Bayesian Linear Regression Modeling to Predict GDP

Katherine Botz, Juan Luis Gómez Chanclón

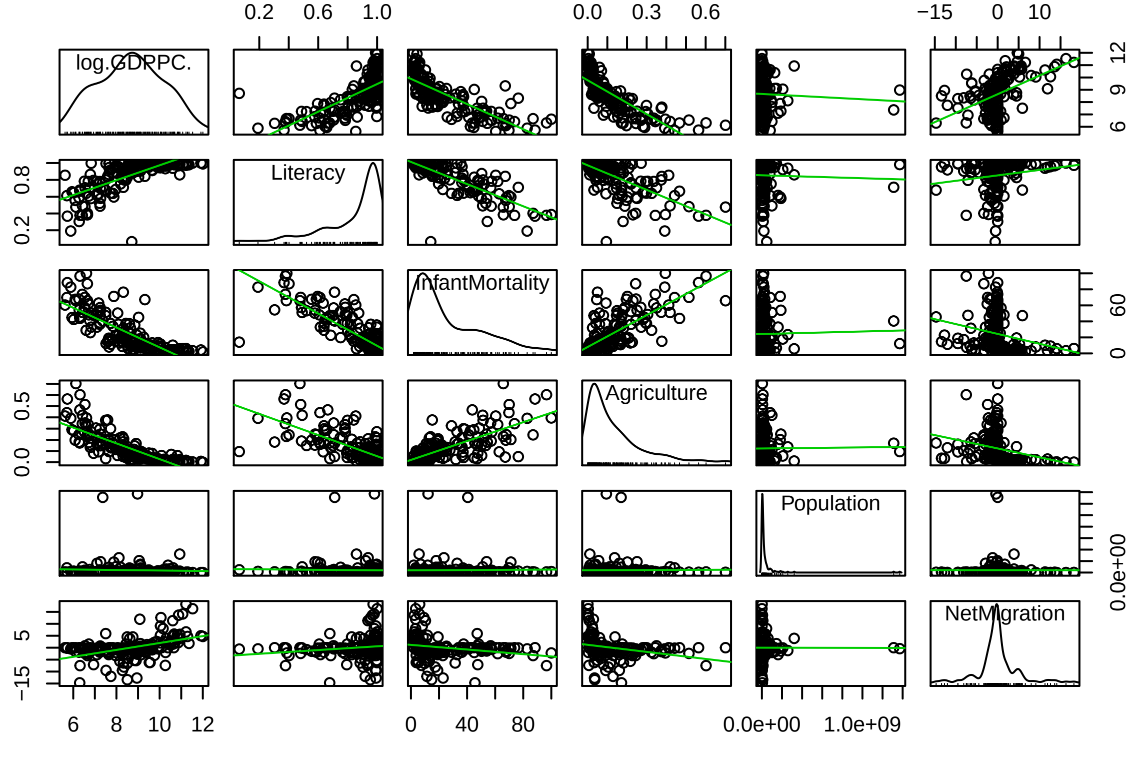
Introduction

The goal of this project is to find a Bayesian Linear Regression Model to predict the GDP for various Countries in the world. Our dataset consists of 170 instances, with information about the Literacy, Infant Mortality per 1000 births, Agriculture as a percentage of the GDP, Population of the Country, and Net Migration. We select the best model in the end by maximizing both the DIC and calculated Accuracy of the GDP prediction in the Test data.

