**Homework 3: SDGB 7840**

**Report**

**Minxia Ji**

1. **Executive summary:**

The paper presents a possible model of literacy rate. It shows that the main determinants of worldwide literacy rate are the percentage of the rural population using improved sanitation facilities , internet users per 100 people and PM 2.5 pollution.It includes raw data clean, model building, assumption check, model validation and relevant hypothesis tests.

1. **Introduction:**

**Define the literacy rate:**Percentage of the population age 15 and above who can, with understanding, read and write a short, simple statement on their everyday life. Generally, ‘literacy’ also encompasses ‘numeracy’, the ability to make simple arithmetic calculations. This indicator is calculated by dividing the number of literates aged 15 years and over by the corresponding age group population and multiplying the result by 100.

**The purpose of the study:**As literacy is a skill, which is the ground of almost all processes of learning and is necessary for people not only to acquire languages and study literature, but also to learn other subjects, we should find out the factors affecting literacy rate, understand these factors and do relevant analysis.

**3.Data**

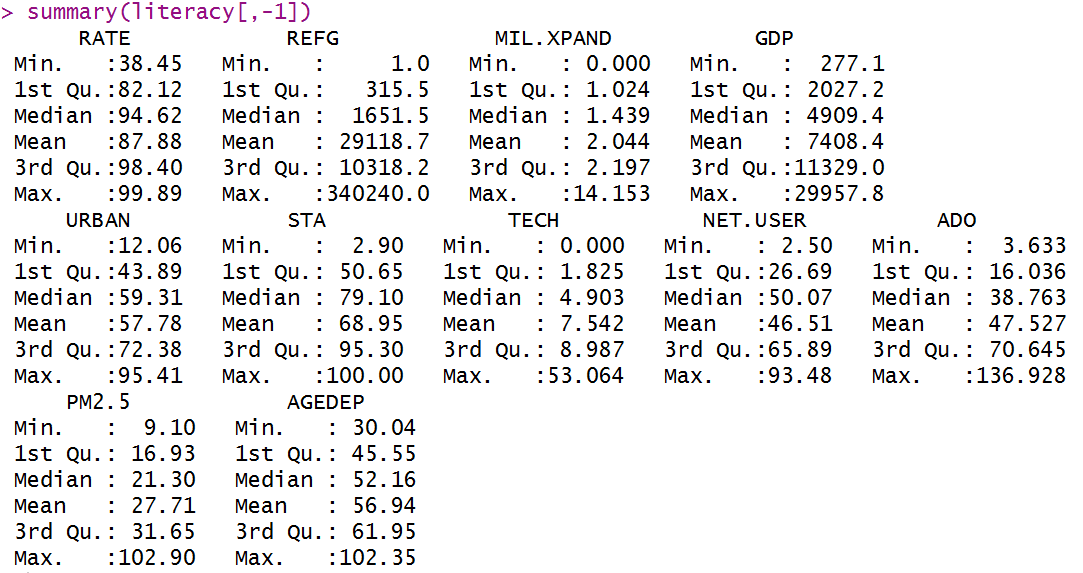
**Source of our data**:***http://www.worldbank.org/***

