AN EXPLANATORY RESEARCH ON LITERACY RATE ACROSS COUNTRIES

SDGB 7840-EXPLANATORY MODELS-SPRING 2017

APRIL 6, 2017

BINGYANG HU(ZOEY)

1.Executive summary

This paper represents an explanatory research on literacy rate across countries, constructing multiple regression model using 11 variables. The methods include descriptive analysis, variable screening method and other validation criterion. The result shows that literacy rate is related to three variables, and GDP is excluded, which is contrary to expectation.

2.Introduction

Literacy refers to ability, with understanding, to read and write a short, simple statement on their everyday life. Generally, 'Literacy' also encompasses 'numeracy', the ability to make simple arithmetic calculations. [1]

Literacy rate is the number of literates divided by corresponding group and then multiplying by 100, used to measure overall literacy ability.

For a country, to some degree this rate is reflection of overall education level, while inverse it will affect the country from many aspects and even individual lives. With the purpose of getting insight, an explanatory research on associations between literacy and other factors are made, hoping to contributing to the prosperity of countries and better life of individuals.

3.Data

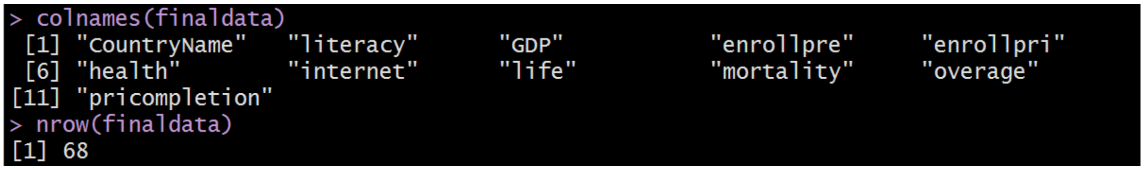
Through this analysis, all data used are from World Bank official website. Below are data description for 1 response variable and 10 explanatory variables considered for the model.

Variable Name	**Variable Description**	**Type**
Country Name	Name of nation	nominal
Literacy rate (%)	Percentage of population age 15 and above who can read and write	continuous
Preprimary enrollment (%)	Gross enrollment ratio in pre-primary education for both sexes	continuous
Primary enrollment (%)	Gross enrollment ratio in primary education for both sexes	continuous
Internet users (/100 people)	Internet users per 100 people	discrete
Life (years)	Life expectancy at birth	continuous
Morality (/1000 people)	Number of neonates dying before reaching 28 days of age.	discrete
Overage (%)	Percentage of population ages 65 and above	continuous
Primary completion (%)	Gross completion ratio in primary education for both sexes	continuous
GDP per capita ($)	Gross domestic product divided by midyear population	continuous
Health expenditure ($)	Health expenditure per capita	continuous
Poverty gap at 1.9$/day	Shortfall in income from poverty line to measure depth of poverty	continuous

As shown above, it is obvious that to some extent literacy rate reflects education. Therefore, enrollment rate and completion rate for basic education are all selected to show education level of a country. Besides, reading and writing ability cannot be fully guaranteed without biological health. From this perspective, life, morality and health are chosen for consideration. What's more, without any doubt both education and health are related to technology and economic, and internet users as well as GDP per capita are proper indication for them respectively.

With the fact that the data set used for model construction should be complete without missing value and sample size is supposed to be big enough, data from year 2014 for all explanatory variables are included while Poverty gap is not used because of too many missing values. After combing them and using complete.cases function to choose complete rows, literacy rate for each country needs to be put in. Since this rate does not change a lot within the period of two years, part of data is borrowed from year 2015 to make dataset bigger.

