Praktična uporaba Bayesove statistike

Gregor Vavdi

3/4/2020

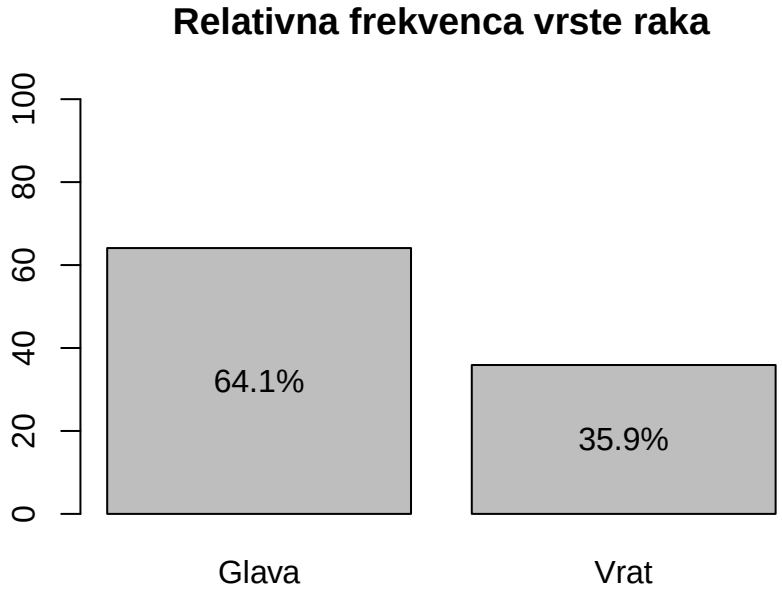
Podatki

Podatke sem pridobil za namene Statističnega svetovanja, ki ga opravljam na Onkološkem inštitu v Ljubljani. Raziskovalca zanima varnostni pas pri obsevanju z radioterapijo. Vsak pacient je deležen svojega obsevalnega načrta, ki zajema različno število frakcij (obsevanje). Teh frakcij je lahko do 35. Pri obsevanju se pacient postavi na mizo, kjer ga s pomočjo slik skalibirajo na teoretično pravilen položaj. Ker pa človek ni togo telo, se vseskozi premika (dihanje, napete mišice, itd.). Zato v ta namen gledajo premike v x, y in z smeri, ki so se zgodili v času ene frakcije od teoretične postavitve, ki bi jo moral pacient dosegati. Ti premiki po oseh določajo varnostni pas obsevanja, da pacientov tumor vseeno v celoti obsevan. Imenujemo jih interfrakcijski razmiki.

V prvem delu se bom osredotočil kaj vpliva na translacijske premike po y-osi (`Lng`). Neodvisne spremenljivke, ki jih bom vključil v model sta: vrsta raka in število frakcij. Model se mi zdi smiseln, saj me zanima ali vrsta raka dejansko pomeni večje translacijske premike pri radioterapiji (pri kateri vrsti raka, se pacienti bolj premikajo) in ali število obsevanj vpliva na napako. Translacij `Vrt` in `Lat`, ter rotacije v model nisem dodajal v model, saj ne gre za neodvisne spremenljivke. Predpostavljam, da se pacient ne more premakniti samo po eni osi.

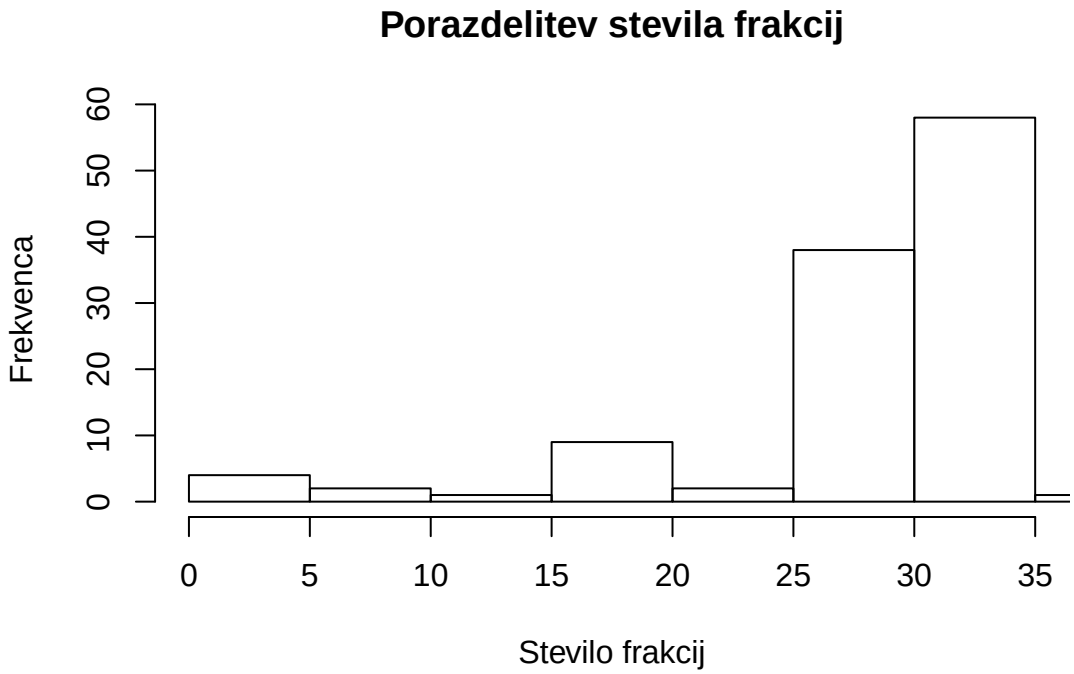
V1	AnonId	PlanId	RefIso	TreatDate	Vrt	Lng	Lat	Rtn
0	Patient000	Plan1	Glava	2014-03-25	-0.5	0.0	0.5	1.0
1	Patient000	Plan1	Glava	2014-03-26	-0.4	-0.1	0.2	0.0
2	Patient000	Plan1	Glava	2014-03-27	0.0	0.3	0.2	0.0
3	Patient000	Plan1	Glava	2014-03-28	-0.2	0.1	0.4	0.1
4	Patient000	Plan1	Glava	2014-03-31	-0.1	0.1	0.3	0.0
5	Patient000	Plan1	Glava	2014-04-01	-0.1	0.1	0.3	0.0

