Assignment 3

Yuwen(Suyi)Wu 5:45-7:45

3/28/2019

i. Executive Summary:

In this research, we will present a possible model of the literacy rate across countries at year 2014 in order to better understand which social economic factors may be related to the literacy rate. The process of this research includes data searching, raw data clean, data transformation, multiple linear regression model building, assumption validation and hypothesis test based on the statistical software R. We could see the conclustion that Female Unemployment rate and Internet using rate are significant related to a country's literacy rate. The dataset is based on World Bank(http://data.worldbank.org/indicator).

ii. Introduction:

The data of Literacy rate we focused on is based on the adult total (% of people ages 15 and above), which is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life.

Why should we study the literacy rate? From the World Bank Website, Literacy rate is an outcome indicator to evaluate educational attainment. Also, this indicator could be used to forecast future educated labor force since country with high educated labor force has higher probability to bring high GDP and high technology improvement. Additional, with the process of globalization, a country with high litteracy rate could effectively participate global issues and gobal markets, which will bring higher interaction between countries. Thus, understanding which factors would affect the literacy rate is helpful for a country and also for culture development.

iii. Data:

My data is provided by the World Bank(http://data.worldbank.org/indicator). In this website, I selected 10 factors which I think are related to literacy rate.

Literacy rate, adult total (% of people ages 15 and above): Adult literacy rate is the percentage of people ages 15 and above who can both read and write with understanding a short simple statement about their everyday life. This is this research's and below models' response variable, which means we want to find factors to explain it.

Individuals using the Internet (% of population): Internet users are individuals who have used the Internet (from any location) in the last 3 months. Internet is a tool to connect to the world. Also, individials who could use internet has high possibility could read and understand news and other infomation around their daily life.

GDP per capita (current US$): GDP per capita is gross domestic product divided by midyear population. This indicator is an popular indicator to estimate or represent economy growth. To some degree, the level of economy growth could connected to a economic entity's literacy rate because it related to the education investment and degree of a country's development.

Immunization, DPT (% of children ages 12-23 months): Child immunization, DPT, measures the percentage of children ages 12-23 months who received DPT vaccinations before 12 months or at any time before the survey. Immunization rate could be an indicator connected to labor force and popualtion. Also higher immunization rate could represent higher civilization degree.

Prevalence of HIV, total (% of population ages 15-49): Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV. Since HIV related to sex education, this might related to Literacy rate. With high rate of Literacy rate, people have more possibility to access the sex education and other health related education.

Population growth (annual %): Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. This is an transparent indicator connected to labor force and population, which, in my opinion, could be used to explain Literacy rate since literacy rate is an percentage of population.

Mortality rate, infant (per 1,000 live births): Infant mortality rate is the number of infants dying before reaching one year of age, per 1,000 live births in a given year. This indicator related to medical level which could be considered as an indicator of economy and civilization, and population, thus it could be connected to Literacy rate.

School enrollment, primary (% gross): Gross enrollment ratio is the ratio of total enrollment, regardless of age, to the population of the age group that officially corresponds to the level of education shown. The primary education is the best way to improve literacy rate of a country because the context of primary school is to educate student understand a short simple statement about their everyday life.

High-technology exports (% of manufactured exports): High-technology exports are products with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery. This indicator could explain total country's education level and also related to investment on Research & Development, which are highly related to population of enginnering and scientists. Therefore, it could be used as explainary factor of Literacy Rate.

Unemployment, female (% of female labor force) (modeled ILO estimate): Unemployment refers to the share of the labor force that is without work but available for and seeking employment. From the result of research paper of Amir H. Mehryar,Akbar Aghajanian, Mohamad Tabibian and Farzaneh Tajdini (Women's Education and Labor Force Participation and Fertility Decline in Iran), improvement in the level of education of women has a clearly negative correlation to fertility. Also, high level of education of women correlated to low unemployment rate of female. Therefore, in my mind, this could be an explainary factor for Literacy rate.

