Kazalo

1 METODA NAJVEČJEGA VERJETJA 1

2 POSPLOŠENA METODA NAJMANJŠIH KVADRATOV 4

2.1 Posplošena metoda najmanjših kvadratov, $\mathbf{\Sigma}$ ni znana 5

2.1.1 Modeliranje nekonstantne variance 5

2.1.2 Uporaba funkcij `varFixed` in `varPower` 8

2.1.3 Uporaba funkcije `varIdent` 26

1 METODA NAJVEČJEGA VERJETJA

Linearni model v matrični obliki zapišemo:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

$\mathbf{y}$ je vektor odzivne spremenljivke dimenzije n, $\mathbf{X}$ je modelska matrika dimenzije $n \times p$, $\boldsymbol{\beta}$ je vektor parametrov modela dimenzije $p = k + 1$. V predhodnih poglavjih je za napake tega modela veljalo, da so neodvisno enako porazdeljene, $\boldsymbol{\varepsilon} \sim iid\ N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, kjer je $\mathbf{I}_n$ identična matrika dimenzije $n \times n$.

Ocene parametrov (cenilke) ter njihovo varianco po metodi najmanjših kvadratov (OLS, *Ordinary Least Squares*) izračunamo takole:

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

$$Var(\boldsymbol{b}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.$$

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

Parametre linearnega modela lahko ocenimo tudi po **metodi največjega verjetja** (*maximum likelihood*, ML). V tem primeru moramo, v nasprotju z metodo najmanjših kvadratov, kjer nismo zahtevali nobenih predpostavk za izračun ocen parametrov modela, za napake $\boldsymbol{\varepsilon}$ vnaprej privzeti normalno porazdelitev.

Če za odzivno spremenljivko linearnega modela velja, da je pogojno na napovedne spremenljivke neodvisno enako normalno porazdeljena: $\mathbf{y} \sim iid\ N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, sledi da so napake tudi neodvisno enako normalno porazdeljene $\boldsymbol{\varepsilon} \sim iid\ N(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Za vzorec velikosti n je funkcija verjetja za $\mathbf{y}$, $L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)$ definirana kot produkt gostot verjetnosti normalne porazdelitve $f(y_i, (\mathbf{X})_i; \boldsymbol{\beta}, \sigma^2)$ v n točkah $(i = 1, ..., n)$:

$$
\begin{aligned}
L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} f(y_i, (\mathbf{X})_i; \boldsymbol{\beta}, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2\right) \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).
\end{aligned} \tag{2}
$$

Zanima nas, pri katerih vrednosti parametrov $\boldsymbol{\beta}$ in σ ima funkcija verjetja L pri danih vrednostih y_i, $i = 1, ..., n$ in $\mathbf{X}$ maksimum. Drugače povedano, iščemo vrednosti parametrov pri katerih so dani podatki najbolj verjetni. Metoda največjega verjetja je najbolj pogosto uporabljena metoda za iskanje cenilk parametrov na različnih področjih statistike. Izkaže se, da imajo cenilke po metodi največjega verjetja v primeru velikega n lepe lastnosti, so asimptotsko nepristrane, normalno porazdeljene okoli prave vrednosti, njihovo varianco izrazimo s pomočjo pričakovane vrednosti drugega odvoda logaritma verjetja, cenilke so asimptotsko učinkovite, kar pomeni, da imajo najmanjšo varianco od vseh alternativnih cenilk. V praksi se pokaže, da pri iskanju cenilk z ML lahko naletimo tudi na težave: funkcija verjetja ima lahko več ekstremov (lokalni maksimumi), lahko se pojavi numerična nestabilnost, če je n majhen, se lahko pojavijo odstopanja od zgoraj naštetih lepih lastnostih cenilk.

Verjetje (2) lažje maksimiramo, če ga pred tem logaritmiramo, ker je tako odvajanje lažje. Z logaritmiranjem naredimo monotono preslikavo funkcije L in zato imata funkciji L in $logL$ maksimum v isti točki. Logaritem verjetja ($logL$) označimo z $l(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)$.

$$
l(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = logL(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3}
$$

Če (3) maksimiramo glede na $\boldsymbol{\beta}$, je enako, kot bi glede na $\boldsymbol{\beta}$ minimirali izraz $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, kar smo naredili pri metodi najmanjših kvadratov. Ocene/cenilke parametrov linearnega modela po metodi največjega verjetja so enake ocenam parametrov po metodi najmanjših kvadratov.

$$
\boldsymbol{b}_{ML} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},
$$

Tudi za varianco cenilk po metodi največjega verjetja dobimo enak izraz kot pri metodi najmanjših kvadratov (izračunamo jo na osnovi pričakovane vrednosti drugega odvoda logaritma verjetja)

$$
Var(\boldsymbol{b}_{ML}) = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}.
$$

Po metodi največjega verjetja izračunajmo še oceno za σ^2. Z odvajanjem (3) po σ^2 dobimo:

$$\frac{\partial}{\partial\sigma^2} l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{4}$$

Naj bo $\hat{\sigma}^2_{ML}$ cenilka za σ^2 po metodi največjega verjetja in $\mathbf{b}$ vektor ocen parametrov $\boldsymbol{\beta}$. Ko (4) izenačimo z 0, dobimo:

$$\frac{n}{2\hat{\sigma}^2_{ML}} = \frac{1}{2\hat{\sigma}^4_{ML}}(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}), \tag{5}$$

iz česar sledi, da je

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{6}$$

Matematična statistika pokaže, da je $\hat{\sigma}^2_{ML}$ pristrana cenilka:

$$E(\hat{\sigma}^2_{ML}) = \frac{n-p}{n}\sigma^2. \tag{7}$$

Pristranost je odvisna od števila parametrov v modelu $p = k + 1$ in velikosti vzorca n. Če je vzorec velik v primerjavi s številom ocenjenih parametrov. Nepristrana cenilka za σ^2 je

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{8}$$

Ta rezultat je enak kot v primeru, ko parametre ocenimo po metodi najmanjših kvadratov.

Pogled nazaj: z uporabo ocenjevanja parametrov modela po metodi največjega verjetja smo se prvič srečali pri uporabi Box-Cox transformacij (9), kjer se ustrezna vrednost parametra λ izračuna na podlagi maksimuma funkcije logaritma verjetja (funkcija `powerTransform` iz paketa `car`).

$$z_i = \begin{cases} \frac{y^\lambda - 1}{\lambda} = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda \neq 0 \\ ln(y) = (\mathbf{X}\boldsymbol{\beta})_i + \varepsilon_i & , \lambda = 0 \end{cases} . \tag{9}$$

kjer velja $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I})$.

V kontekstu izbire ustreznega modela na podlagi kakovosti napovedi modela smo definirali Akaikejev informacijski kriterij (AIC), ki ga izračunamo na podlagi logaritma verjetja $AIC = -2log\hat{L} + 2p$, p je število parametrov v modelu in $log\hat{L}$ je logaritem verjetja normalnega modela ovrednoten z ocenami parametrov modela $\mathbf{b}$ in $\hat{\sigma}^2$:

$$log\hat{L} = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{10}$$

Najboljši je model z najmanjšo vrednostjo AIC, ker je tako izgubljene najmanj informacije, ki jo nosijo podatki.

2 POSPLOŠENA METODA NAJMANJŠIH KVADRATOV

Posplošeno metodo najmanjših kvadratov (GLS, *Generalised Least Squares*) uporabljamo za ocene parametrov regresijskega modela v primerih, ko ne moremo predpostaviti konstantne variance napak in/ali neodvisnosti napak. Pri posplošeni metodi najmanjših kvadratov za napake predpostavimo, da so porazdeljene normalno, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, kjer je $\boldsymbol{\Sigma}$ **variančno-kovariančna matrika napak** dimenzije $n \times n$. Za $\boldsymbol{\Sigma}$ velja, da je simetrična in pozitivno definitna, posledično ima $n(n+1)/2$ različnih elementov. Če so vrednosti na diagonali različne, imamo opravka z nekonstantno varianco napak; če so izvendiagonalni členi različni od 0, obstaja kovarianca med napakami.

