HW5: Team 12

Thomas Fleming, Blaire Li, Marc D. Ryser, Hengqian Zhang

Due March 10, 2016

For this assignment we will explore simulation of data to compare methods for estimation and model selection. To get started, refer to the code from Lab6 and simulate the datasets as described there. Some "guideposts" for when to finish parts are provided within the problem set.

1. Add to the Lab6 code a second set of 100 datasets for testing (prediction) with 25 observations, but where the X's have the same correlation matrix as in the training data. Provide a brief description of the model that generated the data and summary of the simulation study. (ie dimensions, true β etc, number of simulated datasets etc.).

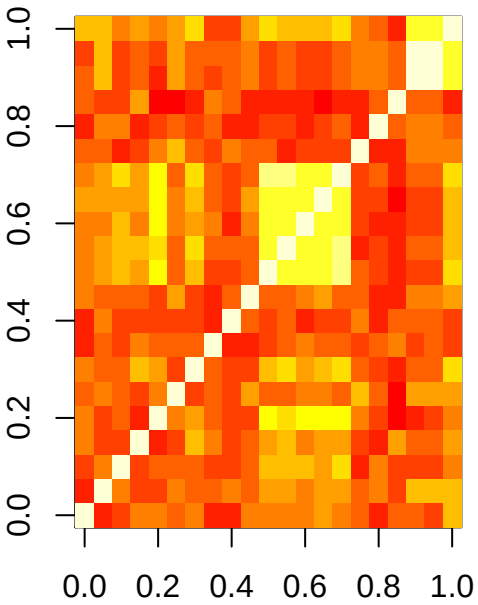
ANSWER: The true betas' we choose is the same as the betas' in lab6 which are {4,2,0,0,0,-1,0,1.5,0,0,0,1,0,0.5,0,0,0,0,-1,1,4}. We simulated 100 datasets and each with dimension 25 by 21. The procedure we use to generate the data is exactly the same as we did in lab 6. Considering the training data set, we wanted to simulate 100 datasets each with dimension 75 by 23 (21 for beta plus Y and mu) For each dataset, first we generated 75 by 10 from standard normal distribution.

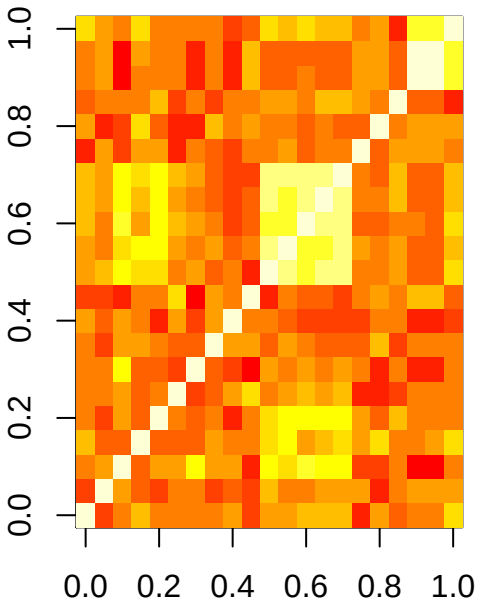
Secondly, we picked up the last 5 columns of the data we generated in step 1 and did matrix multiplication with {0.3,0.5,0.7,0.9,1.1} to get 75$\times$ 1 vector and replicated it 5 times to get X1 with dimension 75 $\times$ 5

Thirdly, we generated 75 $\times$ 4 matrix from standard normal distribution and picked up the 4th column and added error from normal(0,0.1) to get X2 and X3 respectively.

Finally, we used cbind to combine X1, X2, X3 to get desired dataset and also computed Y and mu.

In addition, we can see the similar pattern in both correlation heatmap of training data and testing data. In simulation, there are 5 variables(columns) have correlation to each other. In heat map, there are ranctangles around the diagnal indicating the same patterns.





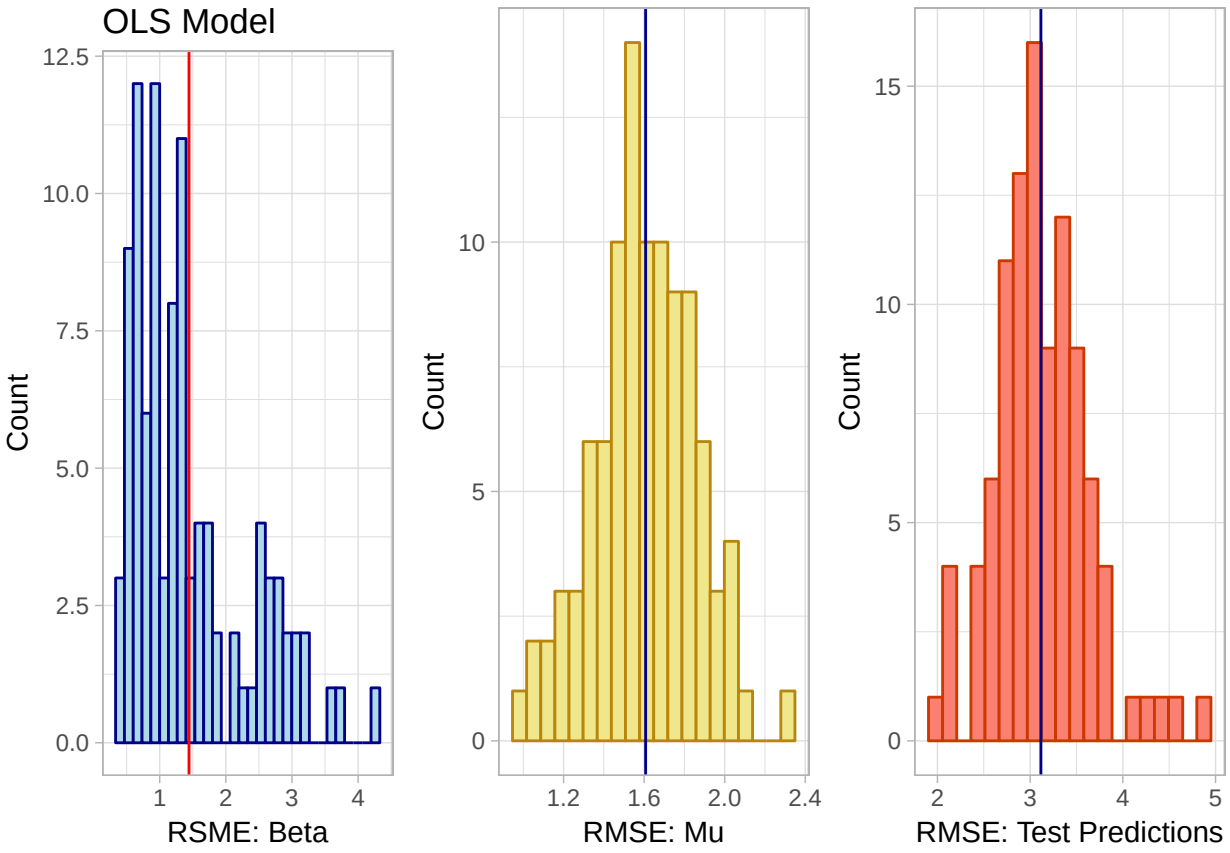
2. Using Ordinary Least squares based on fitting the full model for each of the 100 data sets, Compute the average RMSE for a) estimating β^{true}, b) estimating $\mu^{true} = X\beta^{true}$ and c) out of sample prediction for the test data from the 100 data sets. Present histograms of the RMSEs and show where the average

falls. Note for a vector of length d, RMSE is defined as

$$RMSE(\hat{\theta}) = \sqrt{\sum_{i=1}^{d}(\hat{\theta}_j - \theta_j)^2/d}$$

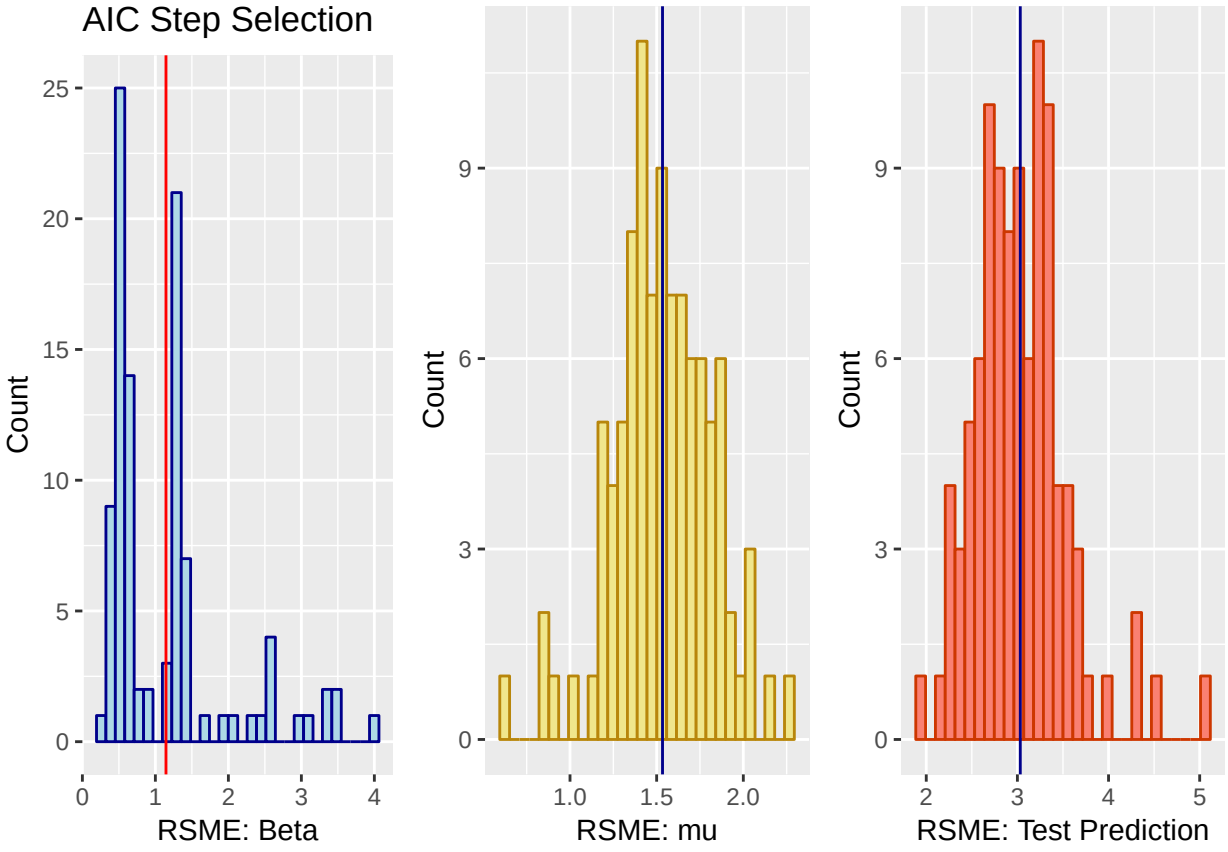
Table 1: Average RMSE's

beta	mu	test
1.442778	1.60756	3.116084



The average RMSE for β^{true}, μ^{true} and out of sample prediction are 1.4427782, 1.6075596, and 3.1160843, respectively.

