Kazalo

1 OCENJEVANE PARAMETROV PO METODI 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 verjetnostno 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 ($i = 1, ..., n$):

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

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 nepristranske, 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 (10) 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 $l(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)$.

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

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 $\hat{\sigma}^2_{ML}$ pristranska cenilka:

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

Pristranskost je odvisna od števila parametrov v modelu $p = k + 1$. Nepristranska 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 (1), kjer se ustrezna vrednost parametra λ izračuna na podlagi maksimuma funkcije logaritma verjetja (funkcija `powerTransform` iz paketa `car`).

$$T_{BC}(y, \lambda) = y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & , \lambda \neq 0 \\ ln(y) & , \lambda = 0 \end{cases} . \tag{9}$$

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:

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

logaritem verjetja je

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

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 nepristranske, $E(\boldsymbol{b}_{GLS}) = \boldsymbol{\beta}$, 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 $\boldsymbol{\Sigma}^{-1}$ lahko vedno najdemo matriko $\boldsymbol{\Gamma}$ dimenzije $n \times n$, za katero velja $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{-1}$ (razcep Choleskega). Potem $\boldsymbol{b}_{GLS}$ lahko izrazimo takole:

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

V (16) je $\mathbf{X}^* = \boldsymbol{\Gamma}\mathbf{X}$ in $\mathbf{y}^* = \boldsymbol{\Gamma}\mathbf{y}$. 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 Metoda tehtanih najmanjših kvadratov

Če so od 0 različni samo diagonalni elementi matrike $\mathbf{\Sigma}$, potem so parametri modela $\boldsymbol{b}_{GLS}$ v (14) ocenjeni po metodi **tehtanih najmanjših kvadratov** (WLS, *Weighted Least Squares*).

Predpostavimo, da poznamo variance napak ε_i $Var(\varepsilon_i) = \sigma_i^2$, ali, da poznamo variance napak izražene z znanimi utežmi w_i, $Var(\varepsilon_i) = \sigma^2 w_i$, $i = 1, ..., n$, kjer je σ^2 neznana. Uteži za variance napak morajo biti pozitivne, zapišemo jih v diagonalno matriko $\mathbf{W}$. Ocene parametrov modela po metodi tehtanih najmanjših kvadratov dobimo z minimiranjem izraza

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} W_{ii}(y_i - (\mathbf{X}\boldsymbol{\beta})_i)^2. \tag{18}$$

Večja utež W_{ii} pomeni, da ima i-ti podatek večji vpliv na oceno parametrov modela. Če primerjamo (11) in (18), vidimo, da je $\mathbf{W} = \mathbf{\Sigma}^{-1}$, če so izven diagonalni členi $\mathbf{\Sigma}$ enaki 0. 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.

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.

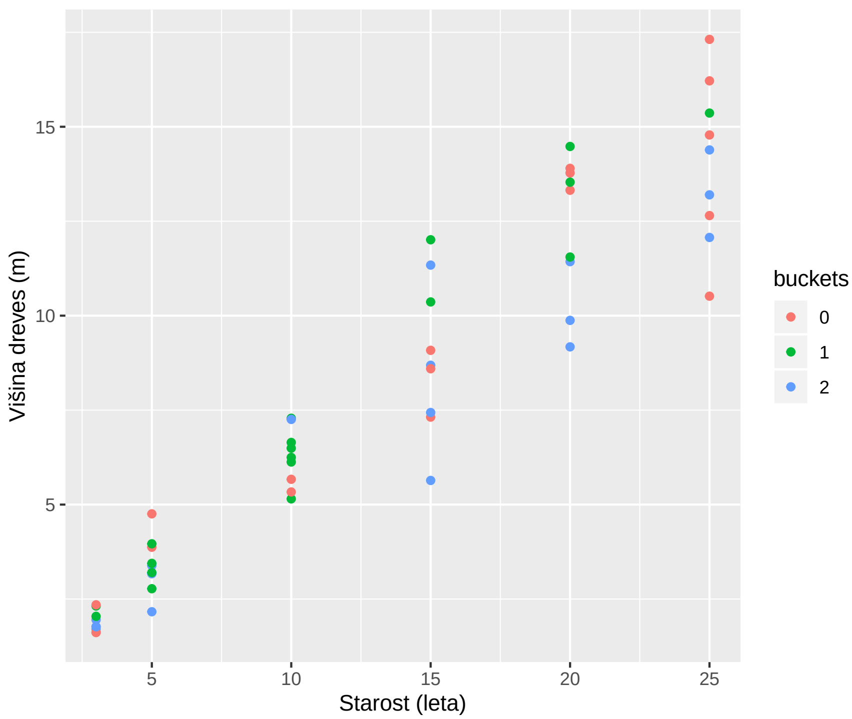
V posameznih primerih so uteži lahko določene tudi na podlagi vrednosti izbranih napovednih spremenljivk (ene ali več). V primerjavi z OLS ocenami parametrov imajo WLS ocene parametrov v splošnem manjšo varianco.

2.1.1 Primer `ANDY`

Drevesničar Andy je želel ugotoviti, kako namakanje vpliva na višino dreves ob upoštevanju njihove starosti. Naredil je večletni poskus, v katerem je drevesa v poletnem času dnevno namakal s tremi različnimi količinami vode (0, 1 in 2 vedra vode). Ob koncu poskusa je za vsako drevo zabeležil starost, višino in količino namakanja. Za vsako starost je imel v poskus vključenih več dreves. Podatki so v podatkovnem okviru ANDY.txt, višina dreves (`height`) je izražena v čevljih (*feets*), starost (`age`) je v letih in namakanje (`buckets`) v številu veder vode.

```
> andy<-read.table("ANDY.txt", header=T)
> str(andy)

'data.frame':        54 obs. of  3 variables:
 $ height : num  5.6 12.7 23.9 28.5 45.6 34.5 5.3 11.1 16.9 18.5 ...
 $ age    : int  3 5 10 15 20 25 3 5 10 15 ...
 $ buckets: int  2 0 1 2 0 0 0 2 1 2 ...

> andy$height<-andy$height/3.2808 # višino dreves izrazimo v metrih
> andy$buckets<-factor(andy$buckets) # buckets naj bo opisna spremenljivka
```

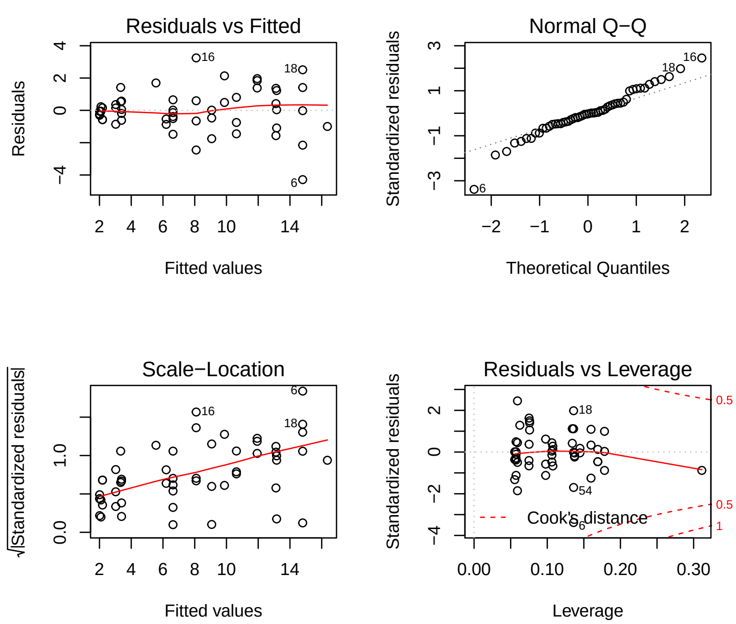
```
> summary(andy)

     height            age         buckets
 Min.   : 1.615   Min.   : 3.0   0:18
 1st Qu.: 3.399   1st Qu.: 5.0   1:18
 Median : 7.270   Median :12.5   2:18
 Mean   : 7.816   Mean   :13.0
 3rd Qu.:11.895   3rd Qu.:20.0
 Max.   :17.313   Max.   :25.0
```



Slika 1: `height` v odvisnosti od `age` in `buckets`

Slika 1 kaže odvisnost `height` od `age` za tri različne količine namakanja (0, 1, 2 vedra vode dnevno). Na sliki se vidi, da se variabilnost podatkov s starostjo dreves povečuje, kar pomeni, da predpostavka o konstantni varianci ni izpolnjena.

```
> mod.OLS<-lm(height ~ age * buckets, data=andy)
> anova(mod.OLS)

Analysis of Variance Table

Response: height
            Df  Sum Sq Mean Sq  F value  Pr(>F)
age          1 1050.60 1050.60 565.1371 < 2e-16 ***
buckets      2   16.51    8.25   4.4395 0.01702 *
age:buckets  2    9.34    4.67   2.5131 0.09163 .
Residuals   48   89.23    1.86
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



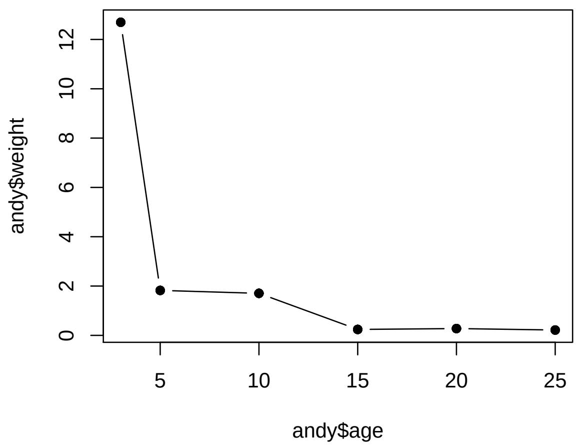
Slika 2: Ostanki za `mod.OLS`

Ker imamo več podatkov pri isti starosti, lahko ocenimo variance za `height` pri posamezni starosti:

```
> var.height <- tapply(andy$height, andy$age, var)
> w <- 1/var.height
> w.df <- data.frame(as.numeric(names(w)), as.numeric(w))
> names(w.df) <- c("age", "weight")
> andy <- merge(andy, w.df, by="age")
```

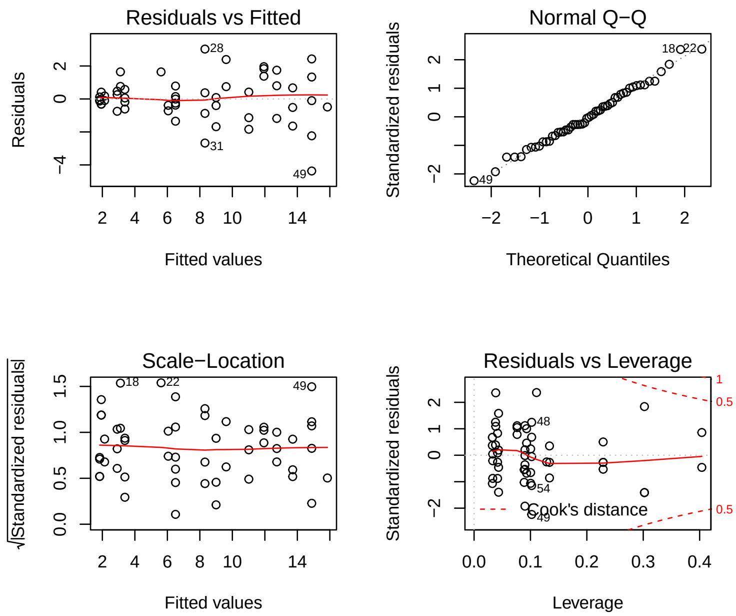
```
> head(andy)
```

```
  age   height buckets   weight
1   3 1.706901       2 12.69631
2   3 2.316508       1 12.69631
3   3 1.615460       0 12.69631
4   3 2.346989       0 12.69631
5   3 1.767861       2 12.69631
6   3 2.042185       1 12.69631
```



Slika 3: Uteži za varianco `height` v odvisnosti od `age`

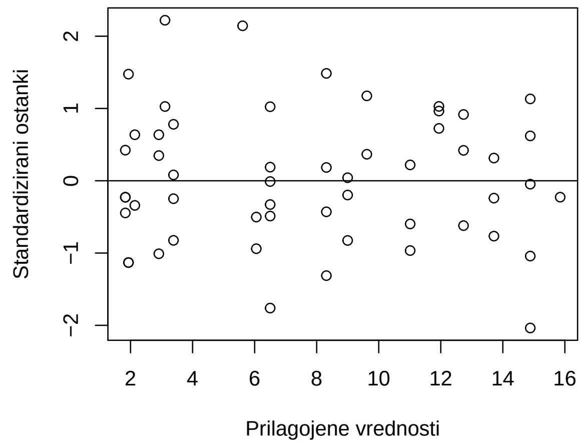
```
> mod.WLS<-lm(height ~ age * buckets, weights = weight, data=andy)
```



Slika 4: Ostanki za `mod.WLS`

Pri modeliranju variance smo malo spremenili ostanke, veliko bolj pa standardizirane ostanke, kar se kaže na Sliki 4 na grafu desno zgoraj in levo spodaj. Na grafu levo zgoraj je še vedno videti nekonstantno varianco v ostankih. Če hočemo grafično prikazati standardizirane ostanke v odvisnosti od prilagojenih vrednosti, moramo to narediti peš (Slika 5).

```
> plot(fitted(mod.WLS), residuals(mod.WLS)/(1/sqrt(andy$weight)),
+       xlab="Prilagojene vrednosti",ylab="Standardizirani ostanki")
> abline(h=0)
```



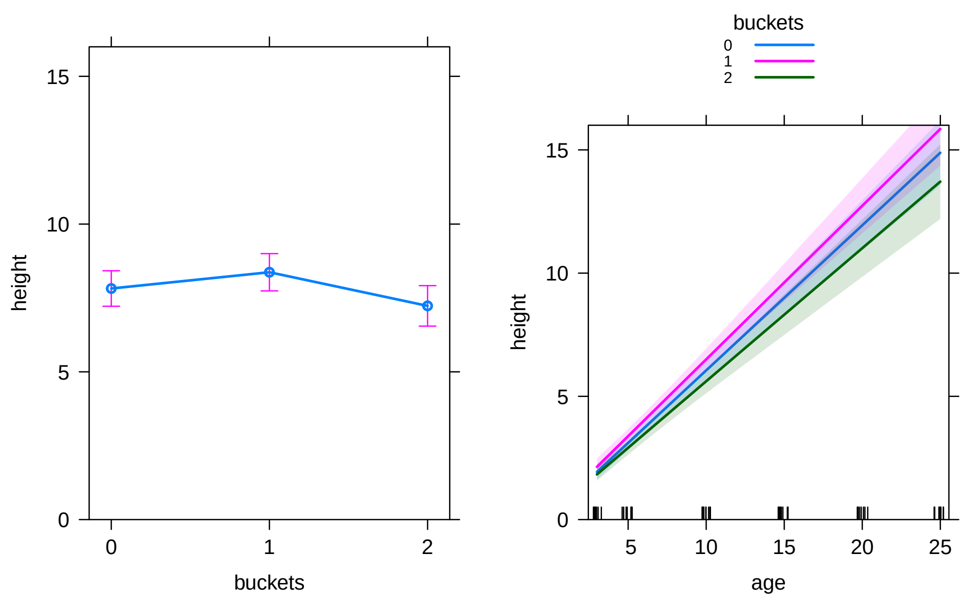
Slika 5: Standardizirani ostanki za `mod.WLS`

```
> anova(mod.WLS)

Analysis of Variance Table

Response: height
            Df Sum Sq Mean Sq  F value Pr(>F)
age          1 829.34  829.34 899.9925 <2e-16 ***
buckets      2   5.12    2.56   2.7792 0.0721 .
age:buckets  2   2.46    1.23   1.3367 0.2723
Residuals   48  44.23    0.92
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

V primerjavi z `mod.OLS` vidimo, da se rezultati sekvenčnega testiranja domnev s funkcijo `anova()` spremenijo. Ko upoštevamo uteži, ki so obratno sorazmerne z varianco `height` pri posamezni vrednosti `age`, količina namakanja `buckets` ob upoštevanju `age` nima več statistično značilnega vpliva na `height`.

```
> summary(mod.WLS)

Call:
lm(formula = height ~ age * buckets, data = andy, weights = weight)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-2.03668 -0.57355 -0.02891  0.63673  2.22073
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.16697    0.20161   0.828    0.412
age            0.58869    0.03158  18.644   <2e-16 ***
buckets1       0.10112    0.32097   0.315    0.754
buckets2       0.04469    0.27642   0.162    0.872
age:buckets1   0.03454    0.04875   0.709    0.482
age:buckets2  -0.04863    0.04747  -1.025    0.311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9599 on 48 degrees of freedom
Multiple R-squared:  0.9498,       Adjusted R-squared:  0.9446
F-statistic: 181.6 on 5 and 48 DF,  p-value: < 2.2e-16
```



Slika 6: Povprečne napovedi `height` glede na `buckets` pri povprečni vrednosti `age` s pripadajočimi 95% intervali zaupanja (levo) in povprečne napovedi `height` glede na `age` pri posamezni vrednosti `buckets` za `mod.WLS`

Analiza pokaže, da je vpliv namakanja na višino dreves ob upoštevanju starosti zanemarljiv. Nakloni premic na Sliki 6 niso statistično značilno različni. Z modelom je pojasnjene 95 % variabilnosti višine dreves.

