Model creation to predict GDP growth of a country

Adiba Mobarak, Autumn Larsen, Sanskriti Giri

Problem

What determines the GDP growth of a country? Is it determined by factors such as the number of active working population or the investment by the government in its economy? To answer our question, we looked for a dataset as such. We found a dataset that included several determinants of economic growth for a sample of 121 countries which had records from the years 1960 - 1985. We decided to use this data to see if we could form an adequate model to predict the GDP growth of a given country. We used the dataset Determinants of Economic Growth (GrowthDJ) from package AER.

Data Description

```
no yes
98  23
```

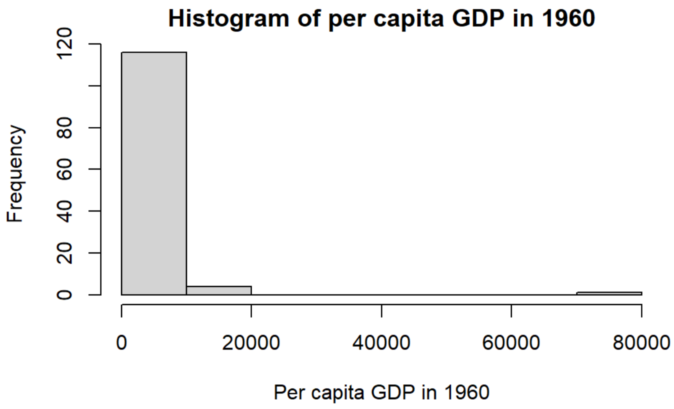
The variable *oil* is a binary variable that indicates whether a given country is oil-producing (*yes*) or not (*no*). In this data, 98 countries are not oil-producing and 23 are.

```
no yes
46  75
```

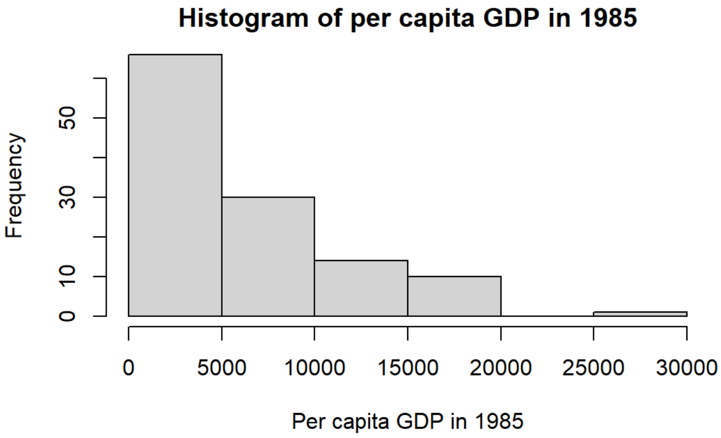
The variable *inter* is a binary variable that indicates whether a given country has better quality data (*yes*) or not (*no*). In this data, 46 countries do not have better quality data and 75 do.

```
no yes
99  22
```

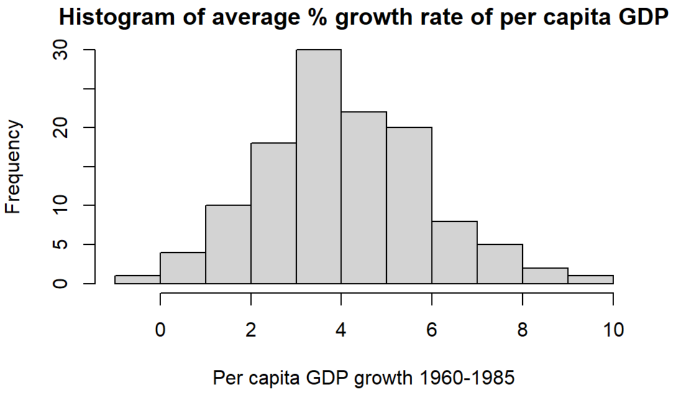
The variable *oecd* is a binary variable that indicates whether a given country is a member of the OECD (Organisation for Economic Co-operation and Development) (*yes*) or not (*no*). In this data, 99 countries are not members, and 22 are.



