ANALYSIS OF STUDENT PERFORMANCES IN MATH

Jeremy Meyer
BYU Department of Statistics
STAT 651 Project

1 Introduction

There are no shortage of explanations in the media on what makes students do well in school. While individual students that have been successful can give anecdotes, these may or may not apply at scale. For this reason, it is necessary to look at the entire student body to see what distinguishes students who excel. In this analysis, we will look at data collected from reports and questionnaires in a Portuguese secondary school [1]. The data can be used by the school to find areas to improve. The data contain several socioeconomic factors from 349 students with student performance in mathematics courses. Student performance was measured on a 20-point scale as a final grade at the end of the year. The goal of the analysis is to determine which socioeconomic factors contribute to high math scores. Specifically, we will determine if factors such as parent education level, extracurricular activities, amount of study time, and many others contribute to a higher final grade at the end of the year. We will perform Bayesian beta regression using Markov chain Monte Carlo (MCMC) to determine significant predictors.

2 The Data

The data contains 10 discrete socioeconomic factors of interest and are listed as follows:

1. Medu: Mother's Education level (0 None - 4 College level)
2. Fedu: Father's Education level (0 None - 4 College level)
3. studytime: Study time [1 ($<$ 2 hours) - 4 ($>$ 10 hours)]
4. activities: Extracurriculars? (Yes / No)
5. romantic: Romantic relationship (Yes / No)
6. famrel: Family relations (1 Very Bad - 5 Excellent)
7. freetime: Free Time after school (1 Very Low - 5 Very High)
8. goout: Time out with friends (1 Very Low - 5 Very High)
9. Walc: Weekend Alcohol consumption (1 Very Low - 5 Very High)
10. health: Health (1 Very Poor - 5 Excellent)

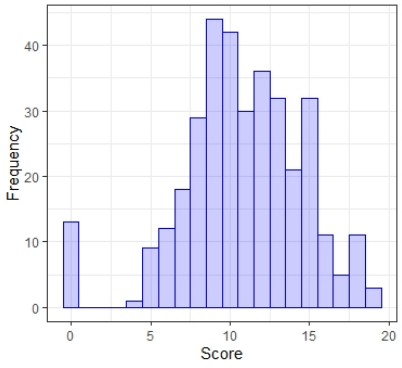

Figure 1: Histogram of scores for the 349 students. Note the 13 outliers at 0.

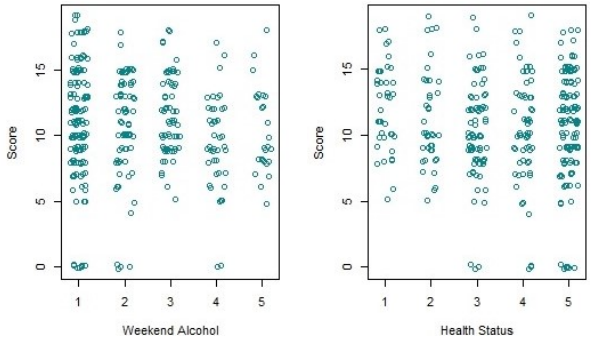

Figure 2: Plots of weekend alcohol consumption and health status against student score. Values have been jittered to see the density of points. No weird non-linear behavior.

We will assume the ordinal variables are continuous for the purpose of the analysis. That is, we will assume the ordinal difference between a 2 and 3 is the same as 4 to 5. The plots in Figure 2 show fairly linear relationships between the response and covariates. Figure 1 shows the distribution of student scores. Most students scored in the middle of the scale, but there some who scored zero. As can also be seen in Figure 2, there are some outliers (13) of students who scored zeros. Later we will explain that these greatly affect the results of the analysis. Although we do not know why these students scored a zero, we will leave these outliers out because we believe they do not fit well with the rest of the data.

3 Methods

To determine which factors are significant we will perform Bayesian regression. We will scale the response to be a proportion (out of 20) so we can perform beta regression of student scores on the factors. We reparameterize Beta(α, β). Let $\mu = \frac{\alpha}{(\alpha+\beta)}$ be the expected value of the Beta distribution. Since the variance has an upper bound of $\mu(1-\mu) < .25$, we will use a dispersion parameter $\phi = \alpha + \beta$ instead. Thus, $\alpha = \phi\mu$ and $\beta = \phi(1-\mu)$. This will make MCMC easier as ϕ is unconstrained for a given μ. ϕ is similar to a precision, in that it has an inverse relationship with the variance. It's an overall measure of how tight the data is packed around μ. Thus, for the i^{th} student's score y_i,

$$E(y_i) = \mu_i \qquad V(y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}. \tag{1}$$