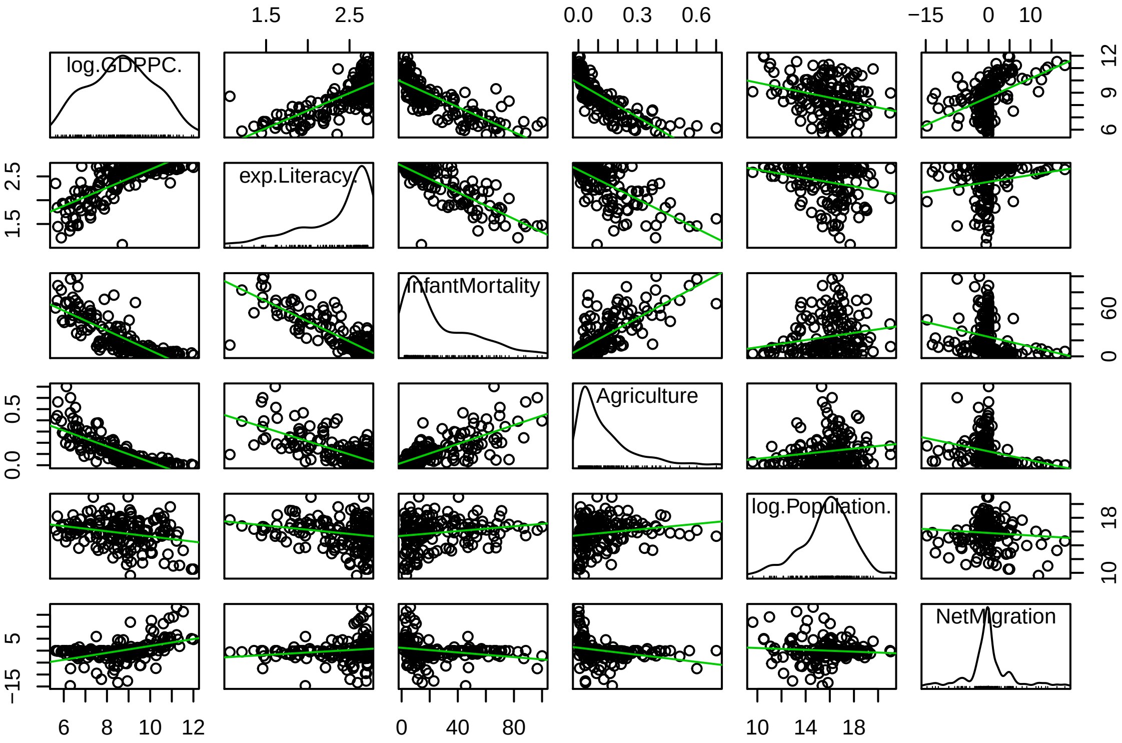
Analysis

First, we analyze the variables in the dataset to see if the relationship between the explanatory variables and the response variables appears to be linear. If not, we will perform a nonlinear transformation to exhibit a more linear relationship.

We will plot the logarithm of the GDP against each of the explanatory variables.



In the end, we find that taking the exponent of Literacy and the logarithm of Population shows a more linear relationship. We use these nonlinear transformations in building the models.



One of our measures of best fit for model selection will be accuracy of the model's predictions. To measure this, we split the dataset into Train and Test data, then train the model using the former and calculate the model accuracy using the latter.

We will first generate a model using the traditional Frequentist approach and to compare to our Bayesian models. Variable selection for the Frequentist approach will be facilitated by the `stepAIC` model selection function.

```
##
## Call:
## lm(formula = log(GDPPC) ~ InfantMortality + Agriculture + log(Population) +
##     NetMigration, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97688 -0.41228 -0.02805  0.39447  2.32226
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     11.278045   0.499773  22.566  < 2e-16 ***
## InfantMortality -0.026765   0.004405  -6.076 1.44e-08 ***
## Agriculture     -4.452608   0.771930  -5.768 6.19e-08 ***
## log(Population) -0.090762   0.031977  -2.838  0.00531 **
## NetMigration     0.073464   0.014813   4.959 2.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.7105 on 122 degrees of freedom
## Multiple R-squared:  0.7897, Adjusted R-squared:  0.7828 
## F-statistic: 114.5 on 4 and 122 DF,  p-value: < 2.2e-16
```

As can be seen in the Summary of the traditional model selected by `stepAIC`, the significant variables are Infant Mortality, Agriculture, Population, and Net Migration. We will compare this model to the Bayesian models later.

