Regression Analysis Project			
Name	MAE ELIZABETH SENADOR LUNGAY	Registration ID	58233099
Course	Regression Analysis	Contact Nos.	+639167831414
Session	Summer 2013	Email	mae.lungay@gmail.com

Maternal Mortality Rates

1|

Introduction

Maternal mortality rate is defined as the number of women who die during pregnancy and childbirth. For the year 2010, The World Bank estimates this figure at 210 per 100,000 of live births. Several factors may be accountable for this ratio including but not limited to: government policies on women's health issues, availability of health programs for women, accessibility and quality of education for the general population and the economic status of the country as a whole.

A national government whose domestic policies include health and education ensures that an enabling environment and well-defined policies are in place to create programs essential to the delivery and success of these agenda. The proportion of the national budget allocated for health and education reflects the level of priority these two issues are to the government. Good health program means quality, affordable and sustainable programs that are available for its citizens especially to those who are vulnerable: women, children and the elderly. Good education means quality and affordable education for the people.

This study attempts to develop a robust regression model for maternal mortality rate for 80 countries from the following explanatory variables representing the government, health, education and economic sectors:

Government
- *Proportion of seats held by women in national parliaments*
- *Labor participation rate of females*
- *Health expenditure*

Health
- *Contraceptive Prevalence*
- *Improved sanitation*
- *Adolescent fertility rate*

Education
- *Secondary school enrolment*
- *Female repeaters in primary levels*

Economy
- *Classification of country by income*

2|

Data and Data Analysis

The data are obtained from The World Bank database found in this website: http://data.worldbank.org/. There are 214 countries listed as part of The World Bank member economies. However, only countries with complete information across all variables are included in the model, thus, reducing the number to 80. The final data are shown in the Data page in the accompanying Excel worksheet. The standard definitions of the variables and the coded variable names are given in the table below:

Table 1. Definition of Variables
Maternal Mortality Ratio (per 100,000 live births) *Code: MMort* Maternal mortality ratio is the number of women who die during pregnancy and childbirth, per 100,000 live births.
Proportion of Seats held by Women in National Parliaments (%) *Code: Parliament* Women in parliaments are the percentage of parliamentary seats in a single or lower chamber held by women.
Labor Participation Rate, Female (% of female population ages 15+) *Code: Labor* Labor force participation rate is the proportion of the population ages 15 and older that is economically active: all people who supply labor for the production of goods and services during a specified period.
Contraceptive Prevalence *Code: Contraception* Contraceptive prevalence is the percentage of women who are currently using, or whose sexual partner is currently using, at least one method of contraception, regardless of the method used. It is usually reported for married or in union women aged 15 to 49.
Health Expenditure, Total (% of GDP) *Code: Health* Total health expenditure is the sum of public and private health expenditure. It covers the provision of health services (preventive and curative), family planning activities, nutrition activities, and emergency aid designated for health but does not include provision of water and sanitation.
School Enrolment, Secondary (% gross) *Code: SecSchEnrol* Total is the total enrollment in secondary education, regardless of age, expressed as a percentage of the population of official secondary education age. GER can exceed 100% due to the inclusion of over-aged and under-aged students because of early or late school entrance and grade repetition.
Repeaters, Primary, Female (% of female enrollment) *Code: Repeaters* Female is the number of female students enrolled in the same grade as in the previous year, as a percentage of all female students enrolled in primary school.
Improved Sanitation Facilities (% of population with access) *Code: Sanitation* Access to improved sanitation facilities refers to the percentage of the population with at least adequate access to excreta disposal facilities that can effectively prevent human, animal, and insect contact with excreta. Improved facilities range from simple but protected pit latrines to flush toilets with a sewerage connection. To be effective, facilities must be correctly constructed and properly maintained.
Adolescent Fertility Rate (births per 1,000 women ages 15-19) *Code: Fert* Adolescent fertility rate is the number of births per 1,000 women ages 15-19.
Classification of Country by Income *Code: Income* *Classification of The World Bank member economies according to 2012 gross national income (GNI) per capital using the World Bank Atlas method. The groups are: low income (Low I), $1,035 or less; lower middle income (Lower), $1,036–4,085; upper middle income (Upper), $4,086–12,615; and high income (High) , $12,616 or more.*