Rural population (% of total population):Rural population refers to people living in rural areas as defined by national statistical offices. In most of countries, people live in the city would have higher education level and by contract to the people live in the rural area. Thus, I select this factor to see whether location connected to Literacy Rate.

Summary Statistics:

Since data set of year 2014 has more full data of these 11 variables, I choose year 2014 for further model building.

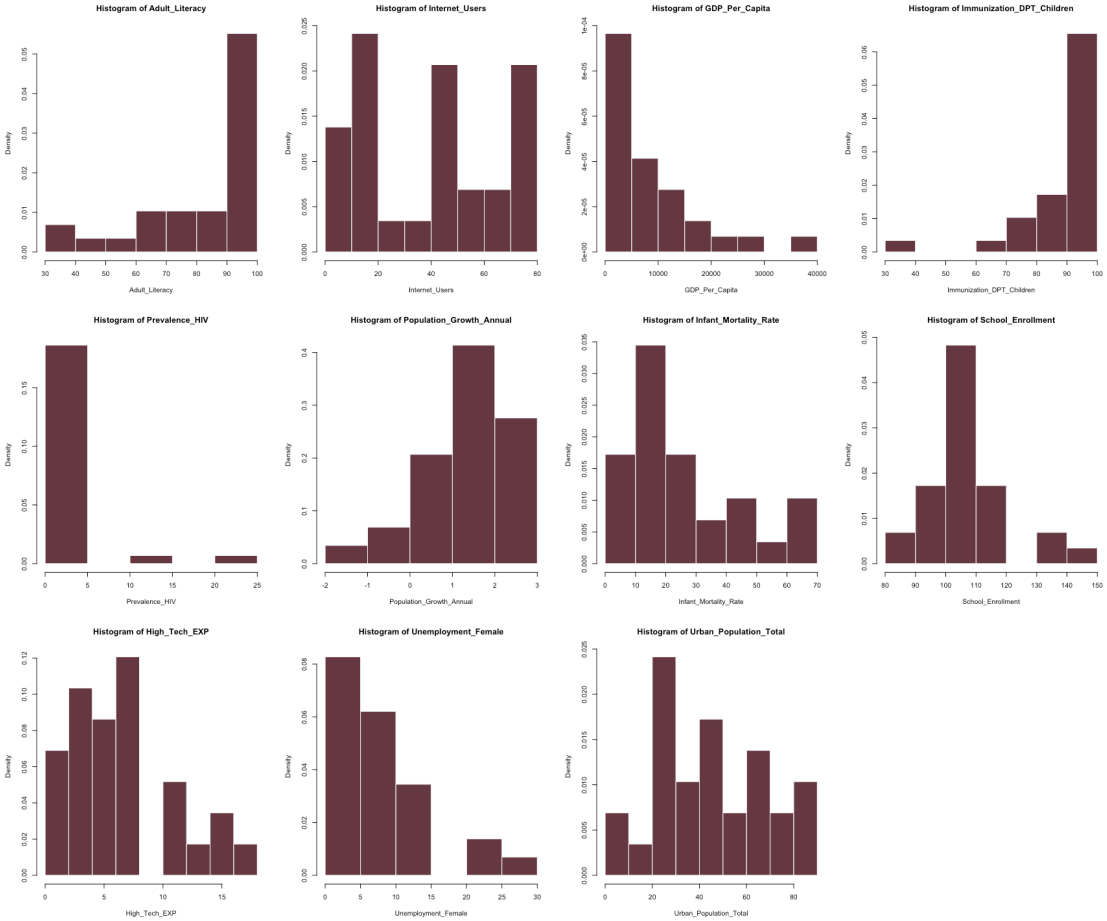
After I cleaned the data and remove all the NA of each line, only 41 countries with full dataset left.

```
##
##          Argentina         Azerbaijan           Barbados
##                  1                  1                  1
##           Botswana             Brazil       Burkina Faso
##                  1                  1                  1
##            Burundi           Cambodia           Colombia
##                  1                  1                  1
##      Cote d'Ivoire Dominican Republic            Ecuador
##                  1                  1                  1
##     European Union            Georgia          Guatemala
##                  1                  1                  1
##             Guinea           Honduras            Hungary
##                  1                  1                  1
##          Indonesia             Malawi             Mexico
##                  1                  1                  1
##           Pakistan               Peru             Rwanda
##                  1                  1                  1
##           Slovenia         South Asia              Spain
##                  1                  1                  1
##            Tunisia            Uruguay
##                  1                  1
```

From above table, we can see that lots of big countries are moved from our data. There are important countries missing from our data and further models. From above, some countries are crowded in one geographical position. This may affect our model since data may not random selected.

