Jon Janelle

MAT 500

Project 2

**Factors Predictive of the 2008-09 U.S. Average Freshman Graduation Rate**

***Introduction***

In the United States, a high school diploma is critically important for finding employment and broadening educational and career options. In 2012, the U.S. Bureau of Labor Statistics (BLS) estimated that the unemployment rate for those will less than a high school diploma was 12.4%, which is particularly alarming when compared to the estimated 8.3% unemployment rate for those with a high school degree (“Earnings and unemployment”, 2013). Further, the BLS estimated that the earnings of someone without a high school diploma in 2012 were an average of $181 per week, or $9512 per year, less than the earnings of someone with a high school degree and arguably below a living wage. These findings clearly paint an unfavorable economic picture for those who fail to complete high school.

Graduation rates are also a major component of current school accountability systems. The 2001 No Child Left Behind Act (NCLB) includes provisions meant to hold schools accountable for the percentage of students who graduate on-time. Schools that fail to meet these requirements face restructuring, often re-staffing, and other corrective measures. The averaged freshman graduation rate is one measure used to compare and evaluate state-reported graduation rates under NCLB.

The averaged freshman graduation rate (AFGR) estimates the proportion of high school students who earn a diploma within four years (Chapman et al., 2013). During the 2008-2009 academic year the AFGR for public school students was estimated at 75.5% (Chapman et al., 2013). In this calculation, and under NCLB, students who leave school before receiving a diploma, graduate in five or more years, or earn a GED are considered “dropouts.” The AFGR has steadily increased over the past several decades, but there are clearly still many at risk of failing to complete high school.

The purpose of this study is to investigate which factors influenced the AFGR during the 2009-2010 academic year. The factors considered are all state-level and include pupil teacher ratio, total revenue per pupil, population density, proportion of students eligible for free or reduced lunch, and the proportion of students with limited English proficiency. The model developed in this analysis is meant only to highlight national trends and cannot be used to make recommendations for particular locales.

***Data and Variable Selection***

All data were obtained from the National Center for Educational Statistics (NCES), which is a division within the U.S. Department of Education. Each year the NCES administers a set of five national surveys to approximately 18,000 public school districts to maintain its Common Core of Data (CCD) repository. The 2009-2010 academic year was the most recent for which complete state-level graduation data was available through the CCD. Washington D.C. and all states except for New York, for which complete data was unavailable, are included in the data.

Many past studies of U.S. graduation rates, for example Green and Winters (2001) and Swanson (2004), have reported significant racial differences in graduation rates. Both studies reported that schools with a majority of white students tend to have higher graduation rates than schools with a majority of non-white students, which suggests that serious opportunity gaps exist within public education. Specific racial and ethnic differences will not be considered in this investigation. However, the percentage of students classified as having limited English proficiency, which may be a considered a vague measure of the ethnic diversity within schools, will be considered. The percentage of students receiving special education services was also considered by Swanson (2004b), but will not be included in the present discussion.

Swanson (2009) reported a significant and substantial difference between the graduation rates in urban schools and those in suburban and rural schools. Jordan and Kostandini (2012) contested this finding and found that there was little difference when factors such as sex, income, race, assets, and peer and family characteristics were considered. State population density is used as an approximation of how many k-12 students, both public and private, are enrolled in urban schools. States with higher population densities tend to have more urban areas, and therefore more students enrolled in urban schools. The 2010 population densities used are measured in people per square mile, and the data were obtained from the U.S. Census Bureau (2012).

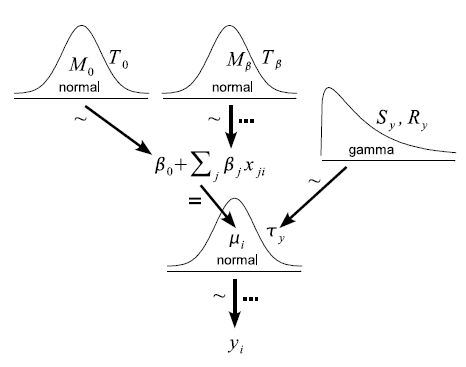
The percentage of students eligible for free or reduced price lunch is a common proxy for measuring the proportion of students in poverty. According to Balfanz and Legters (2006), “poverty is the fundamental driver of low graduation rates” (p. 1), and therefore should be considered in any study related to factors influencing student academic performance. Poverty may interact with other factors, such as geographic region or ethnicity, but such interactions will not be considered.

