BDA - Assignment 9

Anonymous

Contents

Modeling

We are using multiple linear regression to model the relationship between years of service and years since phd (independent variables) with the annual salary of professors(dependent variable). Multiple linear regression takes into account more than one independent variables. In multiple linear regression, we have chosen the additive model. Let x_1 and x_2 be the independent variables and y be the dependent variable. Then, additive multiple linear regression equation is given by,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Here, β_0 is the intercept. β_i is the slope associated with variable x_i.

For observations of salary y, years since phd x_1 and years of service x_2, we are modeling the posteriors of β_0, β_1 and β_2 to obtain the linear regression posterior estimate. The posterior distributions are obtained using two different types of model schemes discussed below.

Hierarchical model:

In hierarchical model, we consider groups based on rank(first model) and discipline(second model) to model the multiple linear regression for each of the group. For each of the group, parameters for the prior distributions of β_k are associated with hyperparameters α_k. In equation,

$$\begin{aligned}
y_{ij} &\sim N(\beta_{0j} + \beta_{1j} x_{1ij} + \beta_{2j} x_{2ij}, \sigma) \\
\beta_{kj} &\sim N(\alpha_k, \tau) \\
\sigma &\sim gamma(v_1, v_2) \\
\alpha_k &\sim N(\mu_k, sd_k) \\
\tau &\sim gamma(a, b)
\end{aligned}$$

Here, j is the index of groups. i the index of observation in a group. k belongs to {0,1,2}. α_k are the hyperpriors for the means of β_k in the normal distribution in each group j. τ is the hyperprior for the standard deviation of β_k. v_1, v_2, a,b,$\mu_k$and sd_k are the values provided depending on the prior used.

Non-hierarchical model:

A pooled model is considered for non-hierarchical model. In pooled model, no group based analysis of observations are done. All the observations are considered to belong one single group and a single set of posteriors of parameters in interest are analysed for all the observations. In equation,

$$
\begin{aligned}
y_i &\sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \sigma) \\
\beta_k &\sim N(\alpha_k, \tau) \\
\sigma &\sim gamma(v_1, v_2)
\end{aligned}
$$

Here, i is the index of observation. Values of α_k, tau, v_1 and v_2 are provided depending on the choice of prior. k belongs to {0,1,2}.

Choice of priors

In all the hierarchical models, two different set of priors have been used for hyperparameters and common standard deviation of groups: one set for the model and one set for sensitivity analysis. For each of the group in hierarchical model, normal distribution is used for the likelihood of the observation. Mean for the normal distribution is given by the multiple linear regression

```r
library(aaltobda)
library(loo)
library(cmdstanr)
library(ggplot2)
library(bayesplot)
library(gridExtra)
set_cmdstan_path('/coursedata/cmdstan')

data <- read.csv("Salaries.csv")
```

Model 1: Hierarchical model with groups on the basis of ranks

Data preparation and sampling:

In this model, three different ranks: Professor(group 1), Associate Professor(group 2) and Assistant Professor(group 3) are used as basis for forming groups of the hierarchical model. The dataset has class imbalance for these ranks. The rank with lowest number of observation has 64 observations. Hence, for class balance, only 64 observations from each group has been considered in the modeling. No preprocessing operations that transform the data have been performed on the data. For posterior predictive distribution data, 10 years of service and 10 years since phd has been provided as input of the independent variables. We considered 10 years to be appropriate value to compare the posterior predictive analysis between ranks.

```r
# Data preparation
prof<- data[data$rank == "Prof",]
Aprof<- data[data$rank == "AssocProf",]
Asprof <- data[data$rank == "AsstProf",]

salary <- data.frame(prof$salary[1:64], Aprof$salary[1:64], Asprof$salary[1:64])
x1 <- data.frame(prof$yrs.since.phd[1:64], Aprof$yrs.since.phd[1:64], Asprof$yrs.since.phd[1:64])
x2 <- data.frame(prof$yrs.service[1:64], Aprof$yrs.service[1:64], Asprof$yrs.service[1:64])

stan_data <- list(
y = salary,
#Number of observations per group
N = nrow(salary),

#Number of groups
J = ncol(salary),
x1 = x1,
x2 = x2,
#For predictive posterior distribution, at 10 years of service and 10 years since phd
x1pred = 10,
x2pred = 10
)
```

The model is run using default parameters setting in cmdstanrR sample() method. The default settings has 4 MCMC chains with 2000 iterations each. Among the 2000 samples, 1000 are discarded as warmup. In total, 4000 samples of the posterior is obtained altogether.

```
mod1 <- cmdstan_model("projectmodel1.stan")
fit_lin1 <- mod1$sample(data = stan_data, refresh = 1000, seed = 1)
```

The stan model code used is given below.

