LINEAR REGRESSION AND ROBUSTNESS

Coursework

Abstract

This report gives an overview of estimating an appropriate model which predicts life expectancy from the other variables.

Sree Lalitha Gorty

Student ID: 201684407

E-mail: mm22slg@leeds.ac.uk

Introduction

Regression is an important element of data analysis with a wide range of real-time applications. In our report, we focus on how linear regression can have a cause-and-effect relationship between the variables. There are different types of regressions that can be performed at different occasions as shown in figure 1.

Figure 1: Types of regression

Out of all the types of regression, linear regression is a basic form of regression algorithm for data analysis which consists of a single parameter and one dependent variable that has a linear relationship. If the number of dependent variables increases, then it is called multiple linear regression. Our data set consists of different variables that throw light on the different factors that affect life expectancy at birth which represents a consolidated overview of various studies conducted by the World Health Organization.

We start our analysis of finding a model fit for our data set which is based on life expectancy of each country. The dataset has 12 variables depicting various attributes which could be used to predict a suitable model that predicts the life expectancy from the other variables.

Task 1: Life expectancy v/s GDP per capita

Our first interest would be to apply a suitable transformation to our dataset to arrive at a linear model that describes a relationship between our variables' life expectancy and GDP per capita. We would first fit the model establishing a relationship between the desired variables. Figure 1 depicts the summary after introducing a linear model.

```
> summary(m)

Call:
lm(formula = life.expectancy ~ GDP.per.capita, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-35.773  -4.019   1.933   5.256  10.944

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.685e+01  3.113e-01  214.76   <2e-16 ***
GDP.per.capita 3.019e-04  1.541e-05   19.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.013 on 729 degrees of freedom
Multiple R-squared:  0.3448,    Adjusted R-squared:  0.3439
F-statistic: 383.7 on 1 and 729 DF,  p-value: < 2.2e-16
```

Figure 1: Summary of our model highlighting the coefficient beta hat and standard error.

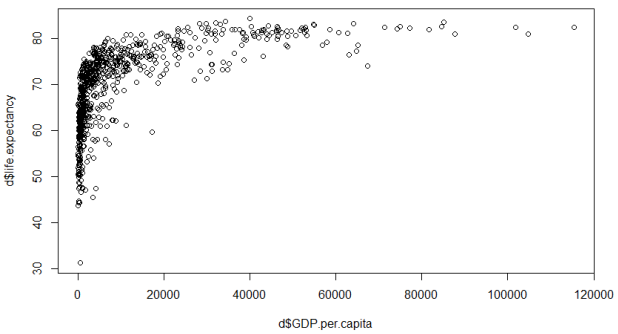


Figure 2: A scatter plot between life expectancy and GDP per capita

A residual plot helps us understand the data behaviour and its nature. Figure 3 shows a residual plot between our variables which clearly shows that the data is non-linear and is very skewed in nature.

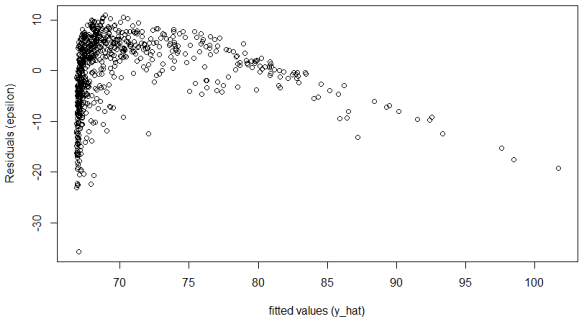


Figure 3: Residual plot between life expectancy and GDP per capita

For the next step, we would like to apply an appropriate transformation that best describes a relationship between our variables. We use a transformation when the data is non-linear in order to linearize the data and draw suitable observations from it. We have different transformations that could be applied to attain the appropriate fit namely, y – log x transformation, y^2 – log x transformation, log y- log x transformation, etc.

In this case, we could use y-log x and log y – log x transformations which could be helpful to make a comparison between them and pick one which fits the best. On applying the transformation and closely observing the values (refer figure 4) the value of R square which gives us a measure of how well it could be a model fit, is lesser for log y – log x transformation. This means that, lesser the R square value, the less it is likely to be an appropriate model fit.

