Key Drivers of Long-Term Economic Growth: A Regression Problem

1. Introduction

Understanding the drivers of long-term economic growth is a central question in economics. Countries differ widely in long-term economic growth due to a range of structural, demographic and institutional factors. To investigate these questions, we rely on a data set which comes from the Sala-i-Martin et al. (2004) paper published in American Economic Review which contains *139 observations and 71 variables.* In this report, we aim to identify the factors most strongly associated with long-term economic growth and evaluate how well different modelling approaches capture these relationships. Our findings show that human capital, wealth of the economy relative to other economies and capital accumulation are the key drivers of economic growth. We begin with a benchmark model, *Linear Regression with 3 selected variables* and compared it to 3 advanced models; *K-Nearest Neighbour, Decision Tree and Linear Regression with variable selection*.

2. Exploratory Data Analysis

2.1. Preprocessing Data

Before analysing our data, it is necessary to clean the data and make any necessary adjustments. When reading in our data, we excluded the 'OBS', 'CODE' and 'COUNTRY' variables as they are meaningless in our investigation. We also changed all our data values to type numeric. When dealing with empty values, we started off by removing observations which had empty values for our outcome variable, *average GDP-per-capita growth rate between 1960-1996 (GR6096)*. To maximise the number of variables we have while keeping the number of observations high, we decided to remove observations with at least 2 empty values and then, remove all variables with at least 1 empty value. Our cleaned data set now has 94 observations and 63 variables. We then performed a Z-score standardisation on the continuous variables. We applied a 70/30 train-test split to test how the model would perform against unseen data. A seed number of 6769 is also set to ensure that results can be replicated.

2.2. Interesting Insights

We decided to explore the relationship between our outcome variable, *GR6096* against some predictors namely *initial GDP 1960 (GDPCH60L), life expectancy in 1960 (LIFE060), primary schooling in 1960 (P60), degree of ethnic diversity (AVELF) and whether the country is East Asian (EAST)*. Based on our findings, for our **continuous** variables, there is a moderately strong positive relationship between *GR6096* and both *LIFE060* (correlation coefficient = 0.5409) and *P60* (correlation coefficient = 0.5726). This is likely due to increase in human

capital leading to increased productivity thus, increasing the economy's output. To visualise this relationship, we plotted a scatter plot of *GR6096* with *LIFE060* (Fig.1) and *P60* (Fig.2).

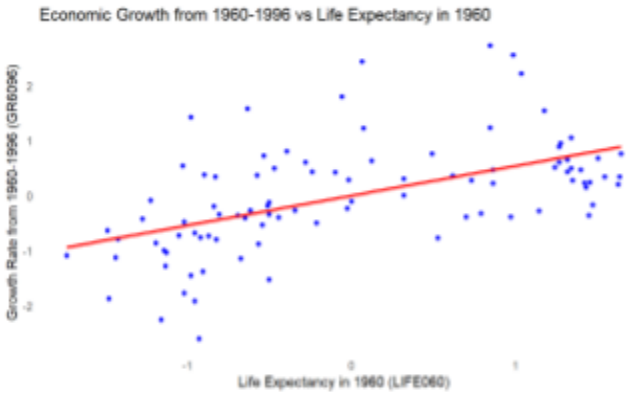

Fig.1. Scatter plot of GR6096 against LIFE60

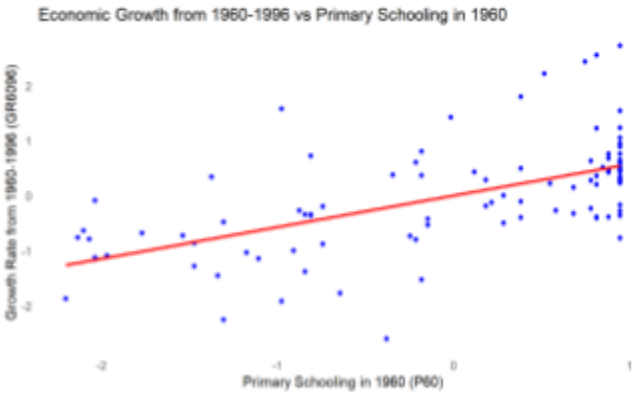

Fig.2. Scatter plot of GR6096 against P60

For the **categorical** variable, EAST, we plotted a boxplot to investigate the relationship with *GR6096* (Fig.3). From Fig.3, we found that East Asian countries had a significantly higher median *GR6096.* This is likely due to the East Asian Miracle (1960-1996) where several East Asian countries experienced rapid economic growth, driven by factors such as improved human capital and increased capital investments.

