Statistika 2 z računalniško analizo podatkov

Neizpolnjevanje predpostavk regresijskega modela

Predpostavke regresijskega modela
(ponovitev)

V regresijskem modelu navadno privzamemo naslednje pogoje:

1. **Člen napake je normalno porazdeljen**
2. **Člen napake ima pogojno matematično upanje 0 pri vseh vrednostih neodvisnih spremenljivk**:

$E(\varepsilon_i \,|X_{1i}, \ldots, X_{mi}) = 0$, *za vsak* i.

3. **Med členi napake ni serijske korelacije:**

$C(\varepsilon_i, \varepsilon_j) = 0$, za $i \neq j$.

4. **Homoskedastičnost:**

$D(\varepsilon_i) = \sigma^2$, za vsak i.

5. **Ničelna kovarianca med členom napake in neodvisnimi spremenljivkami**:

$C(\varepsilon_i, X_{ki}) = 0$, za vsak i in $k = 1, \ldots, m$.

Predpostavke regresijskega modela
(ponovitev)

6. Med **neodvisnimi spremenljivkami ni popolne kolinearnosti** ali **multiple kolinearnosti**.
   - To pomeni, da nobena od neodvisnih spremenljivk ni linearna kombinacija ene ali več preostalih.
   - Multikolinearnost bi obstajala na primer, če bi veljalo $X_1 = 3X_2 - 2X_3$
   - če pa velja na primer $X_2 = X_1^2$, pa o multikolinearnosti ne moremo govoriti, kajti zveza med spremenljivkama ni linearna.
7. Vse **neodvisne spremenljivke morajo imeti dovolj različne vrednosti**, se pravi morajo biti dovolj razpršene. →preverimo porazdelitev
8. **Model je pravilno specificiran** (nastavljen):
   - Vključene so vse relevantne spremenljivke
   - V modelu niso vključene nerelevantne spremenljivke
   - Uporabljena je pravilna funkcijska zveza med spremenljivkami

1-5 Predpostavke o rezidualih (členu napake) (smo že obravnavali pri analizi rezidualov)

- Metode za analizo smo že obravnavali na predavaju 10 (analiza rezidualov)
- Preverjanje:
  - Pregled grafikona, kjer damo na *y* so (standardizirane) reziduale, na *x* os pa (standardizirane) napovedane vrednosti (to so pravzaprav linearne kombinacije neodvisnih spremenljivk in nam kot kake nadomestijo vrednosti vseh neodvisnih spremenljivk (predpostavke 2 in 4)
  - Histogram rezidualov (predpostavka 1).
  - Pregled grafikona, kjer so na *y* osi reziduali, na *x* osi pa zaporedne številke eno (1, 2, ..., *n*) – smiselno samo, če je moč enote urediti v nek smiseln vrstni red (recimo po časovnem zaporedju, po anketarjih, ...) – predpostavka 3 (tega ne nismo/ne bomo delali)
- SPSS: Glejte predavanje 10

1-5 Predpostavke o rezidualih (členu napake) (smo že obravnavali pri analizi rezidualov)

- V primeru, ko so predpostavke izpolnjene v razsevnem grafikonu:
  - V grafikonu ni mogoče opaziti nobenega pravilnega vzorca porazdeljevanja residualov – izpolnjena je predpostavka 2.
  - Variabilnost v rezidualov (os *y*) je pri vseh vrednostih na *x* osi približno enaka – izpolnjena je predpostavka 4 (o homoskedastičnosti).
  - V grafikonu ni velikih odstopanj od povprečja za majhno število enot (outliers). Ekstremene vrednosti sicer formalno kršijo le predpostavko o normalnosti, a lahko povzročijo velike težave pri ocenjevanju.
- S histogramom preverimo predpostavko 1, da se reziduali porazdeljujejo približno normalno.

1-5 Predpostavke o rezidualih (členu napake) (smo že obravnavali pri analizi rezidualov)

- Posledice:
  - Ocene standardnih napak niso zanesljive, zato tudi statistični testi niso zanesljivi
  - V primeru heteroskedastičnosti (kršitve predpostavke 4, o homoskedastičnosti), ocene parametrov niso optimalne (imajo večjo varianco oz. standardno napako, kot bi jo lahko imele)

1-5 Predpostavke o rezidualih (členu napake) (smo že obravnavali pri analizi rezidualov)

- Možne rešitve:
  - Če v grafikonu je mogoče opaziti pravilen vzorca porazdeljevanja residualov – kršene je predpostavka 2 → Kaže na možnost specifikacijske napake (glej prosojnice ...)
  - Heteroskedastičnost ali odstopanje od normalne porazdelitve (kršena je predpostavka 1 ali 4 – ali obe) → Če ne gre za spcifikacijsko napako (najprej razmislimo o tej možnosti), lahko včasih problem rešimo z ustrezno transformacijo odvisne spremenljivke (ne bomo natančneje obravnavali)

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

- To pomeni, da nobena od neodvisnih spremenljivk ni linearna kombinacija ene ali več preostalih.
- Popolna multikolinearnost bi obstajala na primer, če bi veljalo npr. $X_1 = 3X_2 - 2X_3$.
- Če pa velja na primer $X_2 = X_1^2$, pa o multikolinearnosti ne moremo govoriti, kajti zveza med spremenljivkama ni linearna.
- Problem multikolinearnost že, ko se približamo (popolni) multikolinearnosti.

