LITERACY RATE PREDICTION

ZEYANG GONG

Executive Summary

There has been an increasing notice on the education level among all the countries in the world. Literacy rate is one of the measurements that can be used to measure the education level of a country. And people are more interested in find out what factors will have association with the literacy rate. This study investigates a number of factors that may have relationship with literacy rate of a country. The study uses multiple regression model to analyze many different indicators. As a result, three components are selected from initial indicators. The income level, the labor force per capita and the unemployment rate all have positive relationship with the literacy rate of a country.

Data

The data is downloaded from the Word Bank dataset. It initially has 263 countries entry, but 67 was deleted deal to the missing value of the response variable, and 3 was deleted because of too many missing value in explanatory variables. The dataset contains data from 2004-2015. The response variable 'Literacy Rate' is the average value of 2010-2015 literacy rate of the country. All the explanatory variables are the average value of themselves from 2004 to 2014. The reason why the dataset is aggregate in the way is because there are too many missing data in one specific year, and also the mean value is a fair way to represent the variable while investigate the dataset.

The initial dataset contains one response variable, one categorical explanatory variable, and ten numeric explanatory variables.

Response variable: "Literacy", "Literacy3"

Adult literacy rate, population 15+ years, both sexes (100%)

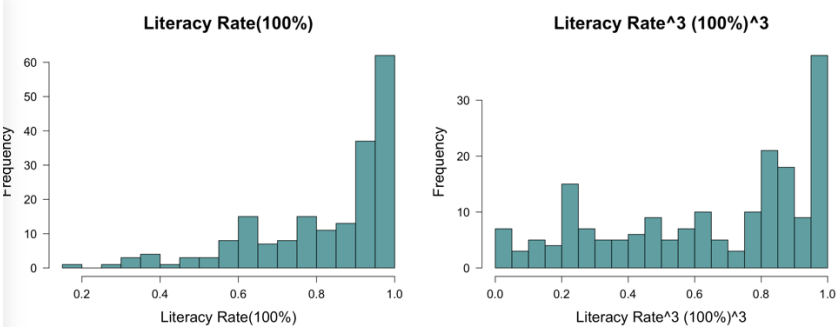


Figure 1- Histogram of "Literacy", "Literacy3" distribution

In the left graph, it is obvious that the data is extremely left-skewed. After tried out many transformations, the cube of the "Literacy" is used as response variable.

Explanatory variables:

1. Income Group: "IncomeGroup"

Income Group is in group defined by the world bank, to represent the income level of a country, which indicates the developing level of a country. There are four income groups, “Low income”, “Lower middle income”, “Upper middle income”, “High income”. In the dataset, the 4 groups are coded as “1”, “2”, “3”, “4”, respectively. In the boxplot, we can roughly see a positive relationship between IncomeGroup and Literacy Rate.

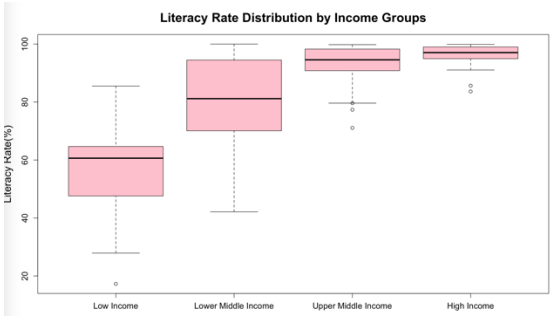


Figure 2- Boxplot of Income Group with Literacy Rate

2. GDP (current US$) (in Billion $)

GDP is the total gross domestic product produced of a country in a year. It is an indicator of the developing level of a country.

3. GDP per capita (current US$): “GDPPC” ($)

GDP per capita is gross domestic product divided by midyear population. It is included because the Literacy Rate may not have correlation with the total GDP, but with the GDP per capita. We will further investigate in the analysis.

4. Government expenditure on education, total (% of GDP): “GEOG” (%)

General government expenditure on education (current, capital, and transfers) is expressed as a percentage of GDP. How sportive the government is on the education may have impact on the literacy rate.

5. Government expenditure per student, secondary (% of GDP per capita): “GEOGPS” (%)

Government expenditure per student is the average general government expenditure (current, capital, and transfers) per student in the given level of education, expressed as a percentage of GDP per capita.

6. Expenditure on education as % of total government expenditure: “EEOGE” (%)

Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.).

7. Labor force, total: “LaborForce” (in Million People)

Total labor force comprises people ages 15 and older who meet the International Labor Organization definition of the economically active population. Labor force level may have some impact on the literacy rate.