Kako ocenimo parametre modela po posplošeni metodi najmanjših kvadratov? Zaenkrat predpostavimo, da je $\boldsymbol{\Sigma}$ znana. Minimirati moramo **posplošeno vsoto kvadratov napak**:

$$\sum_{i=1}^{n} \boldsymbol{\Sigma}_{ii}^{-1}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{11}$$

Ponavadi se pri tem uporablja metoda največjega verjetja. Funkcijo verjetja v tem primeru zapišemo:

$$L(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \Sigma) = \frac{1}{(2\pi det\boldsymbol{\Sigma})^{\frac{n}{2}}} exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right), \tag{12}$$

logaritem verjetja je

$$l(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \Sigma) = -\frac{n}{2}log(2\pi) - \frac{n}{2}log(det\boldsymbol{\Sigma}) - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{13}$$

Logaritem verjetja ima maksimum, kadar je posplošena vsota kvadratov napak minimalna. Če ta člen odvajamo po $\boldsymbol{\beta}$ in parcialne odvode enačimo z 0, dobimo ocene parametrov po posplošeni metodi najmanjših kvadratov (GLS):

$$\boldsymbol{b}_{GLS} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{y}. \tag{14}$$

Matematična statistika pokaže, da so GLS ocene parametrov nepristrane, $E(\boldsymbol{b}_{GLS}) = \boldsymbol{\beta}$, njihova varianca je

$$Var(\boldsymbol{b}_{GLS}) = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}. \tag{15}$$

GLS ocenjevanje parametrov regresijskega modela lahko predstavimo še drugače. Za matriko $\mathbf{\Sigma}^{-1}$ lahko vedno najdemo matriko $\mathbf{\Gamma}$ dimenzije $n \times n$, za katero velja $\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma} = \mathbf{\Sigma}^{-1}$ (razcep Choleskega). Potem $\boldsymbol{b}_{GLS}$ lahko izrazimo takole:

$$\boldsymbol{b}_{GLS} = (\mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Gamma}^{\mathrm{T}}\mathbf{\Gamma}\mathbf{y} = (\mathbf{X}^{*\mathrm{T}}\mathbf{X}^{*})^{-1}\mathbf{X}^{*\mathrm{T}}\mathbf{y}^{*}. \tag{16}$$

V (16) je $\mathbf{X}^* = \mathbf{\Gamma}\mathbf{X}$ in $\mathbf{y}^* = \mathbf{\Gamma}\mathbf{y}$. To pomeni, da je ocenjevanje parametrov po GLS metodi enako OLS ocenjevanju parametrov regresijskega modela:

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \tag{17}$$

ki vključuje transformirane spremenljivke $\mathbf{y}^*$ in $\mathbf{X}^*$.

2.1 Posplošena metoda najmanjših kvadratov, $\mathbf{\Sigma}$ ni znana

V dejanskih primerih seveda kovariančna matrika ostankov $\mathbf{\Sigma}$ ni znana in jo moramo skupaj s parametri modela $\boldsymbol{\beta}$ oceniti po metodi maksimalnega verjetja. $\mathbf{\Sigma}$ ima $n(n+1)/2$ različnih elementov, kar je preveč za ocenjevanje na podlagi n podatkov, zato jo parametriziramo s smiselnim številom parametrov.

V praksi $\mathbf{\Sigma}$ zaradi računskih razlogov zapišemo $\mathbf{\Sigma} = \sigma^2\mathbf{\Lambda}$, pri tem pa matriko $\mathbf{\Lambda}$ izrazimo z dvema preprostejšima in vsebinsko smiselnima matrikama $\mathbf{V}$ in $\mathbf{C}$:

$$\mathbf{\Sigma} = \sigma^2\mathbf{\Lambda} = \sigma^2\mathbf{V}\mathbf{C}\mathbf{V}. \tag{18}$$

V enačbi (18) je $\mathbf{V}$ diagonalna matrika, ki opiše varianco napak, njeni členi so pozitivni. Matrika $\mathbf{C}$ je simetrična z enkami na diagonali, ostali elementi opišejo korelacijo med napakami.

Kadar med napakami obstaja nekonstantna varianca, poiščemo ustrezno strukturo matrike $\mathbf{V}$. Če pa se pojavi serialna ali katera druga korelacija (npr. prostorska), poiščemo ustrezno strukturo korelacijske matrike napak $\mathbf{C}$. Seveda lahko v matriki $\mathbf{\Lambda}$ hkrati nastopata tako heteroskedastičnost kot korelacija napak.

V nadaljevanju bomo predstavili nekaj načinov parametrizacije matrik $\mathbf{V}$ in $\mathbf{C}$. Parametre, ki določajo ti dve matriki, zapišemo v vektor $\boldsymbol{\lambda}$. Pri ocenjevanju parametrov $\boldsymbol{\lambda}$ uporabimo iterativni proces ocenjevanja ocen za $\boldsymbol{\beta}$ in za $\boldsymbol{\lambda}$, saj so le te medsebojno odvisne. V tem procesu na vsakem koraku uporabimo metodo največjega verjetja (ML) ali pa metodo omejenega največjega verjetja (REML, *restricted maximum likelihood*).

2.1.1 Modeliranje nekonstantne variance

V tem poglavju bomo predstavili modeliranje variančno-kovariančne matrike napak $\mathbf{\Sigma}$ za primer nekonstantne variance ob pogoju, da so napake nekorelirane. Za tako situacijo velja, da ima v (18) matrika **C** po diagonali enke, vsi izvendiagonalni členi so enaki 0. Modeliramo varianco napak izraženo z diagonalno matriko **V**.

Varianco napak $Var(\varepsilon_i|\mathbf{b})$ v primeru heteroskedastičnosti modeliramo kot produkt σ^2 in kvadrata **variančne funkcije** $g(\mu, \mathbf{v}_i, \boldsymbol{\delta})$:

$$Var(\varepsilon_i|\mathbf{b}) = \sigma^2 g^2(\mu_i, \mathbf{v}_i, \boldsymbol{\delta}), \quad i = 1, ..., n. \tag{19}$$

Variančna funkcija $g(.)$ ima v splošnem tri argumente: $\mu_i = E(y_i)$, $\mathbf{v}_i$ je vektor t. i. **variančnih napovednih spremenljivk** in $\boldsymbol{\delta}$ je vektor **variančnih parametrov**. V praksi variančno funkcijo $g(.)$ lahko določa en, dva ali pa vsi trije argumenti.

Poglejmo nekaj primerov parametrizacije variančne matrike napak **V**, ki jih najdemo v paketu `nlme` (Pinherio in Bates, 2000: Mixed-Effects Models in S and S-PLUS):

- `varFixed`; varianca napak je funkcija ene **variančne napovedne spremenljivke** v, ki je številska:

  $$Var(\varepsilon_i) = \sigma^2 v_i. \tag{20}$$

  Variančna napovedna spremenljivka je lahko ena izmed napovednih spremenljivk ali pa njena transformacija. Tako variančno strukturo uporabimo, če se varianca linearno spreminja z eno od spremenljivk ali s transformacijo ene od spremenljivk (npr. s časom, z geografsko dolžino, ...). Tu ne ocenjujemo parametra variančne funkcije, temveč na osnovi izbrane variančne napovedne spremenljivke na začetku optimizacije določimo uteži za vrednosti y.

  Na primer, če v primeru modeliranja letne količine padavin predpostavimo, da je varianca napak sorazmerna z geografsko dolžino x, jo zapišemo takole

  $$Var(\varepsilon_i) = \sigma^2 x_i, \quad i = 1, ..., n. \tag{21}$$

  V tem primeru je variančna funkcija enaka

  $$g(x_i) = \sqrt{x_i}. \tag{22}$$

  Ob uporabi variančne strukture `varFixed` v linearnem modelu se za oceno parametrov modela izvede metoda tehtanih najmanjših kvadratov (WLS), uteži so $1/\sqrt{x_i}$. Uteži se med optimizacijo ne spreminjajo.