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

- O multikolinearnosti govorimo, kadar velja med neodvisnimi spremenljivkami popolna linearna zveza, torej ko obstajajo taka realna števila $\lambda_1, \ldots, \lambda_m$, od katerih vsaj eno ni enako 0, da velja (za vse enote):
  $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_m X_m = K$,
  kjer je K neka konstanta
- Seveda v praksi redko naletimo na primer, ko bi zgornja enačba popolnoma veljala. Vendar nastopijo težave že, ko obstaja visoka stopnja multikolinearnost, torej ko velja
  $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_m X_m = v$,
  pri čemer je v nek stohastični člen z majhno varianco.
- Govorimo torej lahko o nižji ali višji stopnji multikolinearnosti.

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

Tipični vzroki za multikolinearnost:

- v model smo kot neodvisni vključili dve spremenljivki, med katerima obstaja močna linearna povezanost, npr. $X_1 = 2X_2$ (plus majhna napaka)
- visoko korelirani so polinomski členi spremenljivke, katere vrednosti so na ozkem intervalu;
- predeterminiran model: v model smo vključili več spremenljivk, kot je velikost vzorca (ali se samo število spremenljivk približuje velikosti vzorca);
- drugo.

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

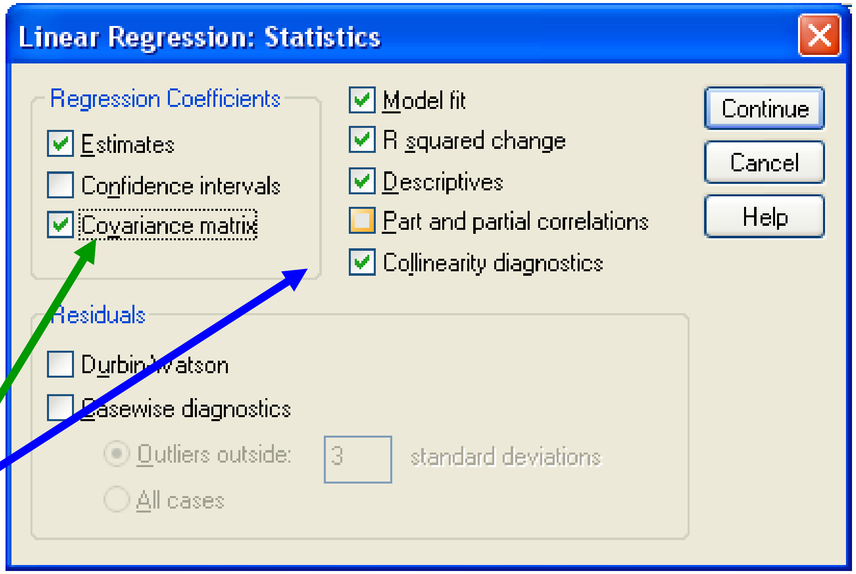
Posledice visoke multikolinearnosti:

- Ocene regresijskih koeficientov imajo velike standardne napake, kar otežuje statistično sklepanje.
- Posledično os ocene regresijskih koeficientov statistično neznačilno različne od 0 in njihovi intervali zelo široki.
- Intervali zaupanja za regresijske ocene so zelo široki, torej so ocene nenatančne.
- Ocene parametrov so močno občutljive na majhne spremembe podatkov (če recimo dodamo ali izpustimo eno "običajno" enoto) in zato nezanesljive.
- Vse omenjene posledice so bolj izražene pri majhnih vzorcih!

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

Ugotavljanje multikolinernosti s programom SPSS:

- V oknu *Linear Regression* izberemo v podoknu *Statistics* (dobimo ga s klikom na gumb *Statistics*) dodatno še *Collinearity diagnostics* in v okvirčku *Regression Coefficients* še *Covariance matrix*



6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

Nekateri indikatorji multikolinearnosti:

- Zelo značilen (majhna stopnja značilnosti) pri *F*-testu za skupno značilnost modela (tabela ANOVA), a izrazito nenatančilni (velike stopnje značilnosti) *t*-testi za regresijske koeficiente (tabela Coefficients). → Nakazujejo samo, da obstaja problem multikolinarnosti, ne pa tudi kje (med katerimi neodvisnimi spremenljivkami)
- Nizke tolerance (tabela Coefficients). Toleranca pomeni delež variabilnosti neodvisne spremenljivke, ki je neodvisen od ostalih neodvisnih spremenljivk. Majhne vrednosti tolerance so blizu 0,1, kar pomeni, da je delež variabilnosti dane spremenljivke, ki je pojasnjen s preostalimi spremenljivkami 0,9. → Spremenljivke, pri katerih je toleranca majhna, so v močno linearno povezane z vsaj nekaterimi ostalimi neodvisnimi spremenljivkami. Razmisliti je potrebno o izključitvi ali zamenjavi kakšne od teh spremenljivk

6 Med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multiple kolinearnosti

Nekateri indikatorji multikolinearnosti:

- Visoke vrednosti (nad 30) ‘indeksa pogojnosti’(‘condition index’, tabela Collinearity Diagnostics), ki pomeni kvadratni koren razmerje med največjo in najmanjšo lastno vrednostjo korelacijske matrike vektorja neodvisnih spremenljivk, ter/ali lastne vrednosti blizu 0 (‘eigenvalues’, tabela Collinearity Diagnostics). → Nakazujejo samo, da obstaja problem multikolinarnosti, ne pa tudi kje (med katerimi neodvisnimi spremenljivkami)
- Visoke korelacije (nad 0,9) med (lahko samo nekaterimi) ocenami regresijskih koeficientov (tabela Coefficient Correlations). → Ocene regresijskih koeficietnov, med katerimi je močna korelacija, so zelo močno povezane (če se spremeni ena, se spremeni tudi druga). Razmisliti je potrebno o izključitvi ali zamenjavi kakšne od teh spremenljivk.