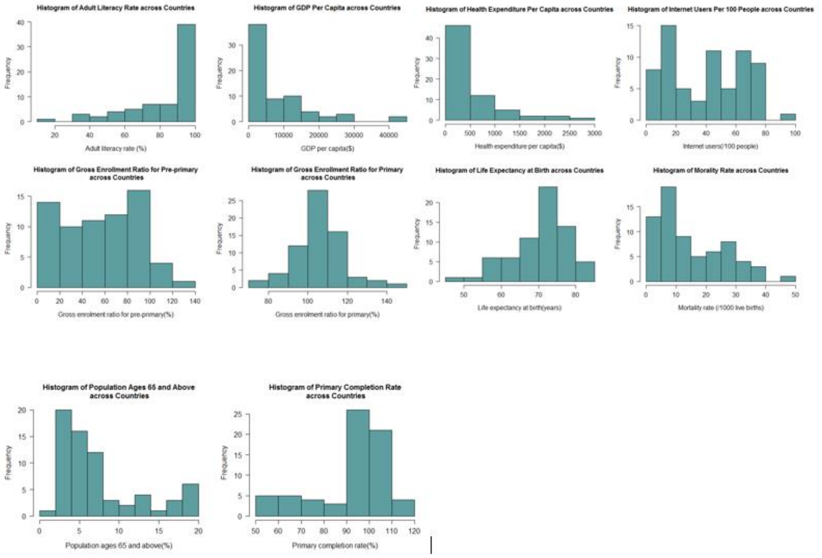
With the deletion of non-country records (region names such as East Asia and Euro Area), 68 countries are left in the end, with each having 10 variables. Those countries include Chile (South America), Indonesia (Southeast Asia), Mexico (North America), Spain (Europe), Kenya (Africa) and so on.

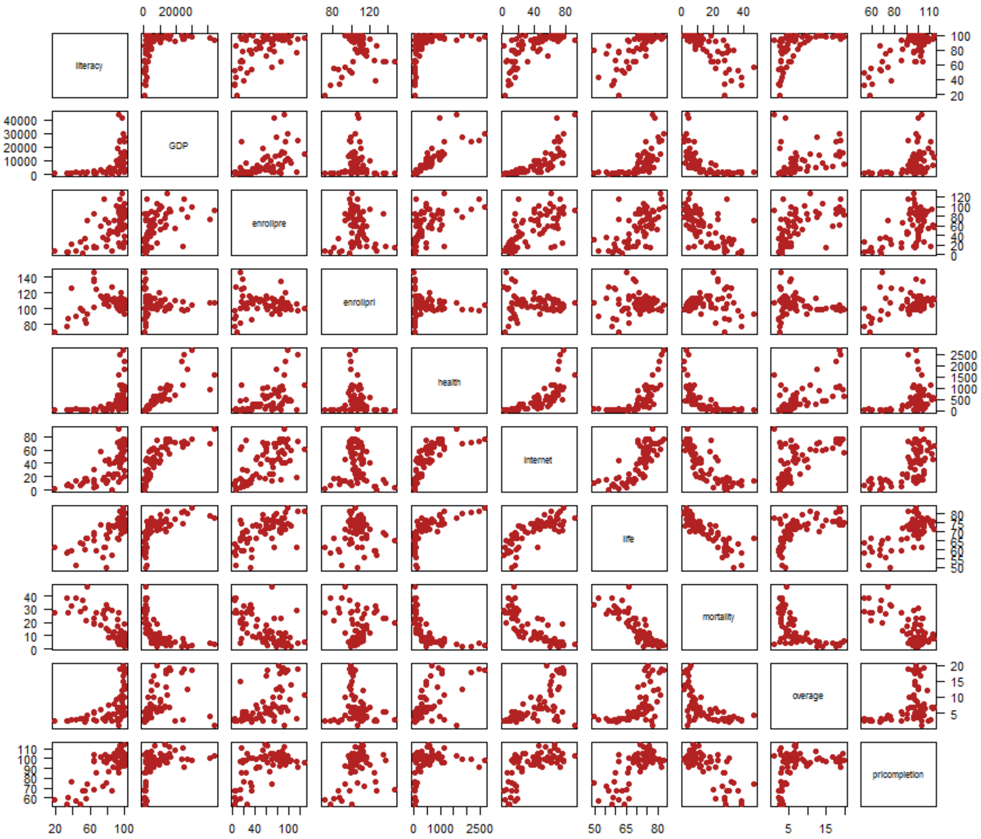


```
> colnames(finaldata)
 [1] "CountryName"   "literacy"      "GDP"           "enrollpre"     "enrollpri"
 [6] "health"        "internet"      "life"          "mortality"     "overage"
[11] "pricompletion"
> nrow(finaldata)
[1] 68
```

