REGRESSION ANALYSIS PROJECT

VARIABLES PREDICTING LITERACY RATE

AND

WORLD'S LITERACY RATE DISTRIBUTION

CHIA-CHIEN CHOU

1 Introduction

1.1 Background for the study

Literacy is typically described as the ability to read and write. The United Nations Educational, Scientific and Cultural Organization (UNESCO) has drafted a definition of literacy as the "ability to identify, understand, interpret, create, communicate, compute and use printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society."[2] "Since the 1990s, when the Internet came into wide use in the United States, some have asserted that the definition of literacy should include the ability to use tools such as web browsers, word processing programs, and text messages."[2]

Many policy analysts consider literacy rates a crucial measure of a region's human capital. This claim is made on the fact that literate people can be trained less expensively than illiterate people and generally have a higher socio-economic status and enjoy better health and employment prospects. "Human capital refers to the stock of skills and knowledge embodied in the ability to perform labor so as to produce economic value. It is the skills and knowledge gained by a worker through education and experience."[3] Policy makers also argue that literacy increases job opportunities and access to higher education.

A case study from India[4] demonstrated that improvements in female literacy have a direct effect on reducing fertility. Biology researcher has also stated "many studies have shown that people with lower socio-economic status have higher mortality rates", "our health care system places high literacy demands on patients, so limited literacy likely impedes access to health care and chronic disease management. Poor understanding of how to take medication or how to manage chronic disease, not to mention being unable to navigate through the complex health care system, could also cause increased mortality."[5]

As a result, we can conclude that all of the five predictor variables: Birth Rate, Death Rate, GDP-per capita(PPP), Unemployment Rate, and Internet Usage Ratio to some extends have correlation with Literacy Rate either positively or negatively.

1.2 Purpose of the study

This study consists of two parts. The main purposes of part I of this study are: (a) to conduct a statistical analysis of the relationship between the five predictor variables and Literacy Rate, (b) to determine how these five variables are significantly affected by Literacy Rate, and (c) to assess the relative importance of all five variables in the full regression model. The main purposes of part II of this study are: (a) to conduct a statistical analysis of the relationship between the six indicator variables and Literacy Rate, and (b) to study the distribution of Literacy Rate between regions around the World.

2 Dataset

2.1 Data Description & Usage

This study draws upon publicly accessible data from The CIA - World Factbook 2009.

2.1.1 Part I of the study

In part I of this study, Literacy Rate is the response variable. The five predictor variables are: Birth Rate, Death Rate, GDP-per capita(PPP), Unemployment Rate, and Internet Usage Ratio.

- Literacy Rate **-** The rate of age 15 and over can read and write. Y = Literacy (%)

- Unemployment Rate **-** It contains the percent of the labor force that is without jobs. Substantial underemployment might be noted. X_1 = Unemployment Rate (%)

- GDP – per capita (PPP) **-** It gives GDP growth on an annual basis adjusted for inflation and expressed as a percent. X_2 = GDP – per capita (PPP)(%)

- Death Rate **-** It gives the average annual number of deaths during a year per 1,000 populations at mid-year; also known as crude death rate. X_3 = Death Rate (%)

- Internet Usage Ratio **-** It gives the percentage of users within a country that access the Internet. X_4 = Internet Usage (%)

- Birth Rate **-** It gives the average annual number of births during a year per 1,000 persons in the population at midyear; also known as crude birth rate. X_5 = Birth Rate (%)

2.1.2 Part II of the study

In part II of this study, Literacy Rate is the response variable. The six indicator variables are: Asia, Oceania, Europe, South America, North America and Africa represented by combination of five dummy variables are X_1, X_2, X_3, $X_{4,}$ and $X_{5,}$ These variables are different from the five predictor variables described in part I of this study.