Since Income is a polytomous factor with four categories, three dummy regressors will be introduced in the regression equation. The coding scheme is as follows:

Table 2. Dummy Regressors for Income			
Category	D_1	D_2	D_3
Low (Low I)	1	0	0
Lower Middle (Lower)	0	1	0
Upper Middle (Upper)	0	0	1
High (High)	0	0	0

Intuitively, some of these variables are dependent on other variables. For instance, those who are in school may be more inclined to adopt a better sanitation facility that those who are not. Also, people who are in school may be more aware of the benefits of family planning and hence may be more inclined to use any form of contraception. In order to identify the extent of these dependencies, the correlations between the explanatory variables are computed. Table 3 shows the results. The correlations will be considered in during the model selection. If two variables are highly correlated, absolute correlation is above 0.5, one of them may be removed from the model.

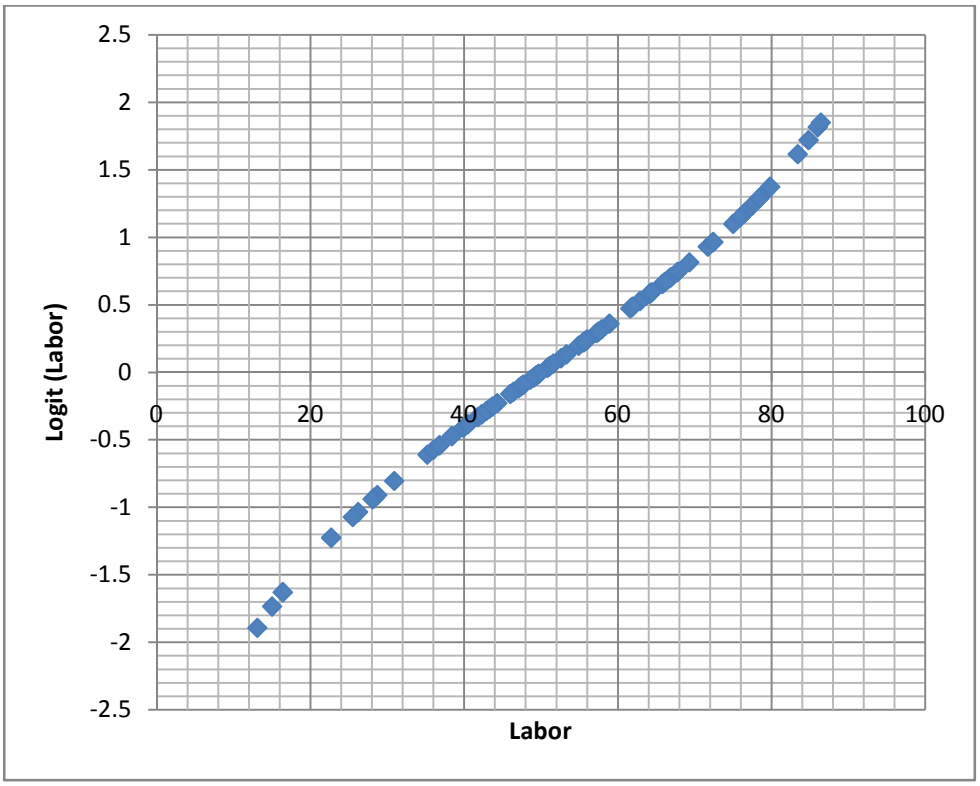
Table 3. Correlations

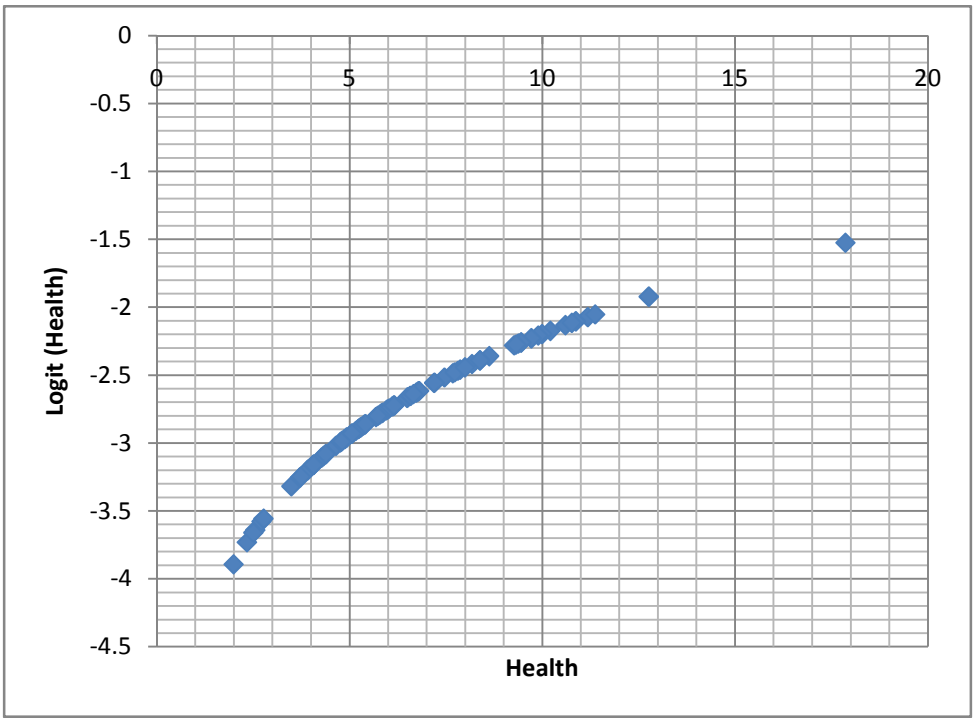
	Logit (Health)	Logit (Labor)	Parliament	Repeaters	Contraception	Sec Sch Enrol	Low I	Lower	Upper	Logit (Fert)	Sanitation
Logit (Health)	1.0000										
Logit (Labor)	-0.0254	1.0000									