4.Methods

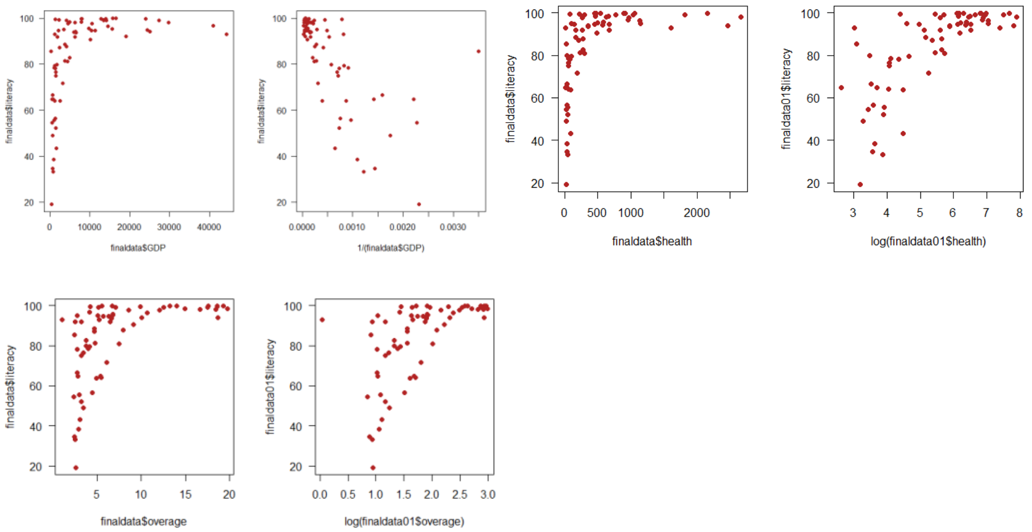
4.1 Relevant plots





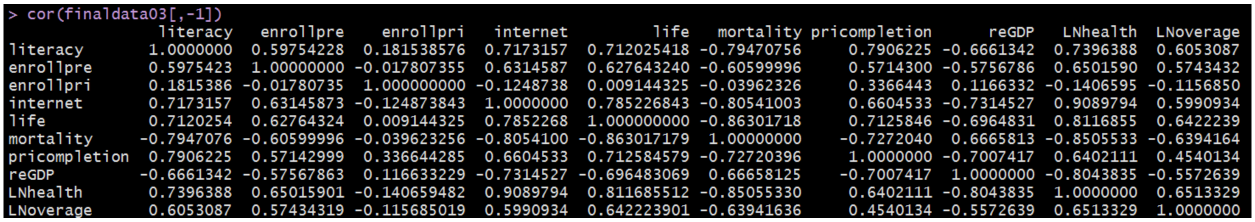
Scatterplot shows that literacy has strong curve relationship with GDP, health and overage Results, so data transformation is needed first. Both health and overage are replaced by log value while GDP is substituted with reciprocal.

Below are the comparisons between before and after transformation.



Meanwhile, it is easily noticed that linear relationship exists between several pairs of explanatory variables (internet vs life, GDP vs health etc.), which alter me about multicollinearity.