- `varPower`; varianca napak je prav tako funkcija ene variančne napovedne spremenljivke v. V tem primeru ocenjujemo parameter δ, ki določa variančno strukturo:

  $$Var(\varepsilon_i) = \sigma^2 |v_i|^{2\delta}. \tag{23}$$

Variančna funkcija je tu enaka $g(v_i, \delta) = |v_i|^\delta$. Parameter δ se v procesu optimizacije spreminja. Tako obliko variančne funkcije lahko uporabimo, kadar je varianca napak sorazmerna z neko potenco pričakovane vrednosti odzivne spremenljivke, v tem primeru variančno napovedno spremenljivko predstavljajo napovedane vrednosti (`fitted(.)`). Za variančno napovedno spremenljivko lahko izberemo katerokoli napovedno spremenljivko ali njeno transformacijo, paziti moramo le, da ta spremenljivka nima vrednosti 0, ker potem utež variančne funkcije ostane nedefinirana;

- `varIdent`; ena napovedna spremenljivka je opisna in ima S vrednosti, torej so enote razdeljene v S skupin, v s-ti skupini je n_s enot, variance po skupinah so različne:

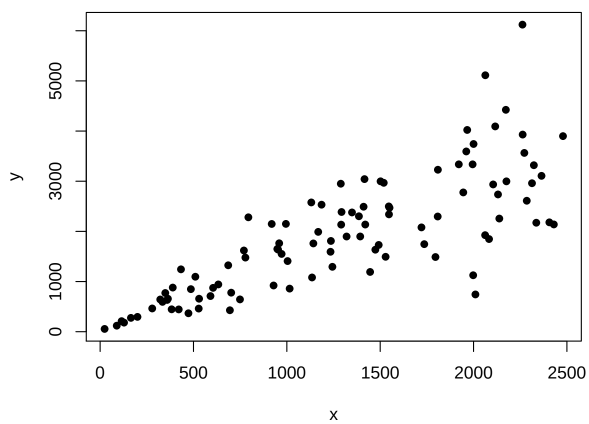
$$Var(\varepsilon_{si}) = \sigma^2 \delta_s^2, \quad s = 1, ..., S, \quad i = 1, ..., n_s. \tag{24}$$

V tem primeru je variančna funkcija $g(s, \boldsymbol{\delta}) = \delta_s$. To pomeni, da moramo za S varianc oceniti $S + 1$ parametrov variančne funkcije: σ^2 in δ_s, $s = 1, ..., S$. Za enolično rešitev moramo postaviti pogoj glede parametrov $\boldsymbol{\delta}$. Za prvo/referenčno skupino določimo, da je $\delta_1 = 1$ in v procesu optimizacije ocenimo ostalih $S - 1$ parametrov δ_s, $s = 2, ..., S$, ki predstavljajo razmerja standardnih odklonov s-te skupine s prvo skupino. Opomba: funkcija omogoča tudi, da ta razmerja določimo vnaprej in se tekom optimizacije ne spreminjajo.

2.1.2 Uporaba funkcij `varFixed` in `varPower`

Vrednosti za x in y generiramo in jih spravimo v podatkovni okvir `primer1`. Vrednosti y generiramo tako, da ima slučajni člen v regresijskem modelu povprečje 0 in standardni odklon sorazmeren z x.

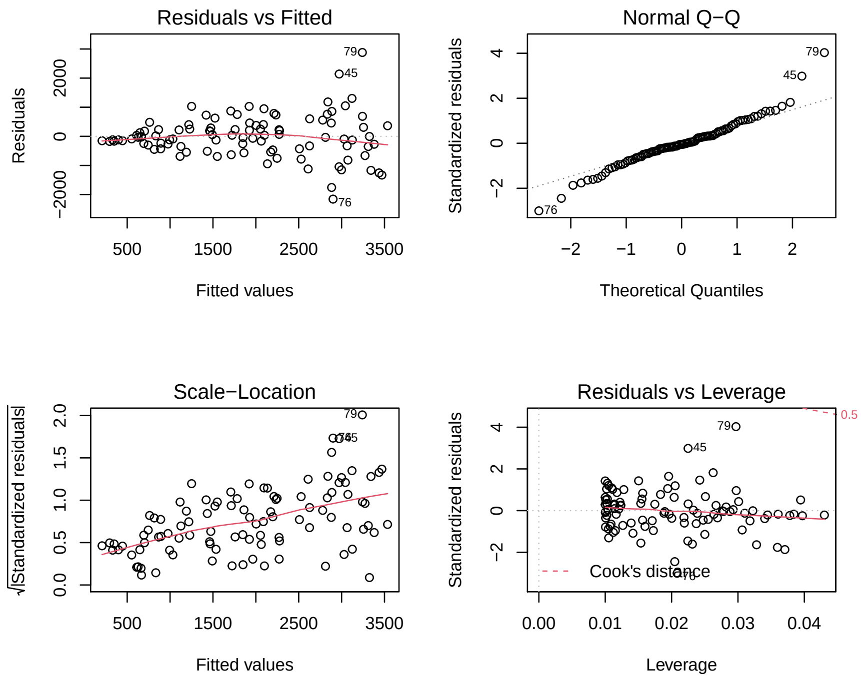
```
> set.seed(777) #zaradi ponovljivosti
> x<-sample(1:2500,100)
> sim<-function(x){10+1.5*x+ rnorm(100,mean=0,sd=0.5*x)}
> y<-sim(x)
> primer1<-data.frame(x,y)
```



Slika 1: Spremenljivka `y` v odvisnosti od `x` za simuliran podatkovni okvir `primer1`

Naredimo `lm` model za y v odvisnosti od x in narišimo ostanke.

```
> mod1.lm<-lm(y~x, data=primer1)
```



Slika 2: Ostanki za `mod1.lm`

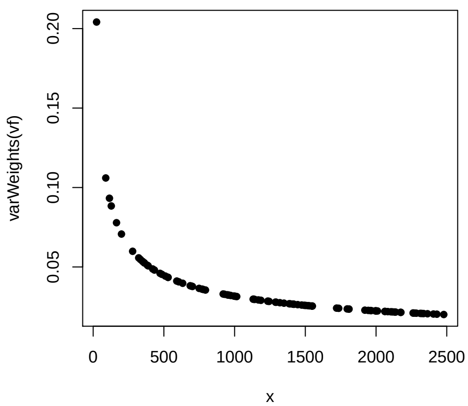
Slika 2 jasno kaže heteroskedastičnost ostankov. Poskusimo jo v modelu upoštevati tako, da uteži za odzivno spremenljivko določimo na osnovi vrednosti spremenljivke `x`. Uporabili bomo variančno strukturo `varFixed` iz paketa `nlme`. Pri modeliranju namesto funkcije `lm` uporabimo funkcijo `gls` iz paketa `nlme`, ki omogoča ocenjevanje parametrov po posplošeni metodi najmanjših kvadratov in s tem tudi uporabo različnih variančno-kovariančnih struktur.

Varianta `varFixed`$(\sim x)$

Predpostavimo, da je varianca napak sorazmerna z `x`. V tem primeru so uteži enake $1/\sqrt{x}$. Za ilustracijo poglejmo inicializacijo za uteži, ki se sicer samodejno izvede na začetku optimizacije pri funkciji `gls`:

```
> library(nlme)
> vf<-varFixed(~x)
> vf<-Initialize(vf, primer1)
> primer1$varW<-varWeights(vf) ### isto kot 1/sqrt(x)
> head(primer1)

     x         y       varW
1   24   55.2185 0.20412415
```

```
2 2138 2255.5988 0.02162699
3  510 1095.8800 0.04428074
4  702  778.1020 0.03774257
5 2131 2736.2776 0.02166249
6  778 1477.5728 0.03585174
```



Slika 3: Uteži variančne funkcije varFixed($\sim x$) v odvisnosti od x

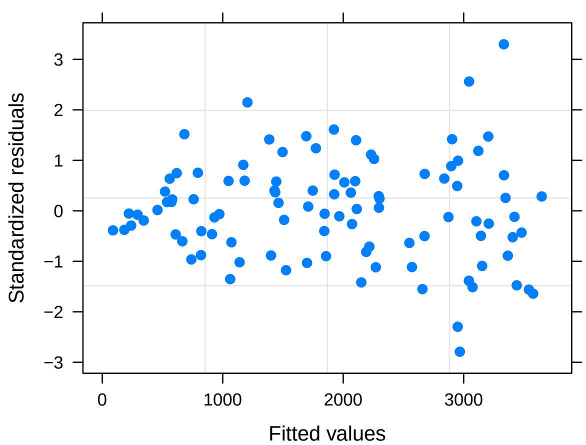
Uteži hitro padajo z x (Slika 3). Ob uporabi funkcije gls z variančno strukturo varFixed($\sim x$) dobimo ocene parametrov linearnega modela po metodi največjega verjetja (method="ML"), ki da pri taki variančni strukturi iste rezultate kot metoda tehtanih najmanjših kvadratov (WLS) z utežmi $1/\sqrt(x)$. V povzetku gls modela vidimo uporabljeno variančno funkcijo (Variance function), poleg ocen parametrov modela se izpiše tudi ocena korelacije med parametroma v modelu (Correlation). Izpisana vrednost za standardno napako regresije za gls model (Residual standard error) ni primerljiva s standardno napako regresije za lm model. Predstavljamo si jo lahko kot standardno napako regresije za model na transformiranih podatkih $\mathbf{y}^*$ in $\mathbf{X}^*$ (16).

```
> mod1.gls1<-gls(y~x, weight=varFixed(~x), data=primer1, method="ML")
> # mod1.lm1<-lm(y~x, weight=1/x, data=primer1)
> summary(mod1.gls1)

Generalized least squares fit by maximum likelihood
  Model: y ~ x
  Data: primer1
       AIC      BIC    logLik
  1557.359 1565.174 -775.6794

Variance function:
```

```
 Structure: fixed weights
 Formula: ~x

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 54.11379  56.70913  0.954234  0.3423
x            1.44954   0.06670 21.731122  0.0000

 Correlation: 
  (Intr)
x -0.661

Standardized residuals:
        Min          Q1         Med          Q3         Max 
-2.79121163 -0.60817450 -0.01939566  0.60630554  3.29795040 

Residual standard error: 17.7801 
Degrees of freedom: 100 total; 98 residual
```