R square value after applying Y – log x Transformation

```
Call:
lm(formula = (life.expectancy) ~ log(GDP.per.capita), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-30.9366  -2.3549   0.6836   3.3606  11.7007

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        32.8442     1.0558   31.11   <2e-16 ***
log(GDP.per.capita)  4.5094     0.1253   36.00   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.198 on 729 degrees of freedom
Multiple R-squared:   0.64,     Adjusted R-squared:  0.6395
F-statistic:  1296 on 1 and 729 DF,  p-value: < 2.2e-16
```

R square value after applying log Y – log x Transformation

```
Call:
lm(formula = log(life.expectancy) ~ log(GDP.per.capita), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68188 -0.03192  0.01082  0.05337  0.18064

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.69029    0.01720   214.6   <2e-16 ***
log(GDP.per.capita)  0.06672    0.00204    32.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08466 on 729 degrees of freedom
Multiple R-squared:  0.5947,    Adjusted R-squared:  0.5941
F-statistic:  1070 on 1 and 729 DF,  p-value: < 2.2e-16
```

R square value of the initial model

```
Call:
lm(formula = life.expectancy ~ GDP.per.capita, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-35.773  -4.019   1.933   5.256  10.944

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.685e+01  3.113e-01  214.76   <2e-16 ***
GDP.per.capita 3.019e-04  1.541e-05   19.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.013 on 729 degrees of freedom
Multiple R-squared:  0.3448,    Adjusted R-squared:  0.3439
F-statistic: 383.7 on 1 and 729 DF,  p-value: < 2.2e-16
```

Figure 4: Comparison of R square value for different transformations

In conclusion for our analysis on finding a suitable with appropriate transformation, we wish to choose the y – log x transformation based on the R square value and the Q-Q plot.

Now that we have our model, we now try and check appropriate diagnostics that can confirm that our model is a suitable fit.

1. We start off with a scatter plot between the response output and the explanatory variables. The scatter plot for the initial model without any transformation is shown in figure 2. The following figure 5 shows us the scatter plot for the response and explanatory variables after applying each of our transformations. As we compare, we can observe a good linearity after applying transformation.

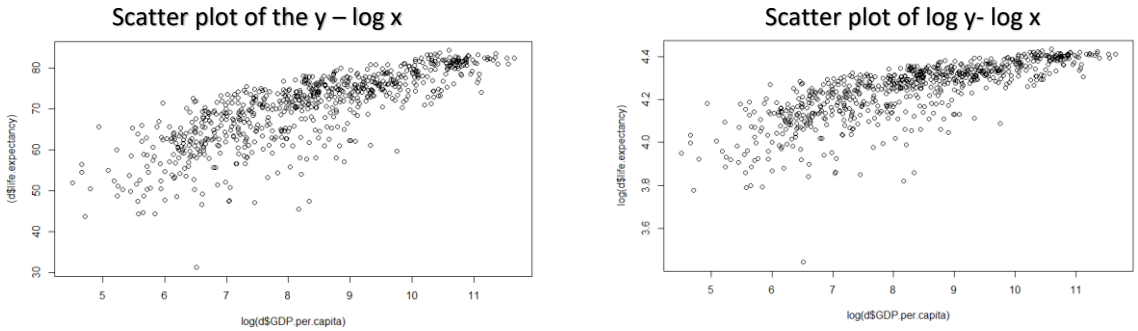


Figure 5: Scatter plot between response and explanatory variables

2. Our interest now would be to plot residuals against each of the explanatory variables. If the values have the same variance around the line epsilon = 0, then we can conclude that it is a well-fitted linear model.

