Statistika 2 z računalniško analizo podatkov

Multipla regresija in polinomski regresijski model

Multipli regresijski model

- Pogosto so vrednosti odvisne spremenljivke linearno odvisne od več kot ene neodvisne spremenljivke.
- Tak tip odvisnosti lahko opišemo z naslednjim modelom:

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon,$

  - kjer je Y odvisna spremenljivka
  - $X_1, \ldots, X_m$ so neodvisne spremenljivke.
  - ε je člen napake
  - $\beta_0, \beta_1, \ldots, \beta_m$ so neznane konstante: β_0 je konstantni člen, členi β_i , $i = 1, \ldots, m$ pa so parcialni regresijski koeficienti.
- Navadno je naloga regresijske analize ocena regresijskih koeficientov in statistično sklepanje o njih.

Dodatno: Matrični zapis

- Kot bomo videli kasneje, izrazi za izračun ocen regresijskih parametrov postanejo pri multipli regresiji precej kompleksni.
- Računanje se izjemno poenostavi, če uporabimo matrični zapis in pravila računanja z matrikami.
- Tako lahko model s prejšnje prosojnice zapišemo kot:

  $y = X\beta + \varepsilon$,

  - kjer je y stolpčni vektor odvisne spremenljivke
  - X je matrika, kjer je prvi stolpec vektor 1 (enic), vsak naslednji stolpec pa predstavlja eno neodvisno spremenljivko.
  - ε je stolpčni vektor napak.
  - β je stolpčni vektor, ki vključuje konstantni člen in regresijske koeficiente.

Dodatno: Matrični zapis

Spodaj lahko vidimo primer definicije matrik in vektorjev za regresijsko analizo.

$y = X\beta + \varepsilon$

$$\mathrm{y}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{bmatrix}, \mathrm{X}=\begin{bmatrix} 1 & X_{1,1} & X_{2,1} \\ 1 & X_{1,2} & X_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & X_{1,i} & X_{2,i} \\ \vdots & \vdots & \vdots \\ 1 & X_{1,n} & X_{2,n} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Predpostavke regresijskega modela

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**:

   $E(\varepsilon_i \mid X_{1i}, \ldots, X_{mi}) = 0$, *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

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, sicer ocenjevanje parametrov ni mogoče.
8. Model je pravilno specificiran (nastavljen):
   - Vključene so vse relevantne spremenljivke
   - V modelu niso vključene nerelevantne spremenljivke
   - Uporabljena je pravilna funkcijska zveza med spremenljivkami

Pomen multiple regresijske enačbe

- Desna stran regresijske enačbe brez člena napake pogojno matematično upanje odvisne spremenljivke glede na dane vrednosti neodvisnih spremenljivk:

  $E(Y|X_{1i}, \ldots, Xmi) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_m X_{mi}$

- Pomen regresijskih koeficientov $\beta_0, \beta_1, \ldots, \beta_m$ je naslednji:
  - Koeficient β_0 je konstantni člen in predstavlja pričakovano vrednost odvisne spremenljivke, če bi bile vrednosti vseh neodvisnih spremenljivk enake 0.
  - Koeficient β_i , za $i = 1, \ldots, m$ predstavlja pričakovano spremembo vrednosti odvisne spremenljivke, če se X_i spremeni za eno enoto, vse ostale spremenljivke pa ostanejo nespremenjene.
  - Absolutna vrednost posameznega regresijskega koeficienta ne kaže nujno na moč vpliva posamezne spremenljivke v modelu. Upoštevati je potrebno še izbor enot in variabilnost pripadajoče spremenljivke.

Ocenjevanje regresijskih parametrov po metodi najmanjših kvadratov

- Vzemimo najpreprostejši multipli regresijski model:

  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

  Model želimo oceniti na osnovi vrednosti spremenljivk na vzorcu velikosti n.

- Formulacija problema

  Znane so vzorčne vrednosti Y_i, X_{1i}, X_{2i}, iščemo pa ocene za parametre β_0, β_1 *in* β_2 in vrednosti napak ε_i. Ocene parametrov označimo z b_0, b_1 *in* b_2, ocene napak pa z e_i, kjer je i indeks enote.

Ideja rešitve po metodi najmanjših kvadratov

- Po metodi najmanjših kvadratov ocenimo parametre tako, da bo vsota kvadratov členov napake $\sum_{i=1}^{n} e_i^2$ čim manjša. Iščemo torej minimum naslednje funkcije parametrov b_0, b_1 *in* b_2:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

- Zgornja funkcija je funkcija treh spremenljivk b_0, b_1 *in* b_2, njen minimum pa najdemo s parcialnim odvajanjem po vseh spremenljivkah.

Reševanje enačbe

- Po tem postopku dobimo naslednji sistem linearnih enačb:

$$\bar{Y} = b_0 + b_1 \bar{X}_1 + b_2 \bar{X}_2$$

$$\sum_{i=1}^{n} Y_i X_{1i} = b_0 \sum_{i=1}^{n} X_{1i} + b_1 \sum_{i=1}^{n} X_{1i}^2 + b_2 \sum_{i=1}^{n} X_{1i} X_{i2}$$

$$\sum_{i=1}^{n} Y_i X_{2i} = b_0 \sum_{i=1}^{n} X_{2i} + b_1 \sum_{i=1}^{n} X_{1i} X_{2i} + b_2 \sum_{i=1}^{n} X_{i2}^2$$

pri čemer $\bar{Y}$, $\bar{X}_1$ in $\bar{X}_2$ pomenijo vzorčna povprečja

Iz prve enačbe dobimo:

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

Reševanje enačbe

- Zapis rešitve zgornjih enačb je precej neroden, nekoliko pa se poenostavi, če vpeljemo naslednji dogovor: $y_i = Y_i - \bar{Y}, x_{1i} = X_{1i} - \bar{X}_{1,} x_{2i} = X_{2i} - \bar{X}_2$
  - mala črka spremenljivke pomeni odklon spremenljivke od vzorčnega povprečja:
- Tedaj se parcialna regresijska koeficienta izražata kot:

$$b_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$b_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

Dodatno: Matrični zapis

Kot smo omenili, so ti izrazi z matričnim zapisom dosti enostavnejši:

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

- kjer je b ocena β,
- X^T pa X transponirano (vrstice in stolpci so zamenjani.
- ter je kot množenje uporabljeno matrično množenje.

Napoved za y (y') pa izračunamo kot:

$y' = Xb$

Primer ocenjevanja regresijskih parametrov

- Analizirali bomo vpliv spremenljivk *prisotnost* „št. obiskanih predavanj“ (X_1) in *knjige* „št. prebranih statističnih knjig“ (X_2) na spremenljivko *ocena* „število točk“ (Y). Zanima nas, ali prisotnost na predavanjih in prebrane knjige vplivajo na oceno.
- Podatki in delni izračuni so na voljo v datoteki „primer-multipla-ocene.xls“.
- Predpostavljamo torej model

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

Primer ocenjevanja regresijskih parametrov

- Na podlagi formule na prosojnici 9 lahko izračunamo:

$$b_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i}\, x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} =$$

$$= \frac{1343{,}8 \cdot 80{,}0 - 459{,}0 \cdot 106{,}0}{713{,}6 \cdot 80{,}0 - 106{,}0^2} = 1{,}823$$

$$b_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i}\, x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} =$$

$$= \frac{459{,}0 \cdot 713{,}6 - 1343{,}8 \cdot 106{,}0}{713{,}6 \cdot 80{,}0 - 106{,}0^2} = 4{,}037$$

Primer ocenjevanja regresijskih parametrov

- Nato pa podlagi formule na prosojnici 8 še:
  $b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 = 63{,}6 - 1{,}283 \cdot 14{,}1 - 4{,}037 \cdot 2{,}0 =$
  $= 37{,}379$
- Med spremenljivkami torej velja naslednja zveza:
  $Y' = 37{,}379 + 1{,}283X_1 + 4{,}037\,X_2$
- Interpretacija:
  - Če se prisotnost (X_1) poveča za en obisk (eno enoto), potem pričakujemo, da se bo število točk (*ocena*) zvečalo za 1,3 točke, če število prebranih knjig ostane nespremenjeno.
  - Če se število prebranih knjig (X_2) poveča za eno knjigo, potem pričakujemo, da se bo število točk (*ocena*) povečalo 4,0 točke, če pristonost ostane nespremenjena.

Primer ocenjevanja regresijskih parametrov

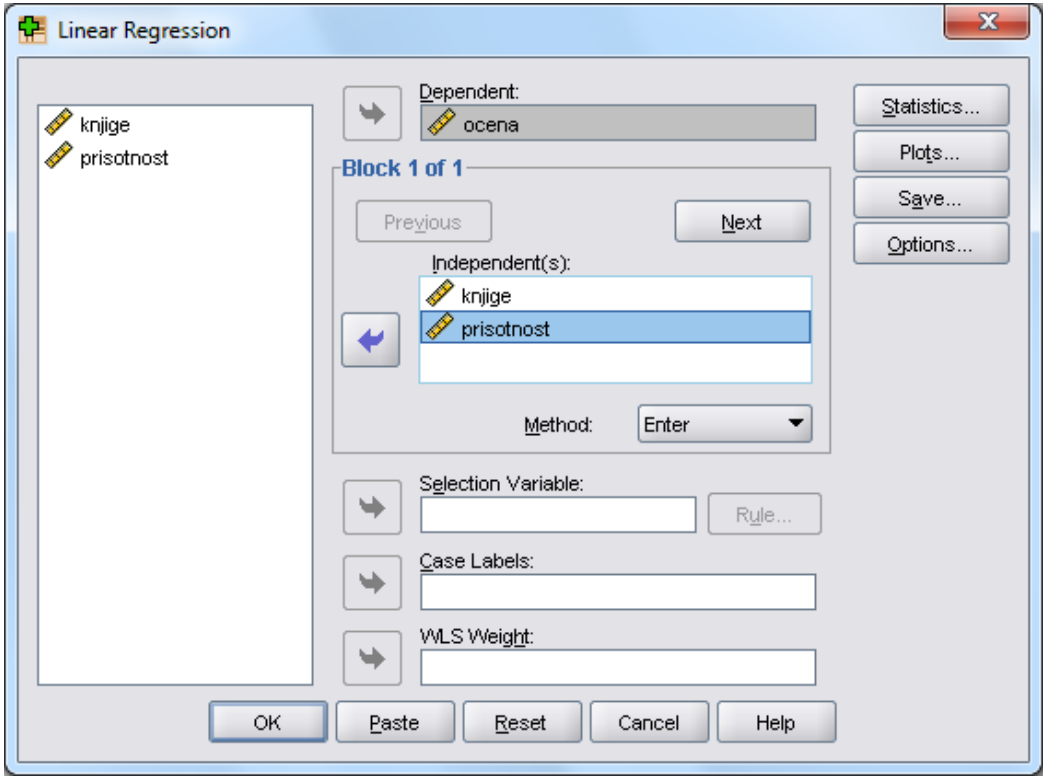
- Kolikšna je pričakovano število točk (Y) za študenta, ki je bil na predavanjih 17-krat (X_1) in je prebral eno statistično knjigo (X_2) ? Vstavimo v regresijsko enačbo in dobimo:

  $Y'(X_1 = 17, X_2 = 1) = 37{,}379 + 1{,}283 \cdot 17 + 4{,}037 \cdot 1 = 63{,}235$

- Za takega študenta pričakujemo, da bo ocenjen z 63-timi točkami.