8. Labor force per capita: “LaborForcePC” (Mppl/100ppl%)

Total Labor force people in 100 people divided by the total population.

9. Unemployment, total (% of total labor force): “Unemployment” (%)

Unemployment refers to the share of the labor force that is without work but available for and seeking employment.

10. Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population): “Poverty” (%)

Poverty headcount ratio at $1.90 a day is the percentage of the population living on less than $1.90 a day at 2011 international prices.

11. GINI index: “GINI”

Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. It represents the degree of inequality income distribution of a country.

Summary Statistics

```
      GDP                GDPPC             GEOG            GEOGPS          EEOGE
 Min.   :    0.160   Min.   :  199.3   Min.   : 0.810   Min.   : 4.43   Min.   : 3.93
 1st Qu.:    9.395   1st Qu.: 1172.6   1st Qu.: 3.225   1st Qu.:13.21   1st Qu.:12.04
 Median :   37.970   Median : 3503.0   Median : 4.140   Median :17.45   Median :15.07
 Mean   : 1197.245   Mean   : 6881.3   Mean   : 4.322   Mean   :20.05   Mean   :15.20
 3rd Qu.:  335.775   3rd Qu.: 7827.5   3rd Qu.: 5.120   3rd Qu.:23.84   3rd Qu.:17.92
 Max.   :20716.620   Max.   :73758.8   Max.   :13.730   Max.   :68.79   Max.   :28.25
                                       NA's   :21       NA's   :69      NA's   :22
   LaborForce        LaborForcePC     Unemployment       Poverty           GINI
 Min.   :   0.040   Min.   :22.92   Min.   : 0.600   Min.   : 0.04   Min.   :21.55
 1st Qu.:   1.570   1st Qu.:37.98   1st Qu.: 5.298   1st Qu.: 2.62   1st Qu.:34.55
 Median :   5.815   Median :43.18   Median : 7.115   Median :12.73   Median :40.72
 Mean   : 147.154   Mean   :42.96   Mean   : 8.669   Mean   :20.95   Mean   :41.13
 3rd Qu.:  38.860   3rd Qu.:47.40   3rd Qu.:10.867   3rd Qu.:36.40   3rd Qu.:46.23
 Max.   :2623.950   Max.   :70.43   Max.   :32.970   Max.   :85.57   Max.   :63.73
 NA's   :6          NA's   :6       NA's   :10       NA's   :61      NA's   :77
```

Figure 3- Summary Statistics of Numeric Explanatory Variables

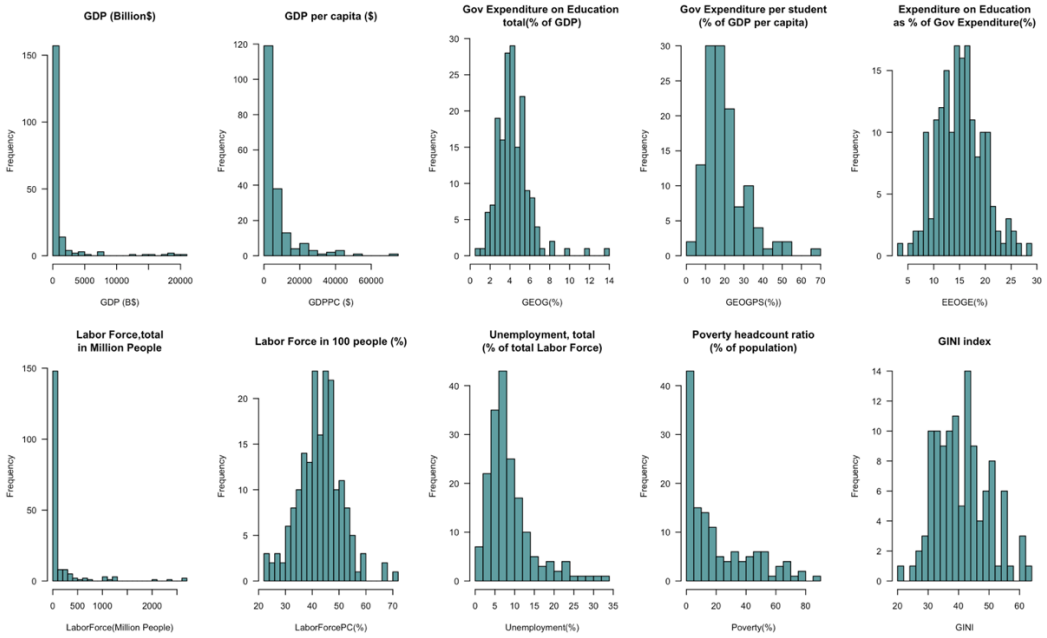
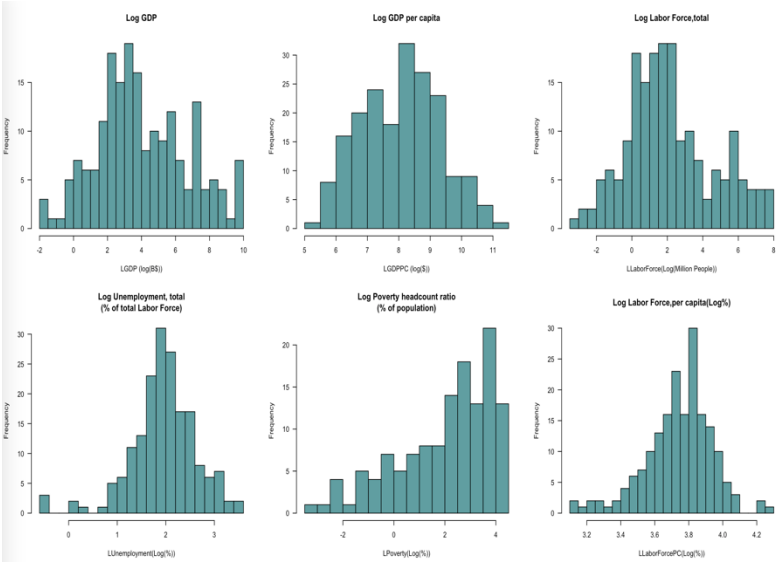


