koeficientih
- Multipla regresija
  - Vključevanje nominalnih in ordinalnih spremenljivk
- Multipla regresija
  - Interakcije med vplivi spremenljivk

Porazdelitev regresijskih parametrov

- Pri statistični analizi rezultatov regresijske analize je pomembno kako so ocene parametrov razpršene okoli populacijskih vrednosti.
- Metoda najmanjših kvadratov zagotavlja njihovo nepristranskost, kar pomeni, da je pričakovana vrednost ocene regresijskega parametra enaka njegovi pravi vrednosti oz.: **$E(b_i) = \beta_i$ za vsak i.**

Porazdelitev regresijskih parametrov

- Vendar pa so koeficienti v splošnem različno razpršeni.
- Razpršenost merimo s standardnim odklon ocene regresijskega parametra, ki mu tudi pravimo **standardna napaka ocene regresijskega koeficienta**.
- Večina statističnih programov pa ta podatek poda skupaj z izračunom koeficientov.

Preverjanje hipotez o regresijskih koeficientih

Pri regresiji z več spremenljivkami je mogoče postaviti več različnih domnev, ki jih lahko preverjamo s statističnimi testi. Omenimo dva najpomembnejša tipa domnev:

- Preverjanje domneve o **vplivu posamezne spremenljivke** oz. o **vrednosti posameznega regresijskega koeficienta** (ali je enak 0).
- Preverjanje domneve **o skupni značilnosti regresijskega modela**. Tu nas zanima, ali je skupen vpliv regresijskih koeficientov statistično značilen, oziroma ali so vsi koeficienti hkrati enaki 0.

Predpostavka o normalnosti člena napake

- Ena od bistvenih predpostavk regresijskega modela je **predpostavka o homoskedastičnosti**. To pomeni, da velja $D(\varepsilon_i) = \sigma^2$ za vsak i, torej da je varianca napake povsod na območju delovanja regresijskega modela približno enaka.
- Ta predpostavka sicer zadostuje za **najboljšo nepristransko oceno po metodi najmanjših kvadratov**.
- Za statistično sklepanje o regresijskih koeficientih pa je potrebna močnejša predpostavka, da se **členi napake povsod porazdeljujejo normalno**. Natančneje: $\boldsymbol{\varepsilon_i : N(0, \sigma)}$.

Standardna napaka ocen regresijskih parametrov – matrični zapis (dodatno)

S pomočjo matričnega zapisa lahko izračunamo standardno napako koeficienta takole:

$$se(b_i) = \sqrt{s_e^2 \left[(X^T X)^{-1} \right]_{i+1, i+1}}$$

, kjer "i + 1, i + 1" predstavlja koordinate matrike, se pravi "i + 1" –ti diagonalni element matrike. Uporabiti moramo i + 1 element in ne i-ti element, ker prvi element ustreza b_0 (konstantnem členu) in ne b_1

Preverjanje domneve o vplivu posamezne spremenljivke

Statistika

$$t = \frac{b_i - \beta_i}{SE(b_i)}$$

se porazdeljuje po Studentovi *t*-porazdelitvi z $n - k$ prostostnimi stopnjami, pri čemer je *n* velikost vzorca, *k* pa je število parametrov, ki jih ocenjujemo v regresijskem modelu - običajno je to število neodvisnih spremenljivk + 1 (konstanta).

Preverjanje domneve o vplivu posamezne spremenljivke

- Če preverjamo domnevo o tem, ali posamezna neodvisna spremenljivka vpliva na odvisno, preverjamo domnevo, da je parcialni regresijski koeficient β_i enak 0:

  $H_0 : \beta_i = 0$

  proti alternativni domnevi

  $H_1 : \beta_i \neq 0$

- Potem izračunamo testno statistiko *t* kot:

$$t = \frac{b_i}{SE(b_i)}$$

- Statistika se porazdeljuje po Studentovi *t*-porazdelitvi z $n - k$ prostostnimi stopnjami, pri čemer je *n* velikost vzorca, *k* pa je število parametrov, ki jih ocenjujemo v regresijskem modelu - običajno je to število neodvisnih spremenljivk + 1 (konstanta).

Preverjanje domneve o vplivu posamezne spremenljivke

- Če je vrednost izračunane *t* statistike večja od eksperimentalne pri izbrani stopnji tveganja in danih stopinjah prostosti, lahko hipotezo zavrnemo.
- Ali, kot ponavadi, **izračunamo natančno statistično značilnost (natančno stopnjo tveganja) *p***, pri kateri lahko zavrnemo ničelno domnevo oz. izračunamo verjetnost, da smo na danem vzorcu ob pravilni ničelni domnevi dobili določeno eksperimentalno vrednost *t* statistike ali večjo (po absolutni vrednosti). **Če je stopnja značilnosti majhna (<0.05), lahko zavrnemo ničelno domnevo**.
- Če zavrnemo ničelno domnevo, da je parcialni regresijski koeficient enak 0 ($\beta_i = 0$), lahko trdimo, da ima obravnavana neodvisna spremenljivka statistično značilen vpliv na odvisno spremenljivko.

Interpretacija rezultatov o vplivu posameznih spremenljivk

- Če nas zanima, katere spremenljivke statistično značilno vplivajo na odvisno spremenljivko, sklepamo o tem na podlagi statističnih značilnosti parcialnih regresijskih koeficientov. Pri izbrani stopnji značilnosti α (običajno 5%) lahko trdimo, da statistično značilno vplivajo na odvisno spremenljivko tiste neodvisne spremenljivke, katerih je stopnja značilnosti njihovih regresijskih koeficientov manjša kot α (npr. 5%).