Vaja: analizo vpliva namakanja na višino dreves ob upoštevanju starosti ponovite z uporabo ustrezne transformacije podatkov.

2.2 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 $\mathbf{C}$:

$$\mathbf{\Sigma} = \sigma^2\mathbf{\Lambda} = \sigma^2\mathbf{VCV}. \tag{19}$$

V enačbi (19) 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 različne načine 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*).

2.2.1 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 (19) matrika $\mathbf{C}$ po diagonali enke, vsi izvendiagonalni členi so enaki 0. Modeliramo varianco napak izraženo z diagonalno matriko $\mathbf{V}$.

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

$$Var(\varepsilon_i|\mathbf{b}) = \sigma^2 g^2(\mu_i, \mathbf{v}_i, \boldsymbol{\delta}), \quad i = 1, ..., n. \tag{20}$$

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 $\mathbf{V}$, 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:
$$Var(\varepsilon_i) = \sigma^2 v_i. \tag{21}$$

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

$$Var(\varepsilon_i) = \sigma^2 x_i, \quad i = 1, ..., n. \tag{22}$$

V tem primeru je variančna funkcija enaka

$$g(x_i) = \sqrt{x_i}. \tag{23}$$

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:

$$Var(\varepsilon_i) = \sigma^2 |v_i|^{2\delta}. \tag{24}$$

  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:

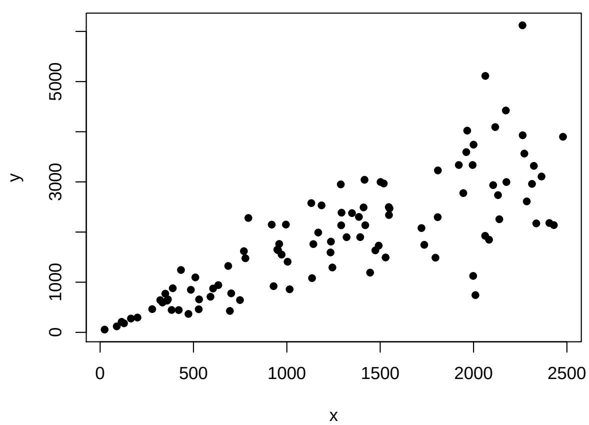
$$Var(\varepsilon_{si}) = \sigma^2 \delta_s^2, \quad s = 1, ..., S, \quad i = 1, ..., n_s. \tag{25}$$

  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.

2.2.2 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.

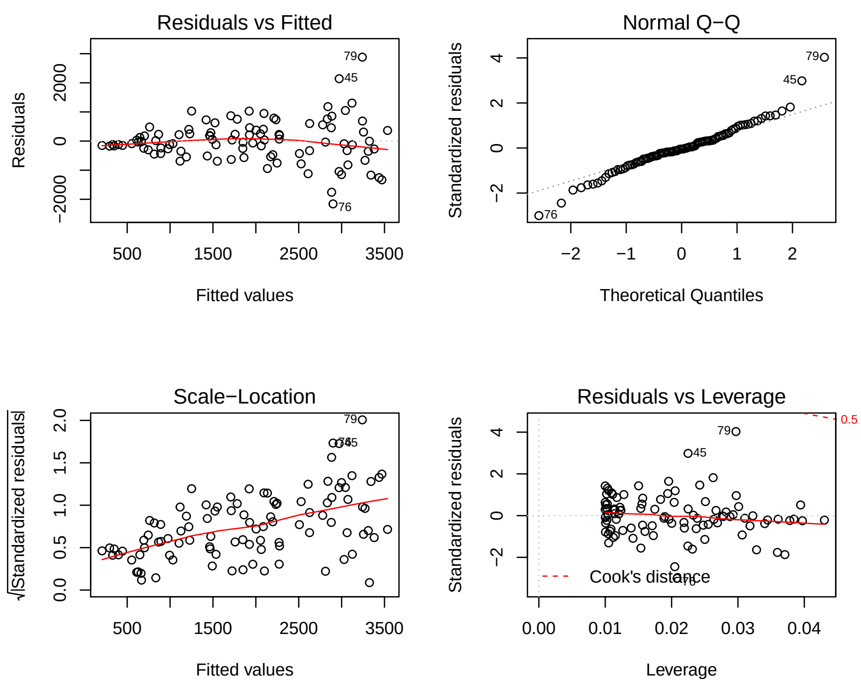
```
> set.seed(777) #zaradi ponovljivosti
> x<-sample(1:2500,100)
> sim<-function(x){10+1.5*x+ rnorm(100,mean=0,sd=0.5*x)}
> y<-sim(x)
> primer1<-data.frame(x,y)
```



Slika 7: 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 8: Ostanki za `mod1.lm`

Slika 8 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.

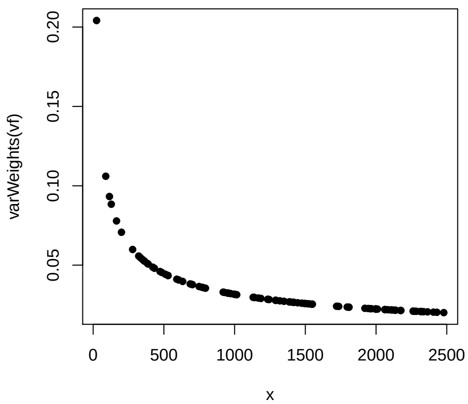
Varianta `varFixed`$(\sim x)$

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 9: Uteži variančne funkcije `varFixed`($\sim x$) v odvisnosti od `x`

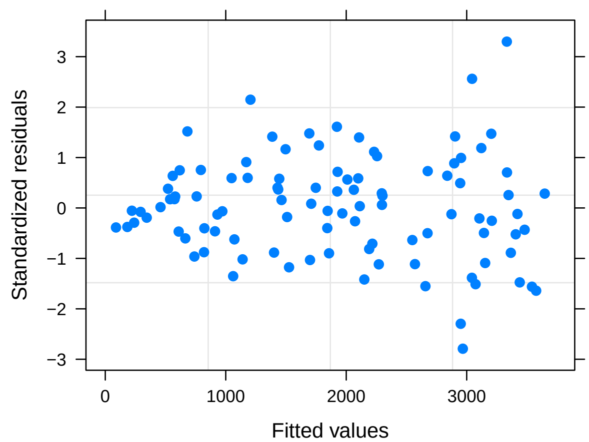
Uteži hitro padajo z x (Slika 9). Ob uporabi funkcije `gls` z variančno strukturo `varFixed`($\sim x$) dobimo ocene parametrov linearnega modela po metodi največjega verjetja (`method="ML"`), ki da pri taki variančni strukturi iste rezultate kot metoda tehtanih najmanjših kvadratov (WLS) z utežmi $1/\sqrt(x)$. 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 10).

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



Slika 10: Ostanki za `mod1.gls1`, variančna funkcija `varFixed`($\sim x$)

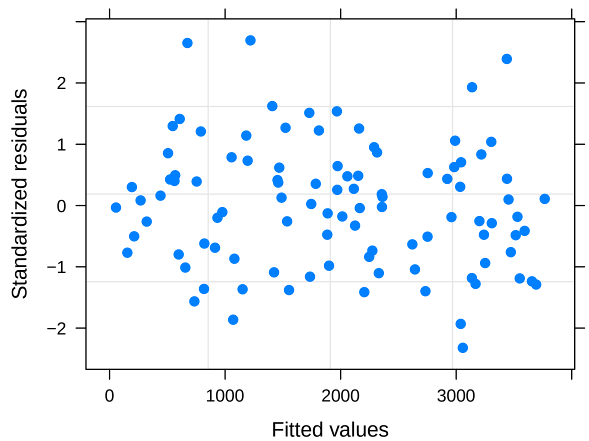
Slika 10 kaže, da z uporabo variančne strukture `varFixed`($\sim x$) heteroskedastičnosti nismo odpravili.

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



Slika 11: 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.

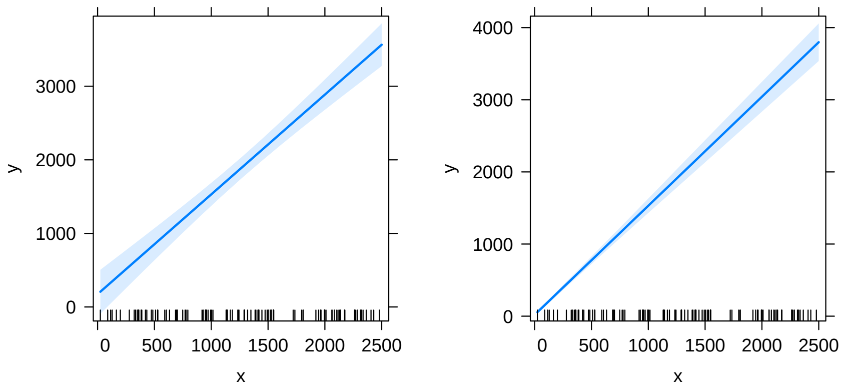
```
> anova(mod1.gls2, mod1.lm)

          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 12). 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 12: Napovedi za `mod1.lm` (levo) in za `mod1.gls2` (desno)

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

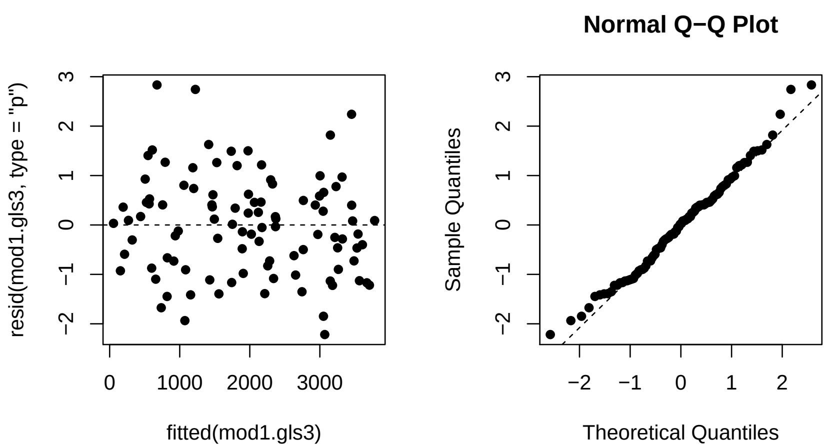
Uporabimo variančno strukturo `varPower`$(form = \sim$ x), kar pomeni, da je varianca sorazmerna $|x|^{2\delta}$. V tem primeru ocenjujemo parameter δ, ki določa diagonalne člene matrike $\mathbf{V}$.

```
> 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 13: Ostanki za `mod1.gls3`, variančna funkcija `varPower`($\sim x$)

Sliki 11 in 13 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.

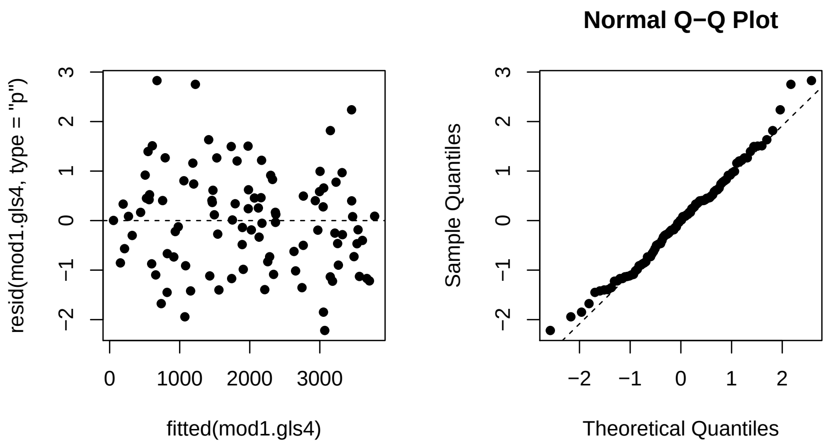
```
> anova(mod1.gls2, mod1.gls3)

          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 14).



Slika 14: 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
```

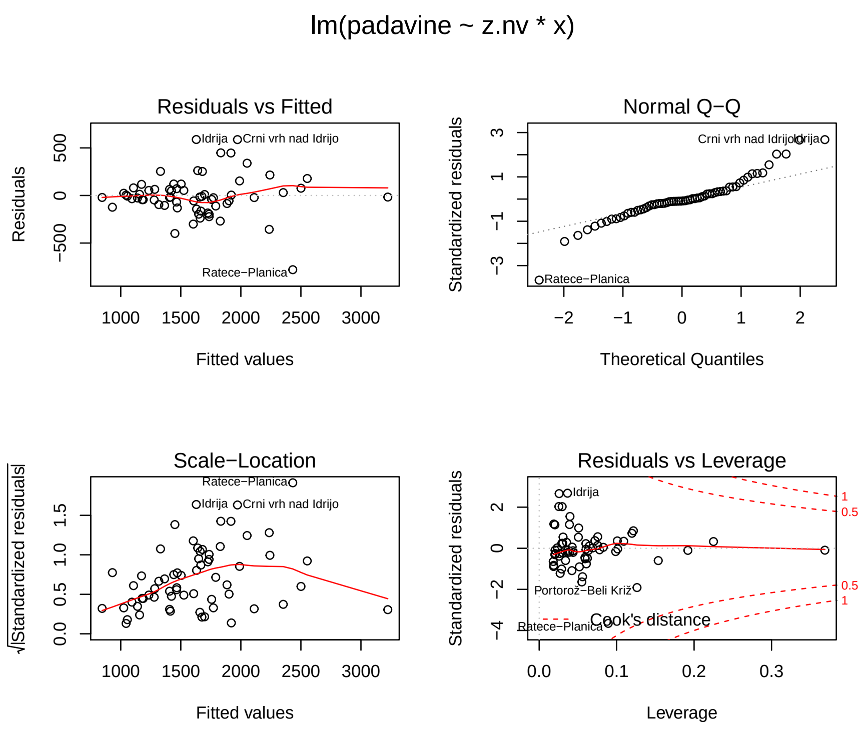
```
> intervals(mod1.gls4)

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

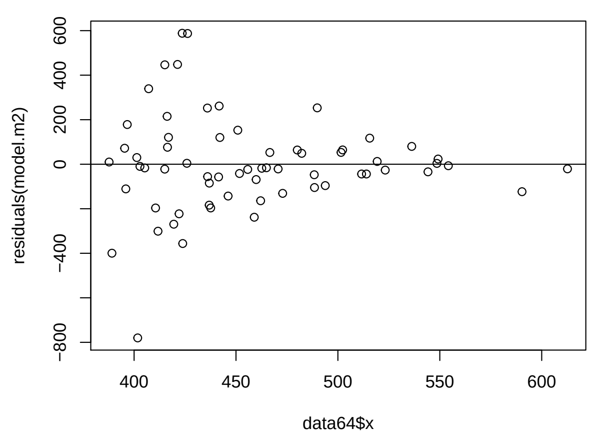
Nadaljujemo analizo primera modeliranja padavin v odvisnosti od geografskih spremenljivk. Najprej povzamemo dobljeni `lm` model, ki je obremenjen z nekonstantno varianco.