Figure 4 – Histogram of Numeric Explanatory Variables



Correlation:

	Literacy	Literacy3
Literacy	1.000000000	0.96835658
GDP	0.104090873	0.07908777
GDPPC	0.416191985	0.45167966
→ GEOG	0.207840978	0.18495036
GEOGPS	-0.174781249	-0.14903094
EEOGE	-0.107380434	-0.15263370
LaborForce	0.005188298	-0.03932278
→ LaborForcePC	0.410188966	0.43744694
Unemployment	0.172969813	0.16364266
→ Poverty	-0.707091616	-0.73356757
GINI	-0.100335818	-0.18118792
Literacy3	0.968356576	1.00000000
LGDP	0.148701966	0.10451275
→ LGDPPC	0.714238167	0.73643304
LLaborForce	-0.136613452	-0.18489154
→ LUnemployment	0.200568236	0.19044029
LPoverty	-0.668964638	-0.73804998
→ LLaborForcePC	0.399751195	0.42360277

Figure 5 – Histogram of Numeric Explanatory Variables after Log Transformation

Initial Variable Selection and Transformation

In Figure 5, “GDP”, “GDPPC”, “LaborForce”, “LaborForcePC”, “Unemployment”, “Poverty” were transformed using log function, because they are right-skewed.

In Figure 3, GEOGPS, Poverty, GINI have too many missing data. Try not to use them unless necessary. As shows in Figure 6, based on the correlation with Response variable, GEOGPS and GINI will be excluded, because they have low correlation. Although Poverty has a high correlation with response variable, it also has a high correlation with GDPPC, which means Poverty can be explained by GDP per capita. Therefore, we exclude Poverty, GINI, GEOGPS from the model.

Initial select some variables based on correlations: “IncomeGroup”, "GEOG", "LaborForcePC", "LGDPPC", "LUnemployment", “LaborForcePC”. Then delete the entries that has missing values on these variables. The dataset moving forward has 163 entries.

Variables	VIF
IncomeGroup	8.090143
GEOG	1.077972
LaborForcePC	53.301132
LLaborForcePC	51.470763
LGDPPC	8.206287
LUnemployment	1.357886

Figure 6 – VIF table of explanatory variables after initial selection

Note that LaborForcePC and LLaborForcePC cannot be used together, because they are highly correlated. Will test to include which one in the model. Same reason with “IncomeGroup” and “LGDPPC”. Both of the two pairs of variable have high VIF value, as shows in Figure 7. This will be handled in the next section.