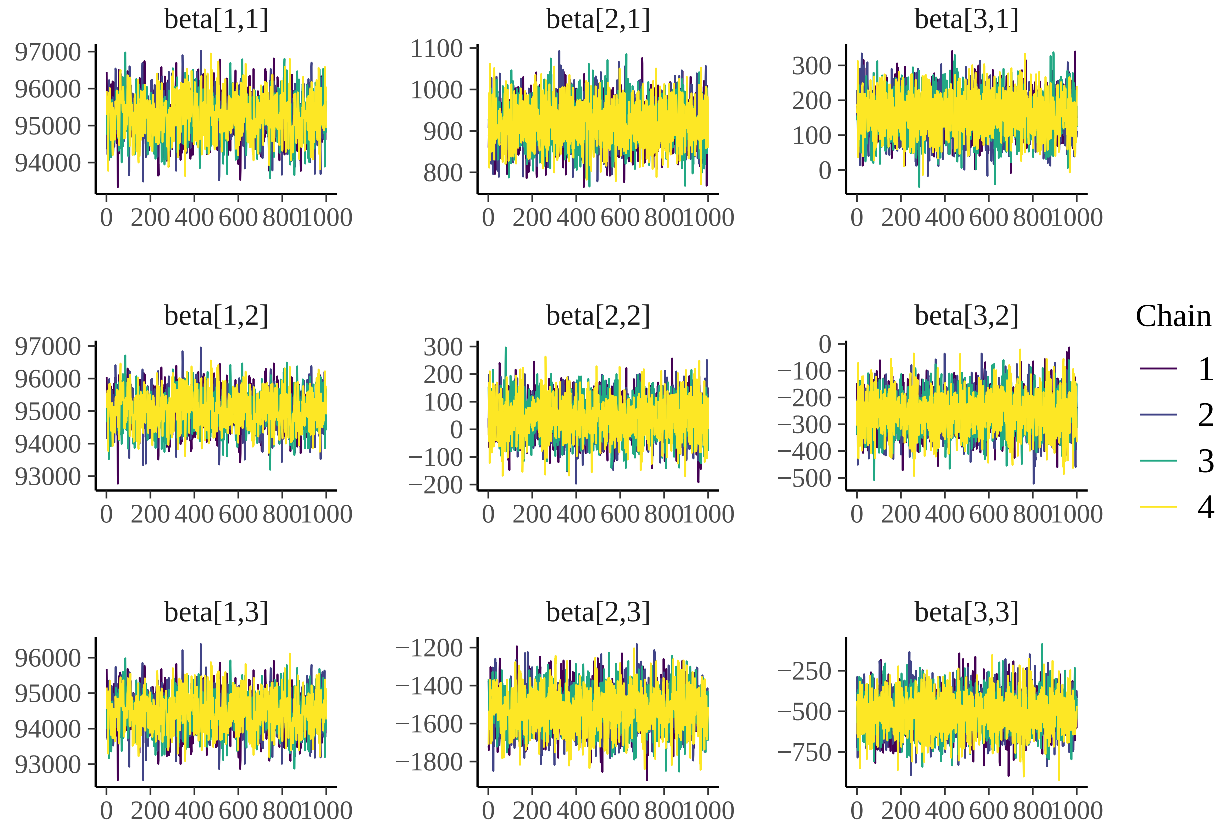
```
mod1$print()
```

```
## //Additive model
## data {
##   int<lower=0> N;
##   int<lower=0> J;
##   vector[J] y[N];
##   vector[J] x1[N];
##   vector[J] x2[N];
##   real x1pred;
##   real x2pred;
## }
##
## parameters {
##   real hyperbeta[3];
##   real<lower=0> hypersigma;
##   vector[J] beta[3];
##   real<lower=0> sigma;
## }
##
## model {
##   hyperbeta[1] ~ normal(70000, 10000);
##   hyperbeta[2] ~ normal(0, 10);
##   hyperbeta[3] ~ normal(0, 10);
##   hypersigma ~ gamma(2,2);
##   sigma ~ gamma(2,2);
##   for (j in 1:J)
##   {
##     beta[1,j] ~ normal(hyperbeta[1], hypersigma);
##     beta[2,j] ~ normal(hyperbeta[2], hypersigma);
##     beta[3,j] ~ normal(hyperbeta[3], hypersigma);
##   }
##   for (j in 1:J)
##     y[,j] ~ normal(beta[1,j] + beta[2,j]*to_vector(x1[,j]) + beta[3,j]*to_vector(x2[,j]), sigma);
## }
##
## generated quantities
## {
##   vector[J*N] log_lik;
##   real ypred[J];
##   int m = 1;
##   for(j in 1:J)
##   {
##     for (i in 1:N)
##     {
##       log_lik[m] = normal_lpdf(y[i,j] | (beta[1,j] + beta[2,j]*x1[i,j] + beta[3,j]*x2[i,j]), sigma);
##       m = m + 1;
##     }
##   }
##
##   for (i in 1:J)
##     ypred[i] = normal_rng(beta[1,i] + beta[2,i]*x1pred + beta[3,i]*x2pred, sigma);
## }
```