- Literacy rate. Y = Literacy Rate

- Six regions around the World represented using dummy variable = (X_1, X_2, X_3, X_4, X_5)

1. Africa
   (X_1, X_2, X_3, X_4, X_5) = (1,0,0,0,0)
2. South America
   (X_1, X_2, X_3, X_4, X_5) = (0,1,0,0,0)
3. Europe
   (X_1, X_2, X_3, X_4, X_5) = (0,0,1,0,0)
4. Asia
   (X_1, X_2, X_3, X_4, X_5) = (0,0,0,1,0)
5. Oceania
   (X_1, X_2, X_3, X_4, X_5) = (0,0,0,0,1)
6. North America
   (X_1, X_2, X_3, X_4, X_5) = (0,0,0,0,0)

2.2 Literature of Analysis Performed

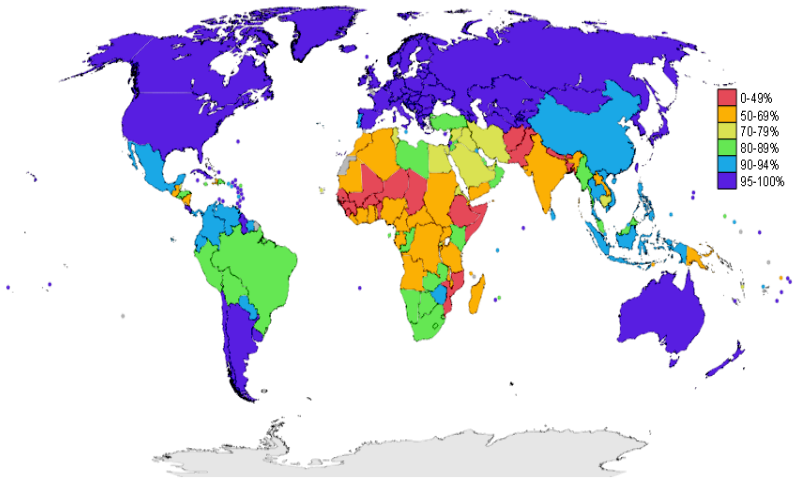
Nowadays, literacy is considered to be a fundamental and essential skill. Literacy skills have been used as both predictors and indicators in various studies. In the introduction section, we have stated that there are research and studies supporting the fact that there is either positive or negative correlation between Literacy Rate and these five variables: Birth Rate, Death Rate, GDP-per capita (PPP), Unemployment Rate, and Internet Usage Ratio.

There is also strong evidence in the literature that suggests that Literacy Rate predicts GDP and Employment Rate in following publications: Lenoir, Gloria., Bellemeur, Jeannette., Illescas-Glascok, Maria Luisa.; and Lim, Soojin. "Factors that Inform International Literacy Rates"[6] and Pant, Mohan. "Does Literacy Predict Economic Growth? A Multiple Regression Analysis"[7]

In current literature indicates that Literacy Rate is related to socio-economic background are key determinants for these five variables: Birth Rate, Death Rate, GDP-per capita (PPP), Unemployment Rate, and Internet Usage Ratio.

This study seeks to examine these variables at the macro-level by looking at the correlation between each of them and Literacy Rate. In addition, it will determine whether or not correlations can be expanded to include the Literacy Rate in each country grouped by region. This study's goal is to add to the literature on educational comparisons among countries by region.

Figure1: World's Literacy Rate Distribution - Literacy rate by country based on CIA World Factbook 2009 data



definition: age 15 and over can read and write
total population: 82%
male: 87%
female: 77%
note: over two-thirds of the world's 785 million illiterate adults are found in only eight countries (Bangladesh, China, Egypt, Ethiopia, India, Indonesia, Nigeria, and Pakistan); of all the illiterate adults in the world, two-thirds are women; extremely low literacy rates are concentrated in three regions, the Arab states, South and West Asia, and Sub-Saharan Africa, where around one-third of the men and half of all women are illiterate (2005 est.)

3 Methodology

3.1 Multiple Regression

Multiple regression analysis involves the formation of an equation with response variable Y and the predictor variables X_i used in the model and then analyzing the significance of each predictor variable in predicting the response variable. The equation for multiple regression model with more than one predictor is given by:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon_i$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters

$X_1, X_2, \ldots, X_{p-1}$ are known constants

ϵ_i are independent $N(0,\sigma 2)$

$i = 1, \ldots, n$

3.2 ANOVA

The analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y.