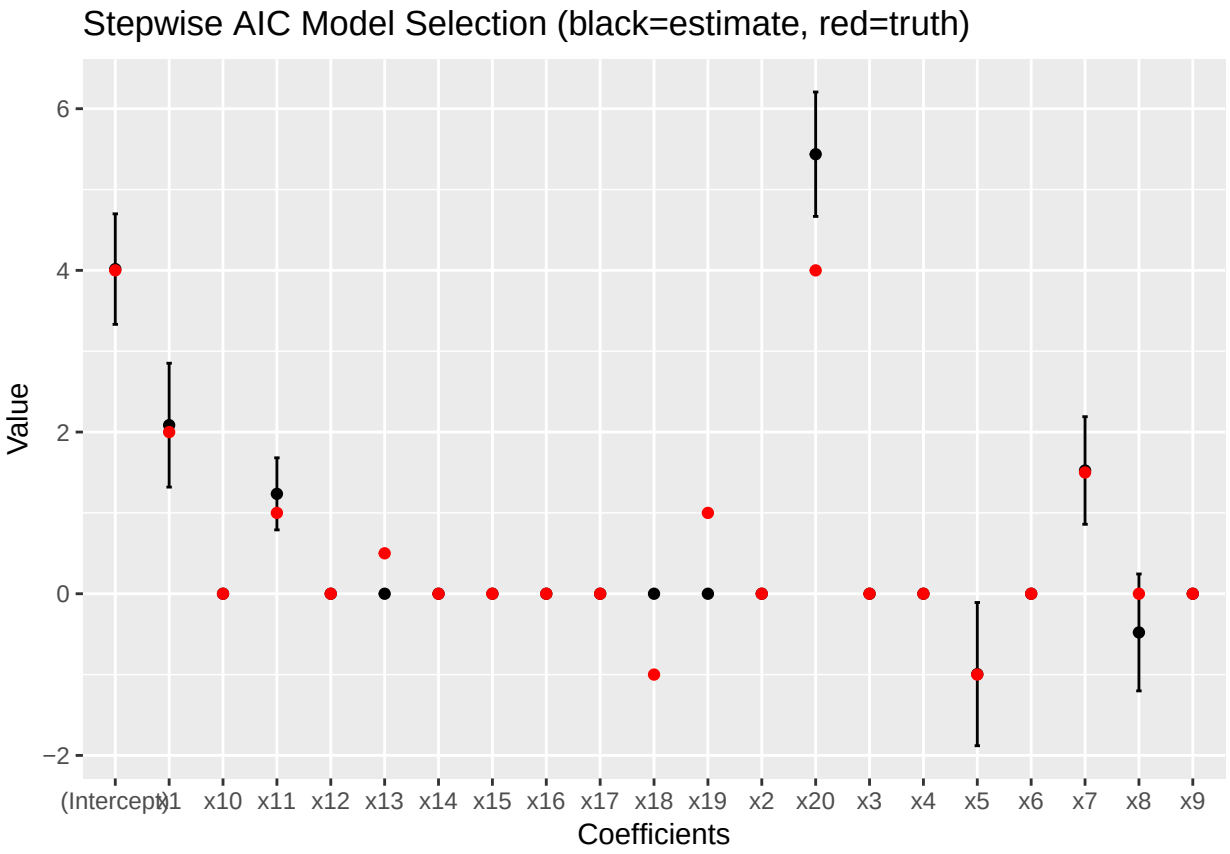
3. Use AIC with either stepwise or all possible subsets to select a model and then use OLS to estimate the parameters under that model. Using the estimates to compute the RMSE for a) estimating β^{true}, b) estimating μ^{true}, and c) predicting Y^{test}. Present histograms of the RMSE, and show where the average RMSE falls on the plot. Also report d) the number of times you select the true model using AIC out of the 100 simulations.




```
## [1] 0
```

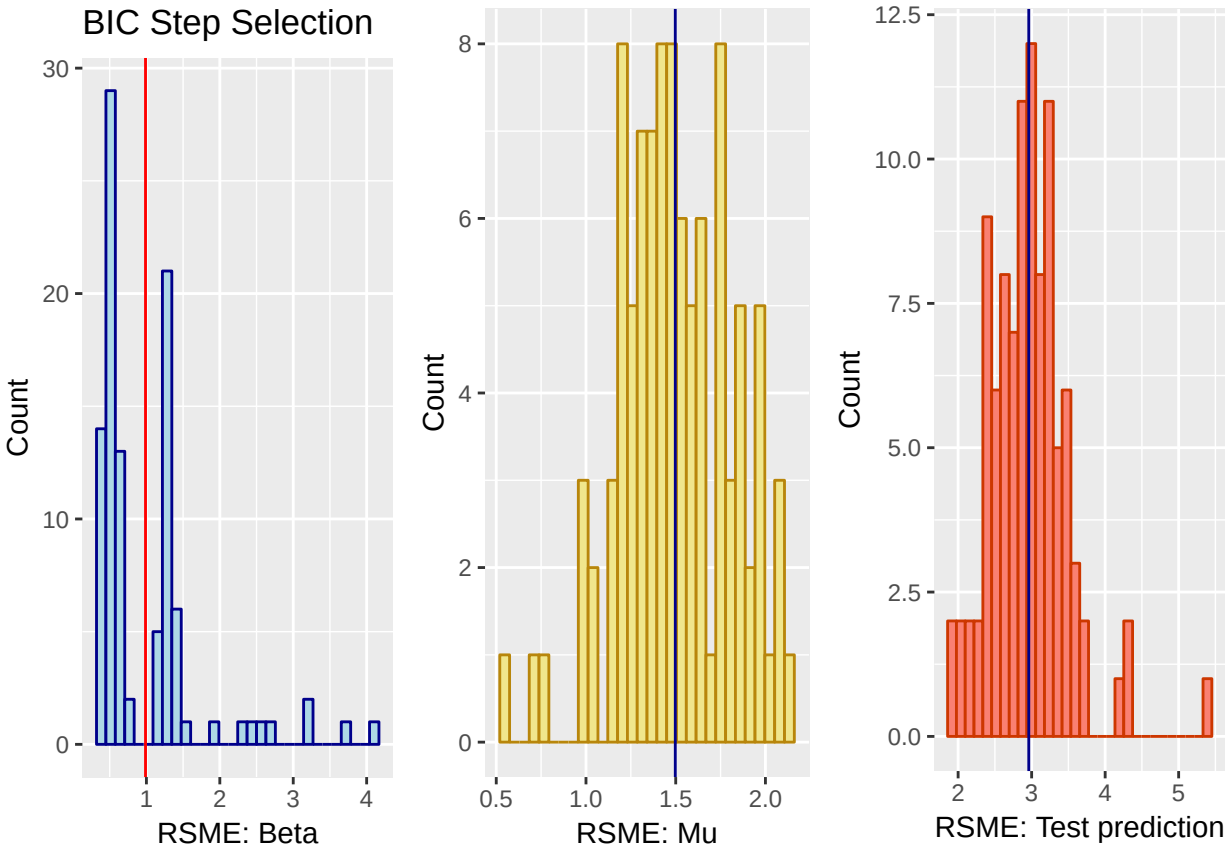
The number of "correct" models is `count_true` which is equal to 0. The average RMSE for β^{true}, μ^{true} and out of sample prediction are 1.1452582, 1.533792, and 3.0315755, respectively. The full distributions are found in the plots above.

4. Take a look at the summaries from the estimates under the best AIC model from the simulation that is equal to your team number. Create confidence intervals for the β's and comment on whether they include zero or not or the true value.



From the figure we see that with the exception of coefficients 13, 18, 19 and 20 the confidence intervals include the true value of β. Coefficients 13, 18 and 19 are not included in the model (set to zero) although their true values are non-zero. Conversely, coefficeint 8 is included and non-zero although the true value is zero.

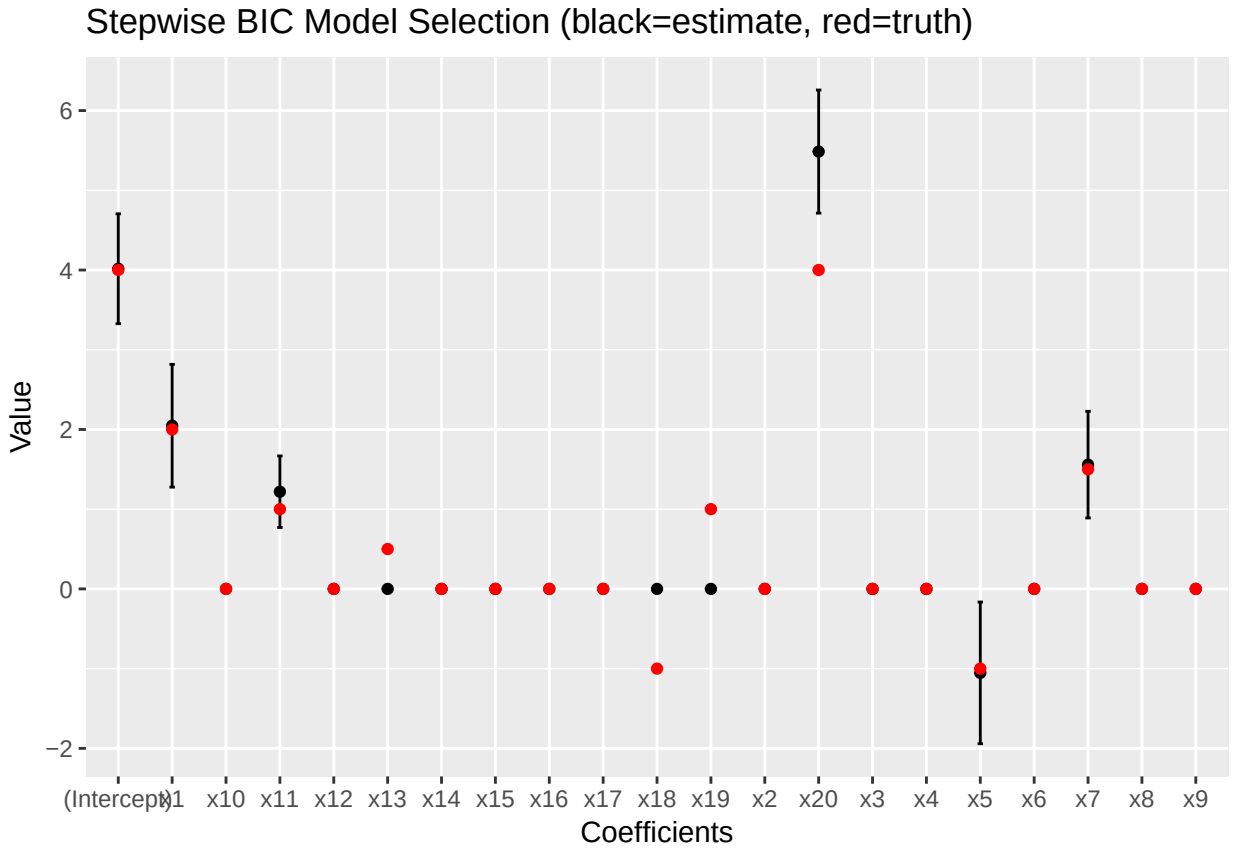
5. Use BIC with either stepwise or all possible subsets to select a model and then use OLS to estimate the parameters under that model. Use the estimates to compute the RMSE for a) estimating β^{true}, b) μ^{true}, and c) predicting Y^{test}. Present histograms of the RMSE, and show where the average RMSE falls on the plot. Also report d) the number of times you select the true model using BIC out of the 100 simulations.




```
## [1] 0
```

The number of "correct" models is `count_true` which is also equal to 0. The average RMSE for β^{true}, μ^{true} and out of sample prediction are 0.9882443, 1.4977411, and 2.9613045, respectively.