6 Ugotavljanje multikolinearnosti - primer

- Uporabili bomo umetne podatke "multikolinearnost.sav". Notri imamo spremenljivke:
  - *x1* – prva neodvisna spremenljivka
  - *x2* – druga neodvisna spremenljivka
  - *x12* – spremenljivka, izračunana kot
    x12 = *x1* + *x2* + *err* (majhna napaka)
    Spremenljivka *x12* je skoraj popolna linearna kombinacija spremenljivk *x1* in *x2*.
  - *y* – neodvisna spremenljivka, kjer je pravilen model
    $y = 100 + 2*x1 - 0.5*x2 + e$ (napaka)

- Ocenimo dva modela:
  - $y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \varepsilon$ (ni multikolinearnosti)
  - $y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x12 + \varepsilon$ (multikolinearnost je)

6 Ugotavljanje multikolinearnosti - primer

- *Zahtevamo linearno regresijo kot običajno (Analyze – Regresion – Linear).*
- Za 1. model v polje *Dependent* prenesemo odvisno spremenljivko *y*, v polje *Independents* pa spremenljivki *x1* in *x2.*
- Za 2. model v polje *Independents* poleg spremneljivk *x1* in *x2* dodamo še spremenljivko *x12.*
- V oknu *Linear Regression* izberemo v podoknu *Statistics* (dobimo ga s klikom na gumb *Statistics*) dodatno še *Collinearity diagnostics* in v okvirčku *Regression Coefficients* še *Covariance matrix*

6 Ugotavljanje multikolinearnosti – primer 1. model (ni multikolinearnosti)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48673,225	2	24336,613	237,064	,000[a]
	Residual	9957,863	97	102,658		
	Total	58631,088	99			

a. Predictors: (Constant), x2, x1

b. Dependent Variable: y

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	99,785	11,125		8,969	,000		
	x1	1,913	,100	,800	19,106	,000	,999	1,001
	x2	-,480	,049	-,409	-9,761	,000	,999	1,001

a. Dependent Variable: y

- Rezultati *F*-testa in *t*-testov se ujemajo (oboji so močno značilni)
- Toleranci sta visoki (blizu 1)

6 Ugotavljanje multikolinearnosti – primer 1. model (ni multikolinearnosti)

Coefficient Correlations[a]

Model			x2	x1
1	Correlations	x2	1,000	,035
		x1	,035	1,000
	Covariances	x2	,002	,000
		x1	,000	,010

a. Dependent Variable: y

- Korelacije med regresijskima koeficeintoma (za spremenljivki x1 in x2) so šibke (blizu 0).

6 Ugotavljanje multikolinearnosti – primer 1. model (ni multikolinearnosti)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	x1	x2
1	1	2,964	1,000	,00	,00	,00
	2	,031	9,801	,03	,89	,08
	3	,005	24,718	,97	,10	,91

a. Dependent Variable: y

- Indeks pogojnosti ni zelo velik (pod 30), čeprav je že kar znaten.
- Najmanjša lastna vrednost ni skoraj 0.

6 Ugotavljanje multikolinearnosti – primer 2. model (multikolinearnost je)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48673,233	3	16224,411	156,414	,000[a]
	Residual	9957,855	96	103,728		
	Total	58631,088	99			

a. Predictors: (Constant), x12, x1, x2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	99,766	11,394		8,756	,000		
	x1	1,903	1,137	,796	1,673	,098	,008	127,852
	x2	-,490	1,129	-,417	-,434	,665	,002	523,123
	x12	,010	1,136	,010	,009	,993	,002	631,749

a. Dependent Variable: y

- Rezultati *F*-testa in *t*-testov se ne ujemajo (*F*-test je značilen, *t*-testi pa niso)
- Tolerance so nizke (blizu 0)

6 Ugotavljanje multikolinearnosti – primer 2. model (multikolinearnost je)

Coefficient Correlations[a]

Model			x12	x1	x2
1	Correlations	x12	1,000	-,996	-,999
		x1	-,996	1,000	,995
		x2	-,999	,995	1,000
	Covariances	x12	1,292	-1,287	-1,282
		x1	-1,287	1,293	1,278
		x2	-1,282	1,278	1,274

a. Dependent Variable: y

- Korelacije med regresijskimi koeficienti (za spremenljivke x1, x2 in x12) so zelo močne (velike, po absolutni vrednosti blizu 1)

6 Ugotavljanje multikolinearnosti – primer
2. model (multikolinearnost je)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	x1	x2	x12
1	1	3,962	1,000	,00	,00	,00	,00
	2	,032	11,088	,01	,01	,00	,00
	3	,006	26,066	,96	,00	,00	,00
	4	7,72E-006	716,217	,03	,99	1,00	1,00

a. Dependent Variable: y

- Indeks pogojnosti je ogromen (močno nad 30)
- Najmanjša lastna vrednost je skoraj 0.