Interpretacija rezultatov o vplivu posameznih spremenljivk

- Če nas zanima, za koliko se spremni odvisna spremenljivka, če se neodvisna spremeni za 1 enoto, gledamo vrednost regresijskega koeficienta (B v SPSS-u). Ta nam pove točno to.
- Če pa nas zanima, katera neodvisna spremenljivka najmočneje vpliva na odvisno, pa ta koeficient ni primeren, ker vsebuje tudi vpliv merskih lestvic. V tem primeru primerjamo standardizirane koeficiente (Beta v SPSS-u). Čim večji je standardizirani koeficient po absolutni vrednosti, večji je vpliv določene spremenljivke.
- Standardizirani regresijski koeficienti so pravzaprav navadni regresijski koeficienti, izračunani na standardiziranih podatkih.

Preverjanje domneve o vplivu posamezne spremenljivke

- V splošnem lahko preverjamo tudi domnevo, da je parcialni regresijski koeficient β_i enak neki drugi vrednosti (ne 0) :

  $H_0 : \beta_i = \beta_{iH}$

  proti alternativni domnevi

  $H_1 : \beta_i \neq \beta_{iH}$

- Potem izračunamo testno statistiko *t* kot:

$$t = \frac{b_i - \beta_{iH}}{SE(b_i)}$$

- Tudi ta statistika se porazdeljuje po Studentovi *t*-porazdelitvi z $n - k$ prostostnimi stopnjami (vse ostalo naprej tako kot prej).

Preverjanje domneve o skupni značilnosti regresijskega modela

- Preverjamo domnevo, da so vsi parcialni regresijski koeficienti hkrati enaki 0:

  $H_0 : \beta_1 = \beta_2 = \cdot \cdot \cdot = \beta_m = 0$

  proti alternativni domnevi

  H_1 : vsaj eden od koeficientov je različen od 0.
- Naivno bi domnevo lahko preverjali kot *m* posameznih enostavnih domnev $H_0 : \beta_i = 0$; za vsak *i*. Toda tak pristop bi bil napačen.
- Vzrok je v tem, da pri neodvisnem testiranju posameznih domnev predpostavimo, da so koeficienti, o katerih domneve testiramo, neodvisni.
- Pri testiranju domneve o skupni značilnosti modela pa so koeficienti lahko (in praktično vedno so) med seboj korelirani.

Test domneve o skupni značilnosti regresijskega modela

- Preverjanje domneve o skupni značilnosti regresijskega modela temelji na razmerju med pojasnjeno in nepojasnjeno varianco (tako kot pri analizi variance), pri čemer upoštevamo še prostostne stopnje pri posameznih izračunih. Testna statistika je:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{1-R^2/(n-k)}$$

pri čemer je:

- *ESS* pojasnjena vsota kvadratov in *RSS* nepojasnjena vsota kvadratov oz. vsota kvadratov rezudialov
- R^2 multipli determinacijski koeficient
- *n* velikost vzorca, *k* (= *m* + 1) pa število parametrov, ki jih ocenjujemo v regresijskem modelu.

$$ESS = \sum (Y_i' - \bar{Y})^2$$

$$RSS = \sum (Y_i - Y_i')^2$$

Test domneve o skupni značilnosti regresijskega modela

- Statistika F se porazdeljuje po porazdelitvi $F(k-1, n-k)$.
- Če je vrednost izračunane *F* statistike večja od eksperimentalne pri izbrani stopnji tveganja in danih stopinjah prostosti, lahko hipotezo zavrnemo.
- Ali, kot ponavadi, **izračunamo natančno statistično značilnost *p***, pri kateri lahko zavrnemo ničelno domnevo oz. izračunamo verjetnost, da smo na danem vzorcu ob pravilni ničelni domnevi dobili določeno eksperimentalno vrednost *F* statistike ali večjo. **Če je stopnja značilnosti majhna (<0.05), lahko zavrnemo ničelno domnevo**.
- Če zavrnemo ničelno domnevo, ki je da so vsi parcialni regresijski koeficienti enaki 0 ($\beta_1 = \beta_2 = \cdot\cdot\cdot = \beta_m = 0$), lahko trdimo, da vsaj ena neodvisna spremenljivka vpliva na odvisno spremenljivko.