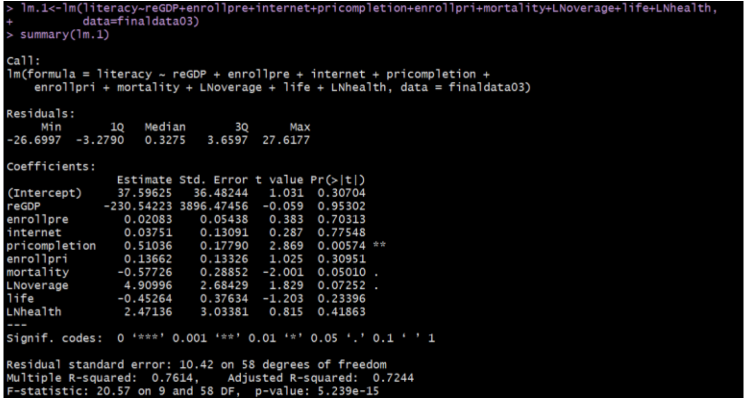
Correlations:



```
> cor(finaldata03[,-1])
              literacy   enrollpre    enrollpri   internet         life   mortality pricompletion      reGDP   LNhealth  LNoverage
literacy       1.0000000  0.59754228  0.181538576  0.7173157  0.712025418 -0.79470756     0.7906225 -0.6661342  0.7396388  0.6053087
enrollpre      0.5975423  1.00000000 -0.017807355  0.6314587  0.627643240 -0.60599996     0.5714300 -0.5756786  0.6501590  0.5743432
enrollpri      0.1815386 -0.01780735  1.000000000 -0.1248738  0.009144325 -0.03962326     0.3366443  0.1166332 -0.1406595 -0.1156850
internet       0.7173157  0.63145873 -0.124873843  1.0000000  0.785226843 -0.80541003     0.6604533 -0.7314527  0.9089794  0.5990934
life           0.7120254  0.62764324  0.009144325  0.7852268  1.000000000 -0.86301718     0.7125846 -0.6964831  0.8116855  0.6422239
mortality     -0.7947076 -0.60599996 -0.039623256 -0.8054100 -0.863017179  1.00000000    -0.7272040  0.6665813 -0.8505533 -0.6394164
pricompletion  0.7906225  0.57142999  0.336644285  0.6604533  0.712584579 -0.72720396     1.0000000 -0.7007417  0.6402111  0.4540134
reGDP         -0.6661342 -0.57567863  0.116633229 -0.7314527 -0.696483069  0.66658125    -0.7007417  1.0000000 -0.8043835 -0.5572639
LNhealth       0.7396388  0.65015901 -0.140659482  0.9089794  0.811685512 -0.85055330     0.6402111 -0.8043835  1.0000000  0.6513329
LNoverage      0.6053087  0.57434319 -0.115685019  0.5990934  0.642223901 -0.63941636     0.4540134 -0.5572639  0.6513329  1.0000000
```

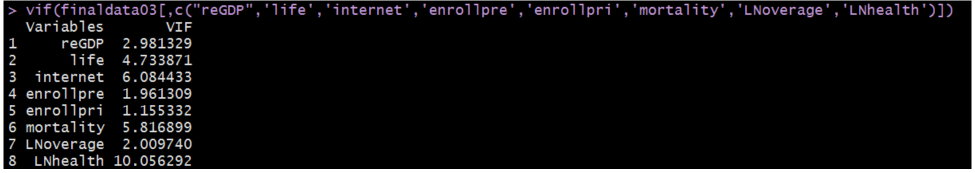
4.2 Model building

Scatterplots and correlation show that on average, literacy has strong linear relationship with each explanatory variable. Then the model 1 is built based on all these variables.



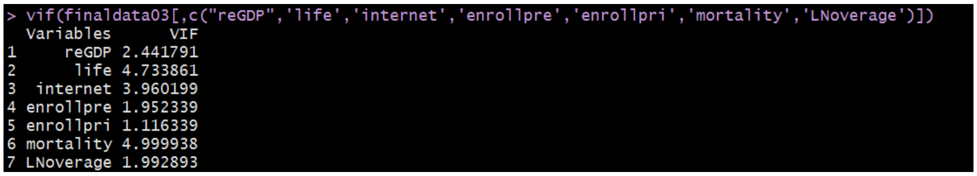
```
> lm.1<-lm(literacy~reGDP+enrollpre+internet+pricompletion+enrollpri+mortality+LNoverage+life+LNhealth,
+          data=finaldata03)
> summary(lm.1)

Call:
lm(formula = literacy ~ reGDP + enrollpre + internet + pricompletion + 
    enrollpri + mortality + LNoverage + life + LNhealth, data = finaldata03)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.6997  -3.2790   0.3275   3.6597  27.6177 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)    37.59625   36.48244   1.031  0.30704   
reGDP        -230.54223 3896.47456  -0.059  0.95302   
enrollpre       0.02083    0.05438   0.383  0.70313   
internet        0.03751    0.13091   0.287  0.77548   
pricompletion   0.51036    0.17790   2.869  0.00574 **
enrollpri       0.13662    0.13326   1.025  0.30951   
mortality      -0.57726    0.28852  -2.001  0.05010 . 
LNoverage       4.90996    2.68429   1.829  0.07252 . 
life           -0.45264    0.37634  -1.203  0.23396   
LNhealth        2.47136    3.03381   0.815  0.41863   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.42 on 58 degrees of freedom
Multiple R-squared:  0.7614,	Adjusted R-squared:  0.7244 
F-statistic: 20.57 on 9 and 58 DF,  p-value: 5.239e-15
```

