NEAS VEE Regression Analysis Student Project Write-Up

By Seah Chun Leong, Leon (AXA Healthcare Management)

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

Introduction

This project aims to examine the relationship between GDP per capita and overall Life Expectancy at Birth of countries around the world. In this report, we will demonstrate the knowledge of Heteroscedasticity, Dummy Variable Regression, Interactions and the application of The Bulging Rule (Mosteller & Tukey, 1977).

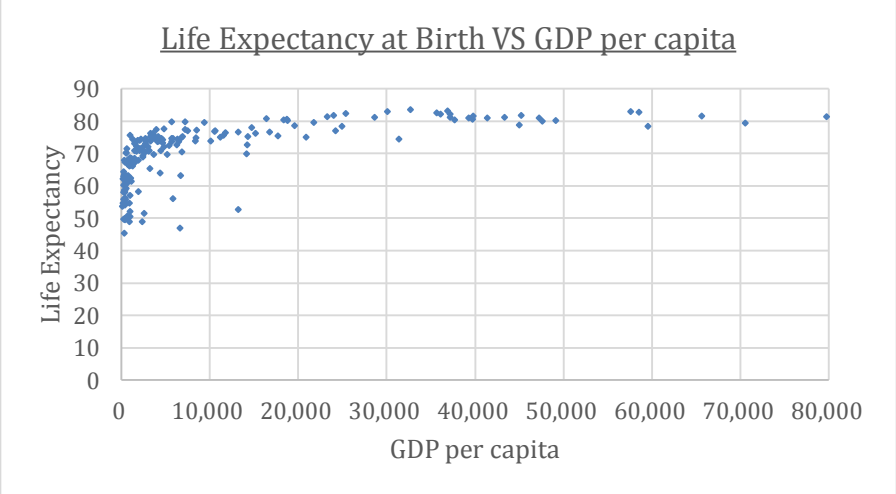
Heteroscedasticity is usually determined in a straightforward fashion. For instance, data may be segregated into "Male" and "Female" or "Smokers and non-Smokers". In this project, we define our own heteroscedastic variable. We will be segregating the countries geographically into "Regions of Africa except the North African region bordering the North Sea" and "The rest of the World". The basis for this segregation is trial and error and also a bit of speculation. We shall not explore the reasons why our method of segregation works, we will instead concentrate solely on the statistical implications.

Data

The data used in this project was taken from the Worldbank website: http://data.worldbank.org/. The geographic categorization of countries was done manually based on the information found on the Internetworldstats website: http://www.internetworldstats.com/list1.htm#geo.

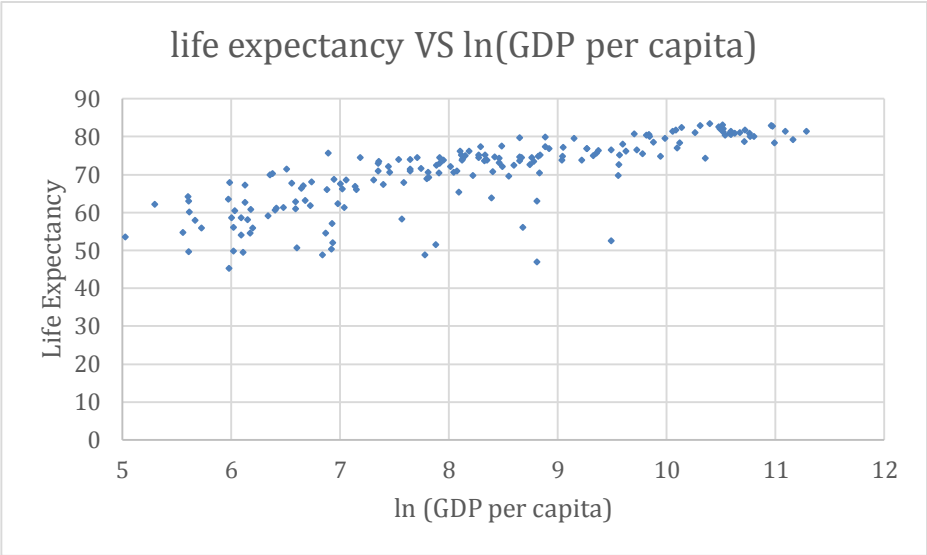
Some rows have missing GDP per capita or Life Expectancy at Birth data as evident from the "Data from World Bank" tab in the accompanying excel spreadsheet. These entries are removed to form the data in the "data cleaned" tab. GDP per capita (*GDP_per_cap*) and Life Expectancy at Birth (*Life_exp*) from a total of 181 countries (after cleaning) is used in this study.

