SDGB 7840 HW 3: Modeling Literacy Rate

Darshan Patel

3/13/2019

## Executive Summary

This paper examines the literacy rate of different countries and seeks to explain it using educational, economical and social factors. It is discovered that by using the number of girls out of primary school and the mortality rate of children under 5, most of the variation in literacy rate of a country can be explained by a linear model. This shows that children have an impact on how the country is pictured on the global scale.

## Introduction

In today’s time, it is very important to know how to speak, read and write in a language in order to communicate with other people. Literacy is one way to measure how educated people are. As defined by UNESCO, a country’s literacy rate is defined to be the fraction of people who are literate in a given age group out of the entire population in that specific country. There are different types of literary rates that can be defined, such as adult literary rate and youth literary rate. These are for different populations of people. Factors that influence whether a person is literate or not range from economical ones such as whether the family is financially well off, to healthwise, whether people are fairly healthy in the country. Big factors that influences a country’s literary rate are completion rates of children in primary school and how they fare out in the work force. It would also be indicative to look at how prominent mobile devices are in different countries, which can allow people to access social media websites such as Facebook.

In this study, several factors that may affect country adult literary rates will be explored. Focusing on socioeconomical, gender and educational variables, a relationship will be sought out to try to explain literacy rates in different countries around the world using ten variables.

## Data

The data for literacy rates for countries come from the World Bank, a firm that provides support and other forms of assistance to developing countries around the world. Their goal is to reduce poverty levels and help underdeveloped nations prosper. Along with the data for adult literary rate, a number of data that could also influence literary rates was also received from the World Bank library of datasets. For this study, the features that will be used to explain literacy rates are

* number of girls out of primary school
* percentage of men over 15 who are in the labor force
* percentage of women who are employed in agriculture
* government expendicture on education, as a percentage of the GDP
* percentage of mobile phone users
* mortality rate of children under 5, per 1000 live births
* percentage of undernourishment in the population
* rate of primary completion for girls
* rate of primary completion for boys
* the ratio of student to teachers in primary school

To best explain literacy rates, it is important to see whether young children are going to school and if they are completing it as well. The size of the classroom can be indicative of whether children get personal attention from teachers which can affect whether students are motivated to actually learn or just there because everyone else is. If there is a high number of girls out of primary school, it could depect whether a country is progressive or regressive towards womens’ rights and their ability to gain an education. The mortality rate of children, as well as percentage of undernourishment seeks to explain the physical features of different countries. Are children receiving adequate health care? Are people able to survive with food and water on a day to day basis? If people are not getting their nutrients everyday, it can be indicative of poor living conditions as well as literary rate. Along with these factors, it is key to look at the labor force, such as the percentage of men who in the labor force along with the percentage of women who are working in agriculture. If there is a high percentage of women working in agriculture, it can show that women are typically not educated enough in school and thus they were not able to achieve other forms of jobs. The percentage of mobile phone users is also looked to draw a connection between peoples’ ability to get on the internet and join popular social media networks. If people are able to do so, it can indicate that people are literate and can read and write in one language.

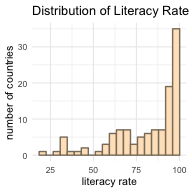
In creating the dataset for analyzing this question, several notions are made. Since not all countries have information for one, or two, particular year, data from 2010-2017 will be used for modeling. This is done so that there is some form of time frame for which information from each country all come from the same period of time while also maximizing how many countries’ information can be received. From this time frame, data for each variable is collected by looking at each countries’ latest record of information. For instance, if the UAE has data on percentage of mobile phone users in 2010, 2013 and 2017 but not other years, the value from 2017 will be used for that country. If a country has no information in this time frame, its data will be null.

Latest literacy rate data is first collected for each country. This forms a baseline of which countries will be used to study literacy rate. Afterwards, data for each of the ten explanatory variables is collected in the similar manner. Each of these variable columns are joined with the literacy rate data so that only countries with full data are retained. For instance, if Algeria does not have data on the pupil to teacher ratio, the country is taken out of the resulting dataset.

A total of countries are found with complete data for all explanatory variables and literacy rate.

## Methods

To explain literacy rate, first consider looking at the distribution of the response variable. A histogram of the literacy rate is shown below.



It is clear that literacy rate is skewed left. The distribution is not symmetric. This will be taken into account. Now, two models will be investigated in explaining literacy rate using a systematic approach. By performing forward stepwise selection, a set of variables will be selected such that a model containing only those variables has the lowest Bayesian Information Criterion compared to models of other variable sizes. From this, a model will be constructed to predict literacy rate as it is. For the second model, transformations will be applied to the first model for better performance.

