Bayesova statisika

Seminarska naloga

Gregor Vavdi

2/23/2020

1 Primerjava frekventističnega in Bayesovega pristopa pri linearni regersiji

1.1 Cilj

Linearna regerisja je eno najbolj osnovnih statističnih orodij. V zadnjem času prihaja do porasta tudi v Bayesovi statistiki. V tej nalogi bom simuliral različne vrste podatkov, pri čemer me je zanimala predvsem razlika v klasičnem (frekvetističnem) prisotpu in Bayesovem pristopu. Podatke sem generiral z namenom, da me bosta zanimala predvsem kako vrsta slučajne napakez vpliva na pristopa in korelacija med podatki.

1.2 Opis simulacije

Simulacijo sem si zamislil na naslednji način. Najprej generiramo podatke, nato pa na istih podatkih apliciramo oba pristopa linearne regresije. To sem ponovil 1000x in nato pa akumuliral na različne statistike.

V podatkih sem spreminjal dva dejavnika:

1. Slučajno napako - spremenljivka s 4 različnimi vrstami napak: $N(0,1)$, $N(0,100)$, χ^2_1 in χ^2_4
2. Korelacija med spremenljivkami- spremenljivka s 3 različnimi vrednostmi: r = $\{0, 0.4, 0.8\}$. S tem sem preverajal ali korelacija vpliva na ocenjevanje bet in kako je pristop robusten na multikolinearnost spremenljivk.

Podatke sem generiral iz multivariatne normalne porazdelitve s povprečji: $\mu = \{3, 3, 4, 1\}$ in variančno - kovariančno matriko Σ:

$$\Sigma = \begin{bmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{bmatrix}$$

pri čemer je r predstavljal spreminjajoč dejavnik - korelacija. Poleg teh podatkov se dodal še eno binarno spremenljivko, ki je bila popolnoma neodvisna od zgornjih podatkov. Generiral sem jo iz $Ber(0.5)$. Velikost vzorca je bila 100.

V okvir simuliranja podatkov sodi tudi teoretične vrednosti za β. Uporabil sem naslednje vrednosti:

$$\beta_0 = -4.5, \quad \beta_1 = 0.8, \quad \beta_2 = 0.6, \quad \beta_3 = 3, \quad \beta_4 = 0, \quad \beta_5 = 1.4$$

Odločil sem se, da je začetna vrednost absolutno nekoliko višja. Močan efekt sem dal spremenljivkama: X_5, X_3, šibek efekt: X_1, X_2 in spremenljivki X_4, nisem dodelil efekta.

Na takšen način sem definiral spremenljivko Y, ki pa sem ji dodal še slučajno napako, ki pa je slučajni dejavnik in ga je bilo moč spreminjati na 4 različnih stopnjah. Vse skupaj sem zapakiral v funkcijo, ki je vidna spodaj:

```
gen.beta.data <- function(n = 100,r=0, napaka ){
  mu <- c(3, 3, 4, 1)
  Sigma <- rbind(c(1, r, r, r),
                 c(r, 1.0, r, r),
                 c(r, r, 1.0, r),
                 c(r, r, r, 1.0))
  podatki <- mvrnorm(n = n, mu = mu, Sigma = Sigma)
  #dodam še binarno spremenljivko
  podatki <- cbind(podatki, rbinom(n, size =0:1, prob=0.5))
  colnames(podatki) <- paste("X", 1:(length(mu)+1), sep="")

  # generiramo ciljno spremenljivko ("odvisno spremenljivko")
  b0 <- -4.5
  b1 <- 0.8
  b2 <- 0.6
  b3 <- 3
  b4 <- 0
  b5 <- 1.4

  y <- b0 + podatki[, "X1"]*b1 + podatki[, "X2"]*b2 +
    podatki[, "X3"]*b3 + podatki[, "X4"]*b4 + podatki[, "X5"]*b5 + napaka
  # zdruzimo podatke
  # data.frame popravi konverzijo formata
  podatki <- data.frame(cbind(podatki, y))
  return(podatki)
}
```

V frekventističnem pristopu sem uporabljal funkcijo `lm`, v Bayesovem pristopu pa sem uporabljal `bayesx` iz paketa `R2BayesX`.

Pri vsaki kombinaciji spreminjajočih se dejavnikov sem shranil podatke o ocenjenih parametrih bet in standardne napake ocen.

1.3 Simulacija

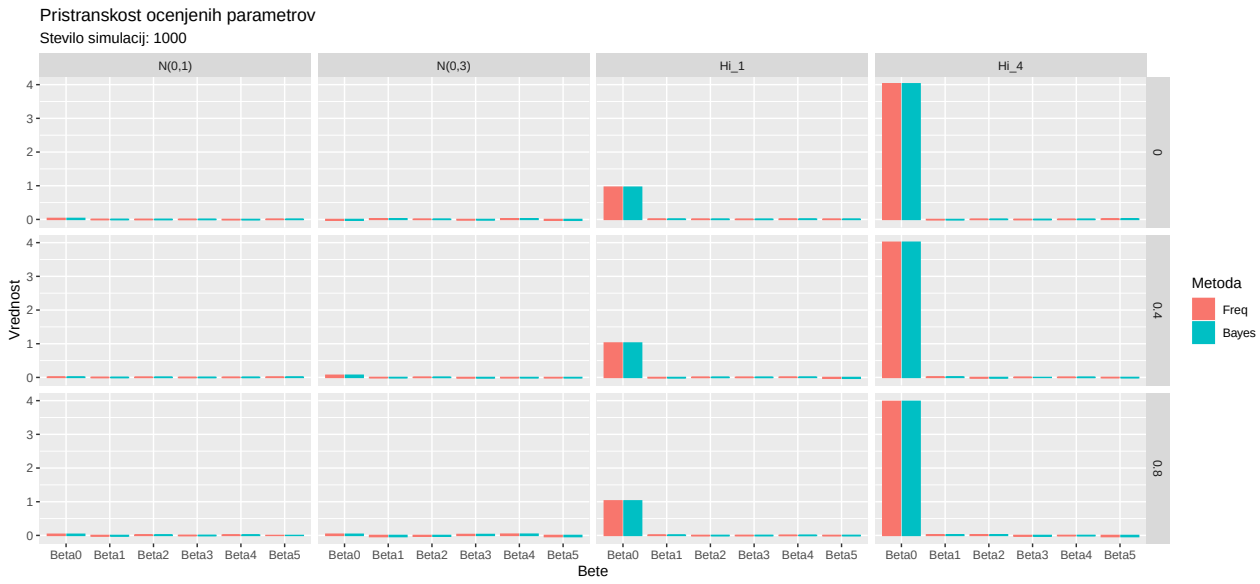
```
pon <- 1000
sampleSize <- 100
vrsta.napake <- c("N(0,1)", "N(0,3)", "Hi_1", "Hi_4")
korelacije <- c(0, 0.4, 0.8)
metoda <- c("Freq", "Bayes")
zasnova <- expand.grid(metoda, vrsta.napake,korelacije)
zasnova.lm <- do.call(rbind, replicate(pon, zasnova, simplify=FALSE)) %>%
  `colnames<-`(c("Metoda","Napaka", "Korelacija"))
zasnova.osnova <- do.call(rbind, replicate(pon,
                                           expand.grid(vrsta.napake,korelacije),
                                           simplify=FALSE)) %>%
  `colnames<-`(c("Napaka",  "Korelacija"))
skupaj.df <- as.data.frame(matrix(NA, nrow = nrow(zasnova.lm), ncol = 9))
table.sd <- as.data.frame(matrix(NA, nrow = nrow(zasnova.lm), ncol = 9))

start.time <- Sys.time()
i <- 1
j <- 1
while(i <= nrow(zasnova.osnova)){
  napaka.i <- as.character(zasnova.osnova[i, "Napaka"])
  if(napaka.i == "N(0,1)"){
    error <- rnorm(sampleSize, 0, 1)
  }
  else if(napaka.i == "N(0,3)"){
    error <- rnorm(sampleSize, 0, 3)
  }
  else if(napaka.i == "Hi_1"){
    error <- rchisq(sampleSize, df = 1)
  }
  else if(napaka.i == "Hi_4"){
    error <- rchisq(sampleSize, df = 4)
  }
  r.i <- zasnova.osnova[i, "Korelacija"]
  #Generiramo podakte
  data.i <- gen.beta.data(n = sampleSize, r= r.i, napaka = error)
  #FREKVENTISTIČNI PRISTOP####
  lm.model <- lm(y~., data = data.i)
  bete.hat <- summary(lm.model)$coef[,1]
  sd.bet <- summary(lm.model)$coef[,2]
  #BAYESOV PRISTOP#####
  bayesx.norm <- bayesx(y ~ X1+X2+X3+X4+X5 , data = data.i,
                        family = "gaussian", method = "MCMC")
```

```
  bayesx.mu <- attr(bayesx.norm$fixed.effects, "sample") #za vsako beto posebaj
  bayes.mu.mean.bet <- apply(bayesx.mu, 2, mean)
  bayes.mu.sd.bet <- apply(bayesx.mu, 2, sd)
  ###SHRANIMO REZULTATE#####
  index.i <- which(zasnova.lm$Napaka == napaka.i &
                     zasnova.lm$Korelacija==r.i& zasnova.lm$Metoda == "Freq")
  index.i.bay <- which(zasnova.lm$Napaka == napaka.i &
                         zasnova.lm$Korelacija==r.i& zasnova.lm$Metoda == "Bayes")
  skupaj.df[j, ] <- cbind(zasnova.lm[index.i.bay, ],
                          "Beta0" = bayes.mu.mean.bet[1],
                          "Beta1" = bayes.mu.mean.bet[2],
                          "Beta2" = bayes.mu.mean.bet[3],
                          "Beta3" = bayes.mu.mean.bet[4],
                          "Beta4" = bayes.mu.mean.bet[5],
                          "Beta5" = bayes.mu.mean.bet[6])
  table.sd[j, ] <- cbind(zasnova.lm[index.i.bay, ],
                         "Beta0" = bayes.mu.sd.bet[1],
                         "Beta1" = bayes.mu.sd.bet[2],
                         "Beta2" = bayes.mu.sd.bet[3],
                         "Beta3" = bayes.mu.sd.bet[4],
                         "Beta4" = bayes.mu.sd.bet[5],
                         "Beta5" = bayes.mu.sd.bet[6])
  skupaj.df[(j+1), ] <- cbind(zasnova.lm[index.i, ],
                          "Beta0" = bete.hat[1],
                          "Beta1" = bete.hat[2],
                          "Beta2" = bete.hat[3],
                          "Beta3" = bete.hat[4],
                          "Beta4" = bete.hat[5],
                          "Beta5" = bete.hat[6])
  table.sd[(j+1), ] <- cbind(zasnova.lm[index.i, ],
                             "Beta0" = sd.bet[1],
                             "Beta1" = sd.bet[2],
                             "Beta2" = sd.bet[3],
                             "Beta3" = sd.bet[4],
                             "Beta4" = sd.bet[5],
                             "Beta5" = sd.bet[6])
  j <- j + 2
  i <- i + 1
}
```

1.4 Rezultati

1.4.1 Pristranskost

```
#Podatki
bete.mean.est <- readRDS("data/simulation_I_bete_mean.RDS")
bete.sd.est <-readRDS("data/simulation_I_bete_sd.RDS")

original.beta <- list("Beta0" = -4.5, "Beta1"= 0.8 ,"Beta2" = 0.6,
                      "Beta3" = 3, "Beta4" = 0, "Beta5" = 1.4)

grupiraj.podatke <- bete.mean.est %>%
  group_by(Korelacija, Metoda, Napaka) %>%
  summarise_all(mean)

#Pristranskost
bias.data <- grupiraj.podatke %>%
  mutate_each(funs(. - original.beta$.), starts_with("Beta")) %>%
  gather(key = "Bete", "Vrednost", -c(Korelacija, Metoda, Napaka))
```



Pri standardno normalni napaki lahko vidimo, da je obe metode zelo dobro opravita nalogo in pristranskosti ni videti. Prav tako, pa tudi povečanje korelacije ne vpliva na pristranskost obeh metod. Tudi ko napaki povečamo standardni odklon na 3, se ne spremeni nič. (V eni od simulacij sem poskusil tudi standardni odklon povečati na 100 in videl, da metodi nista tako stabilni).