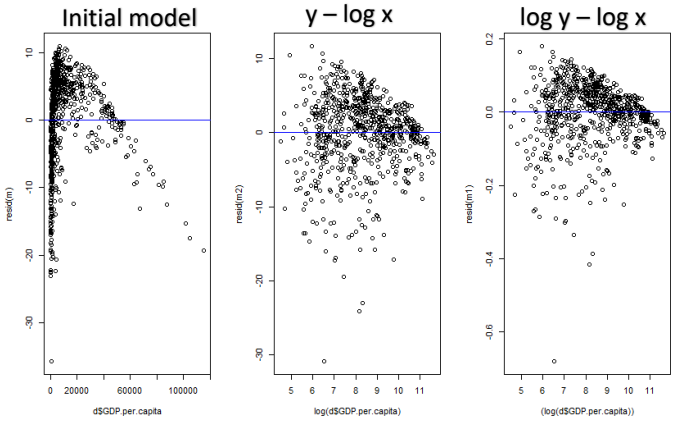


Figure 6: Scatter plot between residual and explanatory variables

3. For the next part, from figure 6, we have a scatter plot between the residual and the fitted values. This plot would help us determine how close the points are scattered near the mean line which signifies linearity. The residual plot of y – log x shows a significant difference and is also in coordination with linearity.

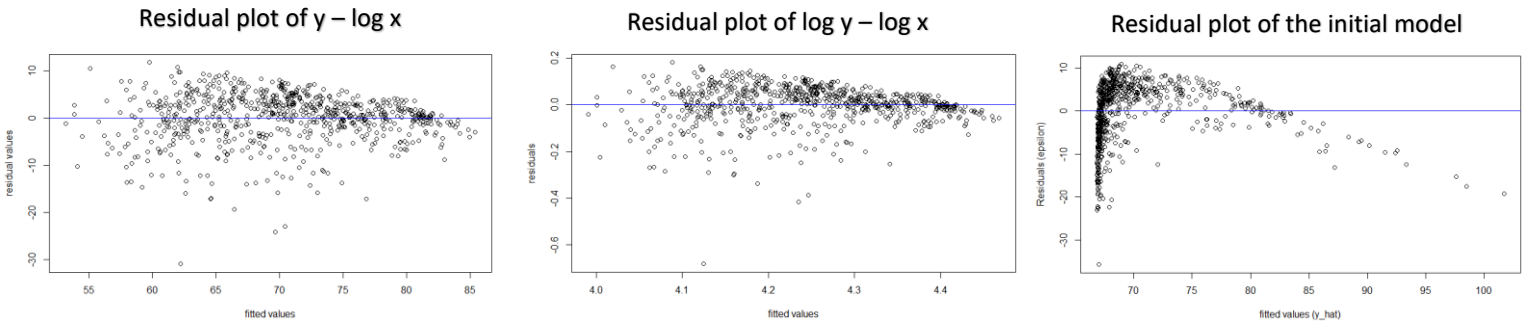


Figure 7: Comparison of residual plot for different transformations

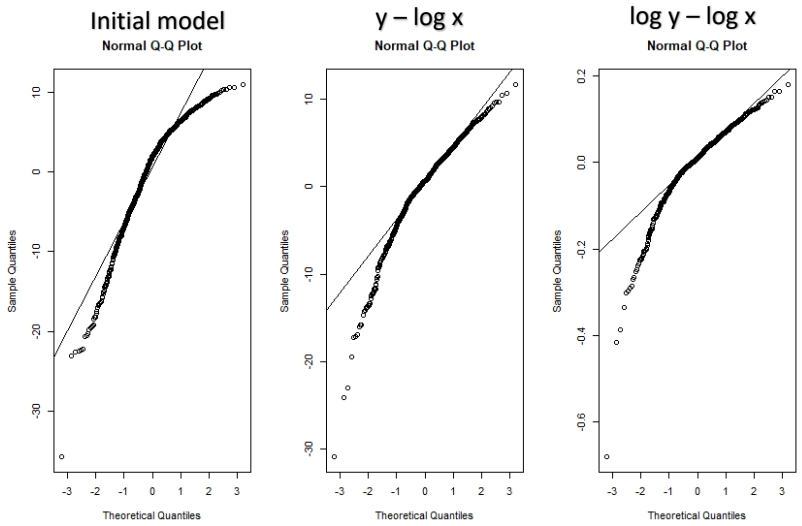


Figure 8: Comparison of Q-Q plot for different transformations

4. A Q-Q plot gives us an idea of the normality of the residuals of the given dataset. The data quantiles are compared to that of the normal distribution ones. For our samples, we see that the points are almost lying on a straight line after applying transformations.

When we plot response and residuals with time, it increases with time. Irrespective of time being an exponential variable according to our sample, it could be a valid estimator in certain cases to fit a model appropriately.