Among 9 variables, only one is significant. And life, which I expected should have positive coefficient, has negative sign. Nearly all T-test are insignificant while he overall F-test is significant, which remind me of the multicollinearity. Then VIF is calculated to check it.



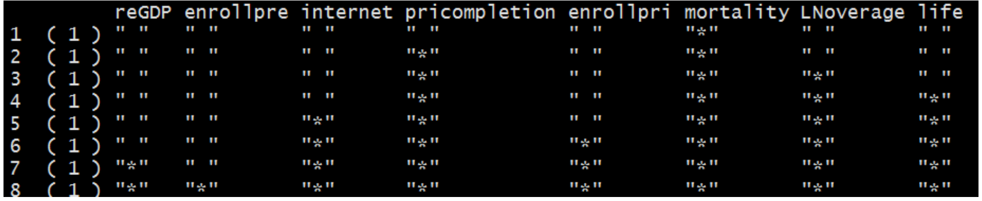
```
> vif(finaldata03[,c("reGDP",'life','internet','enrollpre','enrollpri','mortality','LNoverage','LNhealth')])
  Variables       VIF
1     reGDP  2.981329
2      life  4.733871
3  internet  6.084433
4 enrollpre  1.961309
5 enrollpri  1.155332
6 mortality  5.816899
7 LNoverage  2.009740
8  LNhealth 10.056292
```

VIF of LNhealth (10.056) and internet (6.084) and correlation between them (0.91) prove the existence of multicollinearity, so LNhealth is dropped to reduce the effect of multicollinearity. VIF of rest variables are below:



```
> vif(finaldata03[,c("reGDP",'life','internet','enrollpre','enrollpri','mortality','LNoverage')])
  Variables      VIF
1     reGDP 2.441791
2      life 4.733861
3  internet 3.960199
4 enrollpre 1.952339
5 enrollpri 1.116339
6 mortality 4.999938
7 LNoverage 1.992893
```

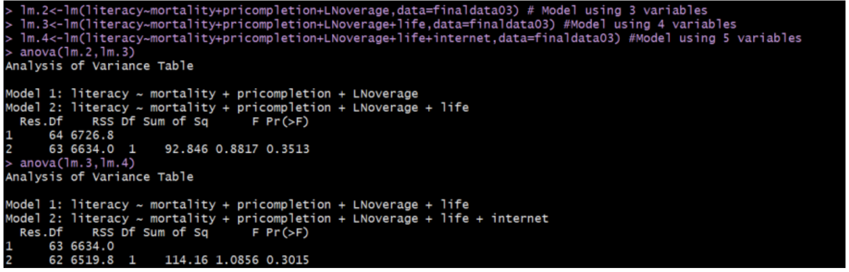
After removing the effect of multicollinearity, forward stepwise regression and Mallow’s C_P value help to build the model.



```
         reGDP enrollpre internet pricompletion enrollpri mortality LNoverage life
1  ( 1 ) " "   " "       " "      " "           " "       "*"       " "       " " 
2  ( 1 ) " "   " "       " "      "*"           " "       "*"       " "       " " 
3  ( 1 ) " "   " "       " "      "*"           " "       "*"       "*"       " " 
4  ( 1 ) " "   " "       " "      "*"           " "       "*"       "*"       "*" 
5  ( 1 ) " "   " "       "*"      "*"           " "       "*"       "*"       "*" 
6  ( 1 ) " "   " "       "*"      "*"           "*"       "*"       "*"       "*" 
7  ( 1 ) "*"   " "       "*"      "*"           "*"       "*"       "*"       "*" 
8  ( 1 ) "*"   "*"       "*"      "*"           "*"       "*"       "*"       "*" 
```

