NEAS VEE Regression Analysis Student Project Write-Up

Namthip Laohemwong

Note: This write-up should be read with its accompanying excel spreadsheet

Introduction

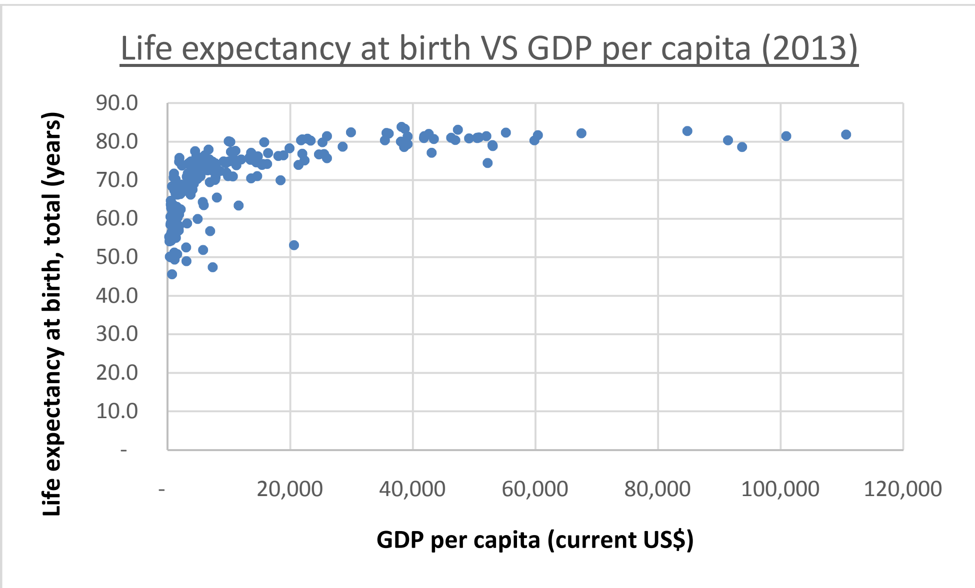
This project is to study the relationship between GDP per capita and overall Expectancy at Birth of countries around the world. It reflects application of ANOVA concepts and The Bulging Rule (Mosteller & Tukey, 1977).

Data

The data used in this project was taken from the Worldbank website: http://data.worldbank.org/. The focus of this project is on the latest data available which is as of 2013. GDP per capita is in million US$, whereas life expectancy at birth is in years.

248 countries are listed out in sheet 'Data' in the accompanying excel spreadsheet. Some countries have missing GDP per capita or Life Expectancy at Birth data. These records are deleted to form a clean list of 213 countries in sheet 'Data_Cut', which is used for the study.

Figure 1: A scattered plot of Life Expectancy against GDP per Capita from 213 countries. With Life Expectancy as the response variable, GDP per Capita as the primary explanatory variable is positively skewed.



Models & Hypothesis

The Mosteller & Tukey Bulging Rule suggests four transformations for such positively-skewed data: ln(X), $X^{0.5}$, Y^2, and Y^3. Therefore, we analyze these transformations as four models listed below. Null hypothesis testing (H_0: β=0) is performed for each model to ensure the strong correlation of Life Expectancy and GDP per Capita. The models are compared for the most suitable one for the specified data. Equations use X representing Life Expectancy and Y representing GDP per Capita.

Model #1: $Y = \alpha + \beta\ln(X) + \varepsilon$

Model #2: $Y = \alpha + \beta X^{0.5} + \varepsilon$

Model #3: $Y^2 = \alpha + \beta X + \varepsilon$

Model #4: $Y^3 = \alpha + \beta X + \varepsilon$

Results & Discussion

Model #1: $Y = \alpha + \beta\ln(X) + \varepsilon$

Regression Statistics	
Multiple R	0.7982372
R Square	0.6371827
Adjusted R Square	0.6354632
Standard Error	5.3934413
Observations	213