Fig.3. Boxplot of GR6096 with EAST

3. Methods

Benchmark model: Linear Regression

We designed a benchmark model, linear regression with 3 predictors – *P60, GDPCH60L and Average Investment Prices from 1960 - 1964 (IPRICE1)*. When using the full model for our linear regression, the coefficients of *P60* (p-value = 0.188)*, GDPCH60L* (p-value = 0.104) *and IPRICE1* (p-value = 0.064) have the top 5 lowest p-value, indicating stronger statistical relationship with *GR6096.* Additionally, these predictors are widely recognized as the core determinants of long-run economic growth. Together they capture human capital, initial income conditions and capital-accumulation distortions, representing the three fundamental channels emphasized by economic theory. Being measured at the start of the period reduces concerns about reverse causality and contemporaneous shocks. Our estimated regression model (Fig.4) achieved a Root-Mean-Square Error (RMSE) of 0.795.

$$\widehat{GR6096} = \underset{(0.0912)}{0.0148} + \underset{(0.1263)}{0.6788} \times P60 - \underset{(0.1396)}{0.3162} \times GDPCH60L - \underset{(0.0875)}{0.3366} \times IPRICE1$$

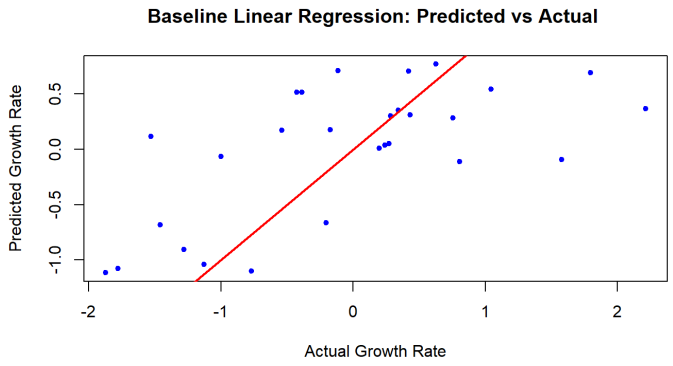

Fig.4. Predicted vs Actual Growth Rates for the Baseline Linear Regression Model

Model 1: K-Nearest Neighbour (KNN)

For the same reasons mentioned in the benchmark model, we utilized the same 3 predictors – *P60, GDPCH60L and IPRICE1* for our KNN model. To decide the optimal K value, we performed leave-one-out cross validation (LOOCV), finding that the best K value is K = 12 (Fig.5). A RMSE of 0.808 was obtained.

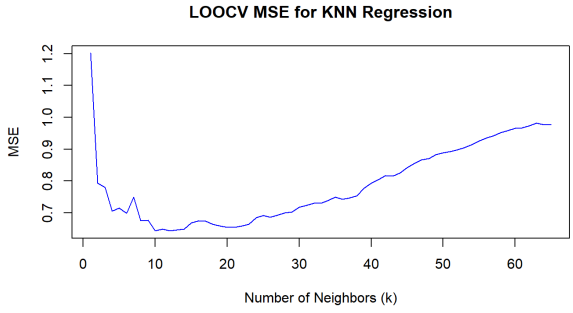

Fig.5. Determining best K value for KNN model using LOOCV

Model 2: Decision Tree

The decision tree works by partitioning the predictor space using recursive binary splits into groups of countries with similar growth rates. Each terminal node of the tree represents a set of countries with comparable long-term growth. When we constructed a full, unpruned tree based on the training data, we obtained a tree with 11 leaves and 21 nodes (Fig.6).

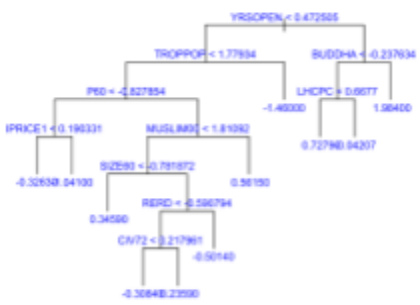

Fig.6. Big Decision Tree

Fig.7. Pruned Tree

We then performed cross-validation to determine the optimal tree size and pruned branches that did not meaningfully improve predictive accuracy. We obtained a tree with 5 leaves and 9 nodes (Fig.7). In this model only 4 predictors are used, *Years open (YRSOPEN), Fraction Population in tropics (TROPPOP), Fraction Buddhist (BUDDHA) and P60*. The pruned tree model achieved a RMSE of 0.839.