6. Take a look at the summaries from the estimates under the best BIC model from the simulation that is equal to your team number. Create confidence intervals for the β's and comment on whether they include zero or not or the true value.



From the figure we see that with the exception of coefficients 13, 18 and 19, and 20, the confidence intervals include the true value of β. Coefficeints 13, 18 and 19 are not included in the best model although their true values are non-zero.

7. Theory (work individually and then combine to add group solution, try to complete by Wednesday before class) For the linear model, assume that the X have been centered so that they all have mean 0. For the linear model

$$Y \sim N(1_n\beta_0 + X\beta, I_n/\phi)$$

using Zellner's g-prior for β with

$$\beta \mid \beta_0, \phi \sim N(0, g(X^TX)^{-1}/\phi)$$

and the improper independent Jeffrey's prior

$$p(\beta_0, \phi) \propto 1/\phi$$

find the a) posterior distriubtion of $\beta \mid Y, g, \phi$, b) posterior distribution of $\mu_i = x_i^T\beta \mid Y, g, \phi$ and c) the posterior predictive distribution of $Y^{test} \mid Y, g, \phi$ as functions of the OLS/MLE summaries. *(you may use results in notes - just quote - or derive)*

ANSWER: While we provide short answers here to #7, #8, and #9, full derivations of each problem are shown at the bottom of the document in the Appendix.

a) For a short answer, we reference the class notes which provide the posterior

$$\beta|Y, \phi, g \sim \mathcal{N}\left(\frac{g}{1+g}\hat{\beta}, \frac{g}{1+g}(X^tX)^{-1}\phi^{-1}\right),$$

and

$$\beta_0 \,|\, Y, \phi \sim \mathcal{N}(\bar{Y}, I_n/n\phi).$$

b) We note that technically $\mu_i = \beta_0 + x_i^t\beta$, but here we are asked to consider only the second term,

$$\mu_i = x_i^t\beta|Y, g, \phi.$$

This is an affine transformation of a Gaussian, and hence by standard results (see, e.g., the book by Christensen on linear models), we know that μ_i is normally distributed as :

$$\mu_i \sim \mathcal{N}\left(x_i\frac{g}{1+g}\hat{\beta}, \phi^{-1}\frac{g}{1+g}x_i(X^TX)^{-1}x_i^t\right).$$

Note: In the derivations below, we replaced "test" by a $*$.

c) Because

$$Y^*|Y, g, \phi = (1_n\beta_0 + X^*\beta + \epsilon^*)|Y, g, \phi$$

is the sum of three Gaussians, it suffices to calculate the mean and variance.

$$E[Y^*|Y, g, \phi] = E[1_n\beta_0 + X^*\beta + \epsilon^*|Y, g, \phi]$$

$$= E[1_n\beta_0|Y, g, \phi] + E[X^*\beta|Y, g, \phi] + E[\epsilon^*|Y, g, \phi]$$

$$= \bar{Y} + x_i\frac{g}{1+g}\hat{\beta} + 0$$

$$= \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}$$

For the calculation of the variance, we note that the β_0 and β are independent random variables conditioned on Y and ϕ, and ϵ^* is independent of any past observation and hence independent of the model coefficients. Therefore, we can write

$$Var[Y^*|Y, g, \phi] = Var[1_n\beta_0 + X^*\beta + \epsilon^*|Y, g, \phi]$$

$$= Var[1_n\beta_0|Y, g, \phi] + Var[X^*\beta|Y, g, \phi] + Var[\epsilon^*|Y, g, \phi]$$

$$= \frac{I_n}{n\phi} + X^*\frac{g}{1+g}\phi^{-1}(X^TX)^{-1}X^{*t} + \frac{I_n}{\phi}$$

$$= \frac{1}{\phi}\left(I_n\left(1+\frac{1}{n}\right) + X^*\frac{g}{g+1}(X^TX)^{-1}X^{*t}\right)$$

8. What are the corresponding distributions in 7) unconditional on ϕ? (hint recall theorem from class) Are β_0 and β still independent? Explain.

ANSWER: No, β and β_0 are no longer independent. In fact, they have a joint dependency on ϕ which is now integrated out and entangles the two.

First we take the shortcut, and then we do the actual calculations.

a) From the theorem for the marginal distribution of a Normal-Gamma, we know that if $\theta \sim N(m, 1/\phi\Sigma)$ and $\phi \sim G(\nu/2, \nu\hat{\sigma}^2/2)$ then $\theta \sim t_\nu(m, \hat{\sigma}^2\Sigma)$. Here we have (see class notes) $\beta\,|\,Y, \phi \sim N(m, 1/\phi\Sigma)$ for $m = \frac{g}{1+g}\hat{\beta}$ and $\Sigma = \frac{g}{1+g}(X^tX)^{-1}$, and $\phi\,|\,Y \sim G(\nu/2, \nu\hat{\sigma}^2/2)$ with $\nu = n-1$ and

$$\hat{\sigma}^2 = (n-1)\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right).$$

It follows from the above theorem that β is a multivariate Student t distribution

$$\beta \,|\, Y \sim t_{n-1}\left(\frac{g}{1+g}\hat{\beta}, \frac{g(n-1)}{1+g}\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right)(X^tX)^{-1}\right).$$

Similarly, we find that for $\beta_0 \,|\, Y, \phi \sim \mathcal{N}(\bar{Y}, \frac{1}{\phi}(\frac{I_n}{n}))$ we obtain

$$\beta_0 \,|\, Y \sim t_{n-1}\left(\bar{Y}, \frac{n-1}{n}\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right)I_n\right).$$

b) We have that $\mu_i \,|\, Y \sim x_i^t\beta \,|\, Y$ and hence by basic results on linear transformations of Student t distributions we find

$$\mu_i \,|\, Y \sim t_{n-1}\left(\frac{g}{1+g}x_i^t\hat{\beta}, \frac{g(n-1)}{1+g}\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right)x_i^t(X^tX)^{-1}x_i\right).$$

c) From 7c) we know that $Y^*|Y, \phi = (1_n\beta_0 + X^*\beta + \epsilon^*)|Y$ has distribution

$$Y^*|Y, \phi \sim \mathcal{N}\left(\bar{Y} + x_i\frac{g}{1+g}\hat{\beta}, \frac{1}{\phi}\left(I_n\left(1+\frac{1}{n}\right) + X^*\frac{g}{g+1}(X^TX)^{-1}X^{*t}\right)\right).$$

Furthermore, $\phi \,|\, Y \sim G(\nu/2, \nu\hat{\sigma}^2/2)$ with $\nu = n-1$ and

$$\hat{\sigma}^2 = (n-1)\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right).$$

It follows from the above theorem that $Y*|Y$ has a multivariate Student t distribution with

$$Y^*|Y \sim t_{n-1}\left(\bar{Y} + x_i\frac{g}{1+g}\hat{\beta}, (n-1)\left(I_n\left(1+\frac{1}{n}\right) + X^*\frac{g}{g+1}(X^TX)^{-1}X^{*t}\right)\left(SSE - \frac{1}{1+g}\hat{\beta}^t(X^tX)^{-1}\hat{\beta}\right)\right)$$

9. Let $\tau = 1/g$ and substitute that in the prior for β

$$\beta \mid \beta_0, \phi \sim N(0, (X^TX)^{-1}/(\tau\phi))$$

If $\tau \sim G(1/2, n/2)$, show that the prior on β is a Cauchy Distribution

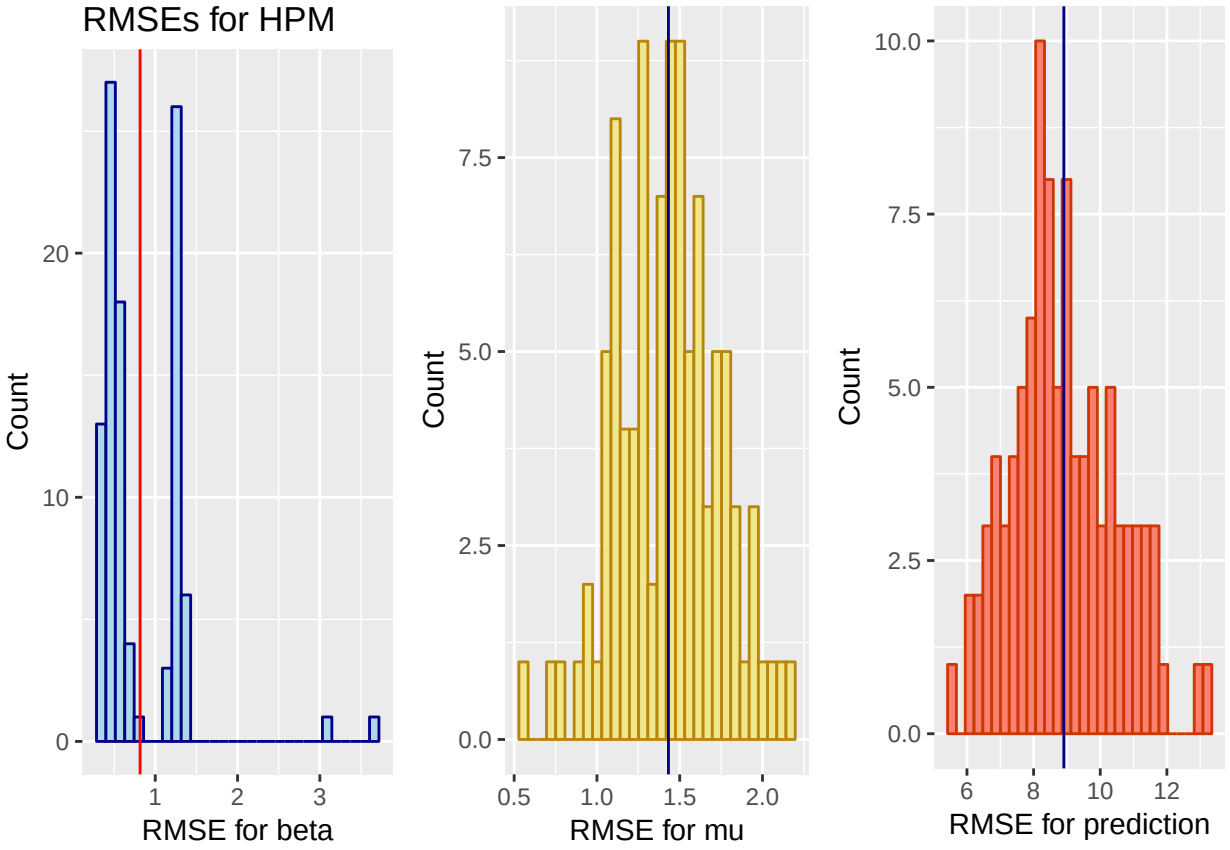
$$\beta \mid \phi, \beta_0 \sim C(0, (X^TX/n)^{-1}/\phi)$$