ANOVA

	df	SS	MS	F	Significance F
Regression	1	10779.295	10779.295	370.5599	2.41592E-48
Residual	211	6137.8232	29.089209		
Total	212	16917.118			

Fitted Model #1: $Y = 27.83 + 4.9220\ln(X) + \varepsilon$

The adjusted-R2 value of 0.6355 shows that 63.55% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient β=4.9220 implies that if ln(GDP per Capita) increases by 1 unit, Life Expectancy will increase by 4.9220. The relationship between GDP per capita and life expectancy is positive. ln(GDP per Capita) has an extremely low p-value (≈0). This means that we can reject the null hypothesis H_0:β=0 and draw the conclusion that $\beta \neq 0$.

<u>Model #2:</u> $Y = \alpha + \beta X^{0.5} + \varepsilon$

Regression Statistics	
Multiple R	0.7216342
R Square	0.5207560
Adjusted R Square	0.5184847
Standard Error	6.1986957
Observations	213

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8809.6898	8809.6898	229.27673	1.50607E-35
Residual	211	8107.4279	38.423829		
Total	212	16917.118			

Fitted Model #2: $Y = 61.42 + 0.0926X^{0.5} + \varepsilon$

The adjusted-R2 value of 0.5185 shows that 51.85% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient β=0.0926 implies that if $(\text{GDP per Capita})^{0.5}$ increases by 1 unit, Life Expectancy will increase by 0.0926. The relationship between GDP per capita and life expectancy is positive. $(\text{GDP per Capita})^{0.5}$ has an extremely low p-value (≈0). This means that we can reject the null hypothesis $H_0:\beta=0$ and draw the conclusion that $\beta\neq0$.

<u>Model #3:</u> $Y^2 = \alpha + \beta X + \varepsilon$

Regression Statistics	
Multiple R	0.6357963
R Square	0.4042369
Adjusted R Square	0.4014134
Standard Error	930.70522
Observations	213

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	124013564	124013564	143.16765	1.59688E-25
Residual	211	182770776	866212.21		
Total	212	306784340			

Fitted Model #3: $Y^2 = 4{,}495 + 0.0383X + \varepsilon$

The adjusted-R2 value of 0.4014 shows that 40.14% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient β=0.0383 implies that if GDP per Capita increases by 1 unit, $(\text{Life Expectancy})^2$ will increase by 0.0383. The relationship between GDP per capita and life expectancy is positive. GDP per Capita has an extremely low p-value (≈0). This means that we can reject the null hypothesis $H_0:\beta=0$ and draw the conclusion that $\beta\neq0$.

Model #4: $Y^3 = \alpha + \beta X + \varepsilon$

Regression Statistics	
Multiple R	0.6654293
R Square	0.4427961
Adjusted R Square	0.4401554
Standard Error	92576.262
Observations	213

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.437E+12	1.437E+12	167.67649	1.31263E-28
Residual	211	1.808E+12	8.57E+09		
Total	212	3.245E+12			

Fitted Model #4: $Y^3 = 306{,}934 + 4.1272X + \varepsilon$

The adjusted-R2 value of 0.4402 shows that 44.02% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient β=4.1272 implies that if GDP per Capita increases by 1 unit, $(\text{Life Expectancy})^3$ will increase by 4.1272. The relationship between GDP per capita and life expectancy is positive. GDP per Capita has an extremely low p-value (≈0). This means that we can reject the null hypothesis $H_0:\beta=0$ and draw the conclusion that $\beta\neq0$.

Conclusion

Summary of regression analysis can be found as follows:

Model	Adjusted R Square	Standard Error
#1	0.6354632	5.3934413
#2	0.5184847	6.1986957
#3	0.4014134	930.70522
#4	0.4401554	92576.262

Model #1 has the highest adjusted-R2 and lowest standard error. Also the primary explanatory variable is statistically significant at 5% level. Consequently, the most suitable model according to transformations of Mosteller & Tukey Bulging Rule is Model #1.

Selected Model #1: Life Expectancy = 27.83 + 4.9220ln(GDP per Capita) + ε