8 Model je pravilno specificiran (nastavljen)

- Možne kršitve te predpostavke so:
  - V modelu manjka relevantna spremenljivka
  - V model je vključena nerelevantna spremenljivke
  - Uporabljena je napačna funkcijska zveza med spremenljivkami
- Specifikacijska napaka je navadno posledica slabe teoretične osnove modela.
- Že pred izvedbo raziskave je namreč potrebno teoretično opredeliti vse relevantne vplive na pojav, ki ga proučujemo.
- V nasprotnem primeru dobimo kot rezultat regresijske analize morda navidezno dobre ocene, ki pa so zaradi napačnega modela zavajajoče.

8 Posledice specifikacijskih napak – manjkajoča relevantna spremenljivka

To pomeni, da smo npr. namesto pravilnega modela:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$

izbrali (ocenili) model

$Y = \alpha_0 + \alpha_1 X_1 + u$

- Če sta spremenljivki X_1 in X_2 korelirani (kar navadno sta), to pomeni, da se bo koeficient α_1 razlikoval od prave vrednosti β_1 in s tem bo ocena vpliva spremenljivke X_1 na odvisno spremenljivko zavajajoča.
- Ne glede na to, ali sta spremenljivki korelirani ali ne, bomo dobili napačno oceno koeficienta α_0.
- Manjša pojasnjevalna moč modela (manjši R^2) in posledično manj natančne napovedi vrednosti odvisne spremenljivke.
- Napako je zelo težko odkriti. Zato je potrebno pred analizo pozorno preučiti teorijo o pojavu, ki ga preučujemo.

8 Posledice specifikacijskih napak – vključena nerelevanta spremenljivka

- Posledice načeloma niso tako hude. Tako napačno specifikacijo pa tudi hitro opazimo iz neznačilnih *t* vrednosti.
- Toda napačno bi bilo vključevati v model spremenljivke brez kakršne koli teoretične osnove. S tem namreč zmanjšujemo učinkovitost ocen parametrov in povečujemo možnost multikolinearnosti.
- POZOR: Prav tako ni dobro izključevati spremenljivk zgolj na podlagi *t*-testov. Spremenljivko izključimo le, če je to teoretično smiselno, drugače pa le, če je nujno (premajhen vzorec, prevelika multikolinearnost,...)

8 Posledice specifikacijskih napak – napačna funkcijska zveza

Indikatorji:

- Nepravilna porazdelitev rezidualov (glej prejšnje točke 1-5 in predavanje 10).

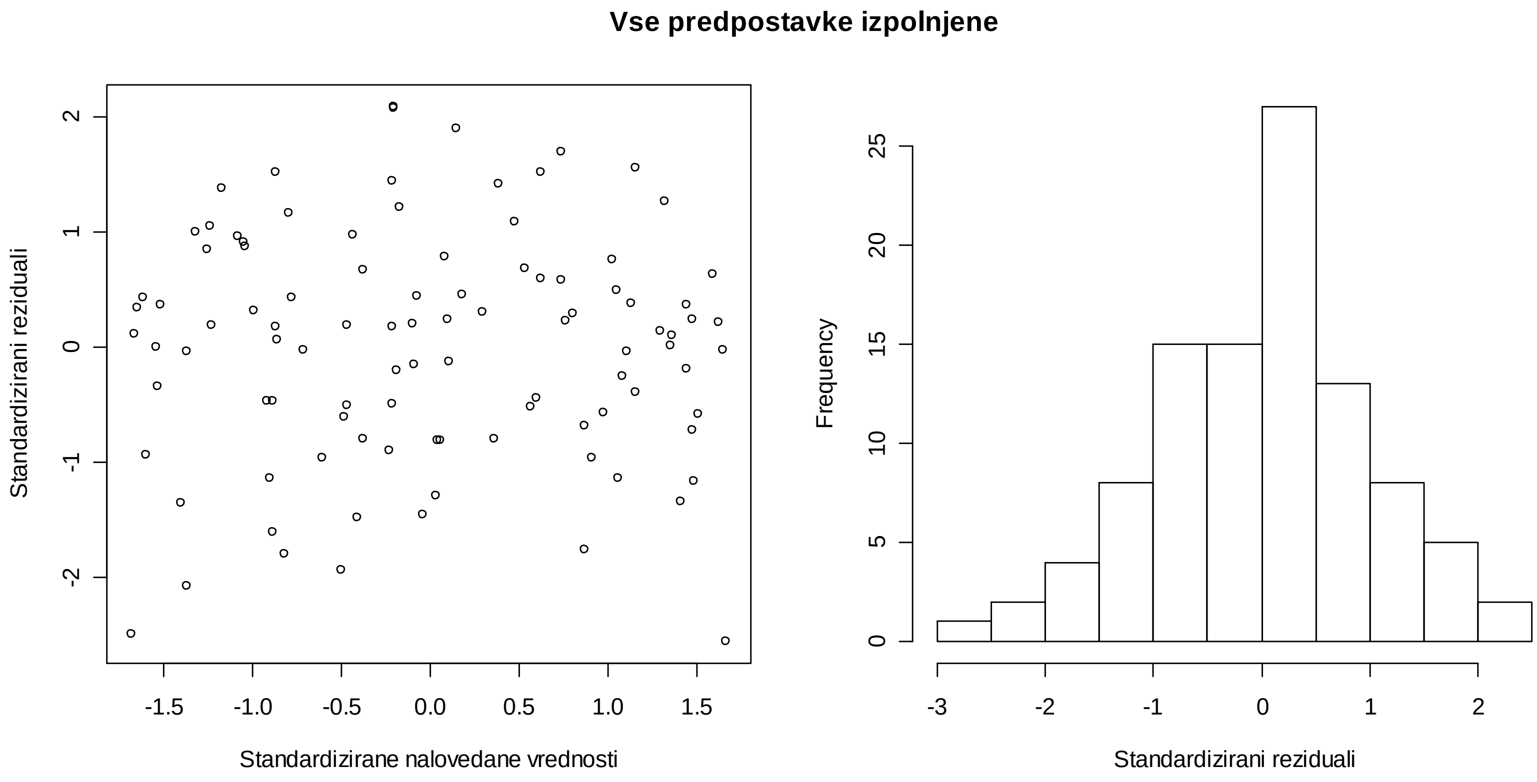
Popravek:

- Poskušamo ugotoviti pravo obliko funkcije (predavanje 9)

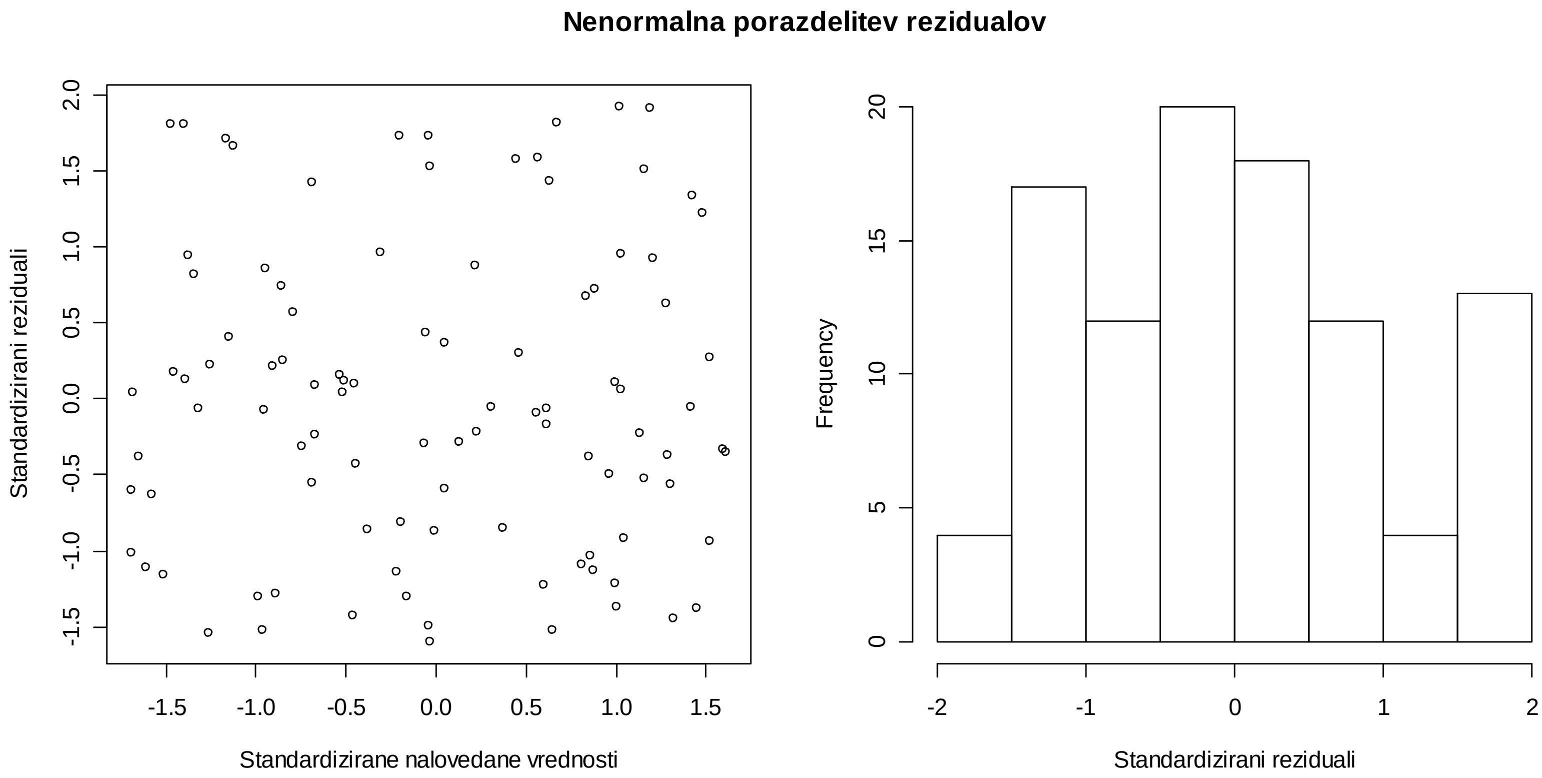
Posledice:

- Napačni sklepi (o vrsti povezanosti med spremenljivkama, moči povezanosti, …).

