Math Proficiency and the Home Environment: A Bayesian Regression Analysis

FYE ID NUMBER 1493
Department of Applied Mathematics and Statistics
University of California, Santa Cruz

Abstract

Using data collected by the Educational Testing Service we build Bayesian Regression Models to explain mathematical proficiency scores of 8th grade students. We consider models that only utilize one covariate that reflects students' accessibility to reading material. Expanding upon our original models, we include a second covariate that indicates the percentage of students that read ten pages daily. Although we implement a g-prior for our expanded models, our results using a non-informative prior with direct sampling are similar. It is evident from posterior-predictive model checks and model comparison that incorporating the second covariate provided better overall models. Residual analysis does indicate that our models do not satisfy the assumption of normality and homoscedasticity. We conclude by considering an alternative Robust model for future work.

KEY WORDS: Bayesian Linear Model, G-Prior, Educational Achievement.

1. Introduction

The motivation for our work is to create, analyze, and compare Bayesian Linear Models of the relationship of Mathematical Proficiency of 8th grade students with their accessibility and amount of daily reading of diverse material in the home environment. (Barton and Coley 1992) An investigation conducted by the Educational Testing Service examined results from the 1990 NAEP Mathematical Proficiency Test (scale 0-500) along with several covariates including family resources, parent involvement, absences from school, television watching, accessibility to at least 3 types of reading materials at home, and daily reading of more than 10 pages by students. For the sake of our investigation we will fit Bayesian Linear Models with the last two covariates. The outline of our paper is as follows: In the remainder of this section we will conduct an Exploratory Data Analysis including classical regression approaches. In section II, we fit a few Bayesian models using only one covariate: the home library. In section III, we expand the model by considering a second covariate: Reading. We discuss a few models, including a g-prior approach, before presenting our final Bayesian Model. In section IV, we perform Bayesian Model Checking and Comparison. Model Checking includes Posterior Predictive P-Values and a Leave-One Out Cross-Validation approach. Model comparison includes Deviance Information Criterion and Posterior Predictive Loss Criteria. Section V provides conclusions and suggestions for future work with the data.

1.1 Exploratory Data Analysis

Our data was collected from 37 U.S. states, District of Columbia, Guam, and Virgin Islands. The response variable, y_i, is the mean score of Mathematical Proficiency by 8th grade students in each territory (scale 0-500). The mean and variance of the 40 scores is 260.95 and 174.51 respectively. There appears to be a few outliers including the Virgin Islands, Guam, and Washington D.C. (218,231,231). The first covariate, x_{1i}, is the percentage of eighth grade students with three or more types of reading materials at home (books, encyclopedias, magazines, newspapers). The second covariate,x_{2i}, is the percentage of the students who read more than 10 pages a day. Guam is well below the mean of 80.4% for students with a diverse home library, while D.C. and the Virgin Islands have a noticeably lower percentage of students reading more than 10 pages a day compared to the mean of 36.85%.
We present our data in 3 scatter-plots in Figures 1-3. For each one we fit a classical least-square linear model to give us a preliminary understanding of any relationships present. We label D.C., Guam, Virgin Islands, and California in Figures 1-2 as we discuss these in our model checking in Section IV.

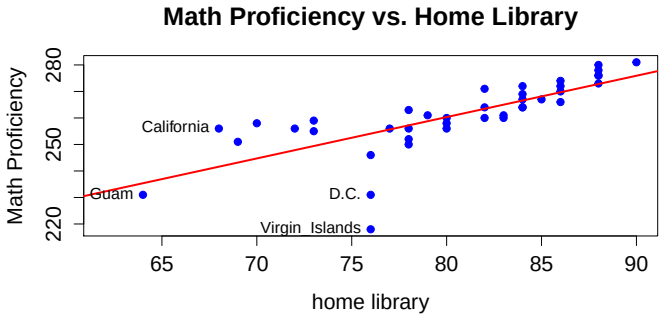

Figure 1: Scatterplot of Math Proficiency and Home Library with least square fit.