(a Cauchy distribution is a Student t with 1 df - see notes for density)

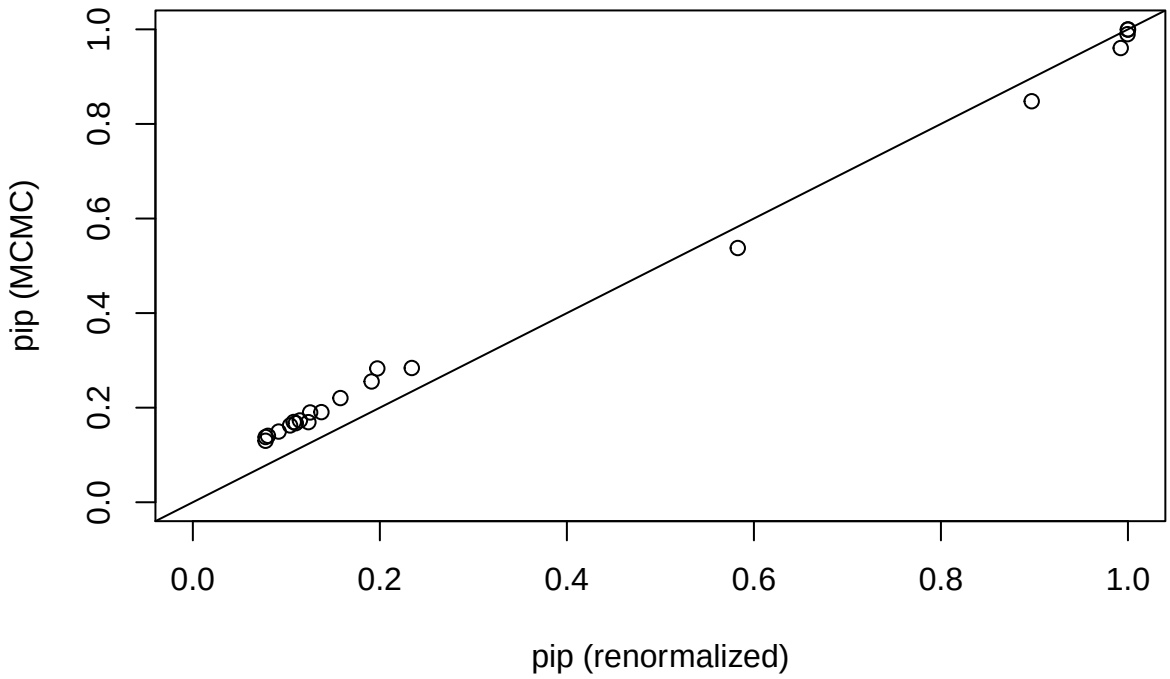
ANSWER: Again we use the theorem on the marginal of a normal-gamma. Here, we have that $\beta \,|\, \beta_0, \phi, \tau \sim \mathcal{N}(m, \frac{1}{\tau}\Sigma)$ with $m = 0$ and $\Sigma = (X^tX\phi)^{-1}$, and $\tau \sim G(\nu/2, \nu\hat{\sigma}^2/2)$ with $\nu = 1$ and $\hat{\sigma}^2 = n$. It follows that the distribution

$$\beta \,|\, \beta_0, \phi \sim t_1(0, n(X^tX\phi)^{-1}).$$

10. Using Bayesian variable selection under the g-prior with $g = n$ and a uniform prior distribution over models, find the highest posterior probability model (HPM) using `bas.lm` from library `BAS` (or other software). (If you use `BAS`, please download `BAS` version 1.4.3 from CRAN). Using the mean of the appropriate posterior distribution under the HPM, find the average RMSE for a) estimating β^{true}, b) estimating μ^{true} and c) predicting Y^{test}. Plot histograms of the RMSE and add the average RMSE to the plots. What proportion of the time did you select the true model? Your answer should describe whether you used enumeration or MCMC, number of iterations or models, etc. If you used MCMC, check diagnostic plots to examin convergence.
Note `BAS` has functions to compute the fitted values `fitted` and predicted values `predict` for the HPM (see the vignette or help files), however, to find the posterior mean for the beta's for a given model, we need to extract the information from the object. The following function can be use to do this.

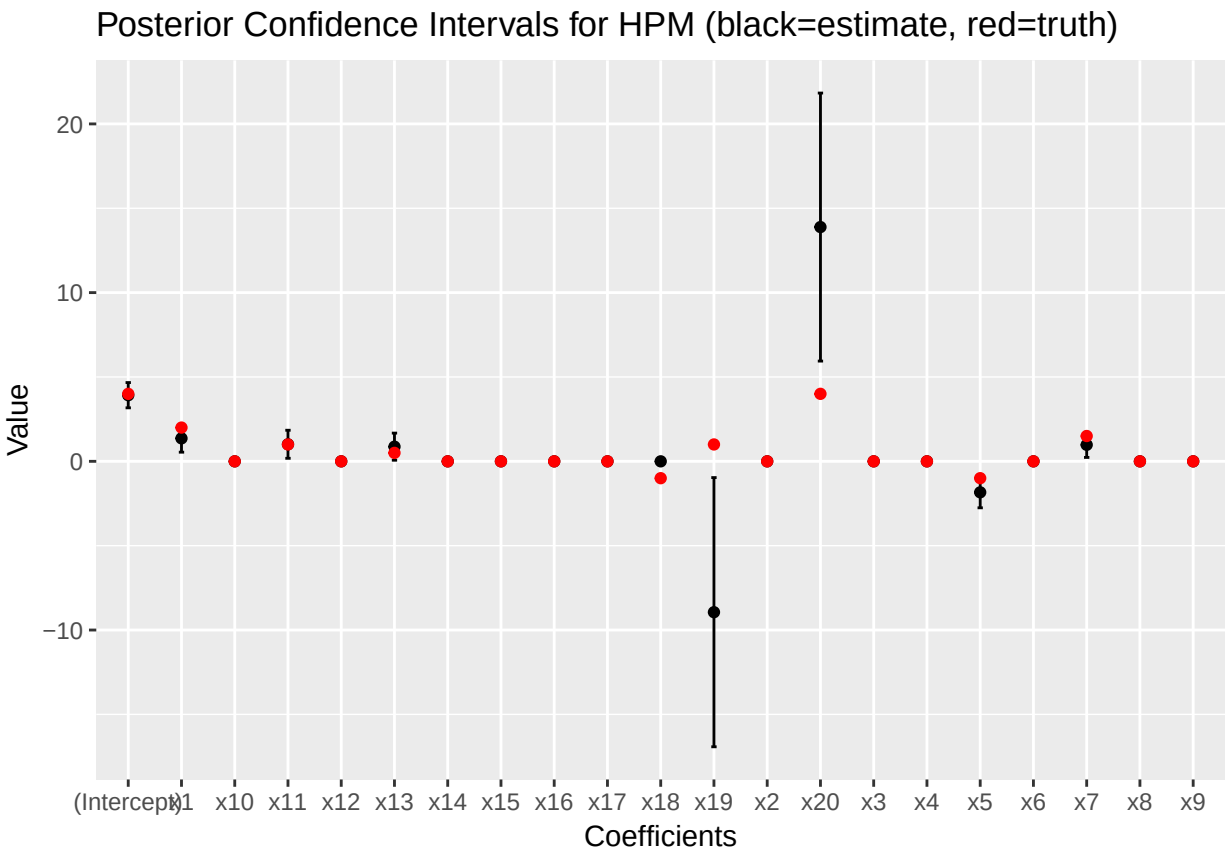


The number of times the "correct" model is selected is 0. The mean RMSEs for β^{true}, μ^{true} and Y^{pred} were 0.8153876, 1.4317516 and 2.8926101, respectively. Since we used MCMC, we checked the 'pip' plots for several datasets. We found them to be satisfatory, see below for the diagnostics of the last data set.



11. Using the simulation that is equal to your team numbers, provide posterior summaries of the coefficient's

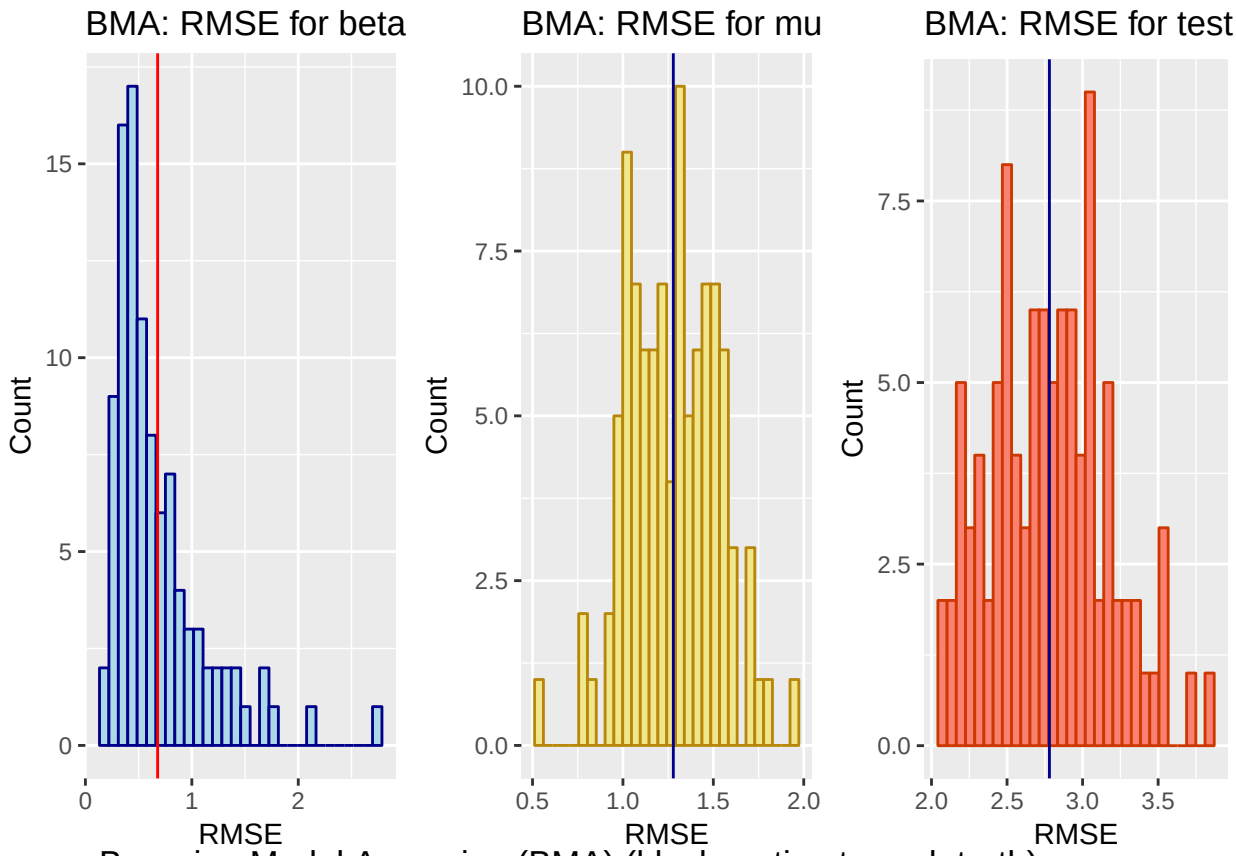
of the HPM, such as Bayesian Confidence intervals.
Comment on whether the intervals contain the true value or zero.