Intervalne ocene regresijskih koeficientov

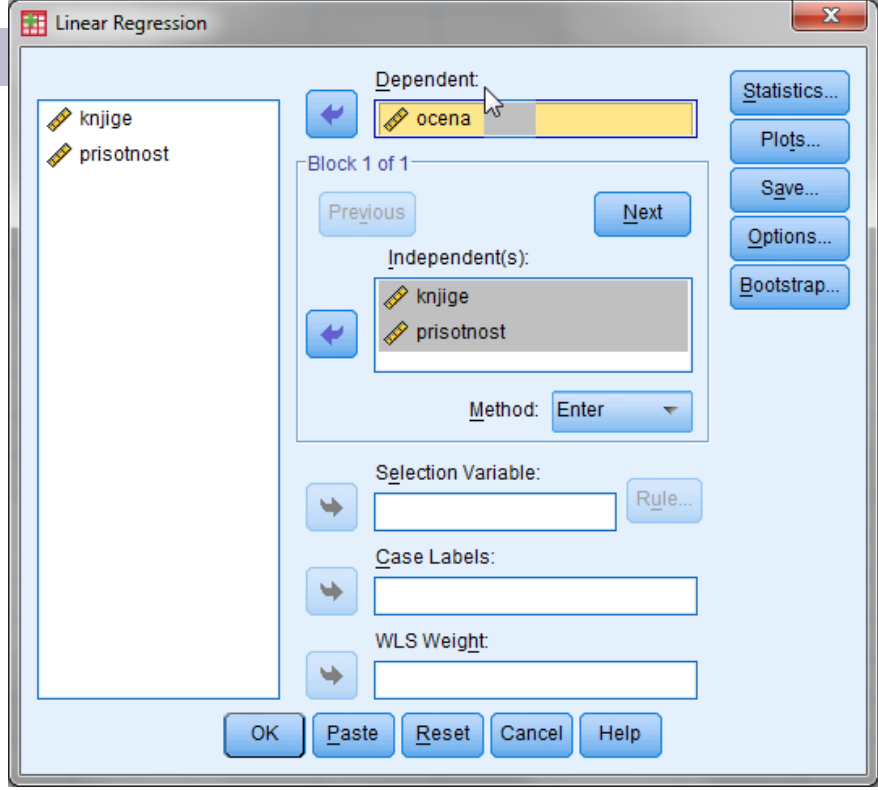
- Drugi način statističnega sklepanja o regresijskih koeficientih je intervalno ocenjevanje. Ob predpostavki, da poznamo porazdelitev vzorčne ocene b_i: $$\frac{b_i - \beta_i}{SE(b_i)} \sim t(n-k)$$ lahko na podlagi stopnje tveganja α določimo interval zaupanja za β_i:

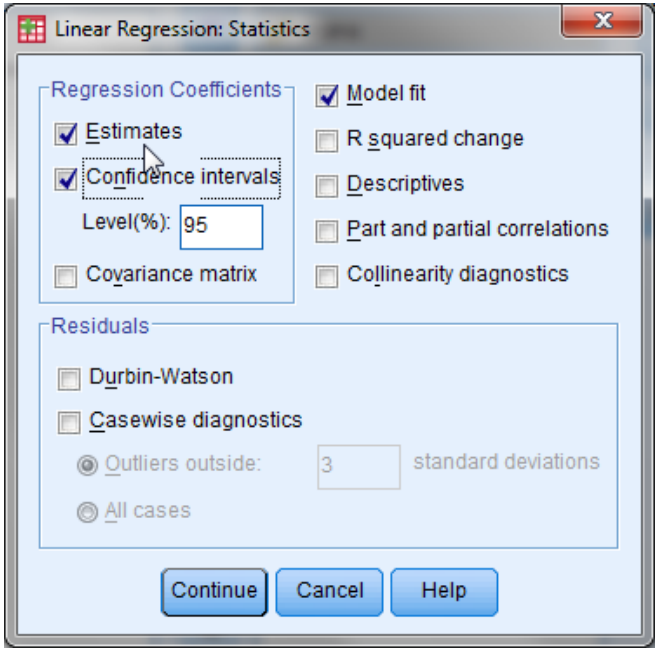
  $$b_i - t_{\alpha/2} SE(b_i) \leq \beta_i \leq b_i + t_{\alpha/2} SE(b_i).$$

- Če dobljeni interval zaupanja ne vsebuje vrednosti 0, tudi lahko sklepamo, da je koeficient β_i statistično značilno različen od 0.

Primer 1 in uporaba SPSS-a

- Uporabimo proceduro *Analyze – Regression – Linear*.
- V polje *Dependent* prenesemo odvisno spremenljivko (*prisotnost*), v polje *Independents* pa vse neodvisne spremenljivke (*knjige, ocena*). Uporabili bomo podatke iz datoteke “primer_regresija_ocene.sav” .
- V podoknu *Statistics* označimo možnosti: *Estimates*, *Confidence intervals* in *Model fit*





Primer - Sklepanje o vplivu posameznik spremenljivk

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	37,379	7,745		4,827	,000	21,687	53,071
	knjige št. prebranih statističnih knjig	4,037	1,753	,346	2,303	,027	,485	7,589
	prisotnost št. obiskanih predavanj	1,283	,587	,329	2,187	,035	,094	2,473

a. Dependent Variable: ocena št. točk

- Vrednost regresijskega koeficienta za spremenljivko *knjige* nam pove, da se odvisna spremenljivka (*ocena*) poveča za 4,037 enot (točk), če se spremenljivka *knjige* poveča za 1 enoto in vse ostale neodvisne spremenljivke ostanejo nespremenjene (pri nas ostane samo še spremenljivka *prisotnost*).
- Variabilnost ocene koeficienta merimo s standardno napako ocene koeficienta, ki za oceno tega koeficienta znaša 1,753. Na podlagi teh dveh vrednosti lahko izračunamo vrednost *t* testa, ki znaša 2,303.
- Na podlagi *t* testa in stopinj prostosti – v našem primeru 37 (= 40 enot – 3 ocenjeni parametri) izračunamo stopnjo značilnosti.