Convergence diagnostics:

We can visually observe that for all the parameters, all the chains have converged.

```
color_scheme_set(scheme = "viridis")
mcmc_trace(fit_lin1$draws("beta"))
```



The r_hat values are <1.01 for all the parameters. This means that the chains have converged.

```
rhat <- fit_lin1$summary()[, "rhat"]
max(rhat, na.rm=TRUE)
```

```
## [1] 1.003139
```

```
min(rhat, na.rm=TRUE)
```

```
## [1] 0.9995296
```

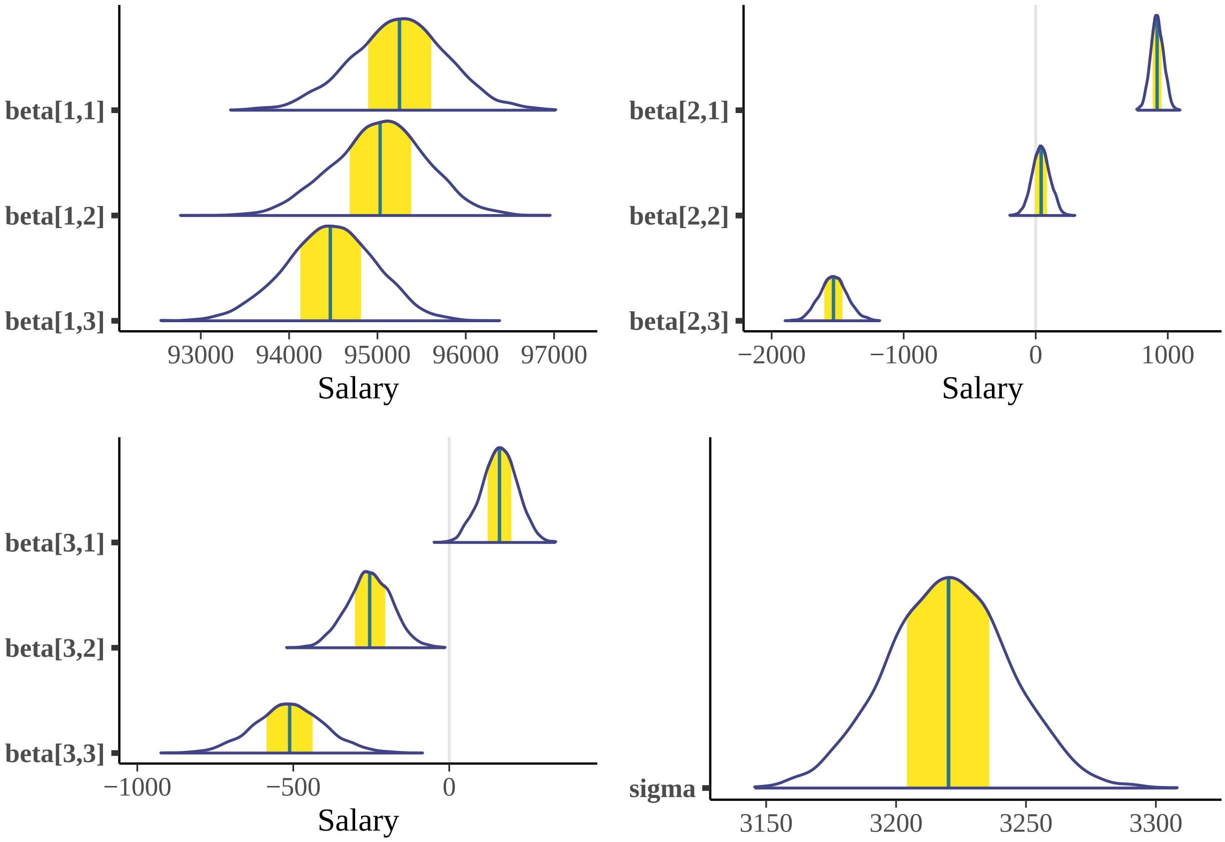
Also, further diagnosis provided by the cmdstanR method cmdstan_diagnose() verifies that the ESS is satisfactory and there are no divergences.

```
fit_lin1$cmdstan_diagnose()
```