3.1 Proposed Model

We model the mean score as a function of the covariates. To keep the mean between 0 and 1, we use a logit link function. Thus, our final model is

$$\begin{gathered} y_i|\phi, \mu_i \sim \text{Beta}(\phi\mu_i,\ \phi(1-\mu_i)),\ \text{where} \\ \mu_i = \text{logit}^{-1}(\beta_0 + \beta_1(x_{1i}) + \beta_2(x_{2i}) + ... + \beta_9(x_{9i}) + \beta_{10}(x_{10i})) \\ \phi \sim \text{Gamma}(.1, .1), \quad \text{each } \beta_j \sim t_4. \end{gathered} \tag{2}$$

Where x_{ji} is the j^{th} covariate for student i and β_j represents the effect of the j^{th} covariate on the score. The covariates are in the same order as they are listed in section 2. We used uninformative prior distributions for ϕ and β_j to let the data guide the results of the analysis. We needed a distribution with positive support for ϕ and a symmetric zero-mean distribution for β_j. We chose a symmetric zero-mean distribution so the prior is unbiased towards any significant effect.

We will discuss the prior choice in more detail in section 4.1. By nature, the magnitude of the β coefficients in beta regression tend to be small (less than 1). Thus, an uninformative prior does not need to have a large variance. After sampling from the posterior and exploring various priors in Table 5, we found that a Gamma(0.1,0.1) and t-distribution with 4 degrees of freedom (t_4) had the best fit.

3.2 Computational Approach

Let $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_{10}, \phi)$. We use the Metropolis algorithm to sample from the posterior of $\boldsymbol{\theta}$. We tried sampling each of the 12 parameters one at a time using a Gibbs sampler. We used a univariate Metropolis algorithm with a Gaussian proposal to sample from the full conditionals. Due to the correlation between the βs, we ran into problems mixing. For example, the worst effective sample size (β_0) after 20000 draws was only 190, less than 1%. For this reason, we decided to sample $\boldsymbol{\theta}|\boldsymbol{y}$ by sampling the parameters in chunks. That is, we first sample the correlated groups of βs, then sample the uncorrelated βs univariately.

By inspection of the correlations, we found several βs to be correlated with β_0 and β_1,β_2 to have high correlations. We sample $\boldsymbol{\theta_1} = (\beta_0, \beta_3, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10})$ together and $\boldsymbol{\theta_2} = (\beta_1, \beta_2)$ together. Thus, $\theta_3 = \beta_4$, $\theta_4 = \beta_5$, and $\theta_5 = \phi$ are each sampled individually. We use Gaussian proposals for both the multivariate and univariate Metropolis updates. The sampling algorithm is as follows:

1. Choose a starting value, $\boldsymbol{d_0}$, for $\boldsymbol{\theta}$. We used 0 for each β_j and 10 for ϕ.
2. For iterations $i = 1, 2, , , , 21000$:
   - Let the current state be denoted $\boldsymbol{d}_{cur}$. Assign $\boldsymbol{d}_{i-1}$ to $\boldsymbol{d}_{cur}$.
   - Using a 7-dimensional multivariate normal (MVN) proposal distribution, generate a proposed value for $\boldsymbol{\theta_1}$.
     - Center on the proposal distribution on the $\boldsymbol{d}_{cur}$ for the respective elements of $\boldsymbol{\theta_1}$.

  - Calculate the inverse hessian matrix generated from the normal approximation at the mode of the conditional posterior. Denote this matrix $\hat{\boldsymbol{\Sigma}}$. The proposal covariance matrix is a scaled version of $\hat{\boldsymbol{\Sigma}}$.
  - Note: The conditional posterior will change as the parameters not in $\boldsymbol{\theta_1}$ are updated. Thus, $\hat{\boldsymbol{\Sigma}}$ will change across iterations depending on the state of parameters not in $\boldsymbol{\theta_1}$. Due to computational costs, we simply generated one $\hat{\boldsymbol{\Sigma}}$ at the beginning using the starting values. This could also be updated infrequently (e.g. for every 1000^{th} draw)
- Calculate the metropolis ratio, denoted α. With probability $\min(1, \alpha)$, update the elements of $\boldsymbol{d_{curr}}$ corresponding to $\boldsymbol{\theta_1}$ with the proposed value.
- Using the same sampling scheme as $\boldsymbol{\theta_1}$, use a bivariate normal proposal for $\boldsymbol{\theta_2}$. Use any updates on $\boldsymbol{d_{curr}}$ from $\boldsymbol{\theta_1}$.
- Update θ_3, θ_4, θ_5 sequentially from their full conditionals using a univariate normal proposal. Again, use any updates to the current state from $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$.
- After cycling through all $\boldsymbol{\theta}_i$, assign $\boldsymbol{d_{curr}}$ to $\boldsymbol{d_i}$. $\boldsymbol{d_i}$ denotes the i^{th} draw from the posterior.