```
> 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 15: Ostanki za `model.m2`

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

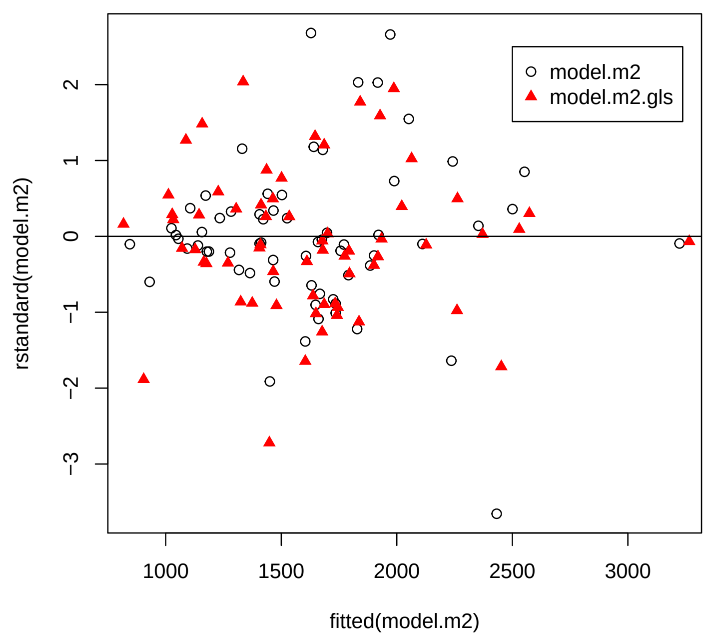


Slika 16: Odvisnost ostankov `model.m2` od geografske dolžine

Sliki 15 in 16 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 17: Ostanki za `model.m2.gls` in `model.m2`

Slika 17 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 $fitted^{(2\cdot 2.203)}$.

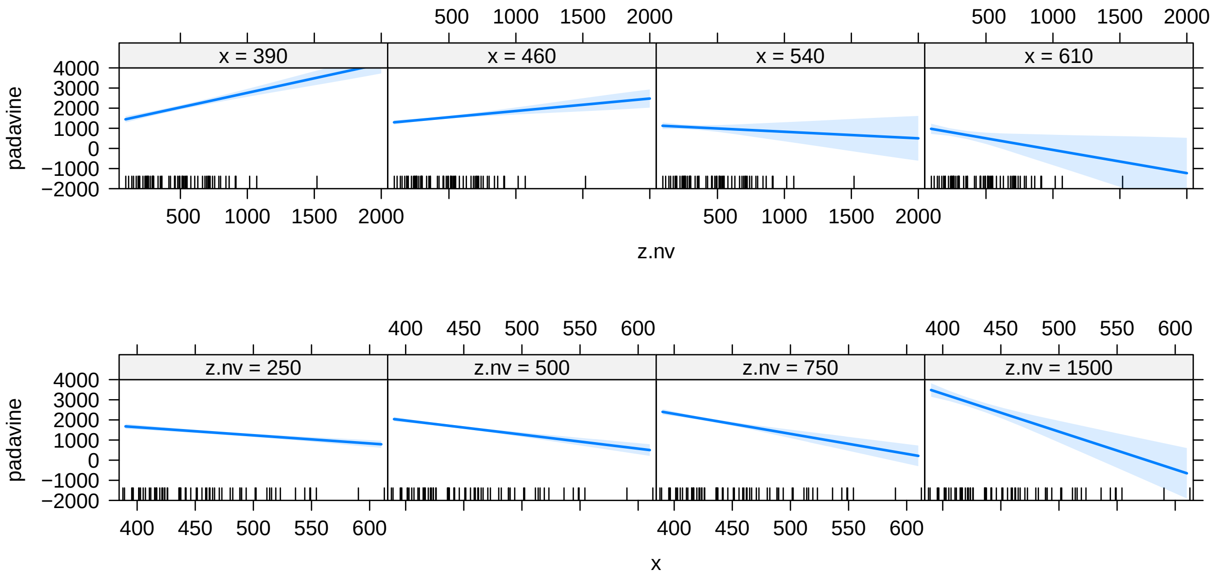
```
> compareCoefs(model.m2,model.m2.gls)

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
```

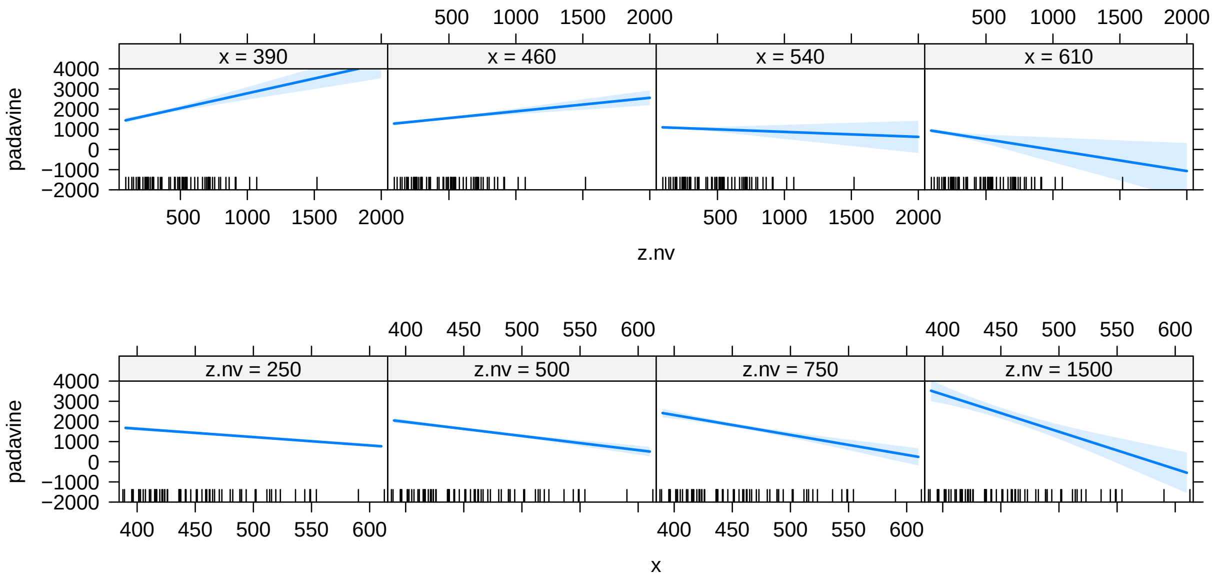
```
> library(multcomp)
> confint(glht(model.m2))  # glht na gls modelu

         Simultaneous Confidence Intervals

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

Quantile = 2.2587
95% family-wise confidence level


Linear Hypotheses:
                 Estimate   lwr        upr
(Intercept) == 0  1.736e+03  7.479e+02  2.725e+03
z.nv == 0         6.052e+00  3.419e+00  8.685e+00
x == 0           -1.078e+00 -3.233e+00  1.077e+00
z.nv:x == 0      -1.181e-02 -1.793e-02 -5.692e-03

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

         Simultaneous Confidence Intervals

Fit: gls(model = padavine ~ z.nv * x, data = data64, weights = varPower(form = ~fitted(.)),
    method = "ML")

Quantile = 2.1732
95% family-wise confidence level


Linear Hypotheses:
                 Estimate   lwr        upr
(Intercept) == 0  1.806e+03  1.146e+03  2.467e+03
z.nv == 0         5.948e+00  3.430e+00  8.465e+00
x == 0           -1.270e+00 -2.599e+00  5.826e-02
z.nv:x == 0      -1.147e-02 -1.690e-02 -6.045e-03
```

Primerjava rezultatov obeh modelov pokaže, da razen intervala zaupanja za presečišče, ki nima vsebinskega pomena, ni bistvenih razlik.

```
> plot2<-plot(Effect(c("x", "z.nv"), model.m2, xlevels=list(z.nv=c(250, 500, 750, 1500))),
+              rows=1, cols=1, main="", layout=c(4,1), ylim=c(-2000, 4000))
> plot1<-plot(Effect(c("z.nv","x"), model.m2, xlevels=list(x=c(390,460,540,610))),
+              rows=1, cols=1, main="", layout=c(4,1), ylim=c(-2000, 4000))
> grid.arrange(plot1, plot2, nrow=2)
```



Slika 18: Napovedane vrednosti za `padavine` za `model.m2`; v odvisnosti od nadmorske višine pri izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj)



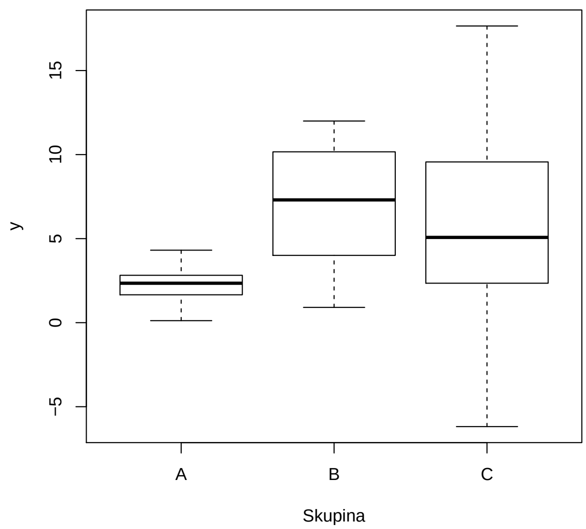
Slika 19: Napovedane vrednosti za `padavine` za `model.m2.gls`; v odvisnosti od nadmorske višine pri izbranih vrednostih geografske dolžine (zgoraj) in v odvisnosti od geografske dolžine pri izbranih vrednostih nadmorske višine (spodaj)

2.2.3 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)
```



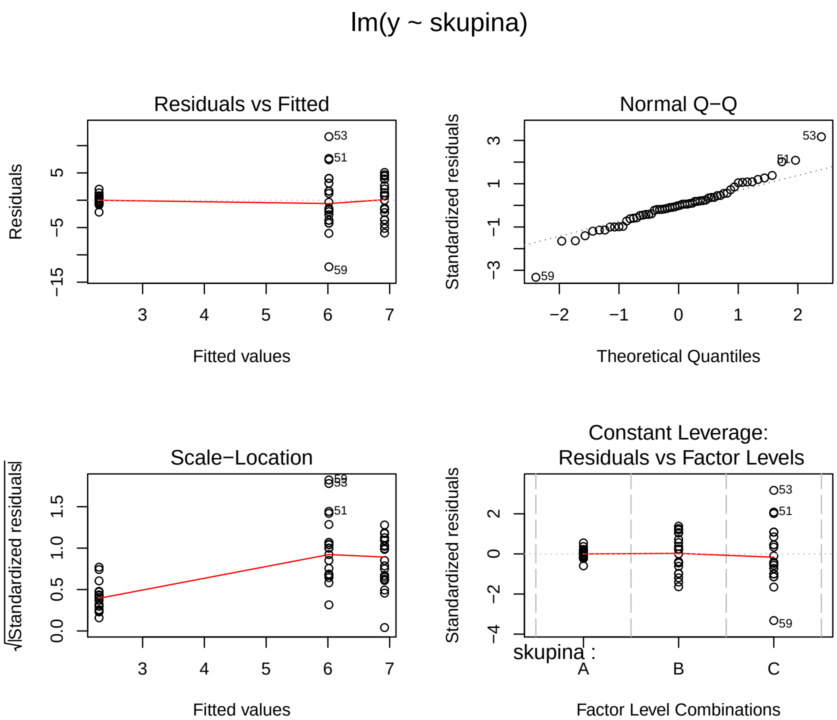
Slika 20: Okvirji z ročaji za tri skupine podatkov

Slika 20 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
```

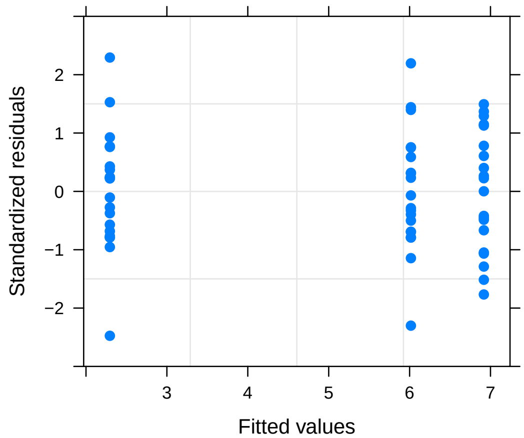
Ocena povprečja v skupini A je 2.2951, v skupini B je 2.2951+4.6227 in v skupini C je 2.2951+3.7213. Poglejmo ostanke za `mod2.lm` (Slika 21).



Slika 21: Porazdelitev ostankov za `mod2.lm`

Slika 21 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 22: Porazdelitev ostankov za `mod2.gls1`

Slika 23 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
```

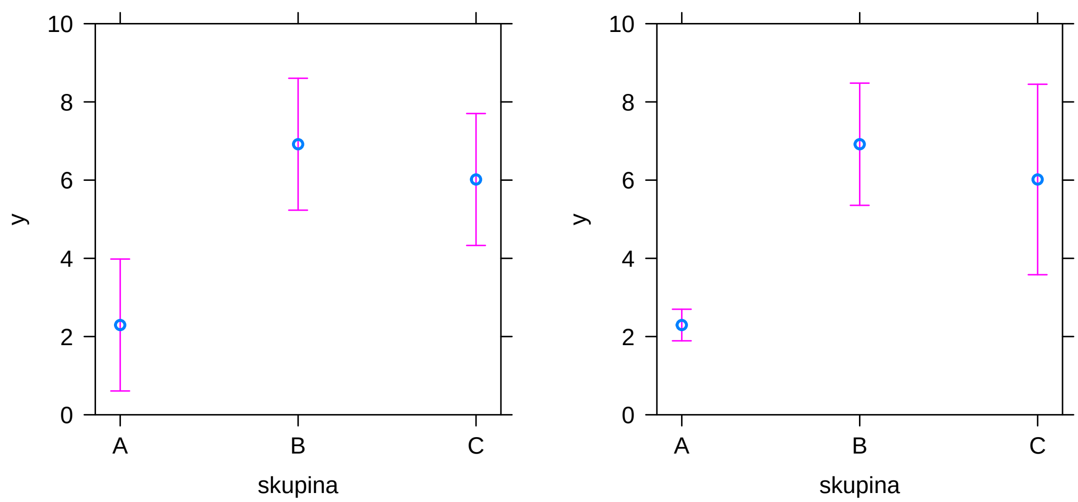
Primerjava modelov `mod2.lm` in `mod2.gls1` pokaže, da je zadnji ustreznejši. Ocene povprečij so enake kot v `mod2.lm`, njihove standardne napake pa se spremenijo. Standardni napaki za A in za `B-A` se zmanjšata, ker na to napako več ne vpliva večja variabilnost v skupini C. Standardna napaka za `C-A` 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)

         Simultaneous Confidence Intervals

Fit: gls(model = y ~ skupina, weights = varIdent(form = ~1 | skupina),
    method = "ML")

Quantile = 2.3146
95% family-wise confidence level


Linear Hypotheses:
         Estimate lwr     upr
B-A == 0  4.6227   2.7577  6.4878
C-A == 0  3.7213   0.8681  6.5744
C-B == 0 -0.9015  -4.2456  2.4426
```

