Chanat Petchsangngam

Regression Analysis Project

Introduction

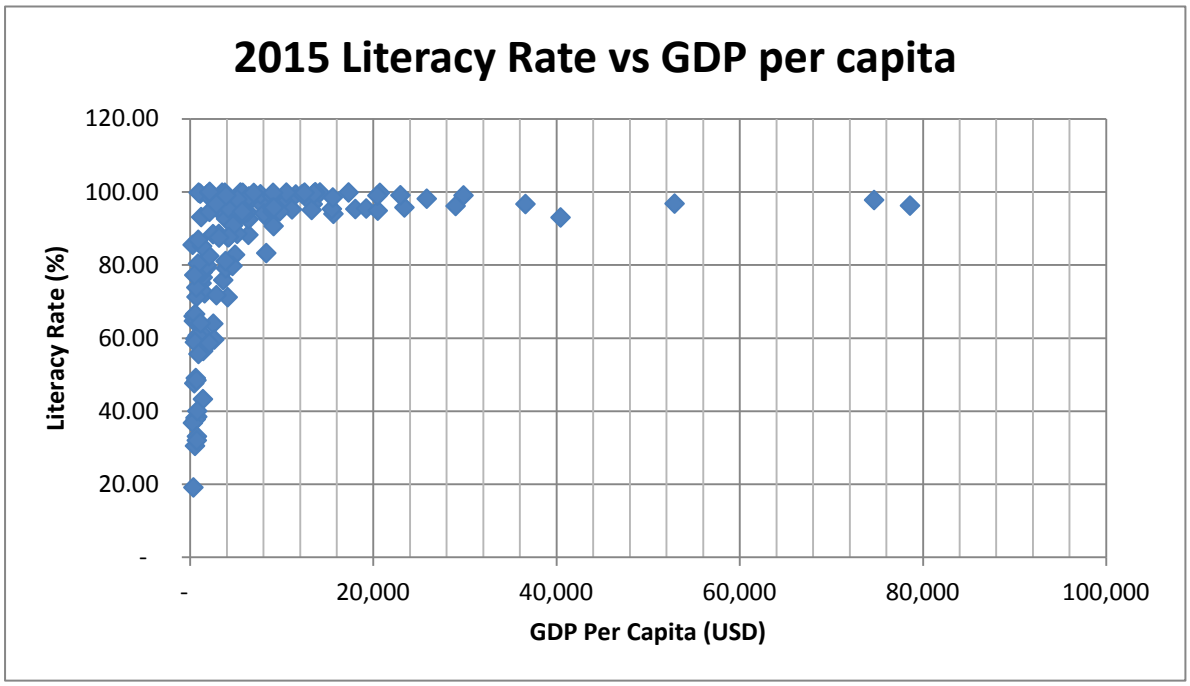
The Regression Analysis Project is written to find the model that best estimate the relation between literacy rate and GDP per capita. The ANOVA is performed together with Tukey and Mosteller's Ladder of Powers to help determine the best models.

Data

The website where the data are obtained is http://data.worldbank.org/. The data required are literacy rate and GDP per capita. The literacy rate is the percentage of population over the age of 15 that could understand, read and write a daily simple statement. The GDP per capita (unit in USD) is the gross domestic product divided by the population collected during midyear.

There are 264 countries listed in the Worldbank database. For literacy rate, there are 156 countries where the data of 2015 are available. For GDP per capita, there are 216 countries where the data of 2015 are available. When we combine the two data available, there are a total of 135 countries in this study.

The scattered plot below shows the relation between literacy rate (independent variable) versus the GDP per capita (dependent variable)



We could see that the data are positively skewed.

