Bayesian Forecasting of C02 Emissions

Mallory Wang, Noah Kochanski

4/13/2022

Introduction

Global warming is the change in the Earth's weather patterns over a extend period of time. Although Earth's climate changing has been a phenomenon that has occurred many times in the past, scientists agree that human activities over the past 100 years are accelerating the speed at which global warming is happening. Carbon dioxide (CO2) is one variable increasing the speed of global warming through the greenhouse gas. Although CO2 is released through natural processes such as breathing and decomposition, humans have been adding C02 to the atmosphere at unprecedented rates through the burning of fossil fuels. In the United States, there has been an increasing demand to understand, predict, and reduce the carbon footprint as a nation. We look to explore the use of both classical and Bayes linear regression and compare the results when forecasting CO2 emissions in the United States.

Data Introduction

Our data set of interest comes from the World Bank and includes the following variables tracked from 1960 to present day:

- Year: annual
- C02 Emissions: sourced from Carbon Dioxide Information Analysis Center, Environmental Sciences Division, Oak Ridge National Laboratory, Tennessee, United States measured in kt
- Population: total population sourced from US Census data
- Gross Domestic Product: in current US dollars
- Gross Domestic Income: derived as the sum of GDP and the terms of trade adjustment
- Net Primary Income: in current US dollars
- Population in Urban Agglomeration: population in urban agglomerations of more than one million is the percentage of a country's population living in metropolitan areas that in 2018 had a population of more than one million people
- Energy Use: kg of oil equivalent per capita
- Net Energy Import: % of total energy used
- Electric Power Consumption: kWh per capita

In Figure 1, CO2 emissions in the US is shown to be increasing with time. The dips in the graph typically correspond to times of economic turmoil in the United States. Much of the auxiliary variables included in Figure 1 also depicts increasing functions over time (except GDP growth). Comparing auxiliary variables to one another, we might determine some correlations between them.

Figure 2 shows a correlation matrix of all variables included in the data. We see many variables have extremely high correlation, as high as 1. As a result, we will want to pick a subset of these variables to avoid singularity when performing matrix calculations.