Parliament	0.2874	0.2100	1.0000								
Repeaters	-0.1701	0.2844	0.1318	1.0000							
Contraception	0.3300	-0.1624	-0.0535	-0.4434	1.0000						
SecSchEnrol	0.2571	-0.3098	-0.1079	-0.7091	0.6987	1.0000					
Low I	-0.1695	0.5333	0.0889	0.4175	-0.4467	-0.5775	1.0000				
Lower	-0.0884	-0.2354	-0.0840	0.0883	-0.0876	-0.1855	-0.3812	1.0000			
Upper	-0.0362	-0.2081	0.0140	-0.1905	0.2688	0.3066	-0.3099	-0.4378	1.0000		
Logit (Fert)	-0.0657	0.3340	0.0669	0.4903	-0.3872	-0.6110	0.3556	0.1505	-0.1077	1.0000	
Sanitation	0.2746	-0.3852	-0.0831	-0.6528	0.6259	0.8570	-0.5137	-0.2009	0.2745	-0.6063	1.0000

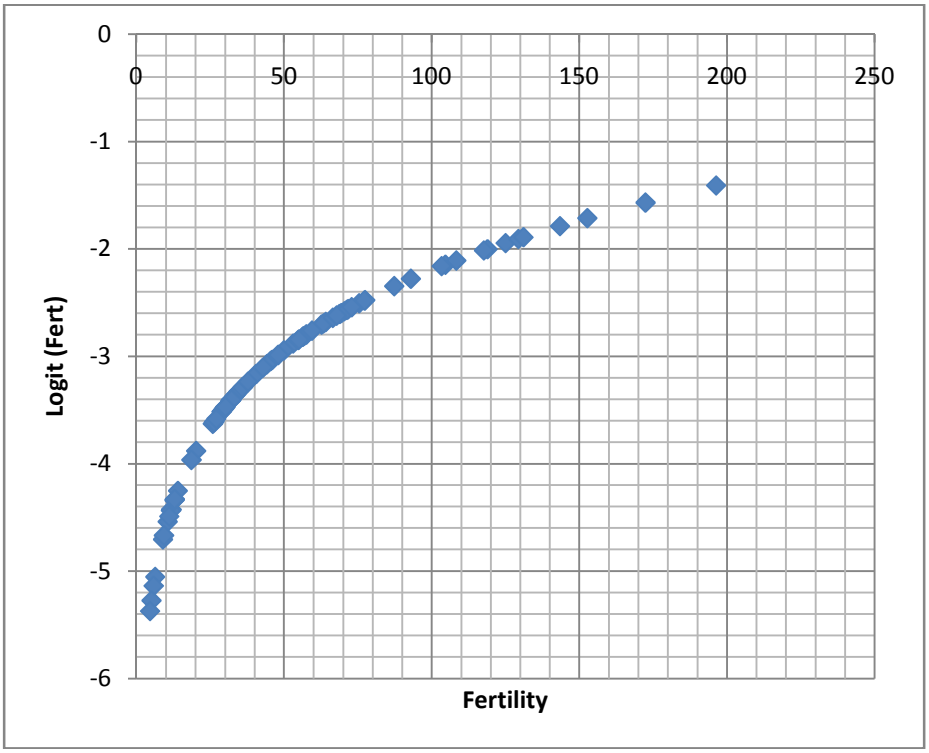
The quartiles of the variables are computed to determine its skewness. See Table 4. Notice that Labor, Health, Fert, Repeaters and MMort are positively skewed while SecSchEnrol is negatively skewed. However, since the variables are given as proportions, power transformations are unhelpful. Instead, logit transformation is used to somehow even out the tails of the distribution and make the resulting quantities symmetric about 0. Out of the six positively and negatively skewed variables, only Labor, Health, Fert and MMort were transformed since Repeaters and SecSchEnrol have values of 0, 1 and above 1.

