An application of Bayesian variable selection to international economic data

By

Xiang Tian, B.A.

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of science
in

Statistics

University of Alaska Fairbanks

June 2017

APPROVED:

Scott Goddard, Committee Chair
Ron Barry, Committee Member
Margaret Short, Committee Member
Julie McIntyre, Committee Member
Leah Berman, Chair
Department of Mathematics and Statistics

Abstract

GDP plays an important role in people's lives. For example, when GDP increases, the unemployment rate will frequently decrease. In this project, we will use four different Bayesian variable selection methods to verify economic theory regarding important predictors to GDP. The four methods are: g-prior variable selection with credible intervals, local empirical Bayes with credible intervals, variable selection by indicator function, and hyper-g prior variable selection. Also, we will use four measures to compare the results of the various Bayesian variable selection methods: AIC, BIC, Adjusted-R squared and cross-validation.

Keywords: GDP, Bayesian statistical methods, Markov chain Monte Carlo, Bayesian variable selection, g-prior.

Table of Contents

1 Introduction

In this paper, we will select models for predicting a nation's gross domestic product (GDP) using a set of candidate predictors. GDP plays an important role in people's lives. For example, standard economic theory predicts that when GDP increases, the unemployment rate will decrease. As many economists have noted, GDP is a flawed measure of economic welfare. Leisure, inequality, mortality, morbidity, crime, and the natural environment are just some of the major factors affecting living standards within a country, and these factors are incorporated imperfectly, if at all, in GDP. It is nevertheless worth modeling because GDP is so important.

Since GDP is so important, many economists have built models to predict it. Huerta and Freitas Lopes (2000) analyzed the Brazilian industrial production index using a Bayesian time series method to fit a Bayesian model and make short-term forecasts. In our paper, posterior estimates and predictors are obtained using Markov chain Monte Carlo (MCMC) methods based on the Gibbs sampler. There are some differences between our project and the Huerta etal's paper. First, we fit a model using data from many nations instead of using just one country$'$s industrial production index. Secondly, whereas Huerta and Freitas Lopes collected the Brazilian$'$s industrial production index from multiple years, 1980 to 1998, we use only the GDP of 2010. As another example, Fang and Miller (2008) analyzed the volatility of real GDP growth for Japan using a time series model.

In addition to modeling GDP, this paper is concerned with comparing several Bayesian variable selection methods. Bayesian variable selection (BVS) has a long history (Zellner 1971; Leamer 1978; Mitchell and Beauchamp 1988). The advent of Markov chain Monte Carlo methods catalyzed the development of Bayesian model selection and Bayesian model averaging in regression models (George and McCulloch 1993; Smith and Kohn 1996; Raftery, Madigan, and Hoeting 1997; Hoeting, Madigan, Raftery, and Volinsky 1999; Clyde and George 2004). Bayesian variable selection is a class of methods within the Bayesian paradigm that is used for selecting important predictors from among a set of candidate predictors. One advantage of BVS over frequentist methods like LASSO and SCAD is having posterior distributions on parameters, which gives us the ability to quantify our uncertainty about them.

Another advantage is having posterior predictive distributions, which gives us the ability to simulate data for prediction and quantify uncertainty in them. The general types of BVS are posterior-based methods, Bayes factors-based methods, and information criteria. In this project, we will use four different posterior-based methods to model the relationship between our response variable (GDP) and ten candidate predictors. The four variable selection methods are: g-priors with credible intervals, empirical Bayesian g-priors with credible intervals, indicator variable selection, and hyper g-priors with credible intervals. These variable selection methods will be compared in order to determine which performs the best.

We have also chosen to use four ways to compare the performance of our selected sets of variables: AIC, BIC, adjusted R-squared, and cross validation.

The remainder of the paper is organized as follows: in Section 2, we introduce the original data; in Section 3, we introduce the four Bayesian variable selection methods and four measures to compare the performance of the various Bayesian variable selection methods; in Section 4, we provide the results of the Bayesian variable selection methods applied to our data set; the last section contains discussion and future work.

2 Data

We obtained our data from the World Bank website (2017). The data set contains information on 217 countries; however, because of missing data, we used the data from only 79 countries. Though several years were included in the original data set, we decided to use the data from just 2010 to conduct our analysis. This year is recent enough to be relevant but also old enough to incorporate data revisions.

In our data set, we have one response variable (GDP) and ten predictors. In Table 1 we introduce the candidate predictors used in our analysis. The units of the response variable (GDP) are U.S. dollars; the first predictor is expenditures on education, expressed as a percentage of total government expenditures; the second predictor is exports of goods and services, expressed as a percentage of GDP; the third predictor is fertility rate, which is the average number of births per woman; the fourth predictor is general government final consumption expenditures, expressed in U.S. dollars; the fifth predictor is gross savings, expressed as a percentage of GDP; the sixth predictor is household final consumption expenditures, expressed in U.S. dollars; the seventh predictor is imports of goods and services, expressed as a percentage of GDP; the eighth predictor is inflation rate, expressed as a percentage; the ninth predictor is military expenditures, expressed as a percentage of government expenditures; the last predictor is population ages 15-64, expressed as a percentage of total population count.

The choice to use these specific candidate predictors in this study is rooted in economic theory. According to economists, there are two ways to partition the total value of GDP (Macroeconomics: A Growth Theory Approach, Alejandro and Mark). One way to obtain GDP is by expenditure (the different ways that our output is “bought”), using the following formula:

GDP= consumption+ investment +government production+ net exports.

The other way is by income:

GDP=wages+profits.

Therefore, according to economic theory, some variables should have a significant relationship with GDP. Among the ten candidate predictors we used were some which would almost certainly have a significant relationship with GDP, such as consumption, exports, imports, and investment. We also included some predictors which we did not expect to have an effect on GDP in order to test the specificity of the various variable selection methods. When we perform the variable selection, those variables ought to be eliminated from our model.

3 Methods

3.1 Multiple linear regression

In this section we describe linear regression methods and build up to the Bayesian variable selection methods. Part of the reason for doing this is to describe the transformations we

Table 1: Candidate predictors used in our analysis

Variable	Variable's name
X_1	Expenditures on education
X_2	Exports
X_3	Fertility rate
X_4	Government consumption expenditures
X_5	Gross savings
X_6	Household consumption expenditures
X_7	Imports
X_8	Inflation
X_9	Military expenditures
X_{10}	Population ages 15-64 (% total)

selected for the data.

In statistics, linear regression is an approach for modeling the relationship between a dependent variable Y and p explanatory variables (or independent variables) denoted X_1, X_2, $\cdots$, X_p. Specifically, we are trying to model $E(\mathbf{Y}|\mathbf{X})$ using the linear function $\mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is a matrix containing column vectors for $\mathbf{X}_1$, $\mathbf{X}_2$, $\cdots$, $\mathbf{X}_p$; $\boldsymbol{\beta}$ is a column vector of coefficients; and $\mathbf{Y}$ is a column vector of responses. Another way to write this is $\mathbf{Y}$=$\mathbf{X}\boldsymbol{\beta}$+$\mathbf{e}$, where we assume that the errors $\mathbf{e}$ have a mean of $\mathbf{0}$ and constant variance and are uncorrelated. Ordinary least squares is a method for estimating the unknown parameters in a linear regression model. The goal of the ordinary least squares method is to minimize the sum of the squares of differences between the observed responses in the given data set and those predicted by a linear function of a set of explanatory variables. That is, we minimize $\mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$.

To begin our analysis, we fit the multiple linear regression model,

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{1}\hat{\beta}_0 + \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \cdots + \mathbf{X}_{10}\hat{\beta}_{10}$$

Here, $\mathbf{1}$ is a column vector of 1s. By fitting a multiple linear regression model, we hoped to begin to see the relationships between our response variable and the explanatory variables. Table 2 summarizes the multiple linear regression model fit. Figure 1 provides a scatterplot matrix that depicts the estimated effects from the multiple linear regression model. Ac-