We implement a Markov Chain Monte Carlo (MCMC) algorithm to obtain samples from the posterior distribution of the model parameters. First we generate a model will all variables, using the nonlinear transformations mentioned before.

```
## 
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000 
## 
##  DIC: 279.9 
## 
##  R-structure:  ~units
## 
##       post.mean l-95% CI u-95% CI eff.samp
## units    0.5059   0.3695   0.6174     1000
## 
##  Location effects: log(GDPPC) ~ exp(Literacy) + InfantMortality + Agriculture + log(Population) + Ne
## 
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC    
## (Intercept)       9.81436  7.69540 11.68532    899.4 <0.001 ***
## exp(Literacy)     0.47361 -0.04554  1.03247   1000.0  0.096 .  
## InfantMortality  -0.02094 -0.03238 -0.01032    962.3 <0.001 ***
## Agriculture      -4.28257 -5.69434 -2.96756   1000.0 <0.001 ***
## log(Population)  -0.07990 -0.14634 -0.01749   1132.5  0.018 *  
## NetMigration      0.07603  0.04526  0.10449   1000.0 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear that Literacy is not significant in this model, so we remove it to train the next model:

```
## 
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000 
## 
##  DIC: 280.6086 
## 
##  R-structure:  ~units
## 
##       post.mean l-95% CI u-95% CI eff.samp
## units    0.5121   0.3957   0.6389     1000
## 
##  Location effects: log(GDPPC) ~ InfantMortality + Agriculture + log(Population) + NetMigration 
## 
##                 post.mean l-95% CI u-95% CI eff.samp  pMCMC    
## (Intercept)      11.27692 10.33943 12.29041     1000 <0.001 ***
## InfantMortality  -0.02679 -0.03513 -0.01797     1000 <0.001 ***
## Agriculture      -4.42955 -5.91774 -2.97914     1000 <0.001 ***
## log(Population)  -0.09089 -0.15384 -0.02962     1000  0.004 **
## NetMigration      0.07385  0.04221  0.09942     1242 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This seems to be a good model, with all significant variables. Let's try to replicate the selected Frequentist model for comparison:

```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: 280.7121
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units    0.5134   0.3951   0.6315     1000
##
##  Location effects: log(GDPPC) ~ log(Population) + InfantMortality + Agriculture + NetMigration
##
##                  post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)      11.26760 10.35318 12.25806    909.4 <0.001 ***
## log(Population)  -0.08977 -0.15747 -0.03368    900.9  0.006 **
## InfantMortality  -0.02706 -0.03526 -0.01796   1000.0 <0.001 ***
## Agriculture      -4.42996 -6.01772 -3.02969   1186.6 <0.001 ***
## NetMigration      0.07316  0.04102  0.09975   1140.0 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Also seems to be a good model, with a similar DIC as the previous one. The following model is one more for comparison; it also has all significant variables, but with a slightly higher DIC.

```
##
##  Iterations = 3001:12991
##  Thinning interval  = 10
##  Sample size  = 1000
##
##  DIC: 286.6805
##
##  R-structure:  ~units
##
##       post.mean l-95% CI u-95% CI eff.samp
## units    0.5458   0.4202   0.6838     1197
##
##  Location effects: log(GDPPC) ~ InfantMortality + Agriculture + NetMigration
##
##                  post.mean l-95% CI u-95% CI eff.samp  pMCMC
## (Intercept)       9.88236  9.68338 10.06996     1000 <0.001 ***
## InfantMortality  -0.02753 -0.03638 -0.01863     1000 <0.001 ***
## Agriculture      -4.60340 -6.22120 -3.07783     1000 <0.001 ***
## NetMigration      0.07273  0.03932  0.09968     1000 <0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