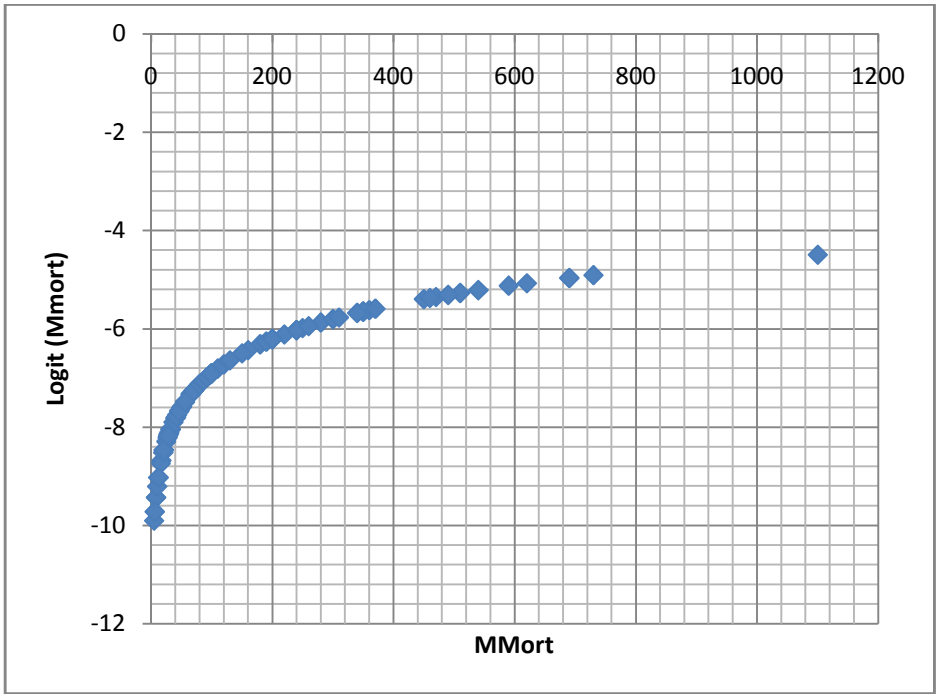
Table 4. Quartiles and Skewness				
Variable	Q_3 (%)	Q_2 (%)	Q_1 (%)	$(Q_3-Q_2)/(Q_2-Q_1)$
Labor	64.75	52.10	43.20	1.42
Parliament	23.90	19.00	12.475	0.75
Contraception	68.10	53.10	35.35	0.85
Health	8.04	6.00	4.80	1.69
Sec Sc Enrol	91.02	79.76	44.59	0.32
Fert	70.59	44.83	26.88	1.44
Repeaters	8.50	3.93	0.34	1.27
Sanitation	95.00	76.50	49.75	0.69
MMort	265.00	94.50	35.75	2.90

The graphs of the logit transformations reveal that the transformations are nearly linear in its median. However, only logit(Labor) appears to be symmetric with its tails evened out. The other graphs reveal the presence of longer tails.