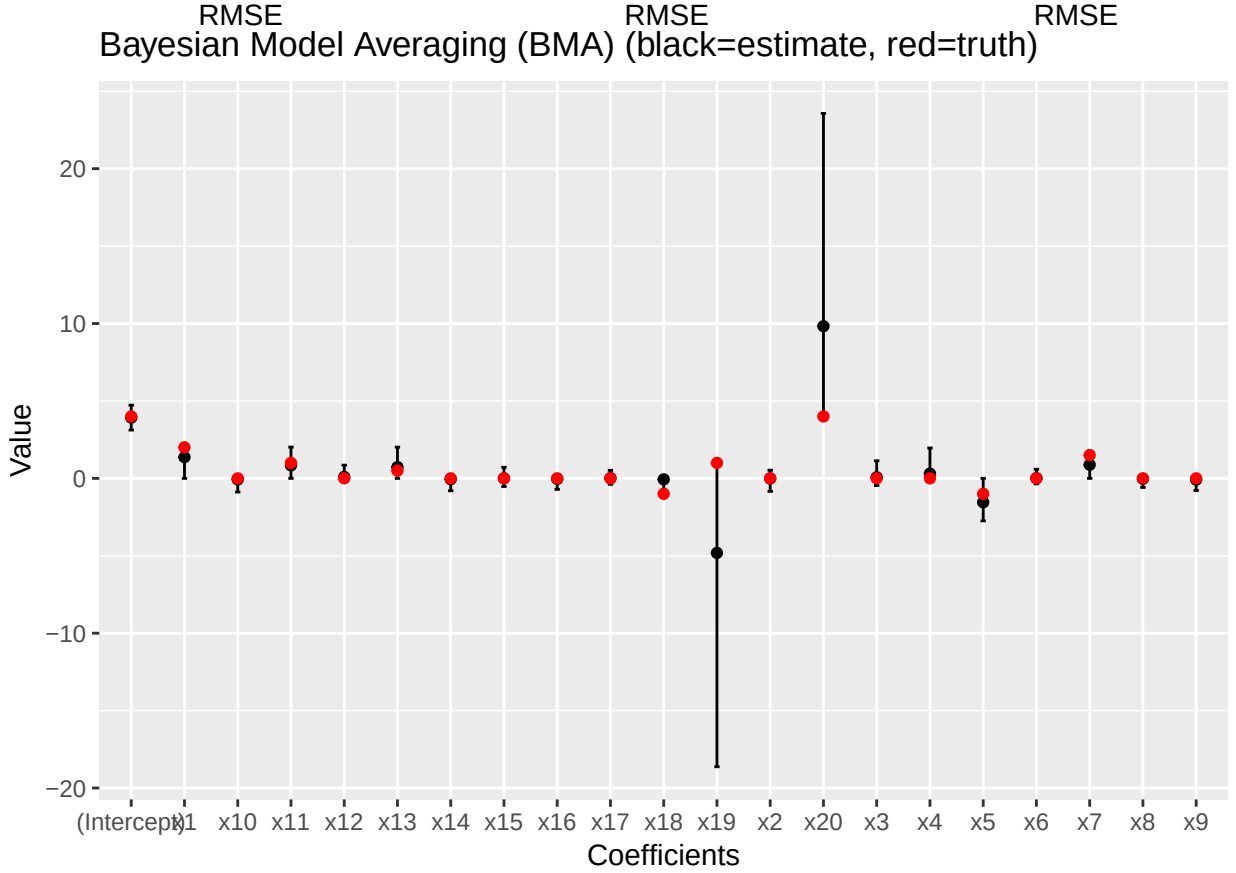


With the exception of x18, x19, and x20, all true coefficients are contained in the CI. With the exception of x18, the HPM included the right coefficients. Of note, x19 and x20 have very large CIs. It looks like there is something wrong here.