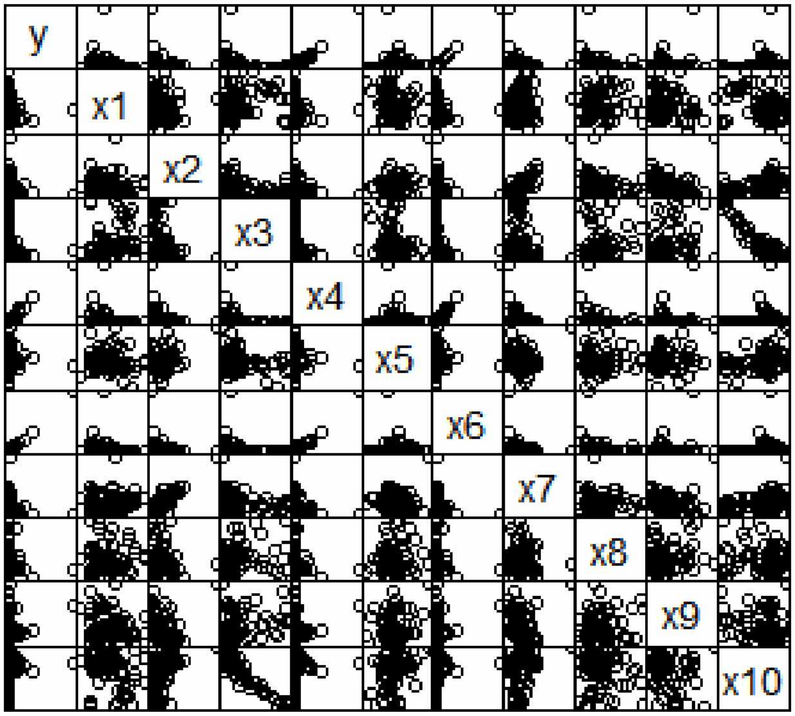


Figure 1: Scatterplot matrix of the raw data

Table 2: Multiple linear regression output of the ordinary least squares fit

Coefficients	Estimate	Std. Error	t value	p-value
Intercept	1.492e+11	2.739e+11	0.545	0.5878
X_1	2.825e+07	1.815e+09	0.016	0.9876
X_2	1.012e+08	7.691e+08	0.132	0.8957
X_3	-1.152e+10	1.643e+10	-0.701	0.4855
X_4	1.920e+00	1.548e-01	12.401	$< 2e - 16$ ***
X_5	3.516e+09	1.088e+09	3.231	0.0019 **
X_6	1.002e+00	4.038e-02	24.822	$< 2e - 16$ ***
X_7	-3.640e+08	8.555e+08	-0.425	0.6719
X_8	8.741e+08	2.379e+09	0.367	0.7144
X_9	1.049e+08	1.563e+09	0.067	0.9467
X_{10}	-2.520e+09	3.667e+09	-0.687	0.4944

cording to this graph, only a few predictors have a linear relationship with our response variable.

It is clear that in the OLS fit, assuming the model's assumptions are satisfied, only three predictors are significant, since the p-values of those predictors are less than 0.05. We will now investigate the model diagnostics to verify the model's assumptions.

3.2 Data transformations and model diagnostics

The nonlinear relationship depicted in the normal probability plot in Figure 2 indicates that the residuals for the multiple linear regression are not normally distributed. Thus, we should use at least one transformation on the data.

The scatterplot matrix in Figure 1 and the curvature plots included in Appendices A.1 and A.2 allow us to assess the linearity of the relationship between the response and predictors. A hypothesis test is also performed to formally test the null hypothesis that a linear mean function is appropriate. The alternative hypothesis states that the mean function is instead curvilinear. We found that the relationship between variable X_4 and our response variable is curvilinear and the relationship between variable X_6 and our response variable is also curvilinear. To see this better, Figure 3 shows the histogram of variable X_6; it shows

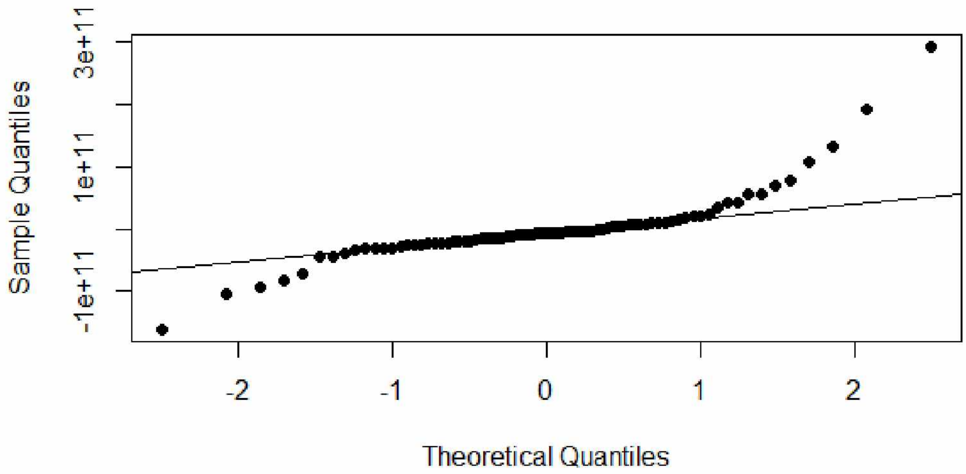


Figure 2: Normal probability plot of residuals before transformation

that this variable has a highly-skewed distribution, which is not good for regression analysis. X_4 has a similar problem. In order to make those two variables' distributions less skewed, we used a log transformation. We also used a log transformation of the response Y in order to improve the normality diagnostics. Finally, we consider the constant variance assumption inherent in multiple linear regression. Figure 4 shows a residual vs. fitted value plot; it suggests that the constant variance assumption is violated. A formal test of non-constant variance also indicates that the variance is non-constant. This problem provides us another justification for the use of log transformations on X_4, X_6, and Y.

We also use Cook's Distance to assess whether the any of the observations have disproportionate influence on the fitted regression model. Cook's distance or Cook's D is a commonly-used estimate of the influence of a data point when performing a least-squares regression analysis. The formula of Cook's distance is:

$$D_i = \frac{(Y_i - \hat{Y}_i)^2}{p \times MSE} \times \frac{h_i}{(1-h_i)^2},$$

where h_i is leverage of i-th observation, $\hat{Y}_i$ is the fitted value of the ith observation, and MSE is the mean squared error of the fit.

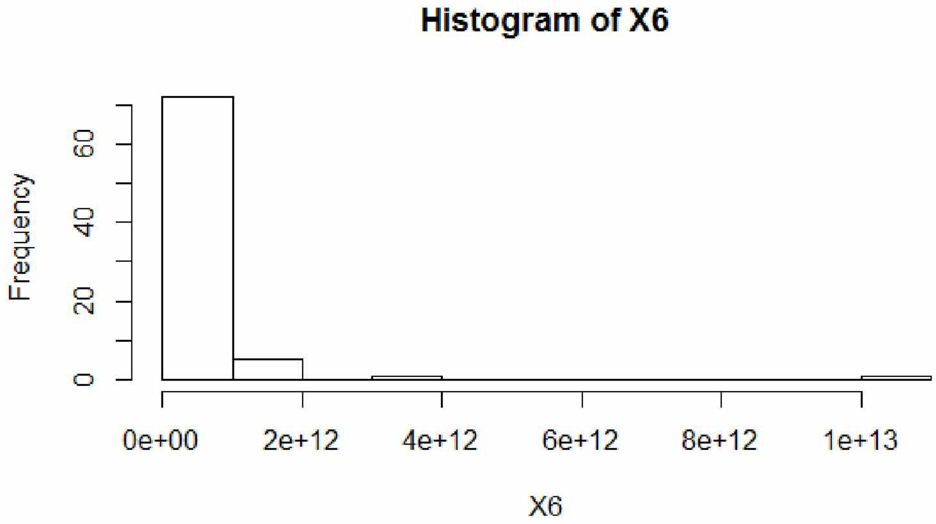


Figure 3: Histogram of the non-standardized variable X_6

There are different opinions regarding what cut-off values to use for spotting highly influential points. A simple operational guideline of $D_i > 1$ has been suggested (Kim 1996). For our data, the largest Cook's distance is 0.491, which is less than 1. We can arrive at the conclusion that none of the observations exert undue influence on our regression analysis. Finally, in order to give the model-fitting algorithms more numerical stability and to put the coefficients on the same scale, we standardized the predictors X_1 through X_{10}. In other words, we subtracted out the mean of each predictor and divided by its sample standard deviation.

Figure 5 and Table 3 depict a scatterplot matrix and multiple linear regression output for the standardized transformed data. Figure 5 indicates that the modeling assumptions for linear regression are better satisfied. Table 3 indicates that all of the predictors except for X_1 (Education expenditures) and X_8 (Inflation) appear to be significant predictors of GDP, and we expect them to appear to our final model.