3.3 Stepwise Regression

Stepwise regression includes regression models in which the choice of predictor variables is carried out by an automatic procedure. Usually, this takes the form of a sequence of F-tests, but other techniques are possible, such as t-tests or Adjusted R-square.

The main approaches are:

- Forward selection, which involves starting with no variables in the model, trying out the predictor variables one by one and including them to test if they are statistically significant.
- Backward elimination, which involves starting with all predictor variables as possible candidate in the model and testing them one by one for statistical significance, deleting any that are not significant.
- Methods that are a combination of the above approaches, testing at each stage for predictor variables to be included or excluded.

4 Analysis

We used R and Minitab to conduct the following analysis.

4.1 Part I of the study

We collected all the data for each of our predictor variables from The CIA – The World Factbook website to form our dataset. We observed that not all countries are covered in each one of the data. As a result, we decided to exclude countries with one or more missing data from the dataset.

After careful selection and elimination, we have finalized our dataset, which contains 171 countries around the World.

4.1.1 Multiple Regression

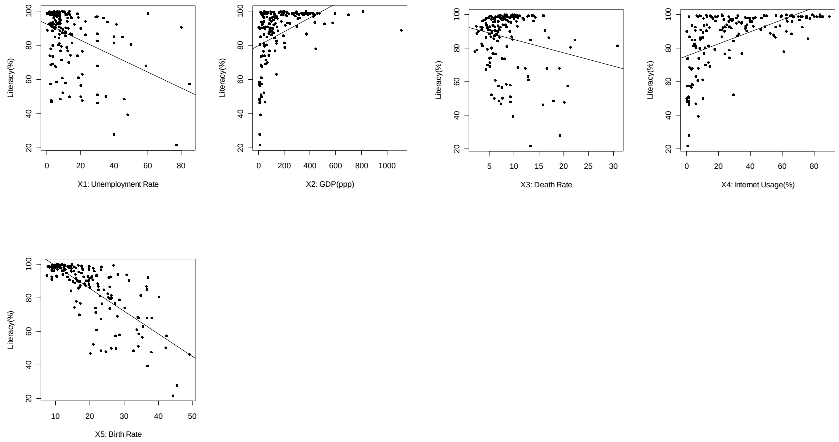
We conducted Multiple Regression Analysis: y versus x1, x2, x3, x4, x5

The regression equation is
y = 111 + 0.0760 x1 + 0.00443 x2 - 0.253 x3 + 0.0323 x4 - 1.30 x5

S = 10.9863 R-Sq = 57.5% R-Sq(adj) = 56.3%

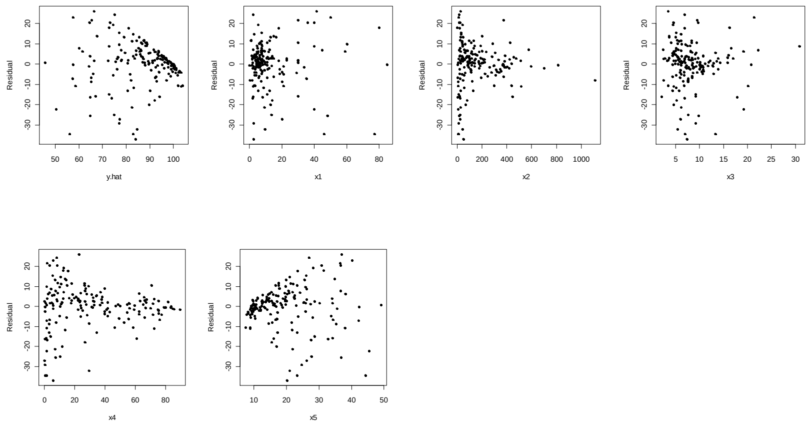
Scatter plot of each predictor variables: Birth Rate, Death Rate, GDP-per capita (PPP), Unemployment Rate, and Internet Usage Ratio against Literacy Rate.

From the plots, we can conclude that there is relationship between Birth Rate and Literacy (%) however we can't say for sure about the other predictor variables.