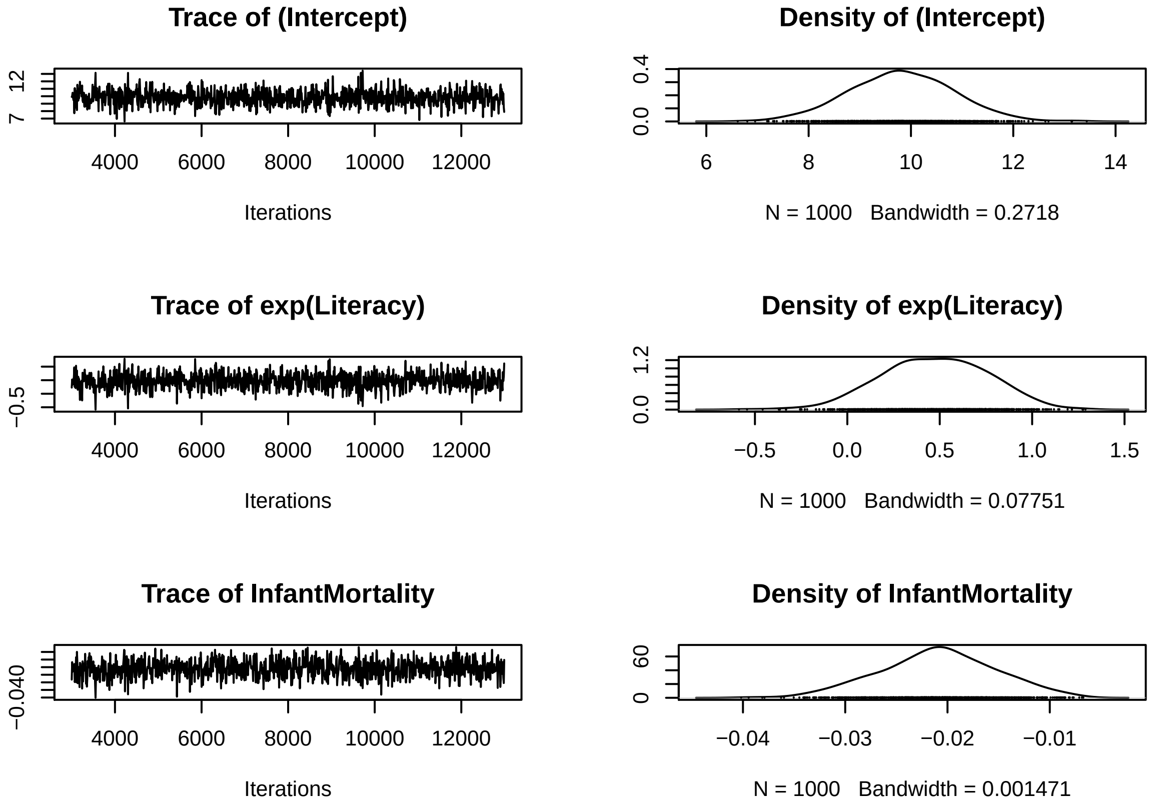
Now that we have four Bayesian MCMC models and one traditional Frequentist model, we calculate the Correlation and Mean Absolute Percentage Error between the predicted values and actual values of the Test dataset as a measure of Accuracy for each model. We calculate the Accuracy as `1-MAPE`, therefore we are trying to maximize this figure. As another measure of accuracy, we calculate the Root-Mean-Square Error with the intention of minimizing this figure. Finally, we show the Deviance Information Criterion (and BIC for the Frequentist model) as another measure of fitness for each model.

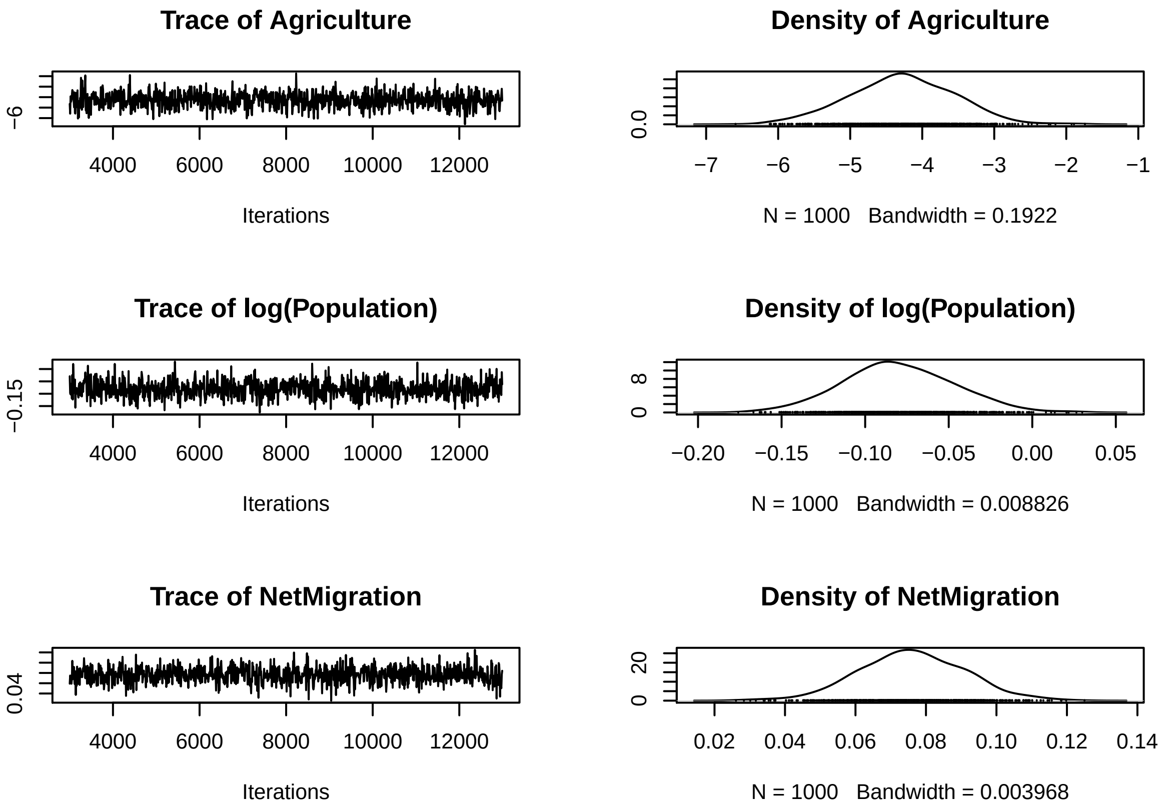
Table 1: Top 5 Models for Fitness and Accuracy