Total revenue per pupil is defined by the CCD as the total revenues allocated to public education received from all sources divided by the number of students enrolled in public schools. Students enrolled in private schools are not included in this figure. In an analysis of the 2007-08 school year graduation rate data, Kabaker (2010) reported a weak positive association between state expenditures per pupil and graduate rates, and she is careful to caution that there is considerably variability in the data. Other researchers such as Baker (2012) have disagreed with this conclusion and have reported “aggregate measures of per pupil spending are positively associated with improved or higher student outcomes” (p. 4).

Findings related to the effects of pupil-teacher ratio on student performance have been mixed in the literature. Hanushek (1996) and Eide and Showalter (1998) concluded that pupil-teacher ratio is insignificant for predicting student success. Krueger and Whitmore (2001) and Angrist and Lavy (1999) reported that smaller class sizes had positive effects on student achievement, particularly for minority students, but their findings are limited to elementary and middle school students and therefore cannot be assumed to translate to secondary schools.

***Data Analysis***

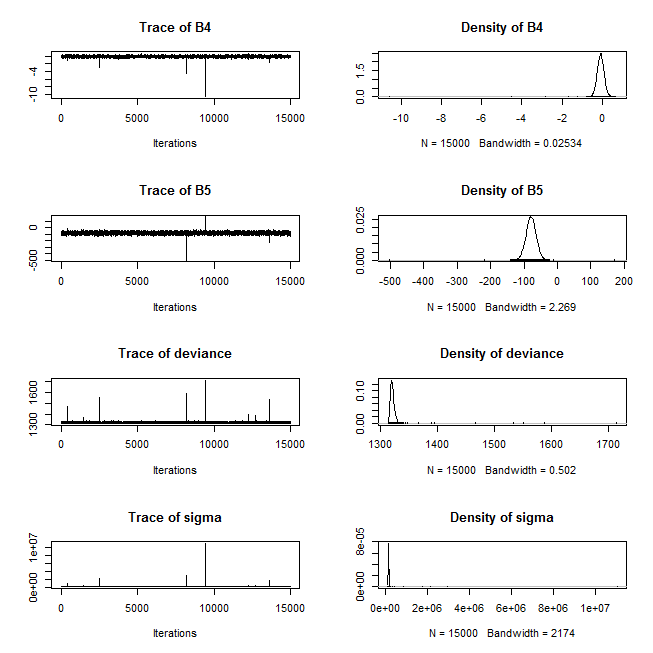
A Bayesian linear model of the form was initially fit to the data. The hierarchical structure used to create this model is borrowed from Kruschke (2010) and is shown in Figure 1. The prior distribution for each parameter estimate is an uninformative normal with mean zero and precision. The data are modeled by a normal likelihood function with mean and precision Gamma(0.01, 0.01).



*Figure 1: Hierarchical Diagram of Multiple Regression Model, Fig. 17.4 in Kruschke (2010).*

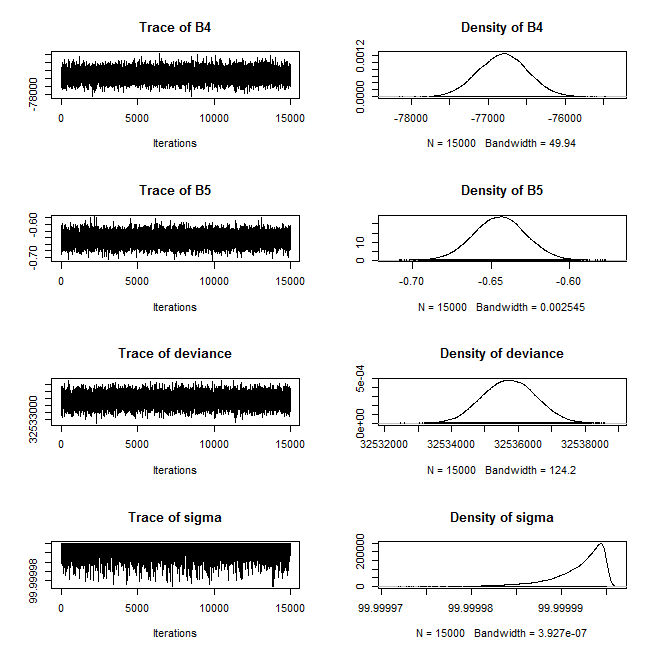
A frequentist multiple regression model was first fit to the data to determine if the data required transformation. A Q-Q plot of the model residuals indicated non-normality, as did the Shapiro-Wilk normality test with p = 0.001. A log-likelihood plot for the Box-Cox transformation parameter suggested may be appropriate. A new model of the form was fit to the data. A scale-location plot indicated that the transformed model had constant residual variance, and a Q-Q plot showed approximate residual normality. The transformed response is therefore used to form the Bayesian regression model.