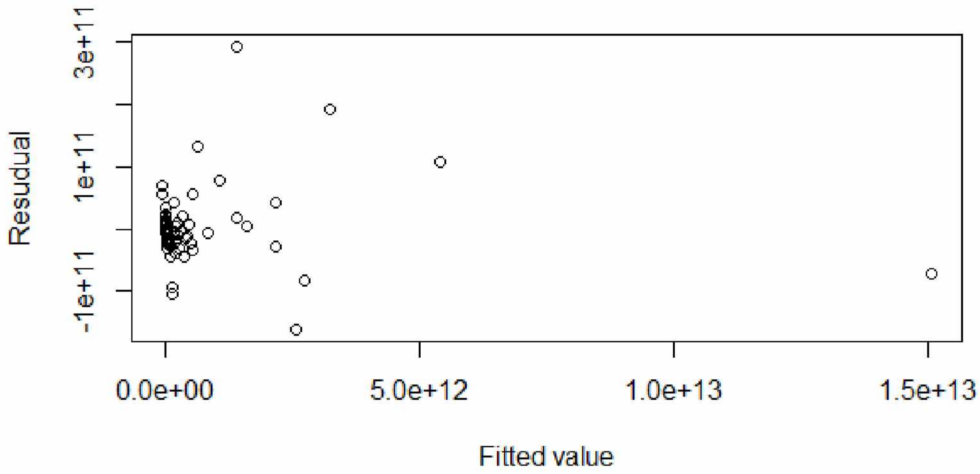


Figure 4: Residuals vs. Fitted values plot using raw, non-transformed data

Table 3: Multiple linear regression output of the ordinary least squares fit of transformed data

Coefficients	Estimate	Std. Error	t value	p-value
Intercept	25.211551	0.005083	4959.605	$< 2e - 16$ ***
X_1	-0.005885	0.006011	-0.979	0.331045
X_2	0.208712	0.016196	12.886	$< 2e - 16$ ***
X_3	0.063403	0.017051	3.718	0.000408 ***
X_4	0.375603	0.032482	11.563	$< 2e - 16$ ***
X_5	0.067602	0.007125	9.489	$4.46e - 14$ ***
X_6	1.580155	0.031692	49.859	$< 2e - 16$ ***
X_7	-0.165789	0.015659	-10.587	$5.04e - 16$ ***
X_8	0.001230	0.005900	0.208	0.835501
X_9	-0.013717	0.006668	-2.057	0.043512 *
X_{10}	0.053378	0.016576	3.220	0.001965 **

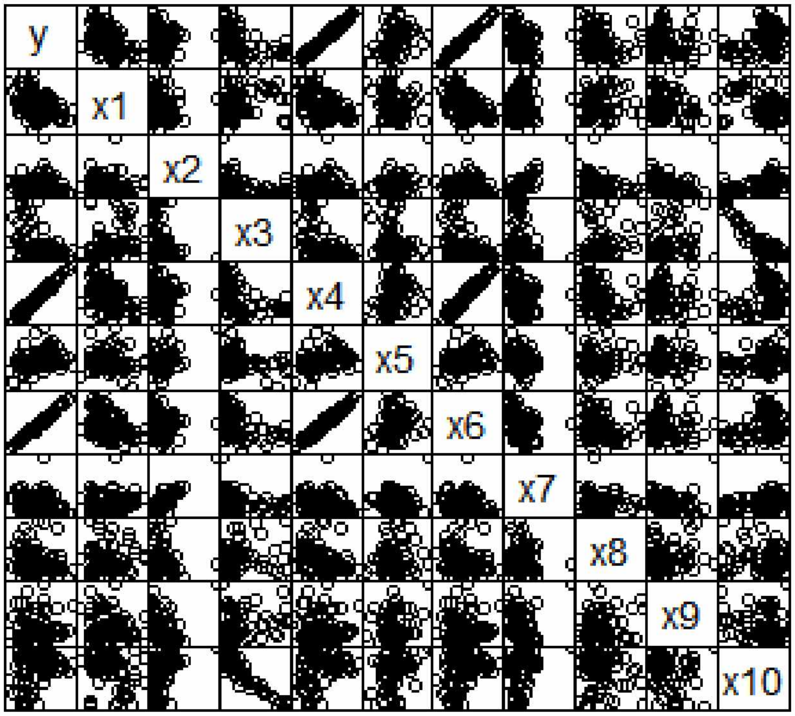


Figure 5: Scatterplot matrix of the transformed data

3.3 Bayesian linear regression

Having applied appropriate transformations to the data set, we now turn to a Bayesian linear model for it. In probability theory and statistics, Bayes' theorem describes the probability of an event, conditional on prior knowledge about associated events.

A prior probability distribution, often simply called the prior, of an uncertain quantity θ or vector of uncertain quantities $\boldsymbol{\theta}$ is the probability distribution that expresses one's beliefs about this quantity before some given data set is taken into account. The prior distribution of $\boldsymbol{\theta}$ is denoted by $\pi(\boldsymbol{\theta})$.

The posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{Y})$, is the conditional probability distribution for $\boldsymbol{\theta}$, conditional on some data. It is constructed by combining the likelihood function and the prior distribution.

In statistics, a likelihood function (often simply the likelihood) is a function of the parameters of a statistical model given data. $L(\mathbf{Y}|\theta)$ is the notation for the likelihood function.

The posterior distribution is proportional to likelihood * prior:

$$\pi(\boldsymbol{\theta}|\mathbf{Y}) = \pi(\text{parameter(s)}|\text{data}) \propto \pi(\boldsymbol{\theta}) * L(\mathbf{Y}|\boldsymbol{\theta})$$

For this project, we will utilize a Bayesian linear regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathsf{e}$, where $\mathsf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. $\mathbf{Y}$ is a 79×1 response vector, $\mathbf{X}$ is the 79×11 design matrix, $\boldsymbol{\beta}$ is a 11×1 vector of coefficients. Thus $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2)^T$, and the likelihood function becomes $L(\mathbf{Y}|\boldsymbol{\beta}, \sigma^2) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp(-\frac{\mathbf{(Y-X\boldsymbol{\beta})^T(Y-X\boldsymbol{\beta})}}{2\sigma^2})$.

In Bayesian linear regression, priors must be set for σ^2 and $\boldsymbol{\beta}$. The g-prior is a certain class of priors for the regression coefficients of a Bayesian multiple regression. The joint prior distribution of σ^2 and $\boldsymbol{\beta}$ factors as

$$\pi(\sigma^2, \boldsymbol{\beta}) = \pi(\sigma^2)\pi(\boldsymbol{\beta}|\sigma^2),$$

which is the product of the prior distribution of σ^2 times the prior distribution of $\boldsymbol{\beta}$ given σ^2. It is common to use an inverse-gamma distribution on σ^2 because it is a conjugate prior. We will prefer the improper prior,

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

which is the limit of an inverse-gamma probability density function parameterized by a and b as a and b approach 0 from above in such a way that $\frac{a}{b}$ is constant. This prior is also conjugate for σ^2. It can be shown that

$$\sigma^2|\mathbf{Y} \sim IG\left(\frac{n}{2}, \frac{1}{2}\mathbf{Y}^T(I - \frac{g}{1+g}P_X)\mathbf{Y}\right),$$

where $P_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The g-prior for $\boldsymbol{\beta}$, conditional on σ^2, is a normal distribution with mean $\mathbf{0}$ and variance $g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. g is a hyperparameter. In other words,

$$\boldsymbol{\beta}|\sigma^2 \sim N(\mathbf{0}, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

We note that $\hat{\boldsymbol{\beta}}_{OLS}$, the ordinary least squares estimator of $\boldsymbol{\beta}$, is $\hat{\boldsymbol{\beta}}_{OLS}$=$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and that var$(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, which serves as partial explanation for why the g-prior is defined the way it is. This prior for $\boldsymbol{\beta}$ is conditionally conjugate, meaning that, conditional on σ^2, both the prior and the posterior are of the same family of density functions. The joint posterior density is likewise the product

$$\pi(\sigma^2, \boldsymbol{\beta}|\mathbf{Y}) = \pi(\sigma^2|\mathbf{Y}) * \pi(\boldsymbol{\beta}|\sigma^2, \mathbf{Y}),$$

which is the product of the posterior distribution of σ^2 and the posterior distribution of $\boldsymbol{\beta}$, given σ^2. It can be shown that the conditional posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\sigma^2, \mathbf{Y} \sim N\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_{OLS}, \frac{g\sigma^2}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

That is, the posterior distribution of $\boldsymbol{\beta}$ given σ^2 is normal with mean $\frac{g}{1+g}\hat{\boldsymbol{\beta}}_{OLS}$ and variance $\frac{g\sigma^2}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}$.

3.4 Bayesian variable selection

The selection of variables in Bayesian linear regression model is related to the prior assumptions made on the model's parameters. Some general strategies for Bayesian variable selection are posterior-based methods, Bayes factor based methods, and information criteria. In this project, we focus on posterior-based methods and consider four such methods which employ the Bayesian linear regression model. In particular, three of the four methods utilize g-priors.

3.4.1 Basic g-prior with credible intervals

For this method g must be set to some value in order to have a well-defined prior distribution for $\boldsymbol{\beta}$. Various values of g have been proposed. According to the risk inflation criterion (Foster and George 1994), g should be set equal to p^2, where p is the number of predictors; according to the unit information prior (Kass and Wasserman 1995), g should be set equal to n, where n is sample size. Initially we tried 5 values to compare their effects: 10, 20, 50, 150, 200. Finally we decided to choose g=20, because it gave the lowest value of mean square prediction error in a pilot study.

Because of the conjugacy of the prior distributions, we are able to simulate samples from the joint posterior distribution using a simple loop in R as follows. For each iteration, first sample

$$\sigma^2|\mathbf{Y} \sim IG\left(\frac{n}{2}, \frac{1}{2}\mathbf{Y}^T(I - \frac{g}{1+g}P_X)\mathbf{Y}\right),$$

then

$$\boldsymbol{\beta}|\sigma^2, \mathbf{Y} \sim N\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}_{OLS}, \frac{g\sigma^2}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

We perform variable selection by calculating the 95% credible interval of each component of $\boldsymbol{\beta}$ using posterior samples of $\boldsymbol{\beta}$, which we obtained as described above. If 0 is inside the interval, it means that 0 is a plausible value for the component, and thus we can eliminate this variable from our model. Once we have selected significant variables, we will fit our regression model by taking the posterior mean of each selected variable′s coefficient as an estimate.