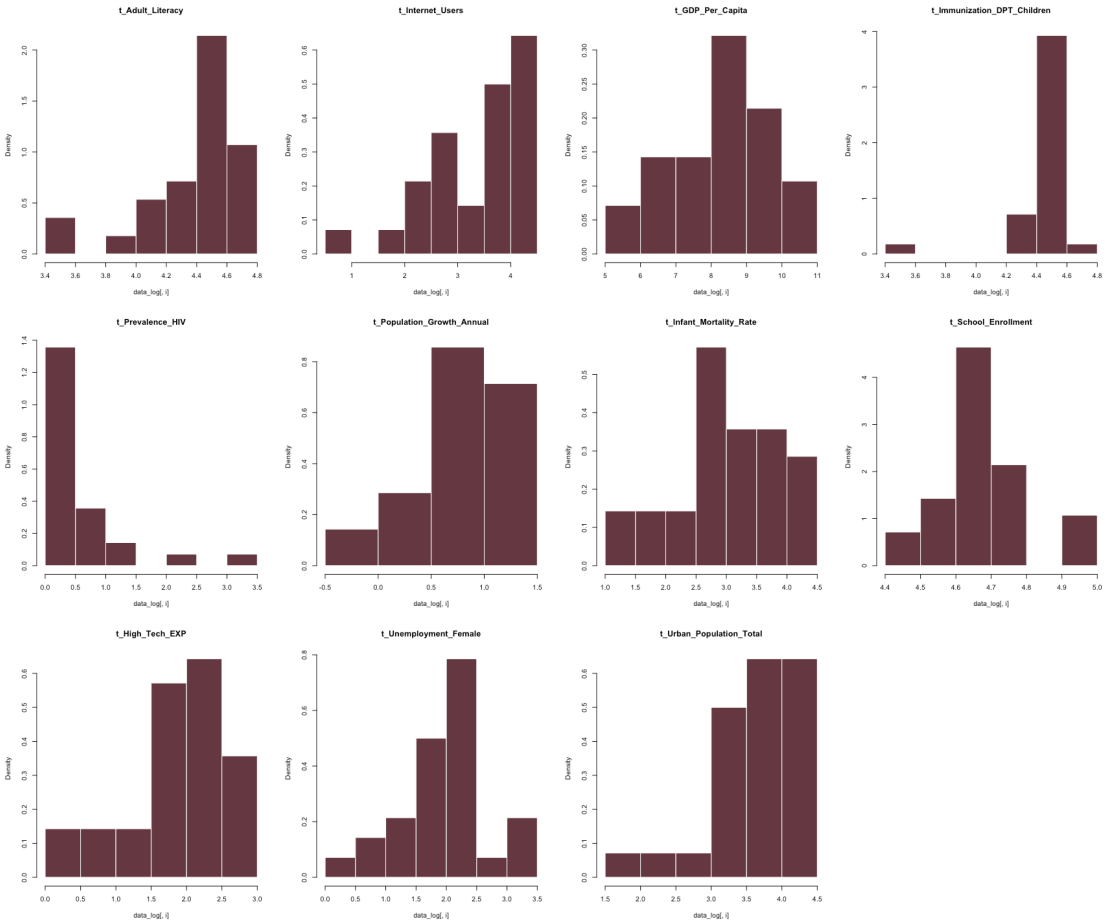
iv. Methods:

First, check for the distribution of 11 variables. We could understand whether they are normal distribution in order to do more data transformation and multi-factor regression in the next step.