From the regression diagnostics, we see that our initial model was far off from linearity and was largely skewed. After applying our transformations, we can decipher that the y – log x transformation exhibited better chances of linearity and being a good model fit for our given sample for a plethora of reasons. The R square value was significantly large, the scatter plot between the response and explanatory variables was fairly exhibiting linearity, the data points approximately distributed themselves near to the epsilon = 0, the Q-Q plot was showing a straight line.

For the next part of our analysis, we would now like to obtain a 95% confidence interval for the mean life expectancy at birth for a country with per capita of 5000$. As we believe that our appropriate model fit would be using the y – log x transformation, we now calculate the model matrix, degrees of freedom and quantiles for the same. The 95% confidence interval for a country with a GDP of 5000 USD is calculated as [70.87, 71.63]

```
       fit      lwr      upr
1 71.25201 70.87039 71.63363
```

Task 2: Life expectancy with all the other variables

In the previous task, our focus was on one explanatory variable. Now, we choose multiple variables based on the weightage they contribute to the data set while selecting an appropriate model fit. This decision is based on examining values like R square adjusted (in this case as there are multiple variables now), the regression diagnostics and so on.

Before we dive into variable selection, we must ensure that the data is clean and does not include null or extreme values. Examining the dataset, we could locate a few NA values if ignored might meddle with our model selection. To mitigate this, we would replace the empty

values with the mean and then obtain the plots required to check for a need for transformation.

As selecting all the variables to estimate an appropriate fit is tedious, we need to narrow it down to a subset of variables that are pivotal to our aim. For this, we have 11 (apart from life expectancy variable) from our dataset out of which, 3 of them are categorical and 9 are numerical. Categorical values could be intervening if not employed in a suitable way to make them numerical. In this regard, we consider the variables that have numerical values and scaled them. Hence, our subset would be the variables-

Population Size, Health, Spending, HIV, alcohol, tobacco, GDP per capita, Life expectancy, and Population Density.

We have excluded the variable Year as it is standardised and has just three unique values which could be considered categorical.

For the next step in narrowing down our required subset, we would employ Exhaustive Search and Stepwise forward/backward selection.

- **Exhaustive Search using regsubsets:** To predict a variable subset, exhaustive search is a technique that could be employed for a dataset with variables less than 20. This algorithm considers all the potential combinations, which is 2^n-1 excluding the life expectancy variable.
- **Stepwise Forward selection:** This technique adds a predictor variable while checking the subsets and it never deletes them. It starts with an empty set and checks even the smaller subsets. Firstly, it predicts N number of variable subsets with each one having a predictor variable and finally selects the one with the higher criterion.
- **Stepwise Backward selection:** This technique is similar to that of the forward one but in this case, all the predictor variables are included before the search begins. Each predictor variable is deleted one at a time until we arrive at a subset with a higher value criterion.

We now go ahead to apply each of these transformations to our selected variables and observe the variables returned after executing each algorithm.

Technique	Initial variables given	Output variables after the algorithm
Exhaustive Search	Population Size, Health, Spending, HIV, alcohol, tobacco, GDP per capita and Population Density	Population size, population density, health, spending HIV, alcohol, tobacco, GDP per capita
Forward Selection	Population Size, Health, Spending, HIV, alcohol, tobacco, GDP per capita and Population Density	GDP per capita, HIV, alcohol, population density, tobacco
Backward Selection	Population Size, Health, Spending, HIV, alcohol, tobacco, GDP per capita and Population Density	Population density, HIV, alcohol, tobacco, GDP per capita

Table 9: Summary of variables obtained for different algorithms

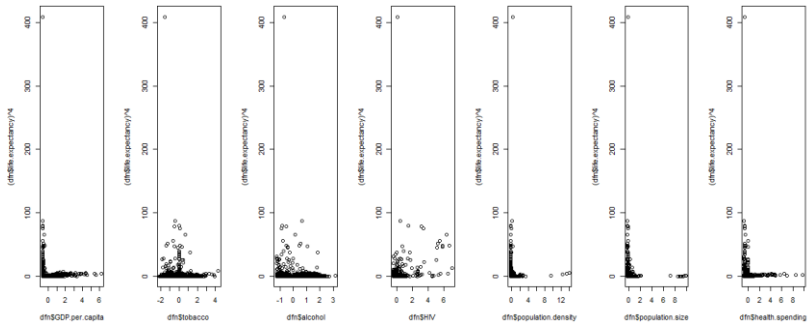
Now that we have obtained our subset of variables, our interest would now be to fit it into a linear model. We should also note that the subset of variables emerged out similar for forward and backward selection. We compare the R-squared adjusted value since we have