Model 3: Linear Regression with variable selection

For this model, we made use of the 'leaps' package in R to perform backward elimination on our linear regression with a full model (62 predictors). For the best in-sample prediction, we identified a model with 47 predictors (Adjusted R-squared = 0.888). Following which, we used this model to predict out-of-sample and obtained a RMSE of 3.447.

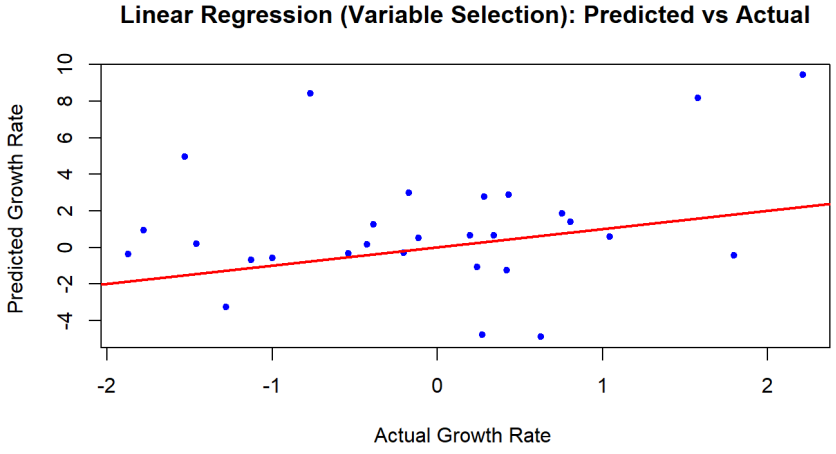

Fig.8. Scatter Plot of Predicted vs Actual Growth Rates for the Linear Regression Model with variable selection

4. Results

We will be using RMSE as our evaluation metric to determine how well the different modelling approaches predict our outcome variable, *GR6096*. Our ranking is as follows:

1. Baseline Linear Regression (RMSE = 0.795)
2. KNN (RMSE = 0.808)
3. Decision Tree (RMSE = 0.839)
4. Linear regression with variable selection (RMSE = 3.447)

Baseline Linear Regression model performed the best, with the lowest RMSE of 0.795. This suggests that *GR6096* follows most closely a linear regression with *P60, GDPCH60L and IPRICE1* as predictors compared to our other models.

5. Discussion

We should note that despite our Baseline Linear Regression model performing better than our KNN and Decision Tree model, it is only by a marginal amount. However, these 3 models perform significantly better than our Linear Regression model with variable selection.

A key limitation for our analysis is the small sample size, constrained by the finite number of countries. This issue is exacerbated by missing data, which further reduces the number of observations and variables post data cleaning. As a result, the number of observations is close to the number of variables increasing the risk of overfitting and unstable model estimates. From an economics perspective, the dataset spans only from 1960-1996 and therefore excludes major events such as the dot-com bubble and recent technological transformations, limiting our model's ability to capture the dynamics of today's economy. As such, we should be cautious when extrapolating the model to other time periods. For example, the *EAST* predictor, despite showing some kind of relationship with *GR6096* (Fig.3), may not be as useful in predicting future economic growth as it was likely due to the East Asian Miracle which only happened during that time period. On closer examination, each modelling approach faces different limitations. Our linear regression with variable selection produced high RMSE due to overfitting, leading to poor out-of-sample performance as the model fitted the noise from the training data rather than underlying relationships. As for the Tree Model, it would underperform in prediction and was unstable as small changes in the training data lead to large shifts in tree size. For KNN, the choice of K = 12 is large relative to sample size (96 observations). Although it reduces sensitivity to noise, it also causes underfitting by oversmoothing predictions which may ignore local patterns.

Despite the likelihood of a model being able to better predict out-of-sample long-term economic growth (eg. KNN with all predictors), we believe that the computational effort needed would outweigh the marginal improvements in our model. As such, it is not meaningful to pursue a model beyond our benchmark model. Additionally, given the performance of our benchmark model, we believe that human capital, wealth of the economy relative to other economies and capital accumulation are the key drivers of economic growth.

6. Appendix

AI DECLARATION (PROMPTS UTILISED):

“How to change all variables in data set to type numeric”
“How to identify variables in data set except some R”
“How to remove gridlines in scatterplot R ggplot2”
“How to make nicer colour for ggplot boxplot R”
“How to sort coefficients table based on p-values”
“How to add a 45 degree reference line in R”
“Check for syntax error”