However, the skewness of each of the variables is greatly reduced as seen in Table 5.

Table 5. Quartiles, Skewness of Transformed Variables

Variable	Q_3	Q_2	Q_1	$(Q_3-Q_2)/(Q_2-Q_1)$
Logit (Health)	-2.44	-2.75	-2.99	1.32
Logit (Fert)	-2.58	-3.06	-3.59	0.91
Logit (MMort)	-5.93	-6.96	-7.94	1.06

Employing these transformations and introducing the dummy regressors for Income, we have the following variables for our regression model:

Table 6. Final Explanatory Variables

Variable	Base Class
Logit (Labor)	N/A
Parliament	N/A
Contraception	N/A
Logit (Health)	N/A
Sec Sc Enrol	N/A
Logit (Fert)	N/A
Repeaters	N/A
Sanitation	N/A
Logit (MMort)	N/A
Low I	High
Lower	High
Upper	High

3|

Regression Analysis and Model Consideration

In this section, we develop an appropriate regression model for maternal mortality. A model containing all 9 explanatory variables will be considered first. Any variable which is not significant will be removed gradually from the model and the model will be tested again. The Regression add-in in Excel will be used in performing the regression. The results of these processes are shown in Models 1 to 5.

Model 1: The Full Model (9-Variable Model)

Regression Statistics	
Multiple R	0.917983
R Square	0.842692
Adjusted R Square	0.817246
Standard Error	0.564177
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	115.9469	10.54063	33.11582	4.89E-23
Residual	68	21.64412	0.318296		
Total	79	137.5911			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-6.29025	0.705114	-8.92089	4.69E-13	-7.69728	-4.88321	-7.69728	-4.88321
Logit(Health)	-0.46563	0.170995	-2.72304	0.008212	-0.80684	-0.12441	-0.80684	-0.12441
Logit(Labor)	0.097464	0.106876	0.911936	0.365024	-0.1158	0.310732	-0.1158	0.310732
Parliament	0.328396	0.697249	0.470988	0.639157	-1.06294	1.719735	-1.06294	1.719735
Repeaters	1.947057	1.628593	1.195545	0.236027	-1.30275	5.196863	-1.30275	5.196863
Contraception	-0.21504	0.416035	-0.51688	0.606916	-1.04523	0.615144	-1.04523	0.615144
SecSchEnrol	-1.85901	0.577185	-3.22082	0.001962	-3.01076	-0.70725	-3.01076	-0.70725
Low I	0.519129	0.310425	1.672317	0.099057	-0.10031	1.138573	-0.10031	1.138573
Lower	0.776345	0.245204	3.166119	0.002313	0.287048	1.265643	0.287048	1.265643
Upper	0.473849	0.223332	2.121727	0.037508	0.028197	0.919501	0.028197	0.919501
Logit(Fert)	0.314642	0.096054	3.27569	0.001661	0.12297	0.506314	0.12297	0.506314
Sanitation	-0.42993	0.457476	-0.93978	0.350656	-1.34281	0.482951	-1.34281	0.482951

The adjusted R^2 indicates that 81.72% of the response variable can be explained by the model. The p-value indicates that several variables are not significant at 5% level of significance. We remove the top 4 of these variables: Logit (Labor), Parliament, Contraception and Sanitation and run a regression analysis again. Notice that Contraception and Sanitation are correlated with each other and with SecSchEnrol.

Model 2: The 5-Variable Model

Regression Statistics	
Multiple R	0.914133
R Square	0.83564
Adjusted R Square	0.81966
Standard Error	0.560438
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	114.9766	16.42522	52.29458055	1E-25
Residual	72	22.61451	0.31409		
Total	79	137.5911			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-6.26247	0.619823	-10.1036	1.90684E-15	-7.49806	-5.02687	-7.49806	-5.02687
Logit(Health)	-0.47657	0.153801	-3.09859	0.002772973	-0.78317	-0.16997	-0.78317	-0.16997
Repeaters	2.218598	1.590204	1.395166	0.167254674	-0.95142	5.388613	-0.95142	5.388613
SecSchEnrol	-2.26348	0.432948	-5.22807	1.60808E-06	-3.12655	-1.40042	-3.12655	-1.40042
Low I	0.605298	0.300818	2.012171	0.047943897	0.005627	1.204968	0.005627	1.204968
Lower	0.750028	0.234053	3.204521	0.002017109	0.283452	1.216605	0.283452	1.216605
Upper	0.436679	0.212844	2.051642	0.043842559	0.012383	0.860975	0.012383	0.860975
Logit(Fert)	0.350967	0.091449	3.837848	0.000264055	0.168667	0.533267	0.168667	0.533267

The adjusted R^2 has not greatly improved even when the four variables were removed. Next, we remove Repeaters from the model. The results are shown in Model 3.