multiple variables to compare between different models. As the values do not show a significant difference in the adjusted R- squared values between different techniques, it would be a wise thing to apply a suitable transformation. In this case, we use (y^2)^2 transformation since the scaled values are negative, the square would make the values positive which gives us a better plot.

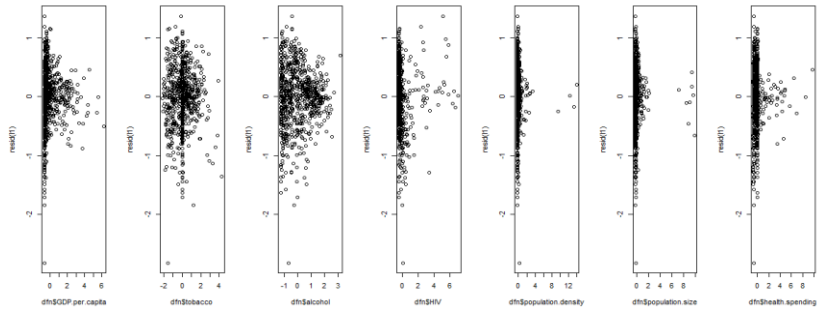
Technique	Adjusted R-squared value before transformation	Adjusted R-squared value after applying a transformation
Exhaustive Search	0.5745	0.725
Forward Selection	0.5737	0.6913
Backward Selection	0.5737	0.6913

Table 10: Summary of adjusted R-squared values for each technique.

Regression diagnostics for the subset of variables obtained from the exhaustive search as adjusted R-squared value is large which means that it could be a good fit.



A scatter plot of the response against each of the explanatory variables obtained from Exhaustive search.



A scatter plot of ε against each of the explanatory variables obtained from Exhaustive search.

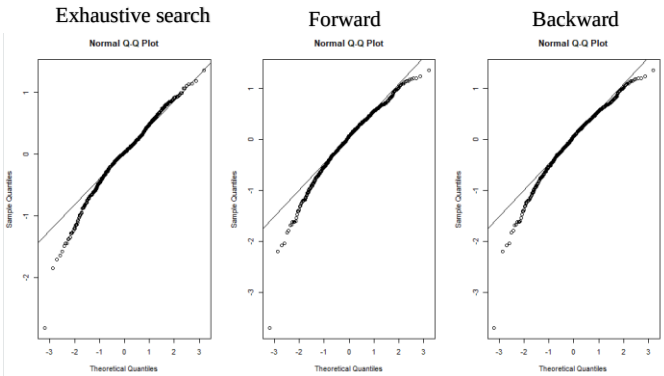


Figure 13: Normal Q-Q plot

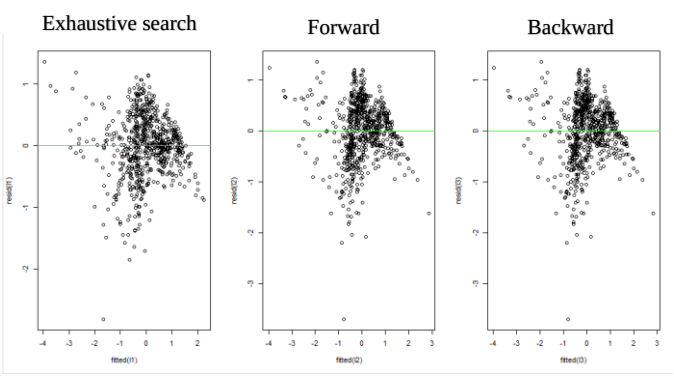


Figure 14: Response vs residual plot

Conclusion:

We can see that the model fitted using the Exhaustive Search results had a very slight advantage over the model fitted without these variables due to the addition of the two input variables, population size and health spending. This suggests that while these factors have a minor impact on the regression procedure, their presence demonstrates a minimal impact on the response variable.