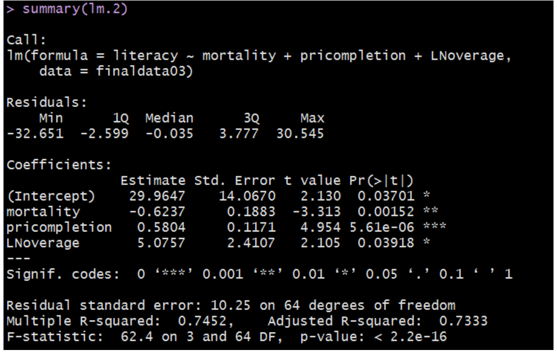
Model	Cp	p+1	
1 variable	26.1	2	24.1
2 variables	4.6	3	1.6
3 variables	2.3	4	-1.7
4 variables	3.4	5	-1.6
5 variables	4.4	6	-1.6
6 variables	5.6	7	-1.4
7 variables	7.2	8	-0.8
8 variables	9	9	0

With the criterion of choosing model in which Cp is low and close to p+1, the options for model are narrowed down to 3 models, with 3, 4 and 5 variables respectively. Since they are nested models, partial F-test is used to decide the final model.



```
> lm.2<-lm(literacy~mortality+pricompletion+LNoverage,data=finaldata03) # Model using 3 variables
> lm.3<-lm(literacy~mortality+pricompletion+LNoverage+life,data=finaldata03) #Model using 4 variables
> lm.4<-lm(literacy~mortality+pricompletion+LNoverage+life+internet,data=finaldata03) #Model using 5 variables
> anova(lm.2,lm.3)
Analysis of Variance Table

Model 1: literacy ~ mortality + pricompletion + LNoverage
Model 2: literacy ~ mortality + pricompletion + LNoverage + life
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     64 6726.8
2     63 6634.0  1    92.846 0.8817 0.3513
> anova(lm.3,lm.4)
Analysis of Variance Table

Model 1: literacy ~ mortality + pricompletion + LNoverage + life
Model 2: literacy ~ mortality + pricompletion + LNoverage + life + internet
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     63 6634.0
2     62 6519.8  1    114.16 1.0856 0.3015
```

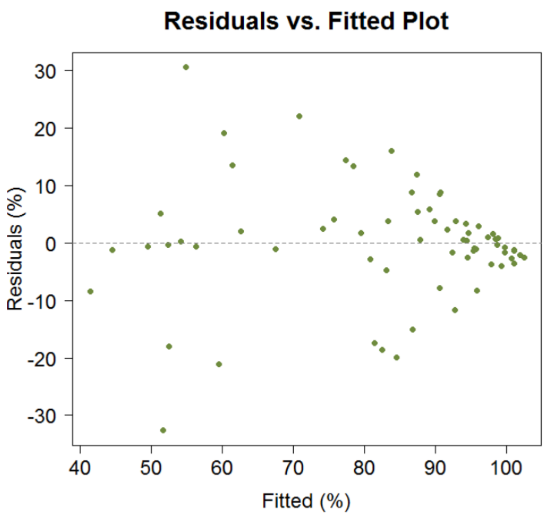
Both partial F-test show insignificant result, and lm.2 is chosen finally.

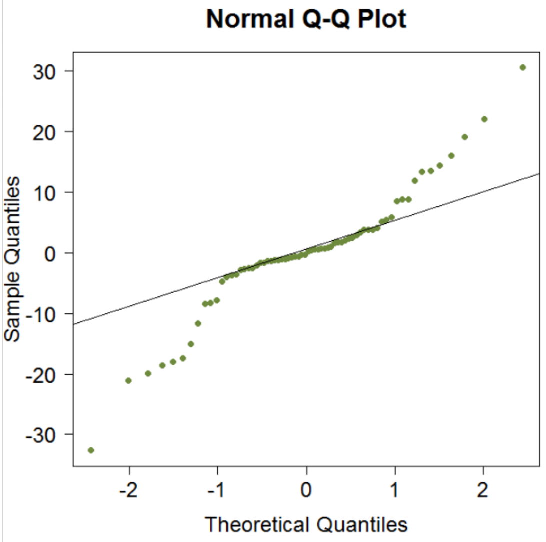


```
> summary(lm.2)

Call:
lm(formula = literacy ~ mortality + pricompletion + LNoverage, 
    data = finaldata03)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.651  -2.599  -0.035   3.777  30.545 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    29.9647    14.0670   2.130  0.03701 *  
mortality      -0.6237     0.1883  -3.313  0.00152 ** 
pricompletion   0.5804     0.1171   4.954 5.61e-06 ***
LNoverage       5.0757     2.4107   2.105  0.03918 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.25 on 64 degrees of freedom
Multiple R-squared:  0.7452,	Adjusted R-squared:  0.7333 
F-statistic:  62.4 on 3 and 64 DF,  p-value: < 2.2e-16
```