12. To incorporate model uncertainty we could use Bayesian Model Averaging, rather than the highest probability model. Repeat 10 and 11 using BMA for estimating the quantities.

```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

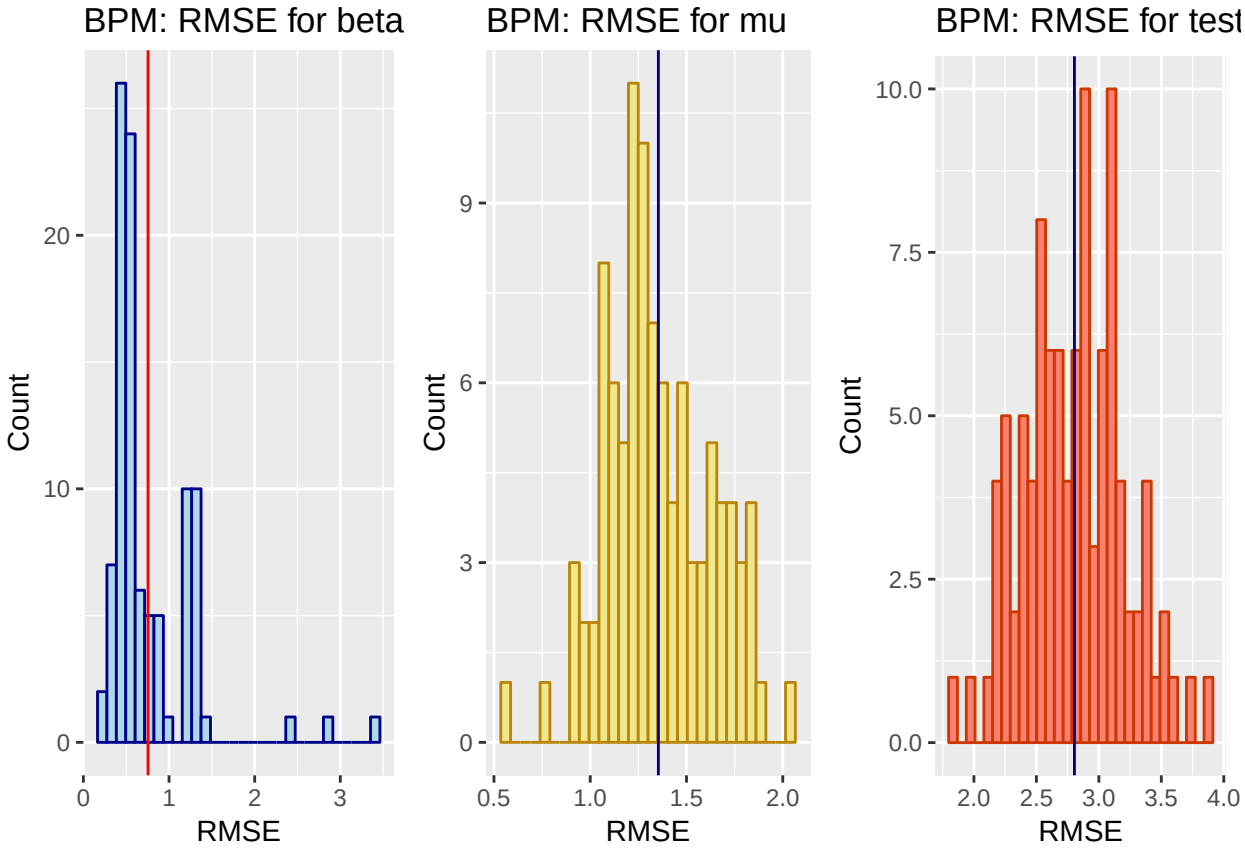


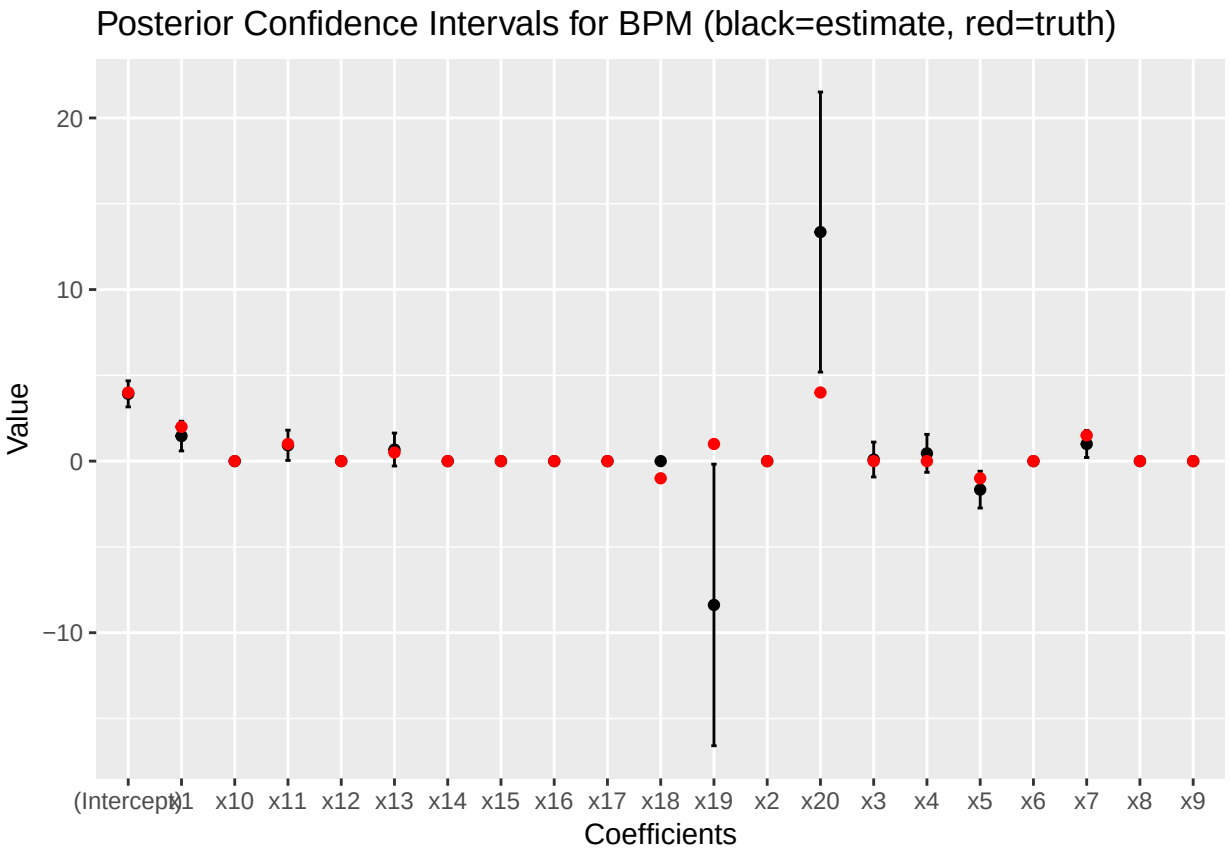


From The figure we see that the true values of all coefficeint are contained in the posterior credible intervals.

19 and 20 are, once more, problematic in the sense that there is a very large credible interval.

13. If we wanted to select the model that is "closest" to BMA, we could use the model whose predictions are closest to BMA using squared error loss. We can find the best predictive model `BPM` from `BAS` using the predict function with `estimator="BPM"`. Repeat 10 and 11 using the Best Predictive Model, `BPM`.

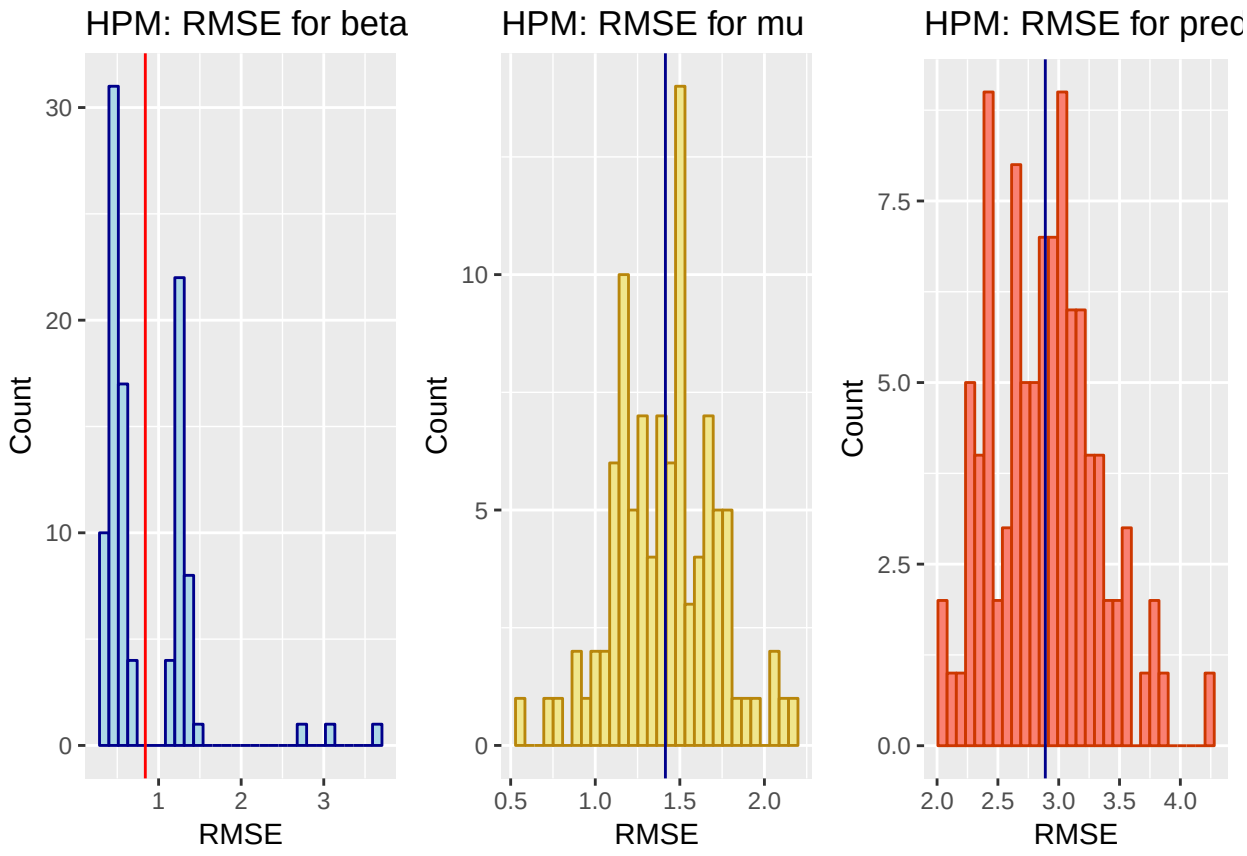


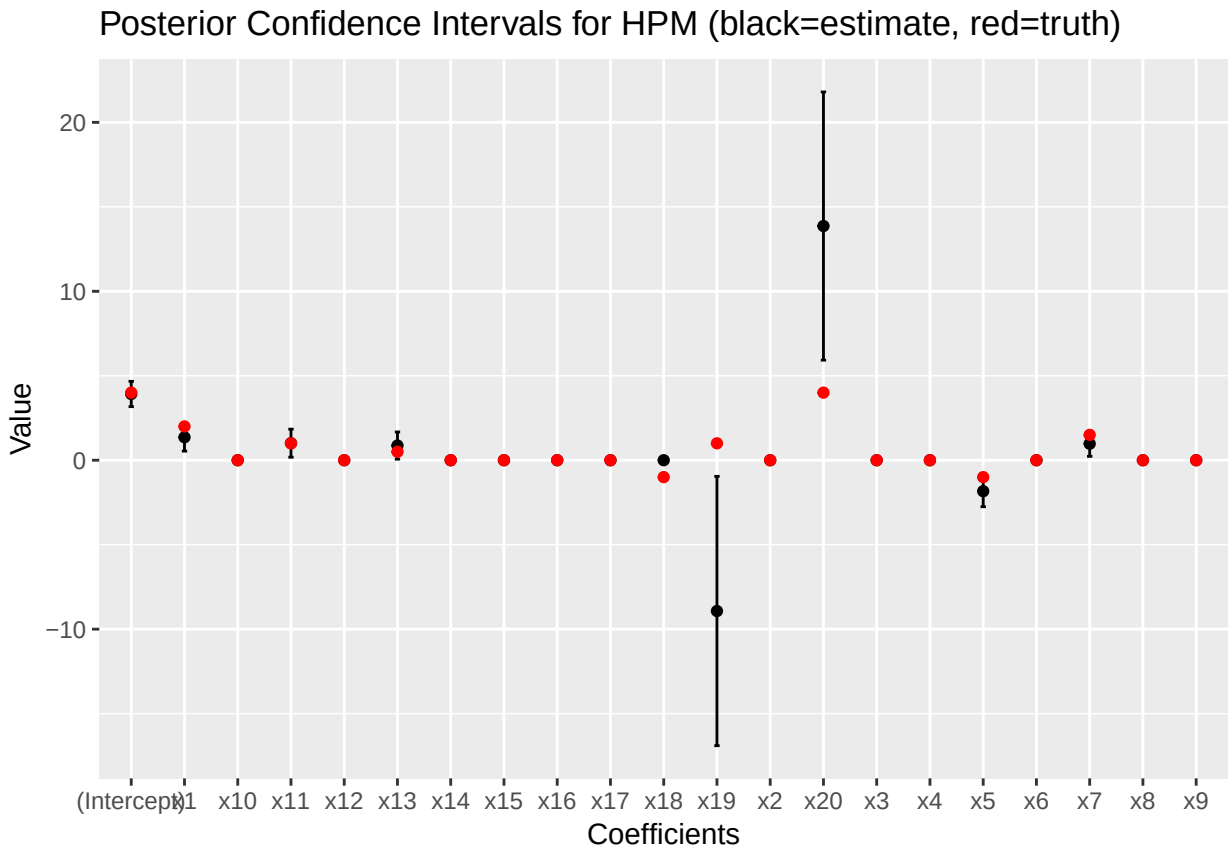


Qualitatively the conclusions to draw here are exactly as in Problem 11.

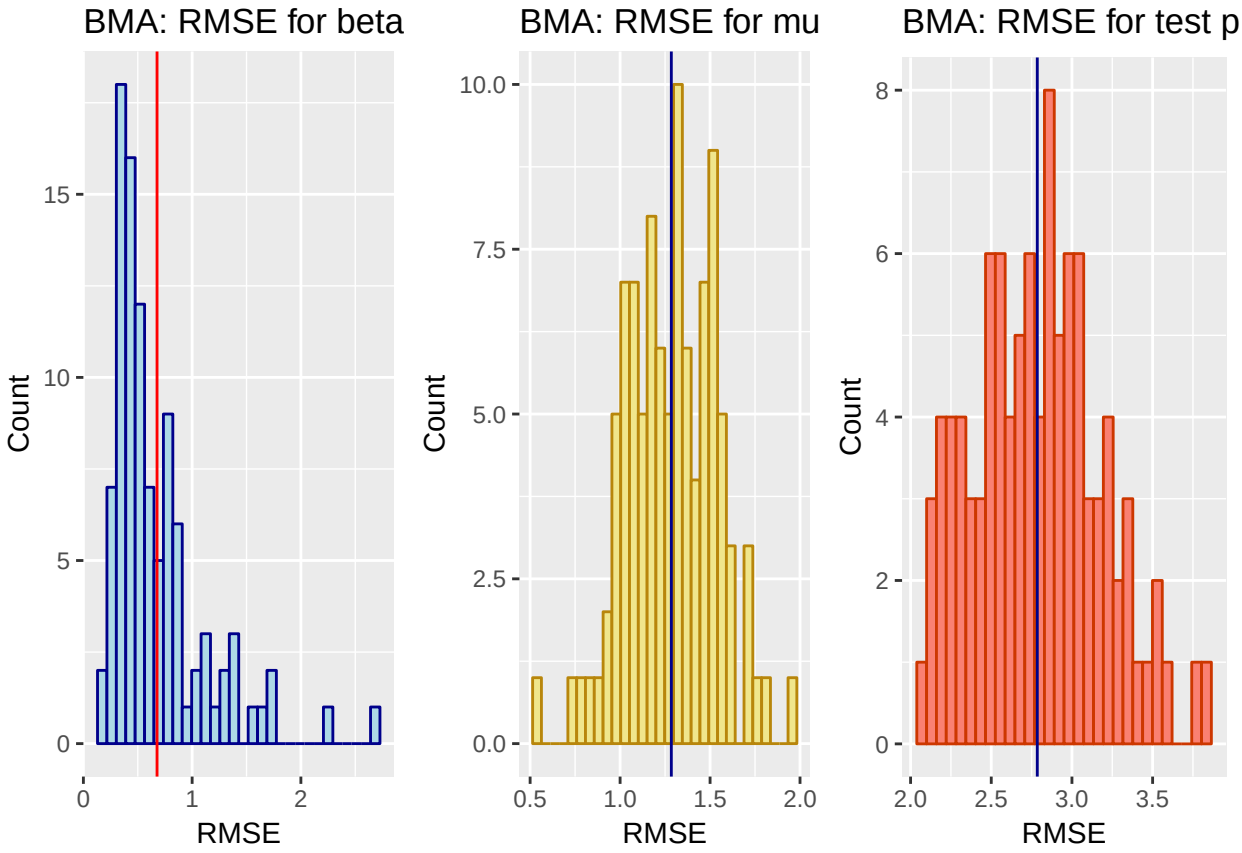
14. Are the Bayesian estimates sensitive to the choice of prior? Try 10-13 using the Zellner-Siow Cauchy prior (option `prior = "ZS-null"` in `bas.lm`)

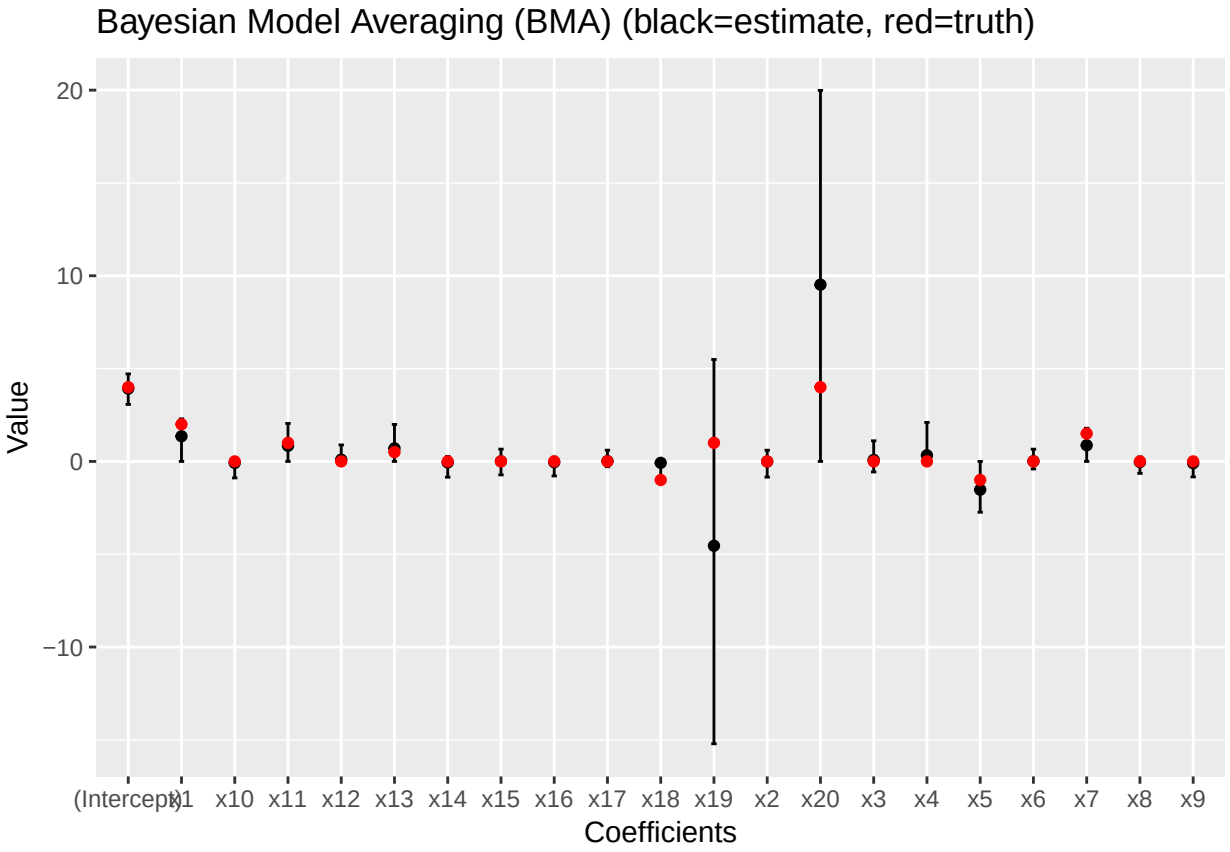
```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