After discarding the first 1000 draws for burn-in, we obtain 20000 posterior draws. We now diagnose MCMC mixing and convergence in the following section.

3.3 Computational Results

As can be seen from the diagnostics in Table 1, the worst effective sample size is about 3.5 times larger than the worst from univariate updates. The ACF plots (Figure 3) are not ideal, but they are not bad either. If larger effective sample sizes are desired, one can run the sampler described in the previous section for more iterations. The sampler takes about 5.86 minutes to generate 20000 draws.

Table 1: Table of Acceptance rates (Acc rate) and Effective sample sizes (Eff size) and $\hat{R}$ from the multiple chains in Figure 4. Bolded βs were sampled univariately.

	Acc Rate	Eff Size	$\hat{R}$
β_0	.258	697	1.0029
β_1	.369	748	1.0056
β_2	.369	1192	1.0015
$\boldsymbol{\beta_3}$	.258	899	1.0023
$\boldsymbol{\beta_4}$	.277	910	1.0011
β_5	.258	1861	1.0014
β_6	.258	971	1.0010
β_7	.258	817	1.0012
β_8	.258	856	1.0017
β_9	.258	861	1.0013
β_{10}	.258	850	1.0007
$\boldsymbol{\phi}$	.439	3848	1.0002

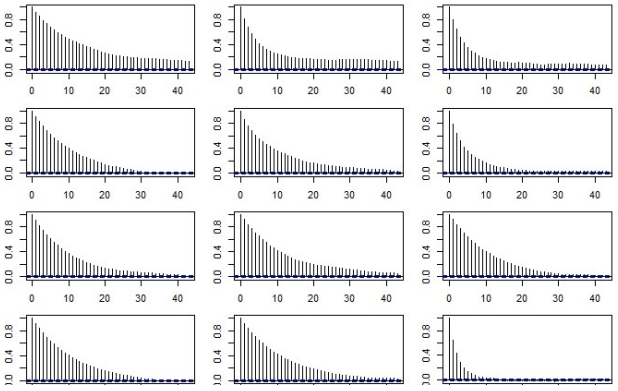

Figure 3: ACF plots. Order is left to right, with β_0, β_1 and β_2 on the top row and ϕ on the bottom right

We now asses convergence. Recall the starting value we used for $\boldsymbol{\theta}$ was $\boldsymbol{d_0} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 10)$, which we will denote on Figure 4 as a black line. We check for convergence by looking at $\hat{R}$ values from Table 1 and trace plots from the following $\boldsymbol{d_0}$'s:

- Initial values low $\boldsymbol{d_0} = (-.25, -.25, ..., 9)$ (red line)
- initial values high $\boldsymbol{d_0} = (.25, .25, .25, ..., 11)$ (cyan line)
- mixed extremes $\boldsymbol{d_0} = (.5, .5, ..., -.5, -.5, ...7)$ (blue line)
- mixed extremes $\boldsymbol{d_0} = (-.5, -.5, ..., +.5, +.5, ...12)$ (green line)

The trace plots in Figure 4 show strong signs of convergence because they started in different locations and ended at the same stationary distribution. The $\hat{R}$ values are also very close to 1, which indicates the chains have converged to the same distribution. We now proceed to evaluating the significant predictors.