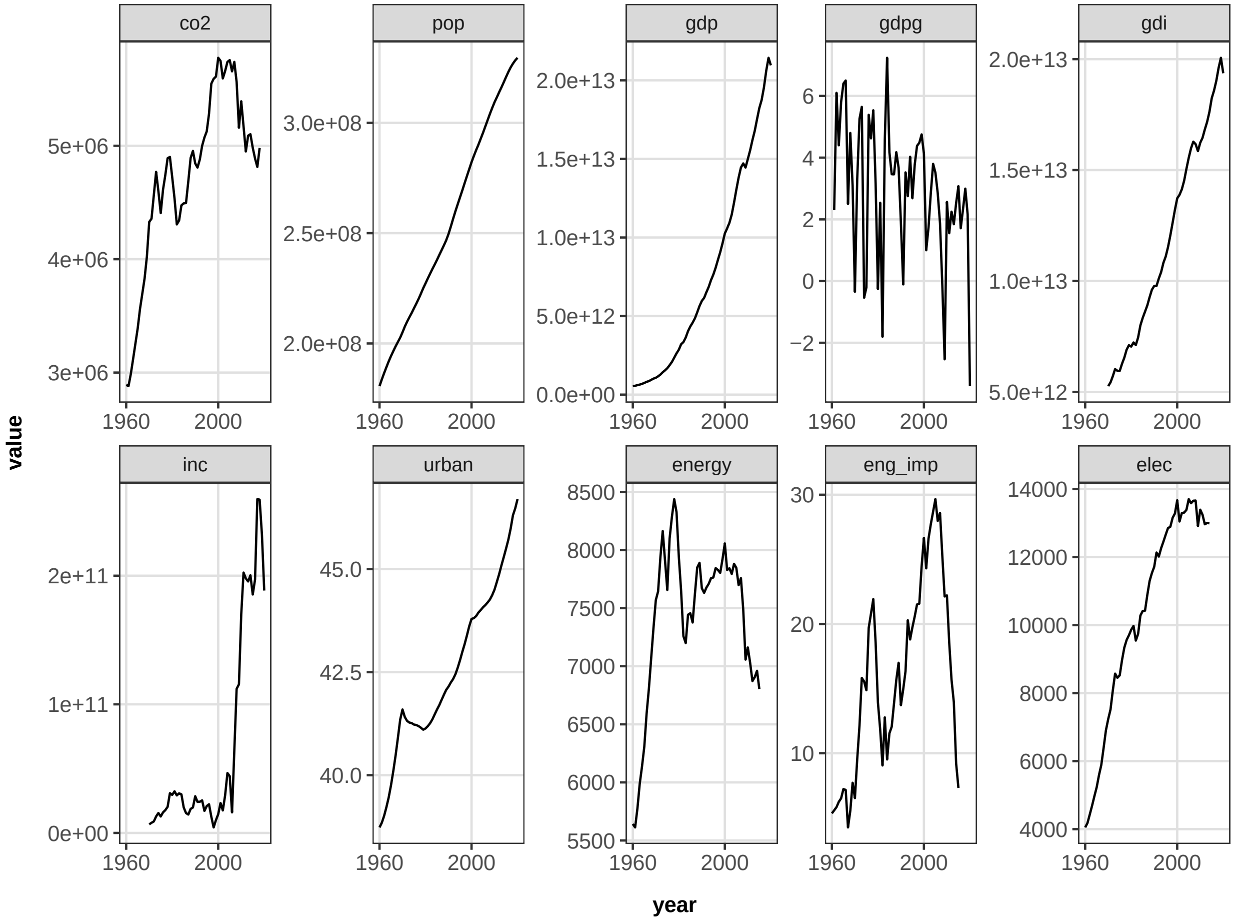

Figure 1: Time series of CO2 and Auxiliary Variables

Figure 2 shows a correlation matrix of all variables included in the data. We see many variables have extremely high correlation, as high as 1. As a result, we will want to pick a subset of these variables to avoid singularity when performing matrix calculations.

In our analysis, we removed population, GDP, GDIncome, and net income due to their high correlation. Additionally, World Bank does not have data for GDI and Income from 1960 to 1970 nor data for Energy Use and Energy Import after 2015. Therefore, the analysis in this report utilizes data from 1970 until 2015.