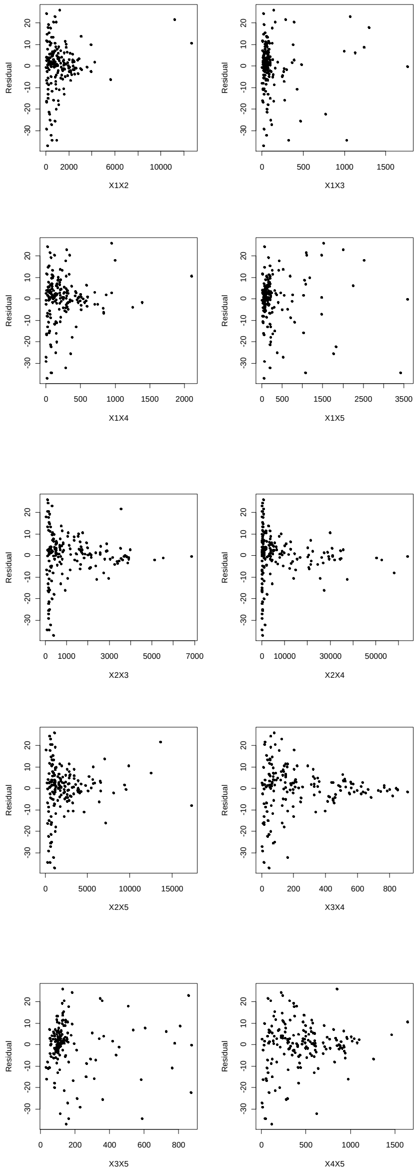
Scatter plot of $\hat{Y}$ against Residual does not suggest any systematic deviations from the response plane, nor that does the variance of the error terms vary with the level of $\hat{Y}$.

Scatter plots of each predictor variables: Birth Rate, Death Rate, GDP-per capita (PPP), Unemployment Rate, and Internet Usage Ratio against Residual; these plots do not show any special pattern, indicating the good fit by the response function and constant variance of the error terms.



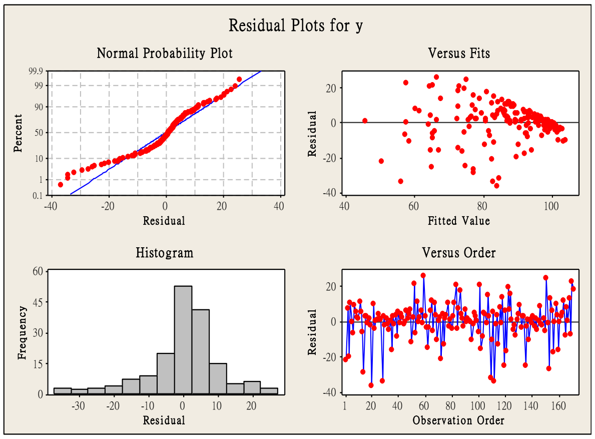
Scatter plot of cross-product term of predictor variables: Birth Rate, Death Rate, GDP-per capita (PPP), Unemployment Rate, and Internet Usage Ratio against Residual.

All of the cross-product term plots do not exhibit any clear systematic pattern; hence, we cannot yet conclude if there is any interaction effects reflected by each of the corresponding model term β X_iXj appear to be present.



Residual plots of Y: Normal Probability Plot of Residual, $\hat{Y}$ (fitted values) against Residual, Histogram of Residual and Observation Order against Residual.

From the Normal Probability Plot, we can observe that the pattern is moderately linear. Histogram chart shows that the residual is normally distributed with mean about 0. This helps to confirm the reasonableness of the conclusion that the error terms are fairly normally distributed.

