STAT 638 HW09

Dennis Leet

November 15, 2021

This is some homework of which I am particularly proud. It was completed for my Fall 2021 Introduction to Applied Bayesian Methods course and comes from ***A First Course in Bayesian Statistical Methods*** *by* ***Peter D. Hoff ISBN: 978-0-387-92299-7*** *chapter 9 on Bayesian Linear Regression. Question 1 involves fitting a linear regression model on four athletes performance and obtaining a posterior predictive distribution on each athlete. In question 2 we fit another regression model with a g-prior, obtain posterior parameter confidence intervals and then perform model averaging.*

1 Question 9.1.a

Extrapolation: The file swim.dat contains data on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

a) Perform the following data analysis for each swimmer separately: i. Fit a linear regression model of swimming time as the response and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds. ii. For each swimmer j, obtain a posterior predictive distribution for Y_j^*, their time if they were to swim two weeks from the last recorded time.

We utilized the process on **HOFF P. 155**. The process involved the following steps:
1) Construct Prior for all Parameters
2) Compute V and m and sample β_{s+1}
3) Compute $SSR(\beta_{s+1}) \sim MVN(V, N)$.
4) Sample $\sigma^{(2)s+1}$ from the appropriate inverse-gamma distribution.
5) Repeat until desired sample size is achieved.
We must use a Gibbs sampler with this process to interrelate $\boldsymbol{\beta}$ and σ^2.

It is given that the swimming times are generally expected to fall between 22 and 24 minutes. We therefore view the expected finishing time prior as normally distributed with $\mu = 23$ and $\sigma^2 = 0.5^2$. This would squarely place 95% of the finishing times within the allotted expectation. The Gibbs Sampler starting values and hyper-parameters follow:

We utilized many $\Sigma_0 = \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$

$\beta_0 = (23, 0)$

$\sigma_0^2 = (\frac{1}{2})^2$

$\nu_0 = 1$

While the values for β_0, Σ_0 and σ_0^2 make sense given the prior information, we tried many different ν_0

values. Unless the number were very small or very large it didn't seem to play much of a role in the posterior probabilities. Eventually, we settled on $\nu_0 = 1$ for the sake of simplicity and to affect our posterior in the least extreme fashion. We then sampled each swimmers normally distributed posterior predictive distributions with the Gibbs Sampled values. Plots of their predicted times are bellow in figure 1.

2 9.1.b

(b) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Using your predictive distributions, compute $Pr(Y_j^* = max\{Y_1^*, ..., Y_4^*\}|Y))$ for each swimmer j, and based on this make a recommendation to the coach.

The probability that the swimmer 1-4 are the fastest respectively follow: 0.66, 0.02, 0.29, 0.03. Based on these times and the plots of the predicted values, I strongly encourage the coach to use Swimmer 1.