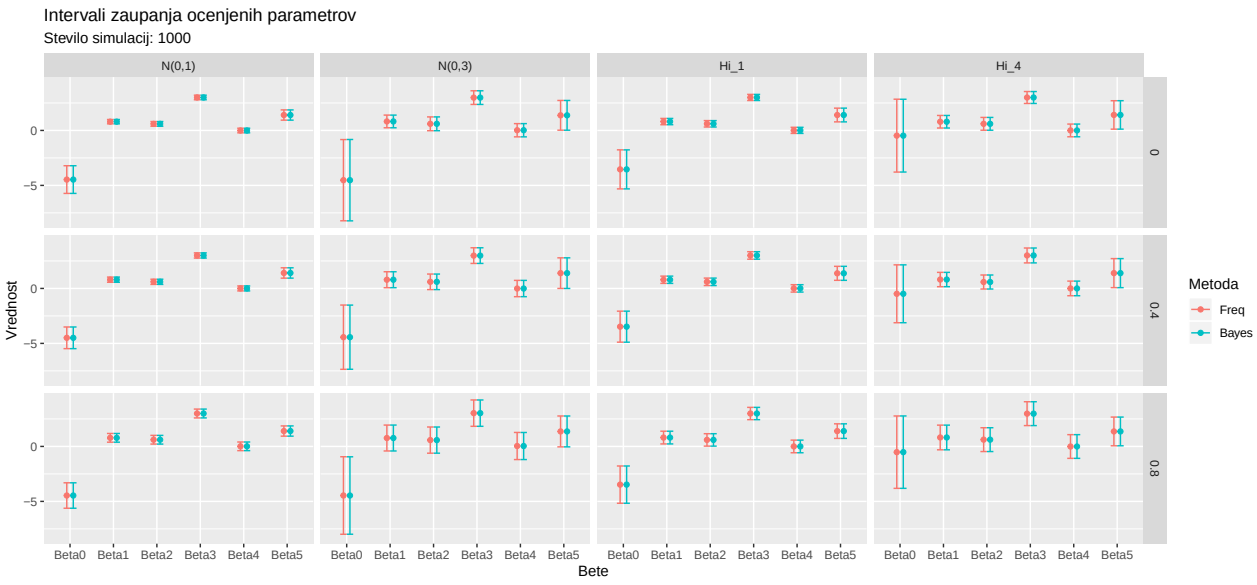
Pri napakah generiranih s χ_2 vidimo, da s povečevanjem parametra povečujemo pristranskost začetne vrednosti - s povečevanjem parametra precenjujemo β_0. Pravtako, pa korelacija ne vpliva na pristranskost metod.

Ko primerjamo obe metodi ocenjevanja parameterov dobimo zanimive in vsaj malo zaskr-

bljujoče podatke. Metodi sta si v vseh scenarijih praktično enako dobri. Manjša odstopanja so pri napaki N(0,100). V poročilu je nisem izbral, ker sem imel sum, da gre za preveliko napako,saj imamo povprečja med 1-10, z standardnim odklonom 100 pa je lahko dobimo zelo naključne rezultate.

1.4.2 Stanardna napaka

```
grupiraj.podatke.sd <- bete.mean.est %>%
  group_by(Korelacija, Metoda, Napaka) %>%
  summarise_all(sd) %>%
  gather(key = "Bete", "SE", -c(Korelacija, Metoda, Napaka))

grupiraj.podatke.mean <- bete.mean.est %>%
  group_by(Korelacija, Metoda, Napaka) %>%
  summarise_all(mean)%>%
  gather(key = "Bete", "Mean", -c(Korelacija, Metoda, Napaka))

podatki.se.mean.join <- grupiraj.podatke.mean %>% left_join(grupiraj.podatke.sd,by = c('
```



Standardne napake narejene na simulacije sem se odločil predstaviti v obliki intervalov zaupanja po normalni porazdelitve, saj sem predpostavil, da velja centralno limitni izrek. Na grafu so tako 95% intrvali zaupanja za povprečne vrednosti koeficientov β pri različni tipih slučajnih napak (stolpci) in korelaciji spremenljivk v podatkih (vrstice).

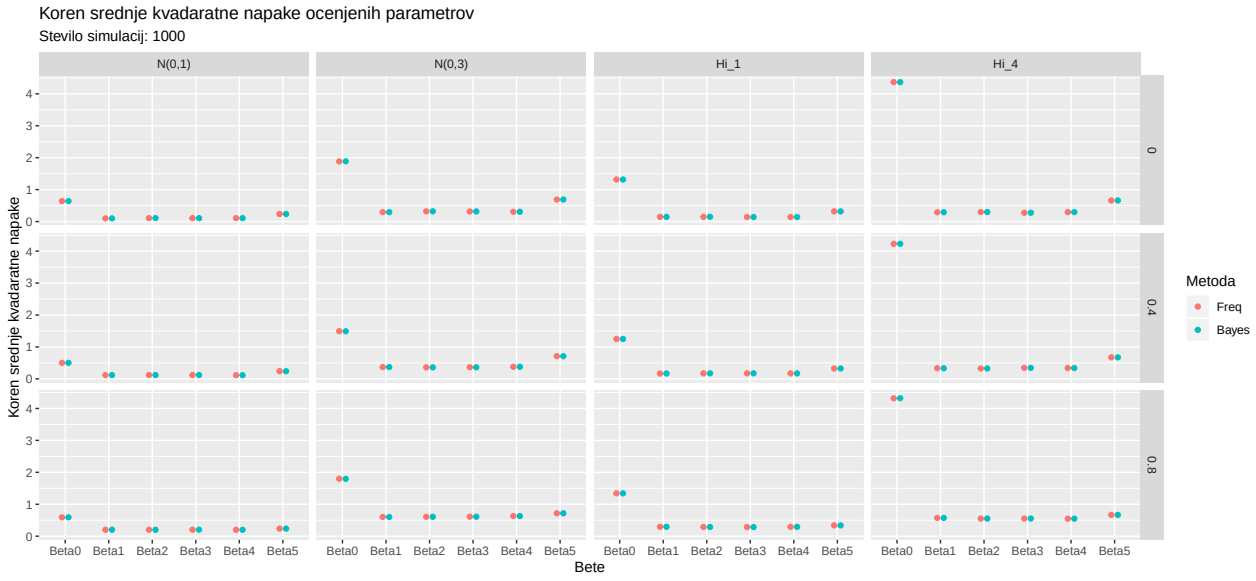
V prvem scenariju - slučajna napaka N(0,1), je opaziti, da so intervali izredno majhni pri vseh 4 povezanih spremenljivkah. Pri spremenljivki X_5 (dihotomna spremenljivka) je standardna napaka nekoliko večja kot pri ostalih. Največja standardna napaka je pri začetnem koeficientu β_0. S povečevanje korelacij oz. kolinearnosti med spremenljivkami X_1 do X_4 vidimo, da se intervali zaupanja nekoliko širšijo. Glede na razlike v prstopih razlik ni moč opaziti. S povečanjem standadnega odklona slučajne napaka, se intevali zaupanja nekoliko razširijo v

primerjavi s scenarijem 1. Še bolj pa k temu pripomore povečanje korelacijskega koeficienta. Standardna napaka je pri β_0 največja.

V drugem delu grafa imamo 2 scenarija χ^2 porazdelitve. Pri prvem imamo 1 stopnjo prostosti in se standardne napake obnašajo podobno kot pri slučajni napaki N(0,1). S povečanjem korelacij se povečuje tudi standardna napaka. Korelacijski koeficient ne vpliva na β_5, kar je pravilno. Pri 4 stopinjah prostosti pa so sicer standardne napake pri ostalih β majhne, vendar pri β_0 teoretična vrednost na zadane intervala zaupanja. Metodi (Bayes in Frekventistčni pristop) imata premajhne razlike, da bi bile opazne.

1.4.3 Koren srednje vrednosti

```
mean.square.error <- function(beta.original, beta.est){
  return(sum((beta.original - beta.est)^2) / length(beta.est))
}
bete.srednje <- bete.mean.est %>% group_by(Korelacija, Metoda, Napaka)%>%
  summarise(Beta0 = sqrt(mean.square.error(original.beta[[1]], Beta0)),
            Beta1 = sqrt(mean.square.error(original.beta[[2]], Beta1)),
            Beta2 = sqrt(mean.square.error(original.beta[[3]], Beta2)),
            Beta3 = sqrt(mean.square.error(original.beta[[4]], Beta3)),
            Beta4 = sqrt(mean.square.error(original.beta[[5]], Beta4)),
            Beta5 = sqrt(mean.square.error(original.beta[[6]], Beta5))) %>%
  gather(key = "Bete", "SrednjeVrednosti", -c(Korelacija, Metoda, Napaka))
```



Koren srednje kvadratne napake sem izračunal na naslednji način:

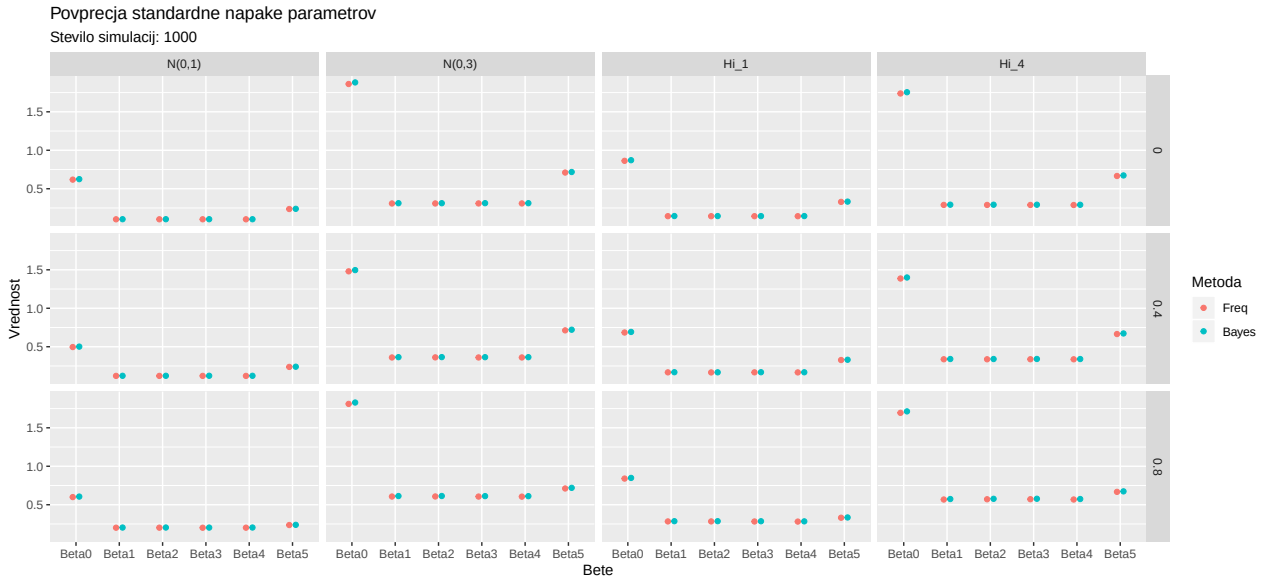
$$SMSE_{\beta_i} = \sqrt{\sum_{j=1}^{1000} (\hat{\beta_{ij}} - \beta_i)^2}$$

Tako sem za vsak koefeicient ter vsako kombinacijo simulacije izračunal SMSE in dobil zgornji

graf. Glavna opazka je, da razlik v pristopih ni in da je največja SMSE pri parametru $beta_0$, ta je še posebno visok pri scenariju z napako χ^2_4. Korelacija ne vpliva na napako SMSE. Pri scenariju z normalno slučajno se pozna vpliv povečanega standardnega odklona, kar se vidi na povečanjiu SMSE pri vseh parameterih. Vpliv neodvisne spremenljivke X_5 - dihotomna spremenljivka, je viden v odskoku vrednosti od ostalih pri scenariju N(0,3).

1.4.4 Povprečje ocenjenih standardnih napak koeficientov

```
grupiraj.podatke.sd <- bete.sd.est %>% group_by(Korelacija, Metoda, Napaka)%>%
  summarise_all(mean) %>%
  gather(key = "Bete", "Vrednost", -c(Korelacija, Metoda, Napaka))
```



Pričakovano podobno se obanašajo tudi povprelja standardnih napak ocenjenih β. Pri začetni vrednosti je povprečje standardne napake največeje in se spreminja z obliko slučajne napake. Pri scenariju N(0,3) se vidi nekolikošna razlika med pristopoma, saj ima Bayesov pristop nekoliko večjo povprečno standardno napako. Pri ostalih scenarijih tega iz grafa ni moč zaznati. Scenarija s χ^2 porazdelitvama imata podobne zaključke kot scenarija s normalno slučajno napako z razliko, da korelacija ne vpliva na velikost povepreči j. Ponovno se pozna efekt neodivsne dihotomne spremenljivke X_5, pri kateri je standardna napaka nekoliko večja kot pri ostalih spremenljivkah. Iz standardnih napak ni razvidno katera od spremenljivk naj bi imela visok ali zmeren efekt.