Vrsta raka



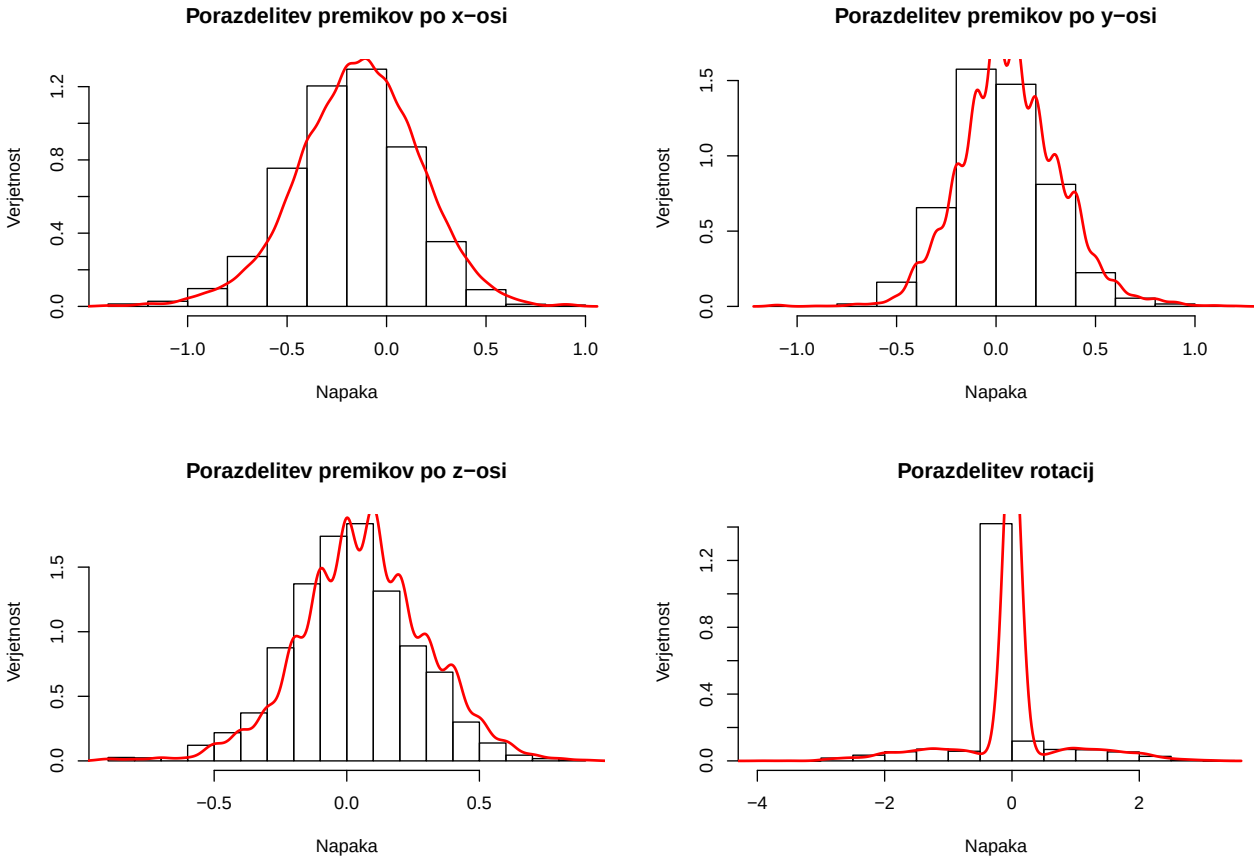
V podatkovju imamo 115 pacientov, ki je skupaj opravilo 3393 obsevanj z radioterapijo. 64 % je imelo raka v glavi, ostali pa na vratu. Obsevanje je potekalo od septembra 2012 do marca 2015.

Število frakcij



Za porazdelitev števila frakcij med pacienti, ki velja za ključno v mojem problemu, je na vzorcu vidna ena velika skupina, ki obsega 84 % pacientov, ki ima med 25 in 35 frakcij. Skoraj 10 % pacientov ima število obsevanj med 15 in 25, medtem ko ima le 6 % pacientov od 1 do 15 obsevanj.

Translacije in rotacije pacientov



	n	mean	sd	median	min	max	skew	kurtosis
Vrt	3393	-0.14	0.30	-0.1	-1.4	0.9	-0.20	0.44
Lng	3393	0.07	0.25	0.1	-1.1	1.2	0.19	0.95
Lat	3393	0.06	0.24	0.1	-0.9	0.9	-0.10	0.57
Rtn	3393	-0.01	0.75	0.0	-3.9	3.2	-0.15	4.12

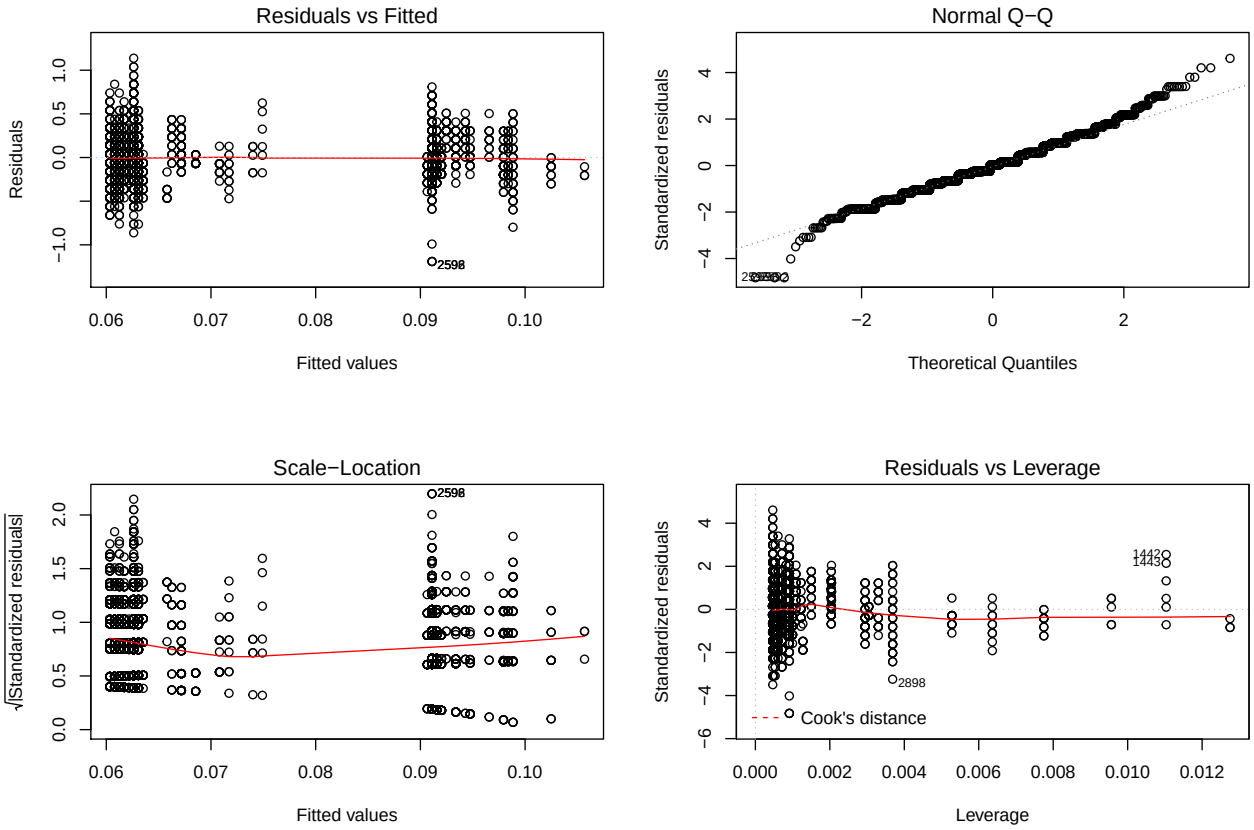
Porazdelitve po oseh so normalno porazdeljene s precej podobnim povprečji. Pri porazdelitvi rotacij ne moremo trditi, da je spremenljivka normalno porazdeljena, saj je prevelik del vrednosti okoli 0, ostale vrednosti pa so minimalno prisotne v negativno in pozitivno smer.