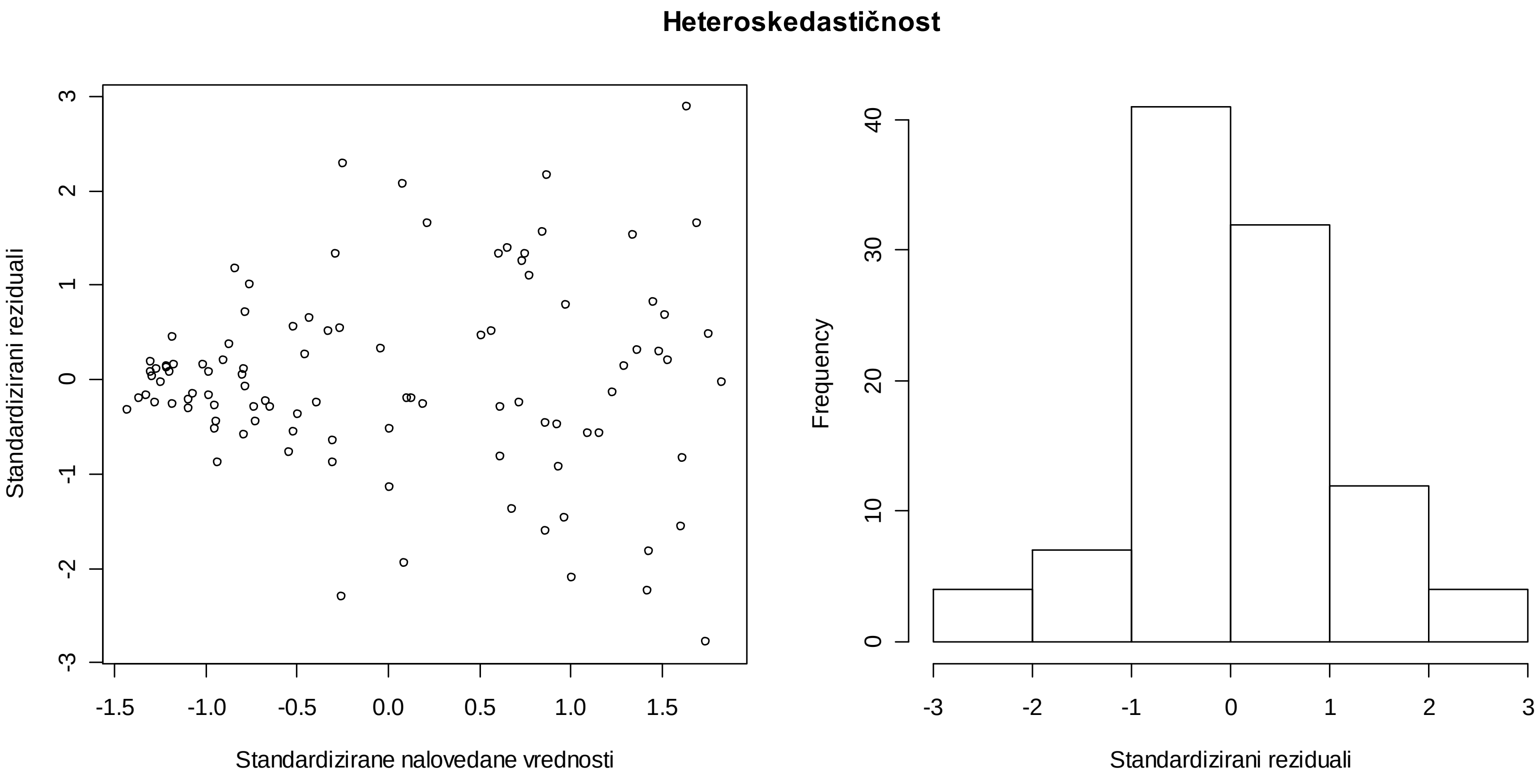
1-5 Predpostavke o rezidualih /specifikacijske napake