Figure 1 : A preliminary plot of *Life_exp* against *GDP_per_cap* from 181 countries. *GDP_per_cap* is positively skewed



Using the Mosteller & Tukey Bulging Rule, we transform GDP_per_cap down the ladder of power. We could have transform Life_exp up the ladder of power too, but we wish to correct for the positively skewed GDP_per_cap data.

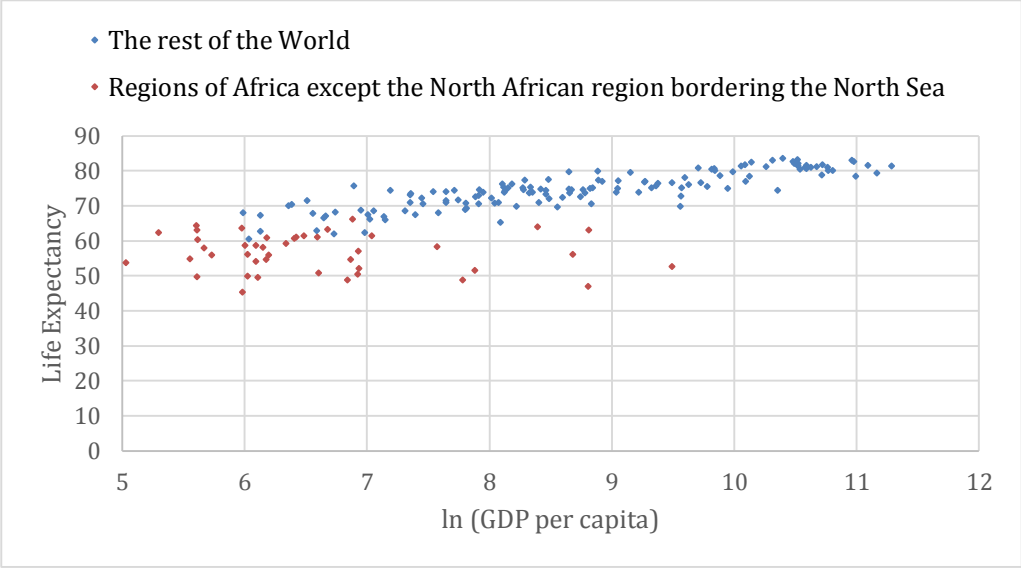
Figure 2 : A plot of Life_exp against ln(GDP_per_cap) from 181 countries. The positive skew in the GDP_per_cap data is corrected. We also get a seemingly linear relationship with a slight hint of Heteroscedasticity.



Models & Hypothesis

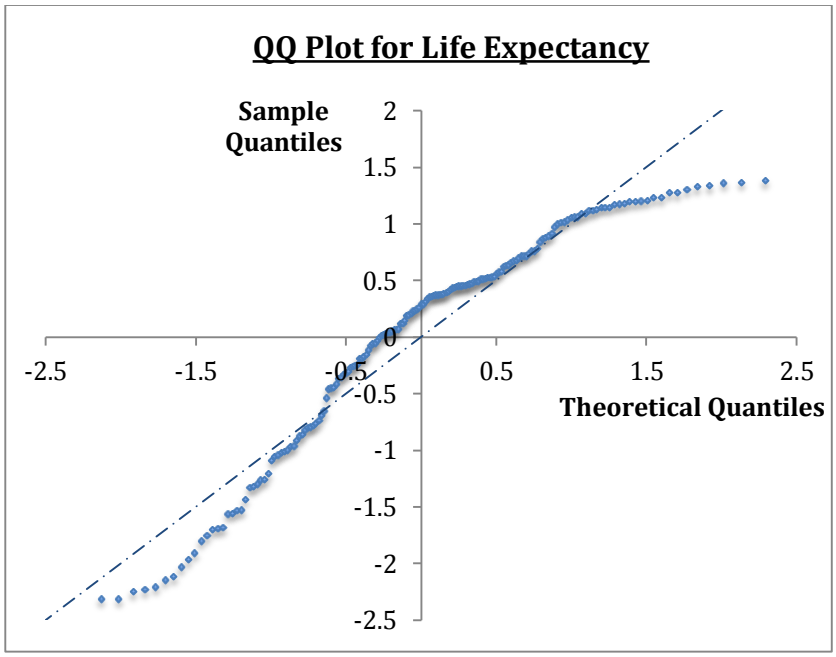
We define the primary explanatory variable as *ln(GDP_per_cap)* and the response variable as *Life_exp*. The secondary explanatory variable (*D*) is heteroscedastic in nature. We define it as a binary variable; 1 if the statistic is for "Regions of Africa except the North African region bordering the North Sea", 0 if the statistic is for "The rest of the World".

Figure 3 : A plot of *Life_exp* against *ln(GDP_per_cap)* from 181 countries taking into account of the Heteroscedastic nature of the data.



First, we examine the QQ plots to determine if the data is distributed appropriately for us to apply classical regression analysis.