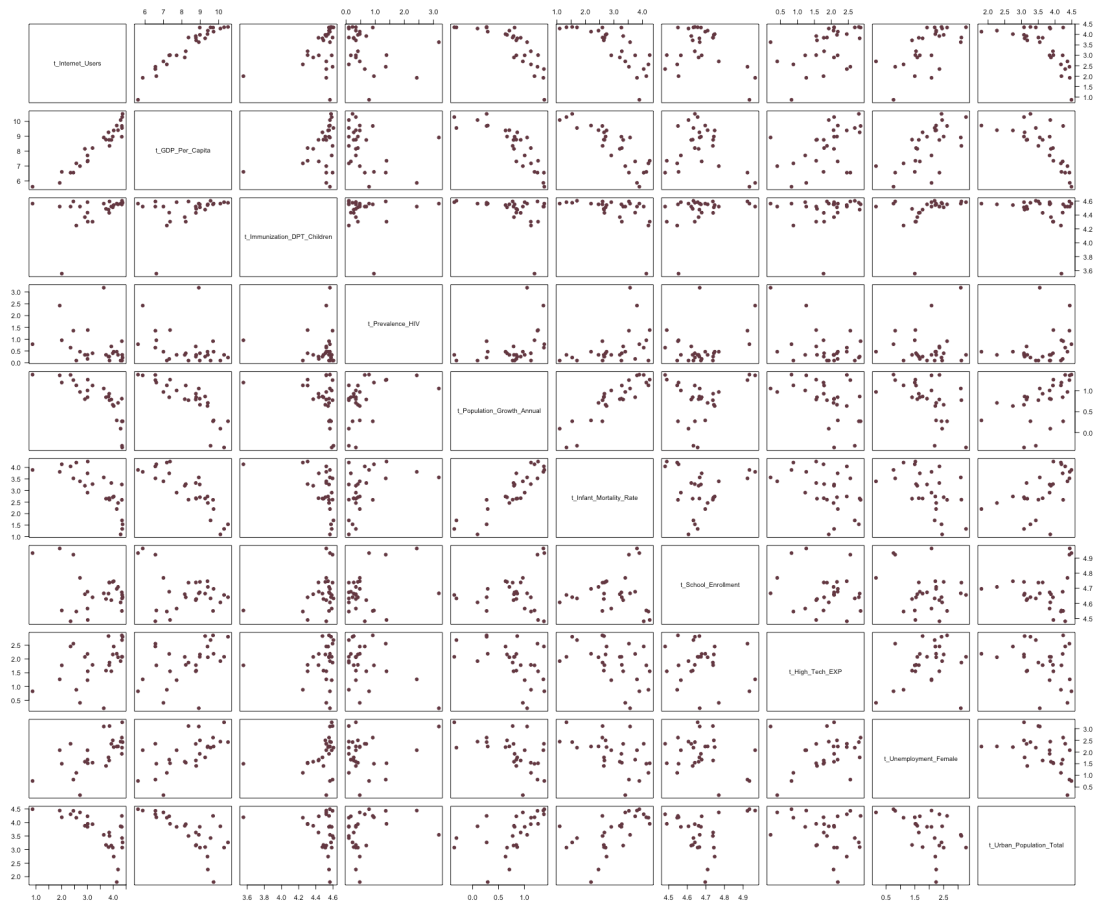


From above histogram plot, we can see that both the response factor and explainary factors are all skewed. The Literacy rate, Immulization, Population Growth Annual, and Internet Use are left skewed and GDP per Capita and other factors are right skewed. Therefore, we need to perform natural log transformations to each of the variables used in our modeling.

Since the data of Population Growth Annual Factor has negative number, and at the mean time, the number of data are large, we choose to use log(1+data) to transform the data.