Regression diagnostic plots for the g-prior variable selection fitted model can be seen in Appendices A.3 and A.4.

3.4.2 The local Empirical Bayes approach

In the second variable selection method we employed, the local Empirical Bayes approach can be viewed as estimating a separate g for each candidate model. Using the marginal likelihood after integrating out all parameters, an empirical Bayes value of g is the maximum (marginal) likelihood estimate constrained to be nonnegative, which turns out to be $\hat{g}_{EBL} = \max(F-1, 0)$, where $F = \frac{R^2/p}{(1-R^2)/(n-1-p)}$ is the usual F statistic for testing $\beta_1, \ldots, \beta_p$ all equal 0.

Once g is set as per above, the variable selection method proceeds just as the g-prior with credible intervals method does.

Regression diagnostic plots for a local empirical Bayes variable selection fitted model can be seen in Appendices A.5 and A.6.

3.4.3 Indicator variable selection method

At present, the computational method most commonly used for fitting Bayesian models with intractable posteriors is the Markov chain Monte Carlo (MCMC) technique (Robert and Casella 2004). Variable selection methods can take advantage of the MCMC framework. Indicator model selection does not use a conjugate prior as do the g-prior based model selection methods. Thus MCMC is required to estimate the posterior distributions of the parameters.

For the purpose of our project, indicators are functions which can either take a value of 1 or 0 in order to indicate whether a predictor variable belongs in the model. We use I_i as our indicator for i=1, 2, . . . , 10. By including the indicator in our model, we can decide which variables should be eliminated from our regression model. The resulting regression model is:

$$Y = \beta_0 + X_1(\mu_1 I_1) + X_2(\mu_2 I_2) + \cdots + X_{10}(\mu_{10} I_{10}) + e,$$

where

$$I_i = \begin{cases} 1 & \text{if variable } i \text{ belongs in the model} \\ 0 & \text{if otherwise} \end{cases}$$

In order to give each variable an equal chance of being eliminated from our model, we assumed that $I_i \sim$ Bernoulli(0.5) for i=1, 2, $\cdots$, 10. Also, we assumed that $\mu_i \sim$Normal(0,τ_i), and $\tau_i \sim$Gamma(1,1). The model was fitted using OpenBUGS, which was called from R using R2OpenBUGS (Sturtz et al. 2010). We performed 10,000 iterations with 100 more iterations for burn-in. We eliminated variables for which the posterior mean of the corresponding indicator was less than 0.5.

Regression diagnostic plots for the variable selection by indicator function method and traceplots of β_0, I_1 and U_1 can be seen in Appendices A.7 through A.11.

3.4.4 Hyper-g prior with credible intervals

This method is discussed in (Liang et al.2008); we summarize it here. The shrinkage factor $\frac{g}{1+g}$ in the conditional posterior of $\boldsymbol{\beta}$ given σ^2 is a factor which adjusts the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{OLS}$. It pulls the maximum likelihood estimator toward the prior mean of 0. Instead of requiring us to pick a value of g, the hyper-g prior method allows the data to pick g by placing a prior distribution on either g or on the shrinkage factor. We assumed that the prior distribution of g is $\pi(g)$=$\frac{a-2}{2}(1+g)^{-a/2}$, g>0, which is proper distribution for a>2. In this case, we assumed that a=3 in deference to the aforementioned authors' suggestion. Hyper-g priors are equivalent to the specification of a Beta prior on the shrinkage factor $\frac{g}{1+g}$; that is $\frac{g}{1+g} \sim$ Beta(1,$\frac{a}{2}$-1), which is a Beta distribution with mean $\frac{2}{a}$.

Once again we must use R2OpenBUGS to call OpenBUGS and estimate the posterior distribution of $\boldsymbol{\beta}$ using MCMC, because the incorporation of g as a parameter to be sampled results in a non-conjugate model. In using a Markov chain Monte Carlo algorithm, we performed 10,000 iterations to simulate the posterior distribution of $\boldsymbol{\beta}$ and took the posterior mean of the coefficients of significant variables. Variable selection was again performed using credible intervals from the posterior distributions of regression coefficients.

Regression diagnostic plot for the hyper-g prior variable selection method and the trace-plots of β_0, β_1 and β_2 can be seen in Appendices A.12 through A.16.

3.5 Measures for comparing the methods' results

We would like to be able to compare the results of the four variable selection methods and determine which does best for this data set. To do this, we will use a cross-validation routine to measure the predictive performance of each variable selection method. In addition, we will also calculate three measures of model quality in order to show what variables would have been selected if these had been applied instead. We describe these measures first followed by the cross-validation.

3.5.1 Akaike information criterion (AIC)

The Akaike information criterion (AIC) is a measure of the relative fitness of various statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of the fit of each model based on the maximum of the model's likelihood function. It provides a means for model selection. The model with the lowest AIC is preferred.

For some candidate model of a given data set, let $\mathcal{L}$ be the log-likelihood function for the model and let p be the number of estimated parameters in the model. Then the AIC value of the model is:

$$AIC = 2p - 2\hat{\mathcal{L}}$$

where $\hat{\mathcal{L}}$ is the log-likelihood function evaluated at the maximum likelihood estimate of θ. We note that small values of $2p$ correspond to parsimonious models, and that small values of $-2\hat{\mathcal{L}}$ correspond to models with good fit ($\hat{\mathcal{L}}$ is correspondingly large). This is why we prefer models with small AIC.

3.5.2 Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) is another criterion for comparing models. The model with the lowest BIC is preferred; it differs from AIC in that it has a different penalty for nonparsimonious models:

$$BIC = \log(n)p - 2\hat{\mathcal{L}}$$

3.5.3 Adjusted R-squared

R-squared is another measure used for model comparison. It is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. The bigger the R-squared value is, the better the model fits the data. The formula for R-squared is:

$$R^2 = 1 - \frac{SSE}{SST},$$

where SSE is the sum of squared errors of the regression model and SST is the sum of squares total for the model.

Every time a predictor in regression analysis is added, R^2 increases. Therefore, the more predictors that are added, the better the regression will seem to "fit" the data. Even the addition of predictors which are insignificant will nevertheless increase the value of R^2.

The adjusted R^2 can instead be used to include a more appropriate number of variables, thwarting the temptation to keep on adding variables to a data set. The adjusted R^2 will increase only if a new predictor improves the regression more than would be expected by chance. So, when adding a new predictor into a regression analysis, we would like to use adjusted R-squared to help us make decisions. The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The bigger the adjusted R-squared value is, the better the model will fit. The formula for adjusted R-squared is

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)}{(n - p - 1)}\right],$$

where n is the sample size and p is the total number of explanatory variables in the model.

3.5.4 Cross-validation

Our chosen measure of predictive performance is cross-validation. Many of the model fit statistics are not a good guide to how well a model will predict: high R^2 does not necessarily mean the model makes good predictions. It is easy to over-fit the data by including too many predictors and thereby inflate R^2 and other fit statistics. For example, in a simple polynomial regression we can just keep adding higher order terms and get better and better fits to the data. But the predictions from the model on new data will usually get worse as higher order terms are added.

One way to measure the predictive ability of a model is to test it on a set of data not used in fitting the model, called a "test set". This is known as cross-validation. The data used for estimation is the "training set". In each cross-validation iteration, we randomly divided our data set into two parts: a testing data set and a training data set. The testing data contained about one third of our original data set (29 observations); the training data set contained about two thirds of our original data set (50 observations).

In each iteration, and for each variable selection method, we used the training data set to perform variable selection. Once the variables were selected, we took the posterior means of the coefficients corresponding to the selected variables, resulting in a fitted model. Then we used these fitted models to predict the test data set responses $\mathbf{Y}$. Finally, when we differenced the observed test variable Y_i and the predicted response $\hat{Y}_i$, we obtained the prediction error, Y_i-$\hat{Y}_i$. Using the mean of the squared prediction errors (MSPE), which is averaged over all squared prediction errors in each cross-validation iteration and then averaged over 100 iterations, we can compare the variable selection methods. If MSPE is large, the method has a poor predictive performance. Likewise, if MSPE is low, the method has a good predictive performance.

4 Results

We now present the results of our study, starting with the variables selected by each method. Table 4 gives the output of four Bayesian variable selection methods. It shows that the g-prior with credible intervals selected exactly one predictor, X_6, to stay in the model. This is concerning, given that Figure 5 shows that X_4 is also strongly linearly associated with the response. We see that the empirical Bayes method selected X_2, X_3, X_4, X_5, X_6, X_7, and X_{10}. Indicator variable selection chose X_4 and X_6 only, while hyper-g prior variable selection chose X_2 and X_6. Though every method selects X_6, there is wide disagreement regarding several of the other candidate predictors.

According to economic theory, at least four predictors (consumption, exports, imports, and investment) should have a significant impact on GDP; and according to the output of Table 4, only the local empirical Bayes method selected at least four predictors. This indicates this method works best according to economic theory.