Primer ocenjevanja regresijskih parametrov s SPSS-om

- V SPSS-u za multiplo regresijo uporabimo isto proceduro kot za bivariatno regresijo. Uporabimo torej 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”.

Primer ocenjevanja regresijskih parametrov s SPSS-om - izpis

- Med drugim dobimo sledeč izpis. Ker je interpretacija enaka kot prej, jo ne bomo ponavljali.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	37,379	7,745		4,827	,000
	knjige št. prebranih statističnih knjig	4,037	1,753	,346	2,303	,027
	prisotnost št. obiskanih predavanj	1,283	,587	,329	2,187	,035

a. Dependent Variable: ocena št. točk

Prileganje regresijskega modela

- V bivariatnem regresijskem modelu smo merili prileganje regresijskega modela z **determinacijskim koeficientom**, ki je meril **delež pojasnjene variance odvisne spremenljivke**.
- Pri Statistiki I smo ga označili z *R* (*v SPSS-u se pri regresiji oznaka R uporablja za koren determinacijskega koeficienta*), včasih tudi r^2, v SPSS-u je označen z $R^{2.}$
- Podoben koeficient meri tudi prileganje multiplega regresijskega modela. Imenujemo ga multipli determinacijski koeficient in ga označimo z R^2.

Izračun (multiplega) determinacijskega koeficienta za trivariatni regresijski model

- Postopek je skoraj enak kot za bivariatni model. Zapišimo regresijsko enačbo v obliki:

  $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i = Y_i' + e_i$

- Pri čemer je Y_i' ocenjena vrednost spremenljivke Y, e_i pa ocena napake.
- Delež pojasnjene variance bomo torej merili (tako tko pri bivariatni analizi) kot kvocient med varianco spremenljivk Y' in Y.

Izračun (multiplega) determinacijskega koeficienta za trivariatni regresijski model

- Z upoštevanjem predpostavk regresijskega modela in ocenjevanja po metodi najmanjših kvadratov dobimo naslednjo enačbo:

  $y_i = b_1x_{1i} + b_2x_{2i} + e_i = y_i' + e_i$

  pri čemer upoštevamo dogovor, da male črke označujejo odklone od povprečij danih spremenljivk (pri napaki (e_i) to ne povzroči nobene spremembe, ker je povprečje napak enako 0).
- Kvadrirajmo obe strani enačbe in seštejmo po vseh enotah:

  $\Sigma y_i^2 = \Sigma(y_i')^2 + \Sigma e_i^2 + 2\Sigma y_i'e_i = \Sigma(y_i')^2 + \Sigma e_i^2$

  pri čemer upoštevamo predpostavko, da sta *Y'* in *e* nekorelirani!
- Z besedami: Vsota kvadratov odklonov odvisne spremenljivke je enaka **pojasnjena vsota kvadratov odklonov** + **nepojasnjena vsota kvadratov odklonov**.

Izračun (multiplega) determinacijskega koeficienta za trivariatni regersijski model

- Nepojasnjeno vsoto kvadratov ocenimo (z nekaj računanja) kot:

  $\Sigma e_i^2 = \Sigma y_i^2 - b_1 \Sigma y_i x_{1i} - b_2 \Sigma y_i x_{2i}$

- Če ta izraz vstavimo v prejšnjo enačbo, dobimo s preureditvijo:

  $\Sigma (y_i')^2 = b_1 \Sigma y_i x_{1i} + b_2 \Sigma y_i x_{2i}$

- Multipli determinacijski koeficient je tedaj enak:

$$R^2 = \frac{\sum (y_i')^2}{\sum y_i^2} = \frac{b_1 \sum y_i x_{1\mathrm{i}} + b_2 \sum y_i x_{2\mathrm{i}}}{\sum y_i^2}$$

Dodatno: Matrični zapis

Izračun multipega determinacijskega koeficienta:

$$R^2 = \frac{(\mathbf{Xb})^T \mathbf{Xb}}{\mathbf{y}^T \mathbf{y}}$$

(Multipli) determinacijski koeficient

- Tako kot navadni determinacijski koeficient, leži tudi multipli determinacijski koeficient na intervalu [0, 1]
- vrednost 1 pomeni, da regresijski model pojasnjuje 100 % (vso) variabilnosti spremenljivke *Y*
- vrednost 0 pa pomeni, da model ne pojasnjuje (čisto nič) variabilnosti spremenljivke *Y*.
- Navadno pa je njegova vrednost nekje med obema skrajnima vrednostnima.

Nadaljevanje primera

- Na podlagi izračunov iz datoteke „primer-multipla-ocene.xls“ lahko izračunamo R^2 na podlagi 1. ali 2. dela enačbe:
- 1. del: $R^2 = \frac{\sum(y_i')^2}{\sum(y_i)^2} = \frac{3577,7}{10883,9} = 0,329$
- 2. del:

$$R^2 = \frac{b_1\sum y_i x_{1i} + b_2\sum y_i x_{2i}}{\sum(y_i)^2} = \frac{1,283 \cdot 1343,8 + 4,037 \cdot 459,0}{\sum(y_i)^2} =$$

$$= \frac{3577,7}{10883,9} = 0,329$$

Nadaljevanje primera

- V linearni regresijski zvezi iz prejšnjega primera smo dobili vrednost multiplega regresijskega koeficienta $R^2 = 0{,}329$. Regresijska zveza torej pojasnjuje 32,9% variabilnosti odvisne spremenljivke *Y*.
- To lahko vidimo tudi iz ustreznega SPSS izpisa:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,573[a]	,329	,292	14,052

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

Povečevanje vrednosti determinacijskega koeficienta

- V multipli regresijski model lahko **vedno dodamo** še katero **neodvisno spremenljivko**.
- S tem **skoraj vedno povečamo determinacijski koeficient**, se pravi prileganje modela.
- **Toda ni nujno, da s tem izboljšamo model** (oz. njegovo pojasnjevalno vlogo)– čim višja vrednost determininacijskega koeficienta ni edini cilj. Zato:
  - Vključujemo v model le spremenljivke, za katere obstaja teoretična osnova, da vplivajo na odvisno spremenljivko.
  - Model poskušamo ohraniti čim enostavnejši, saj obstajajo poleg vplivov neodvisnih spremenljivk na odvisno tudi najrazličnejši medsebojni vplivi neodvisnih spremenljivk. Pri prekompleksnem modelu pravega načina delovanja posameznih spremenljivk ni lahko ugotoviti.

Standardna napaka regresijske ocene

- Natančnost regresijske ocene lahko podamo s standardno napako regresijske ocene.
- Ena od predpostavk regresijskega modela je bila, da ima člen napake pri vseh vrednostih odvisne spremenljivke enako varianco, σ^2 oz. standardni odklon σ. $\sigma = \sqrt{(\sigma^2)}$
- Ta standardni odklon napak imenujemo **standardno napako regresijske ocene**, ki jo označimo z σ_ε, oceno standardne napake ocene pa z s_e.

Standardna napaka regresijske ocene

- S standardno napako regresijske ocene torej merimo odstopanja vrednosti odvisne spremenljivke od regresijskih napovedi.
- Izkaže se, da je nepristranska vzorčna ocena za standardno napako regresijske ocene naslednja količina, izračunana na vzorcu:

$$s_e^2 = \frac{\sum e_i^2}{n-k} = \frac{\sum (Y_i - Y_i')^2}{n-k}, s_e = \sqrt{(s_e^2)}$$

kjer je *n* velikost vzorca, k pa je število parametrov, ki jih ocenjujemo v regresijskem modelu - običajno je to število neodvisnih spremenljivk + 1 (konstanta).

Prilagojeni multipli determinacijski koeficient

- Multipli determinacijski koeficient R smo definirali kot razmerje med pojasnjeno vsoto kvadratov odklonov in skupno vsoto kvadratov odklonov od povprečja.
- Toda v resnici je količina, ki nas zanima razmerje med pojasnjeno in skupno varianco
- Nepristranska vzorčna ocena za skupno varianco je, kot vemo iz klasične statistike, količina:

$$s_Y^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1}$$

Dodatno: Matrični zapis

Izračun standardna napake regresijske ocene in variance odvisne spremenljivke:

$$s_e^2 = \frac{\mathbf{e}^{\mathsf{T}}\mathbf{e}}{n-k} = \frac{(\mathbf{y}-\mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\mathbf{b})}{n-k}, s_e = \sqrt{(s_e^2)}$$

$$s_y^2 = \frac{(\mathbf{y}-\bar{Y})^{\mathsf{T}}(\mathbf{y}-\bar{Y})}{n-1}, s_y = \sqrt{(s_y^2)}$$

Prilagojeni multipli determinacijski koeficient

- V skladu z nepristranskima ocenama za nepojasnjeno varianco in skupno varianco, je delež pojasnjene variance enak:

$$R^2_{pop} = 1 - \frac{s_e^2}{s_y^2} = 1 - \frac{\frac{\sum (Y_i - Y_i')}{n-k}}{\frac{\sum (Y_i - \bar{Y})}{n-1}}$$

- Koeficientu R^2_{pop} pravimo **prilagojeni multipli determinacijski koeficient (adjusted R square)**.
- Meri, koliko variabilnosti pričakujemo, da bi z modelom pojasnili na populaciji, iz katere je bil izbran vzorec

Zveza med R^2 in R^2_{pop}

$$R^2_{pop} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

- Od tod sledi tudi neenakost: $R^2_{pop} \leqslant R^2$
- Koeficient R^2_{pop} podaja nekoliko boljšo oceno deleža pojasnjene variance na populaciji kot R^2, saj upošteva zmanjšanje števila prostostnih stopenj zaradi večanja števila spremenljivk v modelu.
- **Pozor:** Vendar pa je pomen R^2_{pop} malce drugačen od pomen R^2, saj strogo gledano to ni več delež pojasnjene variabilnosti
- Prilagojeni koeficient je torej predvsem koristno uporabljati v modelih (in za primerjavo modelov), v katerih nastopa veliko spremenljivk ali jih ocenjujemo na malih vzorcih.