A Jags MCMC simulation was run with three chains with 10000 iterations each. No thinning was performed, and 5000 burn-in steps were run. No scaling or centering was done to the predictor variables. While the posterior densities for some parameters were well-represented, several extreme values were accepted in the chains of others. Figure 2 shows the MCMC chains and posterior distributions for the parameters . Several extreme-valued spikes are evident in the chains, and this results in the posterior distributions spanning too wide an interval. Consequently, the bulk of the probability mass appears inappropriately compacted into a narrow interval.



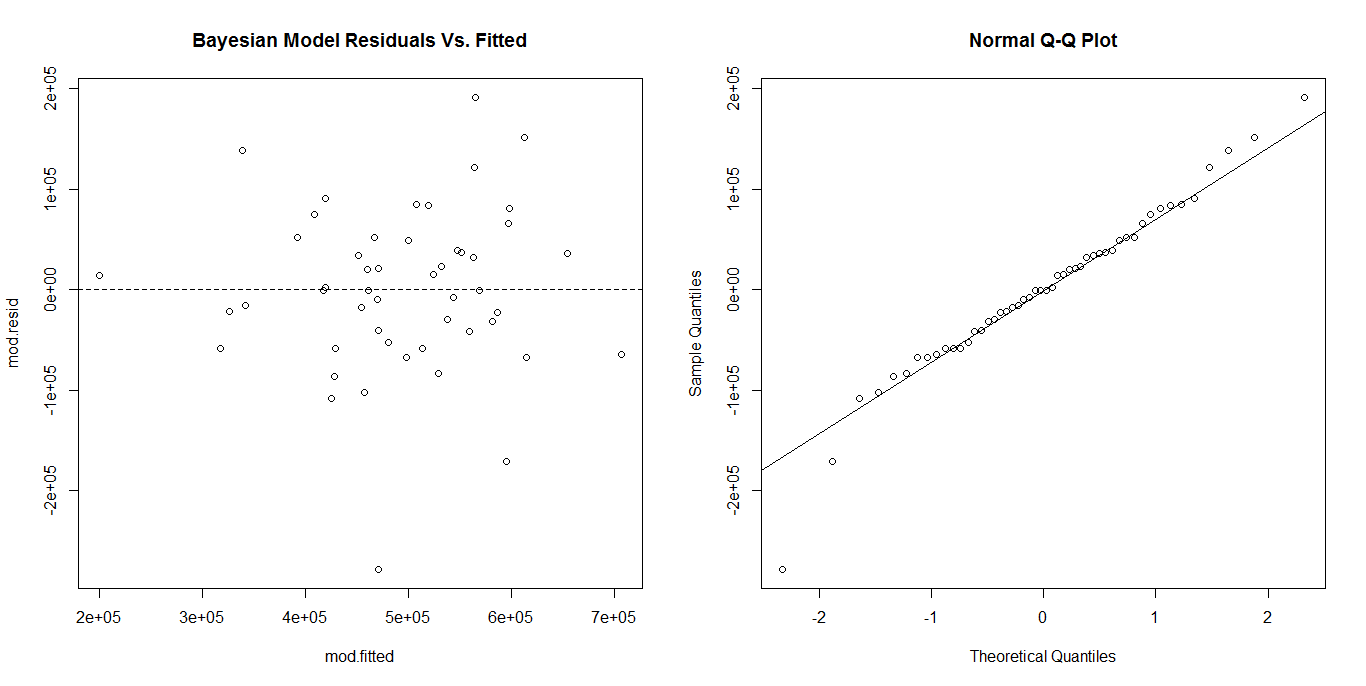
*Figure 2:Posterior Distributions for from Initial Model*