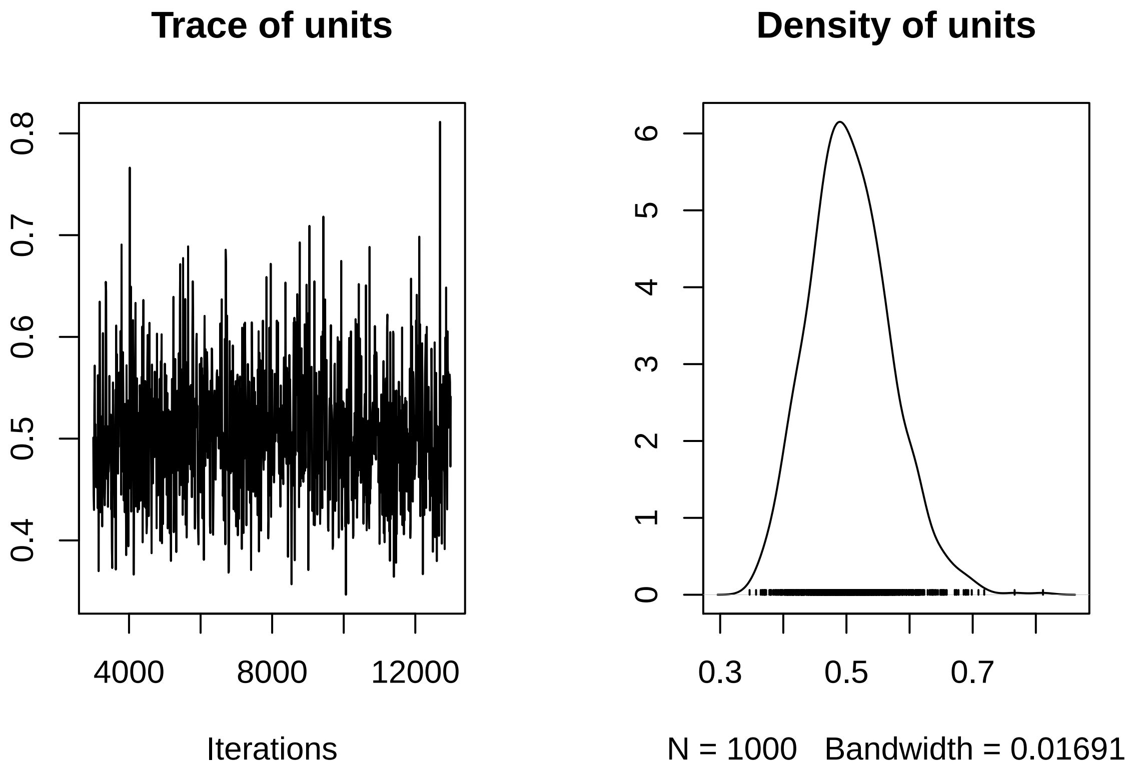
Model	Description	DIC or BIC	Correlation	Accuracy	RMSE
modBIC	Frequentist Model Selected by StepAIC	297.57	86.32%	43.54%	12877
bayes1	Bayes Model with all Variables	279.81	87.01%	47.05%	12634
bayes2	Bayes Model with all Variables except Literacy	280.65	86.33%	43.55%	12872
bayes3	Bayes Model with Same Variables as modBIC	280.61	86.40%	43.56%	13384
bayes4	Bayes Model with Literacy, Infant Mortality, and Agriculture	286.70	90.93%	43.10%	13805

