Predicting Long-Term Economic Growth: A Regression Problem

DSE1101 2024/2025 Semester 1 Final Project

Name: Yanis Bin Enche' Abdol Rahman

Data set: Economic Growth

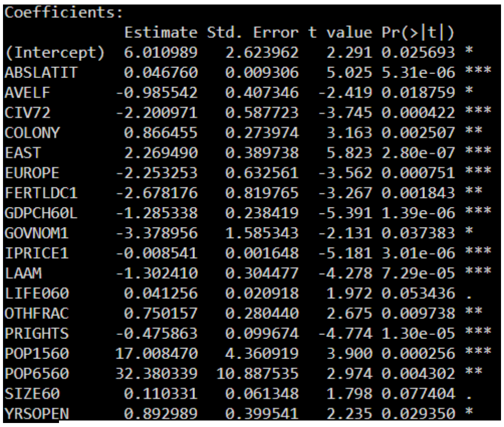
Predicting long-term economic growth has always been a central problem in macroeconomics. In this report, we will be working with the BACE Data used by Sala-I-Martin et al. in their 2004 paper. The data set contains 139 observations, ergo 139 countries, and 71 variables for each country. This data will be used to predict long-term economic growth based on suitable explanatory variables. The response variable used as a measurement of long-term economic growth is the growth of GDP per capita at purchasing power parities between 1960 and 1996. R will be used throughout the project for data analysis and plot making.

Cleaning and Formatting the Data

Before passing a model on the data, it is necessary to clean the data and make any necessary adjustments. Some of the variables in the original data seem to have many missing values. For instance, public investment share (GGCFD3) has 28 missing values. If we were to filter countries that have no value for GGCFD3, this would cause us to lose out on sample size to train our model on. Therefore, I decided to remove variables which have more than 20 missing values, followed by removing countries which have missing values for any of the remaining variables. This leaves us with a final data set with 95 observations of 60 variables.

Linear Regression

Linear regression is a simple and straightforward regression method to create predictions based on a set of explanatory variables. We begin by using the 'leaps' package in R to conduct model selection to find the suitable variables to include in our regression model. When setting 'nvmax' to the maximum number of variables we obtain a linear model, and after removing coefficients with a high p-value, we are left with a linear model with 18 coefficients. As expected, the adjusted R-squared of the model is high and equal to 0.858. Figure 1 shows the coefficients present in the model. When the model is passed on the test data, we get a root mean squared error, RMSE of 1.663. Figure 2 shows the graph of predicted values against actual values in the test data to visualize the performance of the linear model.