Rather than centering and scaling the predictors, which increases the difficultly of interpreting the results, the prior distribution for the likelihood standard deviation was changed to a uniform distribution on the interval [0, 100], and . This modification was first suggested by Gelman (2006) as an effective non-informative prior for the variance of the mean response in Bayesian hierarchical models. This modification dramatically improved the performance of the MCMC chains. Figure 3 shows the resulting chains and posterior distributions for , which clearly include far fewer outliers than the posteriors in Figure 2. The posterior distributions for the other parameters are similarly well-represented, and for each parameter estimate is 1.001, which indicates convergence.



*Figure 3: Posterior Distributions for from Model*

To verify assumptions of residual normality and constant variance, a normal Q-Q plot and a plot of the residual versus fitted values plot were created. These plots are shown in Figure 4 and confirm that the model assumptions are acceptable.

**

*Figure 4: Residual vs. Fitted and Normal Q-Q Plot for Bayesian Model*

The mean estimates and 95% highest density intervals (HDIs) for the posterior distributions of the parameters are shown in Table 1. Parameters that do not contain zero within their HDI are considered statistically significant, and those that contain zero are not considered significant. All parameters were significant and therefore will be retained.

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Mean Estimate** | **HDI** |
|  | 1255003.42 | (1254570.92, 1255436.14) |
|  | -15788.32 | (-15803.01, -15773.78) |
|  | -9.85 | (-9.867, -9.837) |
|  | -874054.89 | (-874407.03, -873699.38) |
|  | -76806.86 | (-77439.72, -76166.24) |
|  | -0.0645 | (-0.068, -0.61) |

*Table 1: Parameter Estimates and 95% HDIs*

Using the mean parameter estimates from Table 1, the final model is:

***Conclusion***

This analysis found that pupil-teacher ratio, total revenue per pupil, the proportion of students receiving free or reduced price lunch, the proportion of students with limited English proficiency, and population density are all significant state level predictors of state average freshman graduation rates. Since these results are based on state characteristics, they represent only broad trends and cannot be used to make predictions or recommendations about individual students, classrooms, schools, or districts. As Jordan and Kostandini (2012) note, a complex mix of factors, including but not limited to student family characteristics, peer characteristics, religious beliefs, geographic location, and race must be considered to understand dropout rates on a smaller scale.

The coefficient suggests that a one point increase in pupil teacher ratio is predicted to correspond with a mean 15788.3 point decrease in the cube of AFGR. This suggests that states with larger class sizes tend to have lower average freshman graduation rates. Given the disagreement that has surrounded class size in the literature, this finding should be cautiously interpreted, although it does merit further investigation.

Total revenue per pupil, a factor contentiously debated in education policy circles, was found to be negatively associated with AFGR in the model. A one dollar increase in educational revenue per pupil is predicted to correspond with a 9.85 point decrease in the cube of AFGR. This result is unexpected as past studies have typically report either no effect or a positive effect on student performance associated with increased educational expenditures. This may be in part due to the effect of areas like Washington D.C. and Alaska where revenue per pupil is considerably above average while AFGRs are below average. As there are many ways in which expenditure per pupil can be measured, a different metric may very well yield a different result. Additional research is needed to clarify this effect.

Each increase of one, or 100%, in the students eligible for free or reduced lunch corresponds with an 874054.9 point decrease in the cube of AFGR, so a one percent increase is predicted to have a 8740.549 point decrease in the cube of AFGR. Schools with a greater proportion of students in poverty tend to have lower AFGRs. This finding was expected and corresponds with the near-consensus studies conducted over the last several decades.