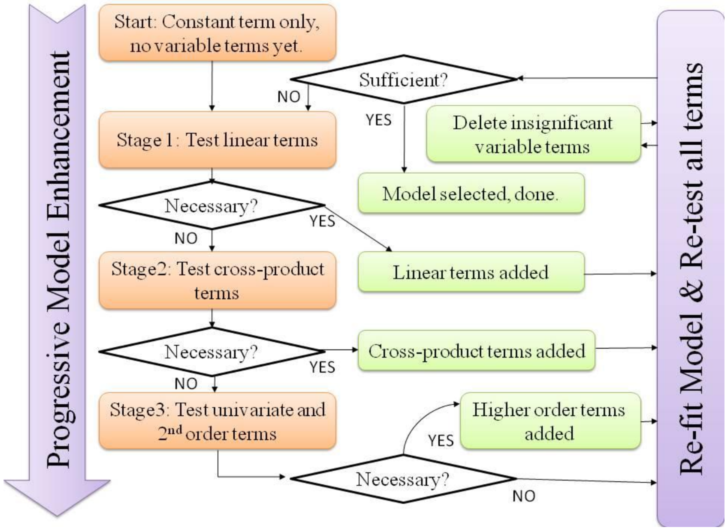


4.1.2 Stepwise Regression

Here is a flowchart of our decision model for Stepwise Regression analysis using forward selection approach with the following condition:

1) To include the term if the alpha value of the new model is less than 0.15

2) To exclude the term if the alpha value of the new model is less than 0.15


Progressive Model Enhancement
Start: Constant term only, no variable terms yet.
Sufficient?
NO
YES
Stage 1: Test linear terms
Delete insignificant variable terms
Model selected, done.
Necessary?
YES
NO
Stage2: Test cross-product terms
Linear terms added
Necessary?
YES
NO
Cross-product terms added
Stage3: Test univariate and 2nd order terms
Higher order terms added
YES
Necessary?
NO
Re-fit Model & Re-test all terms

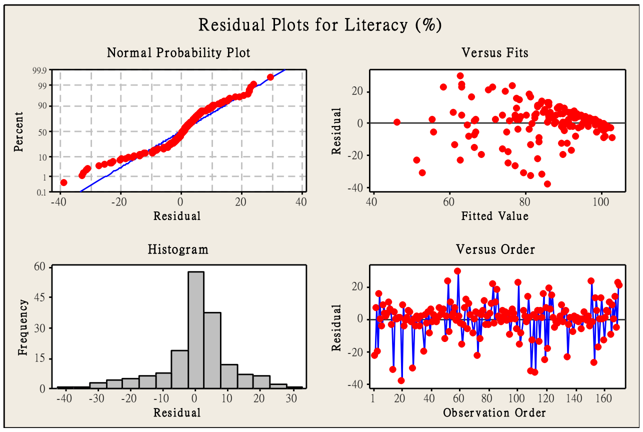
Stage 1: We conducted Stepwise Regression analysis starting with constant term only and no variable terms.

Stage 2: We added test each of the linear terms: x1, x2, x3, x4, and x5; one by one to determine if we will include it into our model.

Here is the result after Stage 2:

```
The regression equation is
Literacy (%) = 113 - 1.35 Birth Rate
```

Note: Statistics Details are in Table 1 of Appendix.

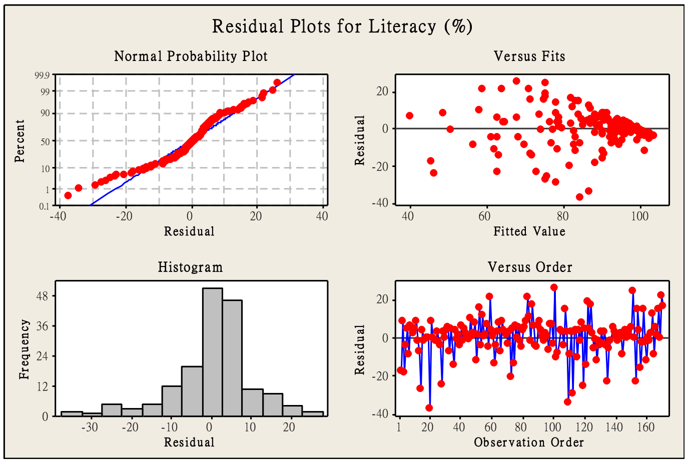


Stage 3: We added test each of the cross-product terms: x1x2, x1x3, x1x4, x1x5, x2x3, x2x4, x2x5, x3x4, x3x5, and x4x5; one by one to determine if we include it into our model.