1.5 Zaključki

V simulaciji sem preveril nekatere statistike (pristranskost, MSME, standardna napaka) dveh različnih pristopov (Freq in Bayes) k ocenjevanju parametrov pri linearni regersiji, pri čemer sem imel dva spreminjajoča dejavnika: korelacijo med spremenljivkami (X_1 do X_4) in različne tipe slučajne napake. Razlik med frekventistčnim in Bayesovim pristopom ni bilo zaslediti. Sam menim, da je to zaradi tega, ker nismo v Bayesovem pristopu upoštevali

dodatnega podatka (prior). Različni tipi slučajnih napak so nakazali, da je to najbolj vpliva na začetno vrednost. S povečevanjem standardnega odklona slučajne napake pri normalni porazdelitvi sem ugotovil, da dokler je standardni odklon v okolici povprečji spremenljivk, linearna regerisja vredu deluje. S povečevanje korealcij oz. kolinearnosti pa povečujemo možnost napak. Pri slučajni napaki tipa χ^2 je opaziti, da korelacija ne vpliva na nobeno od statistik (pristranskost, SMSE, standardna napaka). Omeniti še velja, da pri neodivsni spremenljivki X_5, ki je binarna, SMSE in povprečje standardnih napak povečuje drugače od ostalih spremenljivk, ki so povezana s korelacijskim koeficientom.

2 Primerjava metod za izbor spremenljivk pri linearni regresiji

2.1 Opis

V drugem delu seminarske naloge, bom preveril več različnih metod za izbiro spremenljivk pri linearni regresiji. Predstavil bom 3 metode (metode sem zaradi lažje razumevanja pustil v angleškem zapisu). Prva metoda je *Reversible jump MCMC* s pomočjo programske knjižnice `nimble`. V drugi metodi imenovani Bootstrapped *Augmented Backward Elimination* z pomočjo programske knjižnice `abe`, sem s podobnimi simulacijami pregledoval deleže vrnjenih izbir spremenljivk za različne tipe slučajnih napak, korelacije in metriko (AIC, BIC, p-vrednost). Kot dodatno metoodo pa sem vpeljal še Spike - Slab metodo, pri kateri sem prav tako gledeal delež izbranih spremenljivk.

Podatke sem generiral na enak način kot v prvem delu. Dodal sem še en 5 spremenljivk, s povepreči $\mu = \{6, 17, 22, 12, 5)\}$ in enako variančno - kovariančno matriko kot v prvem delu. Efekt dodatnih spremenljivk sem dodal na 0, saj želimo dobiti bete, katere najbolj opišejo našo odvisno spremenljivko (Y). Vse spremenljivke X, razen binarne X_5 sem centriral, preden sem jo dodal v model.

$$\beta_0 = -4.5, \quad \beta_1 = 0.8, \quad \beta_2 = 0.6, \quad \beta_3 = 3, \quad \beta_4 = 0, \quad \beta_5 = 1.4, \quad \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$$

```
gen.beta.data.2 <- function(n = 100,r=0, napaka){
  mu <- c(3, 3, 4, 1)
  Sigma <- rbind(c(1, r, r, r),
                 c(r, 1.0, r, r),
                 c(r, r, 1.0, r),
                 c(r, r, r, 1.0))
  mu.brez.vpliva <- c(6, 17, 22, 12, 5)
  Sigma.brez.vpliva <- rbind(c(1, r, r, r, r),
                 c(r, 1.0, r, r, r),
                 c(r, r, 1.0, r, r),
                 c(r, r, r, 1.0, r),
                 c(r, r, r, r, 1.0))
```

```
  podatki <- mvrnorm(n = n, mu = mu, Sigma = Sigma)
  podatki <- cbind(podatki, rbinom(n, size =0:1, prob=0.5))
  podatki.brez.vpliva <-mvrnorm(n = n, mu = mu.brez.vpliva,
                                Sigma = Sigma.brez.vpliva)
  podatki <- cbind(podatki, podatki.brez.vpliva)
  colnames(podatki) <- paste("X", 1:ncol(podatki), sep="")
   # generiramo ciljno spremenljivko ("odvisno spremenljivko")
  b0 <- -4.5
  b1 <- 0.8
  b2 <- 0.6
  b3 <- 1.0
  b4 <- 0
  b5 <- 1.4
  b6 <- 0
  b7 <-0
  b8 <- 0
  b9 <-0
  b10 <- 0
  y <- b0 + podatki[, "X1"]*b1 + podatki[, "X2"]*b2 +
    podatki[, "X3"]*b3 + podatki[, "X4"]*b4 + podatki[, "X5"]*b5 +
    podatki[, "X6"]*b6 + podatki[, "X7"]*b7 + podatki[, "X8"]*b8 +
    podatki[, "X9"]*b9 + podatki[, "X10"]*b10 + napaka
  podatki <- data.frame(cbind(podatki,"Y"= y))
  return(podatki)
}
```

2.2 Reversible jump MCMC - NIMBLE

```
grafGG <- function(model, naslov = ""){
  shrani.rez <- as.data.frame(round(samplesSummary(model),2))
  shrani.rez$ime <- rownames(shrani.rez)
  z.izbrani <- shrani.rez %>%
    filter(grepl("z",ime)) %>%
    dplyr::select(Mean) %>%
    mutate(X = paste("X",1:10, sep =""))
  z.izbrani$X <- factor(z.izbrani$X, levels = paste("X", 1:10, sep = ""))
  gg <- ggplot(z.izbrani, aes(x = X, y = Mean)) + geom_bar(stat = "identity")+
    ggtitle(paste("Izbira spremenljivk pri", naslov))
  return(gg)
}

get.rjMCMC <- function(selectX, Y, p=10){
  codeSelect <- nimbleCode({
```

```
    sigma ~ dunif(0, 20)
    psi ~ dunif(0,1)
    beta0 ~ dnorm(0, sd=100)
    for(i in 1:p) {
      z[i] ~ dbern(psi) #indikator za vsak koeficient
      beta[i] ~ dnorm(0, sd = 100)
      zbeta[i] <- z[i] * beta[i]
    }
    for(i in 1:N) {
      y[i] ~ dnorm(beta0 + inprod(X[i, 1:p], zbeta[1:p]), sd = sigma)
    }
  })

  N <- dim(selectX)[1]
  p <- dim(selectX)[2]
  constantsSelect <- list(N = N, p = p)
  initsSelect <- list(sigma = 1, psi = 0.5, beta0 = 0,
                      beta = rnorm(p),
                      z = sample(c(0, 1), p, replace = TRUE))
  dataSelect <- list(y = Y, X = selectX)
  RmodelRJ <- nimbleModel(code = codeSelect, constants = constantsSelect, #isto
                          inits = initsSelect, data = dataSelect)
  confRJ <- configureMCMC(RmodelRJ) #isto
  confRJ$addMonitors('z') #isto
  configureRJ(confRJ,
              targetNodes = 'beta',
              indicatorNodes = 'z',
              control = list(mean = 0, scale = .2))
  RmcmcRJ <- buildMCMC(confRJ)
  CmodelRJ <- compileNimble(RmodelRJ)
  CmcmcRJ <- compileNimble(RmcmcRJ, project = CmodelRJ)
  samplesRJ <- runMCMC(CmcmcRJ, niter = 12000, nburnin = 2000)
  return(samplesRJ)}
```

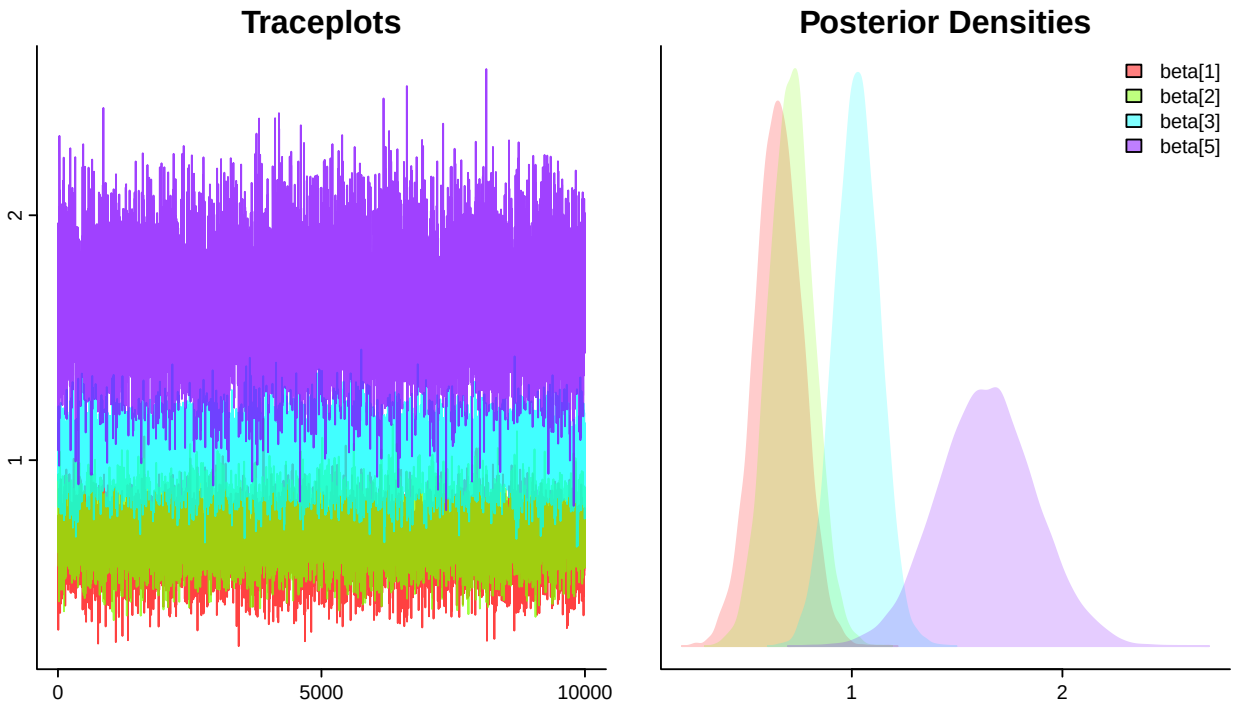


Figure 1: Konvergenca in posteriorne porazdelitve za neničelne bete

2.2.1 Rezultati

Za vsak scenarij bom pregledal konvergenco in posteriorno porazdelitev in pogledal katere spremenljivke bi dodali v model. Konvergenco in posteriorno porazdelitev sem narisal na dveh grafih - tisti, ki imajo ničelni koeficient in tiste, ki imajo neničelne koeficiente na drug graf.

2.2.1.1 Slučajna napaka: N(0,1)

```
data.NIBLE.N1 <- gen.beta.data.2(100, napaka = rnorm(100))
data.NIBLE.N1.selectY <- data.NIBLE.N1$Y
data.NIBLE.N1.selectX <- sweep(data.NIBLE.N1[,-5], 2, colMeans(data.NIBLE.N1[,-5]), FUN=
data.NIBLE.N1.selectX$X5 <- data.NIBLE.N1$X5
data.NIBLE.N1.selectX <- data.NIBLE.N1.selectX[, c(1,2,3,4,10,5,6,7,8,9)]
mod.N1 <- get.rjMCMC(data.NIBLE.N1.selectX, data.NIBLE.N1.selectY)
```

```
## thin = 1: sigma, psi, beta0, beta, z
## |-------------|-------------|-------------|-------------|
## |-----------------------------------------------------|
```

Konvergence so v skladu z pričakovanji in glede na grafičen prikaz za vse parametre model skonvergira. Posteriorne porazdelitve so pri neničelnih betah pravilno porazdelje s povprečji, ki se skladajo s teoretičnimi vrednostmi.

