Final Project Report

Regression Analysis
STAT-741

Demographic data analysis using multiple linear regression

Team Members:
Ajinkya Jadhav
Jay Dhanki

Contents

1. Introduction:

Fertility and Child Mortality are primary determinants of population growth of a country. Each country's Life expectancy index is based on their GDP as well as Fertility, Child Mortality and Population. Therefore, we have taken the demographic data from stat-crunch website for all Countries to analyze the life expectancy around the world. Our dataset contains four continuous predictors namely Fertility, GDP, Population, Child Mortality, and a categorical variable - Geographical region and a response - Life expectancy. Fertility factor describes number of children per couple or individual. In the demography study, Fertility factor can have positive as well as negative values where negative value indicates decreasing fertility in a population of country. Our data set contains no negative value for fertility.

Second predictor – Child mortality (Child Death) has been treated as an index of general development of country in demographic study. Child Mortality Index is measured through child death under the age of five per 1000 live births in a country. Third Predictor, Gross Domestic Product (GDP) is the indicator of economic health or standard of living of a country that is the annual monetary values of all the private, public, government consumption goods, services, investments, and exports occurred within a country's borders in a specific time. In our data set, GDP is the average value of what each people spent or invested throughout the year in a country. Higher value of GDP indicates better standard of living for a population (Fourth Indicator) in that particular country.

Our Response – Life Expectancy is the average number of years to be lived by a population born in the same year. Therefore, we have analyzed the life expectancy of a country in that particular continent through the rate of their various indexes such as Fertility, Child Mortality, GDP and Populations. Our analysis helps to understand which factor causes major changes in life expectancy and how one country can make better life expectancy through improving predictor indexes in their particular continent or demographic region.

2. Plan:

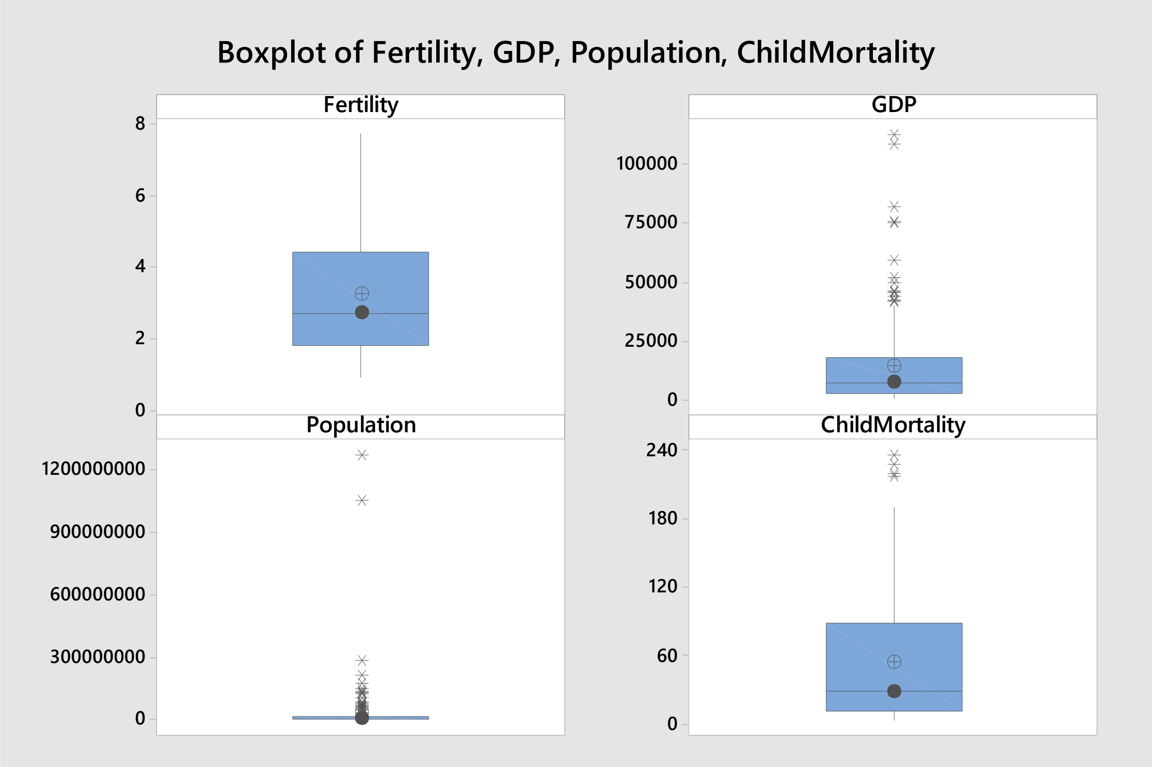
We are going to analyze the data using multiple regression analysis as a tool to determine how much individual predictors weighted in predicting life expectancy for each country in a particular region. This analysis acts as a good input in identifying the broad focus area to improve the life expectancy of a country in the region. We are implementing five number summary to standardize display of data distribution through whisker diagram and ANOVA test to analyze difference in variation in predicting Life Expectancy. Hence, through analysis of variation test, we are estimating how much one predictor is contributing in predicting life expectancy and how they correlate with each other. While doing primary analysis, we found that our data set contains outliers and influential case. Therefore, we will find unusual values of response and predictors in terms of residuals and leverage values. Then we will remove those cases and Re-run the regression model to check whether model accuracy increase or not. We will also check cook's distance to estimate influence data point in our regression model. Our data set contains quantitative variables and a qualitative variable. Therefore, we have divided data in 6 continents and will be implementing dummy variables to check any model discrepancies and ways to correct them. Hence, we can compare life expectancy among continents as well. In the dataset, predictors (GDP & Child Mortality) contain missing values too. Therefore, we will use mean method to replace missing values with mean of the predictor grouped by geographical region. Further, we will discuss results after each analysis on data set.