Nadaljevanje primera

- V našem primeru smo dobili vrednost standardne napake ovene 14,052.
- To pomeni, da je standardni odklon rezidualov/napak 14,052 oz. da se dejanske vrednosti standardno odklanjajo od napovedanih vrednosti za 14,052.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,573[a]	,329	,292	14,052

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

Nadaljevanje primera

- V našem primeru smo dobili vrednost popravljnega R^2 0,292, kar je kar opazno manj kot vrednost R^2, ki je 0,329. Razlika je zaradi majhnega vzorca (40 enot) razmeroma velika kljub majhnem številu neodvisnih spremenljivk.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,573[a]	,329	,292	14,052

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

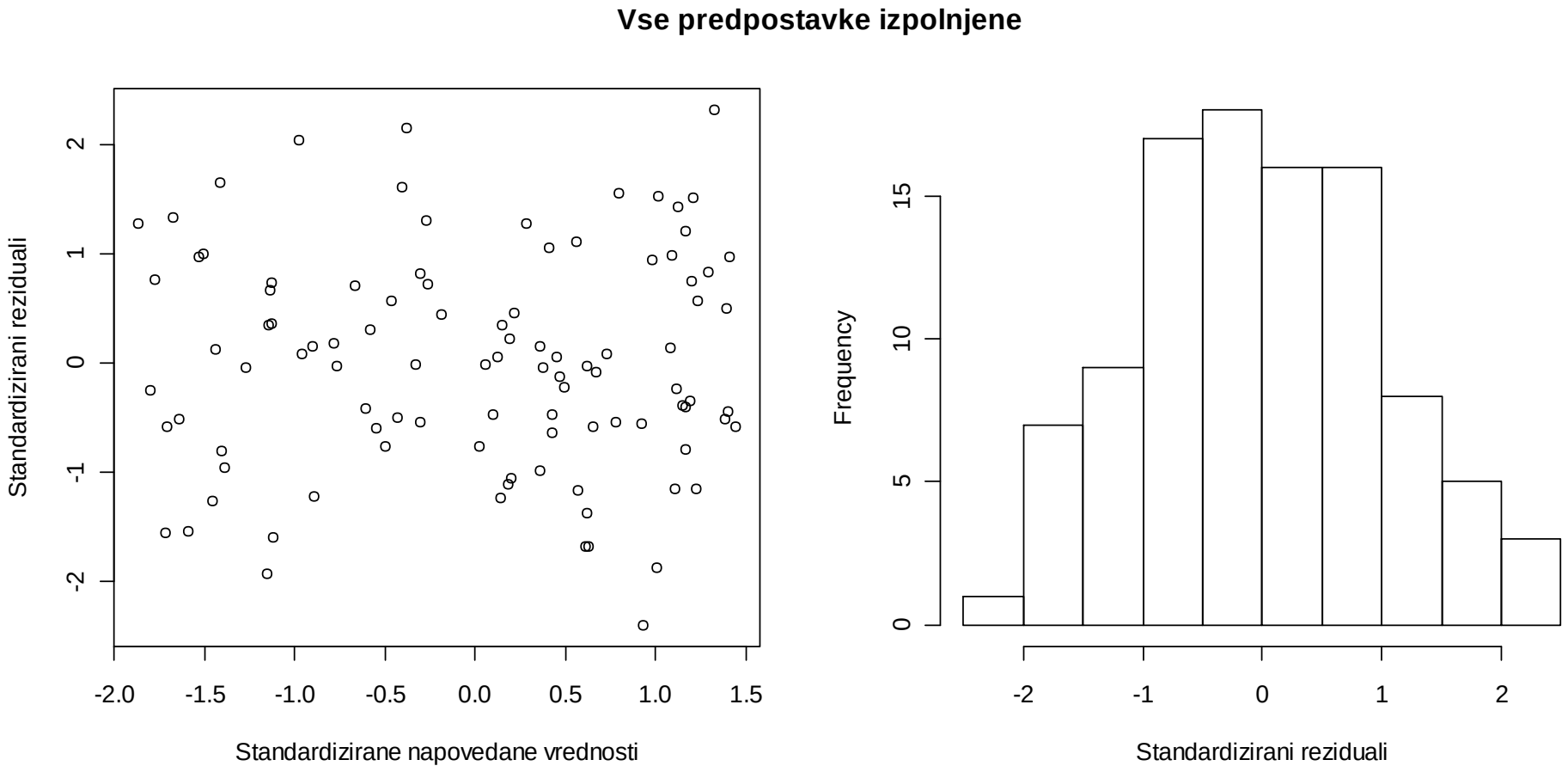
Analiza residualov

- Napake regresijskih napovedi $e_i = Y_i - Y_i'$ imenujemo tudi residuali.
- S pomočjo analize residualov lahko pogosto ugotovimo anomalije v regresijskem modelu, ki so lahko posledica različnih nepravilnosti, ki izhajajo iz neizpolnjevanja predpostavk regresijske analize.
- Najpreprostejše, vendar močno orodje za analizo residualov je grafični prikaz residualov. V dvorazsežnem koordinatnem sistemu prikažemo residuale navadno z razsevnim grafikonom, pri katerem na abscisno (x) os nanašamo regresijske napovedi, na ordinatno (y) os pa residuale pripadajoče tem napovedim.
- Če uporabljamo v regresiji same diskretne spremenljivke in imamo malo neodvisnih spremenljvk, je bolj primerno uporabiti kak drug grafikon.

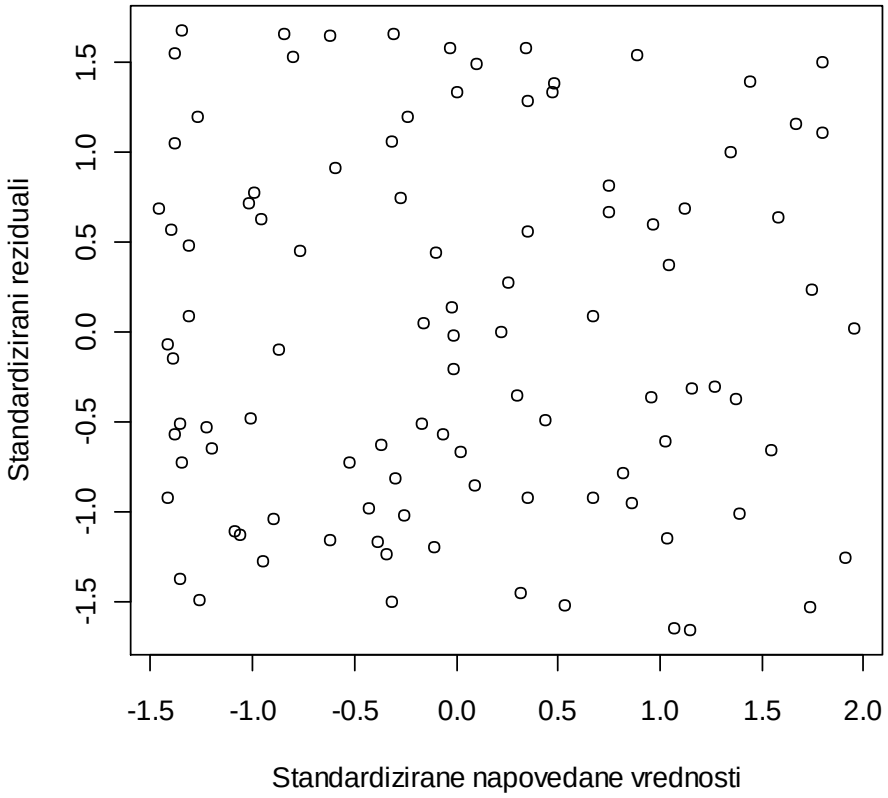
Pravilna porazdelitev residualov

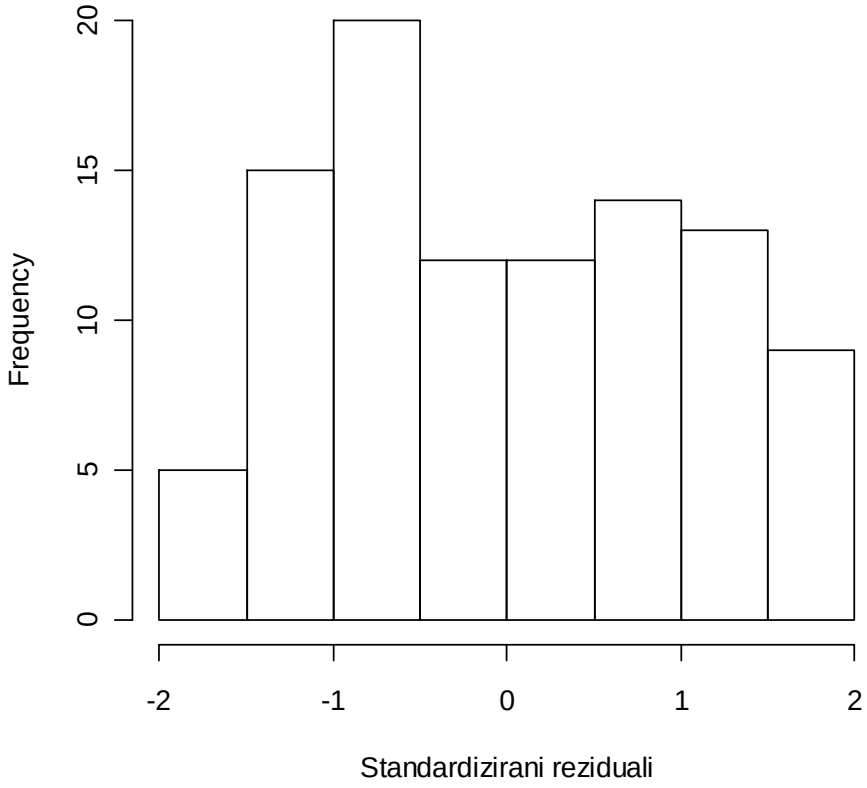
- V primeru, ko so izpolnjene predpostavke regresijske analize, se residuali porazdeljujejo enakomerno okrog 0. Natančneje to pomeni:
  - V grafikonu ni mogoče opaziti nobenega pravilnega vzorca porazdeljevanja residualov – pravilna specifikacija modela.
  - V grafikonu ni velikih odstopanj od povprečja za majhno število enot (outliers).
  - Variabilnost je pri vseh vrednostih X približno enaka – izpolnjena je predpostavka o homoskedastičnosti.
- Poleg tega je koristno analizirati tudi histogram residualov, ki ne sme preveč odstopati od pripadajoče normalne krivulje.