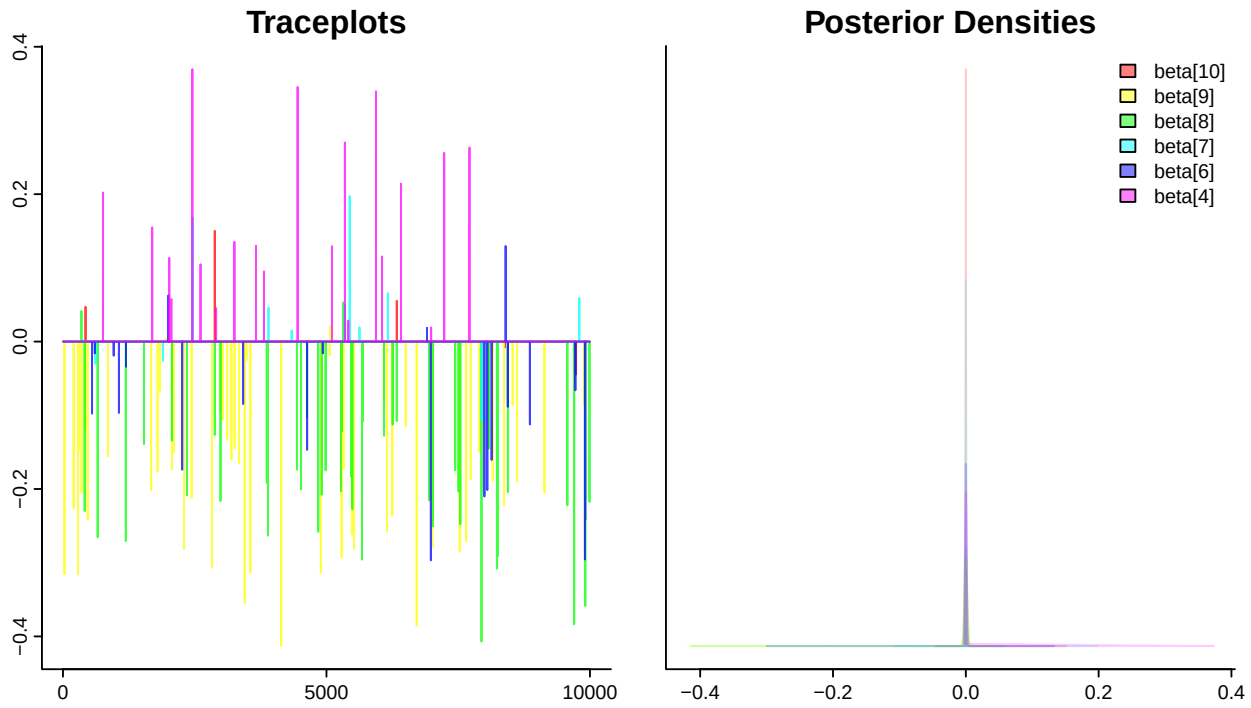
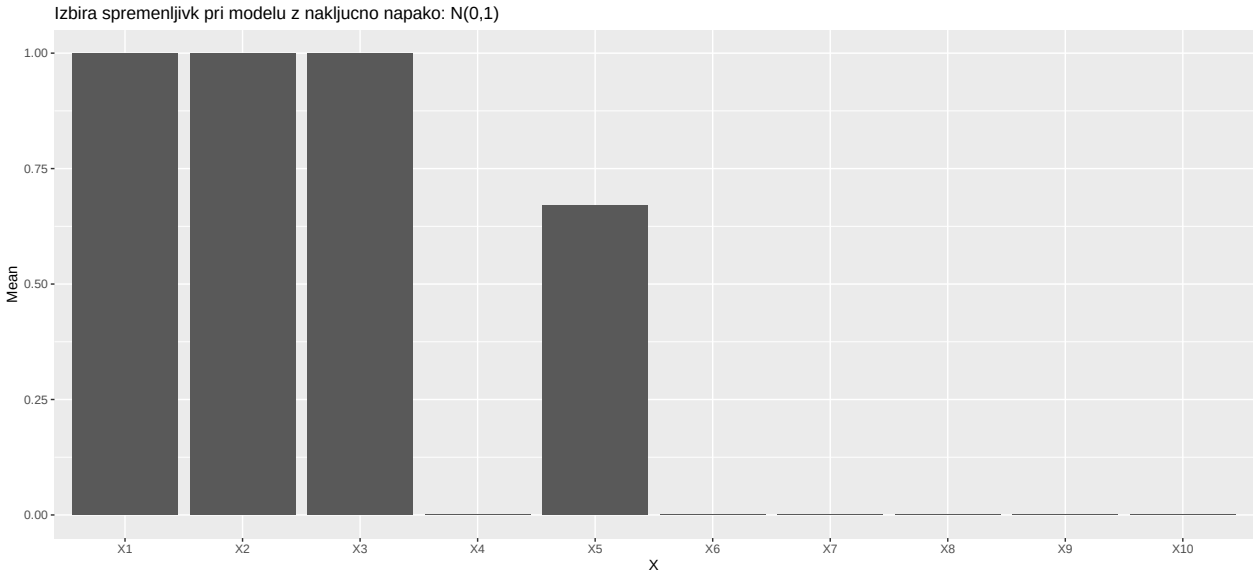


Figure 2: Konvergenca in posteriorne porazdelitve za ničelne bete



Model pravilno izbere vse 4 spremenljivke z zelo visokim deležem verjetja. Tudi ko sem poskusil s povečanjem korealacij, ga to ni zmotilo in se grafa nista razlikovala.

2.2.1.2 Slučajna napaka: N(0,3)

```
data.NIBLE.N3 <- gen.beta.data.2(100, napaka = rnorm(100))
data.NIBLE.N3.selectY <- data.NIBLE.N3$Y
data.NIBLE.N3.selectX <- sweep(data.NIBLE.N3[,-5], 2, colMeans(data.NIBLE.N3[,-5]), FUN=
```

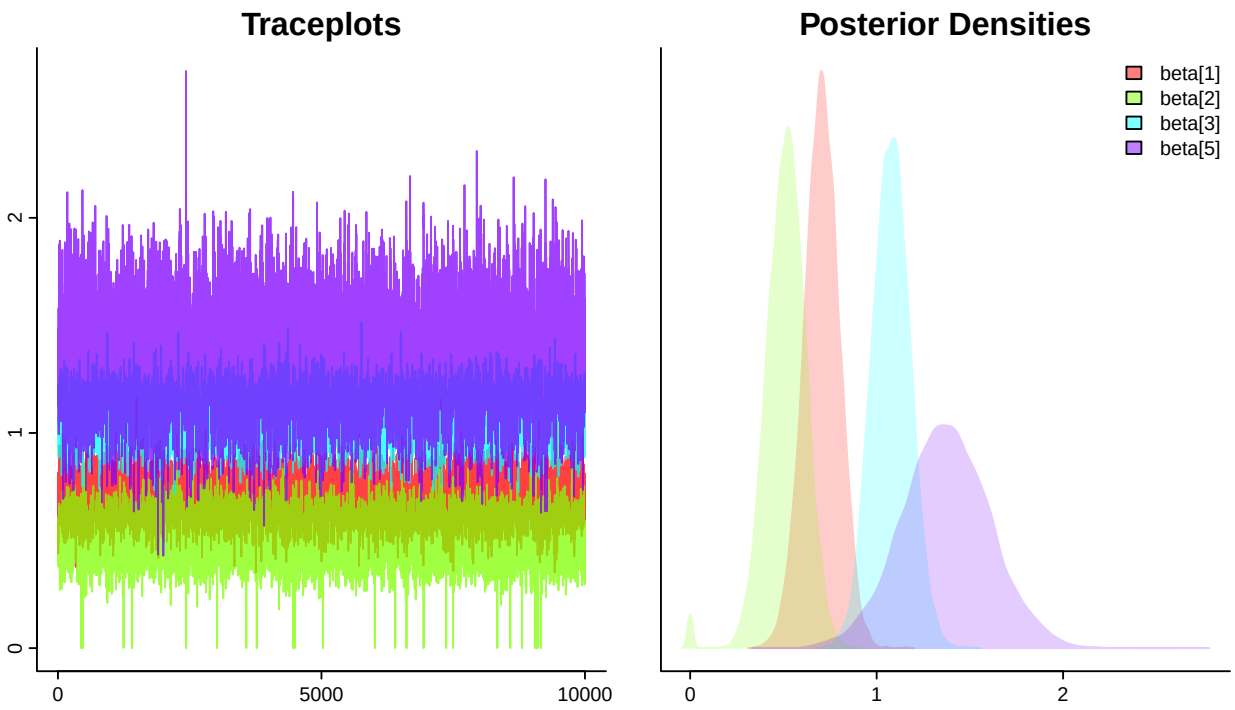
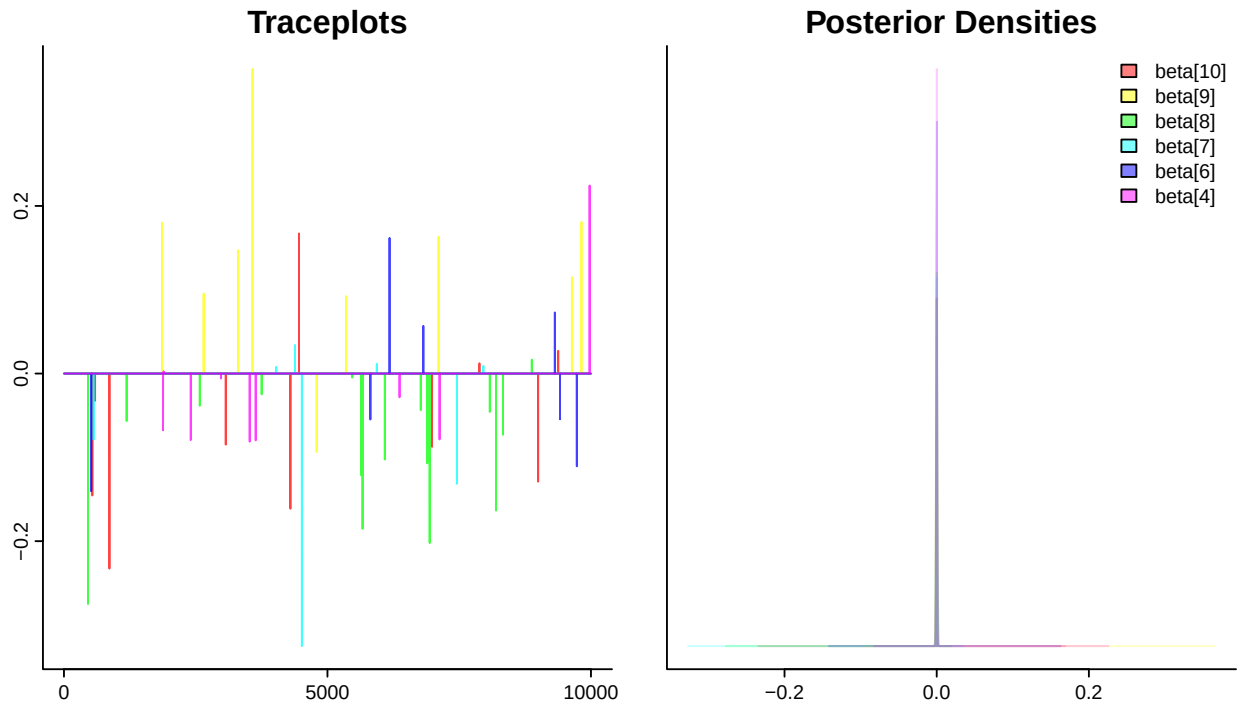


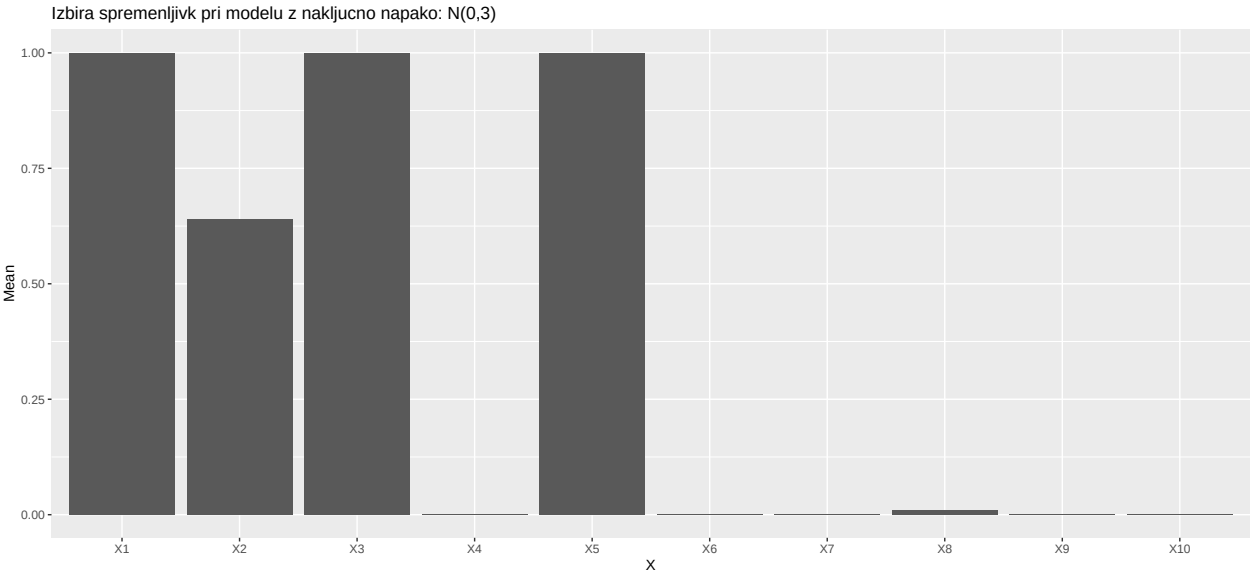
Figure 3: Konvergenca in posteriorne porazdelitve za neničelne bete

```
data.NIBLE.N3.selectX$X5 <- data.NIBLE.N3$X5
data.NIBLE.N3.selectX <- data.NIBLE.N3.selectX[, c(1,2,3,4,10,5,6,7,8,9)]
mod.N3 <- get.rjMCMC(data.NIBLE.N3.selectX, data.NIBLE.N3.selectY)
```

```
## thin = 1: sigma, psi, beta0, beta, z
## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|
```



Konvergence so za ničelne v skladu z pričakovanji, kot tudi posteriorne porazdelitve z zelo visokim verjetjem okoli 0. Težav s konvergenco ni zaslediti tudi pri neničelnih parametrih β.