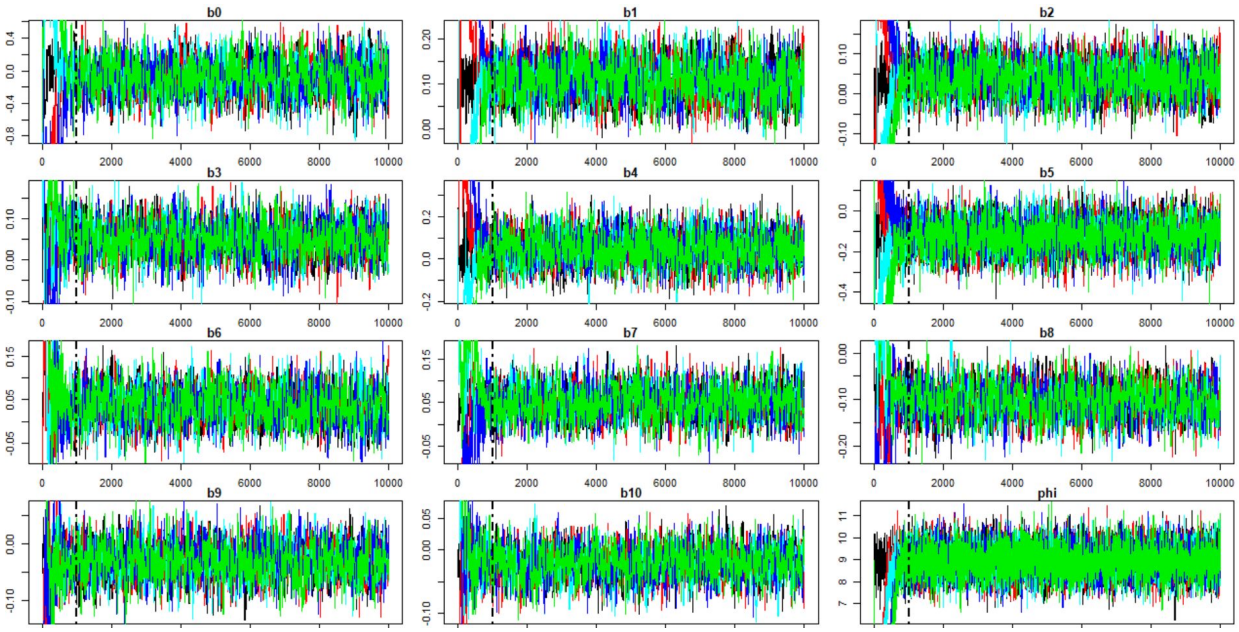

Figure 4: Trace plots at 5 different starting locations: original $\boldsymbol{d_0}$ (black), low (red), high (cyan), extremes (blue and green). The vertical, dashed, black line at 1000 denotes the burn-in period. All trace plots show strong signs of convergence.

4 Results

The posterior estimates of the β coefficients are shown in Table 2. Only two covariates were significant. The level of education for the mother was found to have a positive influence on student math scores, and time out with friends seemed to have a negative impact. All other factors had no significant effect on student performance. Recall that we had dropped the students with a zero score from the analysis. We now discuss impact on the conclusions if we had still included the outliers in the analysis. Leaving the zeros in and after using MCMC to obtain posterior samples, we obtained the posterior estimates for the βs as shown in Table 3

Table 2: 95% Credible intervals and means for the posterior parameters. Only 2 were significant. Students with a zero score were not included.

	2.5%	mean	97.5%
Intercept	-0.52	-0.08	0.35
Medu(+)	0.03	0.11	0.19
Fedu	-0.04	0.04	0.12
studytime	-0.03	0.05	0.13
activities	-0.09	0.05	0.19
romantic	-0.27	-0.12	0.02
famrel	-0.04	0.04	0.11
freetime	-0.02	0.05	0.12
goout(-)	-0.18	-0.10	-0.03
Walc	-0.09	-0.03	0.03
health	-0.07	-0.02	0.03
phi	7.64	8.93	10.24

After including the zeros, both the Mother's education level and time out with friends were still significant, but 4 more covariates appeared significant. The coefficients for studytime, romantic relationships, and health shifted significantly away from zero, and the coefficient for alcohol consumption on weekends switched signs and became significant. If we had used the full data, not only would we report that better health actually has a negative impact on student grades but that increased consumption on weekends has a positive influence on student performance!

From Figure 2, one can get an idea of why this is happening. The students that scored zero also scored low on alcohol consumption and high on health status. Since zero is on the edge of the support for the beta distribution, they needed to have ϵ added to them (we chose .001). The logit function changes much more drastically with inputs close to 0 and 1. For example, $\text{logit}(.001) = -6.91$ while $\text{logit}(4/20) = -1.39$ and $\text{logit}(0.5) = 0$. Thus, the zeros have a strong influence on the beta coefficients.