Za `gls` model z ukazom `plot` dobimo samo eno sliko standardiziranih ostankov glede na napovedane vrednosti (Slika 4).

```
> plot(mod1.gls1, pch=16)
```



Slika 4: Ostanki za `mod1.gls1`, variančna funkcija `varFixed`($\sim x$)

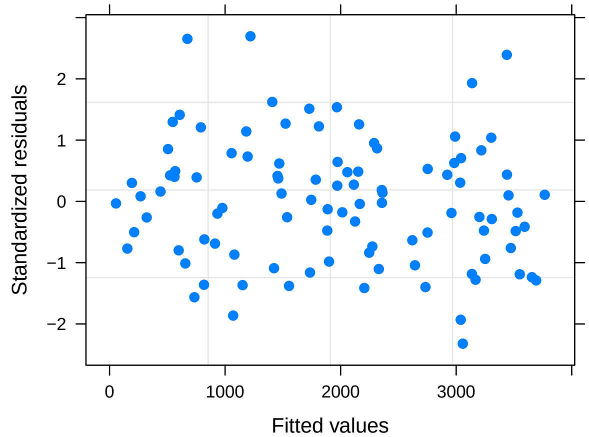
Slika 4 kaže, da z uporabo variančne strukture `varFixed`($\sim x$) heteroskedastičnosti nismo odpravili.

Varianta `varFixed`($\sim x^2$)

Poskusimo z variančno strukturo `varFixed`($\sim x^2$), kar pomeni, da predpostavimo, da je varianca sorazmerna z x^2, oziroma, da uporabimo uteži $1/x$.

```
> mod1.gls2<-gls(y~x, weight=varFixed(~x^2), data=primer1, method="ML")
> # mod1.lm2<-lm(y~x, weight=1/x^2, data=primer1)
> summary(mod1.gls2)

Generalized least squares fit by maximum likelihood
  Model: y ~ x
  Data: primer1
       AIC      BIC   logLik
  1533.448 1541.263 -763.724

Variance function:
 Structure: fixed weights
 Formula: ~x^2

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 19.336304 11.504598  1.680746   0.096
x            1.511457  0.054125 27.925434   0.000

 Correlation:
  (Intr)
x -0.378

Standardized residuals:
          Min            Q1           Med            Q3           Max
-2.3218626474 -0.7630666177  0.0008939755  0.6205430099  2.6949078603

Residual standard error: 0.4959651
Degrees of freedom: 100 total; 98 residual
```

Slika 5 kaže, da smo z uporabo variančne strukture `varFixed`($\sim x^2$) odpravili heteroskedastičnost.



Slika 5: Ostanki za `mod1.gls2`, variančna funkcija `varFixed`($\sim x^2$)

Za primerjavo modela brez variančne strukture `mod1.lm` z modelom `mod1.gls2` uporabimo funkcijo `anova`. V tem primeru se primerjava modelov ne izvede na podlagi F-statistike, kot smo to videli pri primerjavi dveh hierarhičnih `lm` modelov. Izpišejo se vrednosti `AIC`, `BIC` in `-logLik`; če sta modela v hierarhičnem odnosu, se izvede tudi test logaritma razmerja verjetij (*loglikehood ratio test*).

Če modela nista hierarhična, se lahko primerjata na podlagi AIC kriterija (*Akaike information criterion*). AIC kriterij temelji na teoriji informacije in meri relativno izgubo informacije, ko privzamemo, da model opisuje proces, ki generira dane podatke. AIC vrednost za model izračunamo na podlagi maksimalnega verjetja L in števila ocenjenih parametrov p v modelu : $AIC = -2ln(\hat{L}) + 2p$. Manjša je izguba informacije, manjša je vrednost AIC in sprejemljivejši je model.

Modela `mod1.lm` in `mod1.gls2` imata enako število parametrov, razlikujeta se le v tem, da so ocene parametrov pri `mod1.lm` dobljene po OLS, pri `mod1.gls2` pa po GLS, v tem primeru WLS metodi. Ker modela nista v hierarhičnem odnosu, se ne izvede test razmerja verjetij. V funkciji `anova` mora biti `gls` model kot prvi argument, sicer dobimo izpis navadne analize variance `lm` modela, brez primerjave z `gls` modelom.

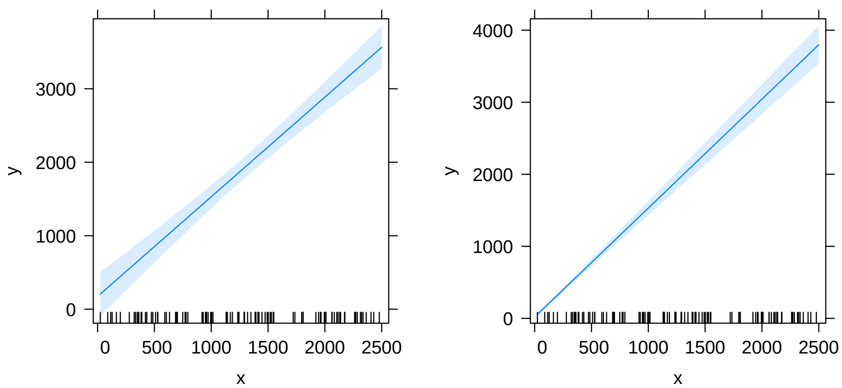
```
> anova(mod1.gls2, mod1.lm)

          Model df      AIC      BIC    logLik
mod1.gls2     1  3 1533.448 1541.263 -763.7240
mod1.lm       2  3 1605.277 1613.093 -799.6386
```

AIC za `mod1.gls2` je manjši kot za `mod1.lm`. Primerjajmo še ocene parametrov in njihove standardne napake ter intervala zaupanje za parametra (za `gls` model dobimo interval zaupanja za parametre modela z ukazom `intervals`).

```
> library(car)
> compareCoefs(mod1.lm, mod1.gls2)

Calls:
1: lm(formula = y ~ x, data = primer1)
2: gls(model = y ~ x, data = primer1, weights = varFixed(~x^2), method =
  "ML")

            Model 1 Model 2
(Intercept)   174.4    19.3
SE            152.8    11.5

x            1.3560  1.5115
SE           0.1045  0.0541

> confint(mod1.lm)

                2.5 %     97.5 %
(Intercept) -128.7672 477.635347
x              1.1487   1.563371

> intervals(mod1.gls2)

Approximate 95% confidence intervals

 Coefficients:
                lower      est.     upper
(Intercept) -3.494196 19.336304 42.166805
x            1.404048  1.511457  1.618865
attr(,"label")
[1] "Coefficients:"

 Residual standard error:
    lower      est.     upper
0.4357262 0.4959651 0.5756851
```

Oceni za odsek na ordinati se relativno na vrednosti spremenljivke `y` malo razlikujeta, večja je razlika njunih standardnih napak in posledično je velika razlika tudi v intervalu zaupanja za presečišče. Oceni za naklon sta primerljivi, standardna napaka pri `mod1.gls2` je dvakrat manjša kot pri `mod1.lm`, kar se pozna na ožjem intervalu zaupanja za naklon za `mod1.gls2`.

Poglejmo še, kako se ocene parametrov poznajo na napovedih modelov `mod1.gls2` in `mod1.lm` ter njihovih 95 % intervalih zaupanja za povprečno napoved (Slika 6). Razlike v napovedih so neznatne, intervali zaupanja za `mod1.gls2` pa so za pri majhnih vrednostih `x` zelo ozki in z vrednostjo `x` naraščajo.