Pravilna porazdelitev residualov

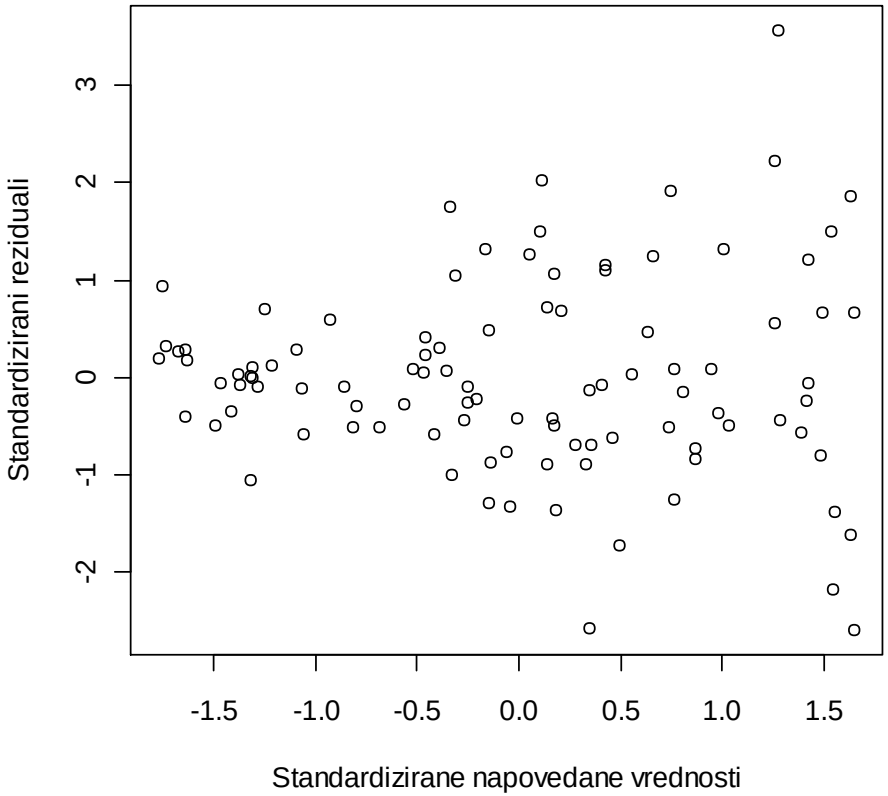


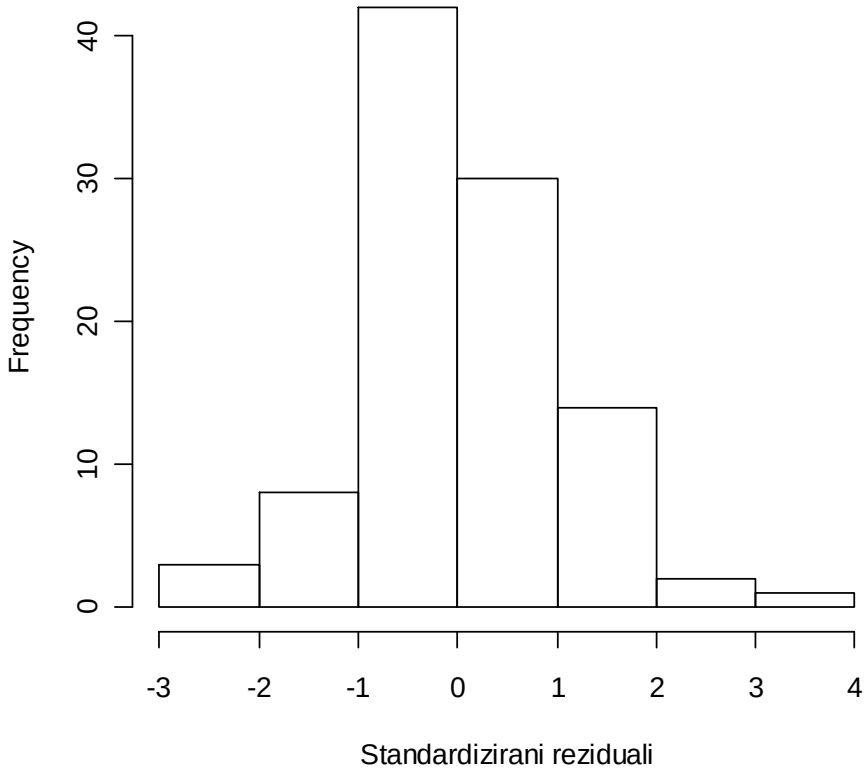
Nenormalna porazdelitev rezidualov



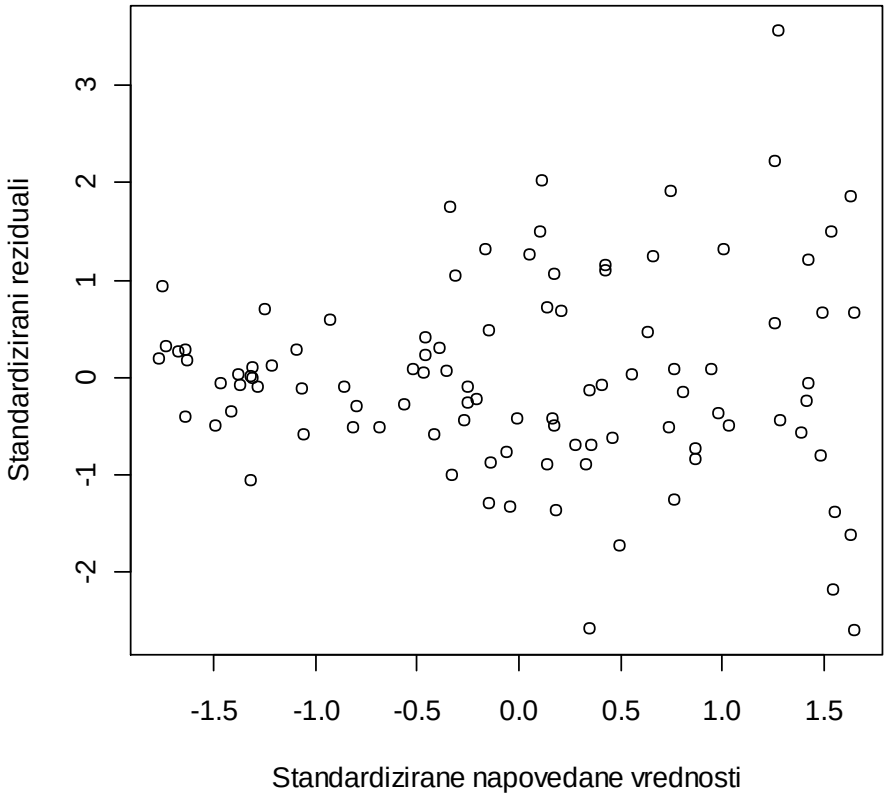


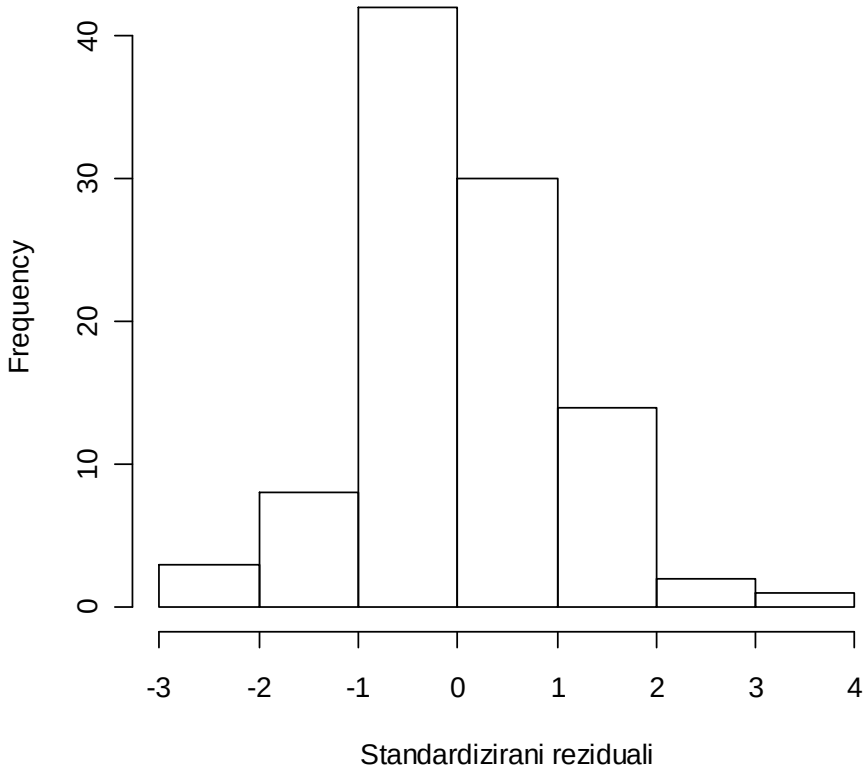
Heteroskedastičnost





Heteroskedastičnost





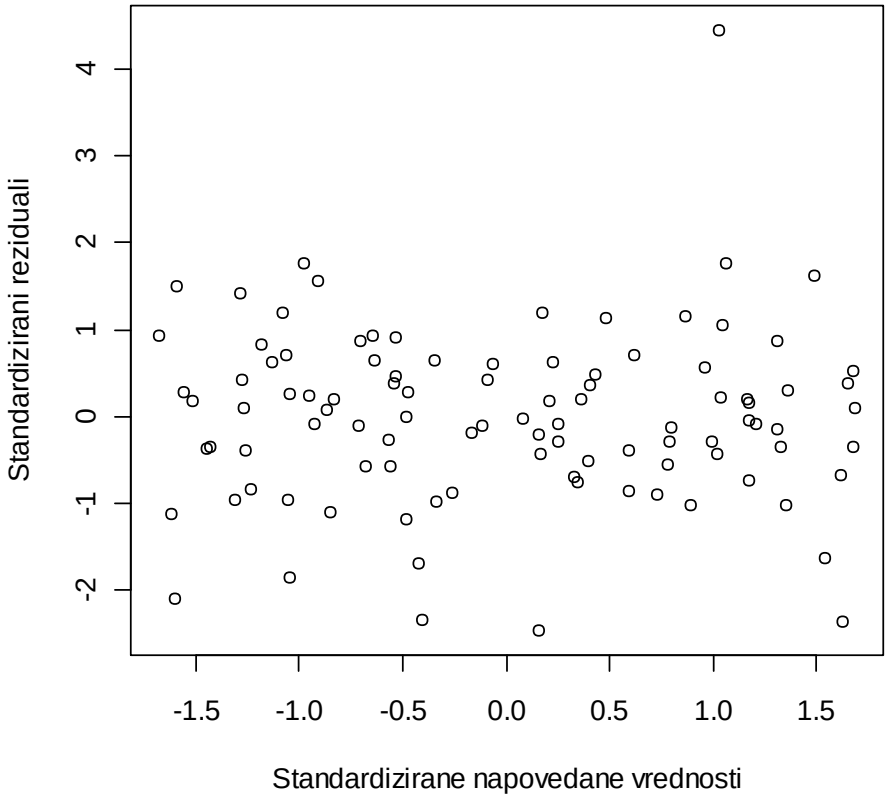
Problem velikih odstopanj (outliers)

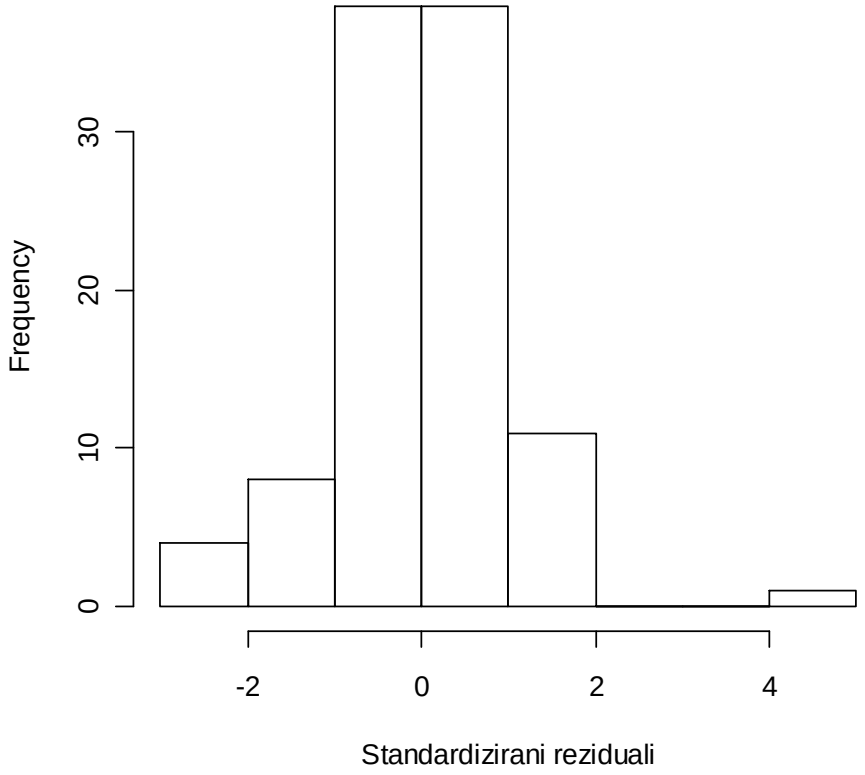
- Ta problem nastopi, ko imamo eno ali nekaj enot, ki izrazito izstopajo iz splošnega vzorca.
- Vzroki za velika odstopanja so različni:
  - napake v podatkih;
  - nelinearna zveza med spremenljivkami;
  - v vzorec je bila slučajno izbrana enota z ekstremnimi vrednostmi ipd.
- Posledica prisotnosti izstopajočih enot pa je slabo prileganje regresijskega modela in posledično tudi zavajajoči izračuni.