Figure 4 : The QQ plot is close enough to the line y = x. We will proceed with the regression analysis.



Because the purpose of this report is to demonstrate the application of regression analysis, we shall delve into the formalities of presenting four different models and present the ANOVA of each. Model #1 regresses *Life_exp* with *ln(GDP_per_cap)*, Model #2 regresses *Life_exp* with *D*, Model #3 regresses *Life_exp* with both *ln(GDP_per_cap)* and *D* and Model #4 is basically Model #3 with consideration of possible interaction between *ln(GDP_per_cap)* and *D*.

1. Model #1 : $Life\_exp = \alpha + \beta \ln(GDP\_per\_cap) + \varepsilon$
2. Model #2 : $Life\_exp = \alpha + \gamma \, \mathrm{D} + \varepsilon$
3. Model #3 : $Life\_exp = \alpha + \gamma \, \mathrm{D} + \beta \ln(GDP\_per\_cap) + \varepsilon$
4. Model #4 : $Life\_exp = \alpha + \gamma \, \mathrm{D} + (\beta + \delta D) \ln(GDP\_per\_cap) + \varepsilon$

Our hypothesis is that *ln(GDP_per_cap)* and D are strongly correlated to *Life_exp* and that the interaction among the explanatory variables is significant, i.e. we expect Model #4 to be the best model.

Results & Discussion

Model #1: $Life\_exp = \alpha + \beta \ln(GDP\_per\_cap) + \varepsilon$

Regression Statistics	
R	0.780270414
R Square	0.608821918
Adjusted R Square	0.606636566
Standard Error	5.80936255
Total number of observations	181

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	1.	9,402.119	9,402.119	278.592	0.E+0
Residual	179.	6,041.016	33.749		
Total	180.	15,443.135			

Fitted Model #1: $Life\_exp = 32.6093 + 4.5881\ ln(GDP\_per\_cap) + \varepsilon$

The adjusted-R^2 value of 0.6088 shows that 60.88% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient $\beta = 4.5881$ implies that if $ln(GDP\_per\_cap)$ increases by 1 unit, $Life\_exp$ will increase by 4.5881. It is obvious that the relationship between GDP per capita and life expectancy is positive. $ln(GDP\_per\_cap)$ 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 $\beta \neq 0$.

Model #2: $Life\_exp = \alpha + \gamma\ \mathrm{D} + \varepsilon$

Regression Statistics	
R	0.82568312
R Square	0.681752614
Adjusted R Square	0.679974696
Standard Error	5.239907817
Total number of observations	181

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	1.	10,528.398	10,528.398	383.456	0.E+0
Residual	179.	4,914.737	27.457		
Total	180.	15,443.135			

Fitted Model #2: $Life\_exp = 74.5502 - 17.9204\ D + \varepsilon$

The adjusted-R^2 value of 0.6818 shows that 68.18% of the Life Expectancy at Birth can be explained by GDP per capita.

The coefficient $\gamma = -17.9204$ implies that the Life Expectancy at Birth in "Regions of Africa except the North African region bordering the North Sea" is, on average, 17.9204 years lower than Life Expectancy at Birth in "The rest of the World". D has an extremely low p-value (≈0). This means that we can reject the null hypothesis $H_0: \gamma = 0$ and draw the conclusion that$\gamma \neq 0$.

Model #3: : $Life\_exp = \alpha + \gamma\ \mathrm{D} + \beta\ \ln(GDP\_per\_cap) + \varepsilon$

Regression Statistics	
R	0.907437045
R Square	0.823441991
Adjusted R Square	0.821458193
Standard Error	3.913825423
Total number of observations	181

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	2.	12,716.526	6,358.263	415.084	0.E+0
Residual	178.	2,726.609	15.318		
Total	180.	15,443.135			

Fitted Model #3: $Life\_exp\ =\ 51.0646\ +\ 2.6952\ ln(GDP\_per\_cap)\ -\ 12.2437\ D + \varepsilon$

The adjusted-R^2 value of 0.8215 shows that 82.15% of the Life Expectancy at Birth can be explained by GDP per capita and whether the country is in continental Africa not bordering the North Sea.

The coefficient $\gamma = -12.2437$ implies that the Life Expectancy at Birth in "Regions of Africa except the North African region bordering the North Sea" is, on average, 12.2437 years lower than Life Expectancy at Birth in "The rest of the World". This is slightly lower than 17.9204 from Model #2. This model has a statistically significant F-statistic of 415.084 with 2 and 178 degrees of freedom, with an extremely low p-value (≈0). We reject the null hypothesis $H_0: \beta = \gamma = 0$.