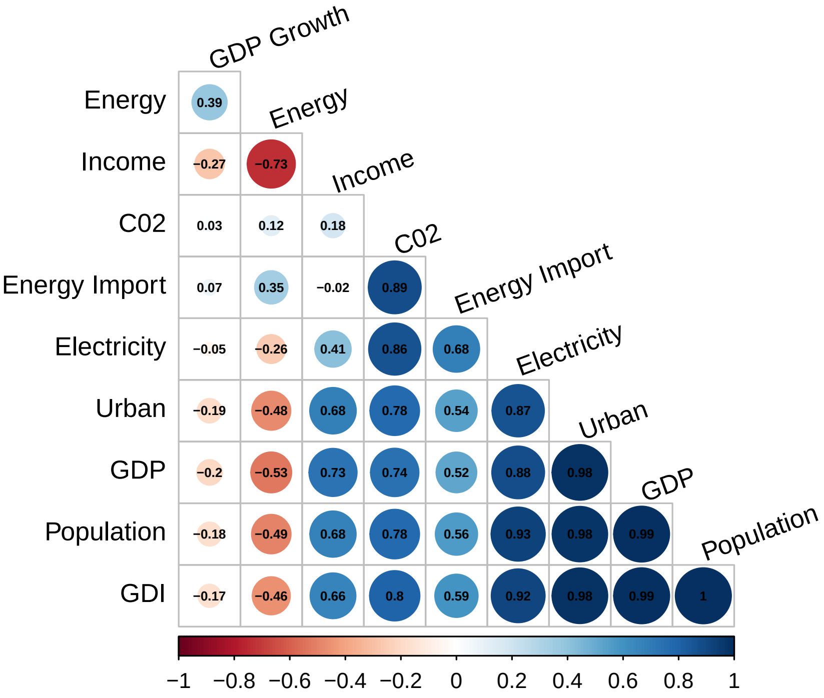

Figure 2: Correlation matrix of all variables

Methods

In this report, we will first demonstrate the feasibility of a Bayesian linear regression by comparing it to a classical linear regression model, modeling CO2 against only `year` predictor variable. Next, build a Bayesian linear regression model with full auxiliary variables and consider time series related attributes in our data, such as lagging auxiliary variables. A comparison of auxiliary parameters with and without lagged components will be considered. Finally, we conclude with varying windows of posterior parameters estimates with our final Bayesian linear regression model. To reiterate, we use World Bank data depicted in the previous section from 1970 to 2015. Additional segmentation of data will be described in their respective sections.

Classical Linear Regression

The first method that we propose to use is by linear regression. Since we are attempting to forecast future CO2 emissions, we use the following model,

$$Y_t = \beta' \begin{bmatrix} X_{t-5} \\ Y_{t-5} \end{bmatrix} + \epsilon_t$$

where Y_t is the observation at time, t, X_{t-5}, Y_{t-5} is the data at time, $t-5$ and $\epsilon_t \sim N(0, \sigma^2)$. To determine the most important variables in predicting CO2 emissions, we perform backward elimination and only keep the 5 most important predictors.

```r
library(leaps)

# lagging the data 5 years
lag5 <- us_df %>% mutate_all(lag, n = 5)
colnames(lag5) <- paste(colnames(us_df), "_5", sep="")
lagged_data <- cbind(us_df, lag5) %>% dplyr::select(-c(3:12)) %>% na.omit()

# fitting the full model
full_model <- lm(co2~. ,data = lagged_data)

# backward elimination
models <- regsubsets(co2~., data = lagged_data, nvmax = 5, method = "backward")
summary(models)
```

```
## Subset selection object
## Call: regsubsets.formula(co2 ~ ., data = lagged_data, nvmax = 5, method = "backward")
## 11 Variables  (and intercept)
##           Forced in Forced out
## year          FALSE      FALSE
## co2_5         FALSE      FALSE
## pop_5         FALSE      FALSE
## gdp_5         FALSE      FALSE
## gdpg_5        FALSE      FALSE
## gdi_5         FALSE      FALSE
## inc_5         FALSE      FALSE
## urban_5       FALSE      FALSE
## energy_5      FALSE      FALSE
## eng_imp_5     FALSE      FALSE
## elec_5        FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: backward
##          year co2_5 pop_5 gdp_5 gdpg_5 gdi_5 inc_5 urban_5 energy_5 eng_imp_5
## 1  ( 1 ) " "  " "   " "   " "   " "    " "   " "   " "     " "      " "      
## 2  ( 1 ) " "  " "   " "   " "   " "    "*"   " "   " "     " "      " "      
## 3  ( 1 ) " "  " "   " "   " "   " "    "*"   " "   "*"     " "      " "      
## 4  ( 1 ) " "  " "   " "   " "   " "    "*"   "*"   "*"     " "      " "      
## 5  ( 1 ) " "  " "   "*"   " "   " "    "*"   "*"   "*"     " "      " "      
##          elec_5
## 1  ( 1 ) "*"   
## 2  ( 1 ) "*"   
## 3  ( 1 ) "*"   
## 4  ( 1 ) "*"   
## 5  ( 1 ) "*"   
```

As seen by the output above, the most important lagged variables which I will include for the following analysis are population, GDI, NPI, population in urban areas, and electric power consumption.