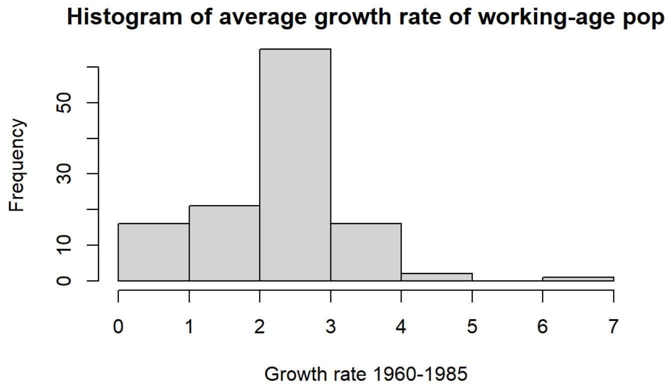
The variable *gdp60* is a continuous variable indicating the per capita GDP (gross domestic product) of a given country in the year 1960. The data is skewed to the right, with an outlier between 70,000 and 80,000.



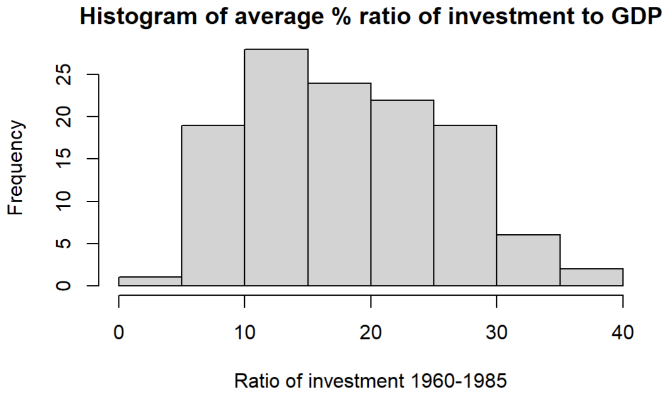
The variable *gdp85* is a continuous variable indicating the per capita GDP of a given country in the year 1985. The data is skewed right with an outlier between 25,000 and 30,000.



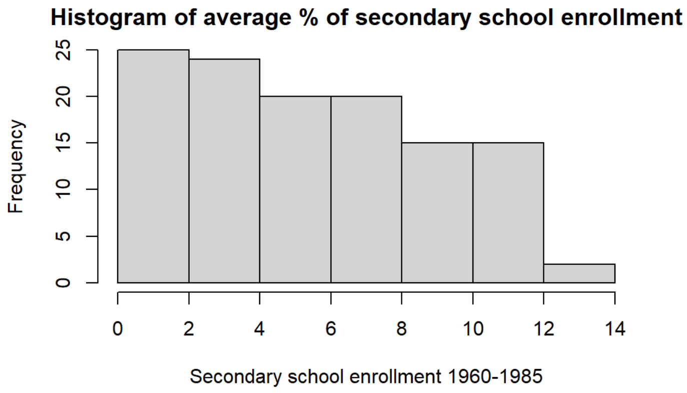
The variable *gdpgrowth* is a continuous variable indicating the average growth rate of per capita GDP of a given country from 1960 to 1985 as a percentage. This data is mostly normally distributed with a slight right skew.