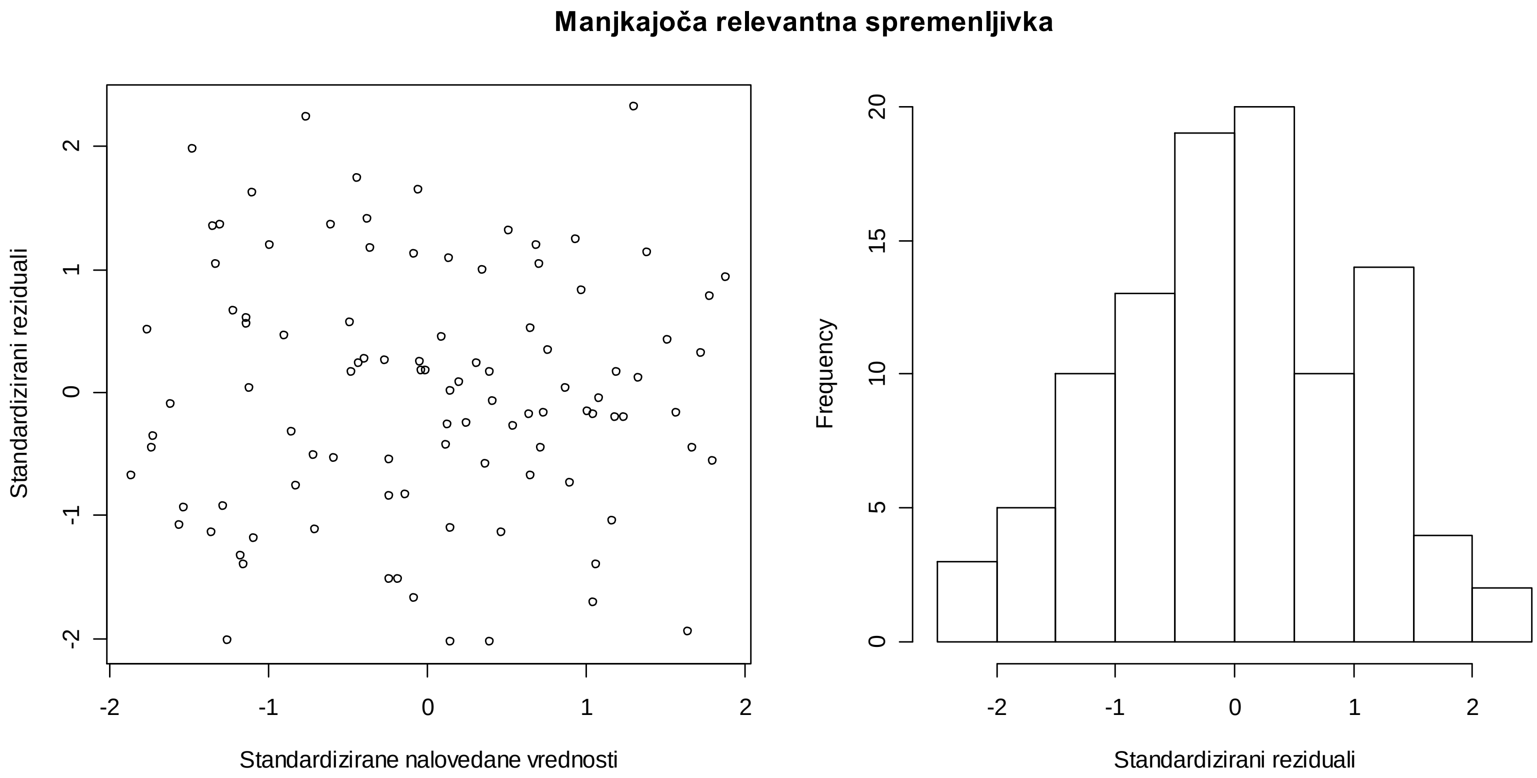
1-5 Predpostavke o rezidualih /specifikacijske napake



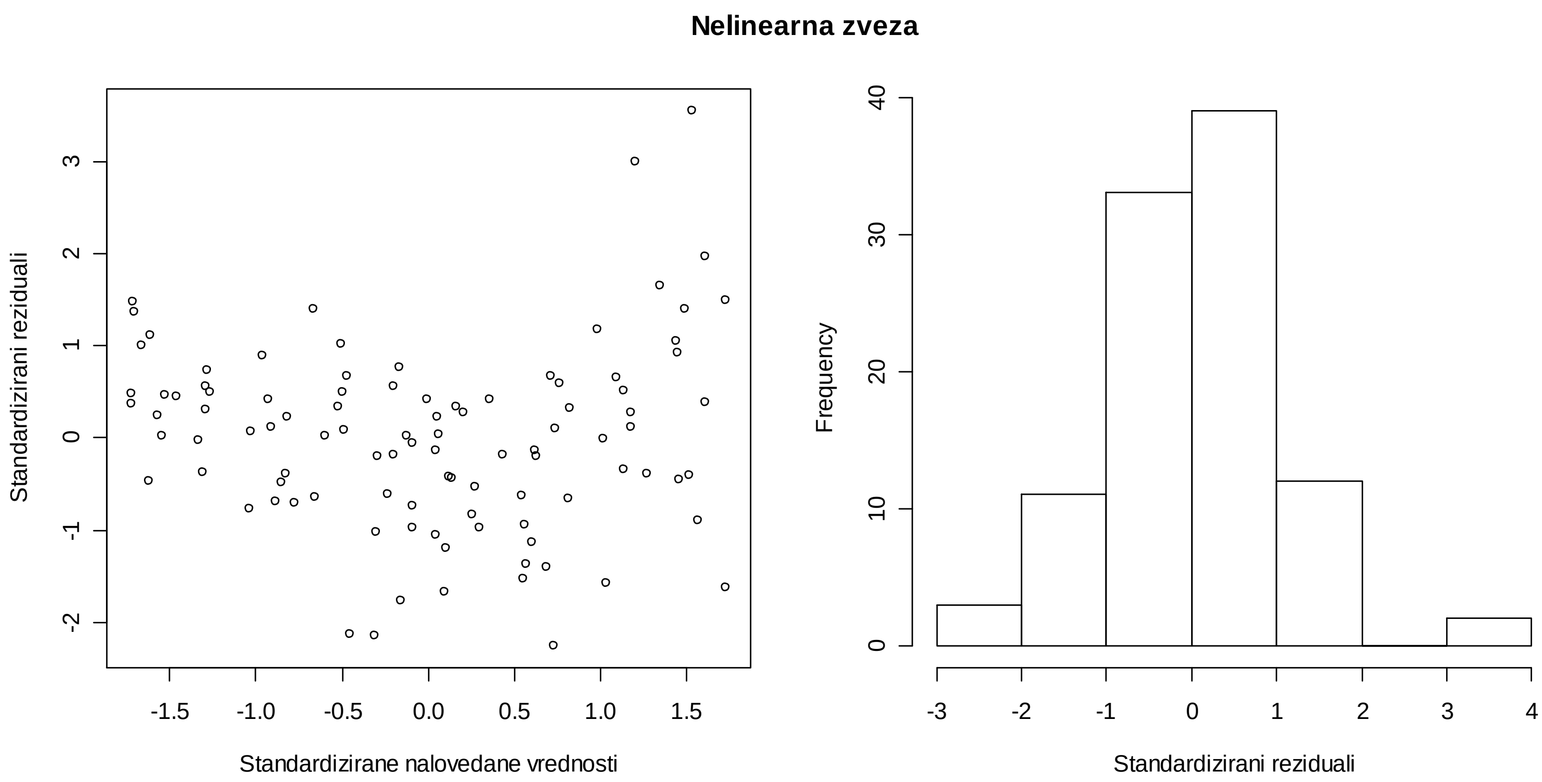
1-5 Predpostavke o rezidualih /specifikacijske napake