We can see from the above histograms that these are looks more like normal distribution, and then we can check their correaltion between each other.




```
## NULL
```

From above scatter plot, we can see that GDP_Per_Capita and Internet_Users have positive correlation. Population_Growth_Annual and GDP_Per_Capita has negative correlation. Therefore, we need to check multicollinearity in our data.

```
##                      Variables       VIF
## 1             t_Internet_Users 13.847529
## 2             t_GDP_Per_Capita 24.929888
## 3  t_Immunization_DPT_Children  1.783070
## 4             t_Prevalence_HIV  2.287392
## 5   t_Population_Growth_Annual  5.733864
## 6      t_Infant_Mortality_Rate  9.936480
## 7          t_School_Enrollment  2.311722
## 8              t_High_Tech_EXP  1.530731
## 9        t_Unemployment_Female  2.479317
## 10    t_Urban_Population_Total  2.466387
```

From the VIF test, we can see that GDP Per Capita and Internet Users are the factor we need to remove since their VIF is larger than 10.

If we regard 0.7 as a high correlation, then Urban_Population_Total and Infant_Mortality_Rate have negative correlation with GDP_Per_Capita. Infant_Mortality_Rate have negative correlation with Internet_Users_Rate.But GDP Per Capita has high correlation to Internet User.

```
##
## Call:
## lm(formula = t_Adult_Literacy ~ ., data = data_model1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16808 -0.07591 -0.01507  0.05851  0.27308
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -5.03930    2.09851  -2.401  0.02805 *
## t_Internet_Users             0.04126    0.10562   0.391  0.70091
## t_GDP_Per_Capita             0.24542    0.09521   2.578  0.01956 *
## t_Immunization_DPT_Children  0.40101    0.17063   2.350  0.03110 *
## t_Prevalence_HIV            -0.06440    0.05506  -1.170  0.25828
## t_Population_Growth_Annual   0.01783    0.13321   0.134  0.89509
## t_Infant_Mortality_Rate      0.06099    0.09408   0.648  0.52546
## t_School_Enrollment          1.13621    0.33032   3.440  0.00313 **
## t_High_Tech_EXP             -0.08513    0.04609  -1.847  0.08220 .
## t_Unemployment_Female       -0.06235    0.05657  -1.102  0.28574
## t_Urban_Population_Total     0.06845    0.06033   1.135  0.27227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.135 on 17 degrees of freedom
## Multiple R-squared:  0.8818, Adjusted R-squared:  0.8123
## F-statistic: 12.69 on 10 and 17 DF,  p-value: 4.999e-06
```

From the Model One, I build a multi factor regression model to include all factors I selected. From the above result, only 2 factors are significant. Considering the VIF test, the factor GDP Per Capita is larger than 10 which should be remove from our model. For other factors VIF under 10, they should be moderate.

```
##
## Call:
## lm(formula = t_Adult_Literacy ~ t_Internet_Users + t_Unemployment_Female +
##     t_Immunization_DPT_Children + t_Prevalence_HIV +
t_Population_Growth_Annual +
##     t_Infant_Mortality_Rate + t_School_Enrollment + t_High_Tech_EXP +
##     t_Urban_Population_Total, data = data_model1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21812 -0.06558 -0.01883  0.06335  0.36029
##
```

```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.365997   1.765579  -0.774  0.44916
## t_Internet_Users              0.260742   0.071631   3.640  0.00187 **
## t_Unemployment_Female        -0.079709   0.064374  -1.238  0.23154
## t_Immunization_DPT_Children   0.327607   0.192820   1.699  0.10653
## t_Prevalence_HIV             -0.021534   0.060155  -0.358  0.72453
## t_Population_Growth_Annual   -0.014410   0.152000  -0.095  0.92552
## t_Infant_Mortality_Rate      -0.074193   0.089520  -0.829  0.41808
## t_School_Enrollment           0.841565   0.355193   2.369  0.02921 *
## t_High_Tech_EXP              -0.085666   0.052823  -1.622  0.12224
## t_Urban_Population_Total      0.005806   0.063284   0.092  0.92792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1547 on 18 degrees of freedom
## Multiple R-squared:  0.8356, Adjusted R-squared:  0.7535
## F-statistic: 10.17 on 9 and 18 DF,  p-value: 2.028e-05
```