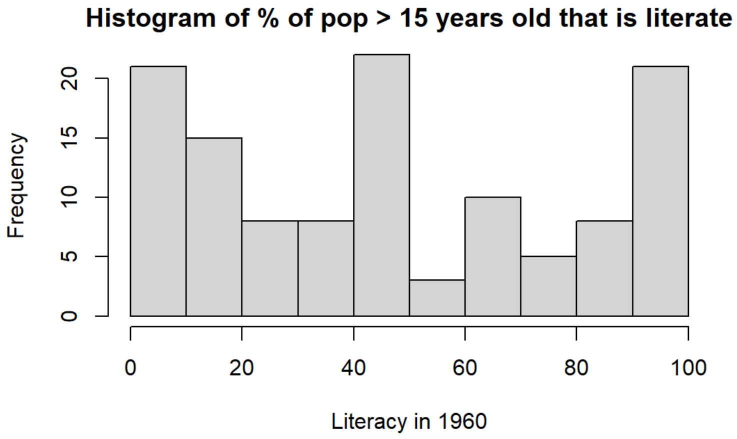
The variable *popgrowth* is a continuous variable indicating the average growth rate of working-age population from 1960 to 1985 of a given country as a percentage. This data is somewhat normally distributed with a right skew and an outlier between 6 and 7.



The variable *invest* is a continuous variable indicating the average ratio of investment (including Government Investment) to GDP of a given country from 1960 to 1985 as a percentage. This data seems normally distributed.



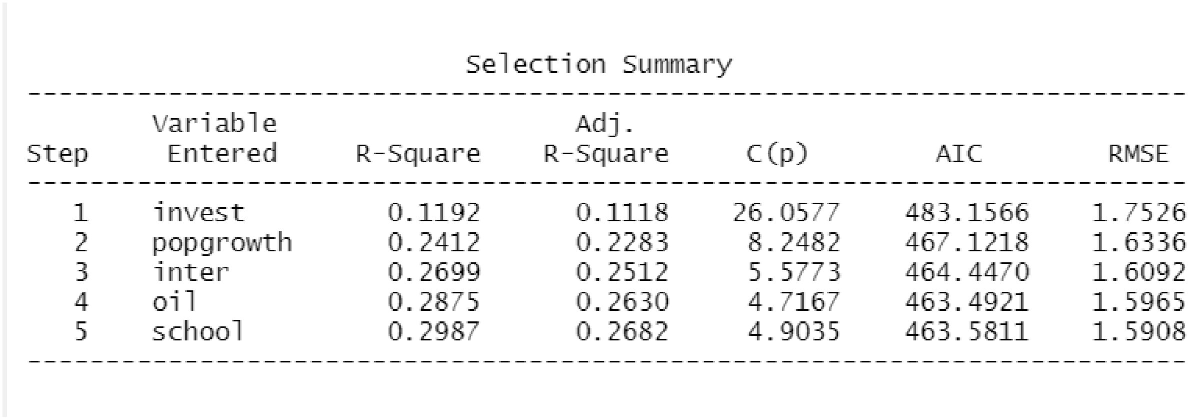
The variable *school* is a continuous variable indicating the average fraction of working-age population enrolled in secondary school of a given country from 1960 to 1985 as a percentage. This data is skewed right.



The variable *literacy60* is a continuous variable indicating the fraction of the population of a given country over 15 years old that is able to read and write in 1960 as a percentage. This data is not normally distributed.

Variable Selection

We used **forward selection method** to choose the most significant variables. Beforehand, we removed the variables *gdp60* and *gdp85* because the variable *gdpgrowth* is a product of those variables, and thus it would be repetitive to include them and would make the model inaccurate. From our analysis, we found these variables to be the most significant:



```
                              Selection Summary
------------------------------------------------------------------------------
        Variable                     Adj.
Step     Entered       R-Square    R-Square     C(p)         AIC         RMSE
------------------------------------------------------------------------------
   1    invest           0.1192      0.1118    26.0577    483.1566    1.7526
   2    popgrowth        0.2412      0.2283     8.2482    467.1218    1.6336
   3    inter            0.2699      0.2512     5.5773    464.4470    1.6092
   4    oil              0.2875      0.2630     4.7167    463.4921    1.5965
   5    school           0.2987      0.2682     4.9035    463.5811    1.5908
------------------------------------------------------------------------------
```

- *oil*
- *school*
- *invest*
- *popgrowth*
- *inter*

Models

Our first step was to create a linear regression model using the 5 selected variables. However, using our knowledge in the field of economics, we suspected that a linear model may not be the best option since variables like invest and oil or school and population growth could interact with each other. Similarly, we also suspected that the effect of population growth and investment on the GDP growth may be exponential rather than linear. To address this problem, we created a model with interaction and quadratic terms.