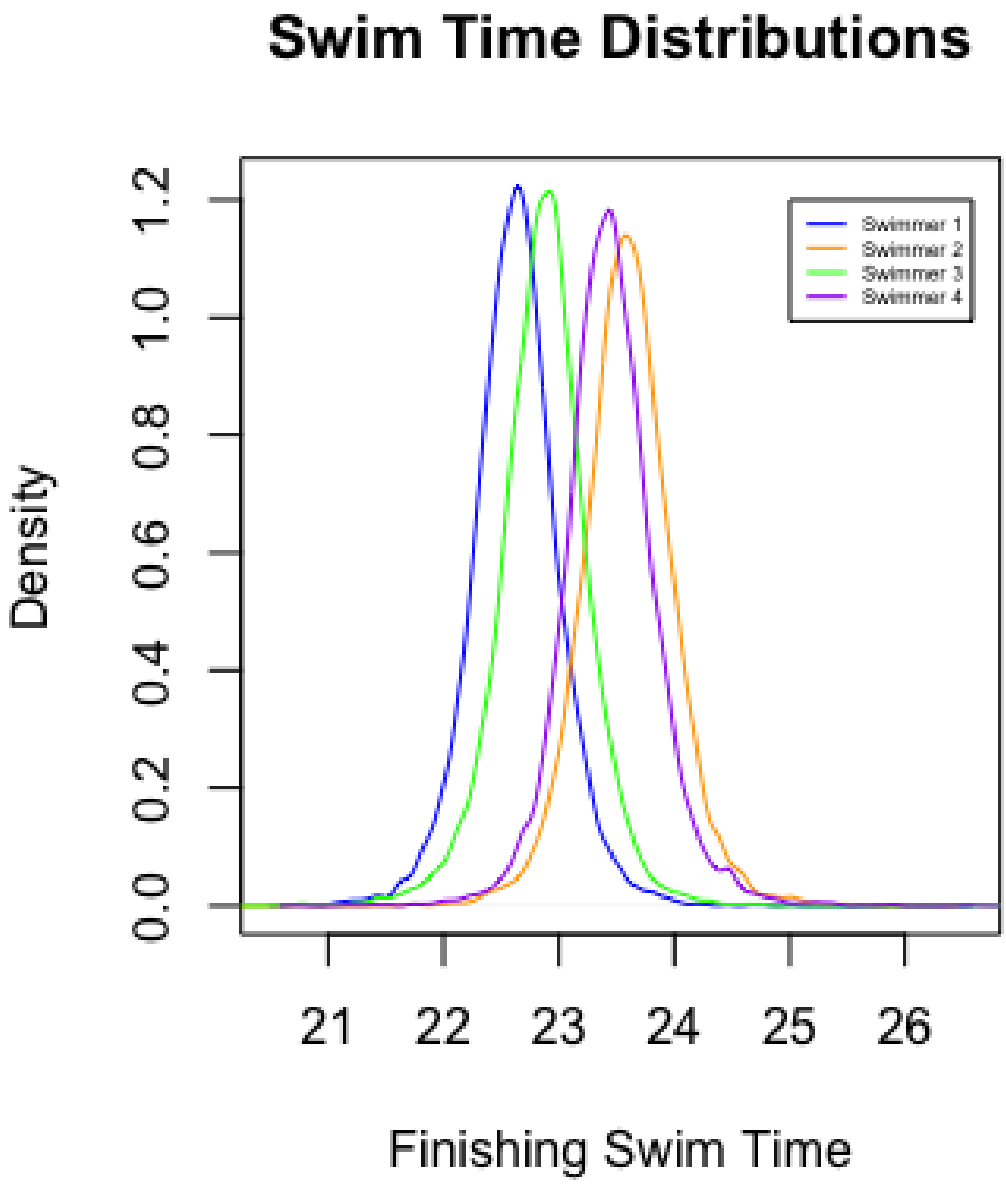

Figure 1:
9.1.a: Predicted Finishing Swim Time Distributions for all Four Athletes.

3 Question 9.2.a

Model selection: As described in Example 6 of Chapter 7, The file azdiabetes.dat contains data on health-related variables of a population of 532 women. In this exercise we will be modeling the conditional distribution of glucose level (glu) as a linear combination of the other variables, excluding the variable diabetes.

a) Fit a regression model using the g-prior with g = n, $\nu_0 = 2$ and $\sigma_0^2 = 1$. Obtain posterior confidence intervals for all of the parameters.

Using the g-prior with Monte Carlo type sampling. The 95% Confidence Intervals are as follows:

β_1 - npreg:(-1.9789894, 0.02737099)
β_2 - bp: (0.4065772, 0.78241667)
β_3 - skin: (-0.2329478, 0.40783650)
β_4 - bmi: (0.7844507, 1.71106005)
β_5 - ped: (5.6971155, 20.69179467)
β_6 - age: (0.6364824, 1.27011993)
σ^2: (797.3927, 1011.7802)

We see that some of the CIs include zero and that many don't however out model building does not stop here. We want to use our new skills of model averaging and selection. Note that we achieved this output with the following code derived primarily from the book:

R CODE:

```
# setting up data for process

y = diab[,2]
y = as.matrix(y)
X = diab[,-c(2,8)]
X = as.matrix(X)
g = length(y)
nu0 = 2
s02 = 1
S = 10000

#creating sampler to build large matrices to calculate quantile statistics from posterior.

n = dim(X)[1] ; p = dim(X)[2]
Hg = (g/(g+1)) * X%*%solve(t(X)%*%X)%*%t(X)
SSRg = t(y)%*%(diag(1,nrow = n) - Hg )

s2 = 1/rgamma(S, (nu0+n)/2, (nu0*s02+SSRg)/2)
Vb = g*solve(t(X)%*%X)/(g+1)
Eb = Vb%*%t(X)%*%y
E = matrix(rnorm(S*p, 0, sqrt(s2)), S, p)
beta = t(t(E%*%chol(Vb))+c(Eb))

CIbeta = matrix(ncol=2, nrow=6)
for (i in 1:dim(beta)[2]) {
CIbeta[i,] = quantile(beta[,i], probs = c(.025,.975))
```

}

We then sampled the **beta** matrix by column in the usual fashion to attain our confidence intervals.

4 Question 9.2.b

b) Perform the model selection and averaging procedure described in Section 9.3. Obtain $Pr(j \neq 0|y)$, as well as posterior confidence intervals for all of the parameters. Compare to the results in part a).

There are a total of 2^6 possible combinations of models available to us. Lets explore which option(s) are best. Using the code from class the posterior probabilities that each $Z_i \neq 0$ are as follows respectively from $\beta_1, ..., \beta_6$: {0.2100, 1.0000, 0.0480, 1.0000, 0.9538, 1.0000}.

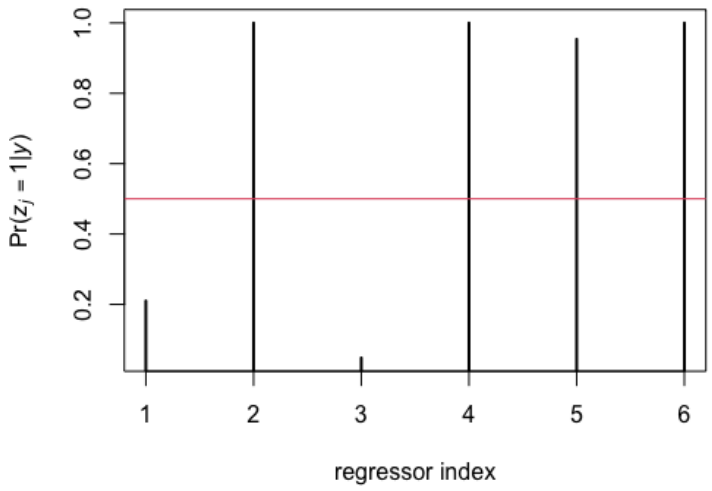

Figure 2:
9.2.b Plot of posterior $P(\beta_i \neq 0)$.

Interestingly our Gibbs sampler for the Z vectors seems to have only generated 8 unique items from across its sample space of $2^6 = 64$ unique elements. Regardless, we can still get a good idea of which parameters tend to be valuable in the regression model.

The 95% posterior confidence intervals for $\boldsymbol{\beta}$ and σ^2 are listed as follows:
CIbetas
β_1: (-1.5952940 , 0.00000000)
β_2: (0.4135406 , 0.79461762)
β_3: (0.0000000 , 0.06940222)
β_4: (0.9923973 , 1.73057512)
β_5: (0.0000000 , 20.92413497)
β_6: (0.5318391 , 1.15576028)

σ^2: (805.0398 , 1020.2481)

It is easy to see the effect of $Z_i = 0$ in the 1st, and 5th β values where the CI is bounded by zero on one side. This creates a situation where the posterior confidence intervals are bounded by zero with none featuring an interval containing both positive and negative numbers. This is in contrast to the CIs in part A where many of the intervals included both positive and negative numbers. The intervals which saw a major change where β_3 and β_5. The CIs for σ^2 remained relatively similar.