Predpostavke niso izpolnjene

Metode odpravljanja kršitev predpostavk linearnega modela v splošnem delimo v dve skupini:

- transformacije
- modeliranje

Predpostavke niso izpolnjene

Pregled metod

Pregled metod:

- Konstantna varianca napak
  - transformacija odzivne spremenljivke
  - modeliranje: tehtana metoda najmanjših kvadratov, interakcijski členi
- Linearnost
  - transformacija odzivne spremenljivke in/ali napovednih spremenljivk
  - modeliranje:polinomska regresija, interakcijski členi
  - modeliranje: zlepki, aditivni modeli (*splines*, *additive models*)
- Vplivnost točk (ni predpostavka modela, lahko posledica neizpolnjenih predpostavk)
  - transformacija napovednih spremenljivk
  - modeliranje: tehtana metoda najmanjših kvadratov, interakcijski členi

Predpostavke niso izpolnjene

Pregled metod

- Normalna porazdelitev ostankov
  - transformacija odzivne spremenljivke
  - modeliranje: posplošeni linearni modeli (GLM) (ne bomo obravnavali)
- Neodvisnost napak
  - diferenciranje (časovne vrste)
  - modeliranje: linearni mešani modeli (longitudinalni, hierarhični)
  - modeliranje: GEE modeli *Generalised Estimating Equations*
  - modeliranje: kopule (*copulas*)

Transformacije
Transformiramo lahko odzivno in/ali napovedne spremenljivke.

Za vsako transformacijo je v procesu modeliranja potrebno ugotoviti, ali smo problem res rešili:

- če nas pri modeliranju zanima zveza med odzivno in napovednimi spremenljivkami ali pa gre za statistično sklepanje, potem naredimo model na transformirani spremenljivki in ponovno izvedemo diagnostiko modela;
- če modeliramo z namenom napovedovanja, potem ustreznost izbire transformacije preverimo na podlagi PRESS-ostankov. Primerjamo vsoto kvadratov PRESS ostankov modela na transformiranih in netransformiranih podatkih (o tem bomo več govorili v poglavju o izbiri modela).

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

Tabela: Najpogosteje uporabljene transformacije pri različnih zvezah med varianco σ^2 in pričakovano vrednostjo $\mathbb{E}(y)$; znak $\propto$ pomeni sorazmernost

Zveza σ^2 do $\mathbb{E}(y)$	Transformacija $f(y)$	Opomba
$\sigma^2 \propto$ konstanta	y	ni transformacije
$\sigma^2 \propto \mathbb{E}(y)$	$\sqrt{y}$	y je frekvenca, Poissonova porazdelitev
$\sigma^2 \propto \frac{\mathbb{E}(y)}{1-\mathbb{E}(y)}$	$arcsin(\sqrt{y})$, $logit(y)$	y je delež, binomska porazdelitev
$\sigma^2 \propto \mathbb{E}(y)^2$	$log(y)$	$y > 0$
$\sigma^2 \propto \mathbb{E}(y)^4$	y^{-1}	$y \neq 0$

Transformacija za delež

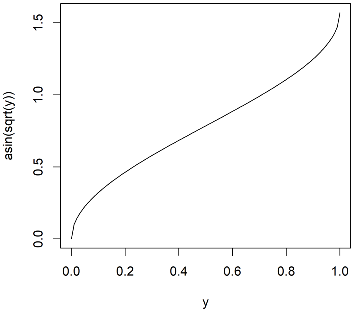
y je delež, omejena zaloga vrednosti na intevalu [0,1]. V ozadju je slučajna spremenljivka z:

$$z \sim b(n, \pi), \quad \mathbb{E}(z) = n\pi, \quad Var(z) = n\pi(1-\pi)$$

Varianca deležev blizu 0 oz. blizu 1 je manjša od variance deležev blizu $1/2$. Z nekonstantno varianco ni težav, če so vrednosti deležev približno na intervalu [0.25, 0.75].

Transformaciji: $asin(\sqrt{y})$ in $logit(y)$.

$$f(y) = asin(\sqrt{y}) \quad y \in [0,1]$$

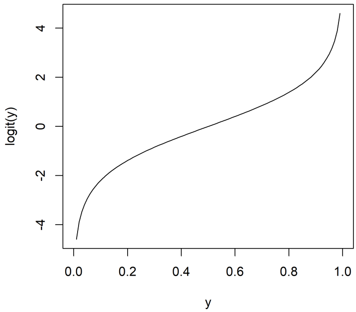


Pokažemo lahko, da je varianca transformiranih vrednosti približno $\frac{1}{4}n$ in je tako neodvisna je od π. To pomeni, da ta transformacija stabilizira varianco.