Frekventistični model

```
lm.mod <- lm(Lng ~ RefIso + st.frakcij, data = podatki.st.frakcij)
summary(lm.mod)
```

```
##
## Call:
## lm(formula = Lng ~ RefIso + st.frakcij, data = podatki.st.frakcij)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.19111 -0.16262  0.00844  0.13875  1.13738
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0762747  0.0286275   2.664 0.007750 **
## RefIsoVrat   0.0307666  0.0089736   3.429 0.000614 ***
## st.frakcij  -0.0004552  0.0009164  -0.497 0.619391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2468 on 3390 degrees of freedom
## Multiple R-squared:  0.003459,   Adjusted R-squared:  0.002871
## F-statistic: 5.883 on 2 and 3390 DF,  p-value: 0.002814
```



Predpostavka o konstantni varianci je izpolnjena, problematični so morda ostanki, ki ne kažejo, da so normalno porazdeljeni. Vseeno bom nadaljeval z analizo.

Pregledam še kolinearnost obeh spremenljivk in vidim, da kolinearnost ni prisotna.

```
kable(vif(lm.mod), "markdown", col.names = "VIF")
```

	VIF
RefIso	1.03184
st.frakcij	1.03184

Pregledali smo osnovne karakteristike linearnega modela, ki jih smatram, da jih moramo narediti, tudi če se odlčamo za Bayesovo statistiko.

Bayesev model

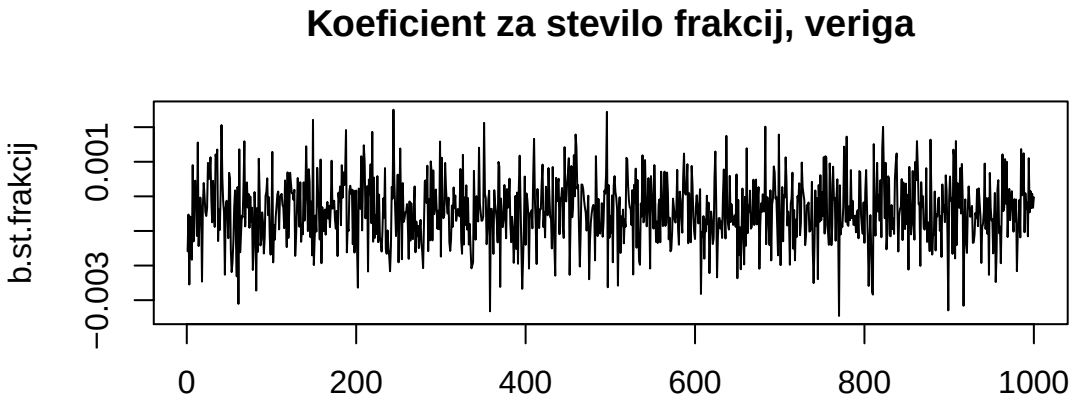
```
fit2.bayesx <- bayesx(Lng ~  RefIso + st.frakcij,
                      data = podatki.st.frakcij,
                      family = "gaussian", method = "MCMC")

b.refIso <- attr(fit2.bayesx$fixed.effects, "sample")[,2]
b.st.frakcij <- attr(fit2.bayesx$fixed.effects, "sample")[,3]

summary(fit2.bayesx)
```

```
## Call:
## bayesx(formula = Lng ~ RefIso + st.frakcij, data = podatki.st.frakcij,
##     family = "gaussian", method = "MCMC")
##
## Fixed effects estimation results:
##
## Parametric coefficients:
##                Mean      Sd    2.5%     50%  97.5%
## (Intercept)  0.0753  0.0291  0.0165  0.0757 0.1350
## RefIsoVrat   0.0307  0.0088  0.0124  0.0305 0.0476
## st.frakcij  -0.0004  0.0009 -0.0023 -0.0004 0.0014
##
## Scale estimate:
##          Mean     Sd   2.5%    50% 97.5%
## Sigma2 0.0610 0.0015 0.0581 0.0609 0.064
##
## N = 3393  burnin = 2000  method = MCMC  family = gaussian
## iterations = 12000  step = 10
```

KOMETNAR

Konvergenca

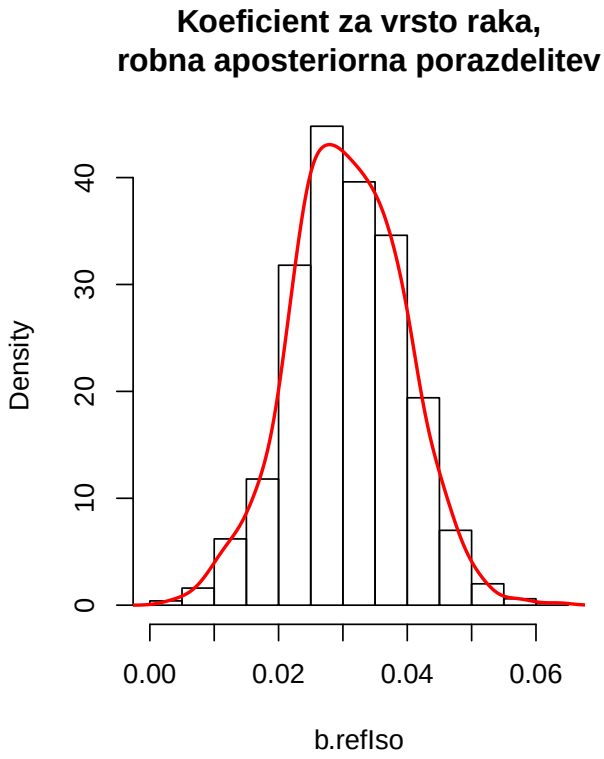
```
par(mfrow = c(2, 1))
plot(b.refIso, type = "l", main = "Koeficient za vrsto raka, veriga",
     xlab = "")
plot(b.st.frakcij, type = "l", main = "Koeficient za stevilo frakcij, veriga",
     xlab = "")
```