3. Data Analysis & Results:

Analysis of the distribution of the data using the Box-plot:

```
Box Plot :

Variable          N       Mean      StDev  SE Mean          95% CI
Fertility       201      3.238      1.741    0.123  (    2.995,     3.480)
GDP             201      14000      18000     1270  (    11496,     16504)
Population      201   30444369  119984452  8463051  (13756112, 47132627)
ChildMortality  201      54.67      56.48     3.98  (    46.82,      62.53)
LifeExpectancy  201     68.329      9.334    0.658  (   67.031,    69.627)
```

It gives us a visual idea of the existing data. We can see the spread of the data and also the least value, 25th percentile, median, 75th percentile and largest value in the data set. We also see some of the outliers in the diagram below. Other details are also provided in the table below.

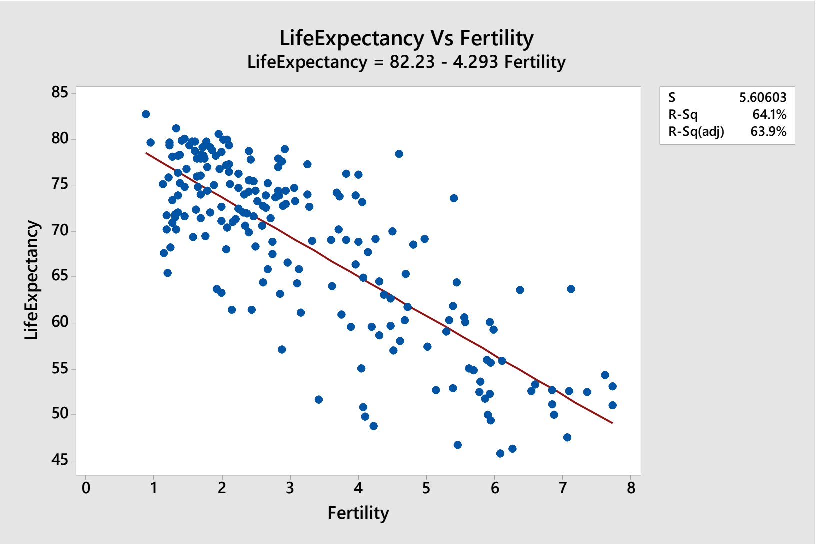


ANOVA test for response vs individual predictor:

Here we are performing the ANOVA analysis and fitting the regression line using individual predictors to get an idea about how each predictor is affecting the response.