Methods

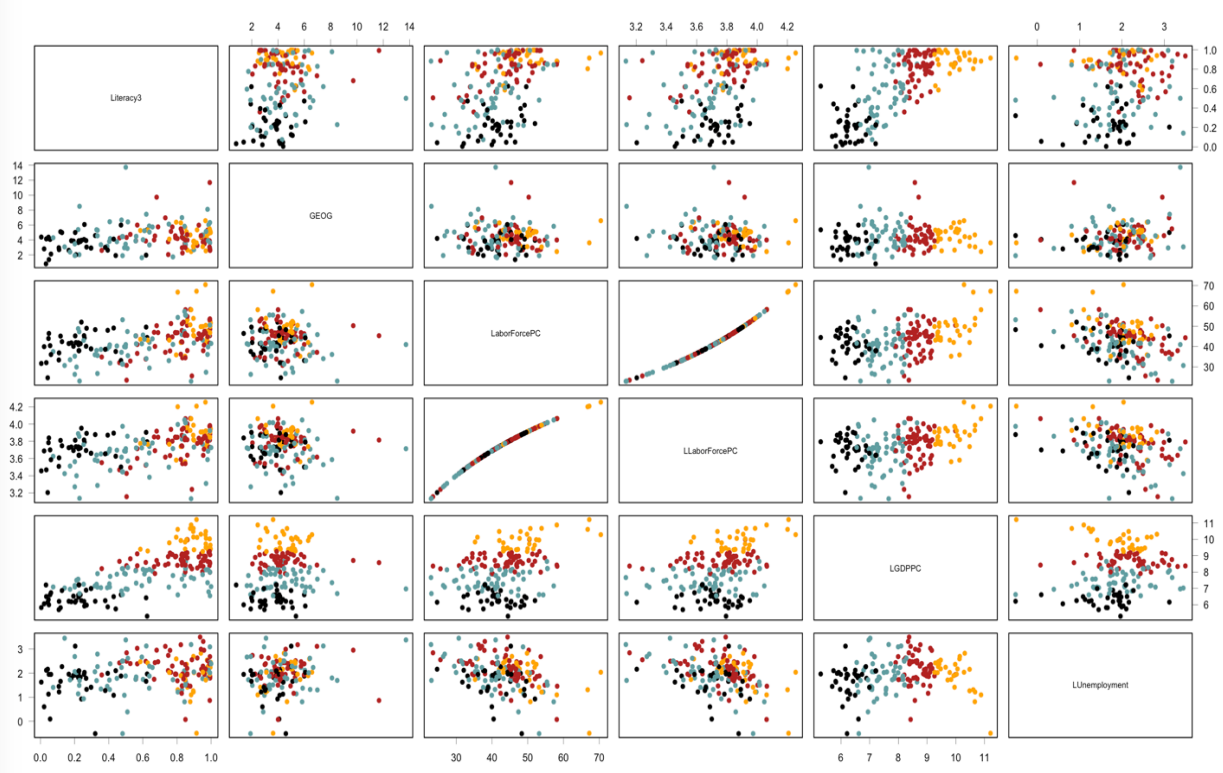


Figure 7 - Pair Scatter Plot of selected variables

Try to build the best model for LGDPPC and IncomeGroup, respectively. And then compare the two model, in order to select a better one for this dataset.

###Best model for LGDPPC###

Use step wise forward method to find the best model for LGDPPC. The R function gives the Cp values of 42.8, 17.2, 5.3, 5.0. Therefore, pick the model for 3 variables and 4 variables, then preform a Partial F-test on them. The p-value of the ANOVA analysis is 0.13, which is large than 0.05. So, null hypothesis is not rejected. The last variable "GEOG" is not helpful to the model, so it will not be included in the model lm.gdp.

Now, we want to determine whether we should put LaborForcePC or Log form of LaborForcePC in the model. The statistics are as follows.

Model	PRESS	RMSE_jackknife	RMSE
Lm.3	5.33	0.182	0.178
Lm.log	5.21	0.182	0.176

There is no much different. Therefore, for interpretability, choose LaborForcePC moving forward.