By calculating the mean squared prediction error of the four methods, over the 100 cross-validation iterations, we can decide which variable selection method performs best empirically for this data set. Table 5 gives a comparison of the four methods in terms of mean square prediction error (MSPE). It is evident that the local empirical Bayes method performs best, since it gives the lowest value of mean squared prediction error. By this measure, g-prior variable selection performs worst. The discrepancy between these two methods may be due to the values of g used by each. In the local empirical Bayes method, g is set to be nearly 10,000; in the g-prior variable selection with credible intervals method, g is set to be 20.

Table 5 also provides the values of AIC, BIC, and adjusted R-squared for the models selected by each method. These are provided merely for reference; we do not imply that the Bayesian variable selection methods can be assessed using such criteria. It is evident that AIC and BIC both favor the same model selected by local empirical Bayes. By the same token, the model selected by g-prior variable selection is not favored by any of these measures.

According to output of the cross-validation as well as using economic theory, we conclude

Table 4: Variable selection by four methods on the full data set

	g-prior VS	EB VS	Indicator VS	Hyper-g prior VS	Theory
X_1	No	N	N	N	
X_2	N	Y	N	N	Y
X_3	N	Y	N	N	M
X_4	N	Y	Y	N	Y
X_5	N	Y	N	N	
X_6	Y	Y	Y	Y	Y
X_7	N	Y	N	N	Y
X_8	N	N	N	N	
X_9	N	N	N	N	
X_{10}	N	Y	N	N	M

Y=Yes(include), N=No(omit), M=Maybe

Table 5: Comparison of the four methods of Bayesian variable selection

Model Selection	Mean MSPE	AIC	BIC	adjusted R^2
g-prior variable selection	0.2951141	-36.77993	-29.67159	0.9913
Local empirical Bayes	**0.1064967**	**-253.9453**	**-232.6203**	**0.9995**
Variable selection by indicator function	0.1104306	-84.98479	-75.50699	0.9953
Hyper-g prior variable selection	0.3195414	-79.64559	-70.1678	0.9950

that the local empirical Bayes method performs best in this analysis.

5 Discussion and future work

In our project, we use four Bayesian variable selection methods to verify economic theory regarding important predictors to GDP. Also, we use four measures to compare the results of the various Bayesian selection methods.

According to classical economics theory, consumption, investment, exports, and imports have a significant impact on GDP. According to the output of Table 4, the local empirical Bayes methods similarly finds that in 2010, consumption, investment, exports and imports do have a significant impact on GDP. Furthermore, and interestingly, the local empirical Bayes method also finds that fertility rate and population ages 15-64 (% total) also have a significant impact on GDP. According to Barro (2001), economic growth is significantly negatively related to the total fertility rate. According to Abegunde (2007), population ages 15-64 (%) has a positive significant relationship with the growth of GDP.

As measured by the cross-validation routine and economic theory, we believe that the local empirical Bayes selection method works best for this data set, and the g-prior variable selection method works worst. In terms of computational time, g-prior variable selection and local empirical Bayes selection run faster than the others. Hyper-g prior variable selection runs the slowest. The easiest method to implement is local empirical Bayes selection and the hardest method to implement is hyper-g prior variable selection, which requires the use of vector-valued function and multivariate distributions in the OpenBUGS model program. The R-code can be seen in Appendices A.17 through A.20.

There are a number of potential ways to extend this work. We could refit the model after doing the variable selection. By doing this, we might obtain a better-fitting model since we eliminate some insignificant predictors from our model. Also, we could compare methods using the deviance information criterion (DIC) in addition to AIC and BIC. Moreover, we could use Bayes factors instead of posterior distributions with credible intervals to do the variable selection. This is a more natural way for g-prior and hyper-g prior variable selection. Furthermore, we might do a sensitivity analysis for our priors and the hyperparameter in the hyper-g prior. Finally, we could try to simulate predictions from the posterior predictive distribution in order to predict training data in cross validation.

Acknowledgements

I would like to thank my committee chair, Scott Goddard of the University of Alaska Fairbanks for a large amount of help for this project, as well as my committee members. Ron Barry, Margaret Short, and Julie McIntyre for coding and course work. The data in this project can be downloaded from http://data.worldbank.org/.

References

[1] Abegunde, Dele O., Colin D. Mathers, Taghreed Adam, Monica Ortegon, and Kathleen Strong. "The burden and costs of chronic diseases in low-income and middle-income countries." The Lancet 370, no. 9603 (2007): 1929-1938.

[2] Barro, Robert J. "Human capital and growth." The American Economic Review 91, no. 2 (2001): 12-17.

[3] Clyde, Merlise, and Edward I. George. "Model uncertainty." Statistical science (2004): 81-94.

[4] Fang, WenShwo, and Stephen M. Miller. "Modeling the volatility of real GDP growth: The case of Japan revisited." Japan and the World Economy 21, no. 3 (2009): 312-324.

[5] Foster, Dean P., and Edward I. George. "The risk inflation criterion for multiple regression." The Annals of Statistics (1994): 1947-1975.

[6] George, Edward I., and Robert E. McCulloch. "Variable selection via Gibbs sampling." Journal of the American Statistical Association 88, no. 423 (1993): 881-889.

[7] Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. "Bayesian model averaging: a tutorial." Statistical science (1999): 382-401.

[8] Huerta, Gabriel, and Hedibert Freitas Lopes. "Bayesian forecasting and inference in latent structure for the brazilian industrial production index." Brazilian Review of Econometrics 20, no. 1 (2000): 1-26.

[9] Kass, Robert E., and Larry Wasserman. "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." Journal of the american statistical association 90, no. 431 (1995): 928-934.

[10] Kim, Choongrak. "Cook's distance in spline smoothing." Statistics and probability letters 31, no. 2 (1996): 139-144.

[11] Leamer, Edward E. "Specification searches: Ad hoc inference with nonexperimental data". Vol. 53. John Wiley and Sons Incorporated, 1978.

[12] Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and Jim O. Berger. "Mixtures of g priors for Bayesian variable selection." Journal of the American Statistical Association 103, no. 481 (2008): 410-423.

[13] Mitchell, Toby J., and John J. Beauchamp. "Bayesian variable selection in linear regression." Journal of the American Statistical Association 83, no. 404 (1988): 1023-1032.

[14] O'Hara, Robert B., and Mikko J. Sillanpää. "A review of Bayesian variable selection methods: what, how and which." Bayesian analysis 4, no. 1 (2009): 85-117.

[15] Raftery, Adrian E., David Madigan, and Jennifer A. Hoeting. "Bayesian model averaging for linear regression models." Journal of the American Statistical Association 92, no. 437 (1997): 179-191.

[16] Smith, Michael, and Robert Kohn. "Nonparametric regression using Bayesian variable selection." Journal of Econometrics 75, no. 2 (1996): 317-343.

[17] Sturtz, Sibylle, Uwe Ligges, and Andrew Gelman. "R2OpenBUGS: a package for running OpenBUGS from R."

URL http://cran. rproject. org/web/packages/R2OpenBUGS/vignettes/R2OpenBUGS. pdf (2010).

[18] World Bank Open Data http://data.worldbank.org/ (accessed April 10, 2016).

[19] Zellner, Arnold, and Claude Montmarquette. "A study of some aspects of temporal aggregation problems in econometric analyses." The Review of Economics and Statistics (1971): 335-342.

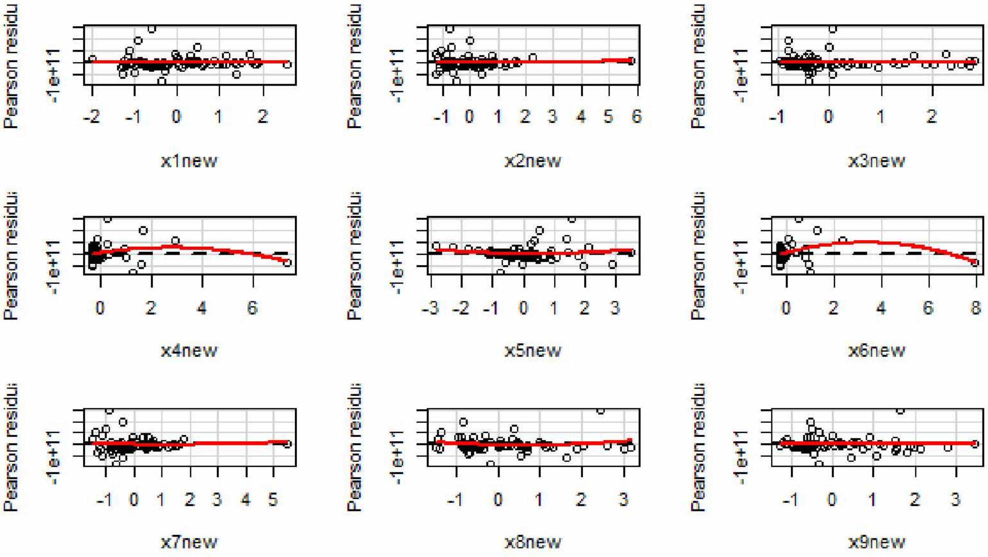
Appendix A

A.1 ncvTest results (Test for curvature)

```
            Test stat Pr(>|t|)
x1new        -0.203    0.839
x2new         0.485    0.629
x3new         0.503    0.617
x4new        -7.241    0.000
x5new         1.131    0.262
x6new        -8.126    0.000
x7new         0.598    0.552
x8new         1.307    0.196
x9new        -0.014    0.989
x10new       -0.267    0.790
Tukey test   -7.892    0.000
```