Primer - Sklepanje o vplivu posameznik spremenljivk

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	37,379	7,745		4,827	,000	21,687	53,071
	knjige št. prebranih statističnih knjig	4,037	1,753	,346	2,303	,027	,485	7,589
	prisotnost št. obiskanih predavanj	1,283	,587	,329	2,187	,035	,094	2,473

a. Dependent Variable: ocena št. točk

- Stopnja značilnosti nam pove, da lahko pri tveganju 2,7% trdimo, je parcialni koeficient za vpliv spremenljivke *knjige* na odvisno spremenljivko (*ocena)* različen od 0.
- Parcialni koeficient za vpliv spremenljivke *pristonost* na odvisno spremenljivko (*ocena*) je statistično značilno različen od 0 pri tveganju 3,5%.
- Pri 5% tveganju (recimo da je to v naprej izbrana stopnja tveganja) lahko trdimo, da obe neodvisni spremenljivki (statistično značilno) vplivata na odvisno spremenljivko.

Primer - Sklepanje o vplivu posameznik spremenljivk

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	37,379	7,745		4,827	,000	21,687	53,071
	knjige št. prebranih statističnih knjig	4,037	1,753	,346	2,303	,027	,485	7,589
	prisotnost št. obiskanih predavanj	1,283	,587	,329	2,187	,035	,094	2,473

a. Dependent Variable: ocena št. točk

- Če primerjamo moč vpliva posameznih spremenljivk, lahko trdimo, da spremenljivka *knjige* minimalno močneje vpliva na odvisno spremenljivko kot spremenljivka *prisotnost*, saj je vrednost njenega standardiziranega parcialnega regresijskega koeficienta (0,346) po absolutni vrednosti malce večja od koeficienta spremenljivke *prisotnost* (0,329).

Primer - Sklepanje o vplivu posameznik spremenljivk

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	37,379	7,745		4,827	,000	21,687	53,071
	knjige št. prebranih statističnih knjig	4,037	1,753	,346	2,303	,027	,485	7,589
	prisotnost št. obiskanih predavanj	1,283	,587	,329	2,187	,035	,094	2,473

a. Dependent Variable: ocena št. točk

- Pri 95% gotovosti (ali 5% tveganju) lahko trdimo, se pravi (populacijski) parcialni regresijski koeficient za vpliv spremenljivke *knjige* na odvisno spremenljivko (*ocena*) nahaja na intervalu med 0,485 in 7,589. Ker interval ne vsebuje vrednosti 0, lahko pri taki gotovosti (ali tveganju) trdimo tudi, da je koeficient statistično značilno različen od 0.
- Pri 5% tveganju lahko trdimo, da se parcialni regresijski koeficient za vpliv spremenljivke prisotnost na odvisno spremenljivko (*ocena*) nahaja na intervalu med 0,094 in 2,473.

Primer in uporaba SPSS-a – sklepanje o skupni značilnosti regresijskega modela

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3577,670	2	1788,835	9,059	,001[a]
	Residual	7306,230	37	197,466		
	Total	10883,900	39			

a. Predictors: (Constant), prisotnost št. obiskanih predavanj, knjige št. prebranih statističnih knjig

b. Dependent Variable: ocena št. točk

- O skupni značilnosti regresijskega modela sklepamo na podlagi analize variance (ANOVA). Tu primerjamo (z regresijskim modelom) pojasnjeno vsoto kvadratov, ki znaša v našem primeru 3577,67 z nepojasnjeno vsoto kvadratov (vsoto kvadratov rezidualov), ki znaša 7306,23. Primerjamo ju s *F* testom.

Primer in uporaba SPSS-a – sklepanje o skupni značilnosti regresijskega modela

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3577,670	2	1788,835	9,059	,001[a]
	Residual	7306,230	37	197,466		
	Total	10883,900	39			

a. Predictors: (Constant), prisotnost št. obiskanih predavanj, knjige št. prebranih statističnih knjig

b. Dependent Variable: ocena št. točk

- O skupni značilnosti regresijskega modela sklepamo na podlagi statistične značilnosti *F* testa. Tu je vrednost F testa 9,059, stopinje prostosti pa so 2 in 37. Statistična značilnost je 0,1%. Pri takem tveganju (0,1%) lahko torej trdimo, da vsaj ena izmed neodvisnih spremenljivk vpliva na odvisno spremenljivko.

- Multipla regresija
  - Statistično sklepanje o regresijskih koeficientih
- Multipla regresija
  - Vključevanje nominalnih in ordinalnih spremenljivk
- Multipla regresija
  - Interakcije med vplivi spremenljivk

Regresija z umetnimi spremenljivkami

- Za spremenljivke, ki nastopajo v regresijskem modelu kot neodvisne spremenljivke, smo predpostavili, da imajo vsaj intervalno mersko lestvico. Spremenljivke s slabšimi merskimi lastnostmi pa lahko vključimo v regresijski model s pomočjo tako imenovanih umetnih spremenljivk (dummy variables).
- Oglejmo si najpreprostejši primer. Denimo, da nas zanima, kako na plače delavcev (Y) poleg izobrazbe (X_2) vpliva spol delavca/ke. Spremenljivka spol ima le dve različni vrednosti. Uvedimo novo spremenljivko X_1, ki ima vrednost 0, če je oseba moškega in vrednost 1, če je oseba ženskega spola. Privzemimo naslednji regresijski model:

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- Denimo, da smo ocenili regresijske koeficiente β_0, β_1 in β_2. Kakšen je njihov pomen?