Bayes Linear Regression

In this section, we first seek to validate our choice of method by comparing the results of the Bayesian linear regression with the results of the classic linear regression. Next, we consider whether to include lagged variables by comparing their respective Bayes factors in model selection. Finally, we discuss the resulting model and the posterior distribution created.

First, we start with a basic model where CO2 is explained only by Year.

$$\begin{aligned} y_i &= \beta + \beta_{year} x_i + \epsilon_i \\ &= \beta^T x_i + \epsilon_i \end{aligned}$$

where y_i is CO2 and ϵ_i are independently and identically distributed normal with mean zero and constant variance (second line is generalized form for more than one predictor). Under these assumptions, we have the generalized conditional

$$Y_i | x_i, \beta, \sigma^2 \sim N(\beta^T x_i + \epsilon_i, \sigma^2)$$

(generalized form will follow from a multivariate normal with mean $X\beta$ and $\sigma^2 I$) and the likelihood of $Y_1, ..., Y_n$,

$$p(y_1, ..., y_n | x_i, \beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(y_i - \beta^T x_i)^2}{2\sigma^2} \right)$$

which is maximized when the sum of squared residuals is minimized, which can be written as

$$SSR(\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$$

Now we consider the semiconjugated prior of β,

$$p(y|X, \beta, \sigma^2) \propto \exp - \frac{1}{2\sigma^2} SSR(\beta)$$

and we have the following relationship

$$p(\gamma | y, X, \beta) \sim \text{inverse-gamma}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + SSR(\beta)}{2} \right)$$

These specifications (and initial values of 1 and 0.5 for ν and σ^2) and were followed to implement a Gibb's sampling to obtain posterior distribution of size 5000. To evaluate these results, we split our data into train and test sets, where 1970 to 2005 were used for train and 2006 to 2015 were used for test. Prediction errors from OLS were compared with the Bayes error (both calculated as sum of squares of test CO2 and CO2 from each method).

Figure 3 shows the distribution of the difference in prediction error between OLS and Bayesian methods. We can see that Bayes consistently has lower error than OLS, at least when Year is used to predict CO2. With these results, we can add the remaining auxiliary variables; we would use the generalized form of the conditional distribution in above derivations.

As noted from Figure 1, our data are time series and may rely on time series analysis. In this section, we will evaluate whether lagged variables should be used in the Bayesian linear regression. On a cursory level, we considered whether our outcome variable has some seasonality. Without a full time series analysis, we tried to CO2 data using the classical seasonal decomposition by moving averages and found no seasonality was determined. Next, to replicate some lagged effect, we generated all auxiliary variables, lagged by 5 years and analyzed correlations like before and dropped (in addition to those we already dropped) 5-years lagged population, GDP, GDI, and Income.

Figure 4 shows the distribution of the difference in prediction error between OLS and Bayesian methods with lagged variables. Though the distribution with lagged variables still demonstrates improvement over classic linear regression, its center is much higher than the distribution without lagged variables.

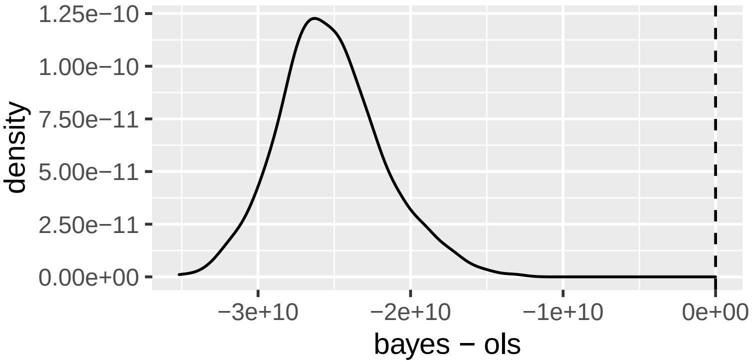

Figure 3: Distribution of prediction error difference between OLS and Bayesian methods