HPM: RMSE for beta
HPM: RMSE for mu
HPM: RMSE for prec
Count
RMSE
30
20
10
0
1
2
3
10
5
0
0.5
1.0
1.5
2.0
7.5
5.0
2.5
0.0
2.0
2.5
3.0
3.5
4.0

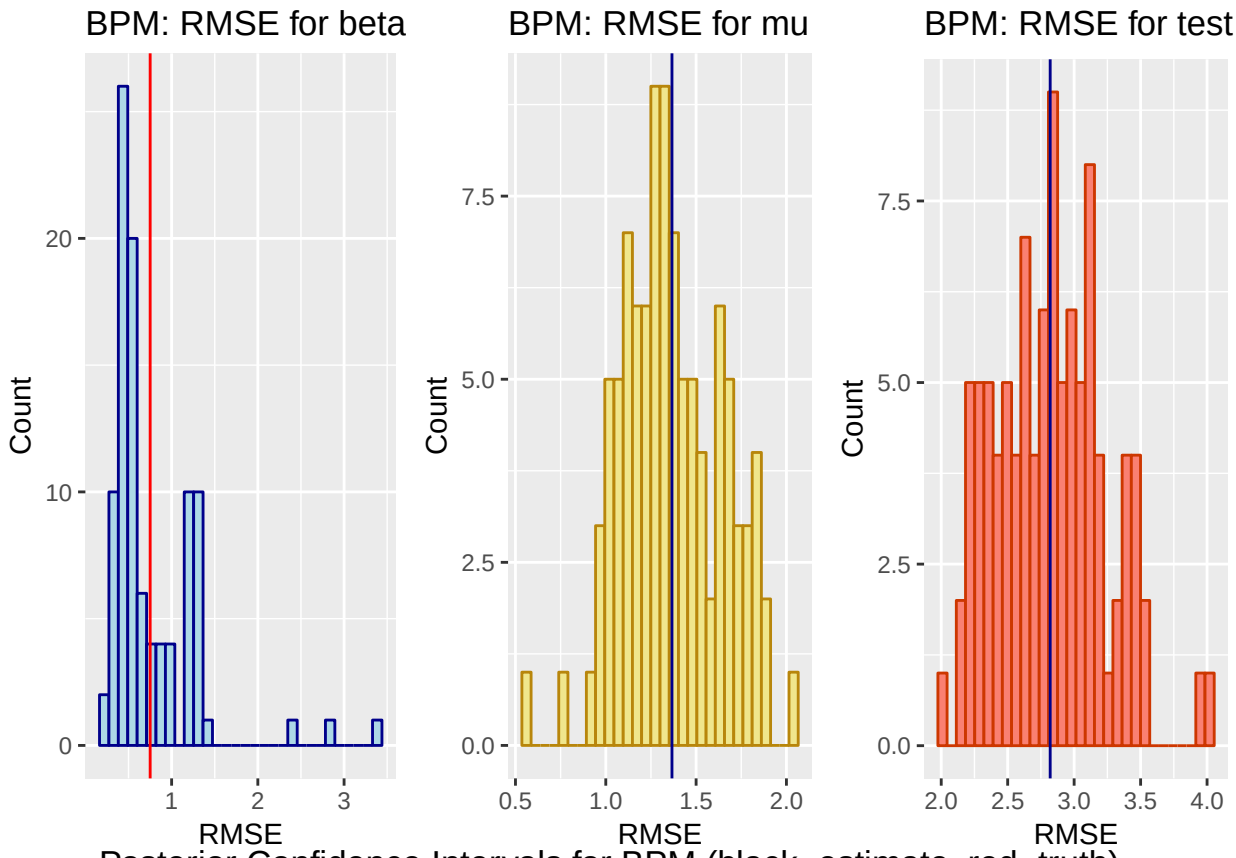

Posterior Confidence Intervals for HPM (black=estimate, red=truth)
Value
Coefficients
20
10
0
−10
(Intercept)
x1
x10
x11
x12
x13
x14
x15
x16
x17
x18
x19
x2
x20
x3
x4
x5
x6
x7
x8
x9

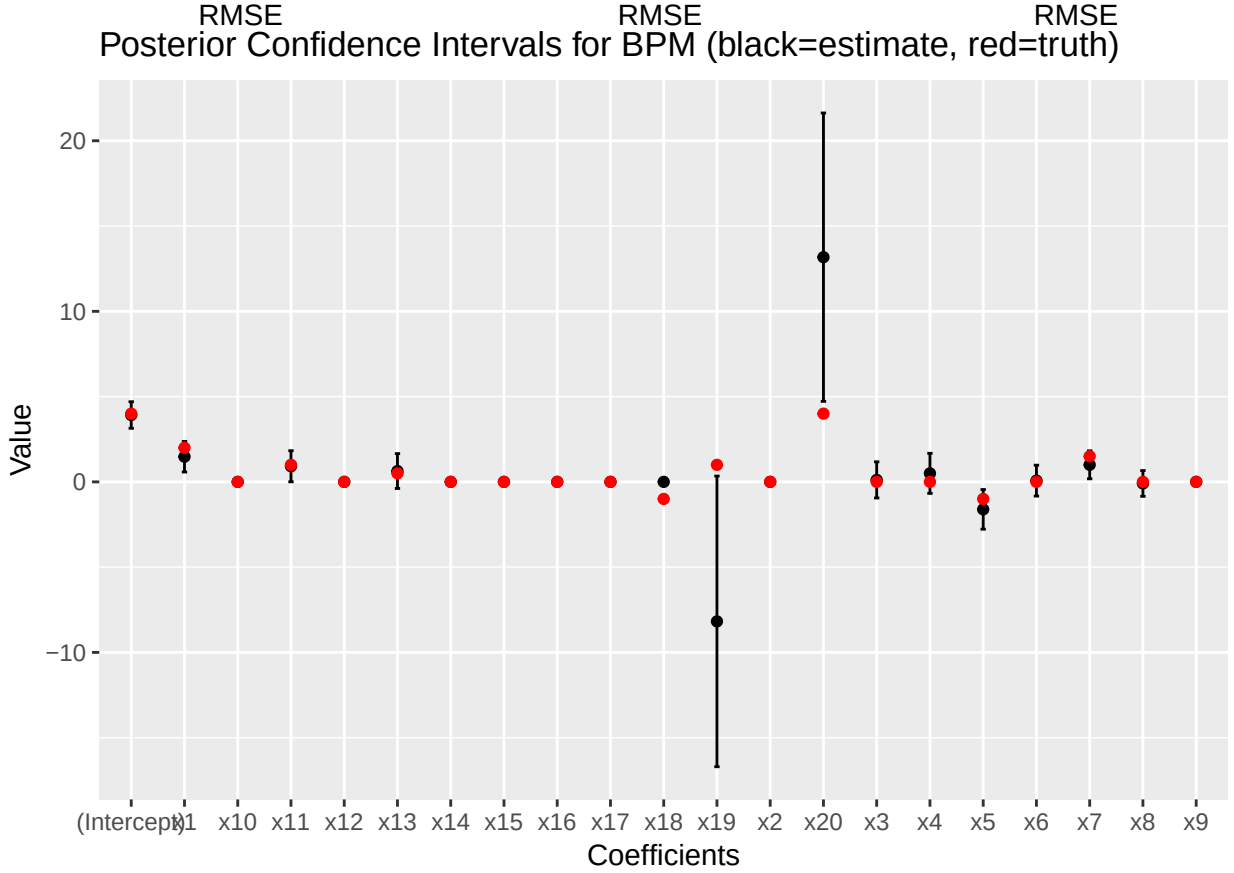
```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```






```
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type data.frame. Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





15. Provide a summary of your simulation findings, with a table for the RMSEs for the different methods

and parameters of interest β, μ, Y^{test} and proportion of time true model was selected. Does any one method seem to do better than the others or are some methods better for one estimation/prediction problem than the others? Explain. For the energetic team - what about coverage?

Table 2: Average RMSE's by Method

	beta	mu	test
OLS	1.4427782	1.607560	3.116084
AIC	1.1452582	1.533792	3.031576
BIC	0.9882443	1.497741	2.961304
HPM.ZG	0.8153876	1.431752	2.892610
BMA.ZG	0.6788914	1.278761	2.780373
BPM.ZG	0.7551912	1.353919	2.804343
HPM.ZS	0.8394703	1.414578	2.889874
BMA.ZS	0.6771972	1.284713	2.785178
BPM.ZS	0.7511921	1.366418	2.819803

Table 3: Proportion of Time True Model Selected

	Proportion
AIC	0
BIC	0
HPM	0
BPM	0
ZS.HPM	0
ZS.BPM	0

ANSWER:

The test RMSE generated by the Bayesian Model Averaging with the Zellner g-prior is the lowest, so we may accept this as the prima facie best method. While none of the sampling models we generated were equal to the true model, the credible interval plot for the BMA Zellner g-prior is as good as any other, as most of the true values are contained within the posterior credible intervals.