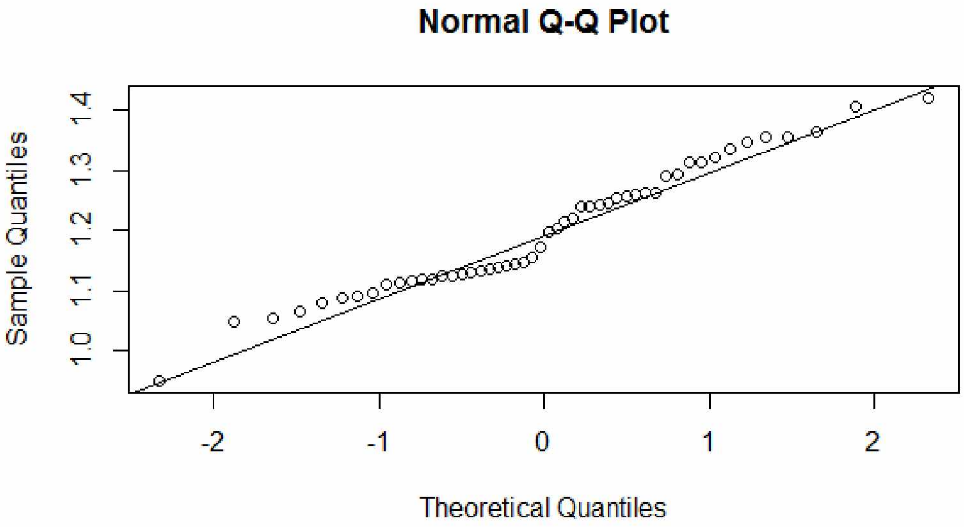
According to the output of ncvTest, we determine that there is some curvature in the relationship between the response and variables X_4 and X_6, because the p-values are less than 0.05. This partly motivates a log-transformation for variables X_4 and X_6.

A.2 Residual Plots (Curvature plots)



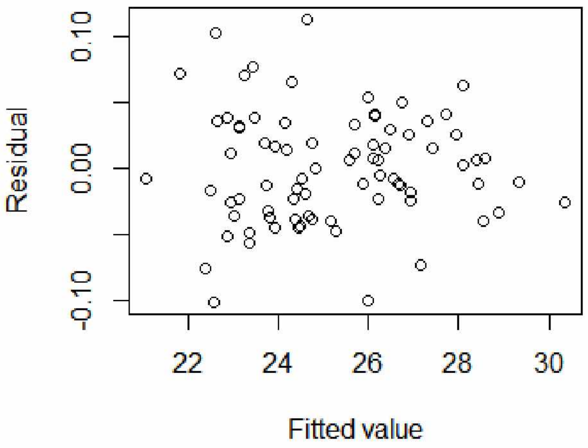
These are marginal residual plots for simple linear regression fit.

A.3 Normal Q-Q plot for g-prior variable selection

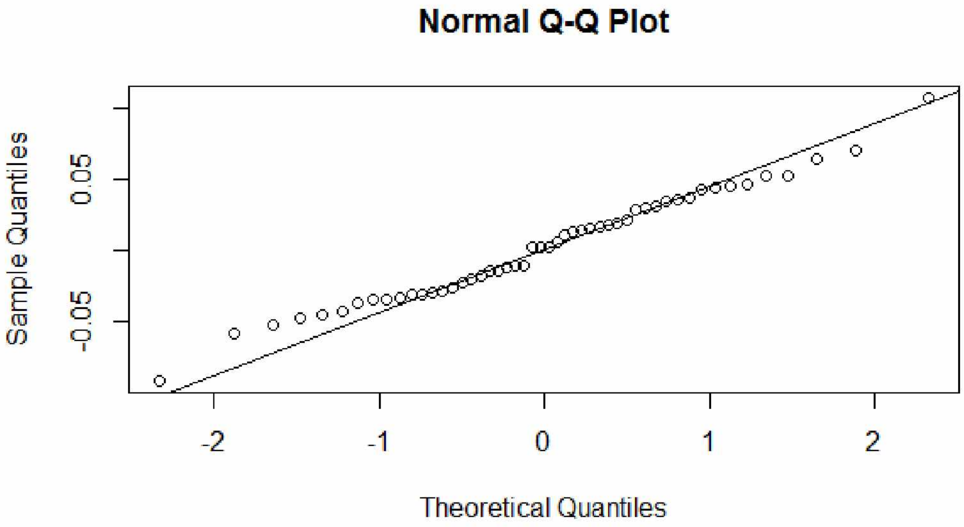


This plot indicates that the Bayesian residuals are normally distributed.

A.4 Bayesian Residuals vs. Fitted values plot using data for g-prior variable selection

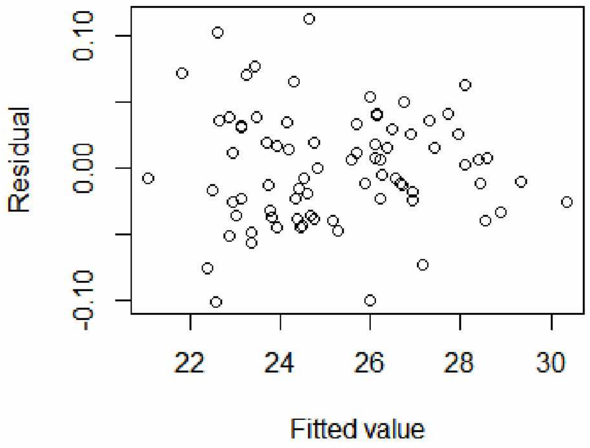


A.5 Normal probability plot of Bayesian residuals for Local empirical Bayes



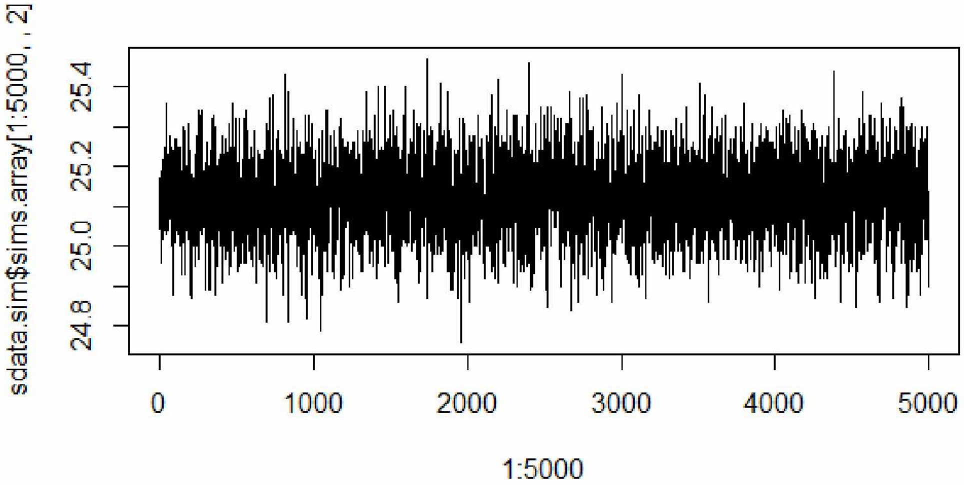
The linearity of this plot indicates that the Bayesian residuals are normally distributed.

A.6 Bayesian Residuals vs. Fitted values plot for Local empirical Bayes



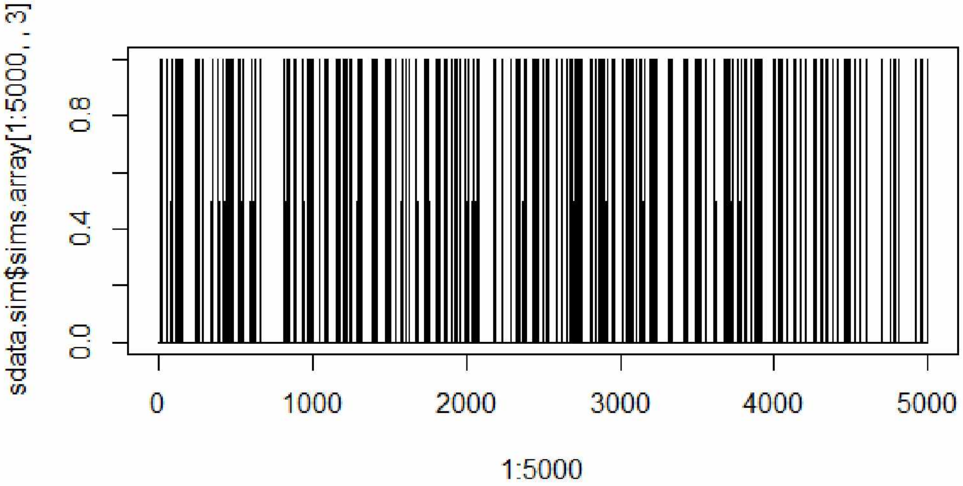
This is Bayesian Residuals vs. Fitted values plot for local empirical Bayes.