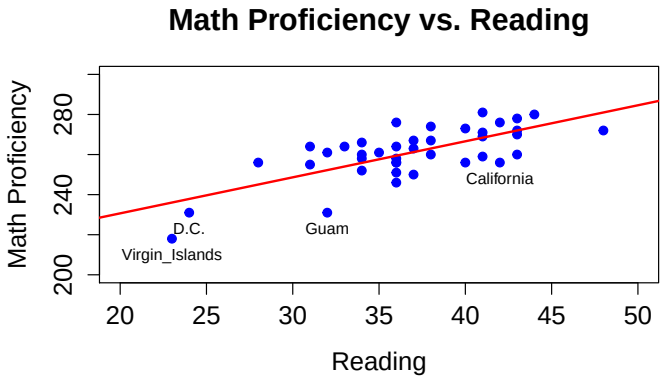

Figure 2: Scatterplot of Math Proficiency and Reading with least square fit.

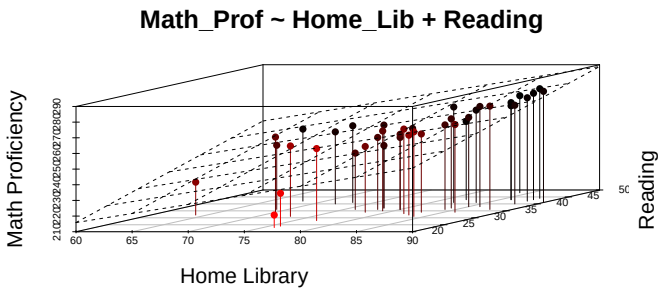

Figure 3: 3D Scatterplot of Math Proficiency, Home Library, and Reading with fitted plane.

In table 1 we present the Least Square Estimates of the following three models along with the R^2. We can see that the first two models using each covariate individually are not good fits. Including both covariates improves the fit, but there is still room for improvement. In the next section we attempt a Bayesian approach to fit a model with just the home library covariate.

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2 I) \quad (1)$$
$$y_i = \beta_0 + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2 I) \quad (2)$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2 I) \quad (3)$$

Table 1: Summary of LSE and R^2 for Classical models 1-3

Fit	*1*	*2*	*3*
β_0	135.6	194.7	120.8
β_1	1.56	—	1.159
β_2	—	1.798	1.274
σ^2	79.49	86.93	41.02
R^2	0.555	0.514	0.777

2. Bayesian Regression with Home Library Covariate

2.1 Model 1

Our first model is a simple linear model to explain the Math Proficiency scores of territory i, y_i. Our covariate is the corresponding home library percentage discussed in the introduction:

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i; \quad \epsilon_i | \sigma^2 \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

In vector notation, the conidtional distribution of y is:

$$y \sim N(X\beta, \sigma^2 I) \quad (4)$$

Where X is the design matrix, $\beta = [\beta_0, \beta_1]$ and I is the identity matrix. Because we have limited background knowledge of our this type of data, we choose to use a non-informative prior distribution that is uniform on (β, log σ):

$$p(\beta, \sigma^2) \propto \sigma^{-2} \quad (5)$$

With this choice of a prior distribution, our joint posterior is:

$$p(\beta, \sigma^2 | y) \propto N(y | X\beta, \sigma^2)\, p(\beta, \sigma^2) \quad (6)$$

By factoring the joint posterior as $p(\beta|\sigma^2)p(\sigma^2)$, we can directly sample from the posterior. As shown in (Gelman et al. 2014):

$$\beta | \sigma^2, y \sim N_k(\hat{\beta}, V_\beta \sigma^2) \quad (7)$$
$$\sigma^2 | y \sim IG\left(\frac{n-k}{2}, \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{2}\right) \quad (8)$$

where $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$ is the familiar maximum likelihood estimate and $V_\beta = (X^T X)^{-1}$, k is the number of explanatory variables, and n is the number of observations. Thus, to directly sample the posterior we follow this outline:

For i = 1 . . . I
Sample $\sigma^{2(i)} | y \sim IG\left(\frac{n-k}{2}, \frac{(y-X\hat{\beta})^T (y-X\hat{\beta})}{2}\right)$
Sample $\beta^{(i)} | \sigma^2, y \sim N_k(\hat{\beta}, V_\beta \sigma^{2(i)})$
end

As Gelman discusses, to ensure propriety of the posterior two conditions must be met: The number of observations must be greater than the number of exploratory variables ($n > k$) and the design matrix must be full rank. The posterior in this model is proper as both conditions are satisfied.