**10 explanatory variables I considered**:

|  |  |
| --- | --- |
| Variable/Name in dataset | Relevant information and Reason for choosing the variable |
| Refugee population by country or territory of origin / **RFEG** | A refugee, generally speaking, is a [displaced person](https://en.wikipedia.org/wiki/Displaced_person" \o "Displaced person) who has been forced to cross national boundaries and who cannot return home safely.Refugee population indicates the stability and safety of a country. |
| Military expenditure(% of central government expenditure) / **MIL.XPAND** | The proportion of military expenditure to total government expenditure.On the one hand, high/low proportion of military expenditure may indicate that the country is unstable/stable.On the other hand, high/low proportion of military expenditure correspond to low/high other expenditures such as education, infrastructure or health insurance, etc., which are indirect contributors of literacy rate. |
| GDP per capita(current US$) / **GDP** | GDP per capita is gross domestic product divided by midyear population.It represent the level of economic development, which contribute to development of the literacy rate. |
| Urban population(% of total) / **URBAN** | The ratio was calculated by people living in urban areas divided by total population.The ratio represent the level of civilization of a specific country.Higher urban population rate should comply with higher literacy rate. |
| Improved sanitation facilities, rural(% of rural population with access) / **STA** | Improved sanitation facilities are likely to ensure hygienic separation of human excreta from human contact. Access to improved sanitation facilities refers to the percentage of the population using improved sanitation facilities.If the hygienic level of rural population could be improved, rural people would be more healthier and focus on working so that they have more energy and money to pay attention to theirs or their kids’ education. |
| High-technology exports(% of manufactured exports) / **TECH** | This reflects the development level both on economy and on technology of a country.High rate means this country more likely to export high-techs and this country should posses high literacy rate.Low rate means this country relies more on manufactured exports(some workers in the manufactured industry even not posses literacy reading skills at all) and this country should possess low literacy rate. |
| Internet user(per 100 people) / **NET.USER** | Internet user per 100 people not only indicate the tech development level of a country, but also indicate the information popularity.Nowadays, more and more people are self-educated through the internet.Thus, the more users we have, we can expect the higher literacy rate. |
| Adolescent fertility rate(births per 1000 woman ages 15-19) / **ADO** | Adolescent fertility rate is the number of births per 1,000 women ages 15-19.In general, young mothers are seldom seen in developed countries.Social development and social norms decide that whether women would be well-educated or not, if not, then the literacy rate should be low correspondingly. |
| PM2.5 air pollution, mean annual exposure(micrograms per cubic meter) / **PM2.5** | Population-weighted exposure to ambient PM2.5 pollution is defined as the average level of exposure of a nation's population to concentrations of suspended particles measuring less than 2.5 microns in aerodynamic diameter.That stands whether a country care about the public healthy. |
| Age dependency ratio(% of working age population) / **AGEDEP** | Age dependency ratio is the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64. The ratio tells us the amount of working force in other way. In my understanding, working force provide the source of finance to support education, therefore influence the literacy rate. |

**Relevant summary information about the explanatory and response variables**:

Response variable is adult literacy rate, percentage of the population age 15 and above who can, with understanding, read and write a short, simple statement on their everyday life. Explanatory variables reflect economy, health, social development and environment situations.I also include some interesting relevant data such as high -tech export and military expenditure.



**Which countries are included in my processed data set**:



**which year(s) are included in my data set:**

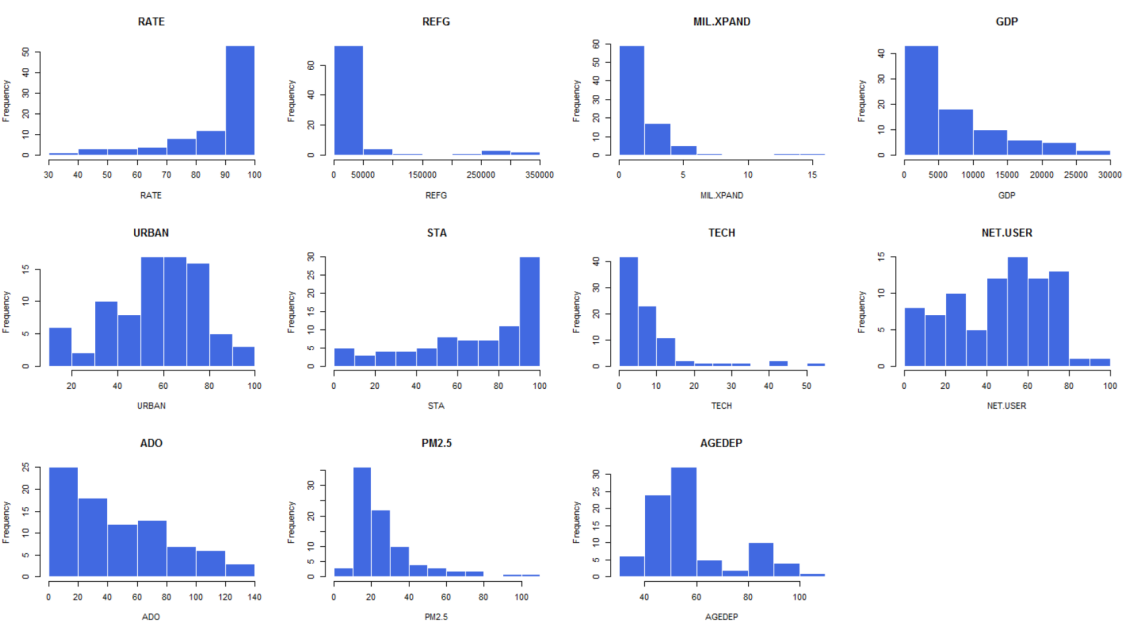
1. Because 2015 has the most of data for literacy rate.