Predpostavke niso izpolnjene

Klasične transformacije odzivne spremenljivke

$$f(y) = logit(y) = ln\frac{y}{1-y} \quad y \in (0, 1)$$



Logit transformacija je osnova logistični regresiji (GLM).

Box-Cox transformacije (Box in Cox, 1964)

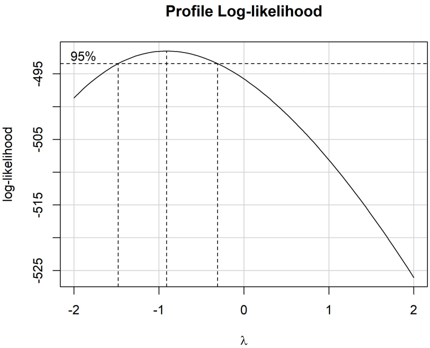
Družino transformacij za odvisno spremenljivko y, ki je funkcija parametra λ. Za i-to točko $i = 1, ..., n$ v tem primeru linearni model zapišemo:

$$z_i = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda \neq 0 \\ ln(y_i) = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda = 0 \end{cases} \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$$

Box-Cox transformacije so definirane za $\mathbf{y} > 0$.

Z metodo največjega verjetja (*maximum likelihood*) hkrati ocenjujemo $\boldsymbol{\beta}$ in λ.

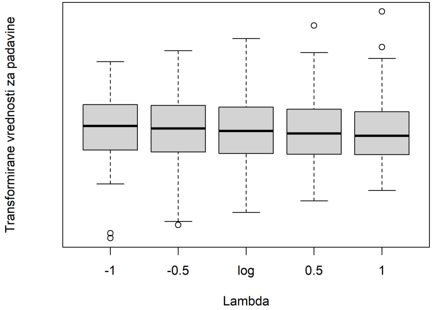
Oceno $\hat{\lambda}$ dobimo z numeričnim postopkom: za različne vrednosti $\hat{\lambda}$ izračunamo verjetje $L(\mathbf{y}, \mathbf{X}, \beta, \lambda)$. Izberemo λ, pri kateri ima logaritem verjetja maksimalno vrednost.



Primer: `postaje`, funkcija `boxCox()`

Predpostavke niso izpolnjene

Box-Cox transformacije



Slika: Okvirji z ročaji za različne transformacije odzivne spremenljivke, približna izbira parametra λ

Primer: `postaje`, funkcija `symbox()`

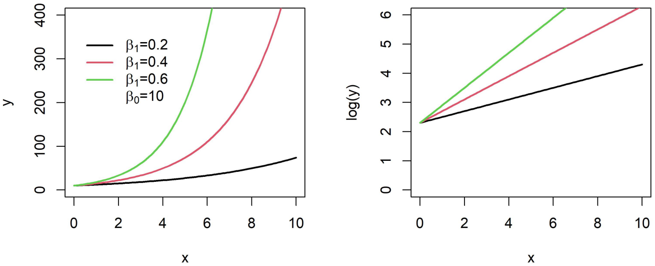
Exponentna zveza

Če zvezo med y in x opišemo z eksponentno funkcijo:

$$y = \beta_0 e^{\beta_1 x},$$

z logaritmiranjem izraza dobimo linearno zvezo:

$$ln(y) = ln(\beta_0) + \beta_1 x = \beta_0^* + \beta_1 x$$



Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, eksponentna zveza

Pomen parametra β_1:

$$\beta_1 = ln(y(x+1)) - ln(y(x)) = ln\frac{y(x+1)}{y(x)}$$

$$\frac{y(x+1)}{y(x)} = e^{\beta_1} \quad \frac{y(x+1) - y(x)}{y(x)} = e^{\beta_1} - 1.$$

Če se x poveča za eno enoto, se y spremeni za $100 \cdot (e^{\beta_1} - 1)$ %.