2.2 Model 2

Referring back to our exploratory data analysis, we plot the residuals of Model 1 vs. our covariate home library in Figure 4. This indicates that there may be nonlinearity present and thus we attempt adding polynomial terms

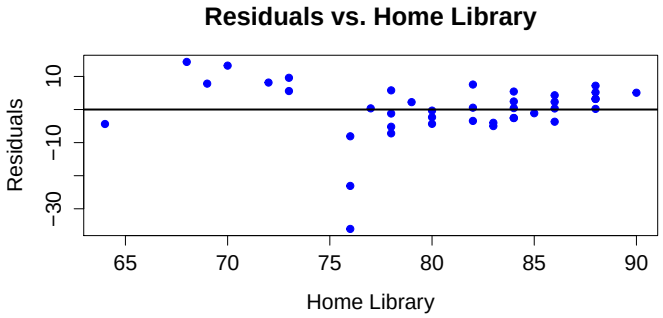

Figure 4: Residuals vs. Home Library from EDA model 1.

After attempting different polynomial models, we select our final model as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \epsilon_i; \quad \epsilon_i|\sigma^2 \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

Using an uninformative prior, this model can be similarly sampled directly as in the previous section. The only difference is we have one more parameter term to our β vector $[\beta_0, \beta_1, \beta_2]$. As in the previous model, our design matrix is full rank, so we are ensured a proper posterior. Model checking and comparison is presented in Section IV.

3. Bayesian Multiple Regression

We expand upon our work in the previous section by including the second covariate: Reading (percentage of students who read more than 10 pages a day). We consider several models and ultimately decide on one that includes log transformation of both covariates and an interaction term. For our prior distribution in our models we consider two options: Zellner's g-prior and a non-informative prior. We elect to use a non-informative prior as the g-prior employed gave very similar results.

3.1 Model 3

For our baseline model we include both covariates without transformation. We will discuss prior specification below.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i; \quad \epsilon_i|\sigma^2 \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

3.2 Model 4

After trying different models, we elect to use a model which logarithmic transform on each of the covariates. In addition to this we include an interaction term while considering the relationship of our covariates (accessibility to reading material and reading daily). Our errors are distributed normally with common variance; $\epsilon_i|\sigma^2 \sim^{i.i.d.} N(0, \sigma^2)$.

$$y_i = \beta_0 + \beta_1 log(x_{1i}) + \beta_2 log(x_{2i}) + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

3.3 Prior Specification

3.3.1 G-Prior

We explore using a g-prior as a means of model selection. As described in (Albert 2009), Zellner's g-prior is way to introduce subjective information into a regression problem. We begin with our prior for σ^2 and the parameter vector, β conditioned on σ^2:

$$\beta|\sigma^2 \sim N_k(\beta^0, g\sigma^2(X'X)^{-1}) \quad (9)$$

$$\sigma^2 \propto \frac{1}{\sigma^2} \quad (10)$$

Where β^0 is an initial guess at the parameter vector and g is a constant that reflects the amount of information in the data relative to the prior. Similar to the direct sampling described in section II, we are able to sample the posterior by considering the form:

$$p(\beta, \sigma^2|y) = p(\beta|y, \sigma^2)\, p(\sigma^2|y) \quad (11)$$

In the case of the g-prior we have:

$$\begin{aligned} \beta|\sigma,^2 y &\sim N_k\left(\frac{g}{g+1}\left(\frac{\beta^0}{g} + \hat{\beta}\right), \frac{g}{g+1}\sigma^2(X'X)^{-1}\right) \\ \sigma^2|y &\sim IG\left(\frac{n}{2}, \frac{SSE}{2} + \frac{(\beta^0 - \hat{\beta})'(X'X)^{-1}(\beta^0 - \hat{\beta})}{2(g+1)}\right) \end{aligned}$$