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.010989   2.623962   2.291 0.025693 *
ABSLATIT     0.046760   0.009306   5.025 5.31e-06 ***
AVELF       -0.985542   0.407346  -2.419 0.018759 *
CIV72       -2.200971   0.587723  -3.745 0.000422 ***
COLONY       0.866455   0.273974   3.163 0.002507 **
EAST         2.269490   0.389738   5.823 2.80e-07 ***
EUROPE      -2.253253   0.632561  -3.562 0.000751 ***
FERTLDC1    -2.678176   0.819765  -3.267 0.001843 **
GDPCH60L    -1.285338   0.238419  -5.391 1.39e-06 ***
GOVNOM1     -3.378956   1.585343  -2.131 0.037383 *
IPRICE1     -0.008541   0.001648  -5.181 3.01e-06 ***
LAAM        -1.302410   0.304477  -4.278 7.29e-05 ***
LIFE060      0.041256   0.020918   1.972 0.053436 .
OTHFRAC      0.750157   0.280440   2.675 0.009738 **
PRIGHTS     -0.475863   0.099674  -4.774 1.30e-05 ***
POP1560     17.008470   4.360919   3.900 0.000256 ***
POP6560     32.380339  10.887535   2.974 0.004302 **
SIZE60       0.110331   0.061348   1.798 0.077404 .
YRSOPEN      0.892989   0.399541   2.235 0.029350 *
```

Figure 2: Coefficients used in the linear regression model

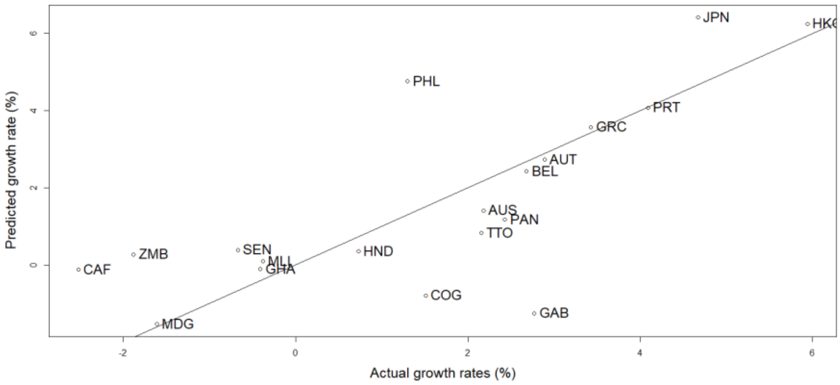


Figure 1: Graph of predicted growth rate values with the linear model against the actual growth rates in the test data. The diagonal line is a perfect fit line where predicted values equal to actual values (i.e. y=x)

K-Nearest Neighbours Regression

A non-parametric model we can use is K-nearest neighbours regression, where the model will predict the growth rate value of an observation in the test set based on the nearest neighbours in the training data. Using leave-one-out cross validation, we can determine the best K value to balance the bias and variance trade-off of the model. We obtain a value of K = 6 which gives us the lowest in-sample MSE. Figure 3 shows us how the MSE decreases as we increase K until we reach a minimum at K = 6 and subsequently increases due to overfitting. Upon passing the model on the test data, we obtain a RMSE = 1.343.

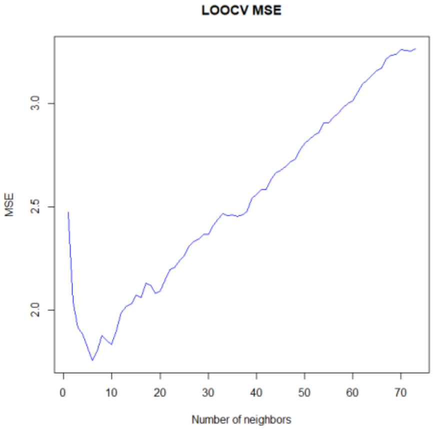


Figure 3: Graph of mean-squared error (MSE) against the number of neighbours (K)

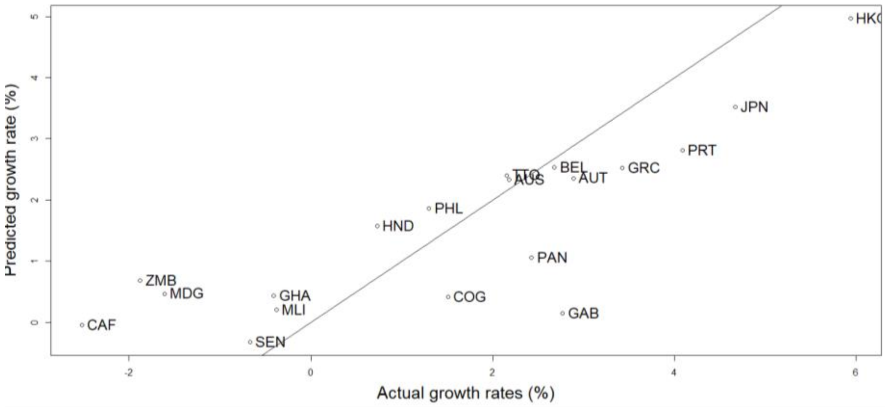


Figure 4: Graph of predicted growth rate values with the KNN regression model against the actual growth rates in the test data. The diagonal line is a perfect fit line where predicted values equal to actual values (i.e. y=x)

Figure 4 shows us the predicted growth rates compared to the actual growth rates. It is worth noting that for countries with higher positive growth rates, the model tends to underpredict while countries with low or negative growth rates, the model overpredicts the growth rates.

Regression Trees

Another useful and visually intuitive method are regression trees where each the value at each leaf represents the predicted GDP per capita growth rate between from 1960 to 1996 in percentages. When an unpruned regression tree built from the training data is passed on the test data, we obtain a RMSE of 1.133. Figure 5 shows us the full tree and its 13 leaves. While this is already lower RMSE compared to previous models, we shall use cross-validation to prune the tree and obtain a smaller tree which has a higher bias, thus giving us less of an overfit to the training data. We end up with a pruned tree containing only 4 leaves.

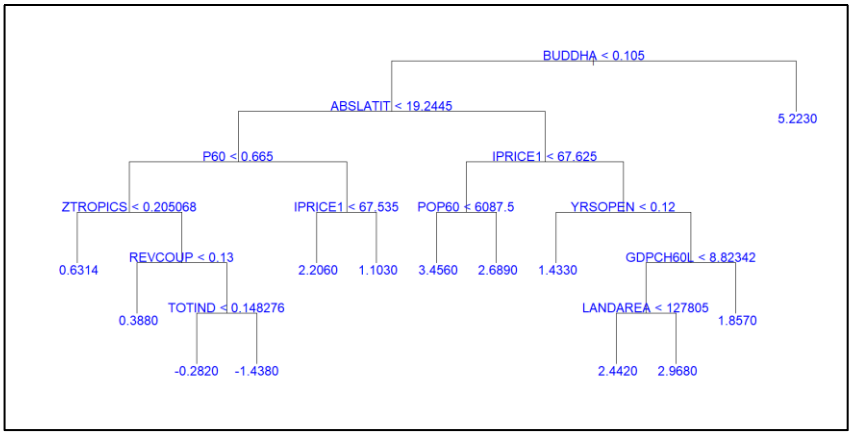


Figure 5: Unpruned regression tree