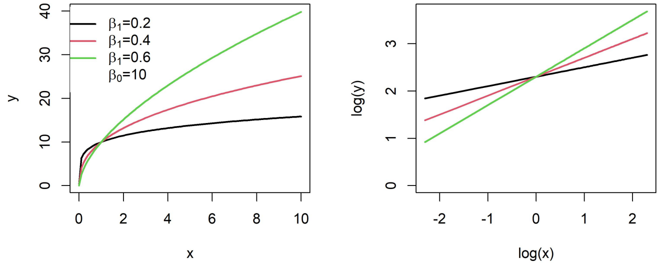
Pri $x = 0$ je povprečna vrednost $\bar{y} = e^{\beta_0^*}$

Multiplikativna zveza

$$y = \beta_0 x^{\beta_1},$$

Lineariziramo z logaritmiranjem:

$$log(y) = log(\beta_0) + \beta_1 log(x),$$



Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, multiplikativna zveza

Pomen parametra β_1:

$$\beta_1 = \frac{dy/y}{dx/x}.$$

Če se x poveča za 1 %, se y spremeni za β_1 %.

Predpostavke niso izpolnjene

Nelinearnost, ki se da linearizirati, multiplikativna zveza

Dva regresorja x_1 in x_2 v multipliktivnem odnosu z y:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2}.$$

$$log(y) = log(\beta_0) + \beta_1 log(x_1) + \beta_2 log(x_2).$$

Če se x_1 poveča za 1 %, se y spremeni za β_1 %, ko je vrednost x_2 konstantna.

Predpostavke niso izpolnjene

Vloga napake v modelu za transformirane spremenljivke

Eksponentna zveza, kjer je napaka v aditivi zvezi z odzivno spremenljivko:

$$y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i.$$

Logaritmiranje v tem primeru ne privede do linearnega modela, ker za $log(\varepsilon_i)$ ne moremo predpostaviti normalne porazdelitve, če normalna porazdelitev velja za ε_i.

$$log(y_i) = log(\beta_0) + \beta_1 x_i + log(\varepsilon_i),$$

Drugače je, če napaka v izrazu nastopa multiplikativno:

$$y_i = \beta_0 e^{\beta_1 x_i + \varepsilon_i},$$

$$log(y) = log(\beta_0) + \beta_1 x_i + \varepsilon_i.$$

Ker v praksi ne vemo, kateri model napake je pravi, je v splošnem nelinearne zveze bolje modelirati z nelinearnimi modeli.

Primer: `kovine`

Linearni model z interakcijo dveh številskih spremenljivk

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

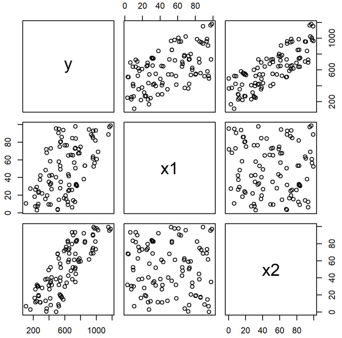
Interakcijske člene v model vključimo iz različnih razlogov:

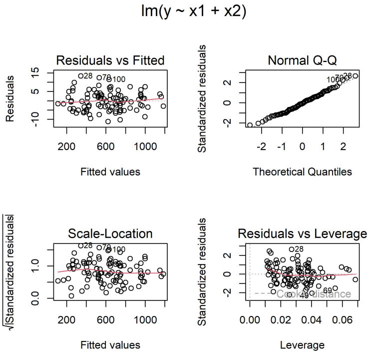
- poznavanje vpliva izbranih dejavnikov na odzivno spremenljivko
- iskanje utreznih regresorjev v linearnem modelu da izponimo predpostavke.

Predpostavke niso izpolnjene

Primer generiranih podatkov **brez interakcije** dveh napovednih spremenljivk

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

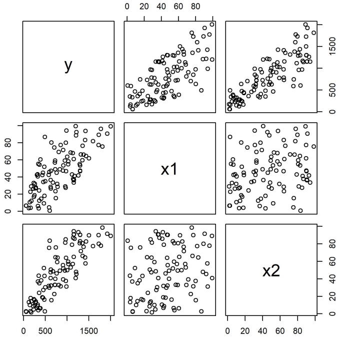




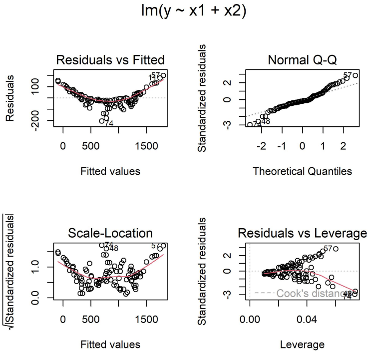
Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$