Because we have limited knowledge of the background of this problem, we set $\beta^0 = 0$ and selected larger values for g, beginning with 40 and increasing to 100 (thus making the prior more and more non-informative). The results we obtained from the g-prior specification were similar to that of a non-informative prior and thus the results we present in section IV assume a non-informative prior.

3.3.2 Non-Informative Prior

Similar to section II, we consider a non-informative prior: $p(\beta, \sigma^2) \propto \sigma^{-2}$. As described in detail in section 2.1 we are able to sample directly from and Inverse Gamma and Multivariate normal for σ^2 and β respectively. Once again, we have more observations than parameters and our design matrices are full rank for all models considered, ensuring propriety.

4. Posterior Results, Model Checking and Comparison

4.1 Posterior Results

For each model we run our direct sampling algorithm for 10,000 iterations in R. Our results are presented in tables 2-5 with posterior means and 95% credible intervals for the parameters of each model. As expected using a non-informative prior, our posterior means for Bayesian Models 1 and 3 are similar to the classical least squares estimators provided in Table 1. The variance of the errors decreases with each model enhancement. However, it appears there is asymmetry in the posterior distribution of this parameter. To check our normality assumption, we will consider residual plots in the next section. It is also worth noting that the credible interval for the quadratic term in Model 2 is close to containing zero, while the interval for β_1 does include zero. The interval for the interaction term in model 4 also almost contains zero.

Table 2: Model 1 Post. Means and 95% Credible Intervals

	Mean	L Bound	U Bound
β_0	135.7	99.24	172.9
β_1	1.557	1.096	2.008
σ^2	84.01	53.03	131.4

Table 3: Model 2 Post. Means and 95% Credible Intervals

	Mean	L Bound	U Bound
β_0	530.0	153.7	910.4
β_1	-8.583	-18.34	1.027
β_2	0.065	0.003	0.127
σ^2	76.68	48.24	121.3

Table 4: Model 3 Post. Means and 95% Credible Intervals

	Mean	L Bound	U Bound
β_0	120.8	93.51	148.3
β_1	1.161	0.796	1.512
β_2	1.271	0.843	1.698
σ^2	43.37	27.29	68.58

Table 5: Model 4 Post. Means and 95% Credible Intervals

	Mean	L Bound	U Bound
β_0	-923.22	-1461.1	-392.41
β_1	182.37	100.42	265.55
β_2	133.22	60.786	207.39
β_3	-0.0317	-0.0587	-0.0053
σ^2	34.175	21.543	54.234

4.2 Model Checking

We conduct model checking in three phases. We begin by performing residual analysis to check our assumptions of normality and homoscedasticity. It is clear that for each of our models that these assumptions are not met.

Our second model check computes posterior predictive p values for all the models for test quantities: mean, median and variance. Comparing the observed mean, median and variance with the distribution of those from replicated data, it appears that all models perform reasonably well under this criteria.

Our last model check is done in a Leave One Out Cross-Validation fashion. It uses the posterior predictive distribution to predict mathematical proficiency in a specific territory after excluding it when training the Bayesian Linear models. The results from this model check suggest that model 4 performs the best. However, all models struggle predicting the math proficiency in the Virgin Islands.

4.2.1 Residual Analysis

We evaluate our fit by checking our error assumptions via residual analysis and Q-Q plots. In each of the four models it appears that the homoscedasticity assumption is violated. Furthermore, examining the Q-Q plot in each, it is clear that the normal assumption is not appropriate considering the left tail of the underlying distribution. As noted in the exploratory data analysis, the data appears to have a few outliers and as a result the distribution of σ^2 has a heavier tail than a normal. Some discussion is provided in the conclusion about an alternative model that could help with this issue.