S povečanjem standardnega odklona slučajne napake se pri izboru spremenljivk ni spremenilo. Izbrane spremenljivke so bile še vedno izbrane pravilno in z visoko vrednostjo verjetja.

2.2.1.3 Slučajna napaka: χ^2_1

```
data.NIBLE.H1 <- gen.beta.data.2(100, napaka = rchisq(100, 1))
data.NIBLE.H1.selectY <- data.NIBLE.H1$Y
data.NIBLE.H1.selectX <- sweep(data.NIBLE.H1[,-5], 2,
                               colMeans(data.NIBLE.H1[,-5]),
```

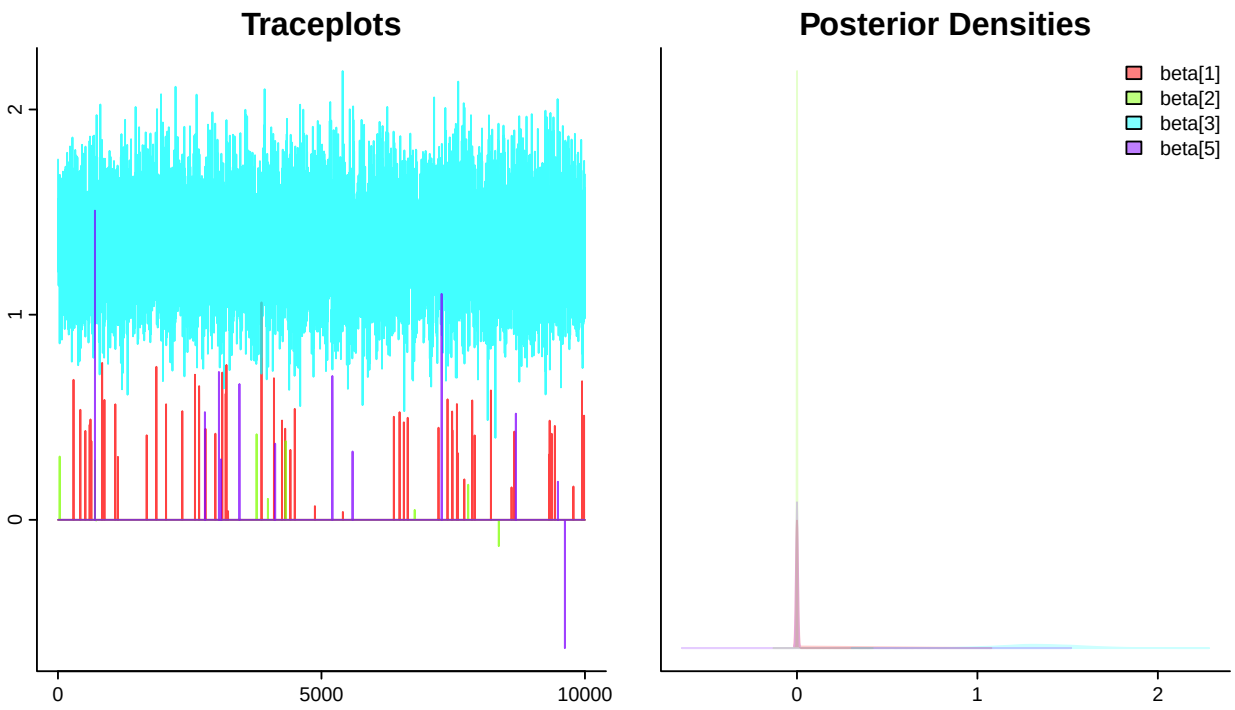


Figure 4: Konvergenca in posteriorne porazdelitve za neničelne bete

```
                                FUN="-")[,-10] #centriramo
data.NIBLE.H1.selectX$X5 <- data.NIBLE.H1$X5
data.NIBLE.H1.selectX <- data.NIBLE.H1.selectX[, c(1,2,3,4,10,5,6,7,8,9)]
mod.H1 <- get.rjMCMC(data.NIBLE.H1.selectX, data.NIBLE.H1.selectY)
```

```
## thin = 1: sigma, psi, beta0, beta, z
## |-------------|-------------|-------------|-------------|
## |-----------------------------------------------------|
```

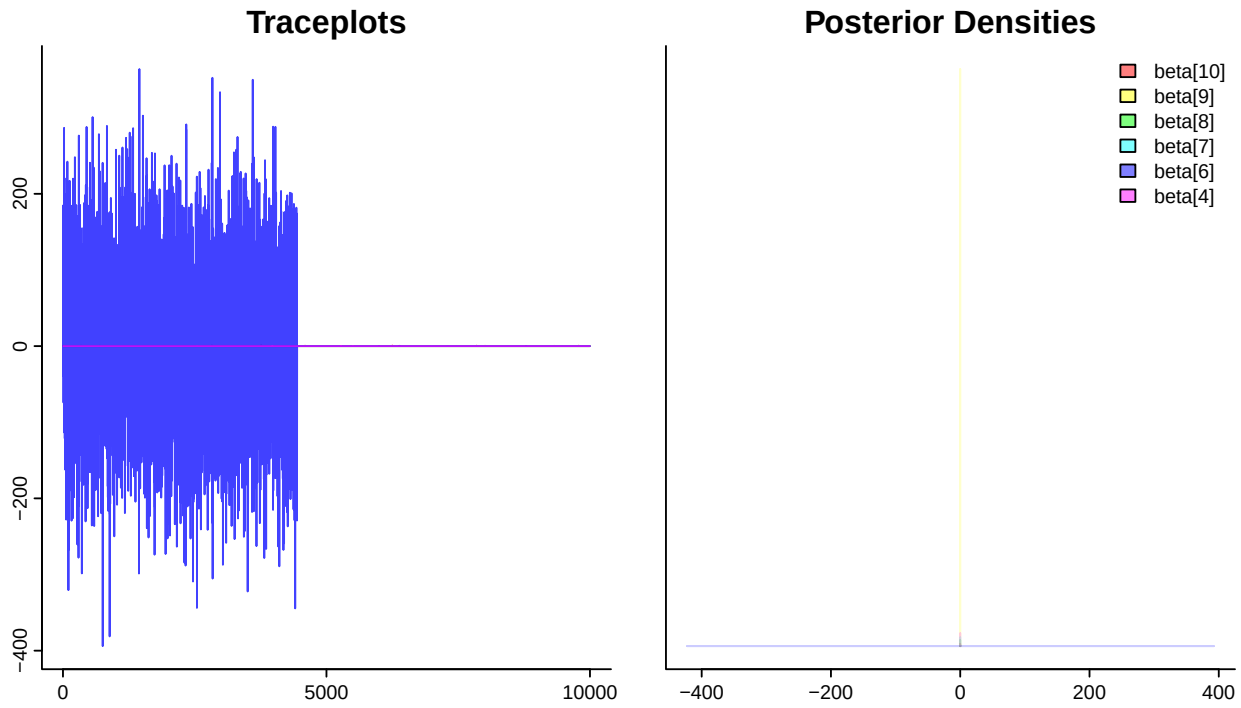
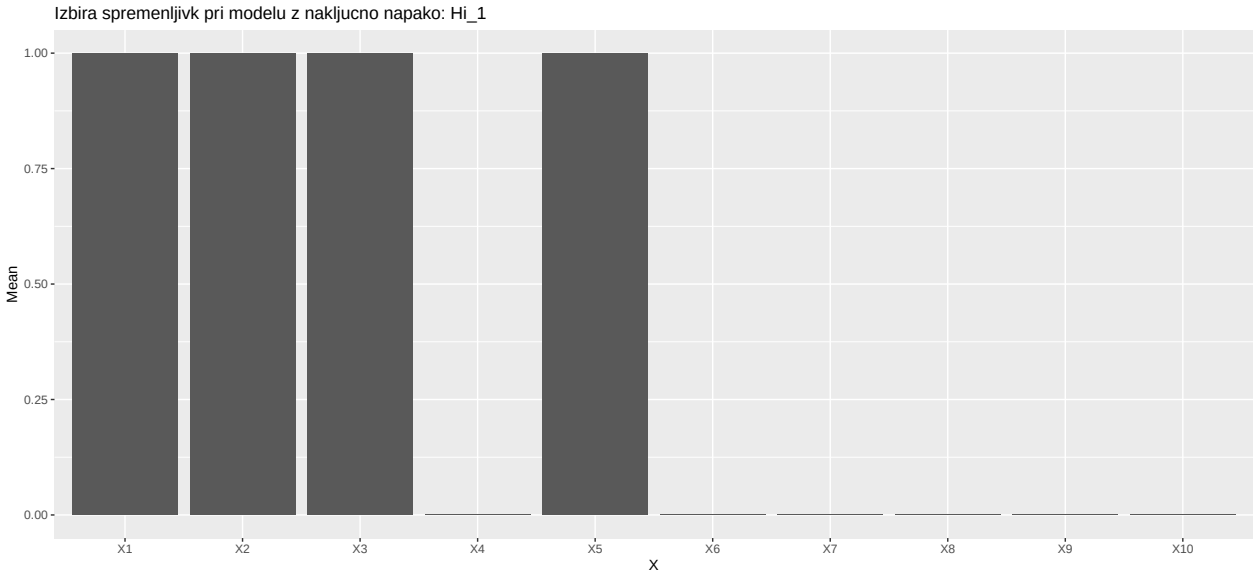


Figure 5: Konvergenca in posteriorne porazdelitve za ničelne bete



Ko spremenimo porazdelitev na χ_1^2 vidimo, da model ni več tako stabilen. Konvergenca je sploh pri spremenljivki X_6 zelo problematična, prav tako pa tudi pri ničenih parametrih spremenljivke X_3. To se pozna tudi na izbiri spremenljivk, ki je zato nepravilna in zelo netočna. Model namreč izbere samo spremenljivko X_3, ki pa ima že prej omenjeno težavo s konvergenco.

2.2.1.4 Slučajna napaka: χ_4^2

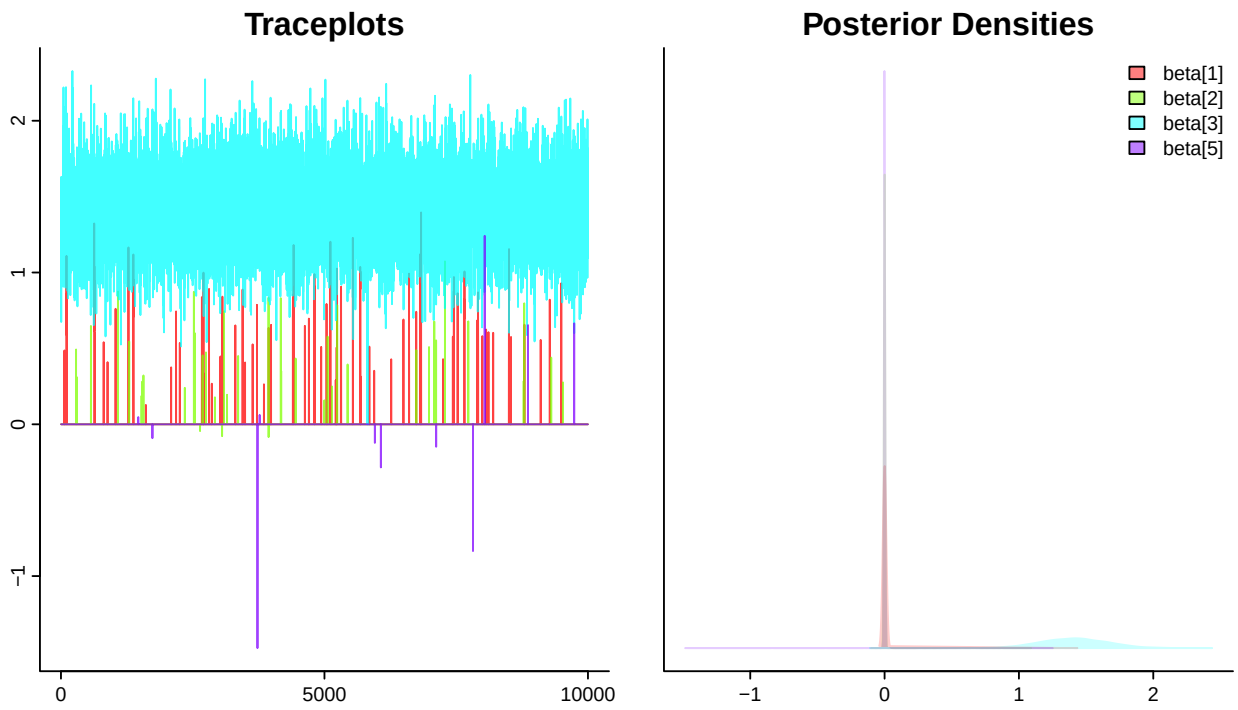


Figure 6: Konvergenca in posteriorne porazdelitve za neničelne bete

```
data.NIBLE.H4 <- gen.beta.data.2(100, napaka = rchisq(100, 4))
data.NIBLE.H4.selectY <- data.NIBLE.H4$Y
data.NIBLE.H4.selectX <- sweep(data.NIBLE.H4[,-5], 2,
                               colMeans(data.NIBLE.H4[,-5]),
                               FUN="-")[,-10] #centriramo
data.NIBLE.H4.selectX$X5 <- data.NIBLE.H4$X5
data.NIBLE.H4.selectX <- data.NIBLE.H4.selectX[, c(1,2,3,4,10,5,6,7,8,9)]
mod.H4 <- get.rjMCMC(data.NIBLE.H4.selectX, data.NIBLE.H4.selectY)
```

```
## thin = 1: sigma, psi, beta0, beta, z
## |-------------|-------------|-------------|-------------|
## |-------------------------------------------------------|
```