Slika 6: Napovedi za `mod1.lm` (levo) in za `mod1.gls2` (desno)

Varianta `varPower`$(form =\sim x)$

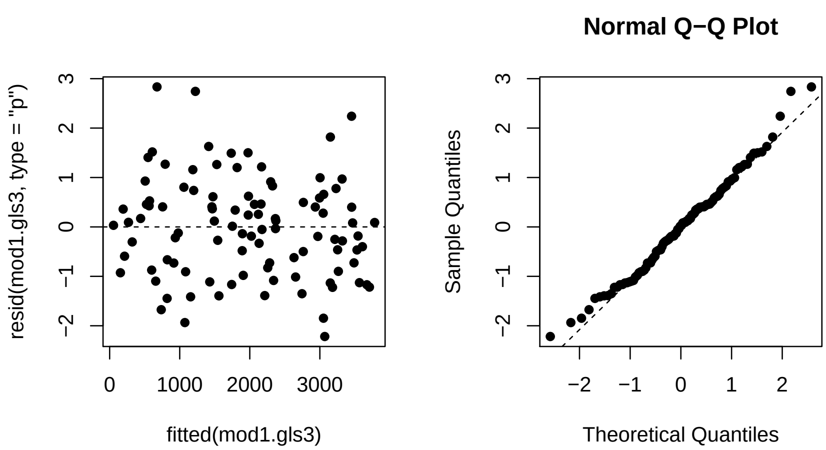
Uporabimo variančno strukturo `varPower`$(form =\sim$ x), kar pomeni, da je varianca sorazmerna $|x|^{2\delta}$. V tem primeru ocenjujemo parameter δ, ki določa diagonalne člene matrike $\mathbf{V}$.

```
> mod1.gls3<-gls(y~x, weight=varPower(form=~x), method="ML")
> summary(mod1.gls3)

Generalized least squares fit by maximum likelihood
  Model: y ~ x
  Data: NULL
       AIC      BIC    logLik
  1534.914 1545.334 -763.4569

Variance function:
 Structure: Power of variance covariate
 Formula: ~x
 Parameter estimates:
```

```
   power
1.078644

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 18.500332  8.800020  2.102306  0.0381
x            1.517991  0.053665 28.286459  0.0000

 Correlation:
  (Intr)
x -0.376

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.21746298 -0.75514060  0.02298281  0.59183995  2.83315697

Residual standard error: 0.2870492
Degrees of freedom: 100 total; 98 residual
```

Ocena za δ je 1.077, kar pomeni, da smo dobili skoraj enake rezultate kot z modelom `mod1.gls2`, saj je $2 \cdot 1.077$ skoraj enako 2.

```
> par(mfrow=c(1,2))
> plot(resid(mod1.gls3, type="p")~fitted(mod1.gls3), pch=16)
> abline(h=0, lty=2)
> qqnorm(resid(mod1.gls3, type="p"), pch=16)
> qqline(resid(mod1.gls3, type="p"), lty=2)
```



Slika 7: Ostanki za `mod1.gls3`, variančna funkcija `varPower`($\sim x$)

Sliki 5 in 7 sta praktično enaki. Isto velja za rezultate primerjave modelov ($p = 0.2835$). Pri primerjavi modela `mod1.gls3` z `mod1.gls2` se izvede test logaritma razmerja verjetij, saj ima `mod1.gls3` en parameter več (δ) in je s tem vzpostavljena hierarhija med modeli.

```
> anova(mod1.gls2, mod1.gls3)

          Model df      AIC      BIC    logLik   Test   L.Ratio p-value
mod1.gls2     1  3 1533.448 1541.263 -763.7240
mod1.gls3     2  4 1534.914 1545.334 -763.4569 1 vs 2 0.5342242  0.4648
```

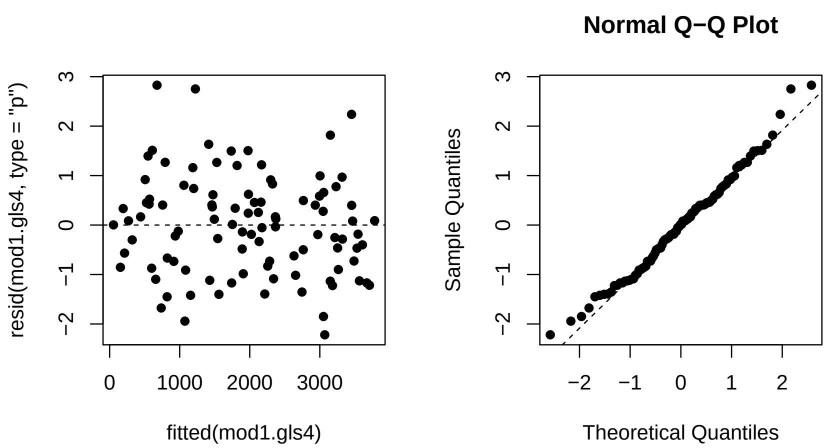
Varianta `varPower`($form =\sim$ `fitted(.)`)

Poskusimo še z uporabo variančne strukture `varPower`($form =\sim$ `fitted(.)`), kar pomeni, da je varianca sorazmerna z absolutno vrednostjo pričakovane vrednosti $E(y)$ na neko potenco. Tudi v tem primeru ocenjujemo parameter δ, ki določa diagonalne člene variančno-kovariančne matrike napak.

```
> mod1.gls4<-gls(y~x, weight=varPower(form=~fitted(.)), method="ML")
> summary(mod1.gls4)

Generalized least squares fit by maximum likelihood
  Model: y ~ x
  Data: NULL
       AIC      BIC    logLik
  1536.026 1546.447 -764.0132

Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
   power
1.093183

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 18.736869 12.185062  1.537692  0.1273
x            1.518353  0.054994 27.609236  0.0000

 Correlation:
  (Intr)
x -0.412

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.21909033 -0.75831994  0.00783506  0.59164129  2.82726641
```

```
Residual standard error: 0.1616636
Degrees of freedom: 100 total; 98 residual
```

Ocena za δ je 1.09, kar kaže, da je varianca napak skoraj sorazmerna z $E(y)^2$. Tudi v modelu `mod1.gls4` je heteroskedastičnost ostankov odpravljena (Slika 8).



Slika 8: Ostanki za `mod1.gls4`, variančna funkcija `varPower(form= fitted(.))`

Sklep: v tem primeru se pokaže, da heteroskedastičnost lahko enakovredno modeliramo na tri načine: `varFixed`($\sim x^2$), `varPower`($\sim x$) ali `varPower(form= fitted(.))`. Spodnji izpis kaže primerjavo ocen parametrov in pripadajočih standardnih napak.

```
> compareCoefs(mod1.lm, mod1.gls2, mod1.gls3, mod1.gls4)

Calls:
1: lm(formula = y ~ x, data = primer1)
2: gls(model = y ~ x, data = primer1, weights = varFixed(~x^2), method =
  "ML")
3: gls(model = y ~ x, weights = varPower(form = ~x), method = "ML")
4: gls(model = y ~ x, weights = varPower(form = ~fitted(.)), method = "ML")

            Model 1 Model 2 Model 3 Model 4
(Intercept)   174.4    19.3    18.5    18.7
SE            152.8    11.5     8.8    12.2

x            1.3560  1.5115  1.5180  1.5184
SE           0.1045  0.0541  0.0537  0.0550
```

Z modeliranjem variančno-kovariančne matrike napak se v primerjavi z `lm` modelom standardne napake ocen parametrov modela zmanjšajo, kar se pozna na intervalih zaupanja za parametre modela. Ocena parametra za naklon se v našem primeru ne spremeni bistveno. Za ilustracijo primerjajmo še intervala zaupanja za `mod1.lm` z intervaloma zaupanja za `mod1.gls4`.

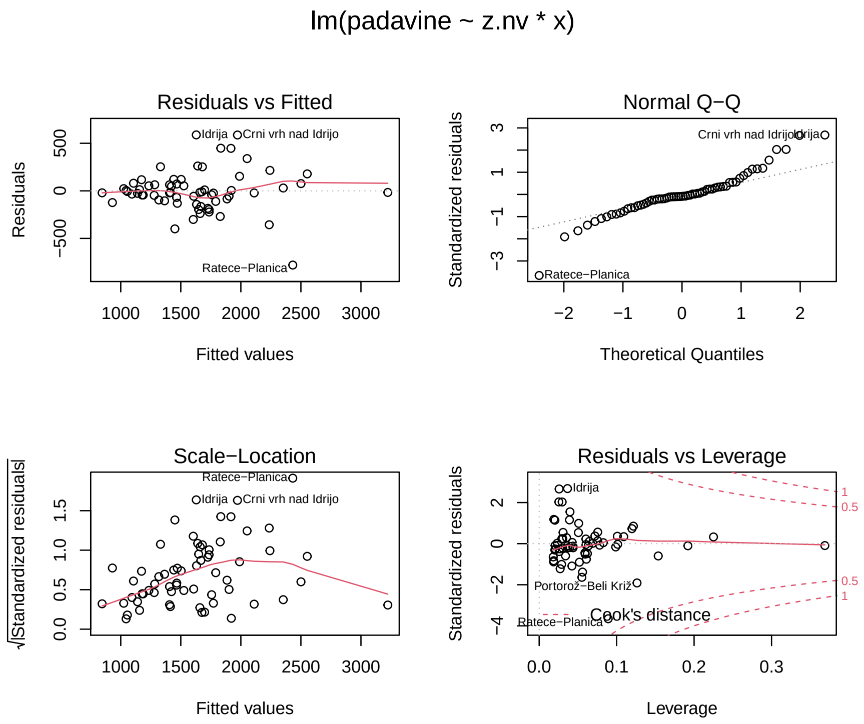
```
> confint(mod1.lm)

                2.5 %     97.5 %
(Intercept) -128.7672 477.635347
x              1.1487   1.563371

> intervals(mod1.gls4)

Approximate 95% confidence intervals

 Coefficients:
                lower      est.     upper
(Intercept) -5.443990 18.736869 42.917729
x            1.409218  1.518353  1.627487
attr(,"label")
[1] "Coefficients:"

 Variance function:
          lower     est.    upper
power 0.8790099 1.093183 1.307356
attr(,"label")
[1] "Variance function:"

 Residual standard error:
     lower       est.      upper
0.03323714 0.16166358 0.78632246
```