A percentage point increase in students with limited English proficiency is predicted to correspond with a 768.069 decrease in the cube of AFGR. Thus, states with higher proportions of LEP students are predicted to have lower AFGRs. Orfield and Swanson (2004) reported a similar result in a study of the nation’s largest 100 school districts. They found English language learners had below average graduation rates, which is likely due to inadequate academic supports. Further work, and perhaps funding, is needed to develop supports for LEP students, which may include interventions such as the training of additional educators in LEP instruction or the development multilingual education programs.

Finally, states with higher population densities, a rough measure of urbanicity, tended to have lower AFGRs. For every additional person per square mile, the cube of AFRG is predicted to decrease by 0.0645 points. This finding corresponds with the findings of Swanson (2004a) that urban areas contain many of the lowest-performing districts, many of which have high proportions of low income and minority students. In a 2009 study, Swanson reported that the average graduation rate in the 50 largest U.S. cities was 51%, while in suburban areas the rate was 71%. This difference has been largely explained by district wealth differences, although there are many other contributing factors that have been suggested in past literature. Additional research is needed to determine effective interventions in low-income urban areas.

***Appendix: R Code***

*require(R2jags)*

*require(MASS)*

*#Read data*

*drop <- read.csv(file.choose())*

*#Convert FR lunch + LEP to proportions*

*for (i in 5:6) {drop[,i]<-drop[,i]/drop[,7]}*

*#Make more readable column names*

*cnames <- c("state","PTR", "TRev","FGrad","FRLun","LEP", "Total","TotalState","PopDens")*

*colnames(drop) <- cnames*

*#Fit traditional MR model to determine if predictor/response transformation is necessary*