A.7 Traceplot of β_0 for variable selection by indicator function



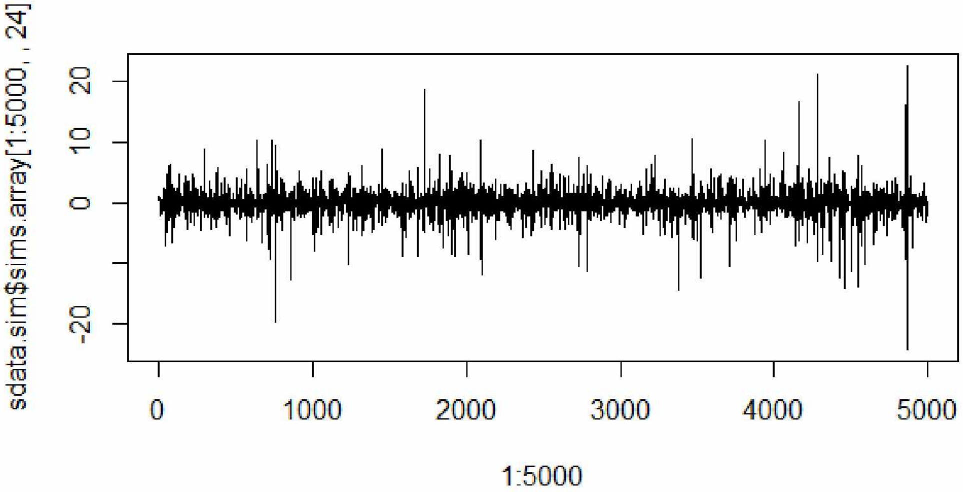
We did 10,000 iterations and used 1 chain to get this plot. The estimated mean of β_0 is about 25.1.

A.8 Traceplot of I_1 for variable selection by indicator function



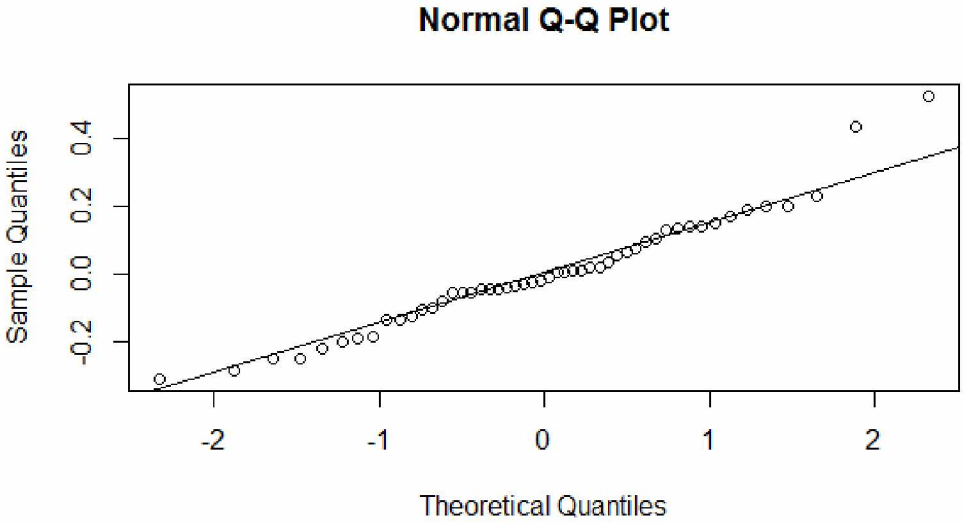
We did 10,000 iterations and used 1 chain to get this plot.

A.9 Traceplot of μ_1 for variable selection by indicator function



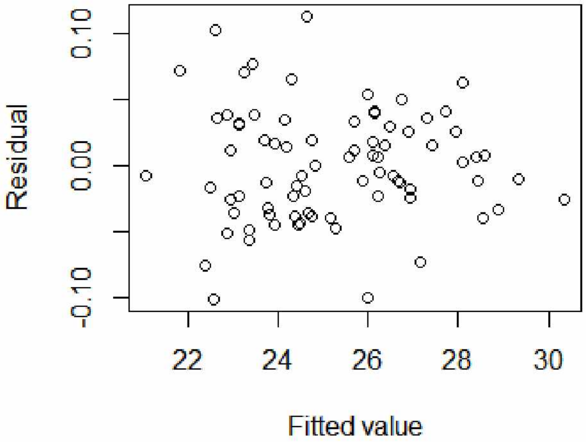
We did 10,000 iterations and used 1 chain to get this plot. The extreme jumps in magnitude of μ_1 are caused by the method's difficulty in estimating μ_1 on iterations where the indicator I_1 is at 0.

A.10 Normal probability plot of Bayesian residuals for variable selection by indicator function



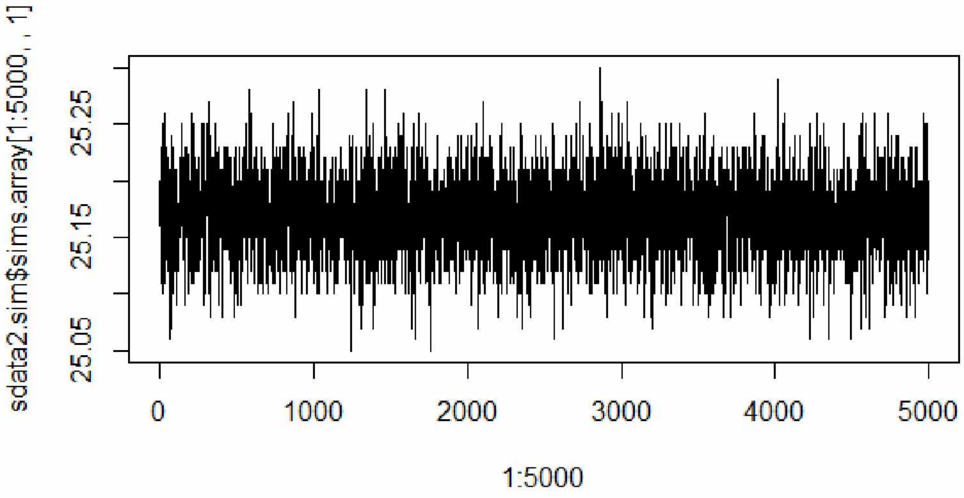
This plot indicates that the Bayesian residuals are essentially normally distributed.

A.11 Bayesian Residuals vs. Fitted values plot using data for Variable selection by indicator function



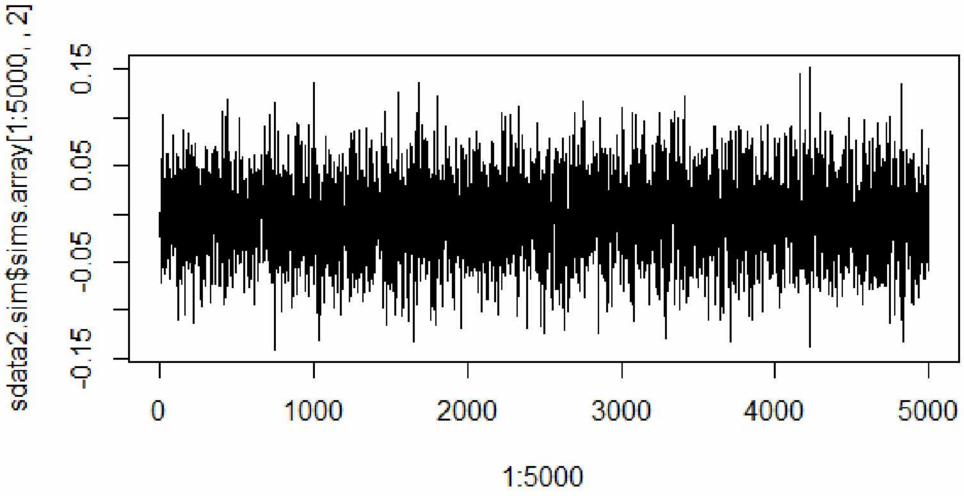
This is Residuals vs. Fitted values plot for Variable selection by indicator function.