lm.2
literacy(%)=29.9647(%)-0.6237(%/(1/1000people))*mortality+0.5804(%/%)*pricompletion+5.0757(%)*LNoverage

4.3 Regression assumptions check

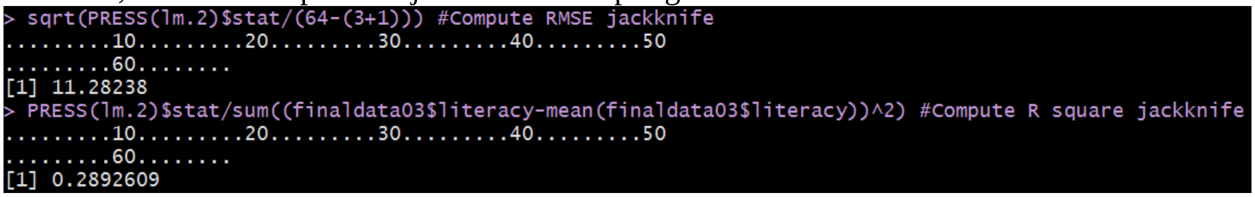




Measurement error and fixed x: Those data are collected by different institutions and collection methods or criterion may vary, so it is likely to have measurement error. However, without any additional information, I assume that the variables have been measured without error. And fixed x is untrue because this value is not set by me who do the analysis.
Constant variance: From the residual plot (on the left), I found that the residuals are heteroscedastic for that the variance of residuals decrease as the fitted values increase.
Linearity: Although residuals are heteroscedastic, there are no other patterns in the residual plot, therefore, the linearity assumption is satisfied.
Normality: The line from normal quantile plot (on the right) shows that residuals are heavy-tailed rather than normally distributed so the normality assumption is violated.
Independence: Considering that the data source is World Bank official website and the countries are randomly selected, I assume that the residuals are independent.
Multicollinearity: VIF has been calculated before and all variables in the model have VIF which is smaller than 5, therefore there is multicollinearity problem.
Outliers: From the residual plot (on the left), there are no obvious outliers. The green on the left top is on the border because the RMSE=10.25 and this point is on the border of 3s distance.

4.4 Model validation

Considering that the sample size is not that big, instead of using data splitting and validating with new data, I chose to implement jackknife resampling method.



```
> sqrt(PRESS(lm.2)$stat/(64-(3+1))) #Compute RMSE jackknife
.........10.........20.........30.........40.........50
.........60........
[1] 11.28238
> PRESS(lm.2)$stat/sum((finaldata03$literacy-mean(finaldata03$literacy))^2) #Compute R square jackknife
.........10.........20.........30.........40.........50
.........60........
[1] 0.2892609
```

R^2	$R^2_{jackknife}$	RMSE	$RMSE_{jackknife}$
0.7333	0.2892	10.25	11.28

The difference between two types of RMSE is acceptable while that between two types of R^2 is substantial, and this is a sign of overfit.

4.5 Relevant hypothesis tests

Roughly, here are 6 hypothesis test used here:

4.5.1 t-test for lm.1

lm.1(literacy~reGDP+enrollpre+internet+pricompletion+enrollpri+mortality+LNoverage+life+LNhealth)

- α=0.05
- hypotheses:

$$H_0: \beta_{reGDP}=0$$
$$Ha: \beta_{reGDP} \neq 0$$

Assuming that other explanatory variables are already in the model.