1. Response vs Fertility:

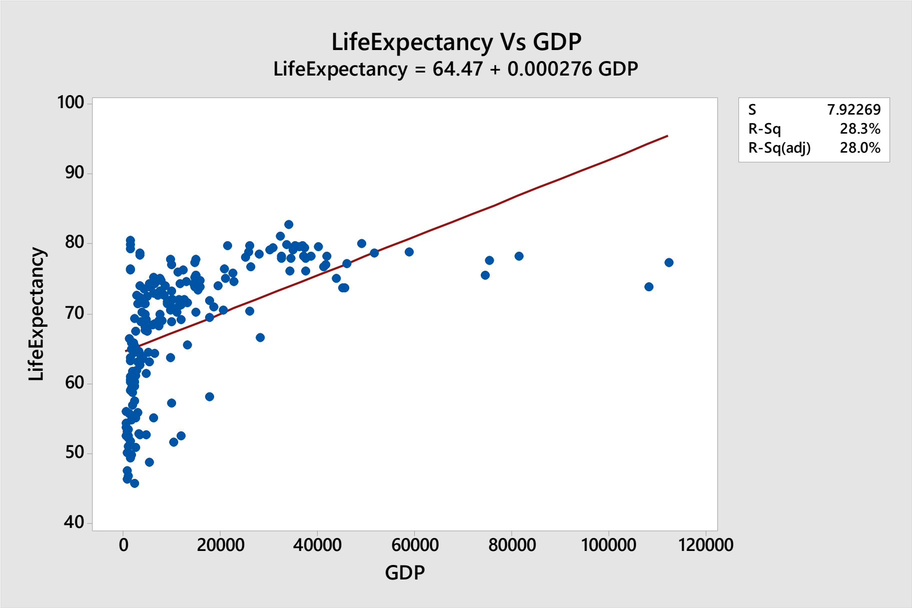
```
Analysis of Variance

Source         DF       SS       MS       F      P
Regression      1  11172.1  11172.1  355.49  0.000
Error         199   6254.1     31.4
Total         200  17426.2
```



Regression function of fertility around the world appears to give a good fit as it has a positive intercept (82.23) and it is statistically significant. Therefore, for one unit of change in fertility, life expectancy decreases by 4.293 units around the world.

2. Response vs GDP:

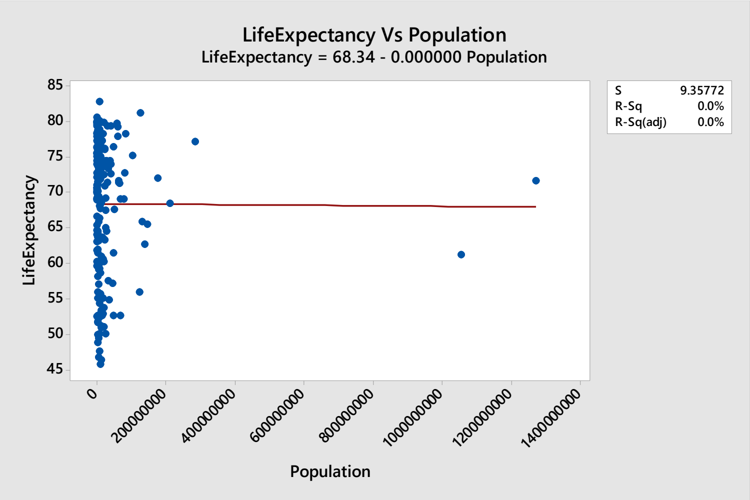
```
Analysis of Variance

Source         DF       SS       MS      F      P
Regression      1   4935.2  4935.15  78.62  0.000
Error         199  12491.0    62.77
Total         200  17426.2
```



Regression function of GDP around the world appears to have a poor fit but it has a positive intercept (64.47) and it is statistically significant. Here, for one unit of change in GDP, life expectancy is increased by 0.000276 units around the world which is negligible. Also the value of R^2 is 28.3% which indicates 28.3% of data variability is explained by the regression line.

3. Response vs Population:

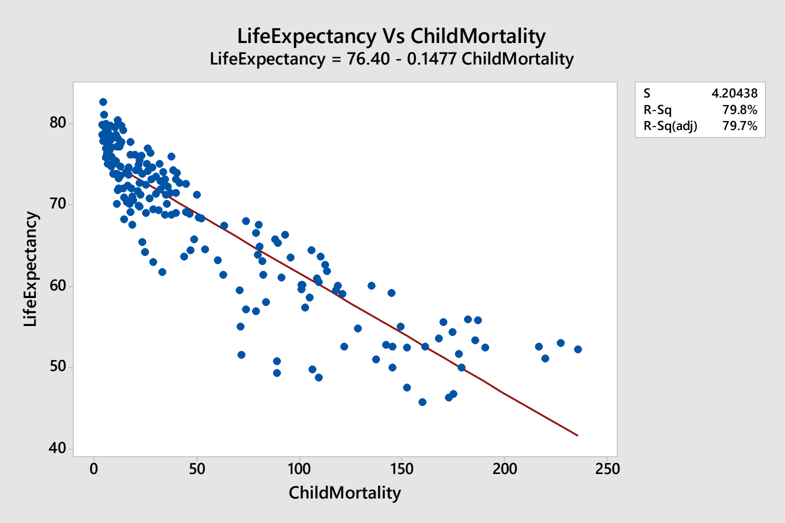
```
Analysis of Variance

Source        DF       SS       MS     F      P
Regression     1      0.4   0.3683  0.00  0.948
Error        199  17425.8  87.5669
Total        200  17426.2
```