Table 4 compares covariate values between students that had zero scores and those that did not. From the table, it can be seen that students with a zero score consumed less alcohol on the weekends and had very good health. Additionally, the proportion of students with romantic relationships was twice as high. Perhaps these students dropped out of the course or had to file incomplete. It's also interesting how ϕ is much lower when including the zeros, which means the overall variance around the mean is much higher than if the zeros are left out. In summary, due to the counterintuitive and drastically different results from including the zero students, we believe they are telling a different story than the rest of the data, and should be examined separately.

Table 3: Results if we had left the zero students in. These results suggest drastically different conclusions! The zero scores have a high influence on the results.

	2.5%	mean	97.5%
Intercept	-0.45	0.11	0.67
Medu(+)	0.04	0.15	0.27
Fedu	-0.11	-0.00	0.11
studytime(+)	0.04	0.15	0.26
activities	-0.08	0.11	0.30
romantic(-)	-0.57	-0.37	-0.17
famrel	-0.14	-0.03	0.07
freetime	-0.05	0.06	0.16
goout(-)	-0.34	-0.24	-0.14
Walc(+)	0.02	0.11	0.19
health(-)	-0.16	-0.09	-0.02
phi	3.33	3.81	4.35

Table 4: Group means comparing students with zeros and students without and their raw differences.

	Zeros	Not Zeros	diff
Medu	2.62	2.81	-0.19
Fedu	2.54	2.55	-0.01
studytime	1.77	2.07	-0.30
activities	0.46	0.53	x0.9
romantic	0.61	0.31	x2.0
famrel	4.38	3.94	0.44
freetime	3.46	3.21	0.25
goout	3.77	3.09	0.68
Walc	1.77	2.28	-0.51
health	4.31	3.55	0.76

4.1 Prior Sensitivity analysis

We also checked if the results changed based on the usage of other uninformative prior distributions. After generating posterior samples from the various priors, we computed DIC values, which are shown in Table 5. Our final prior choice was the model with the lowest DIC. We then checked if any of the conclusions changed from Table 2. All of the prior distributions specified in Table 5 resulted in the same conclusion as stated in Table 2. Thus, the conclusions are not oversensitive to our choice of uninformative priors.

Table 5: DIC values for various prior distributions. The top row was the baseline we used to compare to other values. For example, the DIC for a G(1,1) prior on ϕ also used a t_4 on all β_j

ϕ Prior	DIC	β_j prior	DIC
G(.1, .1)	-288.09	t_4	-288.09
G(.1, 1)	-286.62	t_2	-287.86
G(1,1)	-286.48	N(0,1)	-287.15
Unif(0, 100)	-287.59	N(0, 100)	-286.74

4.2 Comparison to frequentist methods

The `betareg` package in R performs frequentist beta regression. It gives options for the link function and uses the same parameterization as defined in Equation 2, but without the priors. Thus, we can compare our Bayesian analysis to a frequentist method. The estimates are found as follows: given observed covariates $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_{10}}$ and response $\boldsymbol{y}$, where

$$\begin{gathered} y_i \sim \text{Beta}(\phi\mu_i,\ \phi(1-\mu_i)),\ \text{where} \\ \mu_i = \text{logit}^{-1}(\beta_0+\beta_1(x_{1i}) + \beta_2(x_{2i}) + ... + \beta_9(x_{9i}) + \beta_{10}(x_{10i})), \end{gathered} \tag{3}$$

estimate ϕ, $\beta_0, \beta_1, ..., \beta_{10}$ using maximum likelihood. The P-value, coefficients, and confidence interval width is displayed in Table 6. The same 2 covariates are statistically significant, and the coefficients are almost identical between the two methods. Even the interval widths are almost identical. The intercept and ϕ, which do not directly relate to the conclusions, are the only coefficients slightly different. The fact that our results were so similar suggests that our priors were fairly uninformative and the data carries almost all the weight in the Bayesian analysis.

Table 6: Frequentist estimates. the .F refers to frequentist and .B refers to Bayesian. Thus, Coef.F are the frequentist coefficients and Coef.B are the Bayesian coefficients displayed in Table 2.