*#BoxCox transformation suggests lambda = 3*

*modl1=lm(FGrad~PTR+TRev+FRLun+LEP+PopDens,drop)*

*par(mfrow=c(2,2))*

*plot(modl1)*

*boxcox(modl1,lambda = seq(0,4,by = .01))*

*modl2=lm(FGrad^3~PTR+TRev+FRLun+LEP+PopDens,drop)*

*plot(modl2)*

*#Transformed response vector*

*fgrad3 <- drop$FGrad^3*

*# Model attempt 1: No standardization, gamma prior for mean resp. variance*

*reg.mdl=function() {*

*# likelihood*

*for (i in 1:nData) {*

*y[i]~dnorm(mu[i],tau)*

*mu[i] <- inprod(B[],X[i,])*

*}*

*#priors*

*for (j in 1:nParms) {*

*B[j]~dnorm(0,1E-6)*

*}*

*tau ~ dgamma(0.01,0.01)*

*sigma <- 1/sqrt(tau)*

*}*

*reg.data=with(drop,list(y=fgrad3,X=cbind(1,PTR,TRev,FRLun,LEP,PopDens),*

*nData=nrow(drop),nParms=6))*

*reg.inits=function() list(tau=rgamma(1,0.1,0.1),B=rnorm(6,0,10))*

*nChains=3*

*burnIn=5000*

*nIter=10000*

*nThin=1*

*params=c('sigma','B')*

*#Run simulation*

*reg.fit=jags(reg.data,reg.inits,params,reg.mdl,nChains,nIter,burnIn,nThin)*

*print(reg.fit)*

*plot(reg.fit)*

*mcmcChain=as.mcmc(reg.fit$BUGSoutput$sims.matrix)*

*windows()*

*plot(mcmcChain)*

*colnames(mcmcChain)*

*colnames(mcmcChain) <- c("B0","B1","B2","B3","B4","B5","deviance","sigma")*

*# Version 2: uniform prior for error structure*

*reg.mdl=function() {*

*# likelihood*

*for (i in 1:nData) {*

*y[i]~dnorm(mu[i],tau)*

*mu[i] <- inprod(B[],X[i,])*

*}*

*#priors*

*for (j in 1:nParms) {*

*B[j]~dnorm(0,1E-6)*

*}*

*#Try uniform prior for standard deviation as suggested by Gelman*

*sigma~dunif(0,100)*

*tau<- 1/sigma^2*

*}*

*reg.inits=function() list(sigma=runif(1,0,10),B=rnorm(6,0,10))*

*reg.fit=jags(reg.data,reg.inits,params,reg.mdl,nChains,nIter,burnIn,nThin)*

*print(reg.fit)*

*windows()*

*plot(reg.fit)*

*mcmcChain=as.mcmc(reg.fit$BUGSoutput$sims.matrix)*

*windows()*

*colnames(mcmcChain) <- c("B0","B1","B2","B3","B4","B5","deviance","sigma")*

*plot(mcmcChain)*

*betas.mod=cbind(B[,1],B[,2:6])*

*# examine correlation structure of parameter ests.*

*round(cov(betas.mod),3)*

*round(vcov(modl1),3)*

*round(cov(B),3)*

*round(cor(betas.mod),3)*

*round(cor(B),3)*

*# checking model fit - residual analysis*

*beta.bar=colMeans(betas.mod)*

*mod.fitted=cbind(1,as.matrix(drop[,c(2,3,5,6,9)]))%\*%beta.bar*

*mod.resid=fgrad3-mod.fitted*

*windows()*

*plot(density(mod.fitted))*

*par(mfrow=c(2,2))*

*plot(mod.resid~mod.fitted,main='Bayesian Model Residuals Vs. Fitted')*

*abline(h=0,lty=2)*

*windows()*

*par(mfrow=c(1,2))*

*#Compare to frequentist MR fit*

*plot(modl1,which=1)*

*qqnorm(mod.resid)*

*qqline(mod.resid)*

*plot(modl1,which=2)*

***References***

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*,*114*(2), 533-575.

Balfanz, R., & Legters, N. (2006). *The Graduation Rate Crisis We Know and What Can be Done About It Education Week Commentary*, July 12, 2006.

Baker, B. The Albert Shanker Institute, (2012). *Revisiting the age old question: does money matter in education?*. Retrieved from website: http://www.shankerinstitute.org/images/doesmoneymatter\_final.pdf

Cataldi, E. F., & Kewal Ramani, A. (2009). *High School Dropout and Completion Rates in the United States: 2007 Compendium Report*. NCES 2009-064.National Center for Education Statistics.

Chapman, C., Laird, J., & KewalRamani, A. (2013). *Trends in high school dropout and completion rates in the United States: 1972-2009*. BiblioGov.

Eide, E., & Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, *58*(3), 345-350.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, *1*(3), 515-534.

Greene, J. P., & Winters, M. A. (2001). *High school graduation rates in the United States*. Black Alliance for Educational Options.

Hanushek, E. A. (1996). School resources and student performance. *Does money matter*, 43-73

Kabaker, J. (2010, 09). *Examining the data: State per pupil expenditures and state graduation rates*. Retrieved from [http://edmoney.newamerica.net/blogposts/2010/examining\_the\_data\_state\_per\_pupil\_exp en](http://edmoney.newamerica.net/blogposts/2010/examining_the_data_state_per_pupil_exp%09en)ditures\_and\_state\_graduation\_rates-36914

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college‐test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, *111*(468), 1-28

Kruschke, J. (2010). *Doing Bayesian Data Analysis: A Tutorial Introduction with R*. Academic Press

Jordan, J. L., Kostandini, G., & Mykerezi, E. (2012). Rural and Urban High School Dropout Rates: Are They Different?. *Journal of Research in Rural Education*, *27*(12), 1-21.

Orfield, G. & Swanson, C.B. (2004). *Losing our future: How minority youth are being left behind by the graduation rate crisis.*

Swanson, C. B. (2004a). *The real truth about low graduation rates, an evidence-based commentary.*

Swanson, C. B. (2004b). *Who Graduates? Who Doesn't?: A Statistical Portrait of Public High School Graduation, Class of 2001*.

Swanson CB (2009). *Cities in crisis 2009: Closing the graduation gap*. Bethesda, MD: Editorial Project in Education.

U.S. Census Bureau (2012). *State population – rank, percent change, and population density*. Statistical abstract of the United States: 2012 (p. 19). Washington, D.C.

U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "State Dropout and Completion Data File", 2009-10; "State Nonfiscal Public Elementary/Secondary Education Survey", 2010-11. Retrieved from website: <http://nces.ed.gov/ccd/>

U.S. Department of Labor, Bureau of Labor Statistics. (2013). *Earnings and unemployment rates by educational attainment*. Retrieved from website: <http://www.bls.gov/emp/ep_chart_001>.htm