Regression function of Population is not significant. Also the value of R^2 is 0.0% which indicate that population has no effect on the life expectancy of that country or region.

4. Response vs Child mortality:

```
Analysis of Variance

Source        DF       SS       MS       F      P
Regression     1  13908.5  13908.5  786.82  0.000
Error        199   3517.7     17.7
Total        200  17426.2
```



Regression function of Child Mortality around the world appears to give a good fit as it has a positive intercept (76.40) and also statistically significant. So, for one unit of change in Child Mortality, life expectancy is decreased by 0.1477 units around the world.

Multiple Linear Regression:

We have performed multiple regression analysis of response variable Life Expectancy and continuous predictors as Fertility, GDP, Population and Child Mortality. In addition, we have added Geographical Region as a Categorical predictor in our analysis. Therefore, our multiple regression analysis has been performed on 6 Geographical Regions of the world as given below:

1. America
2. East Asia & Pacific
3. Europe & Central Asia
4. Middle East & North Africa
5. South Asia
6. Sub-Saharan Africa

The output from the Minitab window and its analysis is given below:

```
Analysis of Variance

Source                  DF   Adj SS   Adj MS  F-Value  P-Value
Regression               9  14443.1  1604.78   102.75    0.000
  Fertility              1     24.5    24.55     1.57    0.212
  GDP                    1    240.7   240.66    15.41    0.000
  Population             1      0.1     0.08     0.01    0.943
  ChildMortality         1   1223.3  1223.29    78.32    0.000
  Geographical Region    5    251.4    50.27     3.22    0.008
Error                  191   2983.1    15.62
Total                  200  17426.2
```

From the above ANOVA results, we can see that Predictors Fertility and Population are not significant across the regions while deciding life expectancy of any country in a particular region.

Model Summary

```
   S      R-sq       R-sq(adj)  R-sq(pred)
3.95201  82.88%     82.07%      80.87%
```

Here, R^2 of multiple regression model is 82.88%, which indicates that 82.88% of our data points are closer to the fitted regression line, which is a goodness of fit for linear regression analysis. In addition, there is no noticeable difference between the R^2 and R^2 $(\mathrm{adjusted})$ which indicates that there is no unnecessary or redundant predictors present in the model for this data set.

Coefficients

```
Term                              Coef   SE Coef  T-Value  P-Value   VIF
Constant                         76.50      1.10    69.66    0.000
Fertility                       -0.477     0.381    -1.25    0.212  5.62
GDP                           0.000073  0.000019     3.93    0.000  1.45
Population                    0.000000  0.000000     0.07    0.943  1.14
ChildMortality                 -0.1040    0.0117    -8.85    0.000  5.64
Geographical Region
  East Asia & Pacific           -2.277     0.951    -2.39    0.018  1.56
  Europe & Central Asia         -1.414     0.900    -1.57    0.118  1.98
  Middle East & North Africa     -0.83      1.15    -0.73    0.469  1.58
  South Asia                     -2.15      1.63    -1.32    0.190  1.31
  Sub-Saharan Africa             -4.44      1.20    -3.71    0.000  3.40
```