The final model for LGDPPC is as follow:

lm.gdp = lm(Literacy3~ LaborForcePC + LGDPPC + LUnemployment, data=x2)

```
Call:
lm(formula = Literacy3 ~ LaborForcePC + LGDPPC + LUnemployment,
    data = x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37146 -0.12583 -0.00562  0.11282  0.56966

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.307513   0.115243 -11.346  < 2e-16 ***
LaborForcePC   0.014330   0.002266   6.323 2.48e-09 ***
LGDPPC         0.141710   0.012364  11.462  < 2e-16 ***
LUnemployment  0.087635   0.023609   3.712 0.000284 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.178 on 159 degrees of freedom
Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6662
F-statistic: 108.8 on 3 and 159 DF,  p-value: < 2.2e-16
```

Figure 8 – lm.gdp regression summary

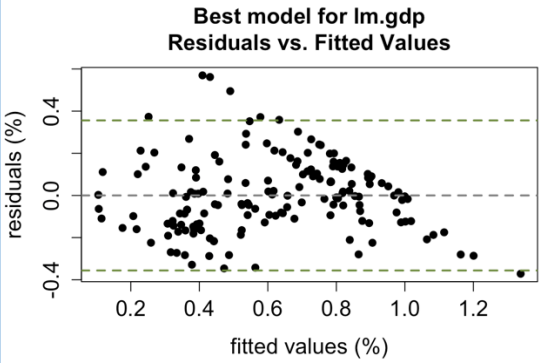


Figure 9 – lm.gdp residual plot

Best model for IncomeGroup

The same approaches are applied in finding the best model for IncomeGroup. The best model is as follows:

lm.income <- lm(Literacy3~IncomeGroup + LaborForcePC + LUnemployment, data=x2)

```
Call:
lm(formula = Literacy3 ~ IncomeGroup + LaborForcePC + LUnemployment,
    data = x2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41131 -0.09230 -0.00967  0.09216  0.50416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.54632    0.10852  -5.034 1.30e-06 ***
IncomeGroup2   0.31863    0.03556   8.959 9.02e-16 ***
IncomeGroup3   0.49648    0.03815  13.015  < 2e-16 ***
IncomeGroup4   0.52830    0.04504  11.729  < 2e-16 ***
LaborForcePC   0.01601    0.00205   7.808 7.71e-13 ***
LUnemployment  0.06894    0.02176   3.168  0.00184 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1612 on 157 degrees of freedom
Multiple R-squared:  0.7346,    Adjusted R-squared:  0.7261
F-statistic: 86.91 on 5 and 157 DF,  p-value: < 2.2e-16
```

Figure 10 – lm.income regression summary

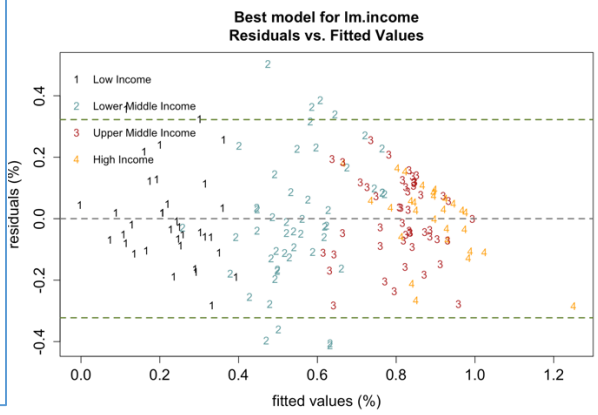


Figure 11– lm.income residual plot

From Adjusted R^2 and the residual plot we can see lm.income is a better model. Let's further validate using some statistics.

Model validation:

Compare lm.gdp with lm.income

Because lm.gdp and lm.income are not nested models, partial F-test cannot be performed here.

Model	PRESS	RMSE_jackknife	RMSE
lm.gdp	5.33	0.182	0.178
lm.income	4.39	0.167	0.161

All the statistics shows lm.income is a better model. Therefore, choose lm.income moving further.