From the Model Two, when the GDP Per Capita has been removed, the School Enrollment factor has become significant. Therefore, here I will remove other non significant factors to check multi factor regression model.

```
##
## Call:
## lm(formula = t_Adult_Literacy ~ t_Internet_Users + t_School_Enrollment,
##     data = data_model1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31331 -0.08922 -0.01168  0.08157  0.35614
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.49680    1.34705  -1.854 0.075643 .
## t_Internet_Users     0.29826    0.03643   8.188 1.53e-08 ***
## t_School_Enrollment  1.25404    0.27881   4.498 0.000137 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1649 on 25 degrees of freedom
## Multiple R-squared:  0.7407, Adjusted R-squared:  0.7199
## F-statistic:  35.7 on 2 and 25 DF,  p-value: 4.709e-08
```

based on α=0.05 and t-test, the model 3 with independent variables of Internet_Users and School Enrollment are the factors in our model at the end.

Higher R square means Higher model representative and higher explain of model. Adjusted R square means R square adjusted with high explaintory factor since more factor at model

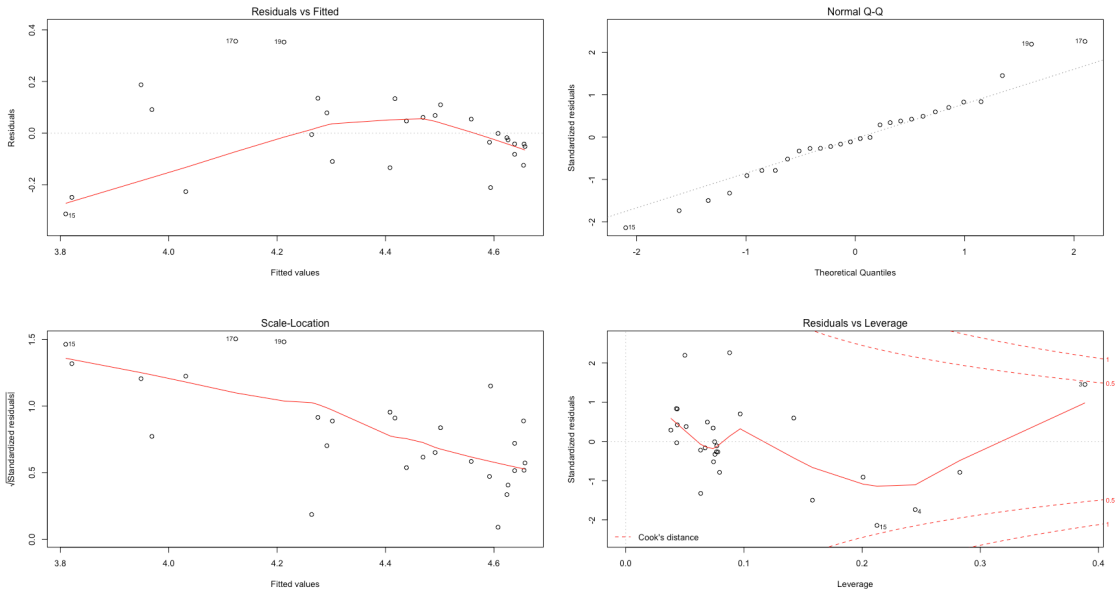
would increase the R square by nature. Adjusted R square is work for decrease this increase by nature.

We can see that, comparing to the R square and adjusted R square at full model which includes all 10 independent variables, in Model 3, both R square and adjusted R square only decreased a little but our model is more simple and clean. This means, the model 3 still has high explaintory and more clean.

Assumption Check:

1. x variables are fixed and measured without error. assumption not satisfied since this is information from surveys, etc. estimated by/reported to the World Bank.