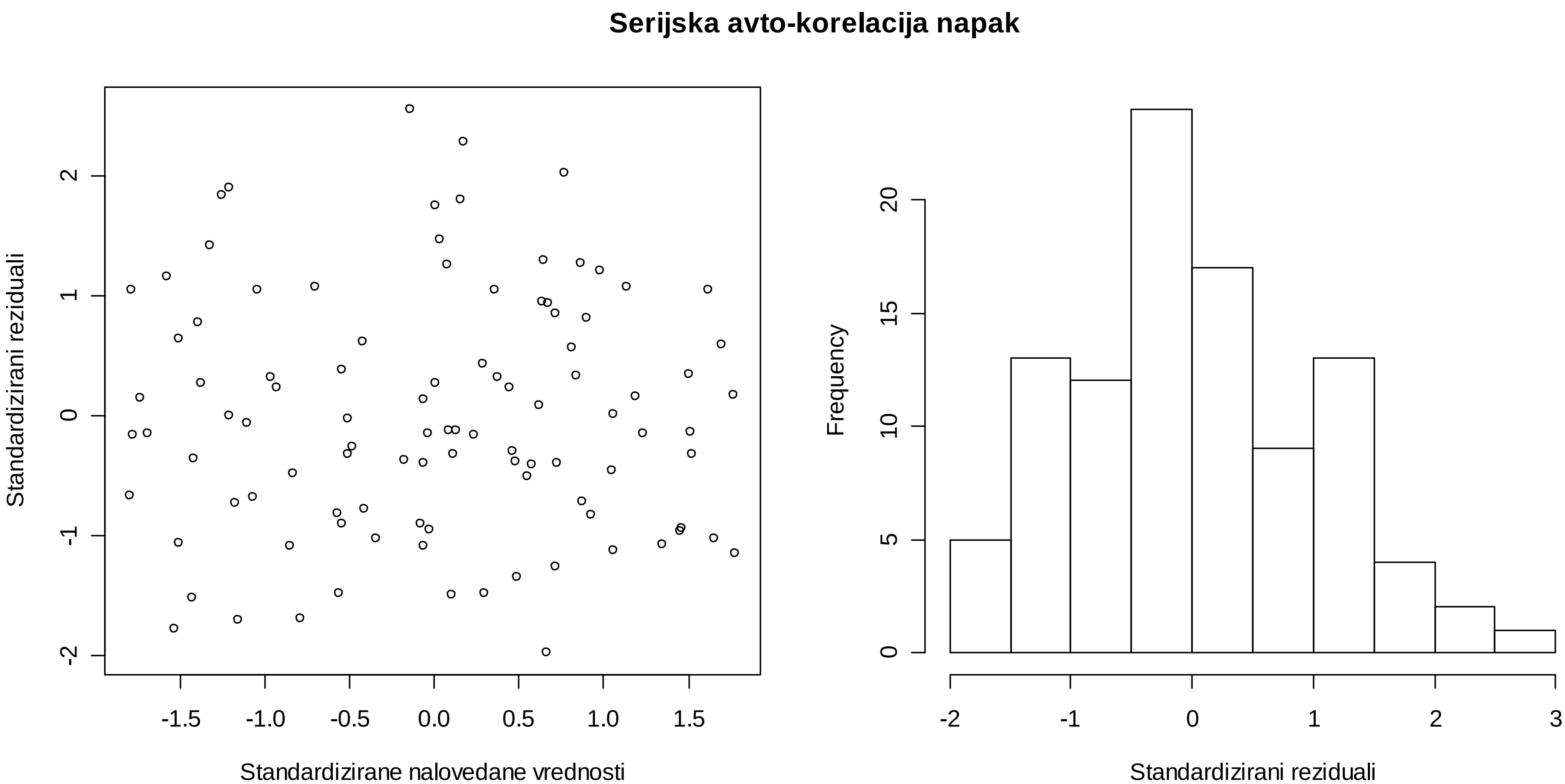
1-5 Predpostavke o rezidualih /specifikacijske napake



1-5 Predpostavke o rezidualih /specifikacijske napake



1-5 Predpostavke o rezidualih /specifikacijske napake



1-5 Predpostavke o rezidualih /specifikacijske napake