Kdaj lahko govorimo o velikih odstopanjih

- Grafični prikaz podatkov nam pomaga vizualno oceniti, ali imamo v podatkih velika odstopanja, vendar je včasih težko odločiti, ali so odstopanja v podatkih velika ali pa so v sprejemljivih mejah.
- Iz normalne porazdelitve sledi, da je pričakovano število enot, ki se odklanjajo za
  - več kot 2 standardna odklona, približno 5 %: torej pričakujemo, da jih od 100 enot za več kot dva standardna odklona odstopa okrog 5;
  - več kot 3 standardne odklone, 0,3 %: med 1000 enotami torej pričakujemo približno tri, ki odstopajo za več kot 3 standardne odklone.

Izstopajoče vrednosti

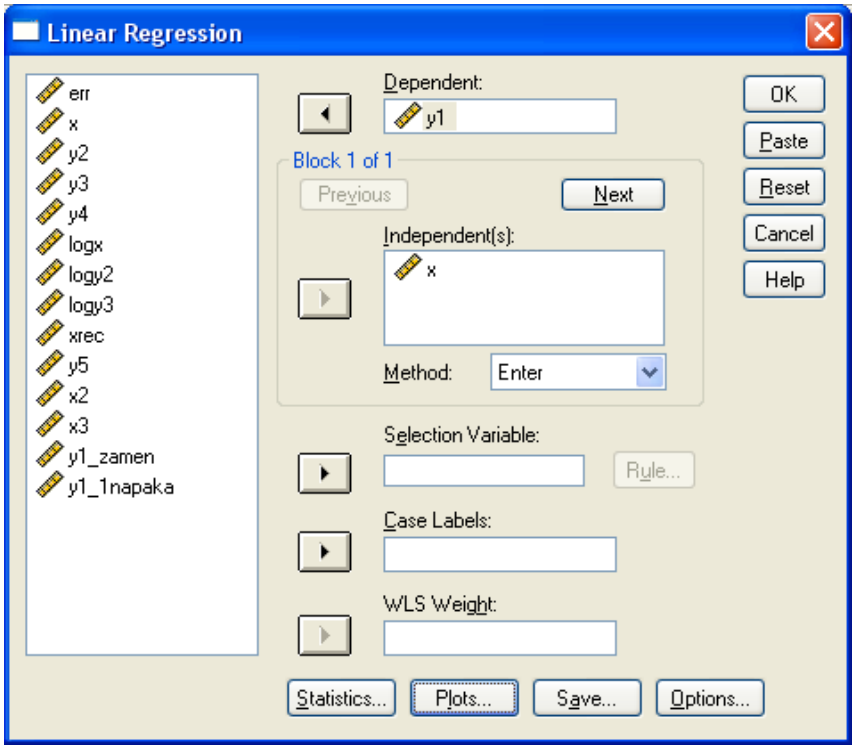


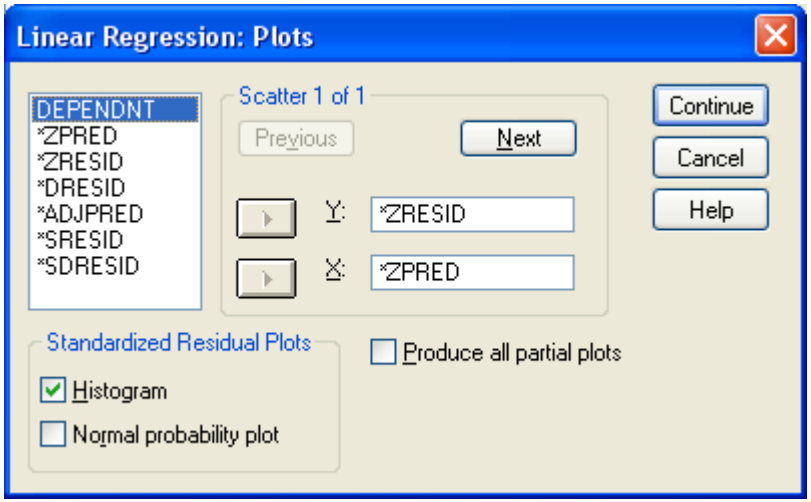


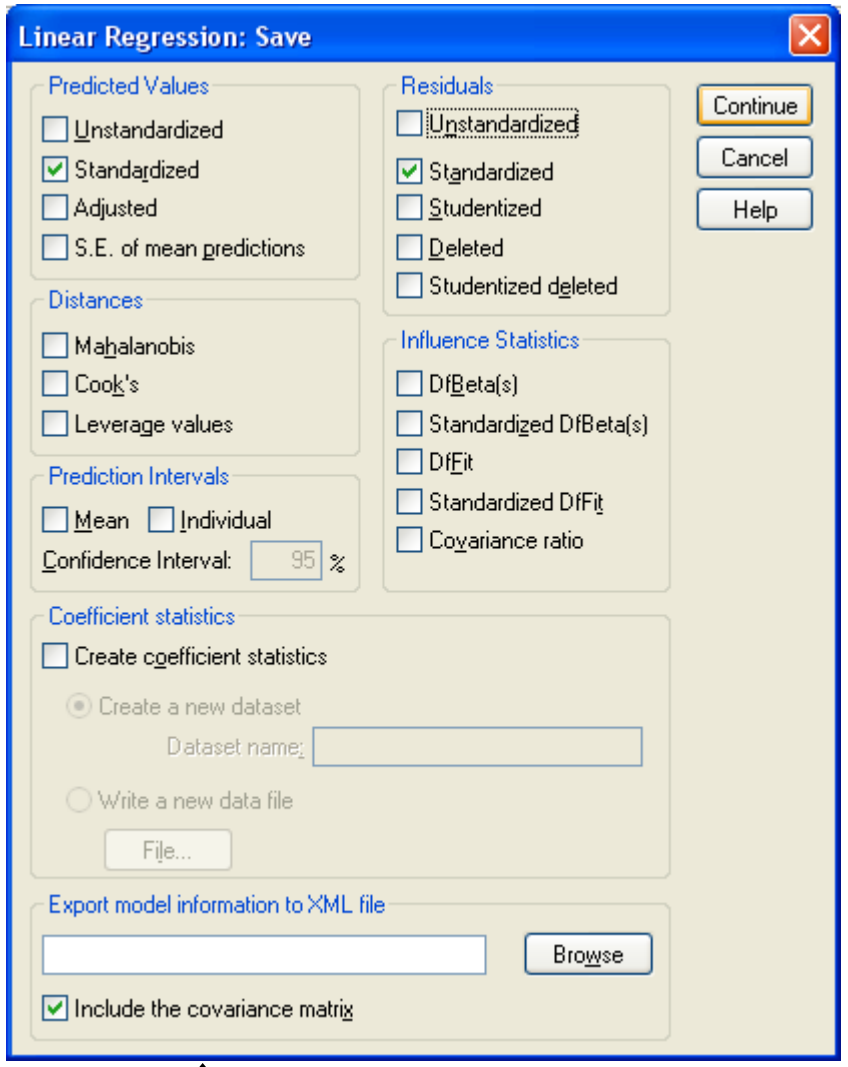
Primer analize rezidualov

- Za primer bomo uporabili umetne podatke (datoteka “nelin-regr-podatki.sav”)
- Zahtevamo linearno regresijo kot običajno (*Analyze* – *Regresion* – *Linear*).
- S klikom na gumb *Plots* odpremo ustrezno pogovorno okno.
- Levo spodaj odkljukamo *Histogram*.
- V zgornjem delu zahtevamo razsevni grafikon z standardiziranimi reziduali na y osi in standardiziranimi napovedanimi vrednostmi na x osi:
  - Iz seznama izberemo **ZRESID* (standardizirani reziduali) in jo prenesemo v polje *Y*
  - Iz seznama izberemo **ZPRED* (standardizirane napovedane vrednosti) in jo prenesemo v polje *X*
- V primeru, da imamo v modelu samo diskretne spremenljivke (in še teh malo), potem ta grafikon ni primeren. V tem primeru kliknemo na gumb *Save* in označimo, kaj naj se shrani v nove spremenljivke (te se dodajo v podatkovno datoteko. V okvirčkih *Predicted values* in *Residuals* odkljukamo *Standardized.*