We now re-express the heteroscedastic regression model as follows:

$D = 0$ as $Life\_exp\ =\ 51.0646\ +\ 2.6952\ ln(GDP\_per\_cap) + \varepsilon$
$D = 1$ as $Life\_exp\ =\ 38.8209\ +\ 2.6952\ ln(GDP\_per\_cap) + \varepsilon$

Model #4: : $Life\_exp = \alpha + \gamma\ \mathrm{D} + (\beta + \delta D)\ \ln(GDP\_per\_cap) + \varepsilon$

Regression Statistics	
R	0.927718943
R Square	0.860662437
Adjusted R Square	0.858300784
Standard Error	3.486705058
Total number of observations	181

ANOVA

	d.f.	*SS*	*MS*	*F*	*p-level*
Regression	3.	13,291.327	4,430.442	364.432	0.E+0
Residual	177.	2,151.809	12.157		
Total	180.	15,443.135			

Fitted Model #4: $Life\_exp =\ 46.0473\ +\ 14.6722\ D + (3.2710 - 3.8900\ D)\ ln(GDP\_per\_cap)\ + \varepsilon$

We have a statistically significant model. This model has an F-statistic of 364.432 with 3 and 177 degrees of freedom, with an extremely low p-value (≈0). Thus we can reject the null hypothesis $H_0: \beta = \gamma = \delta = 0$. This model assumes that countries in "Regions of Africa except the North African region bordering the North Sea" and countries in "The rest of the world" have different slopes and intercepts.

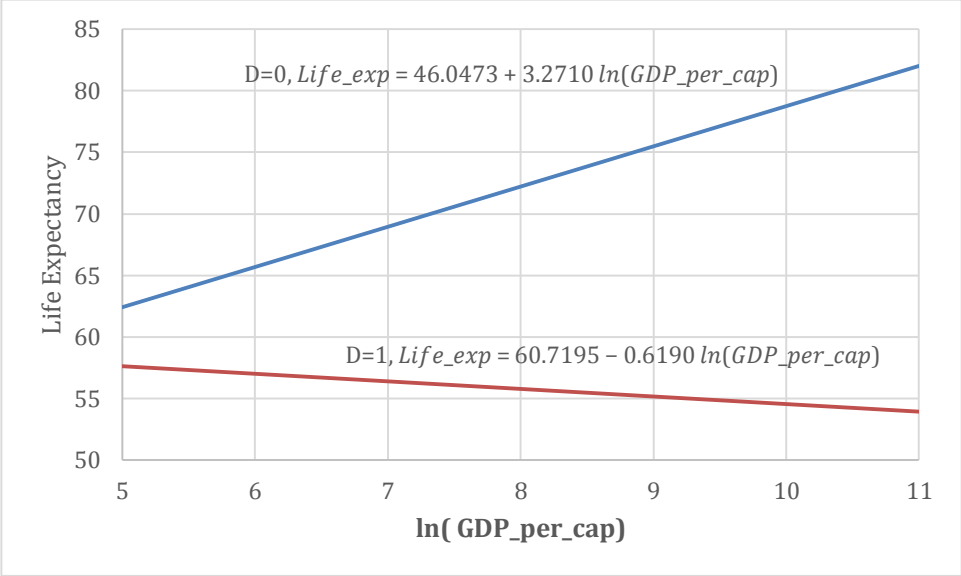
The adjusted-R^2 value is 0.8583, the highest out of all models tested.

The interaction variable $D \ln(GDP\_per\_cap)$ introduced in this model is significant, having an extremely low p-value (≈ 0)

We now re-express the heteroscedastic regression model as follows:

$D = 0$ as $Life\_exp = 46.0473 + 3.2710\ ln(GDP\_per\_cap) + \varepsilon$
$D = 1$ as $Life\_exp = 60.7195 - 0.6190\ ln(GDP\_per\_cap) + \varepsilon$

It is unfortunate that the equation for $D = 1$ has a negative gradient, even though it is close to zero. This means that life expectancy decreases as GDP per capita increase, a counter intuitive conclusion. We attribute this to the influence of outliers. A separate analysis of cook's coefficient could be done, but due to time constraint, we will have to skip it.



Conclusion

The following is a high level summary of our analysis:

Model	Adjusted-R^2	Standard Error
#1	0.60664	5.8094
#2	0.67997	5.2399
#3	0.82146	3.9138
#4	0.8583	3.4867

Model #4 has the highest adjusted-R^2 and lowest standard error. Also all models are statistically significant at 5% level.

Chosen Model #4: $Life\_exp = 46.0473 + 14.6722\ D + (3.2710 - 3.8900\ D)\ ln(GDP\_per\_cap) + \varepsilon$

We note that the data is probably plague with marginal outliers since the gradient for D=1 is negative, i.e. $3.2710 - 3.8900 < 0$. We also note that the model is only good for age ranging from about 55 to 85 and logarithm of GDP per capita ranging from 5 to 11.