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

         Simultaneous Confidence Intervals

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

Quantile = 2.4063
95% family-wise confidence level


Linear Hypotheses:
         Estimate lwr     upr
B-A == 0  4.6227   1.7566  7.4889
C-A == 0  3.7213   0.8551  6.5874
C-B == 0 -0.9015  -3.7677  1.9647
```

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 23: Napovedana povprečja po skupinah s pripadajočimi 95 % intervali zaupanja

2.3 Modeliranje korelacije napak

V tem poglavju se bomo ukvarjali z modeliranjem odzivne spremenljivke $\boldsymbol{y}$, ki predstavlja **ekvidistantno časovno vrsto**, kar pomeni, da so meritve y_t, $t = 1, ..., n$, narejene v enakih časovnih razmikih. V takih primerih običajno ne velja $\varepsilon_t \sim iid\ N(0, \sigma^2)$, temveč so napake linearnega modela medsebojno korelirane. Obstoja korelacija med napako v času t, ε_t, in napako v času $t+s$, ε_{t+s}, imenujemo jo **avtokorelacija** z odlogom s ali **serialna korelacija z odlogom** s.

Predpostavili bomo, da je časovna vrsta napak regresijskega modela **stacionarna**, kar pomeni, da imajo napake konstantno od časa neodvisno pričakovano vrednost in varianco σ^2, kovarianca dveh napak je odvisna samo od časovnega zamika s med napakama:

$$Cov(\varepsilon_t, \varepsilon_{t+s}) = E(\varepsilon_t \varepsilon_{t+s}) = \sigma^2 \rho_s = Cov(\varepsilon_t, \varepsilon_{t-s}), \quad s = 1, ..., n-1. \tag{26}$$

V (26) je ρ_s **koeficient avtokorelacije napak z odlogom** s. V tem primeru ima variančno-kovariančna matrika napak $\boldsymbol{\Sigma}$ naslednjo obliko:

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{C} = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}. \tag{27}$$

V enačbi (27) je **C korelacijska matrika napak**. Ocenjujemo torej σ^2 in $n-1$ koeficientov avtokorelacije ρ_s, $s = 1, ..., n-1$. Brez dodatnih omejitev je to na podlagi n podatkov nemogoče, zato je potrebno podati še dodatne pogoje za strukturo avtokorelacije napak.

2.3.1 Avtoregresijski model

Za stacionarne časovne vrste je osnovni model za korelacijsko matriko napak t. i. **avtoregresijski model prvega reda AR(1)**. V AR(1) velja, da je napaka v času t, ε_t, odvisna od napake v času $t-1$, ε_{t-1}, in slučajnega vpliva v času t, ki ga opiše slučajna spremenljivka w_t:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + w_t. \tag{28}$$

Za člene w_t velja, da so neodvisno enako normalno porazdeljeni $w_t \sim iid\ N(0, \sigma_w^2)$. V teoriji časovnih vrst tako slučajno spremenljivko imenujemo **beli šum** (*white noise*). Če je časovna vrsta ε_t stacionarna, mora v (28) veljati $|\phi_1| < 1$, sicer bi napake s časom neomejeno naraščale.

Pod pogojem stacionarnosti časovne vrste napak in ob upoštevanju, da je pričakovana vrednost napak 0, je varianca napak konstantna in velja:

$$\sigma^2 = Var(\varepsilon_t) = E(\varepsilon_t^2) = Var(\varepsilon_{t-1}) = E(\varepsilon_{t-1}^2).$$

Če (28) kvadriramo in pogledamo pričakovane vrednosti izrazov v enačbi, pridemo do izraza za varianco avtokoreliranih napak σ^2:

$$\begin{aligned} E(\varepsilon_t^2) &= \phi_1^2 E(\varepsilon_{t-1}^2) + E(w_t^2) + 2\phi_1 E(\varepsilon_{t-1} w_t) \\ \sigma^2 &= \phi_1^2 \sigma^2 + \sigma_w^2 + 0, \end{aligned}$$

$$\sigma^2 = \frac{\sigma_w^2}{1-\phi_1^2}. \tag{29}$$

Podobno pridemo do izraza za avtokovarianco z odlogom s. Najprej zapišimo avtokovarianco za $s = 1$:

$$Cov(\varepsilon_t, \varepsilon_{t-1}) = E(\varepsilon_t \varepsilon_{t-1}) = E((\phi_1 \varepsilon_{t-1} + w_t)\varepsilon_{t-1}) = \phi_1 \sigma^2. \tag{30}$$

Koeficient avtokorelacije z odlogom 1 je:

$$\rho_1 = \frac{Cov(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{Var(\varepsilon_t) Var(\varepsilon_{t-1})}} = \frac{\phi_1 \sigma^2}{\sigma^2} = \phi_1. \tag{31}$$

Podobno je avtokovarianca pri odlogu 2 enaka

$$Cov(\varepsilon_t, \varepsilon_{t-2}) = E(\varepsilon_t \varepsilon_{t-2}) = E([\phi_1(\phi_1 \varepsilon_{t-2} + w_{t-1}) + w_t]\varepsilon_{t-2}) = \phi_1^2 \sigma^2 \tag{32}$$

in koeficient avtokorelacije je

$$\rho_2 = \phi_1^2. \tag{33}$$

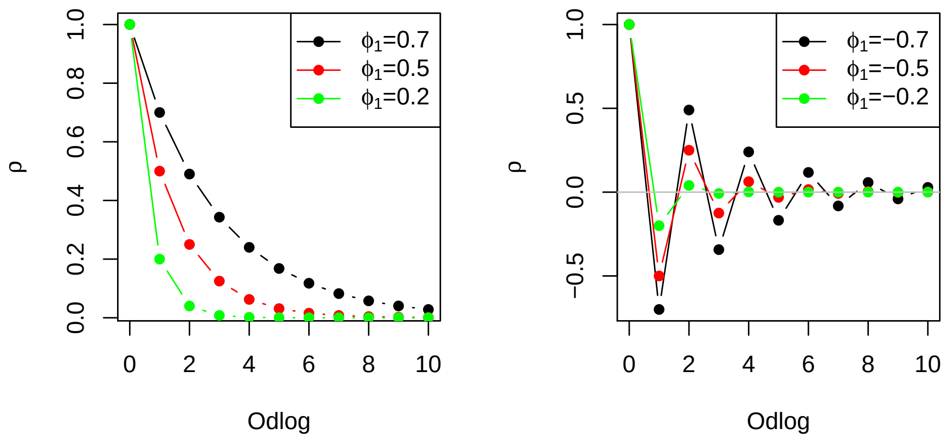
Splošno avtokorelacijsko funkcijo za model AR(1) zapišemo

$$\rho_s = \phi_1^s, \quad s = 1, ..., n-1. \tag{34}$$

Ker mora biti pod pogojem stacionarnosti časovne vrste $|\phi_1| < 1$, avtokorelacijska funkcija ρ_s z večanjem odloga pada eksponentno proti 0. V korelacijski matriki napak $\mathbf{C}$ za model AR(1) ocenjujemo samo en parameter, to je ϕ_1, skupno v variančno-kovariančni matriki napak ocenjujemo dva parametra, poleg ϕ_1 še σ_w^2.

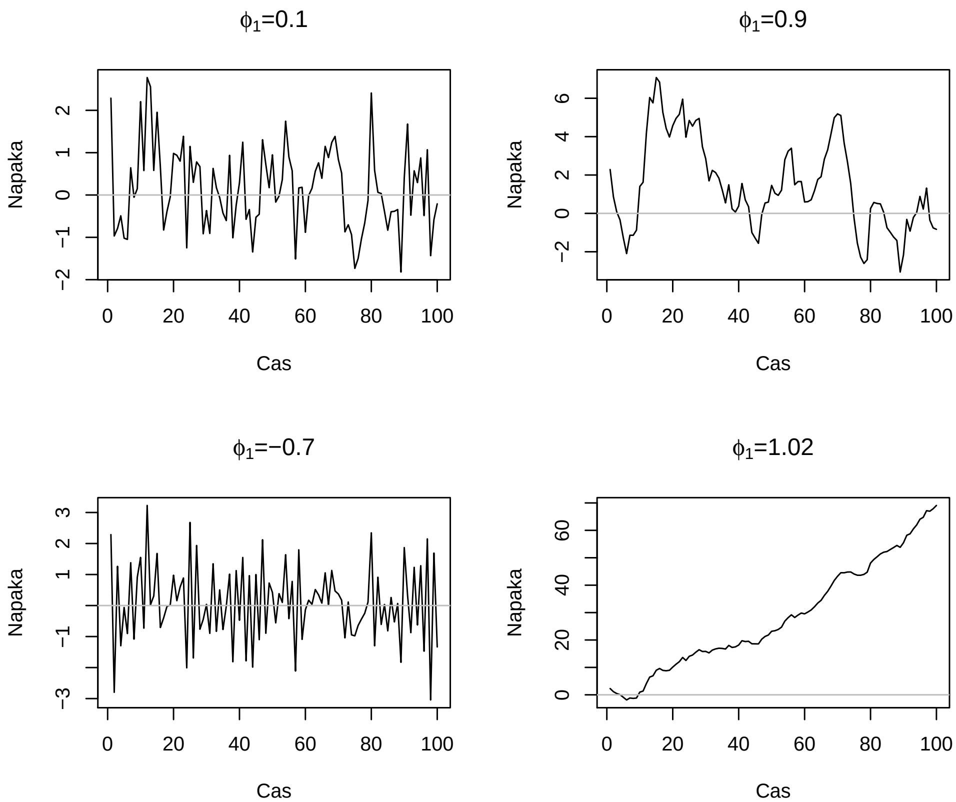
$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{C} = \frac{\sigma_w^2}{1-\phi_1^2} \begin{pmatrix} 1 & \phi_1 & \phi_1^2 & \cdots & \phi_1^{n-1} \\ \phi_1 & 1 & \phi_1 & \cdots & \phi_1^{n-2} \\ \phi_1^2 & \phi_1 & 1 & \cdots & \phi_1^{n-3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi_1^{n-1} & \phi_1^{n-2} & \phi_1^{n-3} & \cdots & 1 \end{pmatrix}. \tag{35}$$

Slika 24 prikazuje primere avtokorelacijskih funkcij za model AR(1) pri različnih vrednostih ϕ_1. Večja vrednost parametra ϕ_1 pomeni počasnejše eksponentno padanje avtokorelacijske funkcije z večanjem odloga. Pri negativnih vrednostih parametra ϕ_1 imajo zaporedne vrednosti avtokorelacijske funkcije nasproten predznak.



Slika 24: Avtokorelacijska funkcija (ACF) za model AR(1) z različnimi pozitivnimi vrednostmi ϕ_1 (levo) in različnimi negativnimi vrednostmi ϕ_1 (desno)

Slika 25 kaže štiri simulirane časovne vrste napak po modelu AR(1). V prvem primeru je $\phi_1 = 0.1$, kar pomeni, da je korelacija med zaporednimi vrednostmi majhna in so zaporedne vrednosti bolj ali manj slučajne. V drugem primeru je $\phi_1 = 0.9$, zaporedne vrednosti so tesno pozitivno korelirane, v tretjem primeru je $\phi_1 = -0.7$, kar pomeni negativno korelacijo. V vseh treh primerih gre za za stacionarne časovne vrste. V četrtem primeru je $\phi_1 = 1.02$ in gre za nestacionarno časovno vrsto, za katero model AR(1) ni ustrezen. V vseh primerih je $w_t \sim N(0, 1)$.



Slika 25: Simulirane časovne vrste napak za model AR(1), $\phi_1 = 0.1$ (zgoraj levo), $\phi_1 = 0.9$ (zgoraj desno), $\phi_1 = -0.7$ (spodaj levo) in nestacionarna časovna vrsta s parametrom $\phi_1 = 1.02$ (spodaj desno)

Posplošitev avtoregresijskega modela prvega reda AR(1) prinese **avtoregresijske modele reda *p*, AR(*p*)**:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + ... + \phi_p \varepsilon_{t-p} + w_t. \tag{36}$$

Splošno je v avtoregresijskem modelu reda p napaka v času t linearna kombinacija napak z odlogom s, $s = 1, ..., p$, in belega šuma v času t, ki ima povprečje 0 in varianco σ_w^2. Model je podoben linearnemu regresijskemu modelu: ϕ_i, $i = 1, ..., p$, so parametri modela, odtod je tudi njegovo ime avtoregresijski. Korelacijsko matriko $\mathbf{C}$ določa vektor parametrov $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)$, ki jih ocenimo na podlagi podatkov.

V splošnem so avtokorelacijske funkcije stacionarnih modelov AR(p) mešanica členov, ki eksponentno padajo in dušenih sinusnih ali kosinusnih nihanj.

Poglejmo še, kakšno vlogo imajo pri AR(p) **parcialne avtokorelacije z odlogom** s, ki izražajo avtokorelacijo med ε_t in ε_{t-s} ob upoštevanju avtokorelacije ε_t z $\varepsilon_{t-s-1}, ..., \varepsilon_{t-1}$.

Koeficient parcialne avtokorelacije z odlogom s, $s = 1, ..., p$, je zadnji parameter avtoregresijskega modela s samo s členi, označimo ga ϕ_{ss}:

$$\begin{aligned}\varepsilon_t &= \phi_{11}\varepsilon_{t-1} + w_t,\\ \varepsilon_t &= \phi_1\varepsilon_{t-1} + \phi_{22}\varepsilon_{t-2} + w_t,\\ &\ldots\\ \varepsilon_t &= \phi_1\varepsilon_{t-1} + ... + \phi_{pp}\varepsilon_{t-p} + w_t.\end{aligned} \tag{37}$$

Pokažemo lahko, da so koeficienti parcialne avtokorelacije z odlogom večjim od p enaki nič. To lastnost uporabimo kot diagnostično orodje pri izbiri ustreznega reda modela AR(p).

2.3.2 Model drsečih sredin

Drugi pogosto uporabljeni model za časovno vrsto napak je **model drsečih sredin prvega reda MA(1)**. V tem primeru je napaka v času t, ε_t, odvisna od belega šuma v času $t-1$ in belega šuma v času t:

$$\varepsilon_t = w_t + \theta_1 w_{t-1}. \tag{38}$$

Avtokorelacijska funkcija za model MA(1) je

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2}, \quad \rho_s = 0 \quad \text{za} \quad s > 1. \tag{39}$$

Korelacijska matrika napak je v tem primeru

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{C} = \sigma_w^2(1+\theta_1^2)\begin{pmatrix} 1 & \frac{\theta_1}{1+\theta_1^2} & & \cdots & 0\\ \frac{\theta_1}{1+\theta_1^2} & 1 & \frac{\theta_1}{1+\theta_1^2} & \cdots & \\ 0 & \frac{\theta_1}{1+\theta_1^2} & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & 1\end{pmatrix}. \tag{40}$$

Za model MA(2)

$$\varepsilon_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2}, \tag{41}$$

je avtokorelacijska funkcija

$$\rho_1 = \frac{\theta_1 + \theta_1\theta_2}{1+\theta_1^2+\theta_2^2}, \quad \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2}, \quad \rho_s = 0 \quad \text{za} \quad s > 2. \tag{42}$$

V splošnem je v modelu **drsečih sredin reda q, MA(q)**, ε_t linearna kombinacija q belih šumov v časih od t do $t-q$, ki so porazdeljeni $N(0, \sigma_w^2)$:

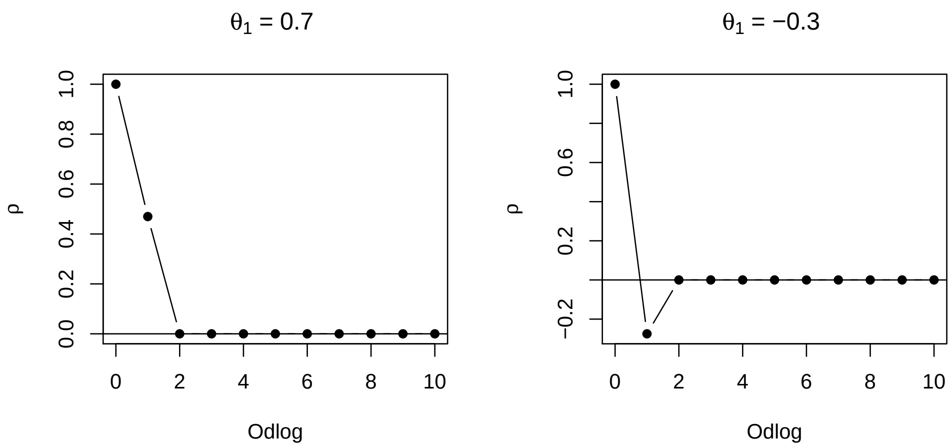
$$\varepsilon_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}. \tag{43}$$

Za avtokorelacijsko funkcijo modela MA(q) velja, da je $\rho_s = 0$ za $s > q$. To lastnost avtokorelacijske funkcije uporabimo kot diagnostično orodje pri izbiri ustreznega reda modela MA(q).

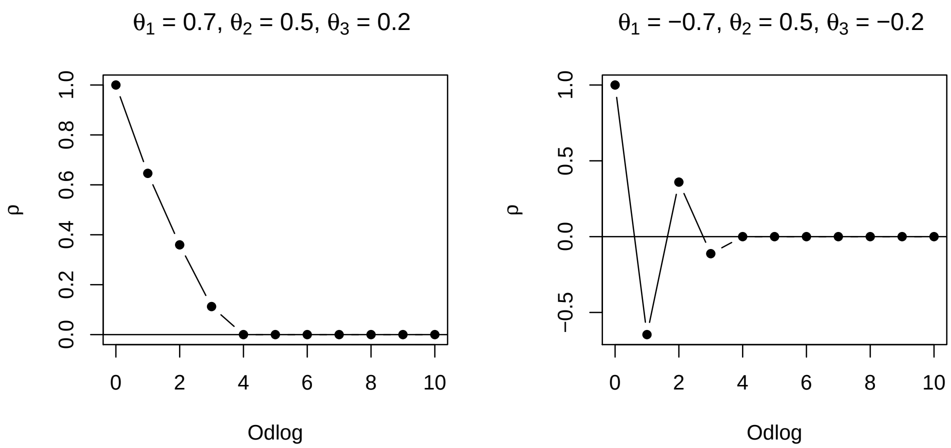
Časovno vrsto napak modeliramo kot neke vrste drsečo sredino belih šumov, od tod tudi njeno ime. Korelacijsko matriko $\mathbf{C}$ v tem primeru določa vektor parametrov $\boldsymbol{\theta}_1 = (\theta_1, ..., \theta_q)$.

Ker je model MA(q) sestavljen iz končnega števila belih šumov, je vedno stacionaren in glede tega ni potrebno postavljati pogojev glede vrednosti parametrov. Pogoji za parametre θ v tem primeru izhajajo iz zahteve po enolično določeni avtokovariančni funkciji. Pravimo, da mora biti model MA(q) **obrnljiv**. Za model MA(1) je obrnljivost zagotovljena, če je $|\theta_1| < 1$.

Za ilustracijo narišimo avtokorelacijsko funkcijo za modela MA(1) in MA(3). Izbrane vrednosti parametrov so razvidne iz Slik 26 in 27.



Slika 26: ACF za dva modela MA(1): $\varepsilon_t = w_t + 0.7w_{t-1}$ (levo) in $\varepsilon_t = w_t - 0.3w_{t-1}$ (desno)



Slika 27: ACF za dva modela MA(3): $\varepsilon_t = w_t + 0.7w_{t-1} + 0.5w_{t-2} + 0.2w_{t-3}$ (levo) in $\varepsilon_t = w_t - 0.7w_{t-1} + 0.5w_{t-2} - 0.2w_{t-3}$ (desno)

2.3.3 Modeli ARMA(p, q)

Pogosto se uporablja kombinacijo modela AR(p) in MA(q), kar imenujemo **model ARMA(p, q)**, avtoregresijski model reda p z modelom drsečih sredin reda q. V primeru ARMA(1,1) je napaka ε_t v času t, odvisna od napake ε_{t-1} v času $t-1$ in od slučajnega vpliva v času t in v času $t-1$:

$$\varepsilon_t = \phi_1 \varepsilon_{t-1} + w_t + \theta_1 w_{t-1}. \tag{44}$$

Ta model časovne vrste napak je stacionaren, če velja $|\phi_1| < 1$ in obrnljiv, če je $|\theta_1| < 1$. Avtokorelacijska funkcija modela ARMA(1,1) je:

$$\rho_1 = \frac{(1+\phi_1\theta_1)(\phi_1+\theta_1)}{1+\theta_1^2+2\phi_1\theta_1}, \quad \rho_s = \phi_1 \rho_{s-1}, \quad \text{za} \quad s > 1. \tag{45}$$

Koeficienti avtokorelacije eksponentno padajo z večanjem odloga. V splošnem v modelu ARMA(p, q) ocenjujemo $p+q$ parametrov $(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ korelacijske matrike $\mathbf{C}$.

2.3.4 Ocene avtokorelacij in parcialnih avtokorelacij

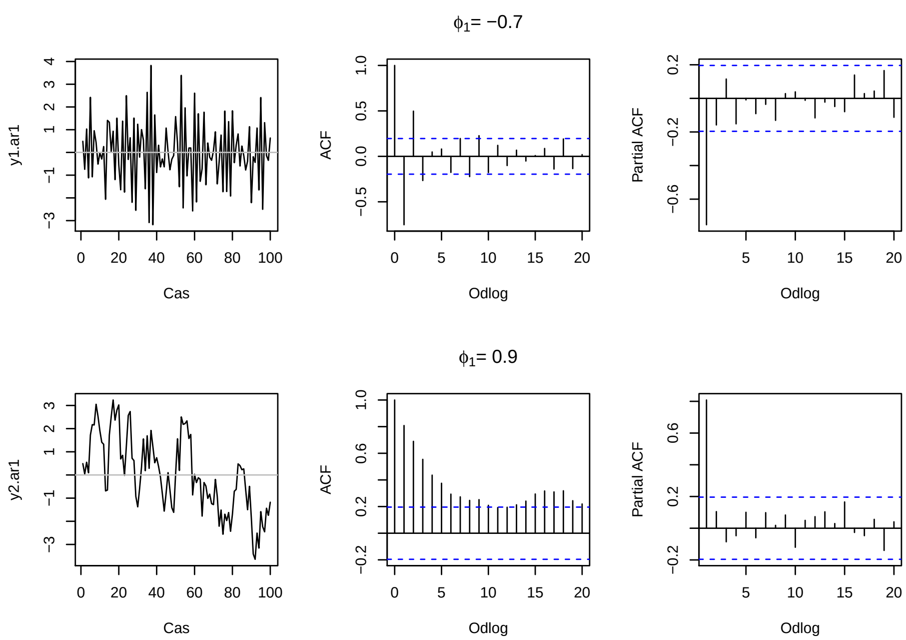
V praksi na osnovi ostankov regresijskega modela ocenimo avtokorelacijo napak z odlogom s, označimo jo r_s. Izračunamo jo kot koeficient korelacije med izhodiščno časovno vrsto ostankov in časovno vrsto ostankov, ki je zamaknjena za s podatkov:

$$r_s = \frac{\sum_{t=s+1}^{n} e_t e_{t-s}}{\sum_{t=1}^{n} e_t^2}. \tag{46}$$

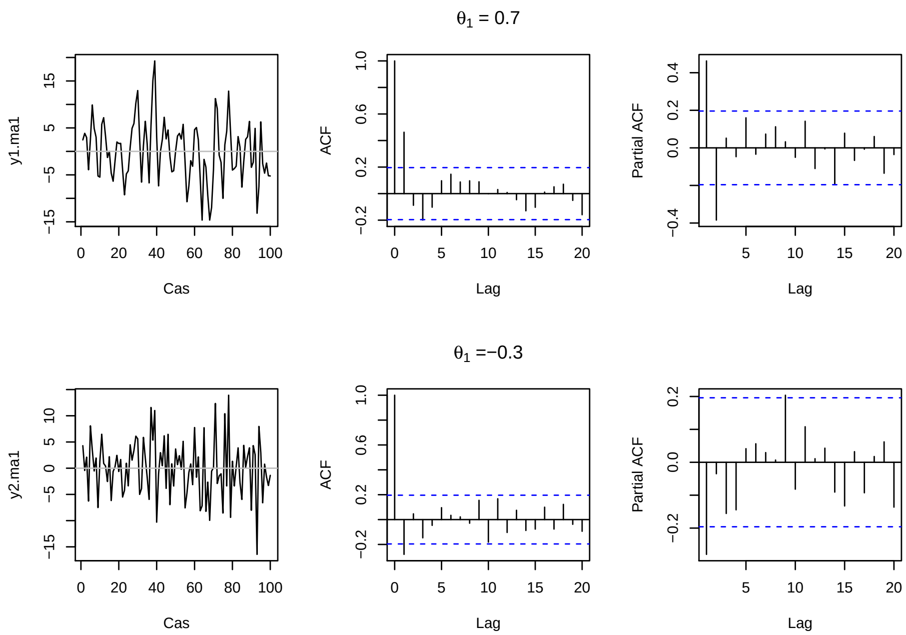
Če bi bili ostanki neodvisni, bi bila standardna napaka koeficienta avtokorelacije r_s pri vseh s aproksimativno $1/\sqrt{n}$. Zato za mejo statistične pomembnosti posameznega koeficienta avtokorelacije r_s pri $\alpha = 0.05$ vzamemo $\pm 1.96/\sqrt{n}$.

Grafični prikaz ocen koeficientov avtokorelacije v odvisnosti od odlogov imenujemo avtokorelogram (ACF). Grafični prikaz na ostankih regresijskega modela ocenjenih parcialnih koeficientov avtokorelacije pa parcialni avtokorelogram (PACF). Za določitev primerne vrednosti za p za model AR(p) uporabimo PACF, p je število zaporednih statistično značilnih koeficientov parcialne avtokorelacije; za določitev primerne vrednosti za q za model MA(`q`) pa ACF, q je število zaporednih statistično značilnih koeficientov avtokorelacije.

Slika 28 kaže dve simulirani časovni vrsti napak po modelih AR(1): $\varepsilon_t = -0.7\varepsilon_{t-1} + w_t$ in $\varepsilon_t = 0.9\varepsilon_{t-1} + w_t$, v obeh primerih je $w_t \sim N(0, 1)$ ter pripadajoča avtokorelograma (ACF) in parcialna avtokorelograma (PACF), ki ju izračunamo in narišemo s funkcijama `acf` in `pacf` iz paketa `stats`. Slika 29 kaže simulirani časovni vrsti za izbrana dva modela MA(1) in pripadajoča avtokorelograma in parcialna avtokorelograma.

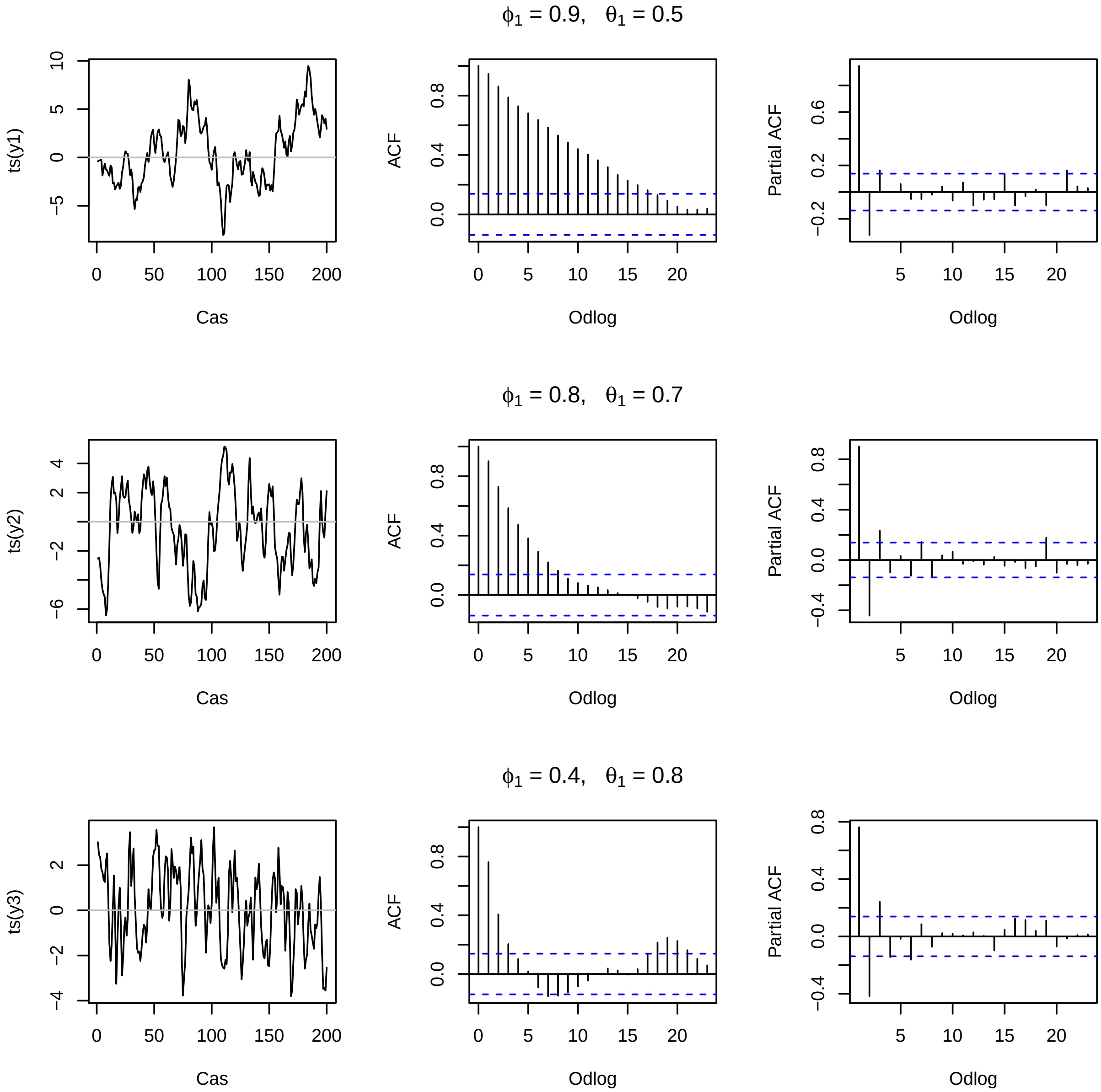


Slika 28: Simulirana časovna vrsta napak za model AR(1) z vrednostjo $\phi_1 = -0.7$ in $\phi_1 = 0.9$ ter pripadajoča ACF in PACF



Slika 29: Simulirani časovni vrsti napak za MA(1) za $\theta_1 = 0.7$ in $\theta_1 = -0.3$ ter pripadajoča ACF in PACF

Primer treh simuliranih časovnih vrst napak za ARMA (1,1) model s pripradajočima ACF in PACF kaže Slika 30. Za ta model določitev p in q na podlagi avtokorelograma in parcialnega avtokorelograma ni več tako jasna. Pokaže se, da z modeli ARMA(p, q) pogosto nadomeščajo modele AR(p) višjih redov.



Slika 30: Tri simulirane časovne vrste napak za model ARMA(1,1) ter ACF in PACF

2.3.5 Durbin-Watsonova statistika

Obstaja statistični test za avtokorelacijo oziroma serialno korelacijo ostankov na osnovi Durbin-Watsonove statistike

$$D_s = \frac{\sum_{t=s+1}^{n} (e_t - e_{t-s})^2}{\sum_{t=1}^{n} e_t^2}, \tag{47}$$

ki ima v splošnem neznano ničelno porazdelitev; če je n velik, velja $D_s \approx 2(1 - r_s)$. Posledično vrednosti Durbin-Watsonove statistike okoli 2 pomenijo majhno avtokorelacijo ostankov, vrednosti pod 2 pozitivno avtokorelacijo, nad 2 pa negativno avtokorelacijo ostankov z odlogom s.

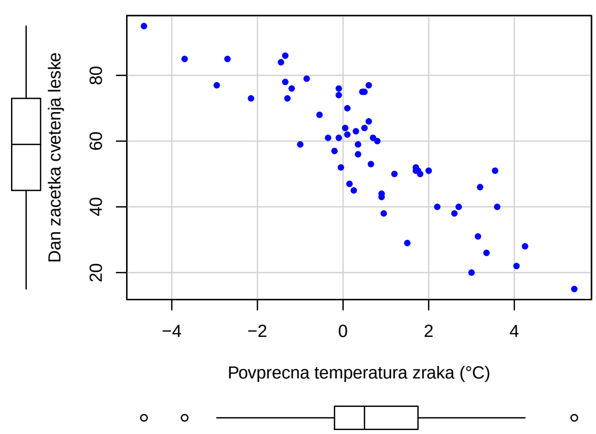
Durbin-Watsonove statistike se izračuna s funkcijo `durbinWatsonTest` iz paketa `car`. Za preverjanje ničelne domneve $\rho_s = 0$ je za oceno p-vrednosti uporabljen bootstrap pristop.

2.3.6 Primer: LESKA

V datoteki LESKA.txt so podatki o dnevu, ko začne cveteti leska (`cvet.dan`, to je zaporedni dan v letu) in o povprečni dvomesečni temperaturi januarja in februarja (`temp`, °C), podatki so za zaporedna koledarska leta od 1955-2011 v Ljubljani (`leto`).