- test statistic: -0.059
- p-value: 0.95302 > 0.05
- conclusion: Cannot reject H_0, reGDP is not needed in the model.

The other variables in the same model implement the same hypothesis.

4.5.2 Overall F-test for lm.1

lm.1(literacy~reGDP+enrollpre+internet+pricompletion+enrollpri+mortality+LNoverage+life+LNhealth)

- α=0.05
- hypotheses:

$$H_0: \beta_{reGDP}=\beta_{enrollpre}=\beta_{internet}=\beta_{pricompletion}=\beta_{enrollpri}=\beta_{mortality}=\beta_{LNoverage}=\beta_{life}=\beta_{LNhealth}=0$$

H_a: at least one slope is not zero

- test statistic: F_c=20.57 with k=9 and 58 degrees of freedom
- p-value: $5.239*10^{15} < 0.05$
- conclusion: Reject H_0, at least one slope in the model is not zero.

4.5.3 Partial F-test for lm.2 and lm.3

lm.2 (literacy~mortality+pricompletion+LNoverage)
lm.3 literacy~mortality+pricompletion+LNoverage+life)

- α=0.05
- hypotheses:

$$H_0: \beta_{life}=0$$
$$Ha: \beta_{life} \neq 0$$

- test statistic: F_c=1.0856, with k-g=1 and 62 degrees of freedom
- p-value: 0.3513 > 0.05
- conclusion: Cannot reject H_0, life is not needed in the model.

4.5.4 Partial F-test for lm.3 and lm.4

lm.3 literacy~mortality+pricompletion+LNoverage+life)
lm.4 (literacy~mortality+pricompletion+LNoverage+life+internet)

- α=0.05
- hypotheses:

$$H_0: \beta_{internet}=0$$
$$Ha: \beta_{internet} \neq 0$$

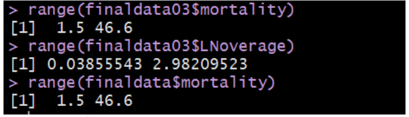
- test statistic: F_c=0.8817, with k-g=1 and 63 degrees of freedom
- p-value: 0.3015 > 0.05
- conclusion: Cannot reject H_0, internet is not needed in the model.

4.5.5 & 4.5.6 T-test and overall F-test for lm.2

lm.2 (literacy~mortality+pricompletion+LNoverage)
And both results are significant.

5.Results:

As I mentioned before, the final model is lm.2, which is
Literacy(%)=29.9647(%)-0.6237(%/(1/1000people))*mortality+0.5804(%/%)*pricompletion+5.0757(%)*LNoverage
The y-interception is 29.9647% and it means when mortality, pricompletion and LNoverage are all 0 (0 of LNoverage which means 100% of population are overage), the literacy rate is 29.9647%. This interpretation to some extent is meaningful because a country with 0 primary completion rate, 0 mortality and 100% overage rate is possible. And 0 is within the range of all variables (maybe a little distance).



```
> range(finaldata03$mortality)
[1]  1.5 46.6
> range(finaldata03$LNoverage)
[1] 0.03855543 2.98209523
> range(finaldata$mortality)
[1]  1.5 46.6
```

The R^2 which is shown before, equals to 73.33%. This means 73.33% of variation in literacy rate can be explained by the combination of mortality, pricompletion and LNoverage. This result is pretty good but when compared to R2jackknife, which is only 0.2892, this value indicates that the model is overfitting and cannot fitted well with new data.
To my surprise, reGDP, the variable that comes to my mind first, is not included in the final model. From my perspective, there are two possible explanations: the correlation between GDP and literacy may be indirect while the intermediary is another variable in the model. Or the result has been changed greatly different from the reality as a result of too many hypothesis tests.
Bigger sample size and larger variety of data would help to make this model better. Except those variables I use the model, there are many other potential relevant variables such as poverty gap and education expenditure (% of GDP). I have to drop them because of too many missing values. Besides, taking time into consideration may also help since all variables included in this model are from year 2014. But in fact, I believe that the data from yesterday will have influence on that of today. For this analysis, autocorrelation should be taken into consideration and therefore it might be much relevant to time series analysis.

6.References

[1] http://data.worldbank.org/indicator/SE.ADT.LITR.ZS?view=chart