### Model 1: Using Important Variables

Using the Bayesian Information Criterion, a model with variables will provide the best ability to predict literacy rate. These variables are number of girls out of primary school and mortality rate of children under . The model is constructed and the output is shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 103.4 | 1.134 | 91.12 | 2.433e-108 |
| **girlsnoschool** | 2.631e-07 | 1.015e-07 | 2.591 | 0.01083 |
| **mortality** | -0.6283 | 0.0249 | -25.23 | 1.772e-48 |

Model 1 - Regressing on Significant Variables

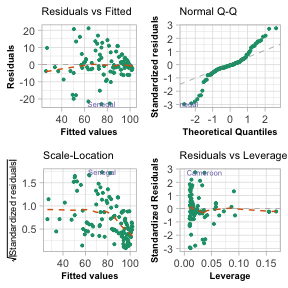
|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 117 | 7.759 | 0.8524 | 0.8498 |

Using only two explanatory variables, the model was able to explain of the variability in the literacy rate in different countries.

By the overall -test, it is found that at the level of , the model’s null hypothesis, namely that all the coefficient estimates are equal to zero, can be rejected since the probability of statistic 329.07 being greater than is less than . This means that at least one of the coefficient estimate is not zero. Therefore the model is adequate. Furthermore, at the level of , the null hypotheses, and can be rejected because the -values associated the estimates are and respectively. The -values associated with these -statistics are and , both of which are less than and so the coefficient estimates are statistically significant. This means that the two variables play a role in determining literacy rate when the other variable is included in the model.

When diagnosing this model, several assumptions are made. It is imperative to check these assumptions are fulfilled so that the model can be deemed valuable. The assumptions are: variables are fixed and measured without error, the mean and variance of the error is and (constant) respectively, the error terms are normally distributed and independent, and finally, the variables are not too highly correlated. For this dataset, the variables are not fixed nor measured without error. This is because surveying data comes with error in certain countries for many of the variables used in this study. In addition, variables are not chosen at fixed intervals. Thus this assumption is not fulfilled, but it should not play a huge role in the analysis of the model. Furthermore, it is found that the correlation between the two explanatory variables is , meaning there is no high level of correlation. Attaining a variance inflation factor value of for both variables, it is reasonable to conclude there is no significant correlation between the two variables, despite coefficient estimates having opposite signs.

After investigating the variables, the error terms can be looked at. Several informative plots are made about the model as shown below.

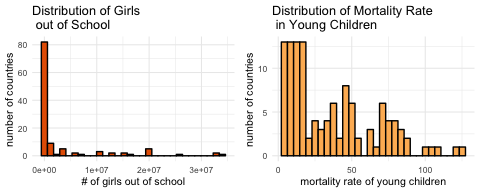


In the residuals vs. fitted values plot, a somewhat curved line is visible around , signifying that the relationship between the literacy rate and the two explanatory variables is not linear. For literacy rates above , the mean error of residuals is but below it is less. Furthermore, as fitted literacy rate increases, the spread of residuals decreases, indicating non-constant variance of the errors. This violation is also witnessed in the scale-location plot where the relationship between fitted values and square root of standardized residuals is nonlinear. Hence the assumption for the mean and variance of the errors being distributed with mean and constant variance is violated. Looking at the distribution of error terms in the normal quantile-quantile plot, the error terms have a heavy tailed distribution. Prediction for the literacy rate in Senegal will be inaccurate using this model as well as for other countries near the ends of this distribution. Thus the normality of the error terms assumption is also violated. In addition, the country of Cameroon has high leverage; its value(s) is/are far away from the mean. Finally, independence of error terms is looked at using a Durbin-Watson test with an level of .

With the null hypothesis that there is no residual correlation and alternative hypothesis that there is positive residual correlation, it is found that the Durbin-Watson statistic is , with a -value of , signifying that the null hypothesis is rejected. Therefore there is some evidence of positive residual autocorrelation in the error terms and hence the independence of error terms assumption is violated. Using time series methods will help to further look at this.

After evaluating the model and checking the regression assumptions, it is clear that this model is not the best model to explain literacy rate. Many assumptions such as normality and independence of error terms, and independence of error terms are broken. Furthermore, the coefficient estimates are of varying signs despite being significant.

### Model 2 - Transformation of Variables

Since so many assumptions are violated, try transforming the variables. First look at the distribution of each of the explanatory variables. 