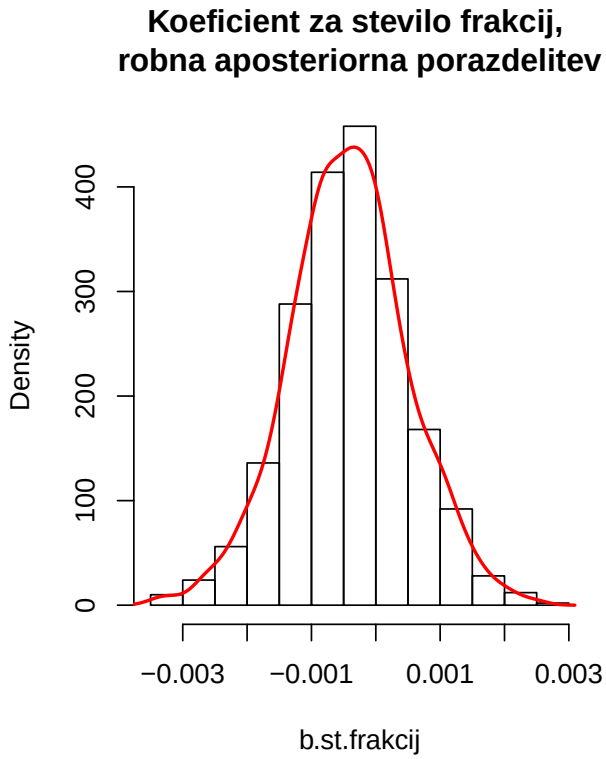


Konvergenca obeh parametrov se mi zdi vredu.

Interpretacija

```
par(mfrow = c(1, 2))
hist(b.refIso, prob = T, main = "Koeficient za vrsto raka,\nrobna aposteriorna porazdelitev")
lines(density(b.refIso), col = "red", lwd = 2)
hist(b.st.frakcij, prob = T, main = "Koeficient za stevilo frakcij, \nrobna aposteriorna porazdelitev")
lines(density(b.st.frakcij), col = "red", lwd = 2)
```





KOMENTAR

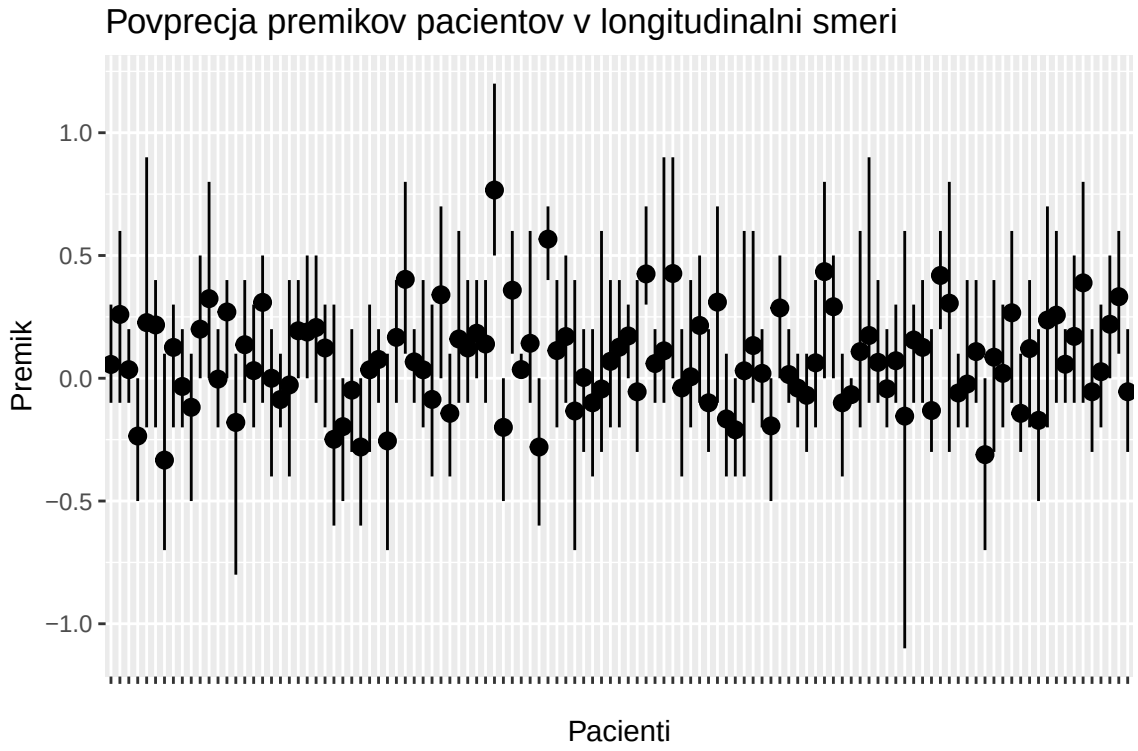
Zaključek

Hierarhični model

Hierarhičen model sem definiral na naslednji način. Zanimala me bo spremenljivka `Lng`, glede na paciente, in kako se razlikuje med njimi. Pri tem bom naredil pogumno predpostavko o tem, da je varianca med posameznimi pacienti enaka. Podatke imamo za 115 pacientov, vsak od njih pa ima do 35 merjenj.

```
pod.Lng <- dt %>%
  group_by(AnonId) %>%
  summarise(povprecje = mean(Lng), n=length(Lng), varianca = var(Lng))
```

```
ggplot(dt, aes(x = AnonId, y = Lng, group = AnonId))+
  stat_summary(fun.ymin = min, fun.ymax = max, fun.y = mean) +
  theme(axis.text.x =element_text(color = "white"))+
  labs(x = "Pacienti", y = "Premik" )+
  ggtitle("Povprecja premikov pacientov v longitudinalni smeri")
```




```
m <- length(pod.Lng$AnonId)
n <- pod.Lng$n
ime.unique <- levels(dt$AnonId)

xMatrix <- matrix(NA, ncol = m, nrow = max(n))
for (j in 1:m) {
  xMatrix[1:n[j],j] <- dt[dt$AnonId == ime.unique[j],]$Lng - mean(dt[dt$AnonId == ime.unique[j],]$Lng)
}
```

Določil sem tudi parametre hiperapriornih porazdelitev:

$$\begin{aligned}\sigma^2 &\sim \text{Inv-Gama}(\nu_0/2, \sigma_0^2\nu_0/2),\\ \mu &\sim \mathcal{N}(\mu_0, \tau_0^2),\\ \eta^2 &\sim \text{Inv-Gama}(\kappa_0/2, \eta_0^2\kappa_0/2).\end{aligned}$$

```
code <- nimbleCode({
  mu ~ dnorm(0, sd = 10);
  eta ~ dunif(0, 100)
  sigma ~ dunif(0, 100)

  for (j in 1:m) {
    muGroups[j] ~ dnorm(mu, sd = eta)
    for (i in 1:n[j]) {
      y[i, j] ~ dnorm(muGroups[j], sd = sigma);
    }
  }
})
```

Ker je želja, da bi bili premiki med opazovanjem čim manjši, kar pomeni, da je pričakovana vrednost radioterapije enaka 0. Za standardni odklon sem preizkusil več vrednosti, nato sem se odločil za 1. Med podatki ni nikoli vrednosti višje od 2, vednar sem želel biti previden in si nisem želel preveč omejevati.