Here is the result after Stage 3:

```
The regression equation is
Literacy (%) = 113 - 1.53 Birth Rate + 0.0312 x4x5 - 0.646 Internet
User (%) + 0.0305 x3x4 + 0.000964 x2x5
```

Note: Statistics Details are in Table 2 of Appendix.

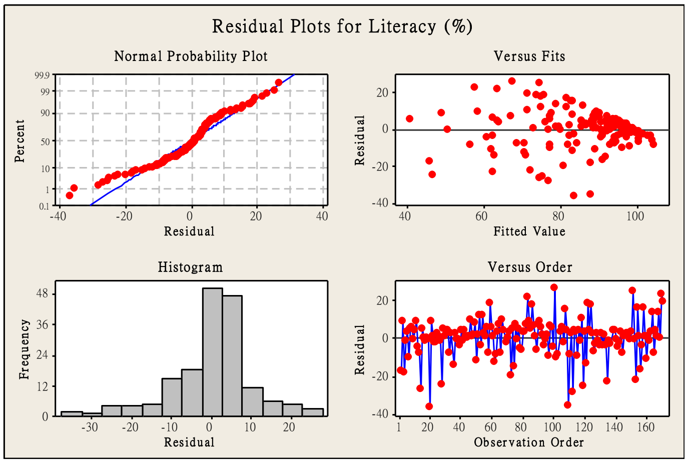


Stage 4: We added test each of the 2nd order terms: x1^2, x2^2,x3^2, x4^2, and x5^2; one by one to determine if we include it into our model.

Here is the result after Stage 4:

```
The regression equation is
Literacy (%) = 110 - 1.46 Birth Rate + 0.0317 x4x5 + 0.00494 x2x3 -
0.000323 x2x4 - 0.395 Internet User (%)
```

Note: Statistics Details are in Table 3 of Appendix.

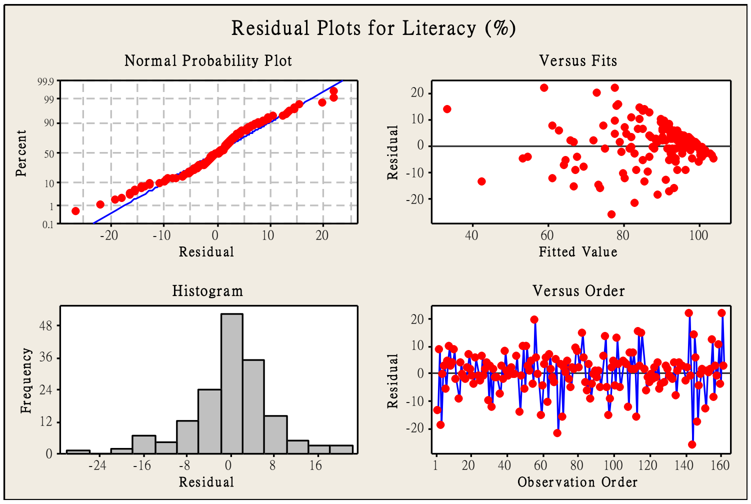


Outlier:

We observed that there are some outliers where these observations which have a large standardized residual. As a result, we are eliminating these observations from our dataset. Again, we conducted Multiple Regression analysis with the selected terms with the new dataset.

```
The regression equation is
Literacy (%) = 91.8 - 0.0222 x5^2 + 0.00140 x1x2 + 0.00350 x1^2 +
0.0150 x4x5 - 0.00341 x4^2 + 0.0253 x3x4 - 0.00795 x1x4 - 0.0129 x3x5
```

Note: Statistics Details are in Table 4 of Appendix.