**How I cleaned my data:**

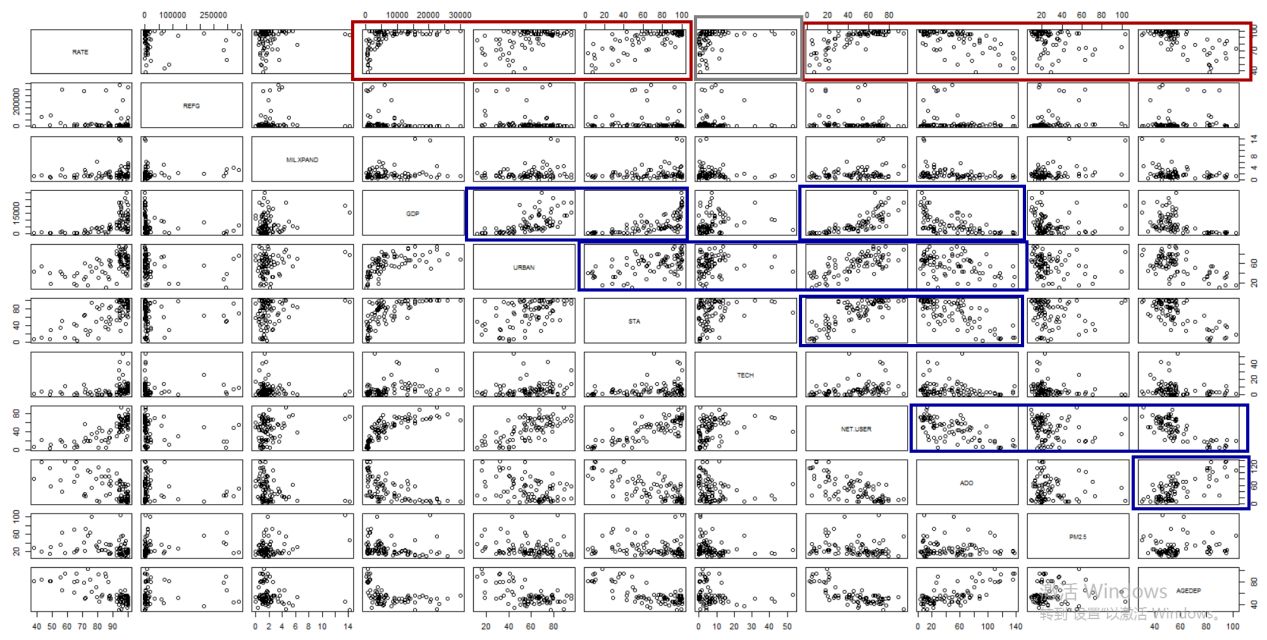
I took out of all 10 variables data for 2015 and combined them with literacy rate for 2015 to form a new csv file.In R studio, I use complete.cases to remove some variables which contains NA.

1. **Methods**

**Relevant plots:**

****

|  |  |
| --- | --- |
| Left-skewed | Literacy rate(RATE), sanitation improvement(STA), |
| Right-skewed | Refugee population(REFG),Military expenditure(MIL.XPAND),GDP,High-technology export(TECH), internet user per 100 people(NET.USER), adolescent fertility rate(ADO),PM2.5 and age dependency ratio(AGEDEP) |
| Roughly symmetric | Urban population(URBAN) |

****

From the scatter plot, we found that:

The relationships between explanatory variables and response variables(red&gray rectangular)

Red rectangular:

There are positive linear relationships between literacy rate and Urban population(% of total)/Improved sanitation facilities, rural(% of rural population with access)/PM2.5 air pollution, mean annual exposure(micrograms per cubic meter).

There are negative linear relationships between literacy rate and Adolescent fertility rate(births per 1000 woman ages 15-19)/Age dependency ratio(% of working age population).We even can say that the relationship between literacy rate and ADO is kinda like negatively curved relationship.