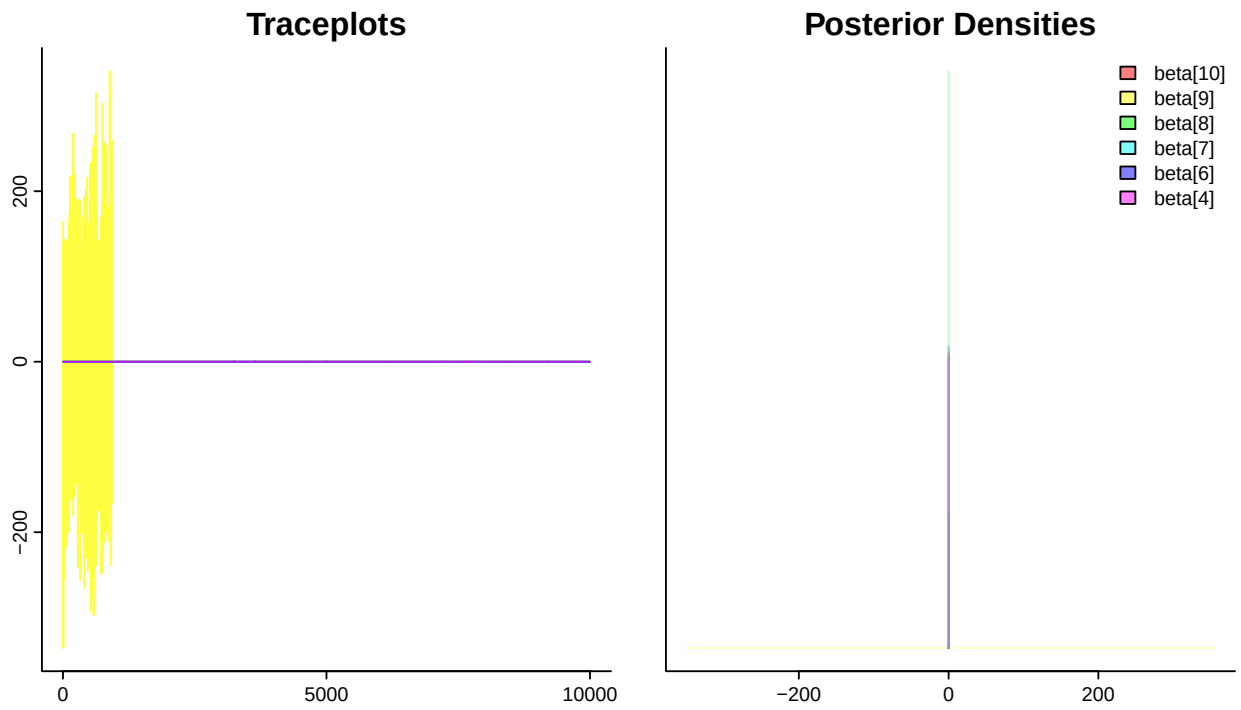
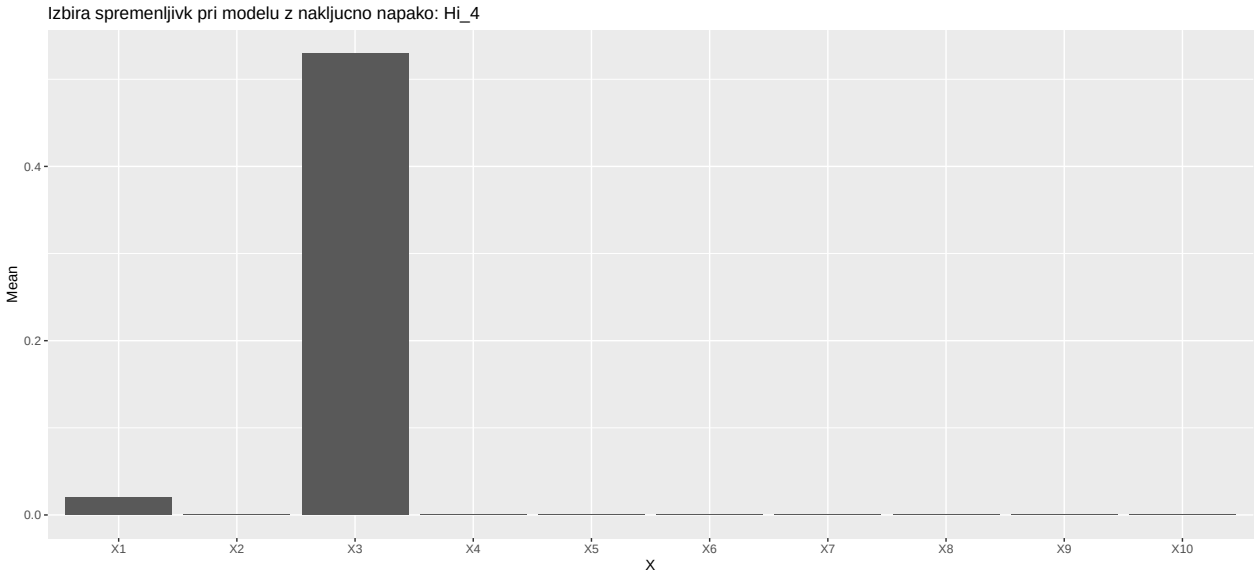


Figure 7: Konvergenca in posteriorne porazdelitve za ničelne bete



Podobna zgodba kot že pri prejšnem primeru, le da je za otenek večja verjetnost pri spremenljivki X_1, kar pa ne popravi kakovosti izbire modela. Pri tem modelu je problematična konvergenca spremenljivke X_9. Model slabo izbere spremenljivke.

2.2.2 Zaključek

Pri metodi Reverse Jump MCMC s programsko knjižnico `nimble`, je bilo moč opaziti, da pri slučajni napaki, ki obsega normalno porazdelitev, model dobro napoveduje izbiro

spremenljivk. S tem ko pa spremenimo vrsto porazdelitve slučajne napake, pa metoda ni tako robustna. Razlog zato vidim v tem, saj smo izbrali veliko število hiperparametrov s normalno porazdelitvijo, kar daje prednost slučajnim napakam s normalno porazdelitivijo. Iz tega tudi izhaja, da pri slučajni napaki χ^2_1 dobimo povsem slabo ocenjen model, s vprašljivo konvergenco.

2.3 Bootstrapped Augmented Backward Elimination - ABE

Za potrebe simulacije sem naredil funkcijo, ki je vrnila model `abe`, imela pa je naslednje vhodne podatke: spremenljivke X, Y, metriko s katero smo določali, kateri model je najboljši, število bootstrap ponovitev.

```
selection.ABE <- function(dataSetSelectX, Y, p = 10, metrika = "AIC", alpha=0.2, num.boo
  data.Skupaj <- cbind(dataSetSelectX, "Y" = Y)
  fit <- lm(Y ~ . , data = data.Skupaj, x = TRUE, y = TRUE)
  if(metrika %in% c("AIC", "BIC")){
    abe.fit.boot <- abe.boot(fit, criterion = metrika, data = dataSetSelectX,
                             tau = Inf, exp.beta = FALSE, num.boot = num.boot,
                             type.boot = "bootstrap")
  }
  else{
    abe.fit.boot <- abe.boot(fit, criterion = metrika, alpha = alpha,
                             data = dataSetSelectX, tau = Inf, exp.beta = FALSE,
                             num.boot = num.boot, type.boot = "bootstrap")
  }
  return(abe.fit.boot)
}
```

2.3.1 Opis simulacije

Za simulacije sem se odločil, saj me je zanimalo odstopanje različnih metod, glede na korelacijo in vrsto slučajne napake. Zato sem se odločil, da za vse 3 metrike naredim 100 ponovitev vseh kombinacij korelacije (r = 0, 0.8) in vrste napak (enake kot v prvem delu).

```
pon <- 100
sampleSize <- 100
vrsta.napake <- c("N(0,1)", "N(0,3)", "Hi_1", "Hi_4")
#vrsta.metode <- c("AIC", "BIC", "alpha")
korelacije <- c(0, 0.8)
zasnova <- expand.grid(korelacije, vrsta.napake)
zasnova <- do.call(rbind, replicate(pon, zasnova, simplify=FALSE)) %>%
  `colnames<-`(c("Korelacija","Napaka"))
matrika.rez <- matrix(NA, nrow = 3*nrow(zasnova), ncol = 14)

i = 1
j = 1
```

```
while(i < nrow(zasnova)){
  vrsta.napake.i <- zasnova[i, "Napaka"]
    if(vrsta.napake.i == "N(0,1)"){
      error <- rnorm(sampleSize, 0, 1)
    }
    else if(vrsta.napake.i == "N(0,3)"){
      error <- rnorm(sampleSize, 0, 3)
    }
    else if(vrsta.napake.i == "Hi_1"){
      error <- rchisq(sampleSize, df = 1)
    }
    else if(vrsta.napake.i == "Hi_4"){
      error <- rchisq(sampleSize, df = 4)
    }
    r.i <- zasnova[i, "Korelacija"]

  data.ABE <- gen.beta.data.2(sampleSize, r = r.i, napaka = error)
  data.ABE.selectY <- data.ABE$Y
  data.ABE.selectX <- sweep(data.ABE[,-5], 2, colMeans(data.ABE[,-5]), FUN="-")[,-10] #
  data.ABE.selectX$X5 <- data.ABE$X5
  data.ABE.selectX <- data.ABE.selectX[, c(1,2,3,4,10,5,6,7,8,9)]
  abe.AIC <-selection.ABE(data.ABE.selectX, Y =data.ABE.selectY,
                              p = 10, metrika = "AIC",num.boot = 500)
  abe.BIC <-selection.ABE(data.ABE.selectX, Y =data.ABE.selectY,
                            p = 10, metrika = "BIC",num.boot = 500)
  abe.alpha <-selection.ABE(data.ABE.selectX, Y =data.ABE.selectY,
                            p = 10, metrika = "alpha",num.boot = 500)

  rez.aic <- summary(abe.AIC)$var.rel.frequencies
  rez.bic <- summary(abe.BIC)$var.rel.frequencies
  rez.alpha <- summary(abe.alpha)$var.rel.frequencies
  matrika.rez[j,] <- c("R" = r.i,  "Metoda" = 1, "Napaka" = vrsta.napake.i, rez.aic)
  matrika.rez[j+1,] <- c("R" = r.i,  "Metoda" = 2, "Napaka" = vrsta.napake.i, rez.bic)
  matrika.rez[j+2,] <- c("R" = r.i,  "Metoda" = 3, "Napaka" = vrsta.napake.i, rez.alpha)
  j <- j + 3
  i <- i + 1
}

rez.df <- na.omit(as.data.frame(matrika.rez))
rez.df.nov <- as.data.frame(apply(rez.df, 2, as.numeric))
rez.df.nov$Metoda <- factor(rez.df.nov$V2, labels = c("AIC", "BIC", "alpha"))
rez.df.nov$Napaka <- factor(rez.df.nov$V3,labels = vrsta.napake)
rez.df.nov$Korelacija <- rez.df.nov$V1
abe.rez <- rez.df.nov[,4:17]
```