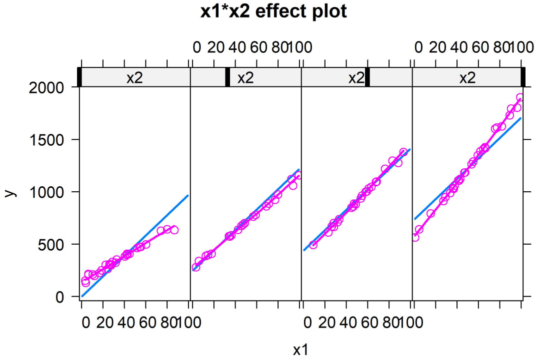
Ostanki modela brez interakcije



Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk

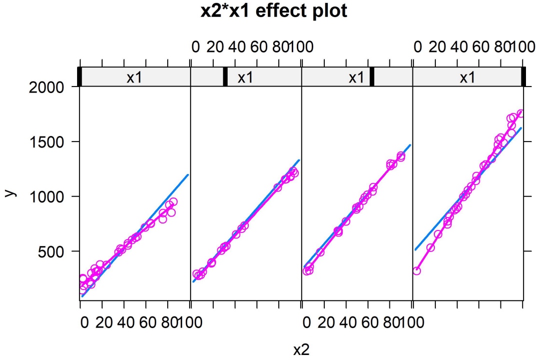
Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_1 pri izbranih intervalih vrednostih x_2



Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model brez interakcije

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_2 pri izbranih intervalih vrednostih x_1

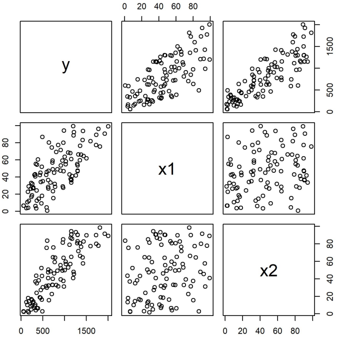


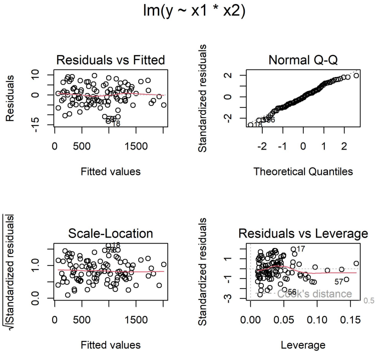
Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk, model vključuje inerakcijski člen

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i$$

Ostanki modela z interakcijo

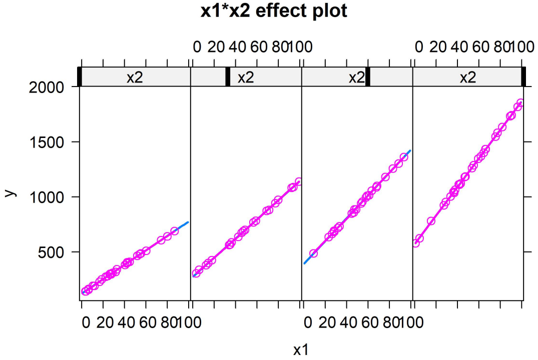




Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model z interakcijskim členom

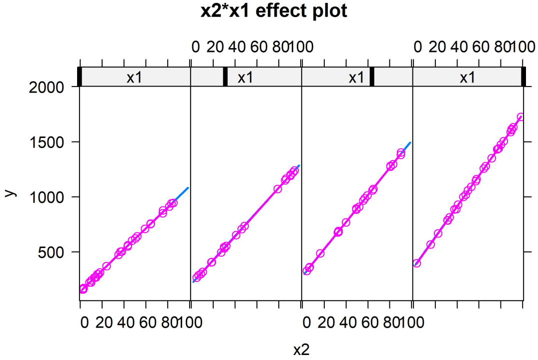
Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_1 pri izbranih intervalih vrednostih x_2



Predpostavke niso izpolnjene

Primer generiranih podatkov **z interakcijo** dveh napovednih spremenljivk, graf parcialnih ostankov za model z interakcijskim členom

Graf parcialnih ostankov za model brez interakcije, y v odvisnosti od x_2 pri izbranih intervalih vrednostih x_1