Conclusions

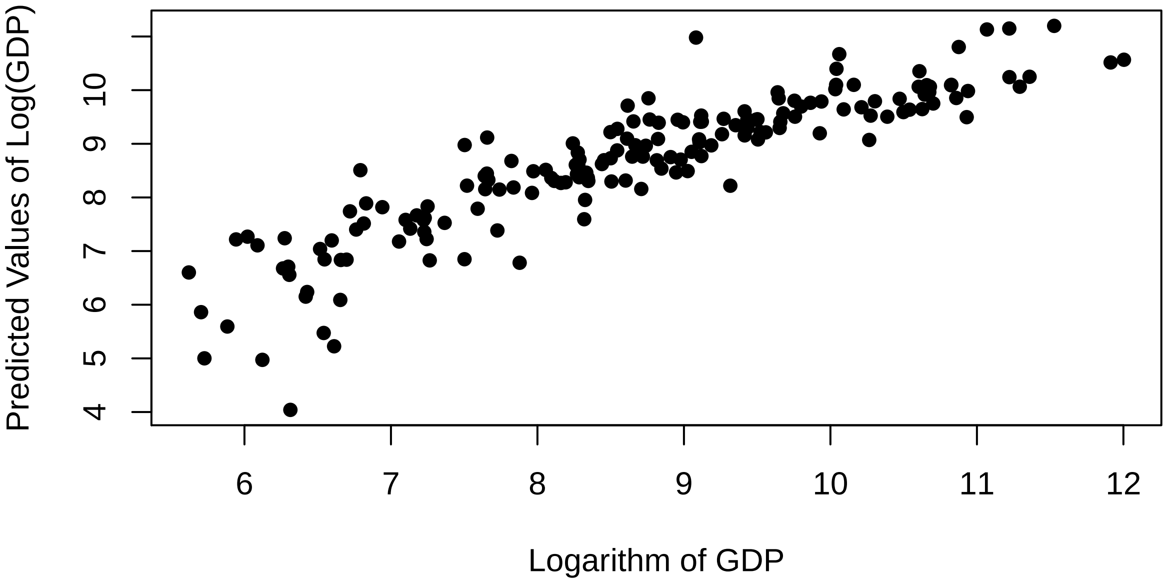
Comparing each measure of Accuracy and DIC, we select the first Bayes Model with all Variables as the best model. We check the Trace Plots for this model and confirm that the posterior distribution is in a stationary state.




Trace of Agriculture
−6
4000
6000
8000
10000
12000
Iterations
Density of Agriculture
0.0
−7
−6
−5
−4
−3
−2
−1
N = 1000 Bandwidth = 0.1922
Trace of log(Population)
−0.15
4000
6000
8000
10000
12000
Iterations
Density of log(Population)
0
8
−0.20
−0.15
−0.10
−0.05
0.00
0.05
N = 1000 Bandwidth = 0.008826
Trace of NetMigration
0.04
4000
6000
8000
10000
12000
Iterations
Density of NetMigration
0
20
0.02
0.04
0.06
0.08
0.10
0.12
0.14
N = 1000 Bandwidth = 0.003968



We also graph the actual GDP values to the predicted values as follows.



In the end, the best model is still not very accurate (47%), but we attribute this issue to the small dataset, with 170 instances. Also, the Frequentist model does not perform better than the Bayesian models, so it is not an issue of the Posterior Distribution or the choice of prior.