Model 3: The 4-Variable Model

Regression Statistics	
Multiple R	0.9117
R Square	0.831196
Adjusted R Square	0.817322
Standard Error	0.564059
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	114.3652	19.06087	59.90917	3.18E-26
Residual	73	23.22588	0.318163		
Total	79	137.5911			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.87541	0.557853	-10.5322	2.67E-16	-6.98721	-4.76361	-6.98721	-4.76361
Logit(Health)	-0.47803	0.154792	-3.08822	0.002847	-0.78653	-0.16953	-0.78653	-0.16953
SecSchEnrol	-2.57206	0.374594	-6.86627	1.84E-09	-3.31863	-1.8255	-3.31863	-1.8255
Low I	0.587389	0.302486	1.941869	0.056012	-0.01547	1.190244	-0.01547	1.190244
Lower	0.726799	0.234969	3.093172	0.002805	0.258507	1.195091	0.258507	1.195091
Upper	0.429409	0.214155	2.005134	0.048657	0.002599	0.856219	0.002599	0.856219
Logit(Fert)	0.364863	0.091492	3.987909	0.000157	0.182519	0.547206	0.182519	0.547206

The adjusted R^2 for this model is 81.73%, not significantly higher than the full model. Next, we try to remove the dummy regressors for Income and see what happens.

Model 4: The 3-Variable Model

Regression Statistics	
Multiple R	0.898839
R Square	0.807912
Adjusted R Square	0.80033
Standard Error	0.58971
Observations	80