Primer analize rezidualov

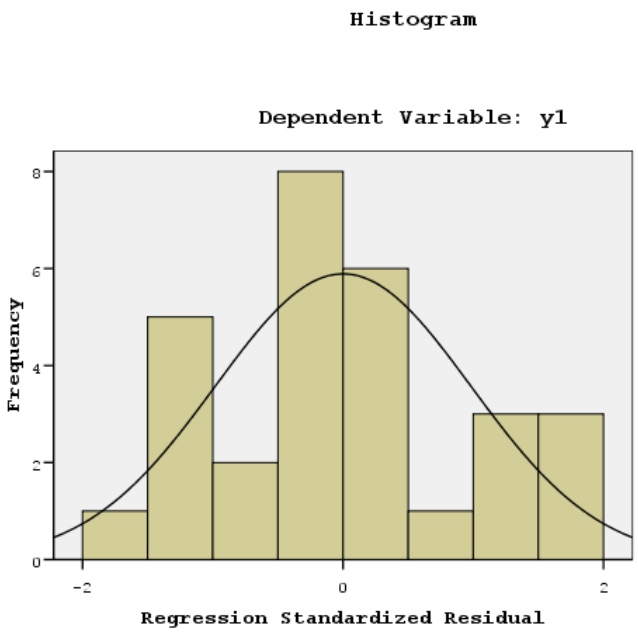


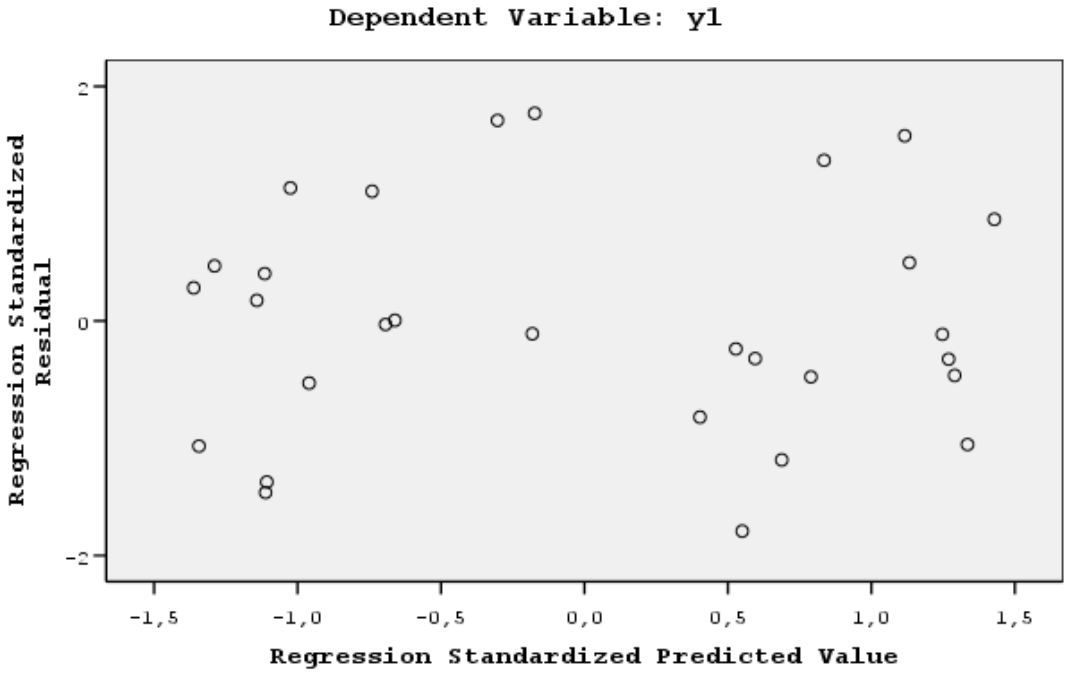




V našem primeru (ker imamo zvezne spremenljivke) to (celo okno) ni potrebno.

Primer grafikonov residualov ob izpolnjenih predpostavkah

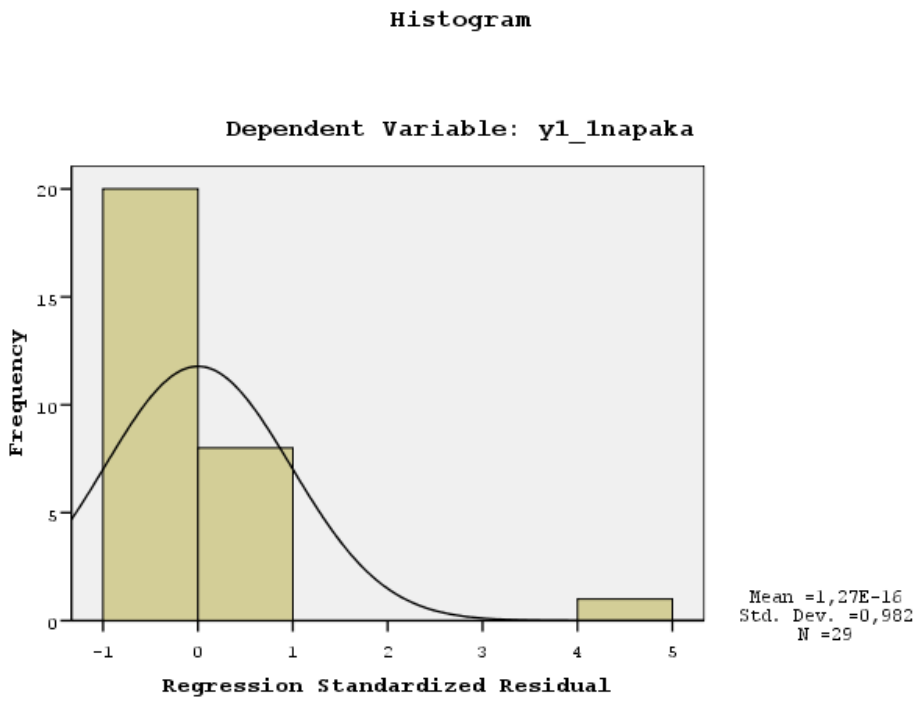


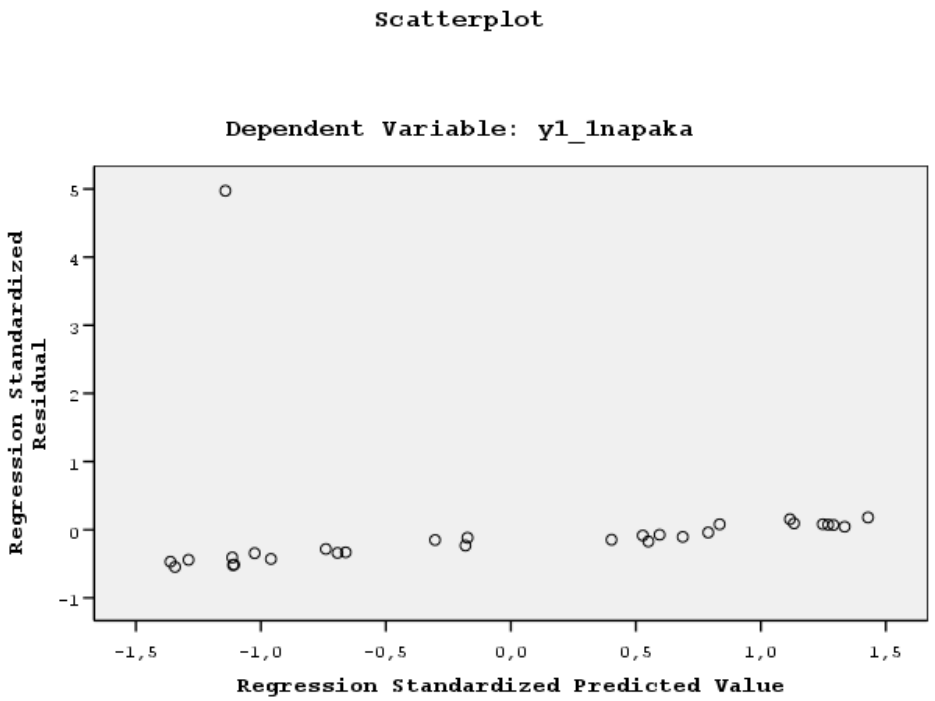


- Zaradi majhnega števila enot so pri histogramu vidna odstopanja od normalne porazdelitve. Pri velikem številu enot bi taka odstopanja že lahko kazala na težave
- Tudi na razsevnem grafikonu se nakazuje “obok”, a to je zgolj slučajna porazdelitev. Če bi imeli veliko enot, se taki vzorci ne smejo pojaviti.

Primer grafikonov residualov ob izstopajočih enotah

- Pri eni enoti smo napačno postavili decimalno vejico – 10-krat prevelika vrednost





Primer – rezultati regresije v prisotnosti izstopajočih enot

Brez izstopajočih enot

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,998[a]	,995	,995	,48735

a. Predictors: (Constant), x Neodvisna spremenljivka

b. Dependent Variable: y1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	2,748	,160		17,196	,000
	x Neodvisna spremenljivka	2,034	,027	,998	76,696	,000

a. Dependent Variable: y1

Z izstopajočo enoto

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,552[a]	,305	,279	8,11847

a. Predictors: (Constant), x Neodvisna spremenljivka

b. Dependent Variable: y1_1napaka

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	6,815	2,662		2,560	,016
	x Neodvisna spremenljivka	1,519	,442	,552	3,439	,002

a. Dependent Variable: y1_1napaka

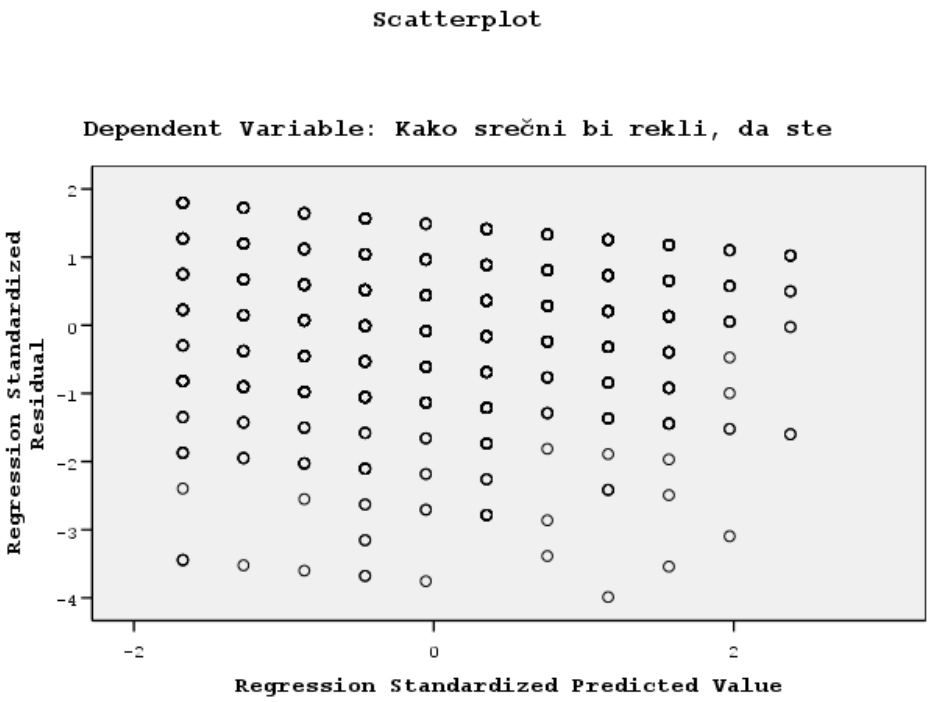
Kako ravnati v primeru, ko imamo izstopajoče enote

Enotnega recepta ni. Ravnamo pa lahko na naslednje načine:

- **Izključimo enote z ekstremnimi vrednostmi.** Ta način se sicer zdi najbolj eleganten vendar ima tudi vrsto pomanjkljivosti. Poleg tega, da zmanjšamo velikost vzorca, izgubimo (morda pomemben) del informacije.
- **Obravnavamo obe regresijski enačbi**, eno z ekstremnimi enotami, drugo pa z izključitvijo le-teh. Slabost te metode so težave pri interpretaciji dveh različnih enačb za isti model.
- **Transformiramo spremenljivko**. S transformacijo (na primer logaritmiranjem) se efekt ekstremnih vrednosti lahko zmanjša. Paziti pa je treba, da ima linearna zveza med transformiranimi spremenljivkami smiselno interpretacijo.
- **Povečamo velikost vzorca**. Če je to mogoče (kar običajno ni), lahko s tem ugotovimo naravo izstopajočih enot. Te včasih samo nakazujejo morebitno nelinearno funkcijsko zvezo.