Analysis

According to Tukey and Mollester's Ladder of Powers, the linear model is $Y^p = A + BX^q + \varepsilon$, where p and q will depend on the shape of the regression function. For the relationship between literacy rate and GDP per capita, the scattered plot shows a positively skewed data; therefore, the following models will be performed:

1) p = 1 and q = 0.5 ; $Y = A + BX^{0.5} + \varepsilon$
2) p = 2 and q = 1 ; $Y^2 = A + BX + \varepsilon$
3) p = 3 and q = 1 ; $Y^3 = A + BX + \varepsilon$
4) natural log of X ; $Y = A + B\ln(X) + \varepsilon$

X = GDP per capita (USD), Y = literacy rate (percentage)
The null hypothesis is H_o: B = 0. The regression statistic will be performed to tell whether to conclude that B equals to zero or not.

1) $Y = A + BX^{0.5} + \varepsilon$

The tables below show the regression results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.567[a]	0.321	0.316	15.9935

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16,100	1	16,100	63	.000b
	Residual	34,045	133	256		
	Total	50,145	134			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	67.249	2.504	
	$X^{0.5}$	0.222	0.028	0.567

The model is $Y = 67.249 + 0.222X^{0.5} + \varepsilon$

The model shows that when the square roots of GDP per capita increases by 1 unit, literacy rate will increase by 0.222.

The p-value is low; therefore, we can reject the null hypothesis and conclude that B≠0.

The adjusted R square shows that 31.6% of literacy rate could be explained by the GDP per capita.

2) $Y^2 = A + BX + \varepsilon$

The tables below show the regression results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.402a	0.161	0.155	2,525.590

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	163,145,079	1	163,145,079	26	.000b
	Residual	848,354,684	133	6,378,607		
	Total	1,011,499,763	134			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	6,659.762	261.957	
	X	0.092	0.018	0.402

The model is $Y^2 = 6{,}659.762 + 0.092X + \varepsilon$

The model shows that when the GDP per capita increases by 1 unit, the square of literacy rate will increase by 0.092.

The p-value is low; therefore, we can reject the null hypothesis and conclude that $B \neq 0$.

The adjusted R square shows that 15.5% of literacy rate could be explained by the GDP per capita.

3) $Y^3 = A + BX + \varepsilon$

The tables below show the regression results:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.414a	0.171	0.165	287,830.998

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2,279,115,633,130	1	2,279,115,633,130	28	.000b
	Residual	11,018,608,878,606	133	82,846,683,298		
	Total	13,297,724,511,736	134			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	585,115.932	29,854.146	
	X	10.907	2.079	0.414

The model is $Y^3 = 585{,}115.932 + 10.907X + \varepsilon$

The model shows that when the GDP per capita increases by 1 unit, literacy rate to the 3rd power will increase by 10.907.

The p-value is low; therefore, we can reject the null hypothesis and conclude that $B \neq 0$.

The adjusted R square shows that 17.1% of literacy rate could be explained by the GDP per capita.

4) $Y = A + B\ln(X) + \varepsilon$

The tables below show the regression results:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
0.724	0.524	0.520	13.400

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	26,263	1	26,263	146	0
Residual	23,883	133	180		
Total	50,145	134			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
ln(X)	10.954	0.906	0.724	12.094	0.000
(Constant)	-6.338	7.545		-0.840	0.402

The model is $Y = -6.338 + 10.954\ln(X) + \varepsilon$

The model shows that when the ln of GDP per capita increases by 1 unit, literacy rate will increase by 10.954.

The p-value is low; therefore, we can reject the null hypothesis and conclude that $B \neq 0$.

The adjusted R square shows that 52.0% of literacy rate could be explained by the GDP per capita.

Conclusion

Comparing the regression results:

Model	Adjusted R Square	Standard Error
1) $Y = A + BX^{0.5} + \varepsilon$	0.316	15.99935
2) $Y^2 = A + BX + \varepsilon$	0.155	2,525.59036
3) $Y^3 = A + BX + \varepsilon$	0.165	287,830.99781
4) $Y = A + B\ln(X) + \varepsilon$	0.520	13.40037

Since the last model has the lowest adjusted R square and the lowest standard error, this model is suitable to model the relationship between literacy rate and GDP per capita.