A.12 Traceplot of β_0 for Hyper-g prior variable selection



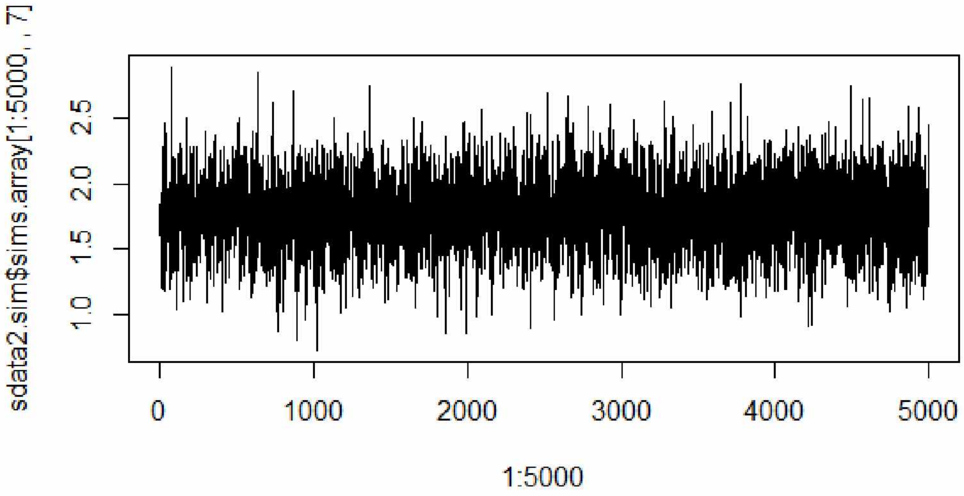
We did 10,000 iterations and used 1 chain to get this plot.

A.13 Traceplot of β_1 for Hyper-g prior variable selection



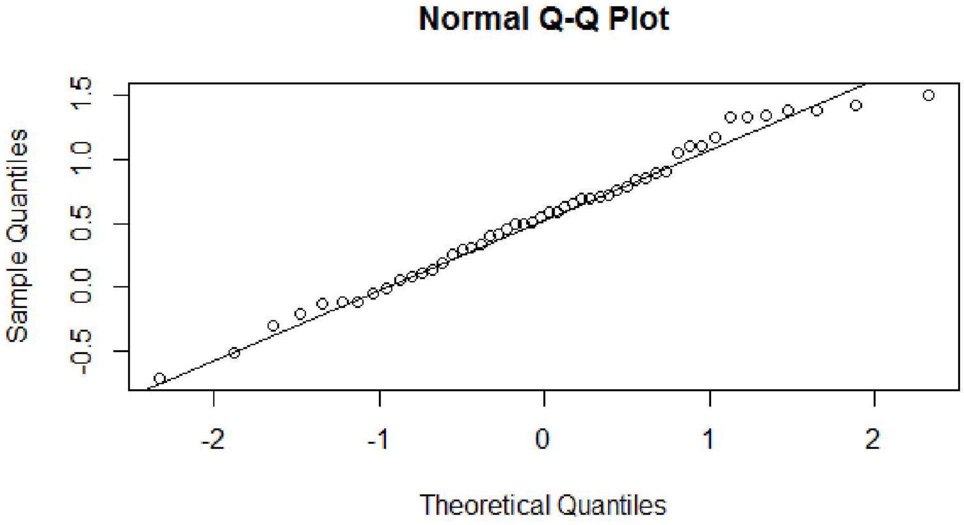
We did 10,000 iterations and used 1 chain to get this plot.

A.14 Traceplot of β_6 for Hyper-g prior variable selection



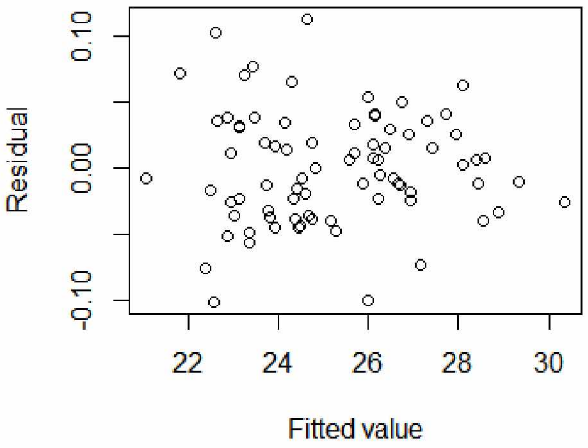
We did 10,000 iterations and used 1 chain to get this plot.

A.15 Normal probability plot of Bayesian residuals for hyper-g prior variable selection



The plot indicates that the Bayesian residuals are normally distributed.

A.16 Bayesian Residuals vs. Fitted values plot for hyper-g prior variable selection



This is Bayesian Residuals vs. Fitted values plot for hyper-g prior variable selection.

A.17 R-code for g-prior and local empirical Bayes variable selection

```
x=model.matrix(trainreg)
P=x%*%solve((t(x)%*%x))%*%t(x)
y=newtry
I=diag(50) # 50=size of training set
msg=0.5*t(y)%*%(I-(g/(g+1))*P)%*%y
umsg=msg/((50/2)-1)
Bmse=(g/(g+1))*as.numeric(umsg)*solve(t(x)%*%x)
mse=diag(Bmse)^0.5
ebeta=trainreg$coefficients
meanbeta=(g/(1+g))*trainreg$coefficients

meanb<-matrix(NA,11,10000)
for(i in 1:10000){

  sigam<-1/rgamma(1,25,rate=msg)
  var=(g/(1+g)*sigam*solve((t(x)%*%x)))

  mean=(g/(1+g)*ebeta)
  meanb[,i]<-rmvnorm(1,mean,var)
}
library(HDInterval) # calculate credible interval
hdi(meanb[1,])
hdi(meanb[2,])
......
hdi(meanb[11,])
```

A.18 R-code for indicator function variable selection

```
inits <- function(){
 list(y.precision=1,beta0=0,I1=0,I2=0,I3=0,I4=0,
 I5=0,I6=0,I7=0,I8=0,I9=0,I10=0,u1=0,u2=0,u3=0,
 u4=0,u5=0,u6=0,u7=0,u8=0,u9=0,u10=0,t0=1,
 t1=1, t2=1, t3=1,  t4=1, t5=1, t6=1, t7=1, t8=1, t9=1, t10=1)
}
```

```
sdata.sim <- bugs(txtdata, inits, model.file = "150.txt",
 parameters = c("y.precision","beta0","I1","I2","I3","I4","I5","I6","I7",
 "I8","I9","I10","t0","t1","t2","t3","t4","t5","t6",
 "t7","t8","t9","t10","u1","u2","u3","u4","u5","u6","u7","u8","u9","u10"),
 n.chains = 1, n.iter = 10000)
```

A.19 OpenBUGS code for indicator function variable selection

```
model{
  for (i in 1:50){


    y[i]~dnorm(n[i],y.precision)
    n[i]<-beta0+x1[i]*beta1+x2[i]*beta2+x3[i]*beta3+x4[i]*beta4+x5[i]*beta5
    +x6[i]*beta6+x7[i]*beta7+x8[i]*beta8+x9[i]*beta9+x10[i]*beta10


}
  beta0~dnorm(0,t0)
  t0~dgamma(1,10)
  y.precision~dgamma(1,10)
  I1~dbern(0.5)
  u1~dnorm(0,t1)
  t1~dgamma(1,1)
  beta1<-I1*u1

  I2~dbern(0.5)
  u2~dnorm(0,t2)
  t2~dgamma(1,1)
  beta2<-I2*u2
......
  I10~dbern(0.5)
  u10~dnorm(0,t10)
  t10~dgamma(1,1)
  beta10<-I10*u10
}
```

A.20 R-code for hyper-g prior variable selection

```
xtr=model.matrix(trainreg)
z=t(xtr)%*%xtr
muo=as.vector(rep(0,11))

y <- matrix(y,50,1)
zzz<-round(z[1:11,1:11],2)
muo=as.vector(rep(0,11))

inits <- function(){
  list(beta=c(0,0,0,0,0,0,0,0,0,0,0),y.precision=1,tau=1,lambda=0.5)
}
```

```
model{
  for(i in 1:50){
    for(j in 1:1){
      y[i,j]~dnorm(ea[i,j],y.precision)
      ea[i,j] <- beta[1]+beta[2]*x1[i]+beta[3]*x2[i]+
      beta[4]*x3[i]+beta[5]*x4[i]+beta[6]*x5[i]+beta[7]*x6[i]+
      beta[8]*x7[i]+beta[9]*x8[i]+beta[10]*x9[i]+beta[11]*x10[i]
}
}
  y.precision~dgamma(1,1)
  beta[1:11]~dmnorm(muo[],precision[,])
  tau~dgamma(1,1)
  lambda~dbeta(1,0.5)
  g<-lambda/(1-lambda)
  for(k in 1:11){
    for(l in 1:11){
      precision[k,l]<-(1/g)*tau*z[k,l]
}
}
}

sdata2.sim <- bugs(txt2data, inits,
model.file = "3-8 test.txt",
 parameters = c("beta","y.precision","tau","lambda"),
 n.chains = 1, n.iter = 5000)

as.integer(hdi(sdata2.sim)[,1][2]*hdi(sdata2.sim)[,1][1]>0)
*sdata2.sim$mean$beta[1]+as.integer(hdi(sdata2.sim)[,2][2]
*hdi(sdata2.sim)[,2][1]>0)
*sdata2.sim$mean$beta[2]*x1te
```