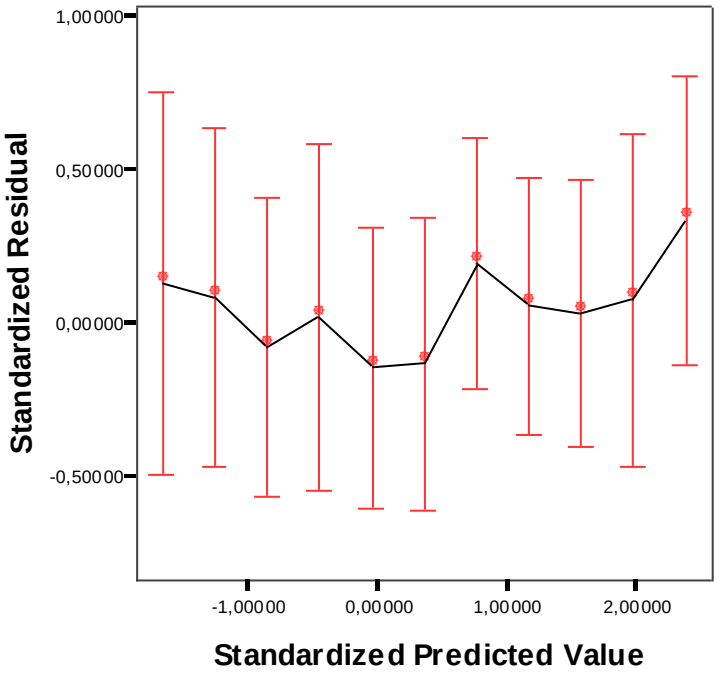
Primer analize rezidualov – diskretne spremenljivke

- Na podatkih ESS za leto 2004 za Slovenijo smo že analizirali odvisnost sreče posameznika (*C1*) od zaupanja ljudem (*A8*).
- Sedaj si poglejmo še analizo rezidualov.
- Ker imamo opravka le z dvema diskretnima spremenljivkama, navaden razsevni grafikon standardiziranih rezidualov glede na standardizirane napovedane vrednosti ni primeren prikaz podatkov.



Primer analize rezidualov – diskretne spremenljivke

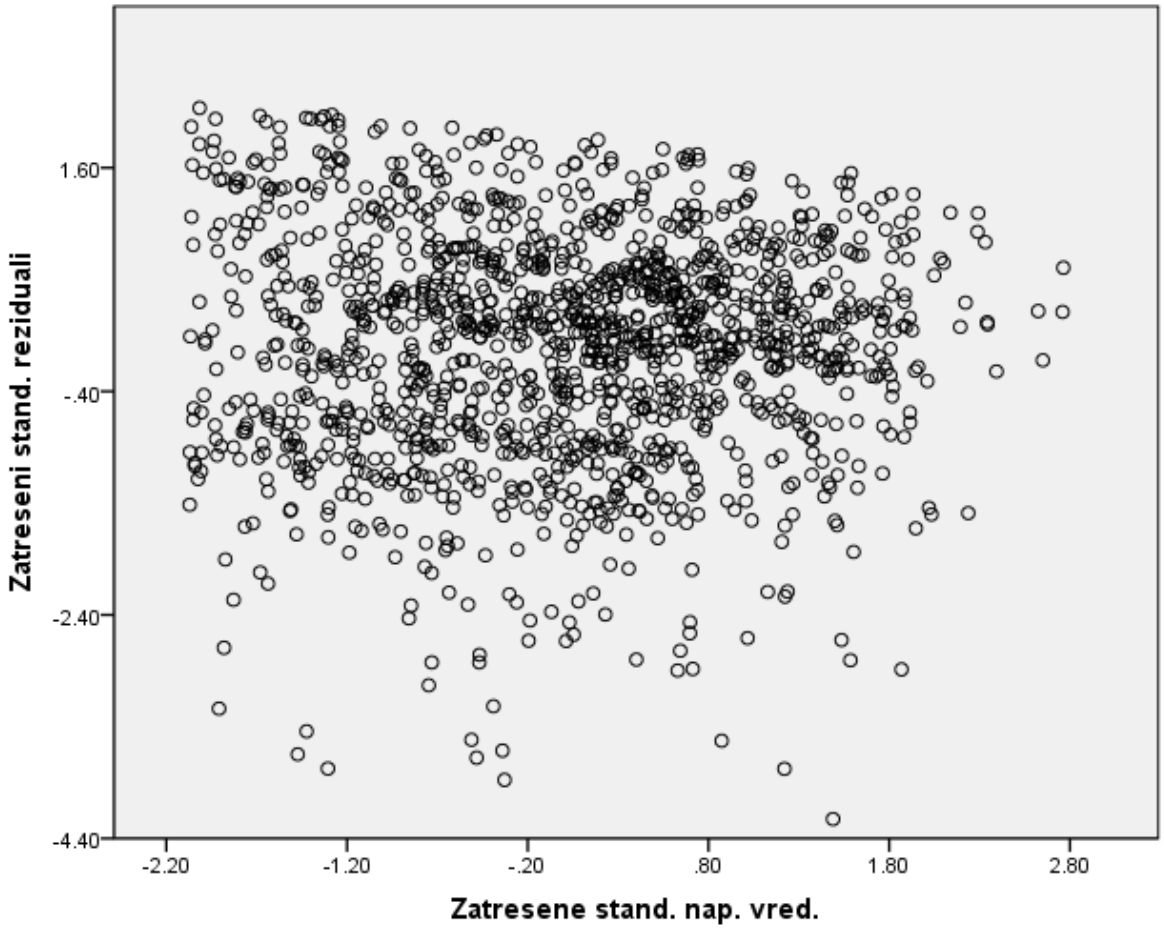
- Zato v oknu *Save* izberemo, da shranimo standardizirane rezidualov in standardizirane napovedane vrednosti.
- Kerirajte se dve novi spremenljivki, *ZPR_1* (standardizirane napovedane vrednosti) in *ZRE_1* (standardizirane rezidualov).
- Z njimi narišemo grafikon povprečju standardiziranih rezidualov glede na standardizirane vrednosti.
- V idealnem primeru bi bili vsi intrevali (rdeče črte) enako dolge, vsa povprečja pa enaka 0.



- V našem primeru odstopanje ni tolikšno in tako sistematično, da bi govorili o kršenih predpostavkah, čeprav se nakazuje “u” krivulja.

Primer analize rezidualov – diskretne spremenljivke

- Zato v oknu *Save* izberemo, da shranimo standardizirane rezidualov in standardizirane napovedane vrednosti.
- Kerirajte se dve novi spremenljivki, *ZPR_1* (standardizirane napovedane vrednosti) in *ZRE_1* (standardizirane rezidualov).



- Spremenljivke zatresemo.
- Narišemo klasični razsevni grafikon.

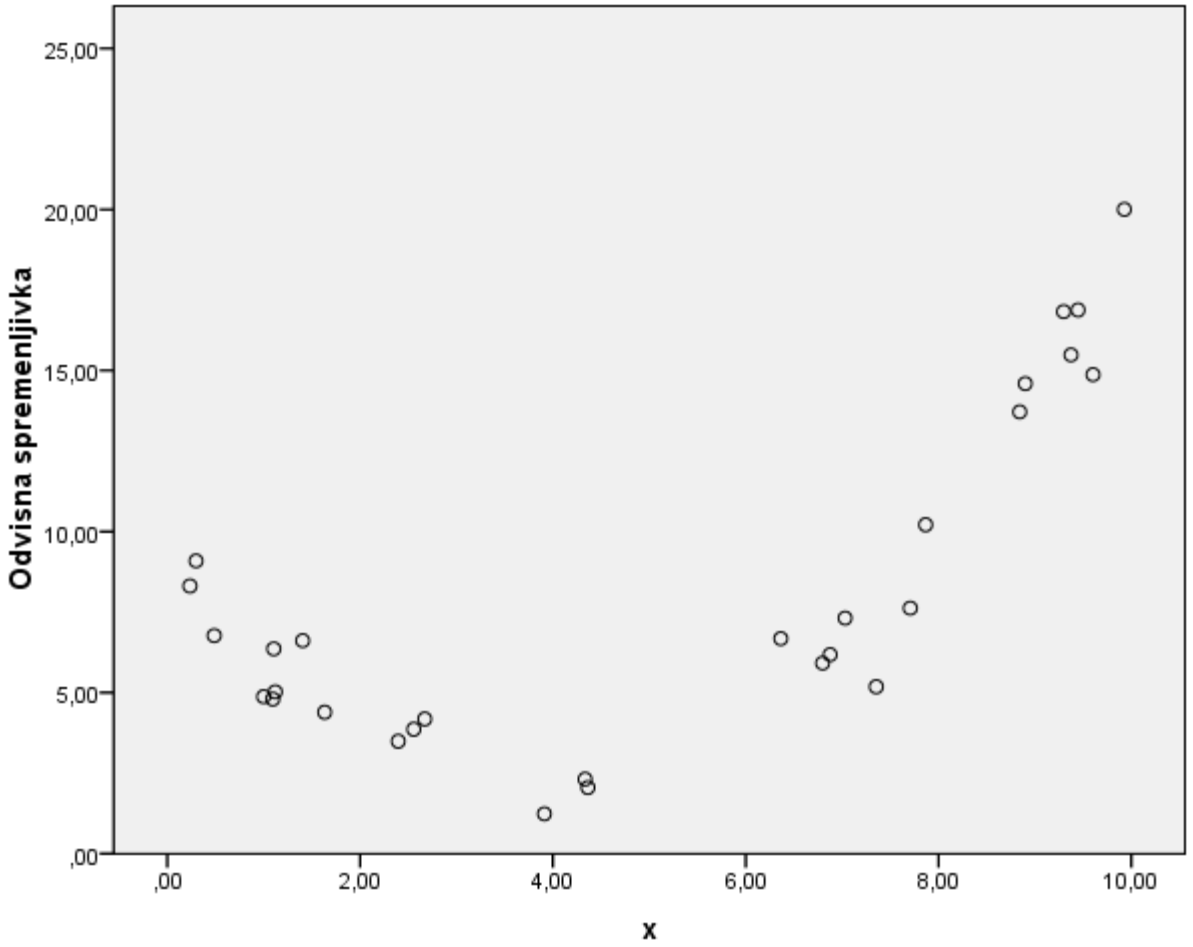
Polinomski regresijski model

- Omenili smo že, da pogosto med spremenljivkami nimamo linearne regresijske zveze, temveč je regresijska funkcija splošnejša. Poseben primer splošnejše regresijske funkcije je polinomska funkcija.
- Ogledali smo si že enega od načinov prevedbe polinomskega modela na linearni, s pomočjo transformacije spremenljivk.
- Vendar pa včasih ni lahko uganiti prave transformacije. V primeru polinomske zveze si lahko pomagamo tudi z multiplo regresijo.
- Matematično lahko opišemo kvadratno zvezo med spremenljivkama X in Y z naslednjo enačbo:

  $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
- Oziroma z njeno stohastično verzijo:

  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
- Tako zvezo imenujemo **polinomska regresija druge stopnje**.