Regresija z umetnimi spremenljivkami

- Če ima spremenljivka X_1 vrednost 0 (moški), dobimo iz zgornjega modela regresijsko enačbo:

  $Y = \beta_0 + \beta_2 X_2 + \varepsilon$,

  če pa ima X_1 vrednost 1 (ženske), se enačba glasi

  $Y = (\beta_0 + \beta_1) + \beta_2 X_2 + \varepsilon$.

- To pomeni, da imamo dve različni regresijski enačbi, eno za moške in drugo za ženske. Pomen regresijskega koeficienta β_1 je torej v tem, da podaja razliko v povprečni plači žensk in moških. Če bi na primer želeli preveriti domnevo, da spol vpliva na višino plače delavca/ke, tudi po tem ko izločimo vpliv izobrazbe, bi preverjali hipotezo

  H0 : $\beta_1 = 0$

  proti alternativni hipotezi

  H1 : $\beta_1 \neq 0$,

  ki jo preverimo po že znani poti.

Primer 2

- Analizirali bomo vpliv spola (*gndr*) in starosti (na podlagi *yrbrn*) na vernost (*C13*).
- Spremenljivka spol ima dve možni vrednosti in sicer 1 – moški in 2 – ženski.
- Ustvarimo novo spremenljivko *gndr_rek* (procedura *Transform – Recode – Into Different Variables*), ki ima vrednost 0, če je oseba moškega spola in 1, če je oseba ženskega spola.
- To novo spremenljivko vključimo v regresijo.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3,956	,200		19,772	,000
	gndr_rek Ženski	,962	,147	,174	6,544	,000
	starost	,009	,004	,061	2,303	,021

a. Dependent Variable: C13 Kako verni ste

- Na podlagi tega lahko napišemo dve regresijske enačbi, eno za moške in eno za ženske:
  - Moški: *C13'* = 3,956 + 0,009**starost*
  - Ženske: *C13'* = 3,956 + 0,962 + 0,009**starost*
    oz. C13' = 4,918 + 0,009**starost*

Primer – interpretacija

- Imamo torej dve enačbi:
  - Moški: *C13'* = 3,956 + 0,009**starost*
  - Ženske: *C13'* = 3,956 + 0,962 + 0,009**starost*
    oz. C13' = 4,918 + 0,009**starost*
- Ženske so v povprečju za skoraj eno enoto (0,962) (na lestvici od 0 do 10) bolj verne kot moški, če izločimo (kontroliramo) vpliv starosti.

Posplošitev modela uvedbe umetnih spremenljivk

Naj bo sedaj X_1 nominalna ali ordinalna spremenljivka, ki ima k možnih vrednosti, ki jih zaradi enostavnosti označimo z 1, 2, ..., k. Uvedimo $k - 1$ umetnih spremenljivk: $P_1, P_2, ..., P_{k-2}, P_{k-1}$, ki imajo le vrednosti 0 oziroma 1 in sicer po naslednjem pravilu:

> Če ima spremenljivka X_1 vrednost i za $i = 1, ..., k - 1$, ima spremljivka P_i vrednost 1, vse ostale P_j za $j \neq i$ pa 0. Če pa ima spremenljivka X_1 vrednost k, imajo vse umetne spremenljivke vrednost 0. Vrednost k imenujemo referenčna vrednost ali referenčna kategorija.

Ni nujno, da je referenčna kategorija (tista, ki nima svoje spremenljivke) ravno zadnja kategorija. Vedeti moramo, da vse ostale kategorije pravzaprav primerjamo s referenčno.

Lahko je katerakoli, važno je le, da natanko ena kategorija nima svoje spremenljivke. Če bi vsaka kategorija imela svojo spremenljivko, bi nastopila multikolinearnost.

Nadaljevanje primera 2 (primer 3)

- Vernost bomo poskusili dodatno pojasniti še s spremenljivko kraj bivanja (*F5*). Spremenljivka ima 5 možnih vrednosti in sicer 1 - veliko mesto, 2 - predmestje ali obrobje velikega mesta, 3 - manjše mesto, 4 – vas in 5 - kmetija ali hiša na deželi.
- Ker ima originalna spremenljivka 5 možnih vrednosti, moramo ustvariti 4 umetne spremenljivke za 4 kategorije, preostala kategorija pa je referenčna kategorija.
- Za referenčno kategorijo sem si izbral “veliko mesto”. Zato sem kreiral spremenljivke *F5_pred* (predmestje ali obrobje velikega mesta), *F5_mm* (manjše mesto), *F5_vas* (vas) in *F5_kmet* (kmetija ali hiša na deželi).
- Sedaj te spremenljivke dodamo k tistim iz prejšnjega primera.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2,901	,303		9,575	,000
	gndr_rek Ženski	,945	,145	,171	6,515	,000
	starost	,011	,004	,073	2,775	,006
	F5_pred Predmestja ali obrobje velikega mesta	,384	,305	,049	1,259	,208
	F5_mm Manjše mesto	,699	,284	,105	2,461	,014
	F5_vas Vas	1,424	,263	,256	5,416	,000
	F5_kmet Kmetija ali hiša na deželi	1,459	,326	,163	4,479	,000