```
> leska<-read.table("LESKA.txt", header=T)
> str(leska)

'data.frame':       57 obs. of  3 variables:
 $ leto    : int  1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 ...
 $ cvet.dan: int  51 77 60 45 52 56 59 77 95 85 ...
 $ temp    : num  1.75 -2.95 0.8 0.25 -0.05 0.35 0.35 0.6 -4.65 -2.7 ...
```

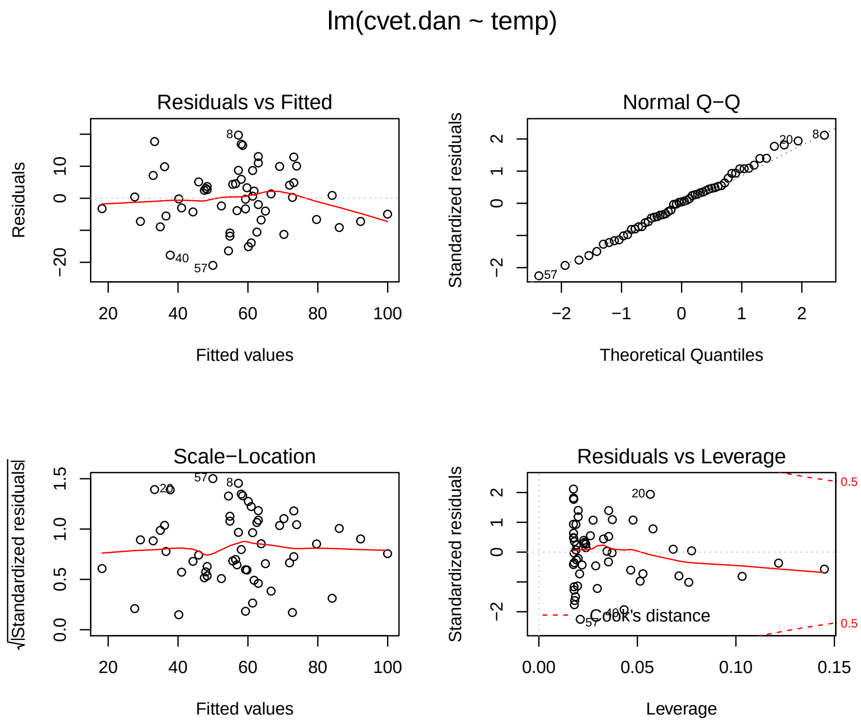
Analizirajmo odvisnost časa začetka cvetenja leske od dvomesečne povprečne temperature za januar in februar (`temp`) (Slika 31). V tem primeru sta odzivna in napovedna spremenljivka časovni vrsti dolžine 57.



Slika 31: Dan začetka cvetenja leske v odvisnosti od povprečne dvomesečne temperature zraka v Ljubljani

Slika 31 kaže, da `cvet.dan` linearno pada s `temp`.

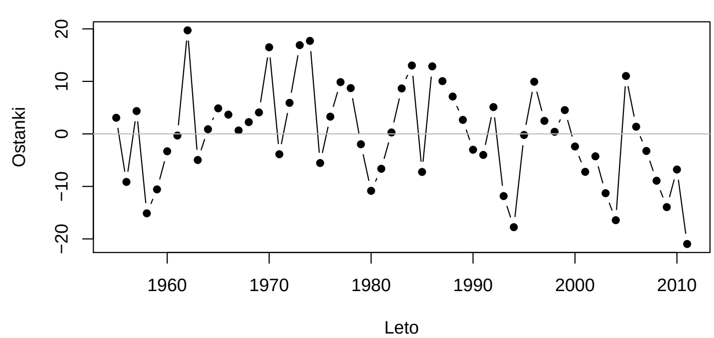
```
> mod.lm<-lm(cvet.dan~temp, data=leska)
```



Slika 32: Ostanki za `mod.lm`

Slika 38 ne kaže, da bi bilo z ostanki kaj narobe. Za upoštevanje časovne dimenzije je dodana Slika 33. Časovna vrsta ostankov ne kaže močne avtokorelacije ostankov, se pa vidi, da v več primerih za nekaj zaporednimi pozitivnimi ostanki sledijo zaporedni negativni ostanki, kar bi lahko bila posledica avtokorelacije.