2. E[ε]=0, assumption satisfied with LSE(Least Squares Estimator)

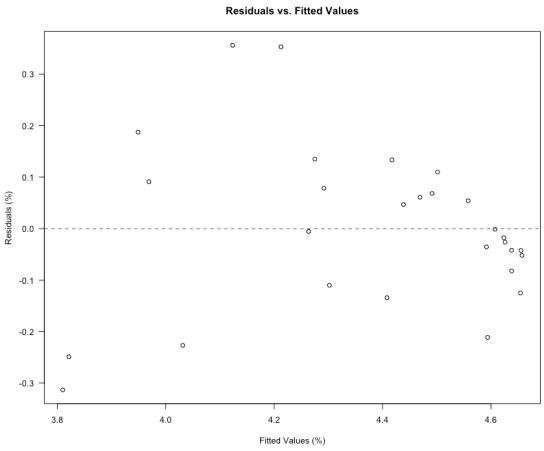


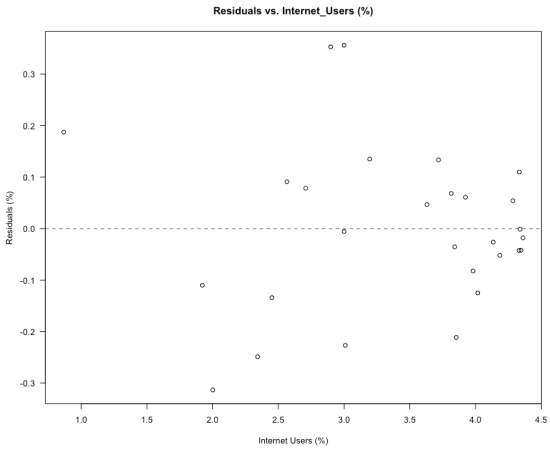
3. Var[ε]=σ2, From above plot, the residual looks clouded and centering with the larger independent valuables.

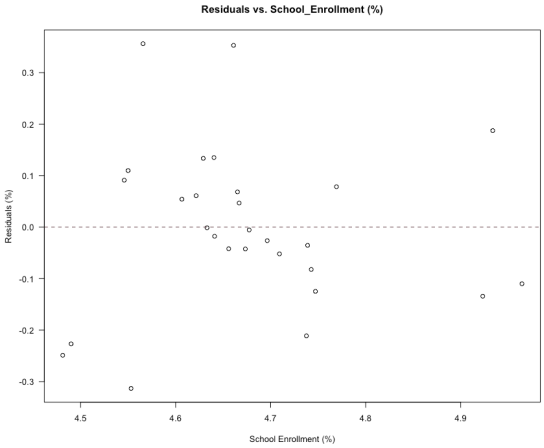
```
## integer(0)

## integer(0)

## integer(0)
```







4. ε's are normally distributed, from Normal Q-Q plot, the residual looks slightly heavy-tailed , thus assumption not satisfied.

5. ε's are independent, from residual plot above with two independent variables, the residual looks not exist centering and clouded condition, thus assumption satisfied.

6. x variables are not too highly correlated (collinearity/multicollinearity), checked by VIF, assumption satisfied.

Also, testing the Positive Autocorrelatio of residuals H0 : no residual correlation, $\rho = 0$ Ha : positive residual autocorrelation, $\rho > 0$

```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 2.0304, p-value = 0.5501
## alternative hypothesis: true autocorrelation is greater than 0
```

From result above, under significance level 5%, we can reject the H0 and the residuals exist negative autocorrelation. However, the p-value for the overall F-statistic in the final model(model 3) is less than $\alpha = 0.05$, so we can say that our overall model is significant.

v. Results:

My final multi-factor regression model is: $\hat{y}$ Adult Literacy Rate = -2.49680 % + 0.29826 % / % x Internet_Users – 1.25404 % / % x School_Enrollment_Rate.

This can be interpreted as: holding all other variables constant, a 1% increase in Internet_Users % is associated with a 0.30% increase in the adult literacy rate and a 1% increase in School Enrollment Rate is associated with a 1.25% increase in the adult literacy rate.

vi. References:

1) The World Bank: http://data.worldbank.org/indicator

2) Women's Education and Labor Force Participation and Fertility Decline in Iran (2002, Amir H. Mehryar, Akbar Aghajanian, Mohamad Tabibian and Farzaneh Tajdini)