Appendix: Derivations for #7, #8, and #9

7.

a) By Bayes' theorem we have

$$p(\beta|Y, g, \phi) \propto p(Y|\beta, g, \phi)p(\beta|g, \phi)$$

Since Y is independent of g and Zellner's g-prior for β is actually independent of β_0, we can plug in the given distributions for Y and β:

$= N(1_n\beta_0 + X\beta, I_n/\phi,)N(0, g(X^tX)^{-1}/\phi)$

$$\propto e^{-((y-(1_n\beta_0+X\beta))^T(I_n/\phi)^{-1}(y-(1_n\beta_0+X\beta)))/2}e^{-\beta^t(g(X^TX)^{-1}/\phi)^{-1}\beta/2}$$

$$\propto e^{-\frac{1}{2\sigma^2}((y-(1_n\beta_0+X\beta))^T(y-(1_n\beta_0+X\beta)))}e^{-\frac{1}{2\sigma^2}\beta^t(X^TX)/g\beta}$$

$$\propto e^{-\frac{1}{2\sigma^2}(y^Ty-2\beta^TX^Ty-\beta^TX^TX\beta-2_n\beta_0^Ty+1_n\beta_01_n\beta_0+\beta^T((X^TX)/g)\beta)} = \beta^TA\beta - 2\beta B + C$$

where

$$A = \frac{(1+g)X^TX}{g\sigma^2}$$

,

$$B = \frac{X^Ty}{\sigma^2}$$

, and

$$c = c(y, \beta_0)$$

Since c is not a function of β, we can leave it out of the following proportion and rewrite the above in terms of A and B as,

$$e^{-\frac{1}{2}(\beta^TA\beta-2\beta B)}$$

While completing the square in this situation is tedious, we can use knowledge of the strategy for the univariate normal, outlined on page 70 of Hoff, where it is easier to calculate the posterior mean and variance. It can be shown that the formulas for the posterior mean and variance also apply to the multivariate case, and so we know that the posterior mean is equal to

$$\frac{A}{B}$$

and the posterior variance is equal to

$$\frac{1}{A}$$

. If we expand these equations out, using the right-hand side of equations A and B, and simplify, we find that we have a normal distribution with:

$$Var[\beta|Y, g, \phi] = \frac{g}{g+1}\sigma^2(X^tX)^{-1}$$

$$E[\beta|Y, g, \phi] = \frac{g}{g+1}(X^TX)^{-1}X^TY = \frac{g}{1+g}\hat{\beta}$$

In summary, we find that

$$\beta|Y, \phi, g \sim \mathcal{N}\left(\frac{g}{1+g}\hat{\beta}, \frac{g}{1+g}(X^tX)^{-1}\phi^{-1}\right)$$

b) We note that

$$\mu_i = x_i^t\beta|Y, g, \phi$$

is an affine transformation of a Gaussian, and hence by standard results (see, e.g., the book by Christensen on linear models), we know that μ_i is normally distributed as :

$$\mu_i \sim \mathcal{N}\left(x_i\frac{g}{1+g}\hat{\beta}, \phi^{-1}\frac{g}{1+g}x_i(X^TX)^{-1}x_i^t\right).$$

c)

$$E[Y^*|Y, g, \phi] = E[1_n\beta_0 + X^*\beta + \epsilon^*|Y, g, \phi]$$

$$= E[1_n\beta_0|Y,g,\phi] + E[X^*\beta|Y,g,\phi] + E[\epsilon^*|Y,g,\phi]$$

$$= \bar{Y} + x_i\frac{g}{1+g}\hat{\beta} + 0$$

$$= \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}$$

For the calculation of the variance, we note that the β_0 and β are independent random variables conditioned on Y and ϕ, and ϵ^* is independent of any past observation and hence independent of the model coefficients. Therefore, we can write

$$Var[Y^*|Y,g,\phi] = Var[1_n\beta_0 + X^*\beta + \epsilon^*|Y,g,\phi]$$

$$= Var[1_n\beta_0|Y,g,\phi] + Var[X\beta|Y,g,\phi] + Var[\epsilon^|Y,g,\phi]$$

$$= \frac{I_n}{n\phi} + x_i\frac{g}{g+1}\phi^{-1}(X^TX)^{-1}x_i^t + \frac{I_n}{\phi}$$

$$= \frac{1}{\phi}\left(I_n\left(1+\frac{1}{n}\right) + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t\right)$$

8.

a)

$$p(\beta|Y,g) = \int p(\beta|Y,g,\phi)p(\phi|Y,g)d\phi$$

$$= \int (2\pi)^{-p/2}\left|\frac{g}{(g+1)\phi}(X^TX)^{-1}\right|^{-1/2} exp(-\frac{(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g}{g+1}\sigma^2(X^TX)^{-1})^{-1}(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})}{2})\frac{(\frac{\nu_n\hat{\sigma^2}}{2})^{\nu_n/2}}{\Gamma(\nu_n/2)} \quad \phi^{\nu_n/2-1}$$

$$\propto \int \phi^{p/2}\phi^{\nu_n/2-1}exp(-\frac{(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g+1}{g}\phi(X^TX))(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})}{2})exp(-\frac{\nu_n\hat{\sigma^2}}{2})d\phi$$

$$\propto \int \phi^{(p+\nu_n)/2-1}exp(-\frac{(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g+1}{g}(X^TX))(\beta-\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\phi)d\phi$$

This is the kernel of a gamma and we know that it is equal to the reciprocal of its proportional constant:

$$= \Gamma(\frac{p+\nu_n}{2})\Big[\frac{(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g+1}{g}(X^TX))(\beta-\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g+1}{g}(X^TX))(\beta-\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[1 + \frac{1}{\nu_n}\frac{(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})^T(\frac{g+1}{g}(X^TX))(\beta-\frac{g}{g+1}\hat{\beta}_{OLS})}{\hat{\sigma^2}}\Big]^{-(p+\nu_n)/2}$$

This is proportional to a Student-T distribution with ν_n degrees of freedom.

b)

$$p(x_i^T\beta|Y,g) = \int p(x_i^T\beta|Y,g,\phi)p(\phi|Y,g)d\phi$$

$$= \int (2\pi)^{-p/2}\left|x_i\frac{g}{(g+1)\phi}(X^TX)^{-1}x_i^t\right|^{-1/2} exp(-\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i\frac{g}{g+1}\sigma^2(X^TX)^{-1}x_i^t)^{-1}(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})}{2})\frac{(\frac{\nu_n\hat{\sigma^2}}{2})}{\Gamma(\nu_n/2)}$$

$$\propto \int \phi^{p/2}\phi^{\nu_n/2-1}exp(-\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i^{-1}\frac{g+1}{g}\phi(X^TX)(x_i^t)^{-1})(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})}{2}\phi)exp(-\frac{\nu_n\hat{\sigma^2}}{2})d\phi$$

$$\propto \int \phi^{(p+\nu_n)/2-1}exp(-\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i^{-1}\frac{g+1}{g}(X^TX)(x_i^t)^{-1})(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\phi)d\phi$$

Again, we recognize this as the kernel of a gamma and we know that it is equal to the reciprocal of its proportional constant:

$$= \Gamma(\frac{p+\nu_n}{2})\Big[\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i^{-1}\frac{g+1}{g}(X^TX)(x_i^t)^{-1})(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i^{-1}\frac{g+1}{g}(X^TX)(x_i^t)^{-1})(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[1 + \frac{1}{\nu_n}\frac{(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})^T(x_i^{-1}\frac{g+1}{g}(X^TX)(x_i^t)^{-1})(\beta - x_i\frac{g}{g+1}\hat{\beta}_{OLS})}{\hat{\sigma^2}}\Big]^{-(p+\nu_n)/2}$$

c)

$$p(Y^{test}|Y,g) = \int p(Y^{test}|Y,g,\phi)p(\phi|Y,g)d\phi$$

$$= \int (2\pi)^{-p/2}\left|\frac{1}{\phi}(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)\right|^{-1/2} exp(-\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{\phi}(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n))^{-}$$

$$\propto \int \phi^{p/2}\phi^{\nu_n/2-1}exp(-\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)^{-1}(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})}{2}\phi)exp(-\frac{\nu_n}{2}$$

$$\propto \int \phi^{(p+\nu_n)/2-1}exp(-\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)^{-1}(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\phi)d$$

As in the previous two derivations, we know this is equal to the reciprocal of the proportional constant of the gamma:

$$= \Gamma(\frac{p+\nu_n}{2})\Big[\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)^{-1}(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)^{-1}(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS}) + \nu_n\hat{\sigma^2}}{2}\Big]^{-(p+\nu_n)/2}$$

$$\propto \Big[1+\frac{1}{\nu_n}\frac{(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})^T(\frac{1}{n} + x_i\frac{g}{g+1}(X^TX)^{-1}x_i^t + I_n)^{-1}(Y^{test} - \bar{Y} + x_i\frac{g}{1+g}\hat{\beta}_{OLS})}{\hat{\sigma^2}}\Big]^{-(p+\nu_n)/2}$$

Again, this is proportional to a student-T with ν_n degrees of freedom.

9.

$$p(\beta|\beta_0,\phi) = \int p(\beta|\beta_0,\phi,g)p(\tau|\phi,\beta_0)p(\phi,\beta_0)d\tau$$

Since τ does not depend on ϕ or β_0, we can use the given unconditional distribution for τ. Plugging in the distributions, we get:

$$= \int (2\pi)^{-p/2}\left|\frac{(X^TX)^{-1}}{\tau\phi}\right|^{-1/2} exp(-\frac{\beta^T(\frac{(X^TX)}{\tau\phi})^{-1}\beta}{2})\frac{\frac{n}{2}^{1/2}}{\Gamma(\frac{1}{2})}\tau^{1/2-1}exp(-\frac{n}{2}\tau)\frac{1}{\phi}d\tau$$

Grouping terms, we get:

$$\propto \int \tau^{(p+1)/2-1}exp(-\frac{(n + (\beta^T\phi(X^TX)\beta))}{2}\tau)d\tau$$

It is easy to see that this is the kernel of a Gamma distribution with $a = \frac{n+1}{2}$ and $b = \frac{n + (\beta^T\phi(X^TX)\beta)}{2}$.

For this expression to be equal to 1, the proportional constant would need to be $\frac{(\frac{n + (\beta^T\phi(X^TX)\beta)}{2})^{(n+1)/2}}{\Gamma(\frac{n+1}{2})}$.

It can be shown that the reciprocal of this proportional constant is what the above proportion is equal to. In mathematical terms:

$$= \Gamma(\frac{p+1}{2})\Big[(\frac{n + (\beta^T\phi(X^TX)\beta)}{2})^{-(p+1)/2}\Big] \propto (n + (\beta^T\phi(X^TX)\beta))^{-(p+1)/2}$$

$$\propto \Big[1 + \frac{\beta^T\phi(X^TX)\beta}{n}\Big]^{-(p+1)/2}$$

So, the prior on β is proportional to a student-T with one degree of freedom, which is also a Cauchy distribution.