a. Dependent Variable: C13 Kako verni ste

- Tako kot v prejšnjem primeru bi tudi tu lahko naredili 5 regresijskih modelov glede na 5 kategoriji spremenljivke *F5*. Če bi podobno želeli storiti še za spremenljivko spol, bi morali navesti 10 regresijskih modelov (5*2).
- Tako tokrat tega ne bomo storili, ampak bomo neposredno iz rezultatov sklepali na razlike v vernosti med različnimi kraji bivanja.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,901	,303		9,575	,000
	gndr_rek Ženski	,945	,145	,171	6,515	,000
	starost	,011	,004	,073	2,775	,006
	F5_pred Predmestja ali obrobje velikega mesta	,384	,305	,049	1,259	,208
	F5_mm Manjše mesto	,699	,284	,105	2,461	,014
	F5_vas Vas	1,424	,263	,256	5,416	,000
	F5_kmet Kmetija ali hiša na deželi	1,459	,326	,163	4,479	,000

a. Dependent Variable: C13 Kako verni ste

- Iz rezultatov vidimo, da so prebivalci predmestij za 0,384 točke (na lestvici od 0 do 10) bolj verni o tistih iz VM (velikega mesta), tisti iz manjšega mesta so za 0,699 točke bolj verni od tistih iz VM, tisti iz vasi za 1,424 točke bolj verni od tistih iz VM in tisti s kmetij za 1,459 točk bolj verni od tistih iz VM.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,901	,303		9,575	,000
	gndr_rek Ženski	,945	,145	,171	6,515	,000
	starost	,011	,004	,073	2,775	,006
	F5_pred Predmestja ali obrobje velikega mesta	,384	,305	,049	1,259	,208
	F5_mm Manjše mesto	,699	,284	,105	2,461	,014
	F5_vas Vas	1,424	,263	,256	5,416	,000
	F5_kmet Kmetija ali hiša na deželi	1,459	,326	,163	4,479	,000

a. Dependent Variable: C13 Kako verni ste

- Izračunamo lahko tudi, da so tisti iz npr. vasi za 1,040 (= 1,424 – 0,384) točke bolj verni od tistih iz predmestja. Podobno lahko izračunamo tudi razlike med ostalimi pari.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,901	,303		9,575	,000
	gndr_rek Ženski	,945	,145	,171	6,515	,000
	starost	,011	,004	,073	2,775	,006
	F5_pred Predmestja ali obrobje velikega mesta	,384	,305	,049	1,259	,208
	F5_mm Manjše mesto	,699	,284	,105	2,461	,014
	F5_vas Vas	1,424	,263	,256	5,416	,000
	F5_kmet Kmetija ali hiša na deželi	1,459	,326	,163	4,479	,000

a. Dependent Variable: C13 Kako verni ste

- Ugotovimo lahko, da razlike v vernosti med prebivalci velikega mesta in prebivalci predmestja ob izločitvi vpliva spola in starosti niso statistično značilne, med tem kot razlike med prebivalci velikega mesta in ostalimi kategorijami so statistično značilne pri 5% tveganju (pravzaprav so značilne pri tveganju 1,4% ali manj).

Primer – rezultati

- O statistični značilnosti razlik v vernosti med ostalimi kategorijami (pari kategorij, kjer ena izmed kategorij ni veliko mesto) na podlagi tega izpisa ne moremo sklepati.
- To najlažje naredimo tako, da spremenimo referenčno kategorijo (vključimo umetno spremenljivko za veliko mesto in izključimo za neko drugo kategorijo), in ponovno izvedemo regresijo.
- Nato lahko enostavno ugotovimo, ali se ostale kategorije statistično značilno razlikujejo od nove referenčne kategorije.

Preverjanje domneve o vplivu skupine spremenljivk

- Preverjamo domnevo, da so (poljubno) izbrani parcialni regresijski koeficienti hkrati enaki 0:

  H_0: $\beta_a = \beta_b = \ldots = \beta_c = 0$

  proti alternativni domnevi

  H_1: vsaj eden od izbranih koeficientov je različen od 0.
- Taka hipoteza je še posebej smiselna, kadar izbrani koeficienti merijo vpliv različnih kategorij iste nominalne (ordinalne spremenljivke)
- V tem primeru pravzaprav primerjamo dva regresijskega modela:
  - Model 1: brez izbranih koeficientov (oz. kjer predpostavljamo, da so izbrani koeficienti enaki 0)
  - Model 2: z izbranimi koeficienti (oz. kjer izbrane koeficiente ocenimo)

Preverjanje domneve o vplivu skupine spremenljivk

- Za oba modela izračunamo *RSS* (nepojasnjeno vsoto kvadratov) in na podlagi tega izračunamo F test.

$$F = \frac{(RSS_1 - RSS_2)/(k_2 - k_1)}{RSS_2/(n - k_2)}$$

pri čemer je:

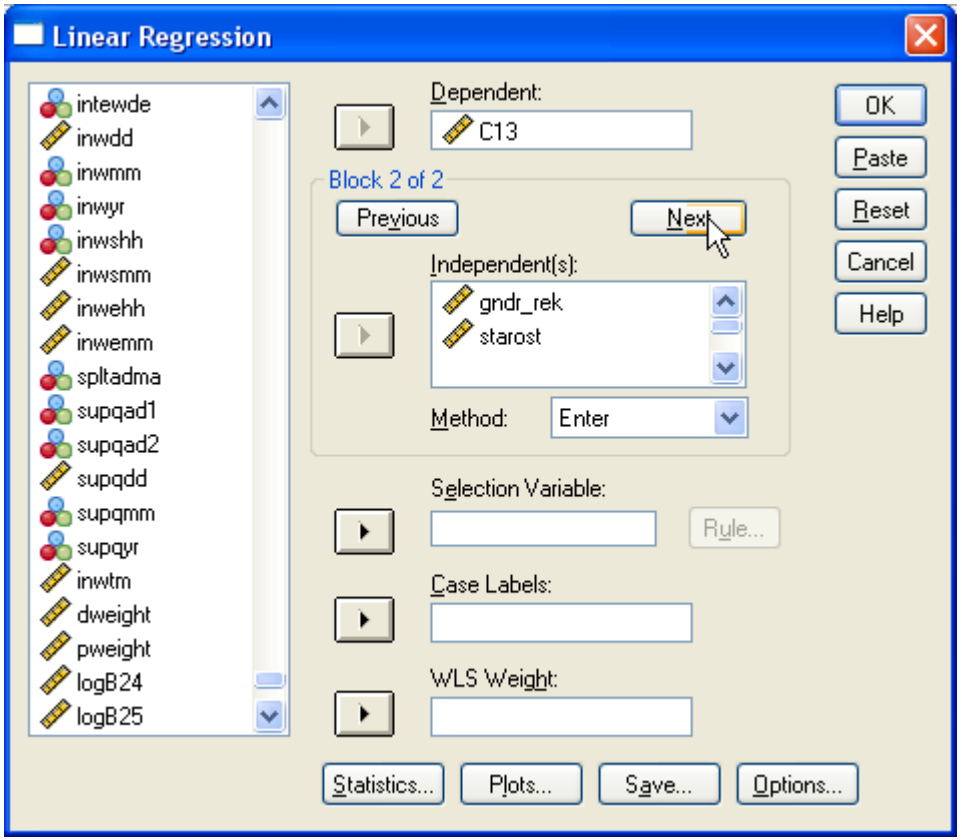
- RSS_1 nepojasnjena vsota kvadratov prvega modela (brez dodatnih koreficientov/spremenljivk)
- RSS_2 nepojasnjena vsota kvadratov drugega modela (z dodatnimi koreficienti/spremenljivkami)
- n velikost vzorca
- k_1 število parametrov, ki jih ocenjujemo v prvem modelu
- k_2 število parametrov, ki jih ocenjujemo v drugem modelu

Preverjanje domneve o vplivu skupine spremenljivk

- Statistika F se porazdeljuje po porazdelitvi $F(k_2 - k_1, n - k_2)$.
- Preverimo značilnost testa kot ponavadi (npr., preverimo, ali je *p* (natančna stopnja značilnosti) manjša od 5%).
- Če zavrnemo ničelno domnevo, da so vsi izbreni parcialni regresijski koeficienti enaki 0, lahko trdimo, da vsaj ena izmed izbranih neodvisnih spremenljivk vpliva na odvisno spremenljivko.
- Test je zelo uporaben za katerikoli sklop spremenljivk, važno je le, da so v drugem (razširjenem modelu) vsi parametri (spremenljivke), ki smo jih ocenjevali že v prvem modelu.

Nadaljevanje primera 3 (in ocenjevanje v SPSS-u)

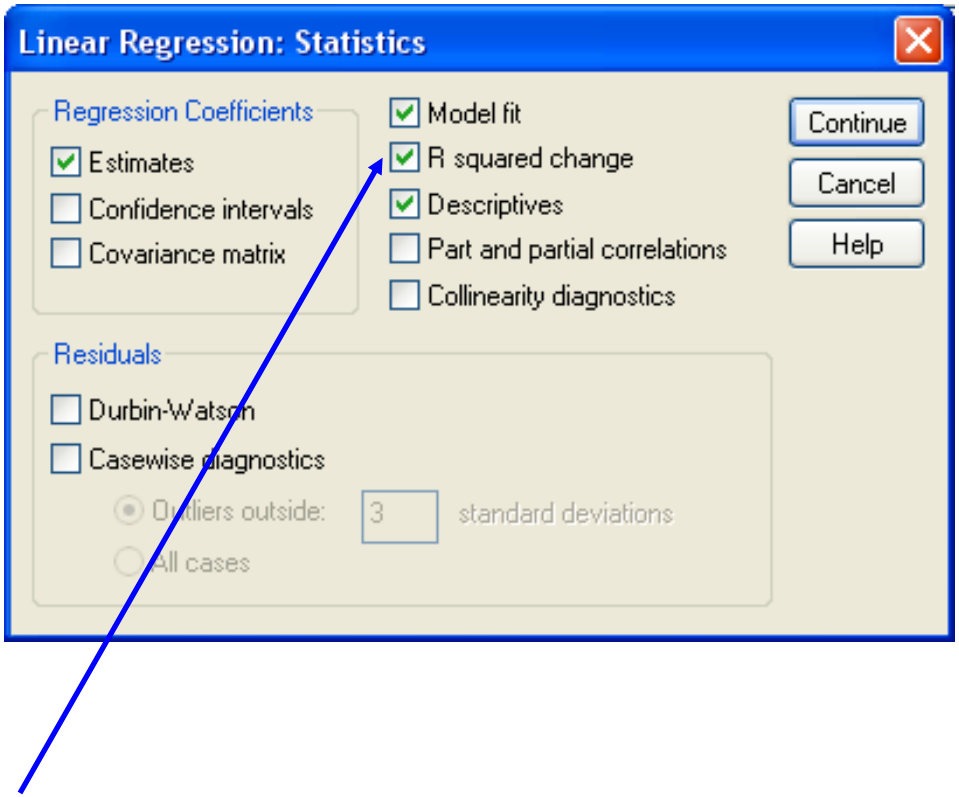
- Za primer 3 želimo ugotoviti, ali je vpliv spremenljivke *F5* “kraj bivanja” (torej katerekoli kategorije) statistično značilen.
- Uporabimo proceduro *Analyze – Regression – Linear*.
- V polje *Dependent* prenesemo odvisno spremenljivko(*C13*), v polje *Independents* pa neodvisne spremenljivke iz primera 2 (torej brez umetnih spremenljivk za spremenljivko *F5*.