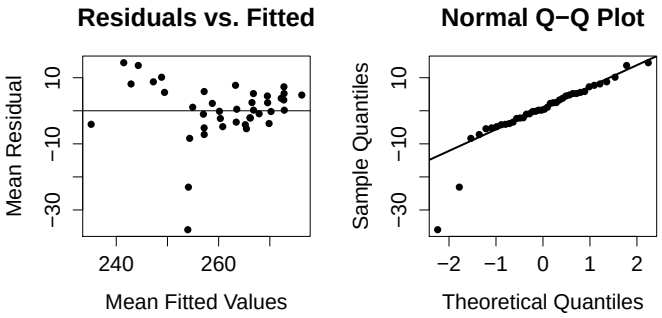

Figure 5: MODEL 1 Residual vs. Fitted and Q-Q plot.

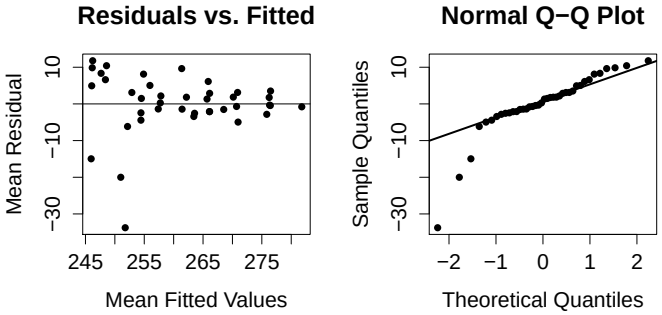

Figure 6: MODEL 2 Residual vs. Fitted and Q-Q plot.

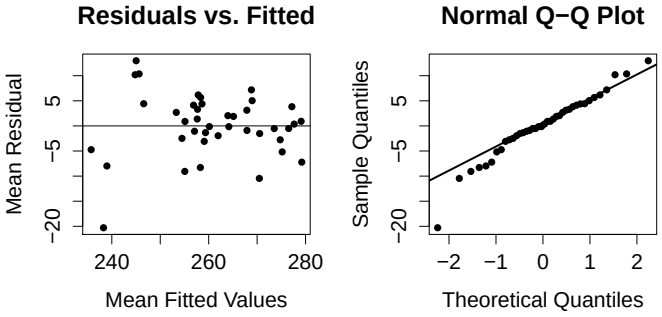

Figure 7: MODEL 3 Residual vs. Fitted and Q-Q plot.

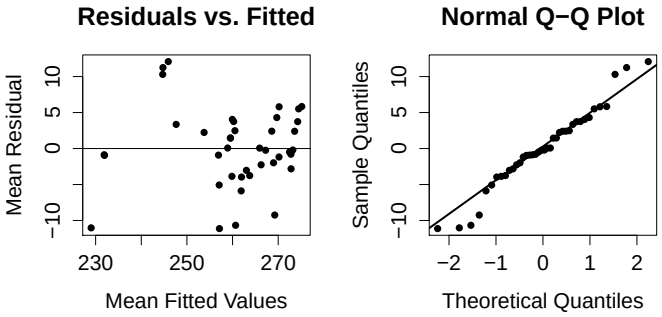

Figure 8: MODEL 4 Residual vs. Fitted and Q-Q plot.

4.2.2 *Posterior Predictive P-Values*

To evaluate the goodness of fit of our models, we created 1,000 replicated data using our posterior predictive distributions and calculated the mean, median, and variance as test quantities. The posterior predictive is given as:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta \tag{12}$$

As discussed in (Gelman et al. 2014), the Bayesian p-value is defined as the probability of replicated data being more extreme than observed data as measured by a test quantity:

$$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y) \tag{13}$$

Where y^{rep} is the replicated data and T is the test statistic. Using our replicated data and test quantities, we present the empirical distributions with their corresponding posterior predictive p-value (PPP value) for Models 1-4 in Figures 9-12. The red lines in each figure represent the observed test quantity. All models perform reasonably well under this model check as we do not have any significant p-values.