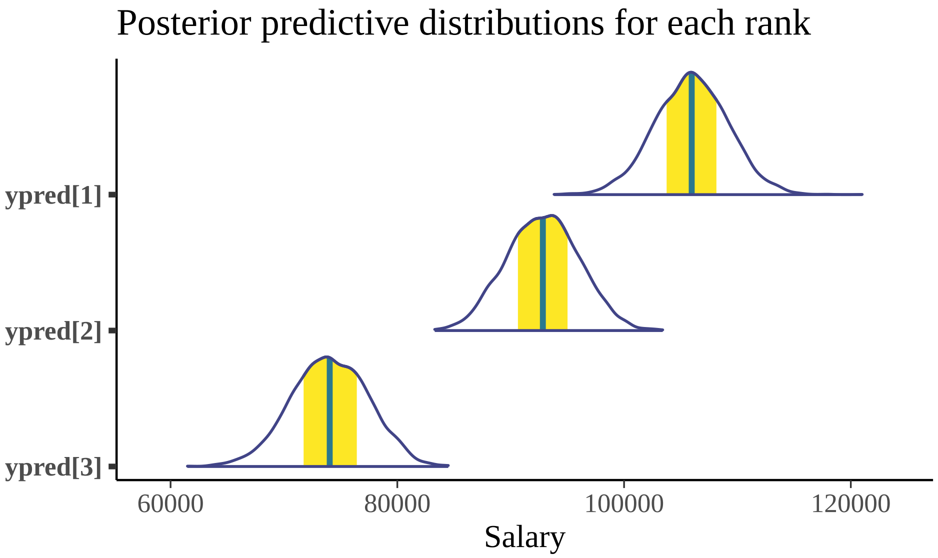
```
## Processing csv files: /tmp/RtmpC3Ul91/projectmodel1-202212032011-1-2d3194.csv, /tmp/RtmpC3Ul91/proje
##
## Checking sampler transitions treedepth.
## Treedepth satisfactory for all transitions.
##
## Checking sampler transitions for divergences.
## No divergent transitions found.
##
## Checking E-BFMI - sampler transitions HMC potential energy.
## E-BFMI satisfactory.
##
## Effective sample size satisfactory.
##
## Split R-hat values satisfactory all parameters.
##
## Processing complete, no problems detected.
```

Results:

The plots for the posteriors of the parameters for each rank are given below.



The plot for the predictive posterior distribution for each rank is for salary at 10 years of service and 10 years since phd.



The mean point estimate(of each parameter) of the linear regression equation for each of the group is:

a) Professor:

$$Salary = 95250 + 919 * years.since.phd + 161 * years.of.service$$

The 90% posterior interval of:

Intercept is [94319, 96135]. Factor of years.since.phd is [838, 1001]. Factor of years.of.service is [64, 253].

b) Associate Professor:

$$Salary = 95030 + 41 * years.since.phd - 255 * years.of.service$$

- The 90% posterior interval of:

  Intercept is [94136, 95868]. Factor of years.since.phd is [-69, 154]. Factor of years.of.service is [-374, -139].

c) Assistant Professor:

$$Salary = 94466 - 1532 * years.since.phd - 512 * years.of.service$$

- The 90% posterior interval of:

  Intercept is [93600, 95293]. Factor of years.since.phd is [-1700, -1360]. Factor of years.of.service is [-701, -321].