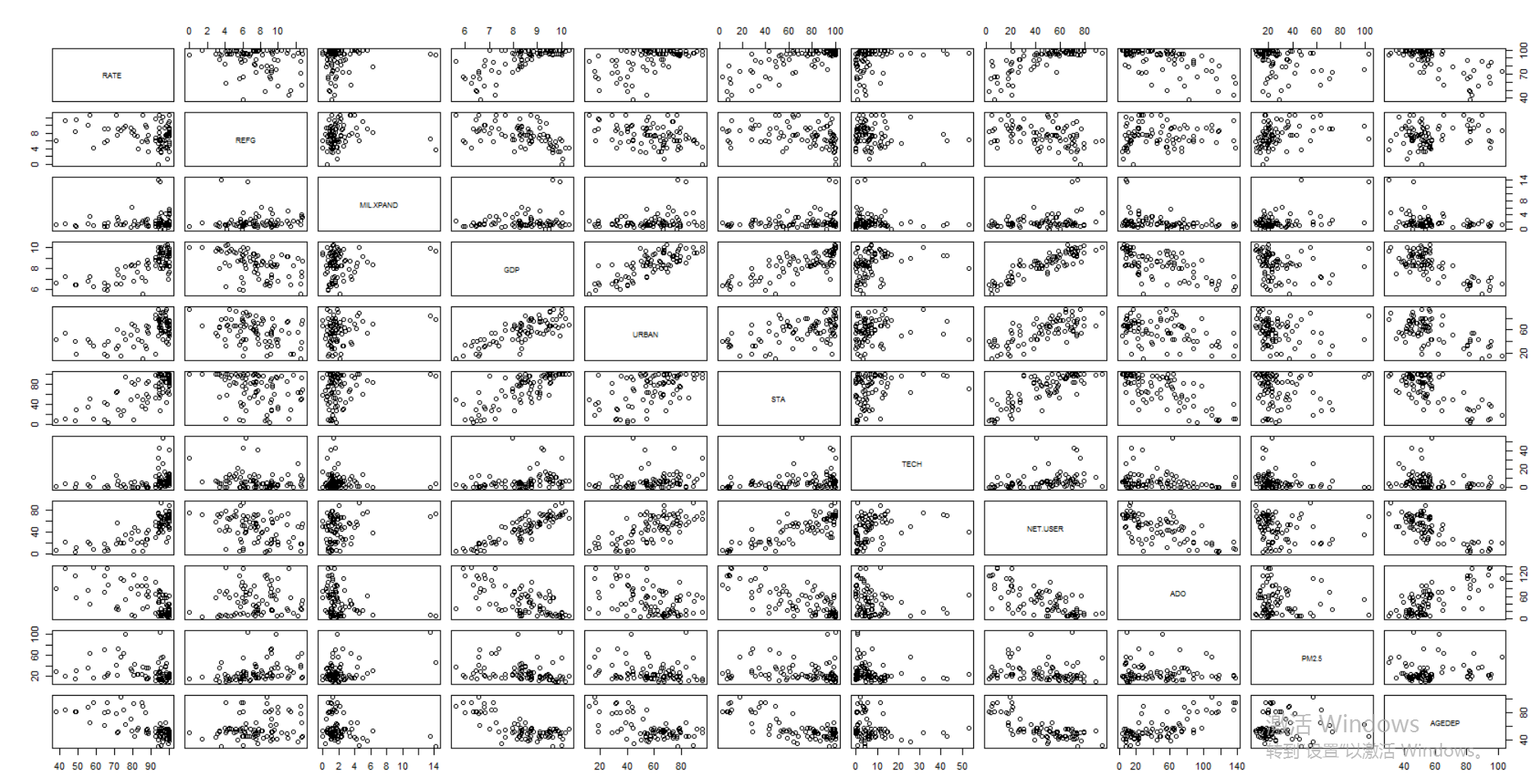
There is a positive curved relationship between literacy rate and GDP/Internet users per 100 people.

Gray rectangular:It is hard to tell the relationship between literacy and high-technology export.We can say that there is a slight positively curved relationship.

The relationships between explanatory variables(blue rectangular):

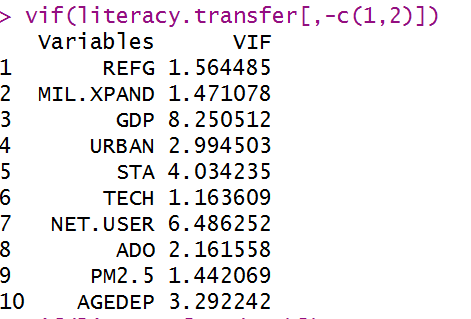
We observed some positive or negative linear relationships even curved relationships Between explanatory variables.So we should be careful and check them in the later analysis steps.