4.1.3 ANOVA

ANOVA output for Multiple Regression Analysis: y versus x1, x2, x3, x4, x5

The regression equation is
y = 111 + 0.0760 x1 + 0.00443 x2 - 0.253 x3 + 0.0323 x4 - 1.30 x5

S = 10.9863   R-Sq = 57.5%   R-Sq(adj) = 56.3%

ANOVA output for Multiple Regression Analysis: y versus Birth Rate, x4x5, x2x3, x2x4, Internet User

The regression equation is
Literacy (%) = 110 - 1.46 Birth Rate + 0.0317 x4x5 + 0.00494 x2x3 - 0.000323 x2x4 - 0.395 Internet User (%)

S = 10.2059 R-Sq = 63.4% R-Sq(adj) = 62.3%

ANOVA output for Multiple Regression Analysis without outliers: y versus x5^2, x1x2, x1^2, x4x5, x4^2, x3x4, x1x4, x3x5

The regression equation is
Literacy (%) = 91.8 - 0.0222 x5^2 + 0.00140 x1x2 + 0.00350 x1^2 + 0.0150 x4x5 - 0.00341 x4^2 + 0.0253 x3x4 - 0.00795 x1x4 - 0.0129 x3x5

S = 7.78190 R-Sq = 71.9% R-Sq(adj) = 70.4%

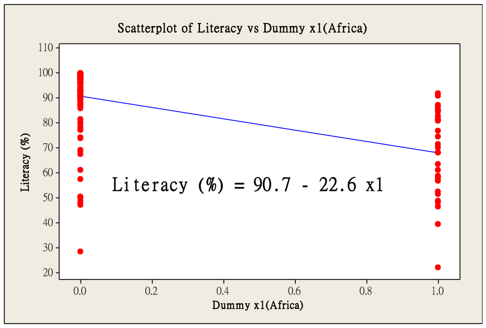
#4(TO DO: Comparison on ANOVA output) eg Adjusted R square…..

4.2 Part II of study

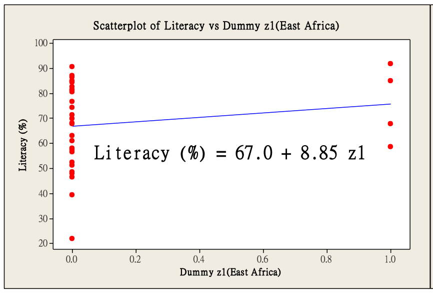
4.2.1 Multiple Regression

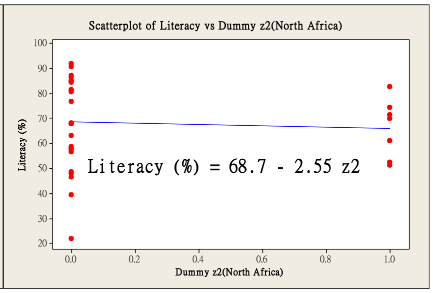
Taking this analysis further, we categorized the data into 6 groups, namely Asia, Oceania, Europe, South America, North America and Africa. From the Chart below, we noticed that the Literacy rate is lowest in Africa and followed by Asia. As a result, we will do further analysis in these two regions to see how it's Literacy Rate compares to the average of all other regions and how Literacy rate is distributed within these regions itself.

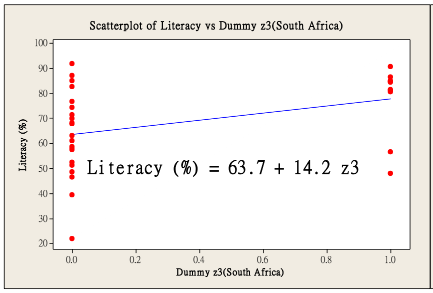
Regions	Average Literacy Rate
Asia	84
Oceania	90.07
Europe	98,44
South America	92.99
North America	91.83
Africa	70.30

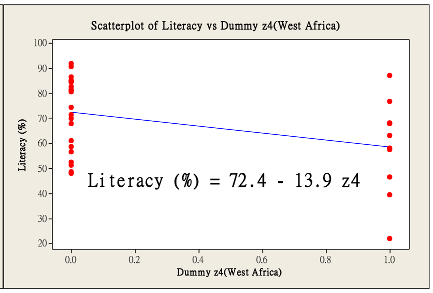


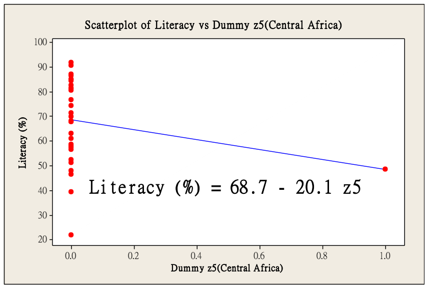
Literacy Rate in Africa is lower than the average of all other regions.