```
> plot(leska$leto, residuals(mod.lm), type="b",pch=16, xlab="Leto", ylab="Ostanki")
> abline(h=0, col="grey")
```



Slika 33: Časovna vrsta ostankov za `mod.lm1`

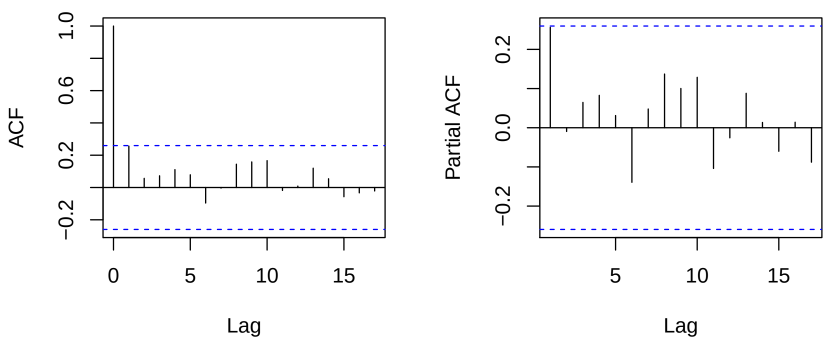
```
> (DWT1<-durbinWatsonTest(mod.lm, max.lag=5))

 lag Autocorrelation D-W Statistic p-value
   1      0.25626565      1.395108   0.016
   2      0.05677640      1.767396   0.422
   3      0.07264266      1.691793   0.330
   4      0.11079895      1.552026   0.216
   5      0.07882800      1.590820   0.286
 Alternative hypothesis: rho[lag] != 0
```

Samo prva DW statistika z odlogom 1 je statistično značilna ($p = 0.016$), tudi PACF ostankov za `mod.lm` (Slika 34) kaže mejno statistično značilen koeficient avtorkorelacije z odlogom 1.

Slika 34 prikazuje avtokorelogram (ACF) za ostanke z odlogi od 0 do 17 let ($10 \cdot log10(n)$, $n = 57$) in parcialni avtokorelogram (PACF) z odlogi 1 do 17 let za ostanke modela `model.lm`, ki jih prikažemo s funkcijama `acf` oz. `pacf` iz paketa `stats`. Če kot argument funkcije dodamo `plot=FALSE`, se izpišejo ocene avtokorelacij oz. parcialnih avtokorelacij.

```
> acf(residuals(mod.lm), plot=FALSE)

Autocorrelations of series 'residuals(mod.lm)', by lag

     0      1      2      3      4      5      6      7      8      9     10
 1.000  0.256  0.057  0.073  0.111  0.079 -0.096 -0.004  0.144  0.159  0.166
    11     12     13     14     15     16     17
-0.018  0.009  0.120  0.054 -0.057 -0.033 -0.021
```

```
> pacf(residuals(mod.lm), plot=FALSE)

Partial autocorrelations of series 'residuals(mod.lm)', by lag

     1      2      3      4      5      6      7      8      9     10     11 
 0.256 -0.010  0.065  0.083  0.031 -0.139  0.048  0.137  0.100  0.129 -0.104 
    12     13     14     15     16     17 
-0.026  0.088  0.013 -0.060  0.014 -0.088 

> par(oma=c(0,0,2,0), mfrow=c(1,2))
> acf(residuals(mod.lm), main="")
> pacf(residuals(mod.lm), main="")
```



Slika 34: ACF in PACF za ostanke za `mod.lm`

Poskusimo model `mod.lm` dopolniti z modeliranjem avtokorelacije napak z modelom AR(1).

```
> mod.gls.ar<-gls(cvet.dan~temp, correlation=corARMA(p=1), data=leska, method="ML")
> anova(mod.gls.ar, mod.lm)

           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.gls.ar     1  4 418.5352 426.7074 -205.2676                        
mod.lm         2  3 421.1650 427.2942 -207.5825 1 vs 2 4.629845  0.0314

> summary(mod.gls.ar)

Generalized least squares fit by maximum likelihood
  Model: cvet.dan ~ temp 
  Data: leska 
       AIC      BIC    logLik
  418.5352 426.7074 -205.2676
```

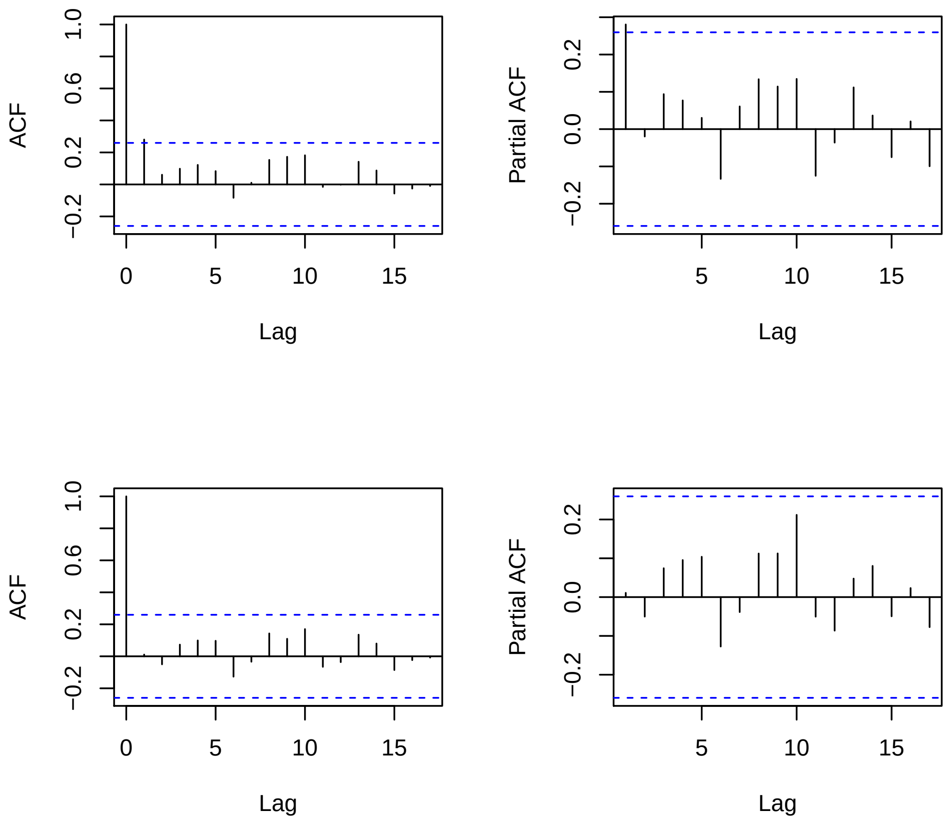
```
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.3042674

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 61.75107 1.7517445  35.25118       0
temp        -7.70622 0.6324062 -12.18555       0

 Correlation:
     (Intr)
temp -0.232

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.2787843 -0.5720084  0.1096345  0.6285807  2.1369420

Residual standard error: 9.299581
Degrees of freedom: 57 total; 55 residual
```

Primerjava modelov na podlagi posplošenega testa razmerij verjetij pokaže, da je boljši `mod.gls.ar` ($p = 0.0314$), ki vključuje modeliranje korelacije napak s funkcijo AR(p=1).

Modeliranje avtokorelacije napak v `gls` v modelu ne spremeni standardiziranih ostankov modela, vpliv avtokorelacije se vidi na t. i. **normaliziranih ostankih**, $r_i = \hat{\sigma}^{-1}(\hat{\Lambda}_i^{-1/2})^T(y_i - \hat{y}_i)$. Če je model, v katerem modeliramo variančno-kovariančno matriko napak, sprejemljiv, velja, da so normalizirani ostanki porazdeljeni $N(0, \sigma^2 I)$. Zato za diagnostične grafične prikaze uporabimo normalizirane ostanke, kar pomeni, da ima argument `type` v funkciji `resid` vrednost `"n"` (Slika 35).

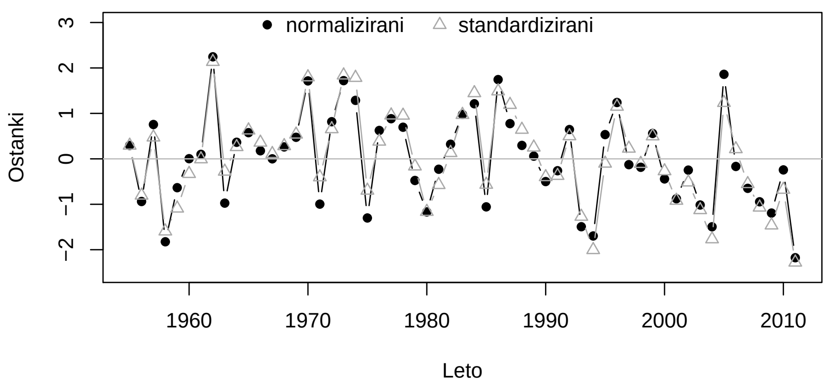
Modeliranje variančno-kovariančne matrike napak spremeni standardne napake ocen parametrov modela. Če je avtokorelacija napak velika, to običajno poveča standardne napake. To je drugače kot pri modeliranju variance napak, kjer se poleg standardnih napake ocen parametrov spremenijo tudi ostanki in.

```
> par(oma=c(0,0,2,0), mfrow=c(2,2))
> acf(residuals(mod.gls.ar, type="p"), main="")
> pacf(residuals(mod.gls.ar, type="p"), main="")
> acf(residuals(mod.gls.ar, type="n"), main="")
> pacf(residuals(mod.gls.ar, type="n"), main="")
```



Slika 35: ACF in PACF za standardizirane (zgoraj) in normalizirane (spodaj) ostanke za `mod.gls.ar`

```
> plot(leska$leto, residuals(mod.gls.ar, type="n"), type="b",pch=16, xlab="Leto",
+        ylab="Ostanki", col="black", ylim=c(-2.5,3))
> points(leska$leto, residuals(mod.gls.ar, type="p"), type="b",pch=2, xlab="Leto",
+        ylab="Ostanki", col="darkgrey")
> legend(x=1965, y=3.5, horiz=T, c("normalizirani", "standardizirani"), box.lty=0, pch=c(16,2),
+        col=c("black", "darkgrey"))
> abline(h=0, col="grey")
```



Slika 36: Časovni vrsti ostankov
in za `mod.gls.ar`

Na podlagi ACF na Sliki 34 bi lahko avtokorelacijo napak modelirali tudi z modelom drsečih sredin MA(1). Pokaže se, da dobimo enakovredne rezultate.

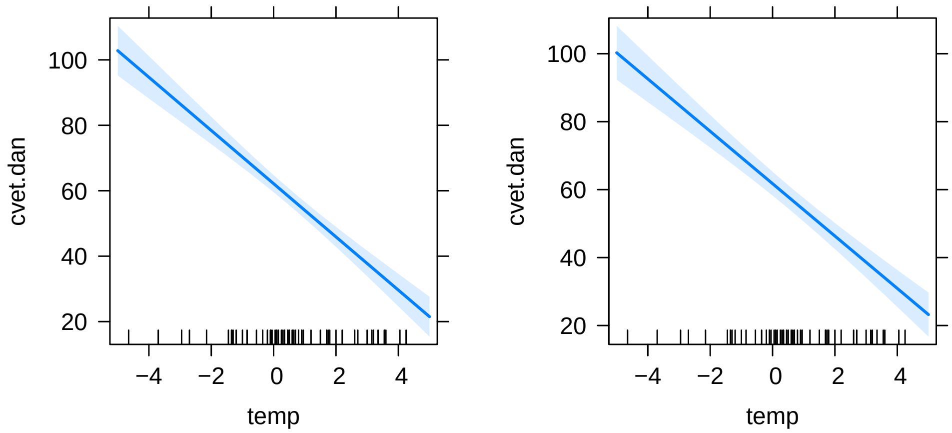
```
> mod.gls.ma<-gls(cvet.dan~temp, correlation=corARMA(q=1), data=leska, method="ML")
> anova(mod.gls.ma, mod.lm)

           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.gls.ma     1  4 418.4744 426.6467 -205.2372
mod.lm         2  3 421.1650 427.2942 -207.5825 1 vs 2 4.690582  0.0303

> anova(mod.gls.ma, mod.gls.ar)

           Model df      AIC      BIC    logLik
mod.gls.ma     1  4 418.4744 426.6467 -205.2372
mod.gls.ar     2  4 418.5352 426.7074 -205.2676

> summary(mod.gls.ma)

Generalized least squares fit by maximum likelihood
  Model: cvet.dan ~ temp
  Data: leska
       AIC      BIC    logLik
```

```
  418.4744 426.6467 -205.2372

Correlation Structure: ARMA(0,1)
 Formula: ~1
 Parameter estimate(s):
   Theta1
0.3242343

Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 61.74802 1.6261059  37.97294       0
temp        -7.60886 0.6310206 -12.05802       0

 Correlation:
     (Intr)
temp -0.251

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.2924377 -0.6081937  0.1088335  0.6064384  2.1293873

Residual standard error: 9.306571
Degrees of freedom: 57 total; 55 residual
```

Primerjajmo ocene parametrov in pripadajoče standardne napake za vse tri modele:

```
> compareCoefs(mod.gls.ar, mod.gls.ma, mod.lm)

Calls:
1: gls(model = cvet.dan ~ temp, data = leska, correlation = corARMA(p = 1),
  method = "ML")
2: gls(model = cvet.dan ~ temp, data = leska, correlation = corARMA(q = 1),
  method = "ML")
3: lm(formula = cvet.dan ~ temp, data = leska)

            Model 1 Model 2 Model 3
(Intercept)   61.75   61.75   62.16
SE             1.75    1.63    1.31

temp         -7.706  -7.609  -8.131
SE            0.632   0.631   0.636
```

Razlike ocen parametrov med `lm` in `gls` modeloma so zelo majhne, standardne napake pa se pri presečišču nekoliko povečajo, pri naklonu pa ni bistvene razlike. To se pozna tudi na nekoliko širših intervalih zaupanja za povprečno napoved za `mod.gls.ar` na Sliki 37.



Slika 37: Napovedi z 95 % intervalom zaupanja za povprečno napoved za `mod.lm` (levo) in za `mod.gls.ar` (desno)

```
> intervals(mod.gls.ar)