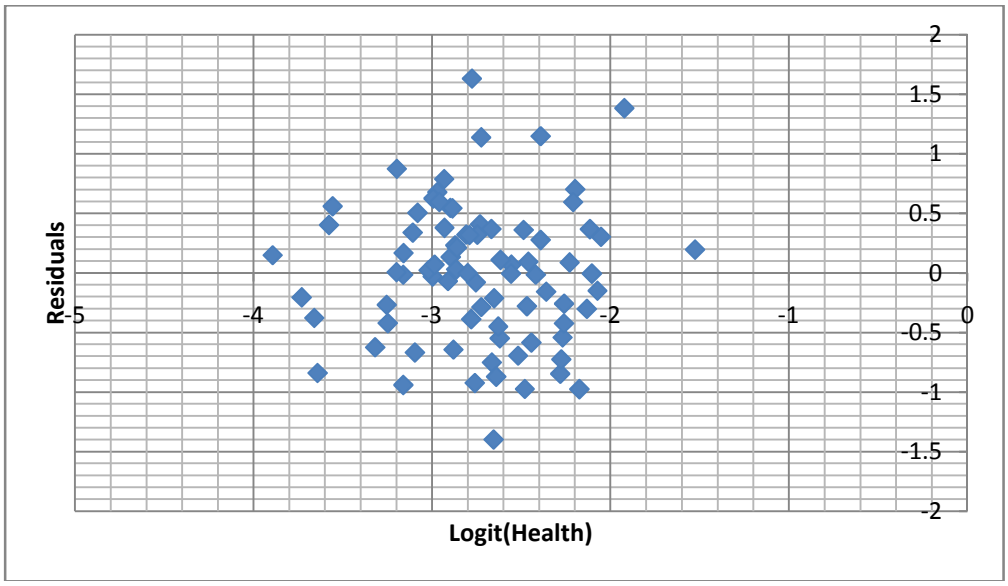
ANOVA

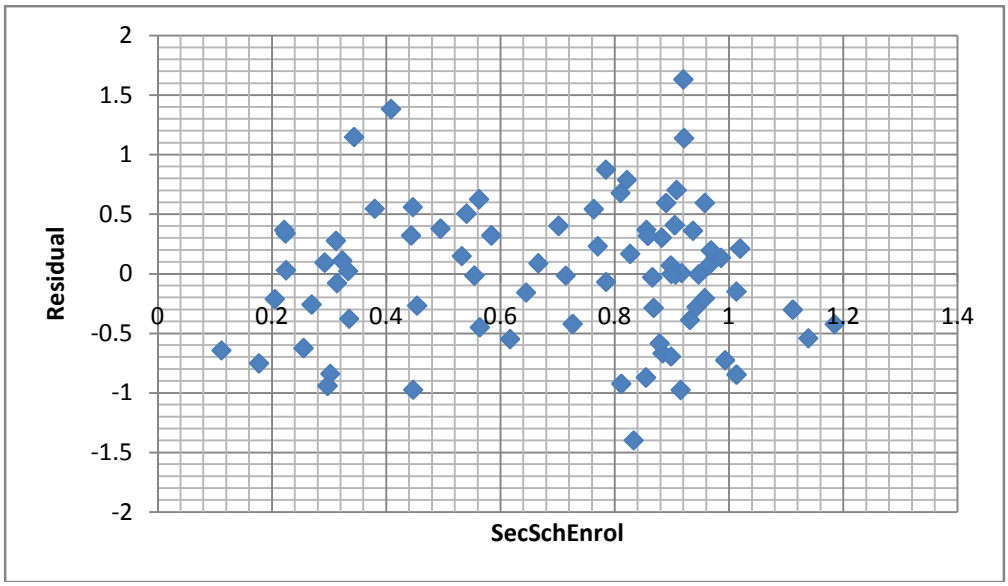
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	111.1615	37.05384	106.5508	3.75E-27
Residual	76	26.42957	0.347757		
Total	79	137.5911			

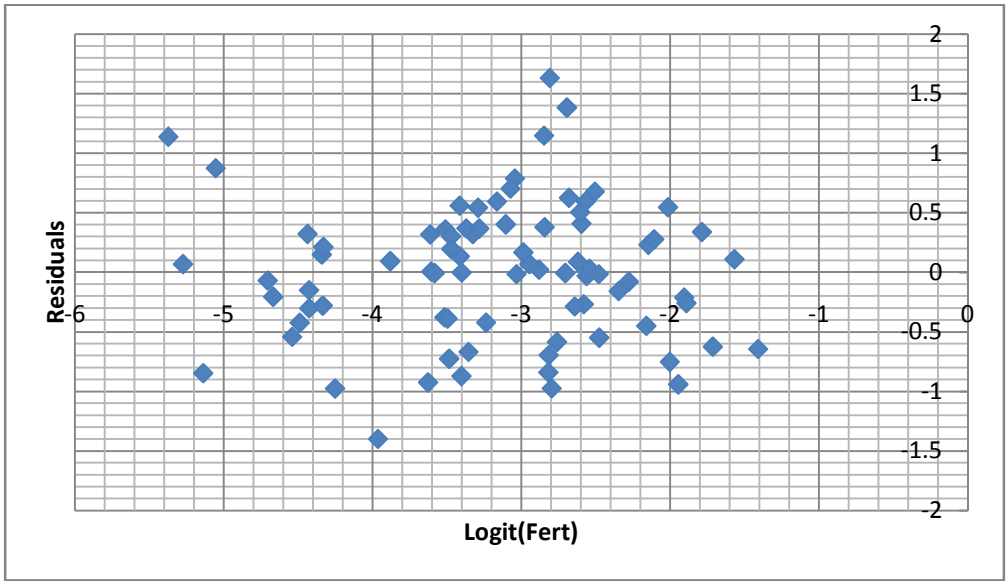
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-5.25296	0.5097	-10.306	4.39E-16	-6.26811	-4.2378	-6.26811	-4.2378
Logit(Health)	-0.58859	0.155168	-3.79327	0.000297	-0.89763	-0.27955	-0.89763	-0.27955
SecSchEnrol	-2.93144	0.310391	-9.44434	1.89E-14	-3.54963	-2.31324	-3.54963	-2.31324
Logit(Fert)	0.423172	0.092684	4.565736	1.88E-05	0.238575	0.607769	0.238575	0.607769

The adjusted R^2 implies that 80% of the variation is accounted by the model. The decrease in the value of the adjusted R^2 from the full model is also negligible considering that that all the variables are now significant at 1% level of significance.