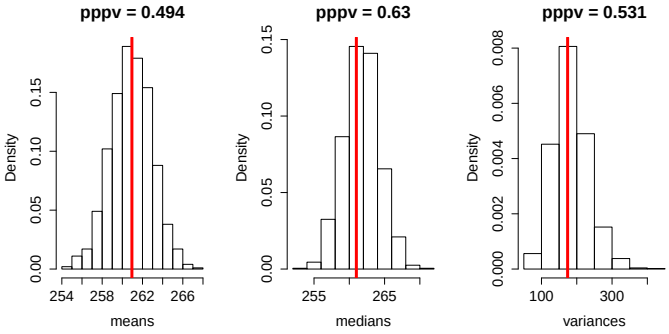

Figure 9: Model 1 PPPV: Distributions of means, medians, and variances from 1000 replicated sets. The red lines represent the observed value.

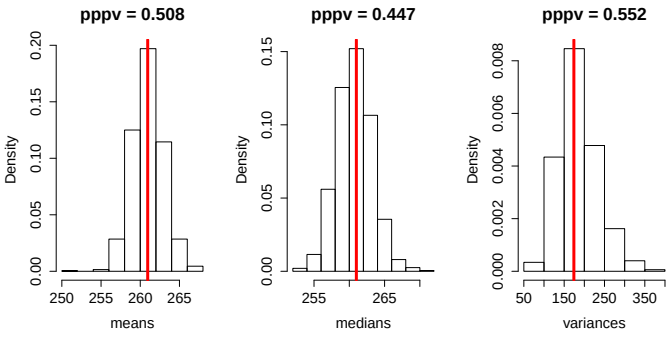

Figure 10: Model 2 PPPV: Distributions of means, medians, and variances from 1000 replicated sets. The red lines represent the observed value.

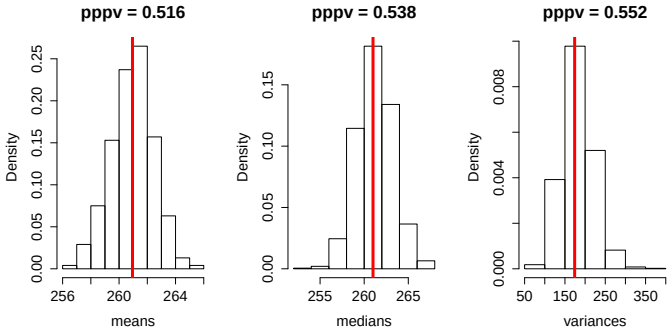

Figure 11: Model 3 PPPV: Distributions of means, medians, and variances from 1000 replicated sets. The red lines represent the observed value.

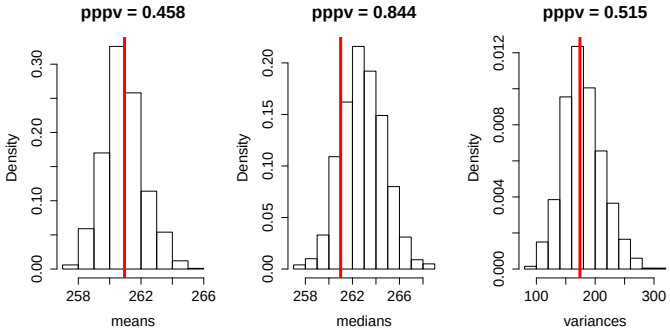

Figure 12: Model 4 PPPV: Distributions of means, medians, and variances from 1000 replicated sets. The red lines represent the observed value.