Approximate 95% confidence intervals

 Coefficients:
                lower      est.     upper
(Intercept) 58.24049 61.751067 65.261641
temp        -8.97359 -7.706219 -6.438849
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
         lower      est.     upper
Phi 0.01697886 0.3042674 0.5451476
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
    lower      est.     upper
 7.583924  9.299581 11.403358
```

Na podlagi `mod.gls.ar` lahko rečemo, da je začetek cvetenja leske v Ljubljani statistično značilno odvisen od povprečne temperature zraka v januarju in februarju. Če je povprečna temperatura zraka v januarju in februarju 0 °C, leska v povprečju zacveti 61.8-ti dan (95 % IZ 58.2, 65.3); če se povprečna temperatura poveča za 1 °C, leska v povprečju zacveti 7.7 dni prej (95 % IZ 6.4, 9.0). Ocena koeficienta avtokorelacije napak z odlogom 1 je 0.30 (95 % IZ 0.02, 0.55).

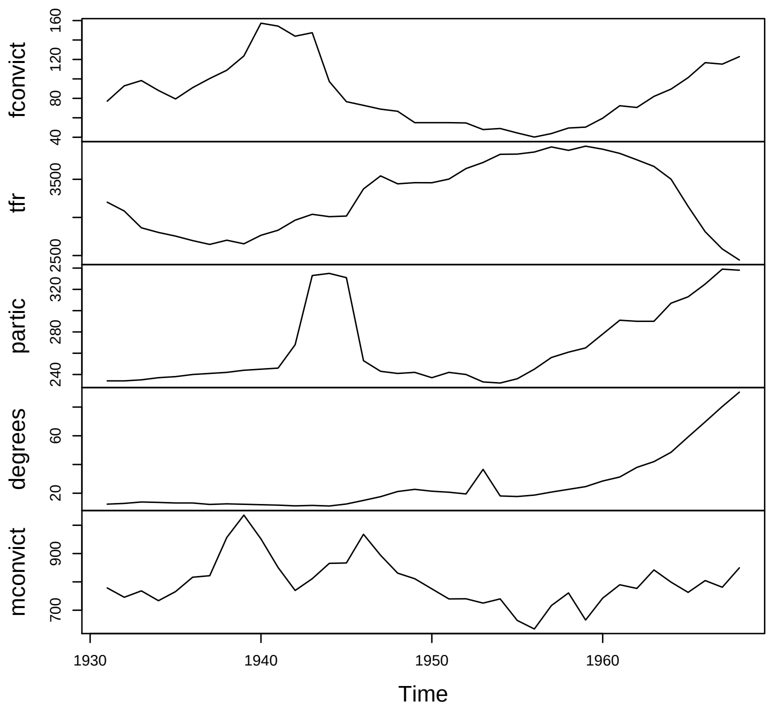
2.4 Primer: Hartnagel

V podatkovnem okviru `Hartnagel` iz paketa `car` so podatki o številu žensk obsojenih za kazniva dejanja na 100 000 žensk starosti 15 let in več v Kanadi v obdobju 1931-1968 (`fconvict`). Hartnagla je zanimalo, kako splošen položaj žensk v družbi vpliva na `fconvict`. Kot napovedne spremenljivke je vzel stopnjo rodnosti (`tfr`, število rojstev na 1000 žensk), stopnjo zaposlenosti žensk (`partic`, število zaposlenih na 1000 žensk), stopnjo visoke izobrazbe med ženskami (`degrees`, število žensk z visoko izpbrazbo na 10 000 žensk) in število za kazniva dejanja obsojenih moških na 100 000 moških (`mconvict`).

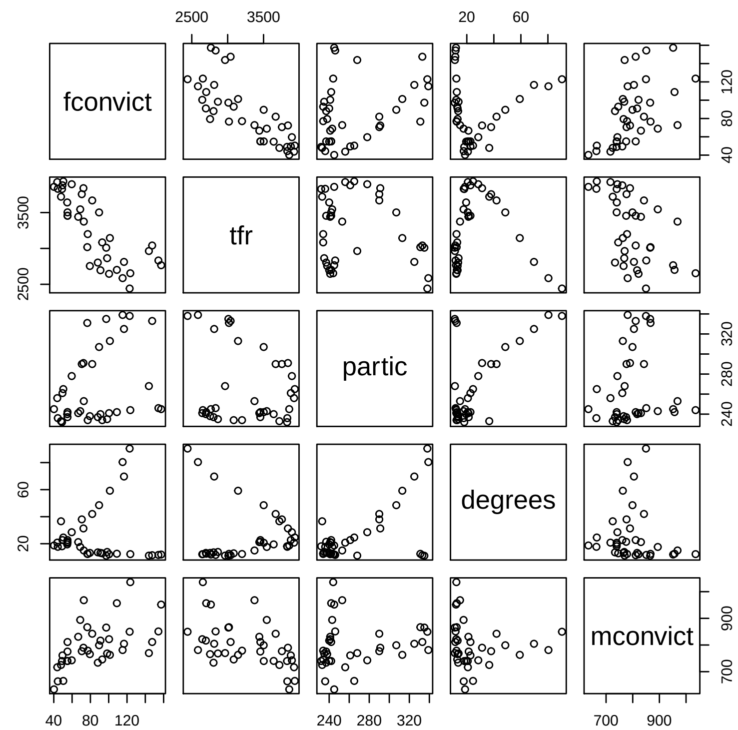
```
> data(Hartnagel)
> str(Hartnagel)
```

```
'data.frame':        38 obs. of  8 variables:
 $ year    : int  1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 ...
 $ tfr     : int  3200 3084 2864 2803 2755 2696 2646 2701 2654 2766 ...
 $ partic  : int  234 234 235 237 238 240 241 242 244 245 ...
 $ degrees : num  12.4 12.9 13.9 13.6 13.2 13.2 12.2 12.6 12.3 12 ...
 $ fconvict: num  77.1 92.9 98.3 88.1 79.4 ...
 $ ftheft  : num  NA NA NA NA 20.4 22.1 22.4 21.8 21.1 21.4 ...
 $ mconvict: num  779 746 768 734 766 ...
 $ mtheft  : num  NA NA NA NA 247 ...
```

```
> plot(cbind(fconvict=ts(Hartnagel$fconvict, start=1931),
+            tfr=ts(Hartnagel$tfr, start=1931),
+            partic=ts(Hartnagel$partic, start=1931),
+            degrees=ts(Hartnagel$degrees, start=1931),
+            mconvict=ts(Hartnagel$mconvict, start=1931)), nc=1, main="")
```

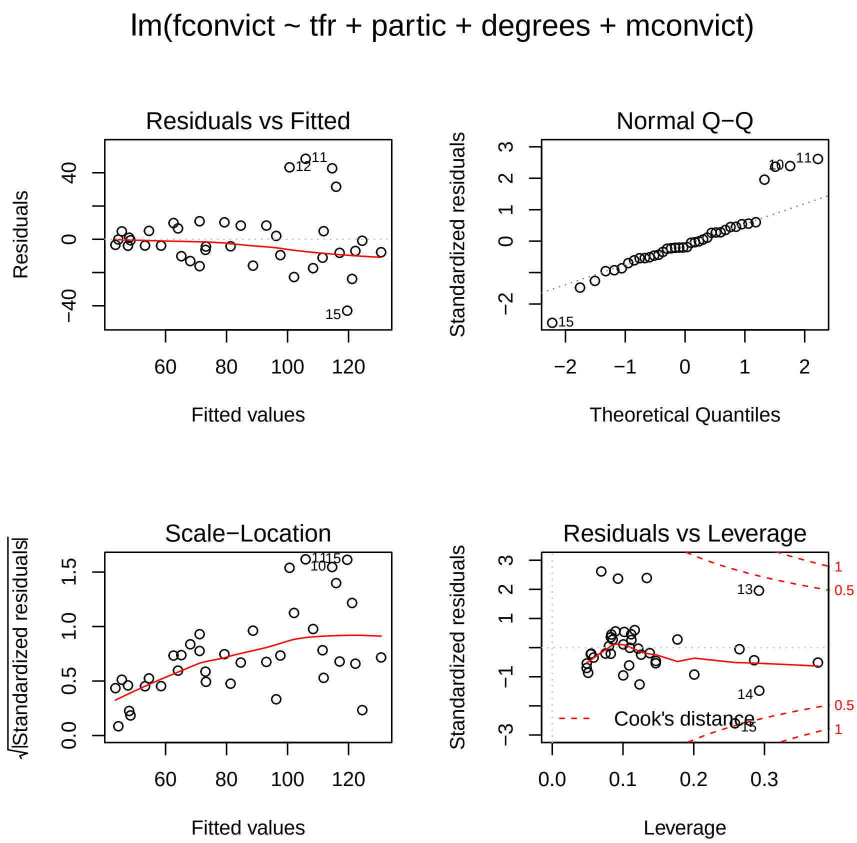


Slika 38: Časovne vrste za spremenljivke v podatkovnem okviru Hartnagel

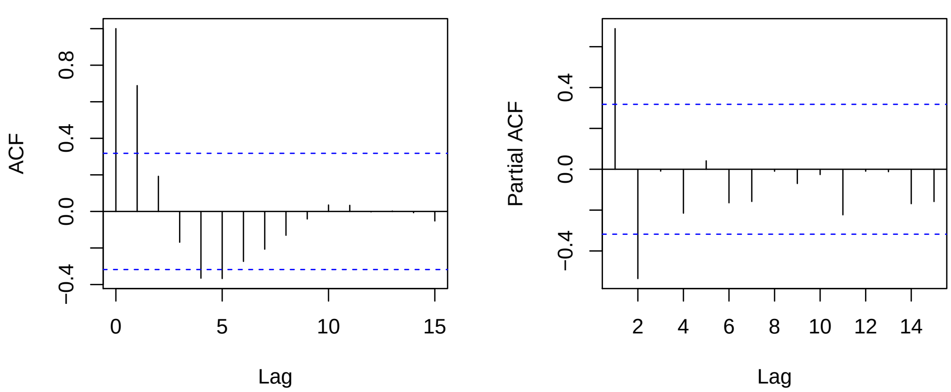


Slika 39: Matrika razsevnih grafikonov za spremenljivke v podatkovnem okviru Hartnagel

```
> mod.H.lm<-lm(fconvict ~ tfr + partic + degrees + mconvict, data=Hartnagel)
> vif(mod.H.lm)

     tfr   partic  degrees mconvict
1.432387 1.757854 1.758885 1.462099

> durbinWatsonTest(mod.H.lm, max.lag=5)

 lag Autocorrelation D-W Statistic p-value
   1       0.6883450     0.6168636   0.000
   2       0.1922665     1.5993563   0.120
   3      -0.1685699     2.3187448   0.272
   4      -0.3652775     2.6990538   0.012
   5      -0.3673240     2.6521103   0.006
 Alternative hypothesis: rho[lag] != 0
```



Slika 40: Ostanki za `mod.H.lm`



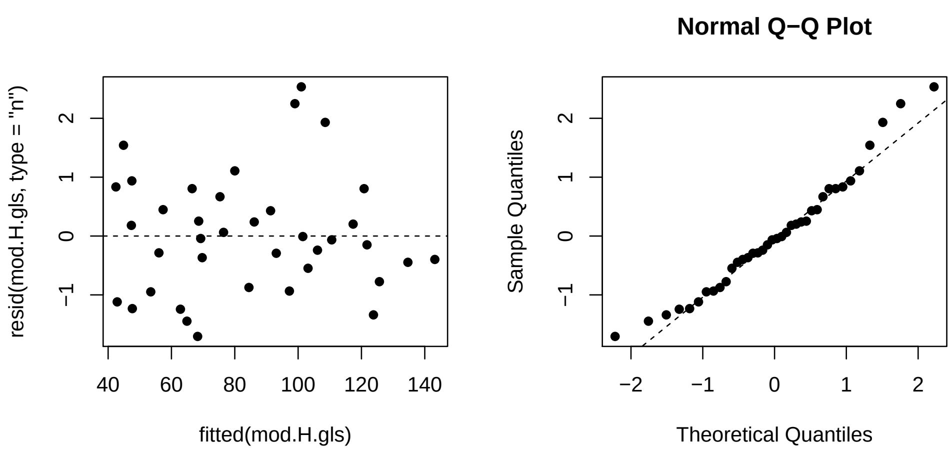
Slika 41: ACF in PACF ostankov za `mod.H.lm`

Modelirajmo najprej varianco napak z variančno funkcijo `varPower`, predpostavimo, da je varianca sorazmerna s potenco pričakovane vrednosti odzivne spremenljivke.

```
> mod.H.gls<-gls(fconvict ~ tfr + partic + degrees + mconvict, data=Hartnagel,
+                 weight=varPower(form=~fitted(.)), method="ML")
> anova(mod.H.gls, mod.H.lm)

          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.H.gls     1  7 307.5281 318.9912 -146.7640
mod.H.lm      2  6 339.0011 348.8266 -163.5006 1 vs 2 33.47304  <.0001
```

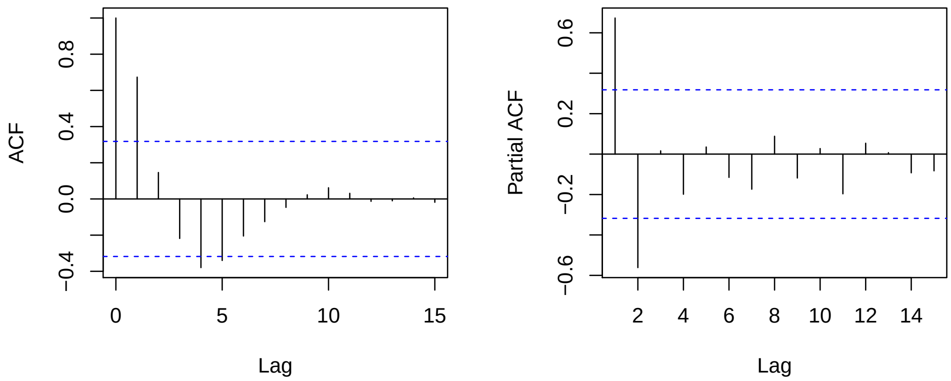
Slika 44 kaže, da je heteroskedastičnost odpravljena, model `mod.H.gls` je statistično značilno boljši od `mod.H.lm`.

```
> par(mfrow=c(1,2))
> plot(resid(mod.H.gls, type="n")~fitted(mod.H.gls), pch=16)
> abline(h=0, lty=2)
> qqnorm(resid(mod.H.gls, type="n"), pch=16)
> qqline(resid(mod.H.gls, type="n"), lty=2)
```



Slika 42: Ostanki za `mod.H.gls`

Slika 45 kaže, da z modeliranjem nekonstantne variance nismo odpravili koreliranosti ostankov.



Slika 43: ACF in PACF ostankov za `mod.H.gls`

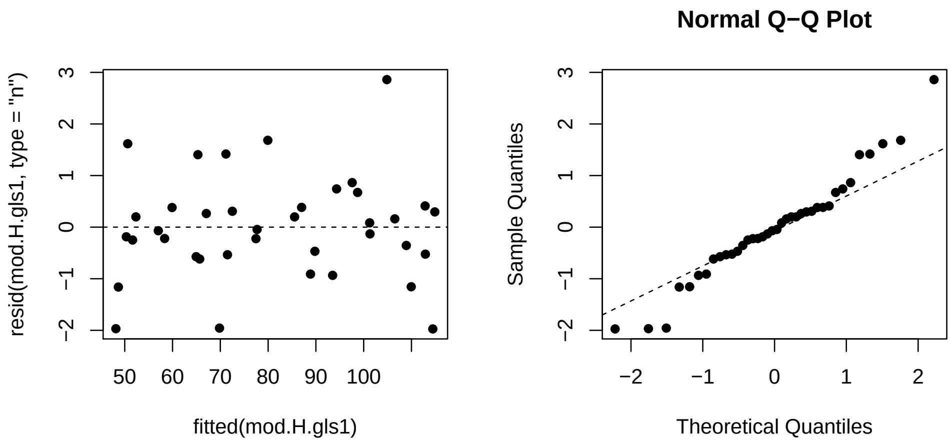
V naslednjem koraku dodajmo še model ARMA(1,1) za avtokorelacijo napak. Za ta model se odločimo na podlagi ACF in PACF (Slika 45): ACF kaže statistično značilen koeficient avtokorelacije z odlogom 1, PACF pa prva dva statistično značilna parcialna koeficienta avtokorelacije. Ta situacija kaže na to, da bi lahko izbrali tudi model AR(2).