- Nad okenčkom za neodvisne spremenljivke kliknemo *Next*

Nadaljevanje primera 3 (in ocenjevanje v SPSS-u)

- Okence za neodvisne spremenljivke se sprazni, nad njim pa piše *Block 2 of 2* (prej je pisalo *Block 1 of 1*)
- V okence ponovno vpišemo vse spremenljivke od prej + umetne spremenljivke za spremenljivko *F5* (torej model iz primera 3).
- Kliknemo na *Statistics* in v oknu *Statistics* odkljukamo poleg *Estimates, Model fit* in *Descriptives* še *R squared chage*



Primer – rezultati

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,189[a]	,036	,034	2,709	,036	25,277	2	1368	,000
2	,266[b]	,071	,066	2,664	,035	12,819	4	1364	,000

a. Predictors: (Constant), starost, gndr_rek Ženski

b. Predictors: (Constant), starost, gndr_rek Ženski, F5_mm Manjše mesto, F5_kmet Kmetija ali hiša na deželi, F5_pred Predmestja ali obrobje velikega mesta, F5_vas Vas

c. Dependent Variable: C13 Kako verni ste

- Poleg običajnih izpisov za oba modela dobimo sedaj še zgornji izpis, kjer nas zanima predvsem del pod *Change Statistics.*
- Drugi *F* test tu primerja drugi model s prvim in torej testira, ali je vpliv vseh dodatnih spremenljivk skupaj značilen.
- Ugotovimo lahko, da lahko pri manj kot 0,05% tveganju trdimo, da vsaj ena izmed dodatnih spremenljivk vpliva na odvisno spremenljivko.

- Multipla regresija
  - Statistično sklepanje o regresijskih koeficientih
- Multipla regresija
  - Vključevanje nominalnih in ordinalnih spremenljivk
- Multipla regresija
  - Interakcije med vplivi spremenljivk

Interakcije med vplivi spremenljivk

- Do sedaj smo vedno predpostavljali, da je vpliv posameznih spremenljivk vedno enak, ne glede na vrednosti ostalih spremenljivk.
- S pomočjo interakcij pa lahko dovolimo, da se vpliv ene neodvisne spremenljivke razlikuje glede na vrednosti druge neodvisne spremenljivke.
- Na primer, lahko predpostavimo, da starost bolj vpliva na vernost pri ženskah kot pri moških.
- Vpliv interakcije ocenimo tako, da v regresijo vključimo zmnožek spremenljivk, za kateri dve mislimo, da sta njuna vpliva povezana.

Interakcije med vplivi spremenljivk

- V primeru, da je katera od spremenljivk intervalna, jo ponavadi pred množenjem centriramo (od vrednosti odštejemo povprečje spremenljivke), da tako zmanjšamo problem multikolinearnosti. Umetnih (dummy) spremenljivk ne centriramo.
- Npr., če imamo spremenljivki *X* in *Z* in želimo oceniti interakcijo med njima, vključimo v regresijo novo spremenljivko $xz = (X - \bar{X})(Z - \bar{Z})$
- Interakcijo še posebej pogosto preučujemo, kadar je vsaj ena spremenljivka umetna (dummy) spremenljivka.
- Interpretiramo jo tako, da pogledamo, kakšen je vpliv ene neodvisne spremenljivke pri različnih vrednostih druge neodvisne spremenljivke.

Nadaljevanje primera 2 (primer 4)

- Preverili bomo, ali starost enako vpliva na vernost žensk in moških.
- S stavkom *Compute* kreiramo novo spremenljivko, ki jo izračunamo kot

  starZen = *(starost* – 45,40)**gndr_rek*

  (45,40 je povprečna starost)
- To spremenljivko vključimo v regresijo poleg tistih iz primera 2

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,392	,274		16,016	,000
	gndr_rek Ženski	,967	,147	,175	6,588	,000
	starost	-,001	,006	-,008	-,189	,850
	starZen	,018	,008	,092	2,322	,020

a. Dependent Variable: C13 Kako verni ste

- Vidimo lahko, da je vpliv nove spremenljivke (*starZen*) statistično značilen, vendar pa je vpliv spremenljivke *starost* postal neznačilen.
- Kaj to pomeni? Starost vpliva samo na vernost pri ženskah, pri moških pa ne.

Primer – rezultati

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,392	,274		16,016	,000
	gndr_rek Ženski	,967	,147	,175	6,588	,000
	starost	-,001	,006	-,008	-,189	,850
	starZen	,018	,008	,092	2,322	,020

a. Dependent Variable: C13 Kako verni ste

- To je lepo vidno, če zopet lahko napišemo dve regresijski enačbi:
  - Moški: *C13'* = 4,392 - 0,001**starost*
  - Ženske: *C13'* = 4,392 + 0,967 - 0,001**starost* + 0,018**(starost* – 45,40)

    oz. C13' = 4,542 + 0,017**starost*