Interval zaupanja za presečišče za `mod1.gls4` je bistveno ožji kot za `mod1.lm`, prav tako je ožji interval zaupanja za naklon, vendar razlika tu ni tako velika. 95 % aproksimativni interval zaupanja za parameter δ je (0.9526, 1.2289) in aproksimativni interval zaupanja za standardno napako regresije je (0.0567, 0.4255); ta interval zaupanja ima pomen zgolj v kontekstu preverjanja, ali je numerična integracija v postopku ocenjevanja parametrov stabilna. Če dobimo nesmiselno širok interval zaupanja za katerikoli parameter v modelu, je potrebno popraviti model.

Primer: modeliranje nekonstantne variance POSTAJE

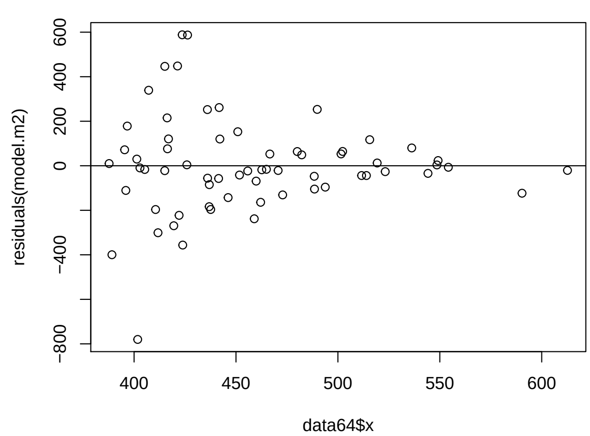
Nadaljujemo analizo primera modeliranja padavin v odvisnosti od geografskih spremenljivk. Najprej povzamemo dobljeni `lm` model, ki je obremenjen z nekonstantno varianco.

```
> data<-read.table("POSTAJE.txt", header=TRUE, sep="\t")
> rownames(data)<-data$Postaja
> data.brez<-subset(data, subset=data$Postaja!="Kredarica")
> data64<-na.omit(data.brez)  ### upoštevajo se samo tisti zapisi, ki so brez NA
> data64$x<-data64$x.gdol/1000
> data64$y<-data64$y.gsir/1000

> model.m2<-lm(padavine~z.nv*x, data=data64)
> summary(model.m2)

Call:
lm(formula = padavine ~ z.nv * x, data = data64)

Residuals:
    Min      1Q  Median      3Q     Max
-780.14  -98.15  -17.35   72.90  588.27

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.736e+03  4.376e+02   3.968 0.000196 ***
z.nv         6.052e+00  1.166e+00   5.192 2.60e-06 ***
x           -1.078e+00  9.541e-01  -1.130 0.262827
z.nv:x      -1.181e-02  2.709e-03  -4.360 5.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 223.6 on 60 degrees of freedom
Multiple R-squared:  0.8001,        Adjusted R-squared:  0.7901
F-statistic: 80.07 on 3 and 60 DF,  p-value: < 2.2e-16
```



Slika 9: Ostanki za `model.m2`

```
> plot(data64$x, residuals(model.m2))
> abline(h=0)
```

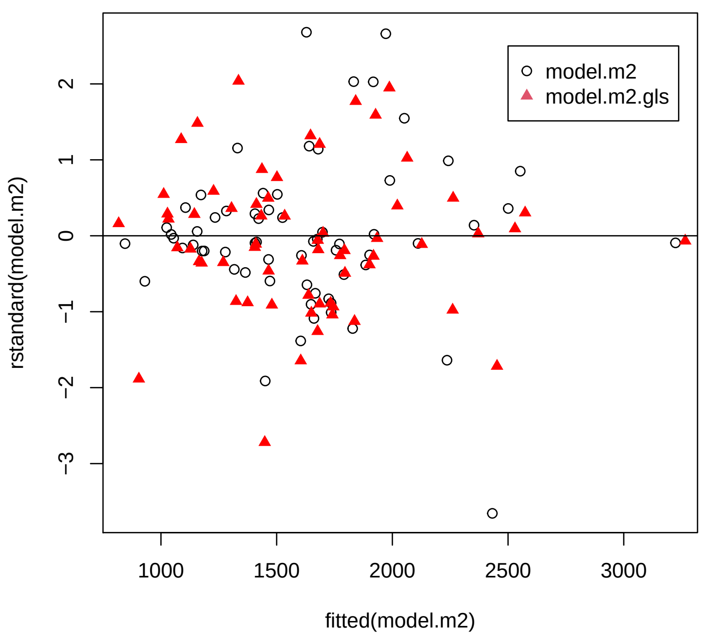


Slika 10: Odvisnost ostankov `model.m2` od geografske dolžine

Sliki 9 in 10 kažeta, da bi za varianco napak lahko predpostavili sorazmernost z geografsko dolžino `x`, ali pa tudi s `fitted(.)`. Ker predpostavljena sorazmernost variance napak s `fitted(.)` zajame hkrati upoštevanje spremenljivk `x`, `z.nv` in njune interakcije v modelu, bomo uporabili variančno strukturo `varPower(form=`$\sim fitted(.)$).

```
> model.m2.gls<-gls(padavine~z.nv*x, weight=varPower(form=~fitted(.)),
+                   method="ML",data=data64)
> anova(model.m2.gls, model.m2)

             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.m2.gls     1  6 854.7796 867.7329 -421.3898
model.m2         2  5 879.9670 890.7614 -434.9835 1 vs 2 27.18739  <.0001
```

```
> # plot(model.m2.gls, pch=16)
> plot(fitted(model.m2),rstandard(model.m2))
> points(fitted(model.m2.gls), residuals(model.m2.gls, type="p"), col="red", pch=17)
> legend(2500, 2.5, legend=c("model.m2","model.m2.gls"),
+        pch=c(1,17), col=(1:2), box.lty = 1)
> abline(h=0)
```



Slika 11: Ostanki za `model.m2.gls` in `model.m2`

Slika 11 kaže, da je heteroskedastičnost v `model.m2.gls` v veliki meri odpravljena.