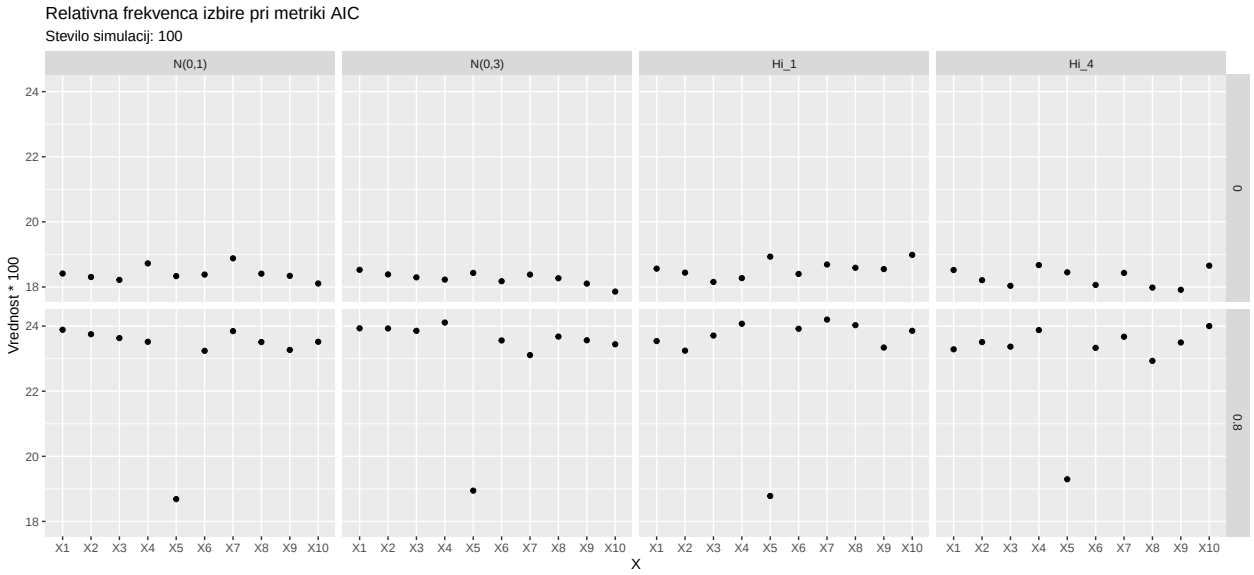
```r
colnames(abe.rez) <- c( paste("X", 0:10, sep=""),"Metoda", "Napaka","Korelacija")
```

2.3.2 Rezultati

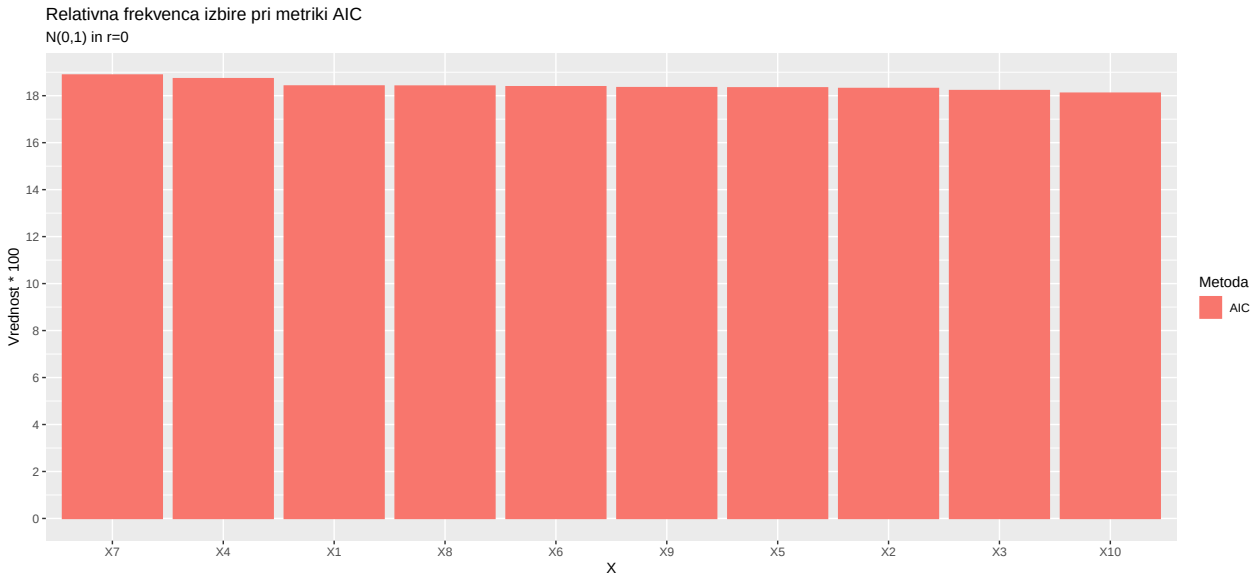
```r
abe.data <- readRDS("data/abe_simulation.RDS")

aba.data.sim <- abe.data %>%
  group_by(Korelacija, Metoda, Napaka)%>%
  summarise_all(mean) %>%
  gather(key = "X", "Vrednost", -c(Korelacija, Metoda, Napaka))  %>%
  filter(X != "X0")

aba.data.sim$X <- factor(aba.data.sim$X, levels = paste("X", 1:10, sep = ""))
```

Pri predstavitvi rezultatov bom izpustil parameter β_0, ki je v prisoten v vsakem modelu.

2.3.2.1 AIC

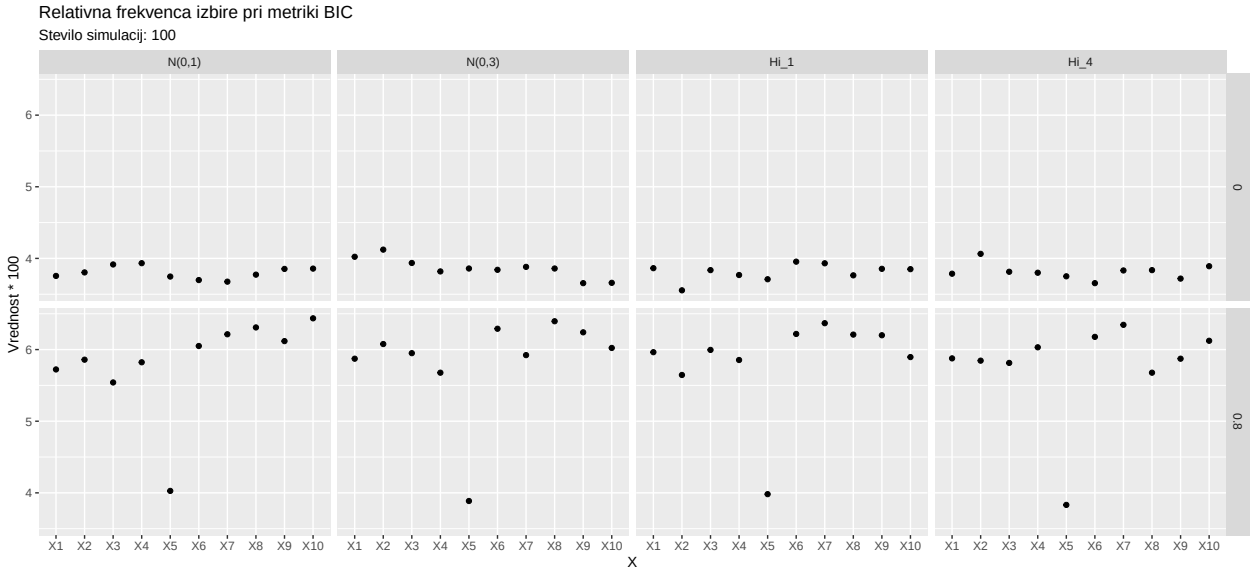


Kot vidimo, so razlike v deležih spremenljivk zelo majhne. Sklepam, da je to posledica boostrapa 500, saj za vsako kombinacijo in vsako metriko naredimo 100 ponovitev po 500 bootstrap vzorcev, kar daje enakomerno porazdelitev vsem spremenljivkam (ali pa je samo slab model). Omeniti še velja, da se pri povečanju korelacije deleži vključevanja v model povečujejo in spremenljivka X_5 se začne vesti drugače - porast vključevanja v model bistveno pade. Zaradi zelo majhnih razlik si poglejmo posebaj primer N(0,1) in r = 0.

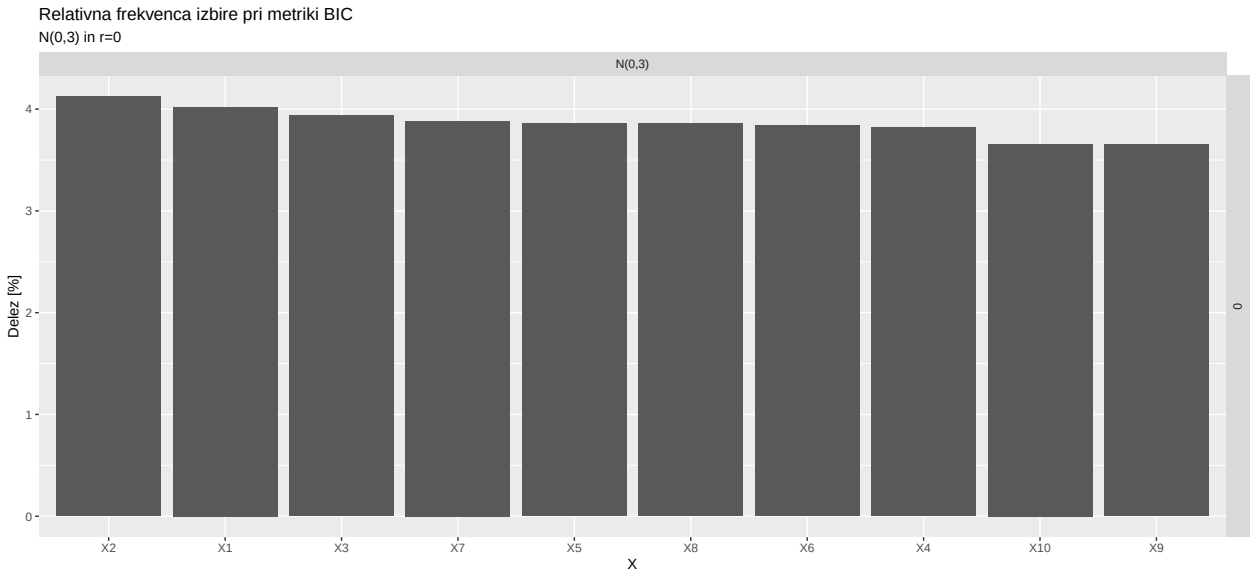


Ko pogledamo deleže, vidimo, da imata najvišji deleže spremenljivke, ki imajo $\beta = 0$, kar nakazuje, da v tem našem primeru model ne bi pravilno izbiral. Med najvišjimi deleži je kar 5 spremenljivk, ki imajo β fiksirano na 0. Podobno je tudi pri drugih scenarijih.

2.3.2.2 BIC

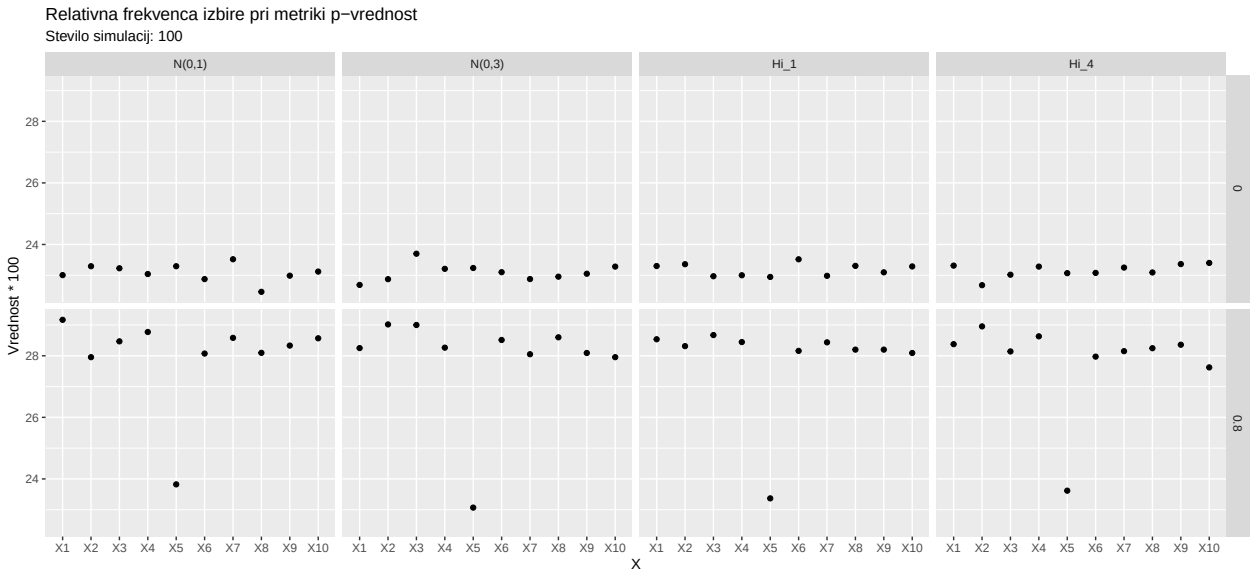


Tako kot pri metriki AIC, so tudi pri BIC razlike v deležih zelo majhne. Razlika, ki jo je moč opaziti je, da imajo spremenljivke X_6 do X_10 nekoliko manjše deleže pri normalni slučajni napaki in korelaciji r = 0. Pri slučajnih napak iz porazdelitve χ^2, se stvari obnašajo čudno in nepravilno (največje deleže imajo spremenljivke s $\beta = 0$). S tem ko povečamo korelacijo med spremenljivkami se nam zgodi, da binarna spremenljivka ne opisje več dobro modela, zato je njen delež daleč najslabši. Še posebno pri slučajni napaki N(0,3) in r = 0 so spremenljivke pravilno selekcionirane, ampak ko povečamo korelacijo, selekcija ni več pravilna, zato ugotavljam, da je korelacija oz. kolinearnost eden od močnih vplivov na izbiro spremenljivk. Vseeno si poglejmo od bližje rezulatate pri slučajni napaki N(0,3) in r = 0.