The distribution of the number of girls out of school is heavily skewed right. This should make sense because this variable is given in integer form without accounting for the countries’ population. Also, the distribution of the mortality rate of young children is also skewed right, albeit not as heavily. Taking this all into account, as well as how the distribution of literacy rate is skewed left, the second model will be made by transforming the literacy rate by taking its square root and using the log of girls not in primary school. The mortality rate variable will be kept as it is since it is not heavily skewed. In addition, the leverage point, Cameroon, will be taken out of the dataset. The output of the model is shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | 9.943 | 0.168 | 59.17 | 7.108e-87 |
| **log\_girlsnoschool** | 0.03734 | 0.01566 | 2.385 | 0.01875 |
| **mortality** | -0.0399 | 0.001778 | -22.44 | 1.92e-43 |

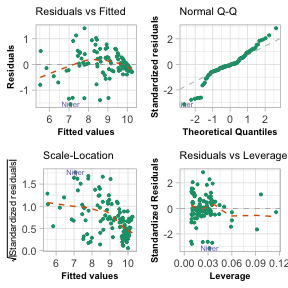
Model 2 - Transformation of Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Observations | Residual Std. Error |  | Adjusted |
| 116 | 0.5036 | 0.8391 | 0.8362 |

The difference between this model and the previous model is not aplenty. The model explains of the variability in the square root of the literacy rate. The previous model explained of the variability of literacy rate. Furthermore, by the overall -test, the model is considered adequate; the model has a -statistic value of 294.61 and a corresponding -value of . Thus at least one of the two coefficient estimates is not zero. Furthermore, at the level of , the null hypotheses, and can be rejected since the -values associated with the coefficient estimates are and , both of which are less than . This makes both coefficient estimates statistically significant and so they play a role on the literacy rate when the other variable is used in the model as well. Note that just like before, the coefficient estimates for both of these variables are of the opposite sign.

Like in the previous case, several assumptions are made when fitting the model. One assumption that is violated is that the values are fixed and measured without error. This is highly inaccurate because the variable values are not fixed by the researcher nor is it completely reliable. Unlike the previous model, the correlation between the two variables depicts a different image. The correlation between the two explanatory variables in this model is which is moderate. In addition, both variables attain a low variance inflation value of . Keeping in mind that the coefficient signs were also of opposite signs, it will be assumed that both variables are not heavily correlated so that it would have an effect on the model.

After investigating the variables, the error terms can be looked at. Plots for the residuals are shown below.



By transforming the variables, several assumptions can now be made that were not possible before. The normal quantile-quantile plot indicates that the distribution of the error terms is only heavy tailed on one end now, not both, therefore almost satisfying the assumption that the error terms are normally distributed. The error terms can be assumed to have mean now but the variance cannot be assumed to be constant since it grows as fitted values decreases. The relationship between fitted value and square root of standardized residual is a bit more linear than before but the fact still remains; the error term does not have constant variance. The relationship between the independent variables and target variable is also not linear here, as shown in the residuals vs. fitted plot. The square root of the literacy rate for Niger is now a leverage point. Lastly, independence of error terms is investigated using the Durbin-Watson test. At an alpha level of , the Durbin-Watson statistic is and the -value is . This means that the null hypothesis is rejected, that there is evidence of positive residual autocorrelation in the error terms. Further investigation using time series analysis will be good here.

After evaluating this model and checking the regression assumptions, it is seen that the model performs just as well as the previous model. In addition, some of the regression assumptions get partially passed in the second model.

## Discussion

The final model that will be used for understanding literacy rate is the second model with transformed variables. Although both models performed performed similarly, using transformations helped to check off a few more regression assumptions. The final model shows some interesting insights on literacy rate. When mortality rate is constant, an increase in of the number of girls out of primary school is associated with an change in the square root of literacy rate on the average. On the other hand, when number of girls out of primary school is kept constant, an increase of one in the mortality rate is associated with a decrease in the square root of literacy rate on the average. Note that squaring these values will not be helpful; it results in only positive changes for the literacy rate which may be accurate. Now, when there is only girl not in primary school, and the mortality rate is , then the country’s square root of literacy rate is . In my opinion, this does not practical meaning.

This model can be used to explain how a country’s literacy rate is impacted by how its children, namely, girls, are brought up. Social cause organizations such as education for girls in third world country can use this model to help their cause. Another social effect will be, by showing people that increasing mortality of young children affects literacy rate, it will help the country get medical attention so that less children die prematurely due to health reasons. By doing so, there will be more children in school, hopefully less girls not in school, and thus the country will prosper intellectually.

Although careful analysis was made when constructing the models, it can be improved using several methods. More variables could be added to the model. Variable selection using lowest BIC was utilized here; alternatively, using the variables that cause the highest could improve the model and satisfy more assumptions. An outside idea for improvement is that instead of using the multiple linear model framework, use regression splines to divide up the distribution of the literacy rate so that each portion is modelled with less error. This would be useful in this case because the distribution of many explanatory variables are skewed.

## References

The definition of literary rate comes from the glossary of the UNESCO website (Source: “Literacy Rate.” UNESCO UIS, 26 Sept. 2018, uis.unesco.org/en/glossary-term/literacy-rate). Various pieces of data come from the World Bank, an organization that helps developing countries financially and non-financially. The data for the adult literary rates comes from one of their publicly available datasets. Amongst the numerous other datasets they have, ten of which have been narrowed down for use in this study (Source: “Indicators.” Indicators | Data, data.worldbank.org/indicator.).