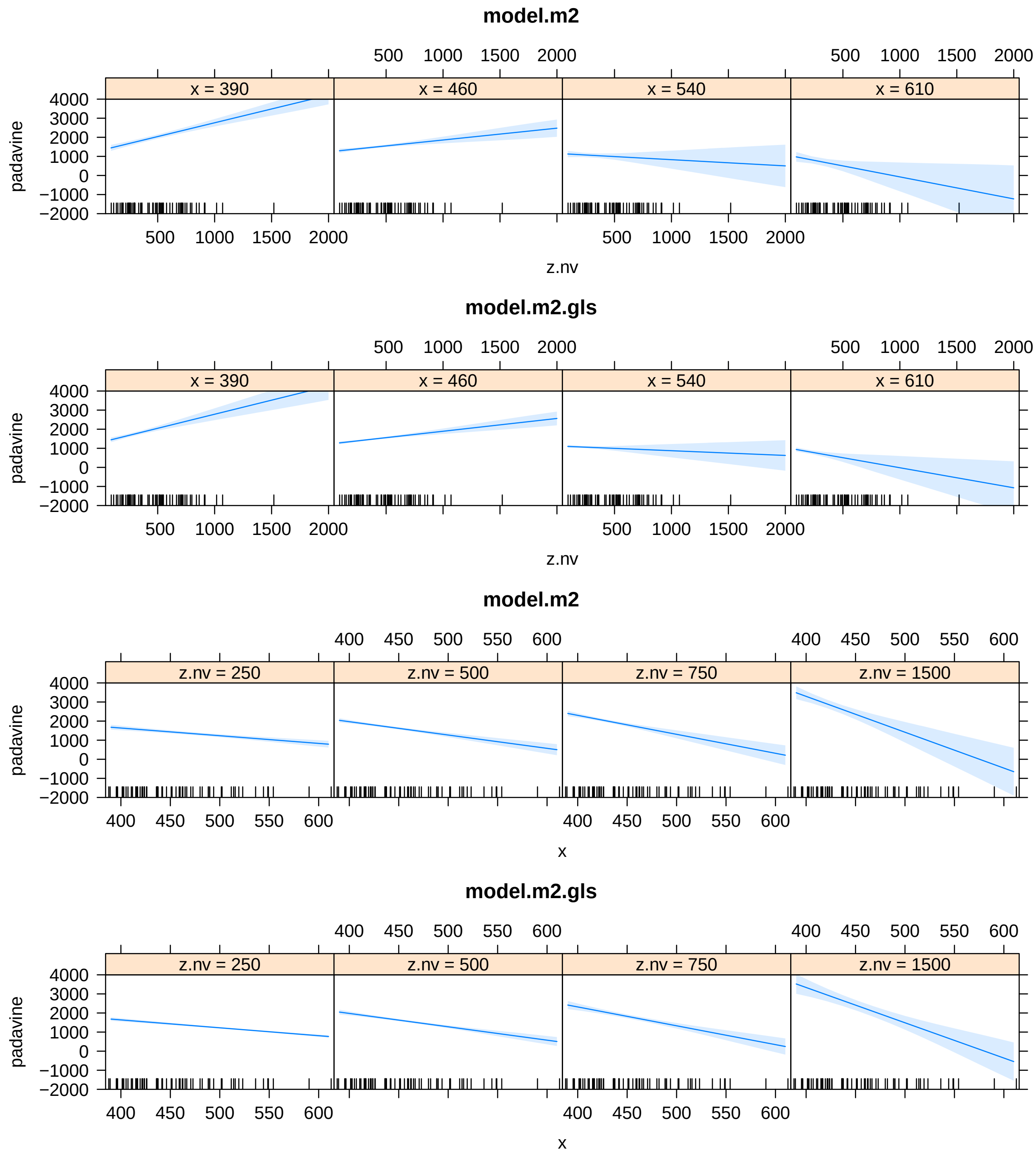
```
> summary(model.m2.gls)

Generalized least squares fit by maximum likelihood
  Model: padavine ~ z.nv * x
  Data: data64
       AIC      BIC    logLik
  854.7796 867.7329 -421.3898

Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
  power
2.20349

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 1806.4456 303.89619  5.944285   0.000
z.nv           5.9477   1.15836  5.134560   0.000
x             -1.2704   0.61139 -2.077920   0.042
z.nv:x        -0.0115   0.00250 -4.593766   0.000

 Correlation:
       (Intr) z.nv   x
z.nv   -0.844
x      -0.989  0.890
z.nv:x  0.812 -0.995 -0.869

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.71966075 -0.56381724 -0.08765019  0.43581852  3.09190945

Residual standard error: 1.584128e-05
Degrees of freedom: 64 total; 60 residual
```

Ocena za δ je 2.203, kar kaže, daje varianca napak sorazmerna s $fitted^{(4.406)}$.

```
> compareCoefs(model.m2,model.m2.gls)

Calls:
1: lm(formula = padavine ~ z.nv * x, data = data64)
2: gls(model = padavine ~ z.nv * x, data = data64, weights = varPower(form =
   ~fitted(.)), method = "ML")

             Model 1  Model 2
(Intercept)     1736     1806
```

```
SE              438      304

z.nv           6.05     5.95
SE             1.17     1.16

x            -1.078   -1.270
SE            0.954    0.611

z.nv:x     -0.01181 -0.01147
SE          0.00271  0.00250

> library(multcomp)
> confint(glht(model.m2))$confint  # glht na gls modelu

                Estimate          lwr           upr
(Intercept) 1736.34007572 747.88540499  2.724795e+03
z.nv           6.05218444   3.41945882  8.684910e+00
x             -1.07842493  -3.23340275  1.076553e+00
z.nv:x        -0.01181133  -0.01793039 -5.692274e-03
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.258737

> confint(glht(model.m2.gls))$confint

                Estimate           lwr           upr
(Intercept) 1806.44564759 1146.01579140  2.466876e+03
z.nv           5.94768690    3.43032116  8.465053e+00
x             -1.27042876   -2.59911659  5.825906e-02
z.nv:x        -0.01147288   -0.01690044 -6.045313e-03
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.173209
```

Primerjava rezultatov obeh modelov pokaže, da se malo spremenijo ocene parametrov modela in pripadajoči intervali zaupanja. To se odraža tudi na napovedih in pripadajočih intervalih zaupanja za povprečno napoved (Slika 12).



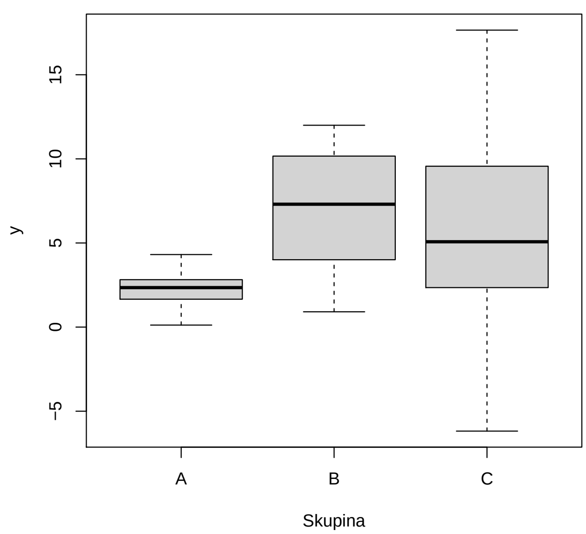
Slika 12: Napovedane vrednosti za `padavine` za `model.m2` in za `model.m2.gls`; v odvisnosti od nadmorske višine pri izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj)

2.1.3 Uporaba funkcije `varIdent`

Imamo različne variance po skupinah A, B in C. Z linearnim modelom želimo napovedati povprečja po skupinah in jih primerjati.

Podatke generiramo v podatkovni okvir `primer2`, $N(\mu_A = 2, \sigma_A^2 = 1)$, $N(\mu_B = 7, \sigma_B^2 = 3^2)$, $N(\mu_C = 6, \sigma_C^2 = 5^2)$. Velikost skupin je 20.

```
> set.seed(777) # zaradi ponovljivosti
> n=20
> ya<-rnorm(n,2,1)
> yb<-rnorm(n,7,3)
> yc<-rnorm(n,6,5)
> y<-c(ya,yb,yc)
> skupina<-rep(c("A","B","C"),each=n)
> primer2<-data.frame(skupina,y)
```



Slika 13: Okvirji z ročaji za tri skupine podatkov

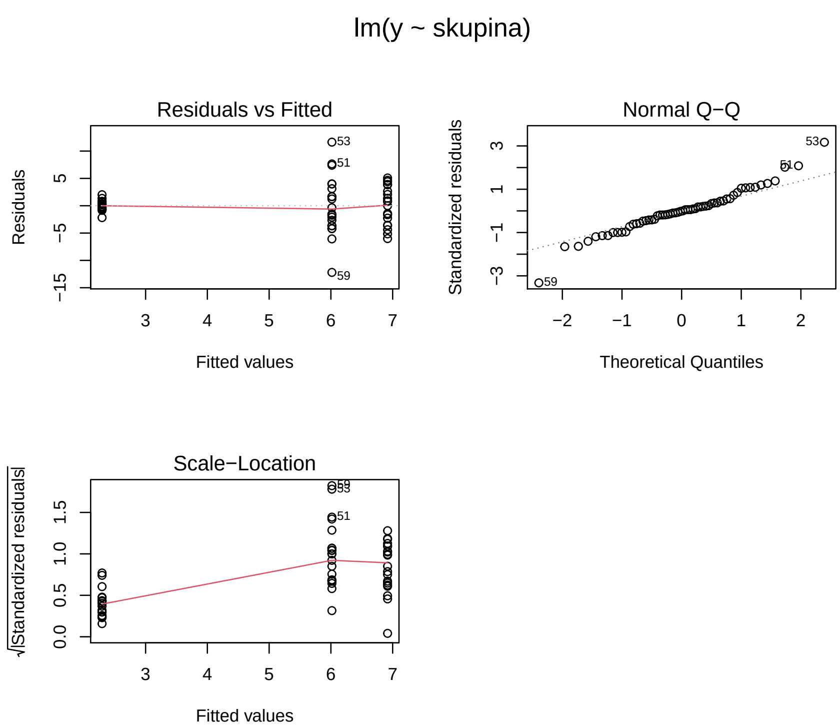
Slika 13 kaže, da je variabilnost podatkov v skupini A veliko manjša od variabilnosti podatkov v skupini C, variabilnost podatkov v skupini B pa je nekje vmes.