```
> mod.H.gls1<-gls(fconvict ~ tfr + partic + degrees + mconvict, data=Hartnagel,
+               weight=varPower(form=~fitted(.)),
+               correlation=corARMA(p=1, q=1), method="ML")
> anova(mod.H.gls1, mod.H.gls)
```

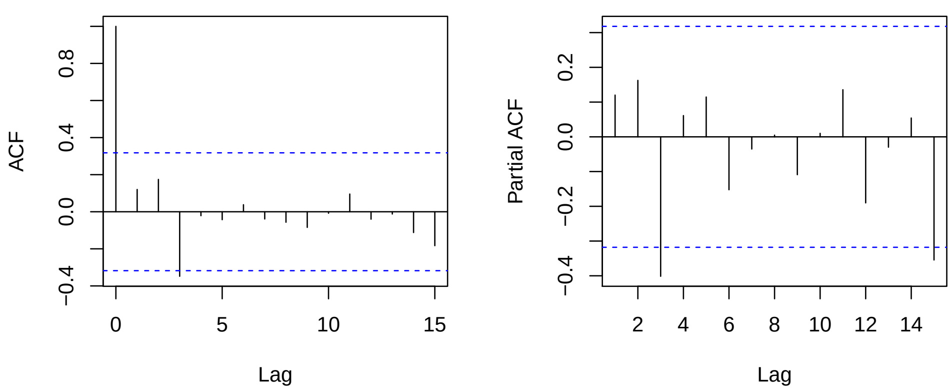
```
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.H.gls1     1  9 276.7085 291.4467 -129.3542
mod.H.gls      2  7 307.5281 318.9912 -146.7640 1 vs 2 34.81962  <.0001
```

Model `mod.H.gls1` je statistično značilno boljši od `mod.H.gls`.

```
> par(mfrow=c(1,2))
> plot(resid(mod.H.gls1, type="n")~fitted(mod.H.gls1), pch=16)
> abline(h=0, lty=2)
> qqnorm(resid(mod.H.gls1, type="n"), pch=16)
> qqline(resid(mod.H.gls1, type="n"), lty=2)
```



Slika 44: Ostanki za `mod.H.gls1`

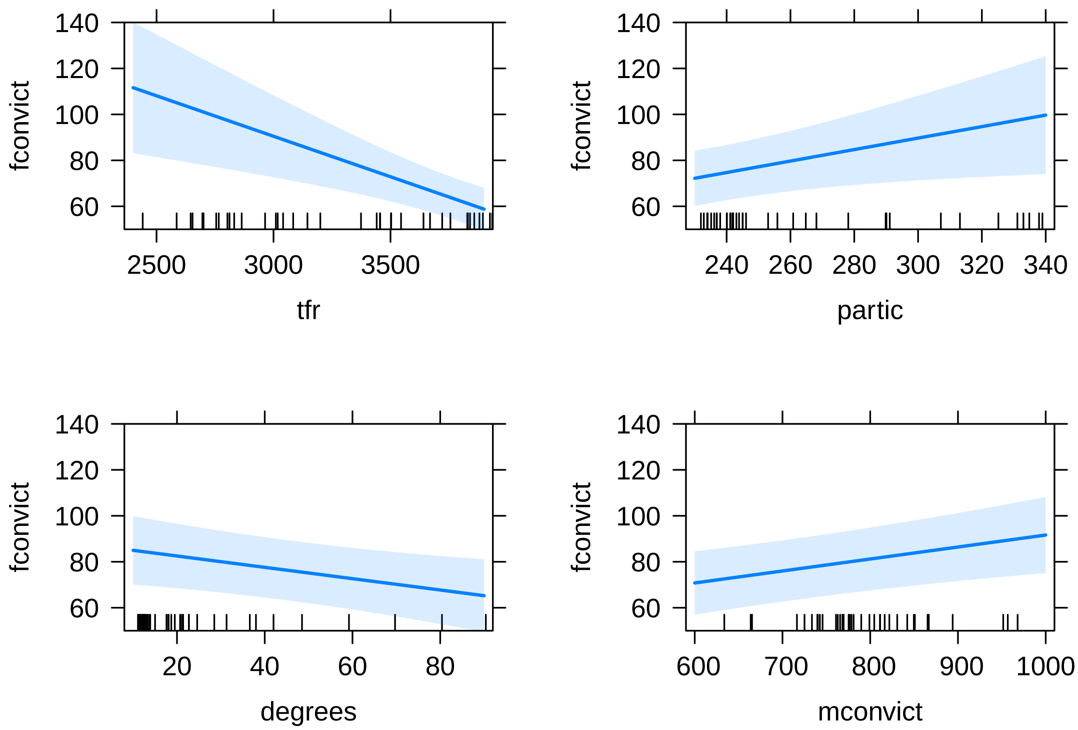


Slika 45: ACF in PACF normaliziranih ostankov za `mod.H.gls1`

```
> intervals(mod.H.gls1)

Approximate 95% confidence intervals

 Coefficients:
                  lower        est.        upper
(Intercept)  2.72883433 94.45282536 186.17681639
```

```
tfr         -0.05459597 -0.03523064  -0.01586531
partic       0.03375011  0.24979122   0.46583234
degrees     -0.44663258 -0.24656652  -0.04650047
mconvict     0.01862156  0.05212330   0.08562504
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
           lower      est.     upper
Phi1   0.1957497 0.7098980 0.9179194
Theta1 0.1132758 0.4088723 0.6379735
attr(,"label")
[1] "Correlation structure:"

 Variance function:
         lower     est.    upper
power 1.780054 2.579503 3.378952
attr(,"label")
[1] "Variance function:"

 Residual standard error:
       lower         est.        upper
4.903151e-06 1.755671e-04 6.286527e-03

> confint(glht(mod.H.gls1))

         Simultaneous Confidence Intervals

Fit: gls(model = fconvict ~ tfr + partic + degrees + mconvict, data = Hartnagel,
    correlation = corARMA(p = 1, q = 1), weights = varPower(form = ~fitted(.)),
    method = "ML")

Quantile = 2.537
95% family-wise confidence level


Linear Hypotheses:
                 Estimate   lwr        upr
(Intercept) == 0  94.452825 -19.926028 208.831679
tfr == 0          -0.035231  -0.059379  -0.011082
partic == 0        0.249791  -0.019610   0.519192
degrees == 0      -0.246567  -0.496047   0.002914
mconvict == 0      0.052123   0.010347   0.093900
```

```
> plot1<-plot(Effect(c("tfr"), mod.H.gls1), ci.style="bands",
+             main="", ylim=c(50,140))
> plot2<-plot(Effect(c("partic"), mod.H.gls1), ci.style="bands",
+             main="", ylim=c(50,140))
> plot3<-plot(Effect(c("degrees"), mod.H.gls1), ci.style="bands",
+             main="", ylim=c(50,140))
> plot4<-plot(Effect(c("mconvict"), mod.H.gls1), ci.style="bands",
+             main="", ylim=c(50,140))
> grid.arrange(plot1, plot2, plot3, plot4, ncol=2, nrow=2)
```



Slika 46: Napovedi za `mod.H.gls1` s pripadajočimi 95 % intervali zaupanja za povprečno napoved glede na posamezno napovedno spremenljivko pri povprečnih vrednostih ostalih napovednih spremenljivk v modelu

Obrazložite rezultate.

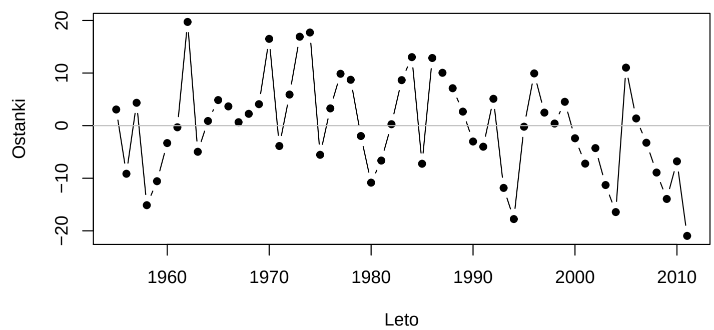
Poskusite avtokorelacijo napak modelirati še z drugačno avtokorelacijsko funkcijo (npr (AR(2))). Primerjajte rezultate.

3 VAJA

3.1 Leska

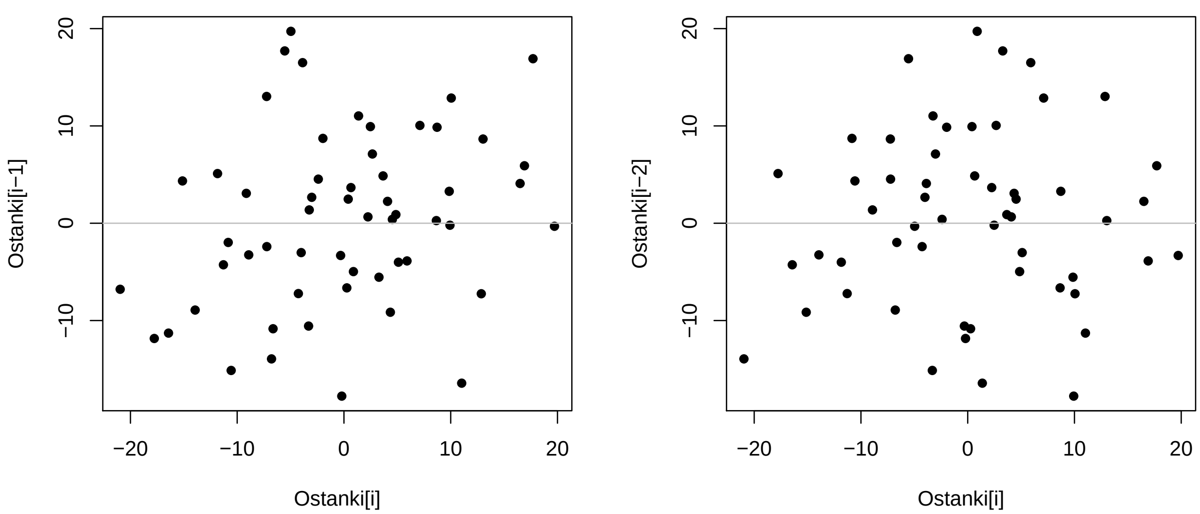
Grafično prikažite avtokoreliranost časovne vrste ostankov `mod.lm` tako, da naredite razsevni grafikon: na vodoravni osi naj bodo ostanki, na navpični osi pa ostanki z odlogom 1. Enako sliko naredite še z ostanki z odlogom 2 in 3 na navpični osi. Izračunajte pripadajoči koeficient avtokorelacije. Primerjajte dobljene rezultatez analizo v primeru LESKA v gradivu.

```
> plot(leska$leto, residuals(mod.lm), type="b",pch=16, xlab="Leto", ylab="Ostanki")
> abline(h=0, col="grey")
```



Slika 47: Časovna vrsta ostankov za `mod.lm`

```
> n<-length(residuals(mod.lm))
> par(mfrow=c(1,2))
> plot(residuals(mod.lm)[2:n],residuals(mod.lm)[1:(n-1)],
+       type="p",pch=16, xlab="Ostanki[i]", ylab="Ostanki[i-1]")
> abline(h=0, col="grey")
> plot(residuals(mod.lm)[3:n],residuals(mod.lm)[1:(n-2)], type="p",pch=16,
+       xlab="Ostanki[i]", ylab="Ostanki[i-2]")
> abline(h=0, col="grey")
> cor(residuals(mod.lm)[2:n],residuals(mod.lm)[1:(n-1)], method="pearson")

[1] 0.2694546

> cor(residuals(mod.lm)[3:n],residuals(mod.lm)[1:(n-2)], method="pearson")

[1] 0.05985431
```



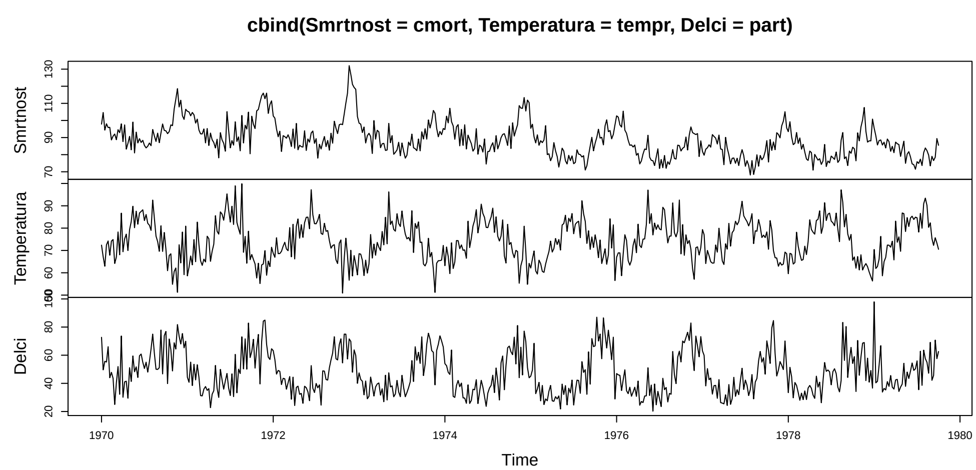
Slika 48: Razsevni grafikon za ostanke modela `mod.lm1` in za ostanke z odlogom 1 (levo) oziroma za ostanke z odlogom 2 (desno)

Na levi sličici Slike 48 se vidi šibka povezanost med ostanki in z odlogom 1 zamaknjenimi ostanki. Pearsonov koeficient korelacije je 0.27, med ostanki in z odlogom 2 zamaknjenimi ostanki ni povezanosti, Pearsonov koeficient korelacije je 0.06 (Slika 48 desno).

3.2 Kardiovaskularna smrtnost

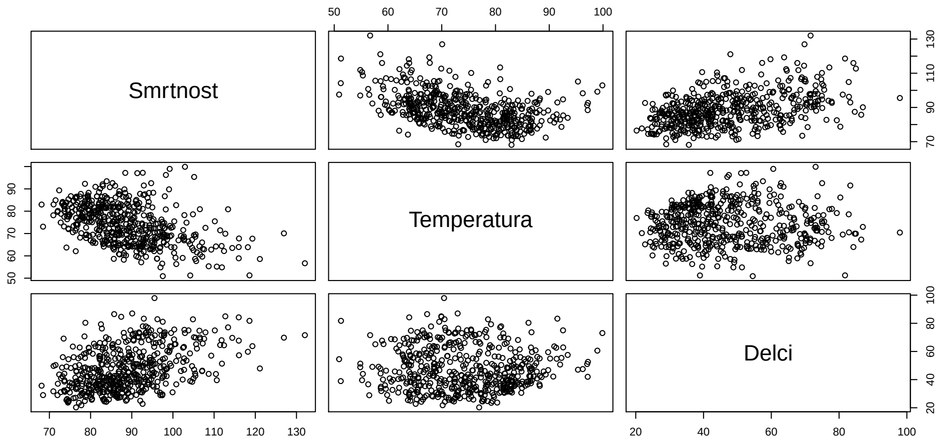
V podatkovnem okviru `lap` v paketu `astsa` je več spremenljivk, ki se navezujejo na smrtnost prebivalcev in onesnaženost zraka v Los Angelesu (*LA Pollution-Mortality Study (1970-1979, weekly data)*). Podatki so zapisani v obliki časovnih vrst (`mts format`). Zanima nas, kako na kardiovaskularno smrtnost (`cmort`) vplivata temperatura zraka (`tempr`) in onesnaženost s prašnimi delci (`part`). Ob upoštevanju teh dveh spremenljivk nas zanima, ali je `cmort` odvisna tudi od časa (čas naj bo izražen v tednih: funkcija `time(cmort)` vrne vrednosti časa izražene v dvainpetdesetinah leta, od tako izraženega časa odštejemo 1970, da se znebimo velikih številk, `time(cmort)-1970`).

```
> library(astsa)
> str(lap)

 Time-Series [1:508, 1:11] from 1970 to 1980: 184 191 180 185 174 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:11] "tmort" "rmort" "cmort" "tempr" ...

> cas<-time(cmort)-1970
> head(cas)

[1] 0.00000000 0.01923077 0.03846154 0.05769231 0.07692308 0.09615385


> plot(cbind(Smrtnost=cmort, Temperatura=tempr, Delci=part))
```



Slika 49: Grafični prikaz časovnih vrst `cmort`, `tempr` in `part`

```
> pairs(cbind(Smrtnost=cmort, Temperatura=tempr, Delci=part))
```



Slika 50: Razsevni grafikoni za `cmort`, `tempr` in `part`

Najprej poiščite najustreznejši model za odvisnost `cmort` od `cas`, `tempr` in `part`. Obrazložite izbiro in naredite diagnostiko modela.

Za izbrani model analizirajte avtokorelacijo ostankov in model ustrezno dopolnite z modeliranjem avtokorelacije napak. Obrazložite izbiro končnega modela.

Obrazložite končni model.