According to the distribution of each variable and also the scatter plot, we should do **log transformation** to these variables:**REFG, GDP, NET.USERS.**



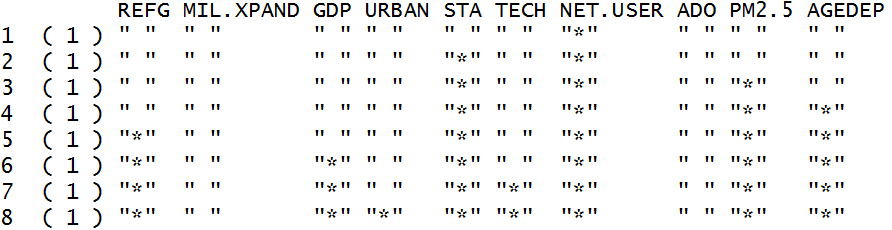
After log transformation, the relationships seems more clear to us.

**Collinearty check:**



Multicollinearity is moderate in GDP and NET.USER.Other variables are fine.

Forward stepwise:



Order of adding variables:

(see explanation of variables in question 3)

NET.USER - STA - PM2.5 - AGEDEP - REFG - GDP - TECH - URBAN

MIL.XPAND and ADO are not included.

**Using anova function to conduct partial F-test.Removing variables based on t-test:**

|  |  |  |
| --- | --- | --- |
| 1 | 2 | 3 |
| 4 | 5 | 6 |
| 7 | 8 | **Hypothesis test when α=0.05 for lm8**  Assume regression assumptions are satisfied;  Overall F-test significant;  Slope of STA is statistically significant;  Slope of NET.USER is statistically significant;  Slope of PM2.5 is statistically significant. |

Choose α=0.05, based on t-test, the variable corresponded to black rectangular means slope of this variable is not statistically significant and should not be included in the next model.

Removing order:**ADO-MIL.XPND - URBAN - TECH - GDP - REFG - AGEDEP**

We can see that both R square and adjusted R square **improved** each time when we removing certain variables in each step except for lm7 to lm8, both R square and adjusted R square **slightly decreased.**Still, we should remove AGEDEP from the model because the both t-test and partial-F test told us the slope of AGEDEP is not statistically significant.

**Partial F-test**

|  |  |
| --- | --- |
| 1 | 2 |
| Hypothesis:  *H*0:β(ADO) =0  *H*1:β(ADO) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.0047  Which has an F distribution with 1 and 73 degrees of freedom  P-value = 0.9453>α = 0.05  Conclusion:Can not reject null hypothesis; β(ADO) is not significantly different from 0, therefore the first model is better than the second model. | Hypothesis:  *H*0:β(MIL.XPAND)=0  *H*1:β(MIL.XPAND) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.0108  Which has an F distribution with 1 and 74 degrees of freedom  P-value = 0.9176>α = 0.05  Conclusion:Can not reject null hypothesis; β(MIL.XPAND) is not significantly different from 0, therefore the first model is better than the second model. |
| 3 | 4 |
| Hypothesis:  *H*0: β(URBAN) =0  *H*1: β(URBAN) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.0227  Which has an F distribution with 1 and 75 degrees of freedom  P-value = 0.8807>α = 0.05  Conclusion:Can not reject null hypothesis; β(URBAN) is not significantly different from 0, therefore the first model is better than the second model. | Hypothesis:  *H*0: β(TECH)=0  *H*1: β(TECH) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.2105  Which has an F distribution with 1 and 76 degrees of freedom  P-value = 0.6477>α = 0.05  Conclusion:Can not reject null hypothesis; β(TECH) is not significantly different from 0, therefore the first model is better than the second model. |
| 5 | 6 |
| Hypothesis:  *H*0: β(GDP) =0  *H*1: β(GDP) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.3142  Which has an F distribution with 1 and 77 degrees of freedom  P-value = 0.5767>α = 0.05  Conclusion:Can not reject null hypothesis; β(GDP) is not significantly different from 0, therefore the first model is better than the second model. | Hypothesis:  *H*0: β(REFG) =0  *H*1: β(REFG) not equals to zero  Choose significance level : α = 0.05  Test statistic: F = 0.2285  Which has an F distribution with 1 and 78 degrees of freedom  P-value = 0.634>α = 0.05  Conclusion:Can not reject null hypothesis; β(REFG) is not significantly different from 0, therefore the first model is better than the second model. |
| 7 | Hypothesis:  *H*0: β(AGEDEP) =0  *H*1: β(AGEDEP) =0 not equals to zero  Choose significance level : α = 0.05  Test statistic: F =3.2322  Which has an F distribution with 1 and 79 degrees of freedom  P-value = 0.07602>α = 0.05  Conclusion:Can not reject null hypothesis;β(AGEDEP) is not significantly different from 0, therefore the first model is better than the second model. |