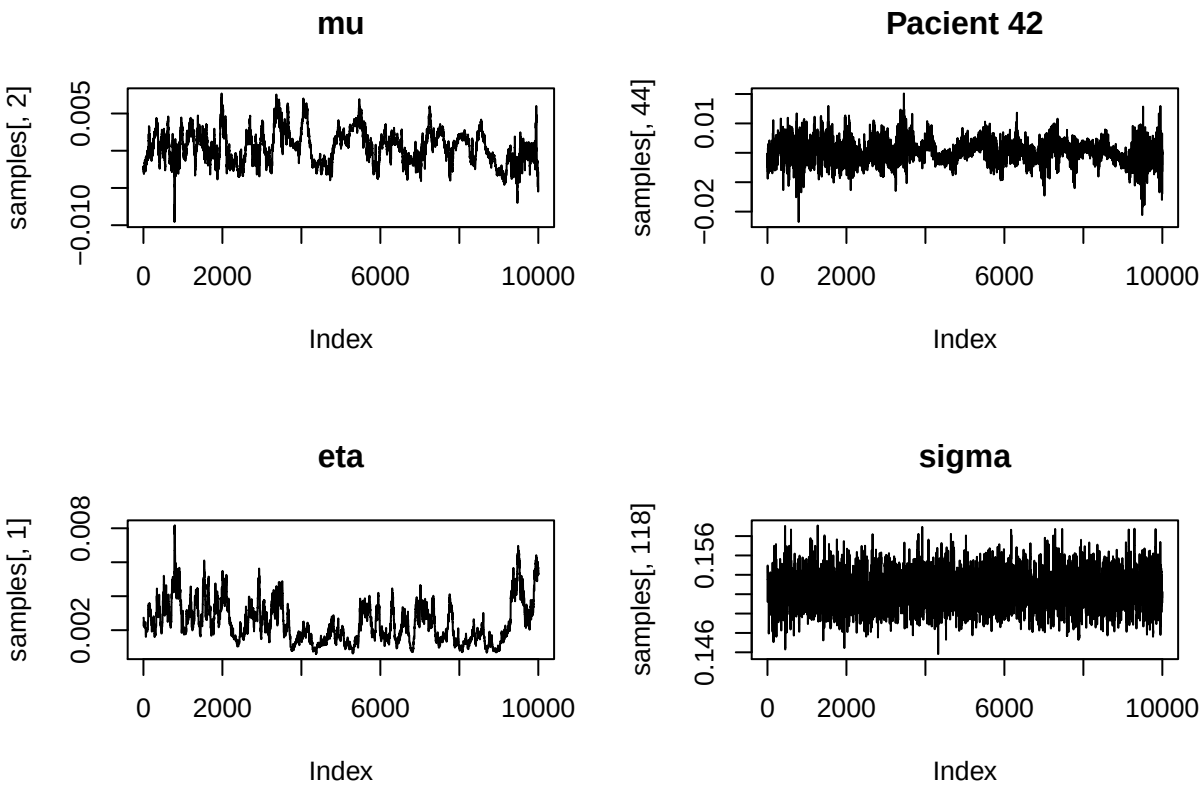
Parametra η in σ sem vzorčil iz enakomerne porazdelitve 0 - 10. Pri tem sem poskusil, tudi širše intervale, vendar so se mi tukaj rezultati zdeli najbolj optimalni.

```
constants <- list(m = m, n = n)

inits <- list(mu = mean(pod.Lng$povprecje),
              eta = sd(pod.Lng$povprecje),
              sigma = mean(sqrt(pod.Lng$varianca)),
              muGroups = pod.Lng$povprecje)

data <- list(y = xMatrix)

Rmodel <- nimbleModel(code = code, constants = constants,
                      inits = inits, data = data)
Rmodel$initializeInfo()

conf <- configureMCMC(Rmodel)
#conf$printSamplers()
#conf$printMonitors()
conf$addMonitors('muGroups')

Rmcmc <- buildMCMC(conf)
Cmodel <- compileNimble(Rmodel)
Cmcmc <- compileNimble(Rmcmc, project = Cmodel)
samples <- runMCMC(Cmcmc, niter = 12000, nburnin = 2000 )
#saveRDS(samples, "data/HieMod.RDS")
```

Konvergenca

Najprej sem preučil konvergenco parametrov in naključno izbranega pacienta.

```
par(mfrow = c(2,2))
plot(samples[,2], type = "l", main = "mu")
plot(samples[,44], type = "l", main = "Pacient 42")
plot(samples[,1], type = "l", main = "eta")
plot(samples[,118], type = "l", main = "sigma")
```

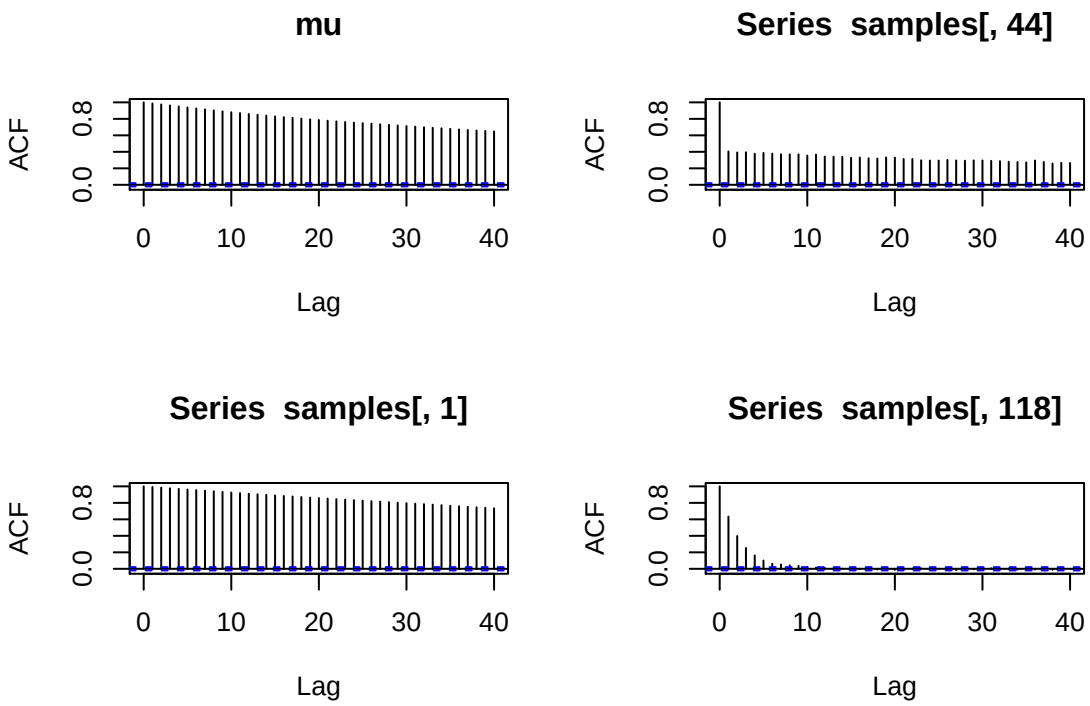


Z konvergenco izgleda vse ok, saj so vrednosti na y-osi dovolj male. Pozorni moramo biti, saj so vrednosti merjenja majhne in da ne pride do prevelikih odstopanj.