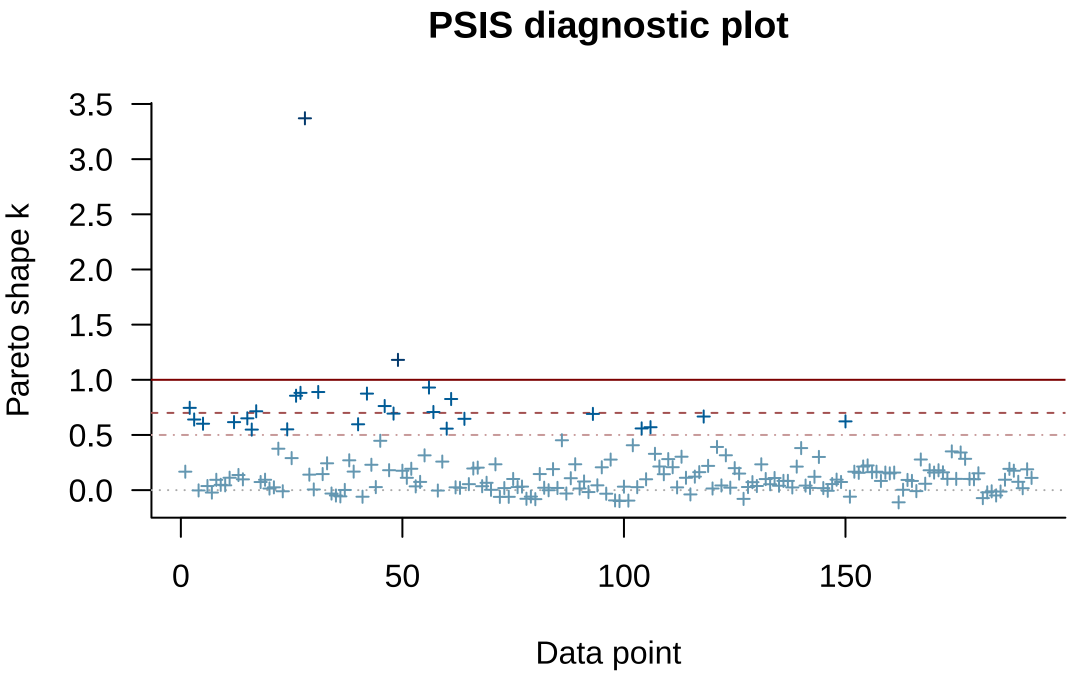
Posterior predictive check:

We use Leave-One-Out Cross Validation(LOO-CV) for posterior predictive check. Some of the observations have k-pareto values are >0.7. This means that the importance sampling estimate isn't reliable. However, these observations are only about 6% of the total observations. So, we will consider the obtained elpd_loo when comparing models.

```
loo <-fit_lin1$loo()
loo
```

```
##
## Computed from 4000 by 192 log-likelihood matrix
##
##          Estimate     SE
## elpd_loo  -5158.9  602.9
## p_loo       230.3   57.5
## looic     10317.8 1205.7
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     165   85.9%   317
##  (0.5, 0.7]   (ok)        15    7.8%   50
##    (0.7, 1]   (bad)       10    5.2%   24
##    (1, Inf)   (very bad)   2    1.0%   1
## See help('pareto-k-diagnostic') for details.
```

```
plot(
  loo,
  diagnostic = c("k"),
  label_points = FALSE,
  main = "PSIS diagnostic plot"
)
```



Sensitivity analysis with respect to prior:

For sensitivity analysis, alternative prior discussed earlier in the section “Choice of priors” is used for the parameters and hyperparameters. The result for each rank with different prior is:

a. Professor:

The point mean estimate and 90% interval are:

For intercept, 95298 and [94554, 96044] respectively.

For factor of years.since.phd, 919 and [850, 990] respectively.

For factor of years.of.service, 160 and [80, 238] respectively.

b. Associate professor:

The point mean estimate and 90% interval are:

For intercept, 95082 and [94358, 95809] respectively.

For factor of years.since.phd, 40 and [-53, 130] respectively.

For factor of years.of.service, -255 and [-353,-157] respectively.

c. Assistant professor:

The point mean estimate and 90% interval are:

For intercept, 94511 and [93804, 95222] respectively.

For factor of years.since.phd, -1538 and [-1686, -1394] respectively.

For factor of years.of.service, -510 and [-672, -343] respectively.

The change in priors don’t have a significant effect on the point estimate of the parameters of the regression equation. Hence, the model isn’t sensitive to the choice of prior.