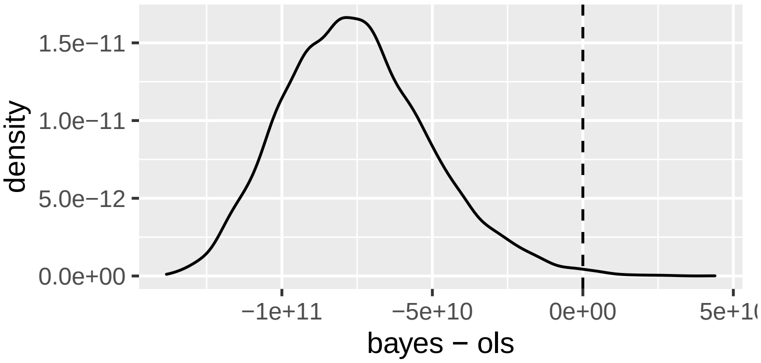

Figure 4: Distribution of prediction error difference between OLS and Bayesian methods with lagged variables

We also want to compare the model with and without lagged variable using Bayesian model comparison. Under the believe that many regression coefficients are equal to zero, generate a matrix Z such that $\beta_j = z_j \times b_j$ where $z_j \in \{0, 1\}$ and our equation

$$y_i = z_1 b_1 x_1 + ... + z_p b_p x_p + \epsilon_i$$

with z_j indicating which variables has non-zero coefficients. We obtain posterior distribution for z by using the g-prior to evaluate $p(y|X, z)$ for each model z such that we obtain

$$p(z|y, X) = \frac{p(z)p(y|X, z)}{\sum_z p(z)p(y|X, z)}$$

Using the g-prior distribution for β, we can compute the marginal probability following

$$\{\beta_z|X_z, \sigma^2\} \sim \text{multivariate noraml}(0, g\sigma^2[X_z^T X_z]^{-1})$$

And we can show the conditional density of (y, γ) given (X, z)

$$p(y|X, z, \gamma) \times p(\gamma) = (2\pi)^{\frac{-n}{2}}(1+g)^{\frac{-p_z}{2}} \times \left[\gamma^{\frac{n}{2}} e^{\frac{-\gamma SSR_g^z}{2}}\right] \times \left(\frac{\nu_0 \sigma_0^2}{2}\right)^{\frac{\nu_0}{2}} \Gamma\left(\frac{\nu_0}{2}\right)^{-1} \left[\gamma^{\frac{\nu_0}{2-1}} e^{\frac{-\gamma\nu_0\sigma_0^2}{2}}\right]$$

where

$$SSR_g^z = y^T \left(I = \frac{g}{g+1} X_z (X_z^T X_z)^{-1} X_z\right) y$$

Here, we set $g = n$ to use the unit information prior for $p(\sigma^2)$ in modeling z. We follow a similar Gibbs sampling scheme to generate $\{z^{(s+1), \sigma^{(s+1)}, \beta^{(s+1)}}\}$ from $z^{(s)}$ by 1) create z matrix 2) in random order sample from $p(z_j|z_{-j}, y, X)$ 3) update $z^{(s+1)}$ 3) sample $\sigma^{2(s+1)}$ from $p(\sigma^2|z^{(s+1)}, y, X)$ and 5) sample $\beta^{(s+1)}$ from $p(\beta|z^{(s+1)}, \sigma^{2(s+1)}, y, X)$.

```
##  [1] 0.2606 0.1473 0.3201 0.2597 0.3384 0.3781 0.1420 0.3088 0.2122 0.1652
## [11] 0.2726
```

```
##                    2.5%         50%        97.5%
## year          556.82545   1451.7069    2299.5480
## gdpg      -426484.99267 -76792.5917  300783.8296
## urban       41569.06335  60323.2619   78647.2698
## energy        105.62945    265.3547     423.6367
## eng_imp     46274.24593  90226.3884  133619.3770
## elec           85.20021    169.8233     255.3297
## gdpg_5      29613.26533 601718.5995 1185320.5202
## urban_5      1335.81383  32148.6015   62939.6302
## energy_5      -49.70218    124.6208     304.6901
## eng_imp_5  -64470.93081  21493.5155  111115.3155
## elec_5        -10.06459    108.4436     227.4202
```