Thinning

Ker v MCMC verigah prevladuje visoka stopnja avtokorelacije, zato je potrebna analiza tudi v mojem primeru.

```
par(mfrow = c(2,2))
acf(samples[,2], main = "mu")
acf(samples[,44])
acf(samples[,1])
acf(samples[,118])
```



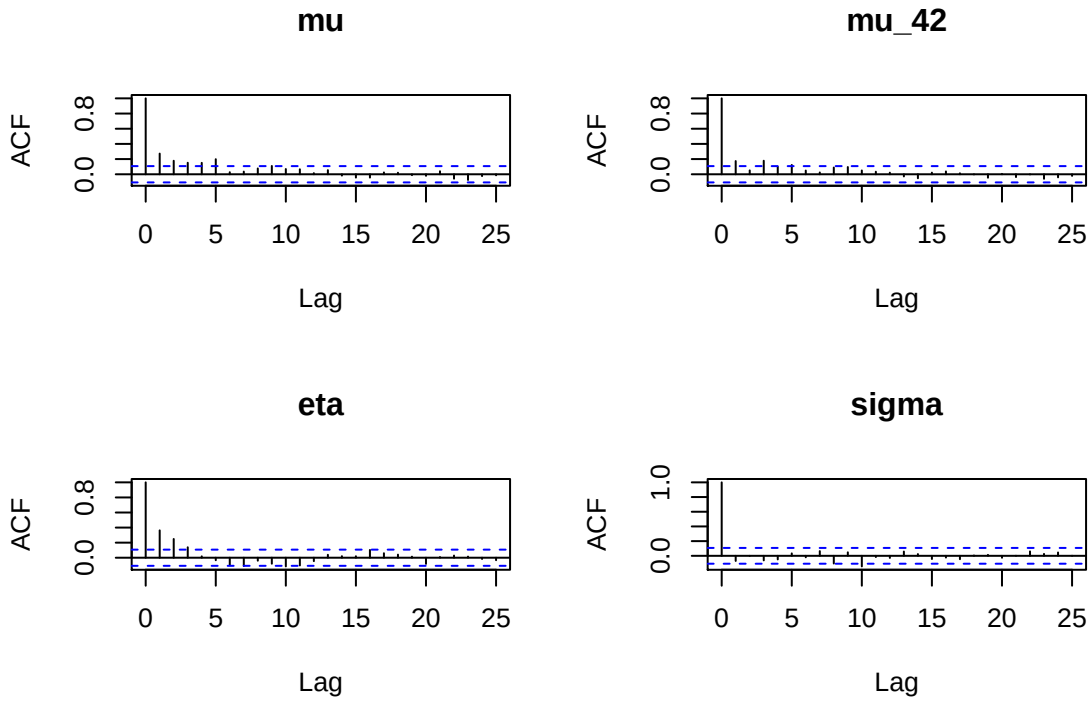
Kot vidimo so podatki v MCMC verigi visoko korelirani, zato moramo uporabiti thinning. Koliko vrednosti bomo spusitli vmes je odvisno od podatkov, zato sem to storil s poskušanjem. Na koncu sem se odločil za 200 in se s tem rešil avtokorelacije. (Ne vem ali je to prevelika številka v praksi in bi moral drugače postopati).

Na novo definiramo model in temu primerno povečamo število iteracij in burn-in.

```
samples.thinning <- samples <- runMCMC(Cmcmc, niter = 120000, nburnin = 20000, thin =300)
#saveRDS(samples.thinning, "data/HieMod_thinning.RDS")
```

in najprej pogledamo thinning:

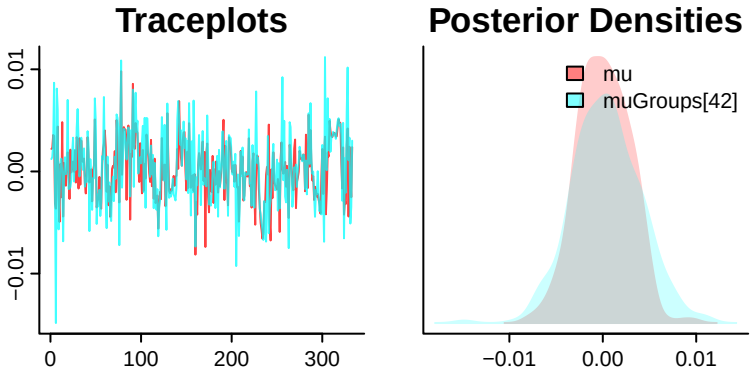
```
par(mfrow = c(2,2))
acf(samples.thinning[,2], main = "mu")
acf(samples.thinning[,44], main = "mu_42")
acf(samples.thinning[,1], main = "eta")
acf(samples.thinning[,118], main = "sigma")
```



Še vedno ni videti vredu, saj nekateri saj so ostanki po lagih še vedno večji od 95 % intervala zaupanja, ki je narisan s črtkano črto. Vseeno nadaljujem z analizo.