**Assumption check:**

|  |  |
| --- | --- |
| R1 | R2 |

1. Measurement accuracy: No measurement error exist in the data, assumption satisfied
2. Linearity:No obvious trend/curve is observed, assumption satisfied.
3. Constance variance:When fitted value increasing, the variance of residuals decreasing;assumption violated.
4. Normality:in normal quantile plot, the points are heavy-tailed; assumption violated.
5. Independence:No time serious data in the dataset, but there are a few group of countries included in the dataset, the assumption might slightly violated.

The regression assumption might **not** satisfied.

**Model Validation:**

Because the data set is too small to split, I use jackknife method to do model validation

|  |  |  |
| --- | --- | --- |
| Adjusted R square | R square | R square(Jackknife) |
| 0.6692 | 0.6568 | 0.3971 |

RMSE:1.34878

RMSE(Jackknife):10.63772

The R square of Jackknife is smaller than R square and adjusted R square of the final model.The RMSE(Jackknife) is bigger than RMSE.The model is fine, however, the R square of Jackknife is much smaller than the R square of the final model, which indicates there might exist overfitting problem when applying the model to test data.

1. **Results:**

**Final model:**

Literacy rate(%) = 46.33261% + 0.17022\*STA(%) + 9.31678\*log(NET.USER(%)) - 0.14366\*PM2.5(micrograms per cubic meter)\*(%/micrograms per cubic meter)

**Interpretation:**

Slope for STA:If the percentage of the rural population using improved sanitation facilities increased 1 percent, holding other explanatory variables log(NET.USER) and PM2.5 constant, the literacy rate would increase 0.17022 percent.

Slope for log(NET.USER):If the percentage of internet users per 100 people increase 1(percentage increase refers to the increase in NET.USER), holding other explanatory variables STA and PM2.5 constant, the literacy rate would increase roughly 9.31678%.

Slope for PM2.5:If the average level of exposure of a nation's population to concentrations of suspended particles measuring less than 2.5 microns in aerodynamic diameter increased 1 micrograms per cubic meter, holding other explanatory variables log(NET.USER) and STA constant, the literacy rate would decrease 0.14366 percent.

Y-intercept: if the the percentage of the rural population using improved sanitation facilities is 0, the internet users per 100 people is 1(minimize:log(NET.USER)=1)and the average level of exposure of a nation's population to concentrations of suspended particles measuring less than 2.5 microns in aerodynamic diameter is 0, the literacy rate would be 46.33261%.

The y-intercept does not have a practical interpretation. Since the minimum values of STA, NET.USER and PM2.5 are 2.9%,2.5 people,9.1 micrograms per cubic meter, the y-intercept does not have a practical interpretation.In addition, it is not possible for a country to have literacy rate equals to 0, the y-intercept does not have a contextual meaning.

**Improvement:**

Need data about poverty for 2015, such as GINI index and poverty gap at $1.9 a day, etc..Here are [Poverty and illiteracy are closely connected](http://www.aecf.org/~/media/Pubs/Topics/Education/Other/EnsuringSuccessforYoungChildrenLiteracy/Ensuring_Success_Early_Literacy.pdf). (More information about literacy rate and poverty are included in the link:

***<https://prezi.com/7_x6oeh8dssh/how-does-literacy-rates-affect-poverty/>***

Need data for enrollment rate and average years of schooling of adults.

It is better to have sufficient data for each variable so that we don’t need to remove a lot of countries from our dataset.

1. **Reference**

Data source:WORLD BANK:***http://www.worldbank.org/;***

Term meaning check:Wikipedia;

Relevant essays reading:**http://scholar.google.com/.**