Figure 5 shows intervals created from samples from the $beta^{(s+1)}$ for each predictor grouped by whether they were significant. We can see that most of the 5-years lagged variables are not significant, but 5-year lagged GDP growth is significant with very large interval. This is compared to the same procedure for Gibbs sampling but without the lagged variables.

Figure 6 shows the intervals created from $beta^{(s+1)}$ for each predictor. Unlike Figure 5, all variables are significant. If we return to the previously described test set (data from 2006 to 2015), we can look at how the predicted and observed values compares in Figure 7. The predcited values follow the trend of observed relatively well but the scale is a little off.

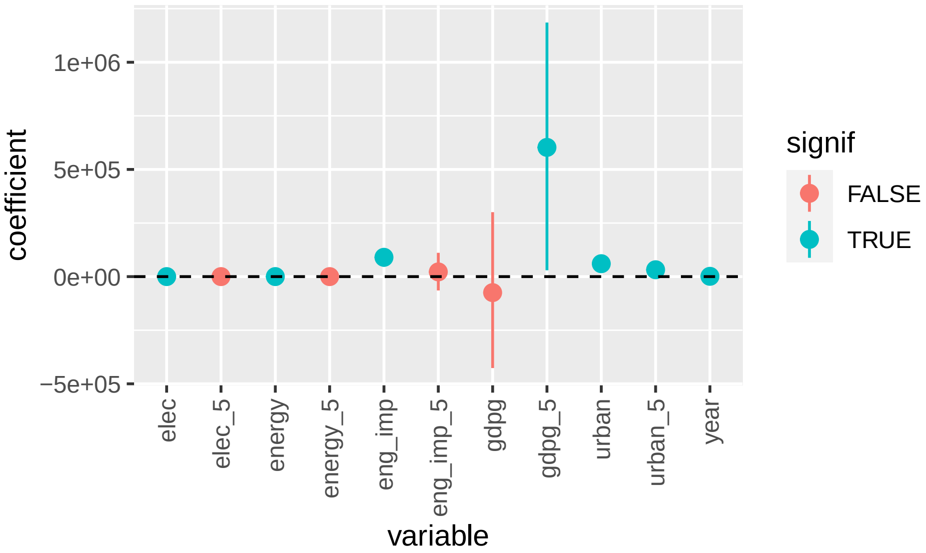

Figure 5: Intervals created from beta(s+1) with lagged variables

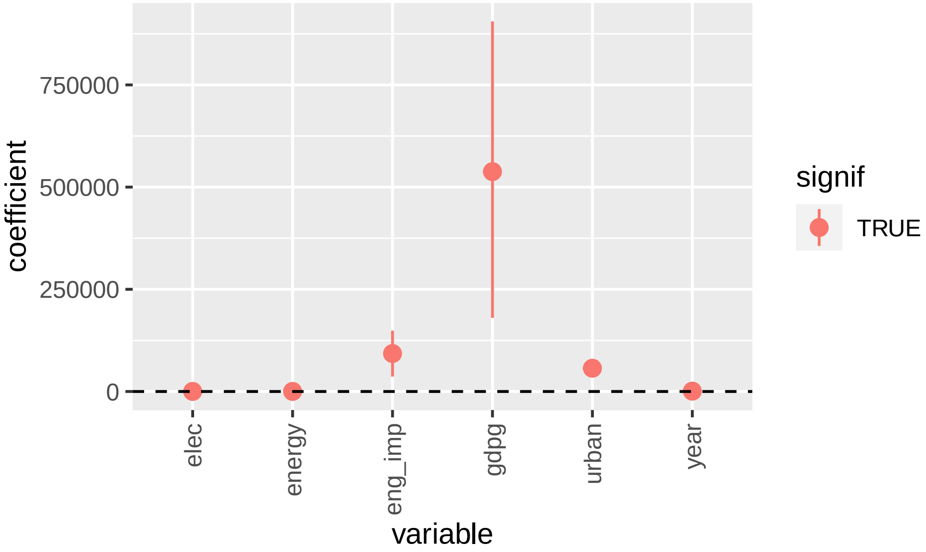

Figure 6: Intervals created from beta(s+1)

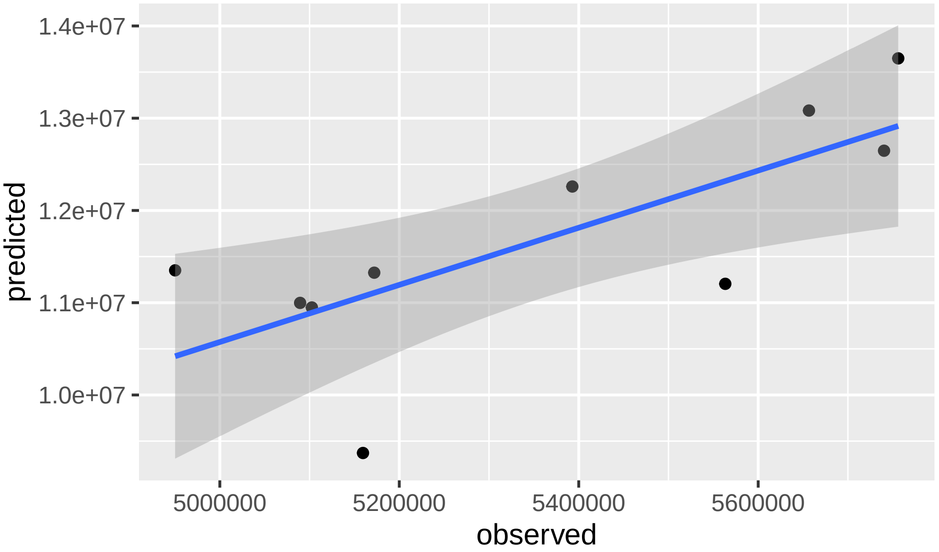

Figure 7: Predicted values vs. observed values in test set

Conclusion

Our final model has the following posterior distribution for each parameter and can be used to predict future values of CO2. From our analysis, we can conclude that the Bayesian approach improves upon the classic linear regression in this application.

```
##                  2.5%         50%       97.5%
## year       420.17340    889.8518   1354.5499
## gdpg    180024.25835 537504.5407 905383.9907
## urban    37569.66770  56956.2070  76117.9716
## energy      96.59824    214.7815    326.8704
## eng_imp  36447.28927  93286.1174 148900.6122
## elec       152.85931    220.7423    289.5448
```

That said, one major limitation of this analysis is its ignorance of time series methodologies and nuances. For example, our linear regression model used only a 5-year lagged variables as predictors. Through the use of moving average (MA) or ARIMA models, we would expect the frequentist approach to improve. Though we did consider basic detection of seasonality and lagged variables, additional consideration of the time series data is probably necessary to make any useful conclusions and to implement in conjunction with a Bayesian approach. Given additional time, more research into time series and Bayesian applications, such as hierarchical regression where some time element could create within group effects, could have contributed significantly to our study, not to mention interesting to learn.

References

Hoff, P. D. (2009). A first course in Bayesian statistical methods. New York: Springer.

Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B.. Bayesian Data Analysis. 2nd ed. : Chapman and Hall/CRC, 2004.