Še enkrat sem pogledal konvergenco za končni model (s thinningom):

```
samplesPlot(samples.thinning, var = c("mu","muGroups[42]"))
```

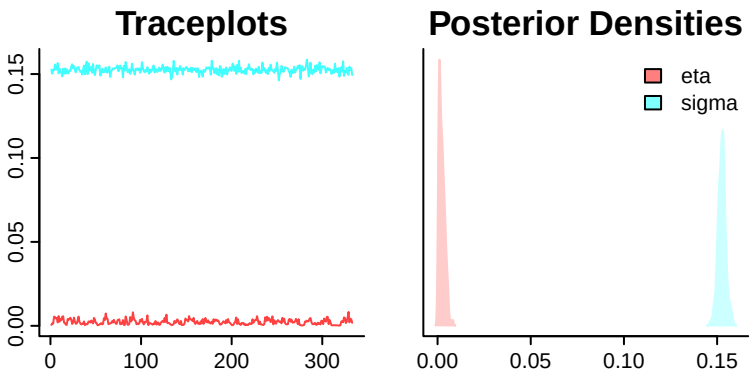



```
kable(samplesSummary(samples.thinning)[c(2, 44), ], "markdown")
```

	Mean	Median	St.Dev.	95%CI_low	95%CI_upp
mu	0.0001099	0.0000950	0.0027880	-0.0051606	0.0050117
muGroups[42]	0.0003875	0.0002561	0.0036447	-0.0066618	0.0075813

Glede na skalo, ki je na y-osi, bi rekel, da je konvergenca spremenljiva, čeprav graf od daleč zgleda da zelo variara.

```
samplesPlot(samples.thinning, var = c("eta","sigma"))
```




```
kable(samplesSummary(samples.thinning)[c(1, 118), ], "markdown")
```

	Mean	Median	St.Dev.	95%CI_low	95%CI_upp
eta	0.0023202	0.0019823	0.0015631	0.0002947	0.0056823
sigma	0.1525250	0.1525196	0.0020404	0.1483986	0.1569827

Za parametra σ in η je konvergenca vredu. Posteriorne porazdelitve η dosegajo vrednosti zelo blizu, medtem ko porazdelitev parametra σ malce odmaknjena od 0, s povprečjem 0.15.

Efective size in standardna napaka

```
efektivni.vzorec<- effectiveSize(samples.thinning)
moj.efect.vzorec <- efektivni.vzorec[c(1,2,44,118)]
sd.vzorec <- apply(samples.thinning[,c(1,2,44,118)], 2, sd)
standardne.napake <- sapply(1:4, function(i){sd.vzorec[i]/sqrt(moj.efect.vzorec[i])})

kable(data.frame("Effective size" = moj.efect.vzorec, "Standardna napaka" = standardne.napake), "markdo
```

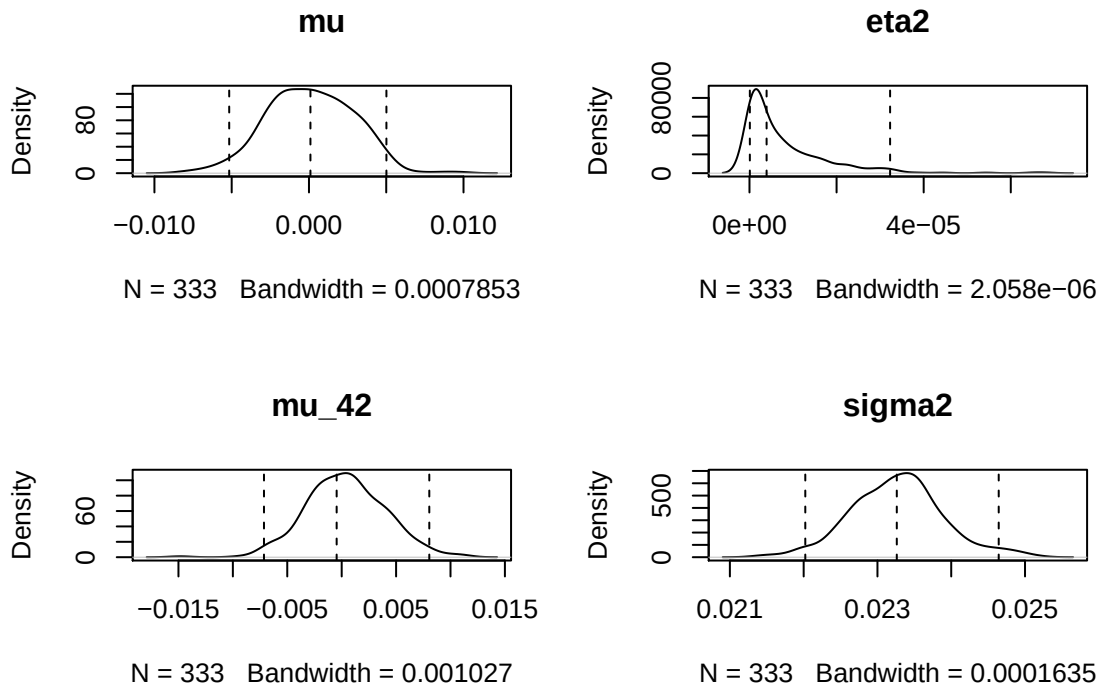
	Effective.size	Standardna.napaka
eta	117.89036	0.0001440
mu	98.10931	0.0002815
muGroups[42]	157.32631	0.0002906
sigma	333.00000	0.0001118

Efektivna velikost vzorca se giblje okoli 100 za posameznega pacienta. Za hiper parameter σ pa okoli 333. Standardne napake so za vse parametre zelo majhne, kar je dober znak za model.

Marginalne aposteriorne porazdelitve

```
par(mfrow=c(2, 2))
plot(density(samples.thinning[ , 2]), type = "l", main = "mu")
abline(v = quantile(samples.thinning[ , 2], prob=c(0.025, 0.5, 0.975)), lty = 2)
plot(density(samples.thinning[ , 1]**2), type = "l", main = "eta2")
abline(v = quantile(samples.thinning[ , 1]**2, prob=c(0.025, 0.5, 0.975)), lty = 2)
plot(density(samples.thinning[ , 44]), type = "l", main = "mu_42")
abline(v = quantile(samples.thinning[ , 3], prob=c(0.025, 0.5, 0.975)), lty = 2)
```

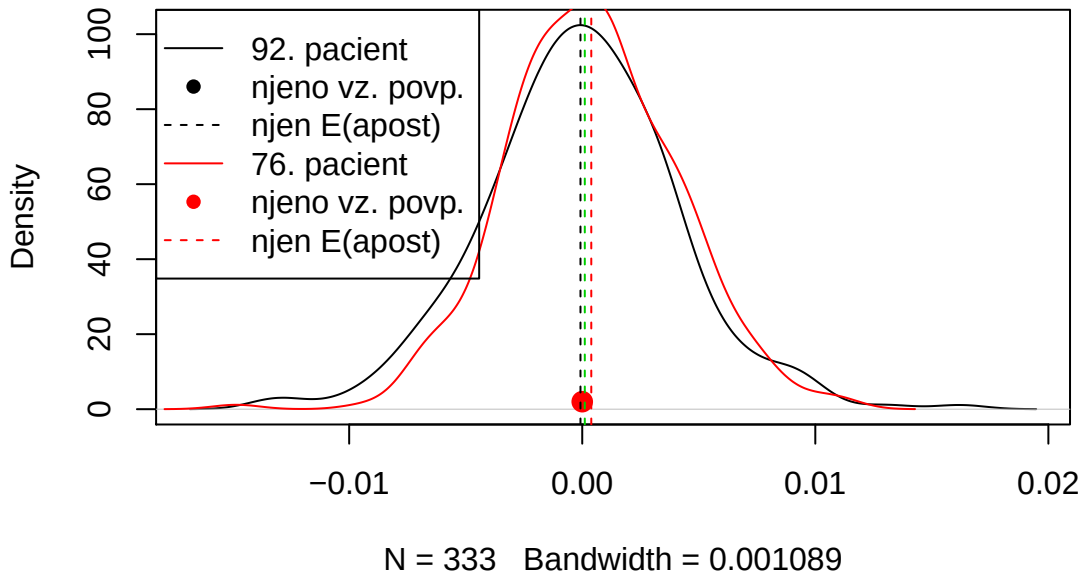
```
plot(density(samples.thinning[ , 118]**2), type = "l", main = "sigma2")
abline(v = quantile(samples.thinning[ , 118]**2, prob=c(0.025, 0.5, 0.975)), lty = 2)
```