Figure 3 shows the pruned regression tree obtained using cross-validation. The only 3 variables used are BUDDHA (Fraction Buddhist), ABSLATIT (Absolute latitude), and P60 (Primary schooling in 1960). When passed on the test data, we obtain a RMSE of 1.02, a small improvement from the unpruned tree. This does imply that even an unpruned tree does a much better job at predicting that a linear regression model with carefully picked coefficients. Figure 4 shows a visualization of the predicted growth rate values using our pruned tree compared to the actual values in the test set. We can observe 4 levels of predicted growth rate values, which comes from the 4 leaves.

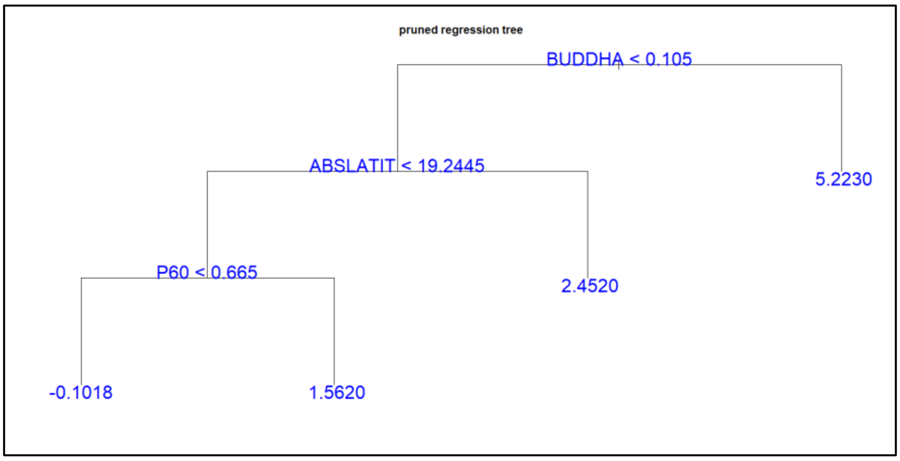


Figure 6: Pruned regression tree

Principal Components Regression

PCR utilizes principal components built via principal component analysis to create a linear regression model. Like determining the appropriate number of Ks to use in KNN regression, an important question in PCR regression is the number of principal components to keep. Using leave-one-out cross validation on the training data as shown in Figure 7, we obtain the number of PCs equal to 17. When the model is evaluated using the test data, we obtain an RMSE of 1.315. Now this is arguably worst than previous models. One reason for this is that the data several categorical variables and PCA is designed for continuous data. However, I find that the presence of some dummy variables such as EAST (East Asian Dummy) or LAAM (Latin American Dummy) important in the context of predicting economic growth rates as geography can play an important role in determining economic growth.

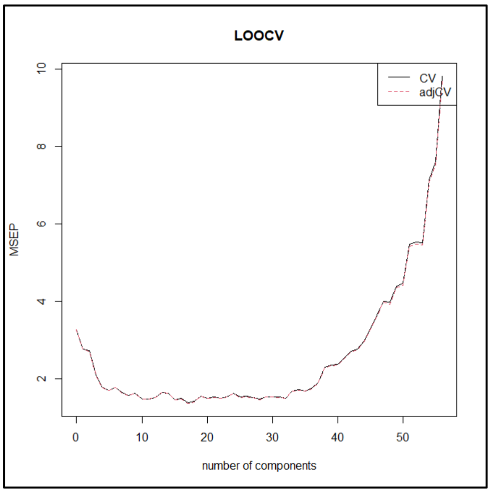


Figure 7: Graph of MSE against number of principal components obtained using LOOCV

Conclusion

Using RMSE as a metric of model performance, my analysis ranks the models as follows:

1. Regression Tree (RMSE = 1.02)
2. Principal Components Regression (RMSE = 1.315)
3. K-Nearest Neighbours Regression (RMSE = 1.343)
4. Linear Regression (RMSE = 1.663)

Regression trees performed the best on this dataset. The presence of many variables that interact with one another or are correlated with one another such as geographic or demographic predictors is a probable reason why linear regression is unsuitable to be built on such non-linear relationships. It is also worth noting that the very limited number of observations in the filtered data set does impose a limitation on how well one can train the models. Even then, I still chose to split the data into training and test set and conduct cross-validation on the training set only even if that meant an even smaller sample size. This is because I value the performance of the model on completely unseen data more as this can help us in solving the penultimate question as to what determines long-term economic growth on any country or region in the world. This also brings me to the importance of interpreting the models. We should be cautious when extrapolating the models to other time periods. For instance, one should not expect to see any decent performance when using any of the models used above the predict the GDP per capita growth rate of Nigeria in 2026. This is because contextual information about the economic and political conditions of the world captured in the data is specific to the 1960 to 1996 period. I believe that long-term economic growth boils down to more fundamental factors such as the strength and nature of a country's institutions.