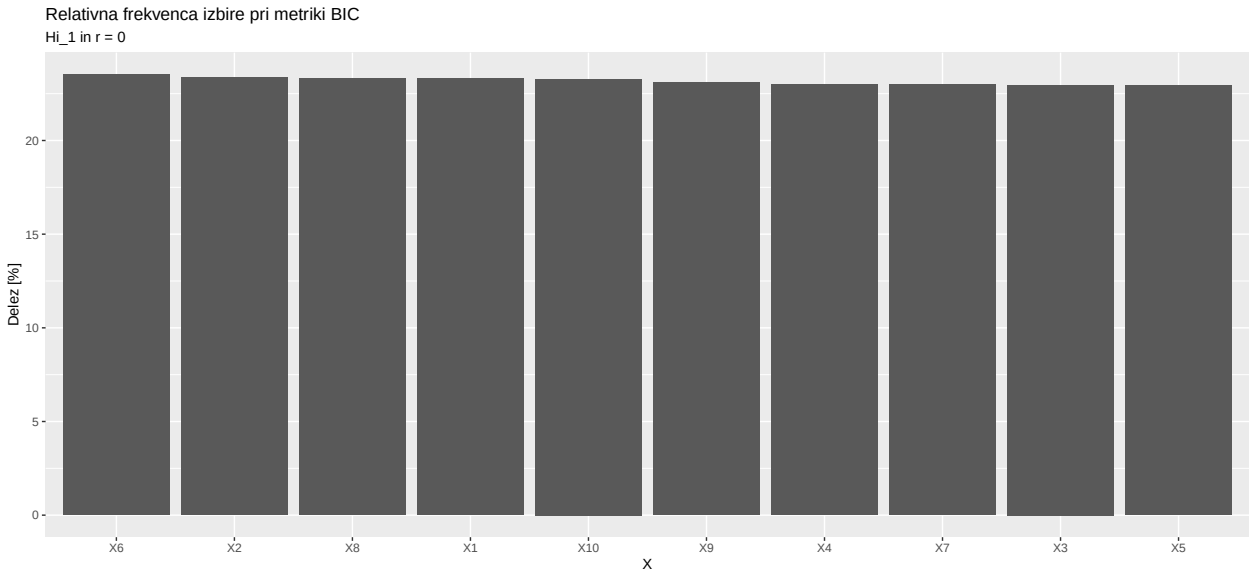


Kot vidimo je vrstni red pravilen, le da gre pri X_7 in X_5 za minimalne razlike.

2.3.2.3 p-vrednost



Pri metriki p-vrednosti sem alpha fiksiral na 0.2, dobimo zanimive rezultate. Pri scenarijih z slučajno napako iz normalne porazdelitve so selekcije spremenljivk boljše pri večjem standardne odklonu, pri korealcijskih koeficientu r = 0. Ko povečamo korealcijski koefiecient se delež pravilno napovedanih spremenljivk znatno zmanjša. Pri scenarijih iz χ^2 porazdelitve, da spremenljivke z največjim deležem izbire imajo teoroetičen koeficien enak 0. S povečanjem korealcij se ponovno opazi razliko na spremenljivki X_5. Tokrat si podrobneje oglejmo χ^2_1 pri korelaciji r = 0.



Dejstvo je, da so razlike tako majhne, da se delež največkrat izbrane spremenljivke v primerjavi z najmanjšim deležem izbrane spremenljivke razlikuje za 0.06. Težko rečemo, da bi v tem primeru izbrali pravilne spremenljivke.

2.3.3 Zaključek

Ko pogledamo vse tri različne metrike, lahko rečemo, da so razlike pri nekaterih metrikah tako majhne, da se nam hitro lahko zgodi da vzmemo nevplivno spremenljivko. Simulacija je vseeno pokazala, da je BIC metrika, dajala najbolj zanesljive rezultate. Rezultati so bili dokaj natančni samo za slučajno napako normalne porazdelitve, pri hi-kvadrat porazdelitvi, so bile pravilne spremenljivke redkeje izbrane. Korelacija je pomemnbno vplivala na izbor spremenljvik, tako, da v tem primeru priororčam izbor glede na metriko p-vrednosti.

2.4 Spike and Slab - BoomSpikeSlab

Spike and Slab regresija se Bayesovi statistiki uporablja predvsem za selekcijo spremenljivk, ko imamo več spremenljivk kot statisitčnih enot. Bralec si lahko več prebere **tukaj** ali **tukaj**

2.4.1 Predstavitev simulacije

```
selection.spike.slab <- function(dataSetSelectX, Y){
  dataX <- cbind(rep(1, nrow(dataSetSelectX)),
                 as.matrix(dataSetSelectX))
  prior <- IndependentSpikeSlabPrior(dataX, Y,
                                     prior.df = .0)
  dataX <- dataX[,-1]
  model <- lm.spike(Y ~ dataX,  niter = 1000, prior = prior)
  return(model)
}
```

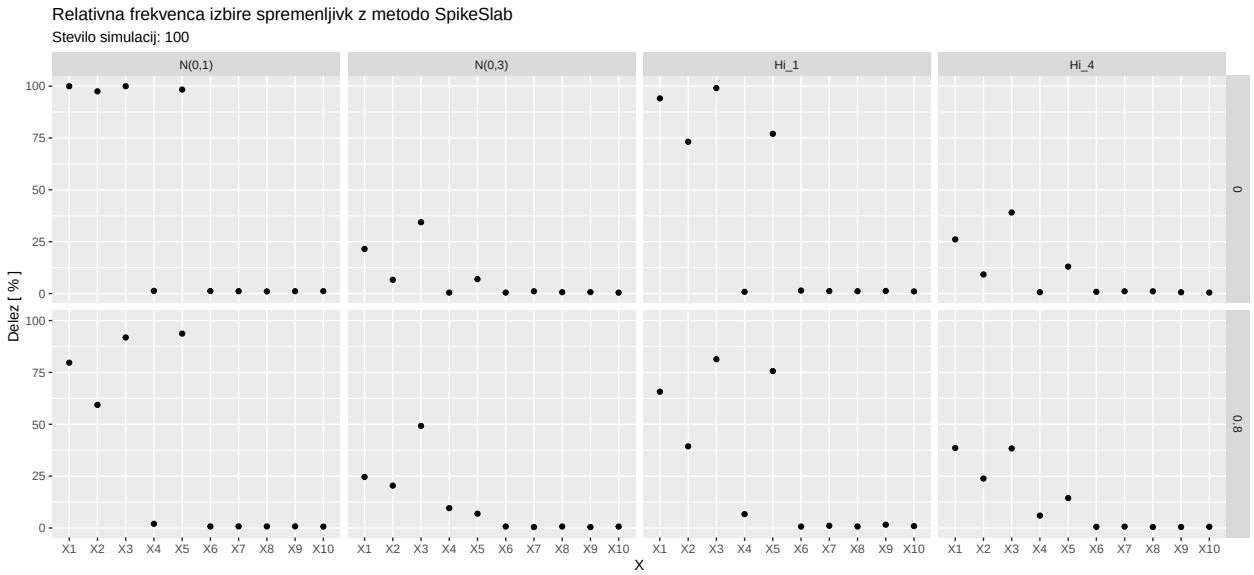
V enostavni simulaciji bom naredil 100 ponovitev zgoraj opisane funkcije, pri čemer me je zanimala frekvenca in izbor izbranih spremenljivk. To sem naredil za vse vse 4 vrste napak in dve različni vrednosti korelacije.

```
pon <- 100
sampleSize <- 100
vrsta.napake <- c("N(0,1)", "N(0,3)", "Hi_1", "Hi_4")
#vrsta.metode <- c("AIC", "BIC", "alpha")
korelacije <- c(0, 0.8)
zasnova <- expand.grid(korelacije, vrsta.napake)
zasnova <- do.call(rbind, replicate(pon, zasnova, simplify=FALSE)) %>%
  `colnames<-`(c("Korelacija","Napaka"))
matrika.rez <- matrix(NA, nrow = nrow(zasnova), ncol = 13)
i = 1

while(i <= nrow(zasnova)){
  vrsta.napake.i <- zasnova[i, "Napaka"]
  if(vrsta.napake.i == "N(0,1)"){
    error <- rnorm(sampleSize, 0, 1)
  }
  else if(vrsta.napake.i == "N(0,3)"){
    error <- rnorm(sampleSize, 0, 3)
  }
  else if(vrsta.napake.i == "Hi_1"){
    error <- rchisq(sampleSize, df = 1)
  }
  else if(vrsta.napake.i == "Hi_4"){
    error <- rchisq(sampleSize, df = 4)
  }
  r.i <- zasnova[i, "Korelacija"]

  data.Spike <- gen.beta.data.2(sampleSize, r = r.i, napaka = error)
  data.Spike.selectY <- data.Spike$Y
  data.Spike.selectX <- sweep(data.Spike[,-5], 2, colMeans(data.Spike[,-5]), FUN="-")[,-
  data.Spike.selectX$X5 <- data.Spike$X5
  data.Spike.selectX <- data.Spike.selectX[, c(1,2,3,4,10,5,6,7,8,9)]

  mod.spikeSlab <- selection.spike.slab(dataSetSelectX = data.Spike.selectX, Y =data.Spi
  inc.prob <- summary(mod.spikeSlab)$coef[,"inc.prob"]
  inc.prob.urejen <- inc.prob[order(factor(names(inc.prob), levels = paste("dataXX", 0:1

  matrika.rez[i,] <- c("R" = r.i,"Napaka" = vrsta.napake.i, inc.prob.urejen)
  i <- i + 1
}
```

```
spikeSlab.df <- as.data.frame(matrika.rez[,-13])
spikeSlab.df$V2 <- factor(spikeSlab.df$V2,labels = vrsta.napake)
colnames(spikeSlab.df) <- c("Korelacija", "Napaka", paste("X", 1:10, sep =""))
```

2.4.2 Rezultati

```
spikeSlab.data <- readRDS("data/spikeSlab_simulation.RDS")

spike.data.sim <- spikeSlab.data %>%
  group_by(Korelacija, Napaka)%>%
  summarise_all(mean) %>%
  gather(key = "X", "Vrednost", -c(Korelacija, Napaka))
spike.data.sim$X <- factor(spike.data.sim$X, levels = paste("X", 1:10, sep = ""))
```



Metoda SpikeSlab se je izkazala kot najbolj uspešna metoda v iskanju najboljši spremenljivk za opis spremenljvike Y. Pri korealcijskem koeficientu r = 0, namreč za vse vrste slučajnih napak pravilno napove največji delež za spremenljvke, ki imajo neničelni koeficient β. Če primerjamo po posameznih porazdelitvah vidimo, da pri normalni sluačjni napaki, večji standardni odklon nekoliko pokvari rezultate, vendar nikoli ni vključil modela, ki bi vseboval ničelni koeficient β. Pri hi-kvadrat porazdelitvi s eno prostostno stopnjo imajo teoretično pravilne spremenljivke visok delež izbire, s povečavanjem prostostnih stopenj na štiri, pa se ta delež nekoliko zmanjša. Metoda pravilno zazna tudi binarno spremenljivko X_5. S povečavo korelacijskega koeficienta (r = 0.8) se deleži nekoliko zmanjšajo, vendar to nespremeni bistveno rezultatov, razen pri slučajni napaki - N(0,3), kjer je delež za spremenljivko X_4 ($\beta_4 = 0$) nekoliko večji kot X_5 ($\beta_5 \neq 0$).

2.5 Zaključek

Kot najboljša metoda se mi je zdela SpikeSlab metoda, saj je delovala najbolj robustno, tudi na slučajnih napakah porazdeljene iz χ_2. Na drugo mesto bi postavil Reverse jump MCMC metodo, ki je za normalne porazdelitve delovala zelo dobro, zmanjkalo jo je pri drugi vrsti porazdelitve. Pri bacward bootstrap elimination se mi zdi, da je vseeno najboljše izbira bila BIC kriterij. Možno, je da je kakšna napaka, saj so deleži preveč enakomerno porazdeljeni po spremenljvkah, tudi v primerih ko je šlo za zelo nizko slučajno napako.