Model za parameter μ daje zelo optimistične napovedi, saj pravi, da bo skupno povprečje vseh pacientov znašalo zelo blizu 0. Paramater η je zelo blizu 0. 95 % interval zaupnja za σ^2 je 0.022 in 0.248, kar kaže na to, da bodo odstopanja od povprečja zelo majhne pri vseh pacientih v modelu. Tako se izkaže tudi pri primeru enega od pacientov, ki ima povprečje pri 0.0003.

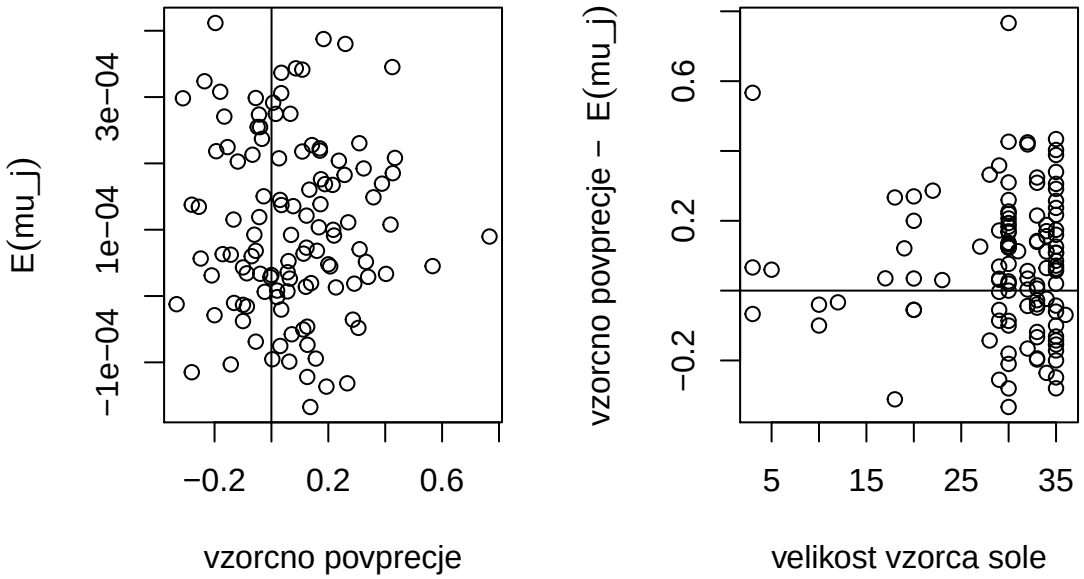
Posebaj sem pogledal primer za dva pacienta. Izbrana sta pacienta, ki sta imela največjo in najmanjšo razliko med vzorčnim povprečj in 0. To sta: največjo (76. pacient) in najmanjšo (92.pacient).

```
plot(density(samples.thinning[,19]), type="l", main="")
points(pod.Lng[19,]$povprecje, 2, pch=16, cex=1.5, col="red")
abline(v = mean(samples.thinning[,19]), lty=2)
lines(density(samples.thinning[,44]), type="l", col="red")
points(pod.Lng[44,]$povprecje, 2, pch=16, cex=1.5, col="red")
abline(v = mean(samples.thinning[,44]), lty=2, col="red")
abline(v = mean(samples.thinning[,2]), lty=2, col="green3")
legend("topleft", c("92. pacient", "njeno vz. povp.", "njen E(apost)",
                    "76. pacient", "njeno vz. povp.", "njen E(apost)"),
       col=c("black","black","black","red","red","red"), lty=c(1,NA,2,1,NA,2),
       pch=c(NA,16,NA,NA,16,NA))
```



Posteriorni porazdelitvi se nekoliko razlikjueta, njuna modelska povprečja $\mu_4 3$ in μ_6 omejujeta modelsko skupno poveprečje (zelena črtkana črta). Vzorčno povprečje za $\mu_4 3 = 0.7$ je tako daleč stran, modelske napovedi, da ju na grafu ni mogoče narisati.

```
pod.Lng$EMuGroup <- colMeans(samples.thinning[,3:117])

par(mfrow=c(1,2))
plot(pod.Lng$povprecje, pod.Lng$EMuGroup,
     xlab = "vzorcno povprecje", ylab = expression(E(mu_j)))
abline(a = 0, b = 1)

plot(pod.Lng$n, pod.Lng$povprecje - pod.Lng$EMuGroup,
     xlab = "velikost vzorca sole",
     ylab = expression(paste("vzorcno povprecje - "," ",E(mu_j), sep="")))
abline(h = 0)
```



Na levi sliki je predstavljeno pričakovana vrednost model za vsakega pacienta v primerjavi z vzorčnim povprečjem vsakega pacienta. Sam menim, da model ne predstavlja dobro podatkov, saj o ocenjuje napako

zelo blizu 0 (povečanje apriorne porazdelitve ne pomaga kaj dosti), vzorčna pa se raztezajo od -0.2 do 0.8. Črta predstavlja kako dobro se vzorčna povprečja ujemajo z modelom.

Desna slika predstavlja vpliv velikosti vzorca na razliko vzorčnege povepreča j-tega pacienta z njegovo pričakovano vrednostjo modela. Vidimo, da je število obsevanj ne vpliva na minimiziranje razlik med podatki in modelom. To je seveda logična posledica, saj so merjenja med seboj popolnoma neodivsna in več merjenj ne bo dalo manjših premikov pacienta. Po drugi strani pa pacienta med obsevanji večkrat slikajo in nato prilagodijo njegov novi položaj, kar posledično pomeni, da bi se napaka morala zmanjševat. Vendar podatkov o tem, kdaj mu na novo izračunajo položaj nimam.

Zakaj je prišlo do tega? Po mojem mnenju zato, ker sem model gradil na osnovni predpostavki, da so variance med pacienti enake. Mislim, da je to glavni razlog zakaj se modelske napovedi ne ujemajo z vzorčnimi povprečji.

Zaključek