4.2.3 *Leave One Out Cross-Validation Model Check*

While conducting our exploratory data analysis in Section 1 we indicated that there were a few possible outliers: Guam, Virgin Islands, and Washington D.C. We labeled these data points in Figures 1 and 2 along with California as a frame of reference. For our final model check we iteratively build our Bayesian Models while excluding the data from one of these territories at a time. Once we have built our model, we utilize the posterior predictive distribution to predict math proficiency in the territory we excluding from our model in that iteration. This is a Leave-One Out Cross-Validation approach. Taking 1,000 samples, we create distributions of each posterior prediction and plot them along with the observed value for math proficiency in the given area. Figures 13-16 display the results from this model check. As can be seen in Figure 13, Model 1 has difficulty with D.C. and the Virgin Islands. While none of the models have success predicting the Virgin Islands, model 4 seems to be the closest at predicting it while performing well on the other 3 territories in this model check.

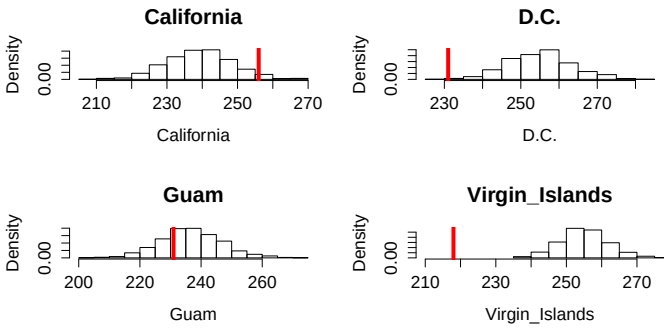

Figure 13: MODEL 1: LOO Cross-Validation Model Check.

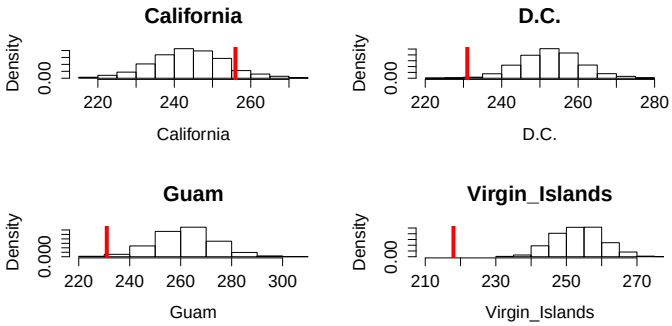

Figure 14: MODEL 2: LOO Cross-Validation Model Check.

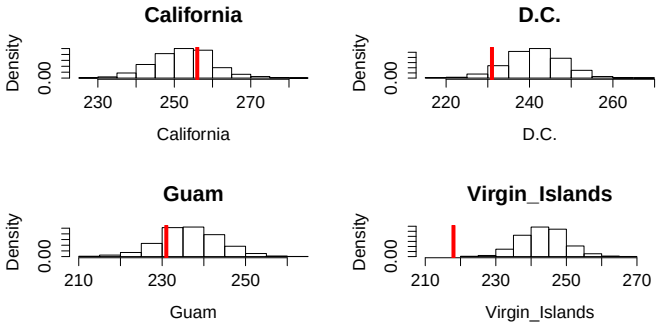

Figure 15: MODELl 3: LOO Cross-Validation Model Check.

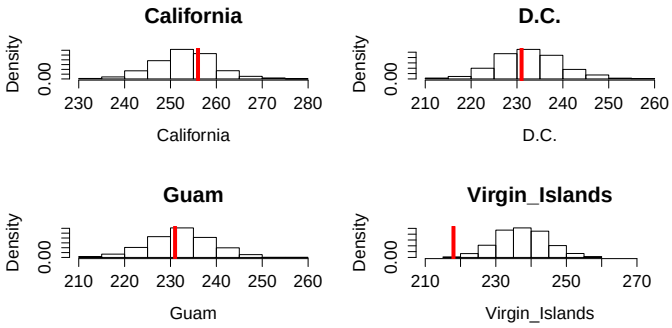

Figure 16: MODEL 4: LOO Cross-Validation Model Check.

4.3 Model Comparison

To compare our models we will use two different methods: Posterior Predictive Loss Criteria (PPLC) as proposed by (Gelfand and Ghosh 1998) and Deviance Information Criterion (DIC) as proposed by (Spiegelhalter et al. 2002).