Predpostavke niso izpolnjene

Interakcija dveh številskih napovednih spremenljivk

Zamislimo si pričakovano vrednost tega modela v točki (x_{01}, x_{02}).

$$E(y|x_{01}, x_{02}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_3 x_{01} x_{02},$$

in v točki $(x_{01}, x_{02} + 1)$

$$\begin{aligned} E(y|x_{01}, x_{02} + 1) &= \beta_0 + \beta_1 x_{01} + \beta_2 (x_{02} + 1) + \beta_3 x_{01} (x_{02} + 1) \\ &= \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \beta_2 + \beta_3 x_{01} x_{02} + \beta_3 x_{01}. \end{aligned}$$

Razlika:

$$E(y|x_{01}, x_{02} + 1) - E(y|x_{01}, x_{02}) = \beta_2 + \beta_3 x_{01}.$$

Primer: `postaje`

Trasformacije, ki zmanjšajo vplivnost točk

Večja vplivnost točk je lahko povzročena zaradi splošnega kršenja predpostavk linearnega modela:

- nekonstantna varianca
- nelinearne zveze med odzivno spremenljivko in regresorji
- nenavadne vrednosti regresorja

S transformacijo odzivne spremenljivke ali regresorjev lahko dosežemo **zmanjšanje** ali pa tudi **povečanje** vplivnosti posameznih točk v modelu.

Primer: `mammals`

Metoda tehtanih najmanjših kvadratov
(WLS, *Weighted Least Squares*)

Predpostavimo:

- napake ε_i so neodvisne in normalno porazdeljene
- varianca napak ni konstantna:
  - $Var(\varepsilon_i) = \sigma_i^2$, $i = 1, ..., n$
  - $Var(\varepsilon_i) = \sigma^2 w_i$, $i = 1, ..., n$, w_i so uteži

Uteži w_i so znane in pozitivne, zapišemo jih v diagonalno matriko $\mathbf{W}$ dimenzije $(n \times n)$.

Predpostavke niso izpolnjene
WLS

Ocene parametrov modela po metodi WLS dobimo z minimiranjem izraza

$$S(\boldsymbol{\beta}, \mathbf{W}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ii}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2.$$

Večja utež W_{ii} pomeni, da ima i-ti podatek večji vpliv na oceno parametrov modela.

Povezava med utežmi in varianco napak v normalnem linearnem modelu, ko ta ni konstantna:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{V})$$

$\mathbf{V}$ je diagonalna **variančna matrika napak** dimenzije $n \times n$. Vrednosti na diagonali so različne.

Za oceno parametrov modela in komponent variančne matrike napak uporabimo **metodo največjega verjetja**. Metodo imenujemo tudi **posplošena metoda najmanjših kvadratov** (GLS, *Generalized Least Square*)

$$L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi det(\mathbf{V}))^{\frac{n}{2}}} exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

Predpostavimo, da je $\mathbf{V}$ znana. Maksimiranje zgornjega izraza je enakovredno minimiranju **posplošene vsote kvadratov napak**:

$$S(\boldsymbol{\beta}, \mathbf{V}^{-1}) = \sum_{i=1}^{n} \mathbf{V}_{ii}^{-1}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Če primerjamo izraz, ki smo ga dobili po WLS in izraz pri GLS , vidimo, da je $\mathbf{W} = \mathbf{V}^{-1}$.

Utež za posamezen podatek je torej obratno sorazmerna z njegovo varianco, kar pomeni, da damo podatku z večjo varianco manj pomembnosti pri ocenjevanju parametrov modela.

Rešitev je

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}$$

- Če poznamo variance σ_i^2 ali uteži w_i ali če podatki omogočajo, da variance oziroma uteži ocenimo, je ocenjevanje parametrov z WLS primernejše kot transformacija podatkov. Podatki ostanejo v osnovnih enotah, kar omogoča lažjo interpretacijo dobljenega modela.
- V posameznih primerih so uteži lahko določene tudi na podlagi vrednosti izbranih napovednih spremenljivk (ene ali več).
- V primerjavi z OLS ocenami parametrov imajo WLS ocene parametrov v splošnem manjšo varianco.
- Tudi za WLS ocene parametrov modela velja, da so najboljše linearne nepristranske cenilke za $\boldsymbol{\beta}$ (BLUE, *Best Linear Unbiased Estimator*), njihova varianca je najmanjša (Gauss-Markov izrek).

Primer: `andy`