Naredimo linearni model za oceno povprečij y po skupinah A, B in C, referenčna skupina je A.

```
> mod2.lm<-lm(y~skupina, data=primer2)
> summary(mod2.lm)
```

```
Call:
lm(formula = y ~ skupina, data = primer2)

Residuals:
     Min       1Q   Median       3Q      Max
-12.1996  -1.8229  -0.0431   1.6563  11.6347

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2951     0.8422   2.725 0.008526 **
skupinaB      4.6227     1.1911   3.881 0.000272 ***
skupinaC      3.7213     1.1911   3.124 0.002803 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.767 on 57 degrees of freedom
Multiple R-squared:  0.229,	Adjusted R-squared:  0.202
F-statistic: 8.465 on 2 and 57 DF,  p-value: 0.0006039
```

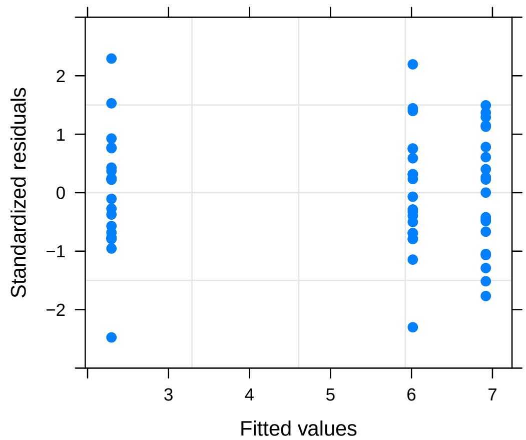
Ocena povprečja v skupini A je 2.2951, v skupini B je 2.2951+4.6227 in v skupini C je 2.2951+3.7213. Poglejmo ostanke za `mod2.lm` (Slika 14).



Slika 14: Porazdelitev ostankov za `mod2.lm`

Slika 14 kaže na prisotnost heteroskedastičnosti. Variabilnost ostankov v skupinah B in C je veliko večja kot v skupini A, zato bomo v modelu uporabili variančno strukturo `varIdent`.

```
> mod2.gls1<-gls(y~skupina, weight=varIdent(form=~1|skupina), method="ML")
> summary(mod2.gls1)

Generalized least squares fit by maximum likelihood
  Model: y ~ skupina
  Data: NULL
       AIC      BIC    logLik
  292.7907 305.3568 -140.3954

Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | skupina
 Parameter estimates:
       A        B        C
1.000000 3.868242 6.029660

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 2.295062 0.2016815 11.379640  0.0000
skupinaB    4.622750 0.8058000  5.736845  0.0000
skupinaC    3.721256 1.2326813  3.018831  0.0038

 Correlation:
         (Intr) skupnB
skupinaB -0.250
skupinaC -0.164  0.041

Standardized residuals:
        Min          Q1         Med          Q3         Max
-2.47522529 -0.69192654 -0.03356865  0.75099325  2.29372268

Residual standard error: 0.8791091
Degrees of freedom: 60 total; 57 residual
```



Slika 15: Porazdelitev ostankov za `mod2.gls1`

Slika 16 ne kaže več nekonstantne variance.

```
> anova(mod2.gls1,mod2.lm)

          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod2.gls1     1  6 292.7907 305.3568 -140.3954
mod2.lm       2  4 334.3371 342.7145 -163.1686 1 vs 2 45.54643  <.0001

> intervals(mod2.gls1) # izračun intervalov zaupanja za parametre gls modela

Approximate 95% confidence intervals

 Coefficients:
               lower     est.    upper
(Intercept) 1.891202 2.295062 2.698923
skupinaB    3.009163 4.622750 6.236336
skupinaC    1.252854 3.721256 6.189658
attr(,"label")
[1] "Coefficients:"

 Variance function:
    lower     est.    upper
B 2.49556 3.868242 5.995969
C 3.88994 6.029660 9.346365
attr(,"label")
[1] "Variance function:"

 Residual standard error:
```

```
    lower      est.     upper
0.6448201 0.8791091 1.1985246
```

```
> compareCoefs(mod2.lm, mod2.gls1)
```

```
Calls:
1: lm(formula = y ~ skupina, data = primer2)
2: gls(model = y ~ skupina, weights = varIdent(form = ~1 | skupina), method
  = "ML")

            Model 1 Model 2
(Intercept)   2.295   2.295
SE            0.842   0.202

skupinaB      4.623   4.623
SE            1.191   0.806

skupinaC       3.72    3.72
SE             1.19    1.23
```

Primerjava modelov `mod2.lm` in `mod2.gls1` pokaže, da je zadnji ustreznejši. Ocene povprečij so enake kot v `mod2.lm`, njihove standardne napake pa se spremenijo. Standardni napaki za A in za `B-A` se zmanjšata, ker na to napako več ne vpliva večja variabilnost v skupini C. Standardna napaka za `C-A` pa se posledično poveča. Ocena za razmerje σ_B/σ_A je 3.87 in za σ_C/σ_A je 6.03. Intervala zaupanja za razmerji vsebujeta vrednosti 3 in 5, ki sta bili uporabljeni v simulaciji.

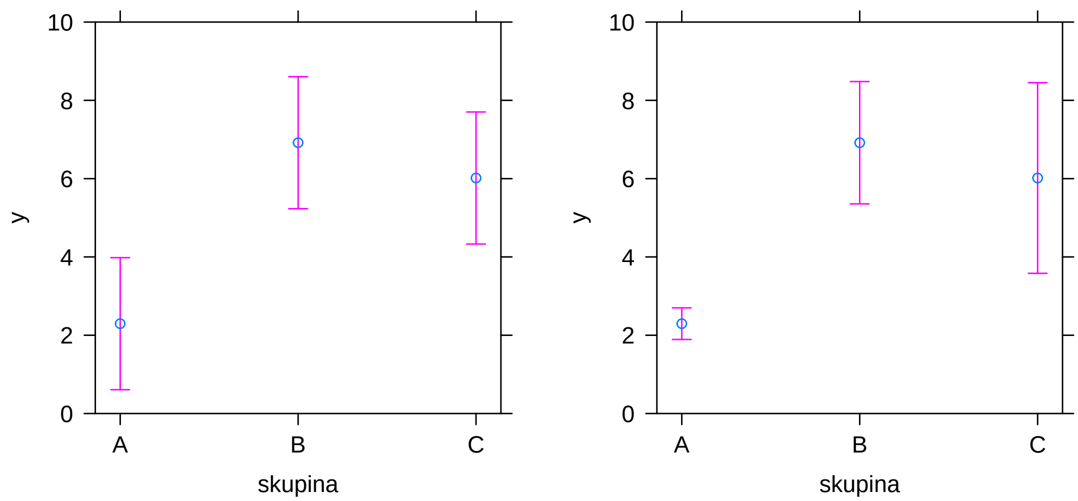
Kot pri `lm` modelu tudi pri `gls` modelu za popravljanje p-vrednosti pri hkratnem testiranju več domnev uporabimo funkcijo `glht` iz paketa `multcomp`. Za ilustracijo izračunajmo intervale zaupanja za razlike povprečij treh skupin za `mod2.gls1` in in jih primerjajmo s tistimi, ki jih dobimo z `mod2.lm`.

```
> library(multcomp)
> C<-rbind(c(0,1,0), c(0,0,1),c(0,-1,1))
> rownames(C)<-c("B-A", "C-A", "C-B")
> test<-glht(mod2.gls1, linfct=C)
> confint(test)$confint
```

```
      Estimate        lwr      upr
B-A  4.6227497  2.7576824 6.487817
C-A  3.7212559  0.8681489 6.574363
C-B -0.9014938 -4.2455762 2.442589
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.314554
```

```
> test.lm<-glht(mod2.lm, linfct=C)
> confint(test.lm)$confint

      Estimate        lwr      upr
B-A  4.6227497  1.7565880 7.488912
C-A  3.7212559  0.8550942 6.587418
C-B -0.9014938 -3.7676556 1.964668
attr(,"conf.level")
[1] 0.95
attr(,"calpha")
[1] 2.40628
```

Interval zaupanaja za razliko `B-A` je v primeru `gls` modela ožji, za razliko `C-A` približno enak, za razliko `C-B` pa širši kot v primeru uporabe `lm` modela.



Slika 16: Napovedana povprečja po skupinah s pripadajočimi 95 % intervali zaupanja