Covariate	P-value	Coef.F	Coef.B	Width.F	Width.B
(Intercept)	0.666	-0.11	-0.08	1.00	0.87
Medu	**0.007***	0.11	0.11	0.16	0.15
Fedu	0.315	0.04	0.04	0.16	0.16
studytime	0.212	0.05	0.05	0.17	0.16
activities	0.478	0.05	0.05	0.28	0.28
romantic	0.082	-0.13	-0.12	0.29	0.29
famrel	0.348	0.04	0.04	0.16	0.15
freetime	0.165	0.05	0.05	0.14	0.14
goout	**0.004***	-0.10	-0.10	0.14	0.15
Walc	0.360	-0.03	-0.03	0.12	0.12
health	0.419	-0.02	-0.02	0.10	0.10
ϕ	—	9.19	8.93	2.64	2.59

4.3 Simulation Study

We now perform a simulation study to check if our sampling algorithm and model can recover known coefficient values. That is, after simulating synthetic response data ($\tilde{y_i}$) using known β_j values, we will see if the model can recover the similar βs using the same computational procedure done for the analysis. Using the model specified in Equation 2, we simulated the synthetic response data as follows:

1. Generate "true" β_j coefficients by sampling from the prior (t_4). To get the coefficients on a similar scale as the estimates in Table 2, we scaled these values by 1/5.
2. Generate a "true" ϕ parameter by simulating from the prior (Gamma(.1, .1)).
3. In regression, the predictors are conditioned on. Thus, we will use the same predictor data as observed in the Portuguese school data for the simulation.
4. Generate $\tilde{\mu_i}$'s by taking the inverse logit of a linear combination of the predictor data and β_j coefficients as in Equation 2. Do this for all $i = 1, ..., 349$.
5. Generate the synthetic responses y_i's by taking a draw from a Beta(α_i, β_i) with parameters $\alpha_i = \phi(\tilde{\mu_i})$ and $\beta_i = \phi(1 - \tilde{\mu_i})$.
6. Since the original dataset was discrete on a scale from zero to 20, multiply the $\tilde{y_i}$'s by 20, then round to the nearest integer. Scale these values back to a proportion out of 20. Adjust generated 0 or 20 values by adding/subtracting $\epsilon = .001$ noise.

After obtaining the synthetic data, we now perform the same procedure that was used earlier to analyze the student data. We wish to see how the true parameters compare to the estimates. The results are displayed in Table 7. All but one of the 12 parameters was covered in the 95% credible interval. Since we only used a 95% credible interval, it is not unreasonable for this to happen. This gives us confidence that under the correct assumptions, we can recover true values for the parameters and thus make correct inference with our posterior estimates.

Table 7: The true parameter values, posterior means, and 95% credible intervals for the estimated coefficients. All but β_5 were contained in the 95% credible interval.

	Truth	Estimate	2.5%	97.5%
β_0	-0.05	-0.09	-0.66	0.52
β_1	0.25	0.26	0.15	0.37
β_2	-0.57	-0.56	-0.67	-0.45
β_3	0.14	0.10	0.00	0.20
β_4	-0.04	-0.05	-0.23	0.14
β_5	0.12	-0.09	-0.27	0.10
β_6	-0.08	-0.06	-0.15	0.04
β_7	-0.00	0.00	-0.09	0.09
β_8	0.15	0.12	0.03	0.21
β_9	0.16	0.21	0.12	0.28
β_{10}	0.01	-0.00	-0.07	0.06
ϕ	4.82	5.19	4.51	5.93

5 Conclusion

We performed Bayesian beta regression of student scores on 10 socioeconomic factors. We dropped students who scored zeros because they strongly influenced the results and were dissimilar with the rest of the data. We sampled from the posterior by jointly sampling the parameters that were correlated and univariately sampling the uncorrelated parameters. The algorithm produced better effective sample sizes and mixed better than a Gibbs sampler. Convergence was assessed with trace plots and scattered initial values. Only two of the 10 covariates were significantly different from zero. We concluded that the education of the mother had a positive impact on student performance in math courses and students who spent too much time out with friends tended to have lower math scores. The exact same conclusions were reached under a frequentist analysis and for several different uninformative priors. We also performed a simulation study to check if under the correct assumptions, we could recover true values for the parameters. We were able to successfully recover 11 out of the 12 parameters.

Some future work would include finding out why some students were marked a zero in their math course. Instead of throwing these observations out because they strongly influence the results, knowing why some students received a zero could help determine if they should be included in the analysis or dropped because they are irrelevant. If the data was missing and was labeled a zero, for example, then including it as a zero in the analysis would not make sense. Another area of interest is to see if these results apply elsewhere. Since the data was only for one school, it may or may not apply at scale. Further studies can be done in different schools, subjects, and countries to see if similar conclusions hold.

References

[1] Paulo Cortez and Alice Silva. Using data mining to predict secondary school student performance. *EUROSIS*, 01 2008.