The residual plots for Model 4 are generated below:







Note that the residual plots show a fairly random pattern which means that linear regression is an appropriate analysis tool for the data.

Lastly, let us try to further simplify the model by removing the variable Health. The results are shown below.

Model 5: The 2-Variable Model

Regression Statistics	
Multiple R	0.878376
R Square	0.771545
Adjusted R Square	0.765611
Standard Error	0.638926
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	106.1577	53.07883	130.0232	2.06E-25
Residual	77	31.4334	0.408226		
Total	79	137.5911			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-3.55149	0.262274	-13.5411	4.58E-22	-4.07375	-3.02924	-4.07375	-3.02924
SecSchEnrol	-3.25474	0.323368	-10.0651	1.09E-15	-3.89865	-2.61083	-3.89865	-2.61083
Logit(Fert)	0.38119	0.099701	3.823329	0.000266	0.18266	0.57972	0.18266	0.57972

Notice that the adjusted R^2 is reduced to 76.55%, an almost 4% decline from that of Model 4. At this point, the model has been reduced to its simplest form, with all the variables significant at 1% level of significance. Nevertheless, since Model 4 has a higher adjusted R^2 than Model 5, we will use it as the final regression model.

Thus, Maternal Mortality can be predicted using the following model:

$$Logit(MMort) = -5.25296 - 0.58859\, Logit(Health) - 2.93144\, SecSchEnrol + 0.423172\, Logit(Fert)$$

or

$$\ln\frac{MMort}{1-MMort} = -5.25296 - 0.58859 \ln\frac{Health}{1-Health} - 2.93144\, SecSchEnrol + 0.423172 \ln\frac{Fert}{1-Fert}$$

4|

Conclusion

The final regression model depicts the effects of health, secondary school enrolment and fertility on maternal mortality. The model,

$$\ln\frac{MMort}{1-MMort} = -5.25296 - 0.58859 \ln\frac{Health}{1-Health} - 2.93144\, SecSchEnrol + 0.423172 \ln\frac{Fert}{1-Fert}$$

shows that, with all other factors being kept constant:

- Odds(Health) has a negative effect on the odds(MMort). Recall that $\text{odds(Health)} = \frac{Health}{1-Health} = \frac{success}{failure}$ or the ratio between the statistic Health and its complement. This implies that the higher the percentage of GDP allotted for health expenditures, the lower the odds(MMort) will be. And the lower the percentage of GDP allotted for health expenditures, the higher the odds(MMort) will be. In particular, if the allocation for health exceeds 50% of the GDP, odds(Health) exceeds 1, ln odds(Health) is positive, and odds(MMort) will be reduced by the factor $\left(\frac{Health}{1-Health}\right)^{-0.58859}$

- Likewise, SecSchEnrol has a negative effect on odds(MMort). This means that a higher number of secondary enrolment reduces the odds(MMort) by the factor $e^{-2.93144\, SecSchEnrol}$

- On the other hand, odds(Fert) has a positive effect on the odds(MMort). This means that the lower the number of births, the lower the odds(MMort) will be. In particular, if the odds(Fert) is less than 1, the odds(MMort) will be reduced by the factor $\left(\frac{Fert}{1-Fert}\right)^{0.423172}$

The resulting model supports our intuition above. High spending on health mitigates the risk of death during labor. This may be due to the provision of better facilities, better health related programs for mothers, more intensive educational campaigns related to women's health, pregnancy, and motherhood and higher outreach especially in rural areas for these campaigns. This may also translate to the availability of doctors, nurses and midwives in government hospitals and public health centers especially in rural areas.

Education is also an important factor in maternal mortality. Most secondary schools have a subject which discusses about the importance of healthy diet and nutrition, exercise, hygiene, pregnancy and its risks. According to UNESCO Education for All report (Education Counts Brochure 2011), a child whose mother can read is 5% more likely to live past age 5. This simply shows that proper education gives the person the ability to make informative decisions and better judgments. Also, education also opens the way for better opportunities in life which may lead to an improved economic and financial status.

Fertility rate is also another significant factor in maternal mortality. Higher number of children per mother exposes the mother to complications arising from pregnancy, gives the body less time to heal. More children in the family may also translate to poorer household conditions, diet and environment.