PPLC utilizes the posterior predictive distribution to compare with the observed data to evaluate the fit of the model. Let y denote the observed data of dimension n and y^{rep} denote posterior predictive replications. The PPLC is defined as the sum of a goodness of fit term and penalty term:

$$PPLC = \sum_{i=1}^{n}(E(y_i^{rep}|x) - y_i)^2 + \sum_{i=1}^{n} var(y_i^{rep}|x) \quad (14)$$

It is evident from this definition that we wish to utilize a model that minimizes this loss criteria; we seek a model where the expected value of our replicates is as close to our observed values as possible (goodness of fit term) while the variability in our replicates (penalty term) is as minimal as possible.

The second method we used to compare models is the Deviance Information Criterion. We first define the deviance of a given sample as $D(\theta) = -2\log p(y|\theta)$. Calculating the mean of the deviances of all posterior samples, $\bar{D}$, and the effective number of parameters, $p_D = \bar{D} - D(\bar{\theta})$, where $\bar{\theta}$ is the posterior mean, we can calculate DIC as:

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta}) \quad (15)$$

Similar to PPLC, we have a goodness of fit and penalty term. The effective number of parameters is our penalty term. Likewise, smaller values of DIC indicate a better overall model.

Table 6: Model comparison.

	p_D	DIC	PPLC
Model 1	3.100	292.9	6572
Model 2	4.172	290.4	5987
Model 3	4.172	267.6	3392
Model 4	5.125	259.1	2701

As we can see from the calculated DIC and PPLC the models that include both covariates perform better. Model 2 which only included the first covariate with an additional quadratic term only saw a mild improvement over the original home library covariate model.

5. Concluding Remarks

We performed Bayesian Regression to explore the relationship between students' mathematical proficiency and two covariates: home library and reading. Our original analysis focused on Bayesian models that only considered the first covariate. We expanded upon these models by introducing the second covariate, reading, to our models. In both cases we ultimately used uninformative priors, although g-priors were considered for the models containing two covariates. Our posterior results for the simpler Bayesian models coincided with the classical least squares results presented in the exploratory data analysis.

We used two criterion to compare our models. Model 4 is the best of those created considering the calculated DIC and PPLC. Although this is the case, it is important to note that

all models struggled when checking the model assumptions of normality and homoscedasticity via residual plots. For future work, one could consider trying a Robust Bayesian Regression approach to handle the few outliers. As (Albert 2009) discusses, the error distribution could instead be assumed as a student-t with a few degrees of freedom. By introducing latent variables, one can rewrite the model as a scale-mixture of Normals. With this approach, one can write a Gibbs sampler and utilize the latent variables to identify outliers.

If data is available that includes multiple observations per state, rather than just the single mean value (e.g. scores measured over several years or within counties of the states), it is possible to expand our investigation with Bayesian Hierarchical Models as well. It is also possible in future work to add other covariates that are discussed in (Barton and Coley).

REFERENCES

Albert, J. (2009), Bayesian Computation with R, Use R!, Springer.

Barton, Paul E., and Coley, Richard J. ETS Policy Infprmation Center. America's Smallest School: The Family. Princeton, N.J.: Educational Testing Service, 1992.

Gelfand, Alan E., and Sujit K. Ghosh. "Model choice: a minimum posterior predictive loss approach." Biometrika 85.1 (1998): 1-11.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2014), *Bayesian Data Analysis*, Chapman and Hall, Boca Raton.

Liang, F., Paulo, R., Molina, G., Clyde, C.A. and Berger, J.O. (2008), Mixtures of g Priors for Bayesian Variable Selection", Journal of the American Statistical Association, Vol. 103, No. 481, 410-423.

Spiegelhalter, David J.; Best, Nicola G.; Carlin, Bradley P.; van der Linde, Angelika (2002). "Bayesian measures of model complexity and fit (with discussion)". Journal of the Royal Statistical Society, Series B 64 (4): 583639.