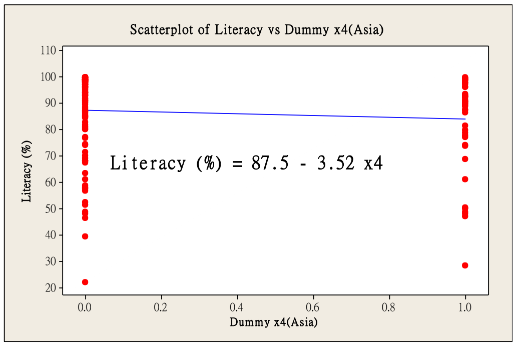




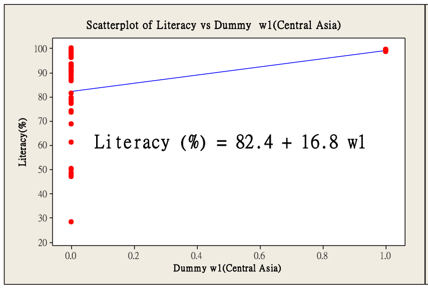


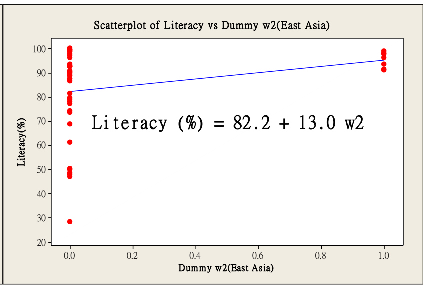


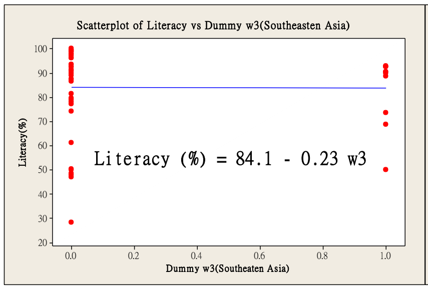


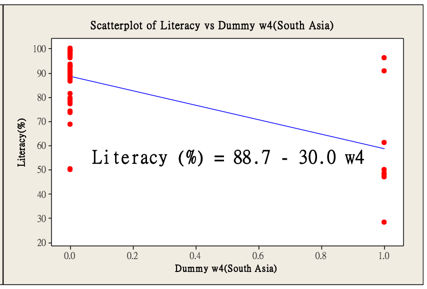


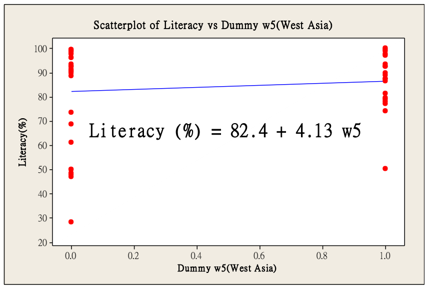
Literacy Rate in Asia is lower than the average of all other regions.






Scatterplot of Literacy vs Dummy w3(Southeasten Asia)
Literacy(%)
100
90
80
70
60
50
40
30
20
Literacy (%) = 84.1 - 0.23 w3
0.0
0.2
0.4
0.6
0.8
1.0
Dummy w3(Southeaten Asia)


Scatterplot of Literacy vs Dummy w4(South Asia)
Literacy(%)
100
90
80
70
60
50
40
30
20
Literacy (%) = 88.7 - 30.0 w4
0.0
0.2
0.4
0.6
0.8
1.0
Dummy w4(South Asia)


Scatterplot of Literacy vs Dummy w5(West Asia)
Literacy(%)
100
90
80
70
60
50
40
30
20
Literacy (%) = 82.4 + 4.13 w5
0.0
0.2
0.4
0.6
0.8
1.0
Dummy w5(West Asia)