Using a partial F-test we decided to move forward with this model:

log(gdpgrowth)=invest+popgrowth+inter+oil+school+oil*popgrowth+I(invest^2)

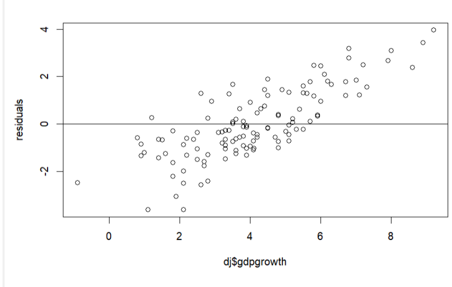
Multicollinearity

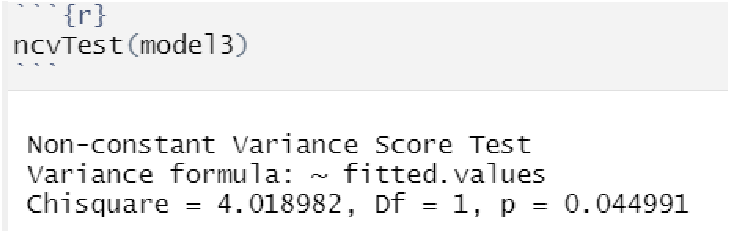
We checked whether there was a significant correlation between our predictors. There were indications that invest had a significant correlation with the other predictors (indicated by the VIF exceeding 10). We chose to keep this predictor in our model and we acknowledge that our model cannot be used to establish a cause-and-effect relationship between *gdpgrowth* and our predictors.

Model Assumptions

We tested our model for:

1. **Constant Variance Assumption**





~~~
```{r}
ncvTest(model3)
```

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.018982, Df = 1, p = 0.044991
~~~

Looking at the plot and the results of our ncvTest we found that our model violated the assumption of constant variance.

2. **Influential Observations**

After conducting a cook's distance test, we found that our model did not have any influential observations.

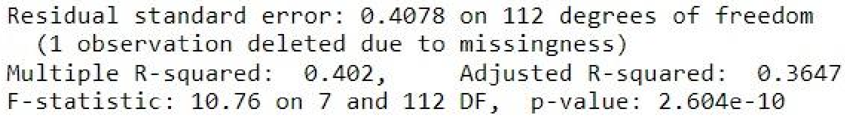
Remedy

To fix the assumption of heteroscedasticity, we used the natural log transformation on the dependent variable GDP growth.

As we can see in the plot, the data points are randomly scattered around the line and do not show a pattern. Our problem is now fixed.

Model Adequacy

We used a global F test to check the adequacy of our model. We assumed a significance level of 1%. We used the summary() function:



```
Residual standard error: 0.4078 on 112 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.402,     Adjusted R-squared:  0.3647
F-statistic: 10.76 on 7 and 112 DF,  p-value: 2.604e-10
```

We found that for our test statistic, F, which was equal to 10.76, the p-value was 0.00000000002604. We thus concluded at 1% significance that our model is statistically useful in predicting the average *gdpgrowth*.

Working Model

Based on our results, our final working model is:

$ln(gdp\hat{g}rowth) =$ -0.517745 $+$ 0.1139176* $invest$ $+$ 0.2079831* $popgrowth$ $+$ 0.3298147* $inter$ $+$ 0.6818305* oil - 0.0191433* $school$ - 0.0022086* $invest^2$ - 0.1792946* $popgrowth * oil$