Primer – razsevni grafikon polinomska regresija 2. stopnje



Polinomska regresijo *k*-te stopnje

- Podobno lahko definiramo tudi polinomsko regresijo *k*-te stopnje:

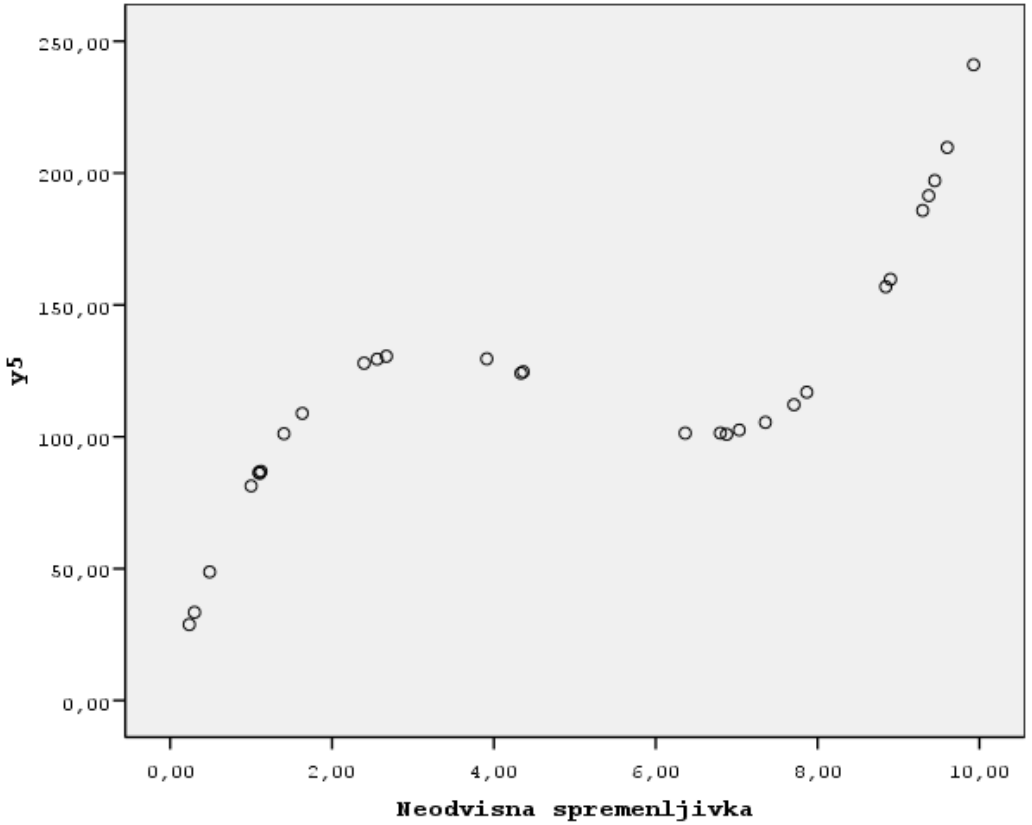
$$Y = \beta_0 + \beta_1 X + \ldots + \beta_k X^k + \varepsilon$$

- Ocenjevanje polinomskega regresijskega modela poteka podobno kot ocenjevanje običajnega multiplega regresijskega modela, pri čemer štejemo potence spremenljivke X za različne spremenljivke.
- Postavlja se vprašanje, ali so še vedno izpolnjene predpostavke regresijskega modela; predvsem zahteva o multipli nekolinearnosti.
- Res je, da so spremenljivke med seboj povezane, celo funkcijsko. Vendar pa povezanost ni linearna, zato (če ne pretiravamo s številom členov) lahko predpostavimo, da med njimi ni multikolinearnosti.
- Polinomski regresijski model lahko uporabimo tudi kot del regresijskega modela, ki vsebuje še druge spremenljivke.

Primer polinomske regresije

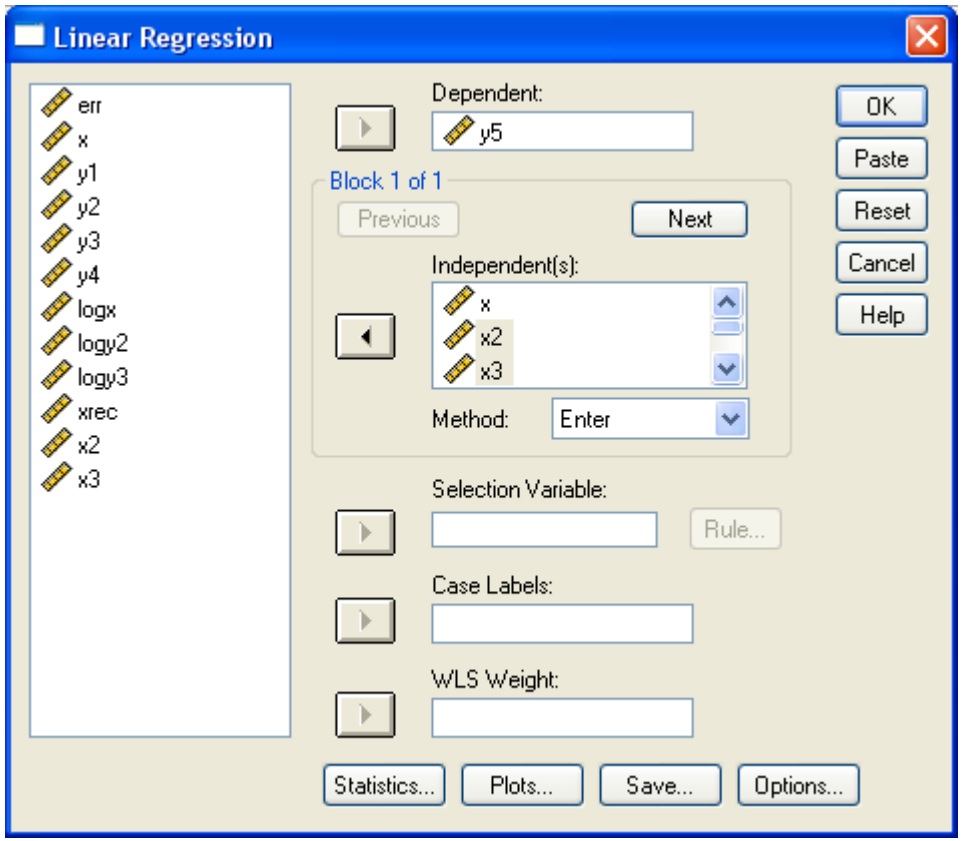
- **Analizirajmo odnos med spremenljivkama *y5* in *x* v datoteki “nelin-regr-podatki.sav”.**
- **Najprej narišemo razsevni grafikon.**

- Nakazuje se polinomska zveza 3. stopnje med spremenljivkama.
- Stopnja zveze mora biti vsaj enaka “št. zavojev” + 1



Primer polinomske regresije

- Predpostavimo, da med spremenljivkama obstaja največ polinomska zveza 3. stopnje. Zato (s proceduro *Transform – Compute*) kreiramo dve dodatni spremenljivki, $x2 = x^2$ in $x3 = x^3$.
- Ocenimo multipli regresijski model z odvisno spremenljivko *y5* in neodvisnimi spremenljivkami *x, x2* in *x3.*



Primer polinomske regresije – rezultati ob pravilni specifikaciji

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1,000[a]	1,000	1,000	,48418

a. Predictors: (Constant), x3, x Neodvisna spremenljivka, x2

b. Dependent Variable: y5

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	69253,254	3	23084,418	98470,817	,000[a]
	Residual	5,861	25	,234		
	Total	69259,115	28			

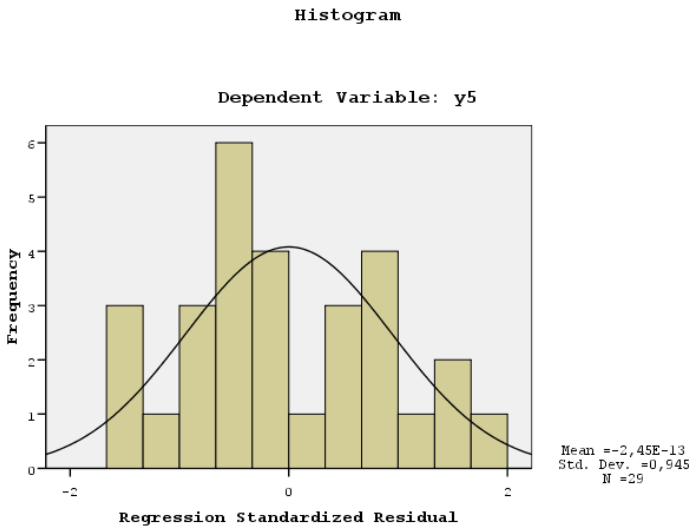
a. Predictors: (Constant), x3, x Neodvisna spremenljivka, x2

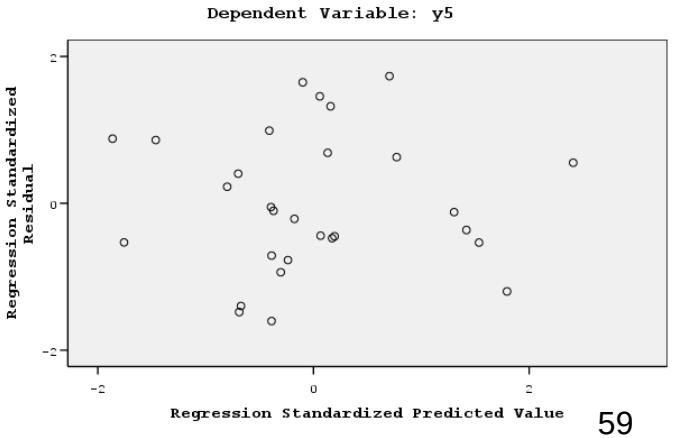
b. Dependent Variable: y5

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7,354	,313		23,518	,000
	x Neodvisna spremenljivka	94,504	,309	6,599	306,185	,000
	x2	-22,108	,071	-15,800	-310,468	,000
	x3	1,507	,005	10,262	328,954	,000

a. Dependent Variable: y5





Primer polinomske regresije – rezultati ob napačni (linearni) specifikaciji

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,782[a]	,611	,597	31,58049

a. Predictors: (Constant), x Neodvisna spremenljivka

b. Dependent Variable: y5

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	42331,278	1	42331,278	42,445	,000[a]
	Residual	26927,838	27	997,327		
	Total	69259,115	28			

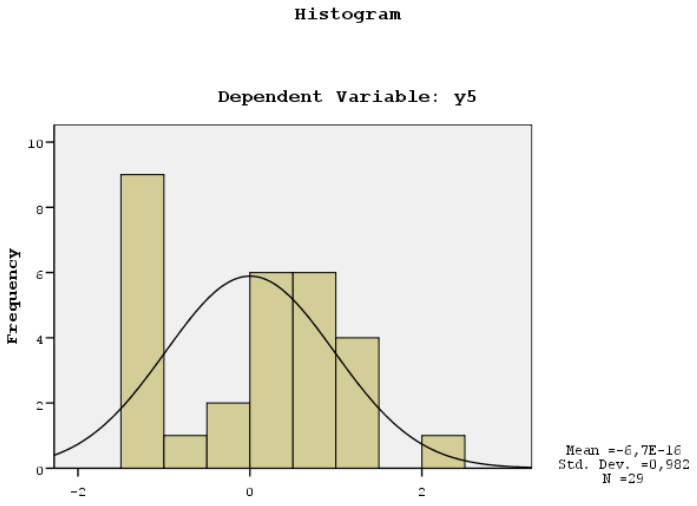
a. Predictors: (Constant), x Neodvisna spremenljivka

b. Dependent Variable: y5

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	65,467	10,354		6,323	,000
	x Neodvisna spremenljivka	11,196	1,718	,782	6,515	,000

a. Dependent Variable: y5



Scatterplot