Check Assumptions

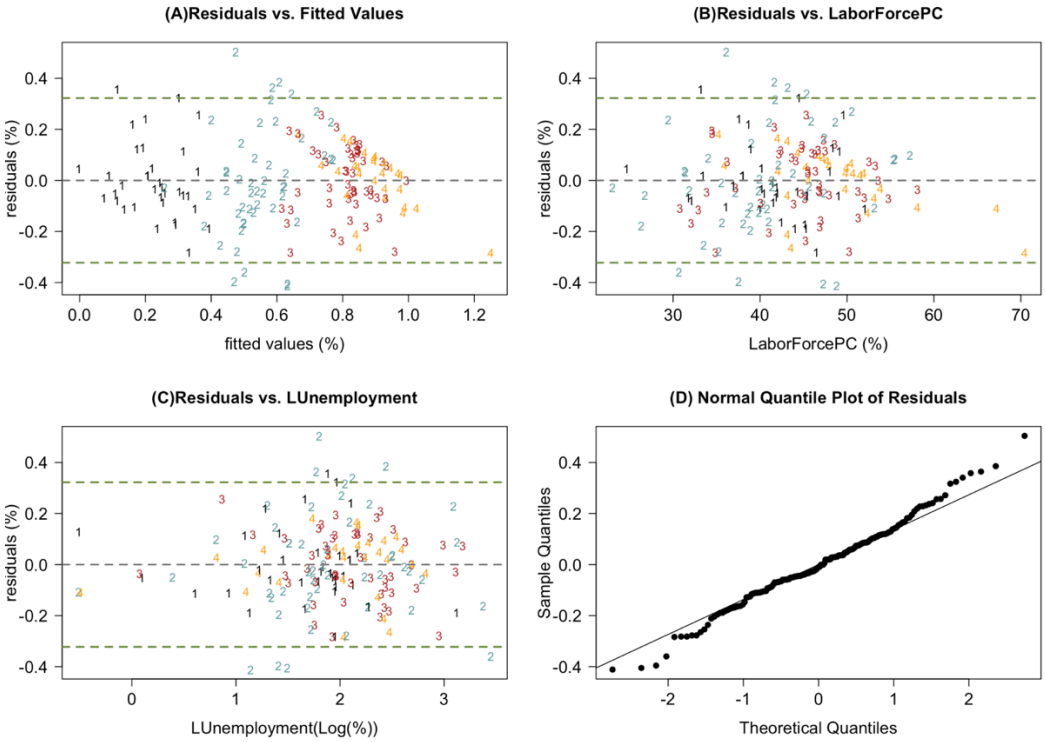


Figure 12 – Residual plots and Normal Quantile plot

1. Measurement Accuracy
   We assume that the World Bank collected the data in a fair way.
2. Linearity
   In Figure (A), (B) and (C), there are no curves, no trends. The linearity assumption is met.
3. Constant error Variance
   In (A), (B) and (C), the points are roughly clustered in the center. This assumption is fine.
4. Normality of residual
   In (D), the QQ plot, there is a little S-shape in the figure, which means the data is heavy tail. But it is fine for the analysis.
5. Independence of residual
   Because the data is not time series, it is the aggregation of 10 years. Therefore, this assumption is met.
6. Multicollinearity
   For all the variables, there is no vif value which is greater than 5. Therefore, no multicollinearity issue in the model.
7. Outliers, influential points
   Use the cook.distance measure. No influential points.

Results

Final regression model

"Low Income" countries:

$Literacy^3 = -0.54(100\%^3) + (100\%^3)/(Mppl/100ppl*\%)0.016*LaborForcePC + (100\%^3) /(\log(\%)) 0.06*Log(Unemployment)$

"Lower Middle Income" countries:

$Literacy^3 = -0.23(100\%^3) + (100\%^3)/(Mppl/100ppl\%)0.016*LaborForcePC + (100\%^3) /(\log(\%)) 0.06*Log(Unemployment)$

"Upper Middle Income" countries:

$Literacy^3 = 0.26(100\%^3) + (100\%^3)/(Mppl/100ppl\%)0.016*LaborForcePC + (100\%^3) /(\log(\%)) 0.06*Log(Unemployment)$

"High Income" countries:

$Literacy^3 = 0.78(100\%^3) + (100\%^3)/(Mppl/100ppl\%)0.016*LaborForcePC + (100\%^3) /(\log(\%)) 0.06*Log(Unemployment)$

Result discussion

With 1 person of labor force increases in 1000 people of a country, the cube of literacy rate will increase by of 1.6%. With 1 percent of unemployment rate increases in the country, the cube of the literacy rate will increase by 6%. $-0.54(100\%^3)$ is the value of cube literacy rate for a 0 people of labor force and 0 people of unemployment for low income countries. The y-intercept is an extrapolation. For lower middle income countries, an extra $0.31(100\%^3)$ is added on the cube of literacy rate. For upper middle income countries, an extra $0.49(100\%^3)$ is added on the cube of literacy rate. For high income countries, an extra $0.52(100\%^3)$ is added on the cube of literacy rate.

High literacy rate countries are associate with higher labor force per capita and higher income level. It is surprisingly to see that unemployment rate has a positive relationship with literacy rate for a country. The reason may be that higher unemployment rate means the country has better social welfare system, which allows more people to stay unemployed while having a normal life condition. The unemployed people doesn't have to be low-educated people, they may just choose not to work.

Improvement

Distribution of Response variable: heavily left-skew, have to use cube to transform y data, which makes it very hard to interpret. Better way of transformation will help to generate better result.

References

http://data.worldbank.org/indicator