Regression Equation

Geographical Region

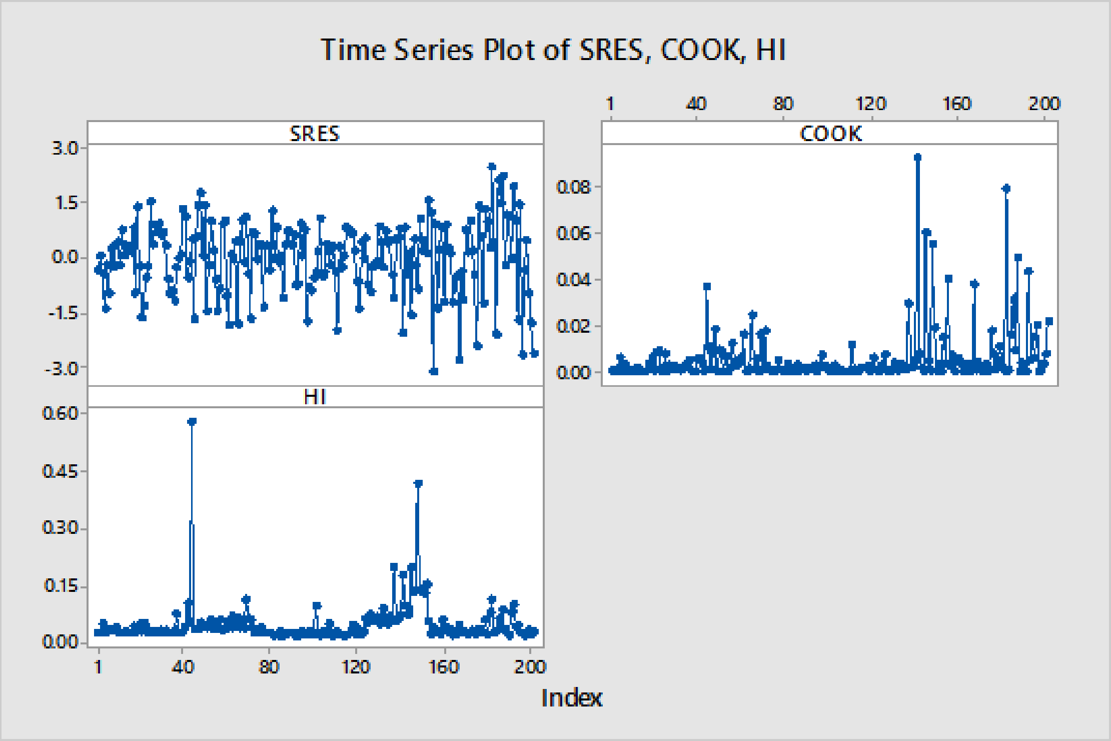
```
America
LifeExpectancy =
76.50 - 0.477 Fertility + 0.000073 GDP + 0.000000 Population -
 0.1040 ChildMortality

East Asia & Pacific
LifeExpectancy =
74.22 - 0.477 Fertility + 0.000073 GDP + 0.000000 Population -
 0.1040 ChildMortality

Europe & Central Asia
LifeExpectancy =
75.086 - 0.477 Fertility + 0.000073 GDP + 0.000000 Population -
 0.1040 ChildMortality

Middle East & North Africa
LifeExpectancy =
75.67 - 0.477 Fertility + 0.000073 GDP + 0.00 Population- 0.1040 ChildMortality

South Asia
LifeExpectancy =
74.35 - 0.477 Fertility + 0.000073 GDP + 0.000000 Population -
 0.1040 ChildMortality

Sub-Saharan Africa
LifeExpectancy =
72.06 - 0.477 Fertility + 0.000073 GDP + 0.000000 Population -
 0.1040 ChildMortality
```

We have six different regression equations for each geographical region here, which only differs in terms of the intercept terms. We can also see that coefficient of the population term is zero here which indicates that it has no effect in deciding the life expectancy in any region. This also sounds logical because even if a country is small and developed then it is not necessary that it will have lower or higher life expectancy compared to other smaller country.

Fits and Diagnostics for Unusual Observations

```
Obs  LifeExpectancy    Fit   Resid  Std Resid
 44           71.50  70.19    1.31       0.51     X
137           77.20  81.08   -3.88      -1.10     X
141           73.80  81.20   -7.40      -2.07  R  X
145           51.00  56.49   -5.49      -1.55     X
148           61.10  63.74   -2.64      -0.87     X
152           72.40  71.95    0.45       0.12     X
155           51.60  63.75  -12.15      -3.14  R
167           49.30  60.07  -10.77      -2.79  R
175           49.80  59.21   -9.41      -2.42  R
182           78.29  69.02    9.27       2.49  R
184           55.00  63.16   -8.16      -2.10  R
185           53.00  44.80    8.20       2.14  R
187           78.61  70.08    8.54       2.26  R
196           48.70  59.05  -10.35      -2.66  R
201           50.80  61.08  -10.28      -2.64  R

R  Large residual
X  Unusual X
```

We also found that there are some values of Y with large residuals and unusual predictor values. To investigate it more we prepared the time series plots of standardized residuals, cook's distance and predictor values with high influence. We see that some predictor values are having a high influence and cook's distance for the given model. We can also observe some values of Y with large residual values. Plot is shown below:



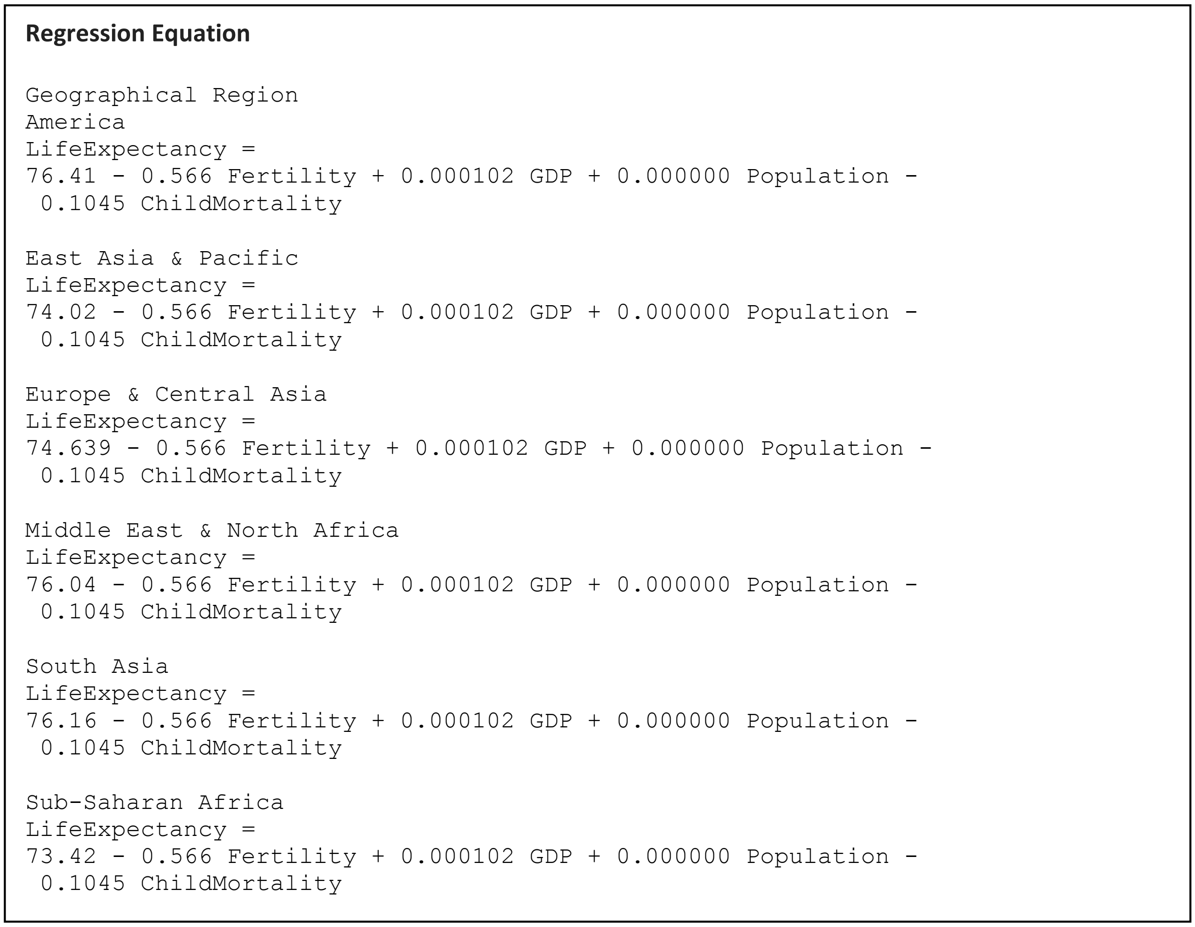
Elimination of the abnormal values of X and Y to check its effect on the model:

We have noticed some unusual values of X and large residuals after performing the multiple linear regression. We also plotted the time series plot for the same. Now to check how the model is affected by these abnormal values, we eliminated them and reanalyzed using the multiple regression analysis, and plotted time series graphs of residuals, High influence variables and cook's distance to get the following plot. We can see now the value of the highest influencing variable is around 0.3, much less than the previous value of 0.59. In addition, the highest cook's distance has improved to 0.06 from 0.08.

We can also notice that there are too many peaks in the current time series graph of high influence and cook's distance values, which were masked earlier by the unusual values.

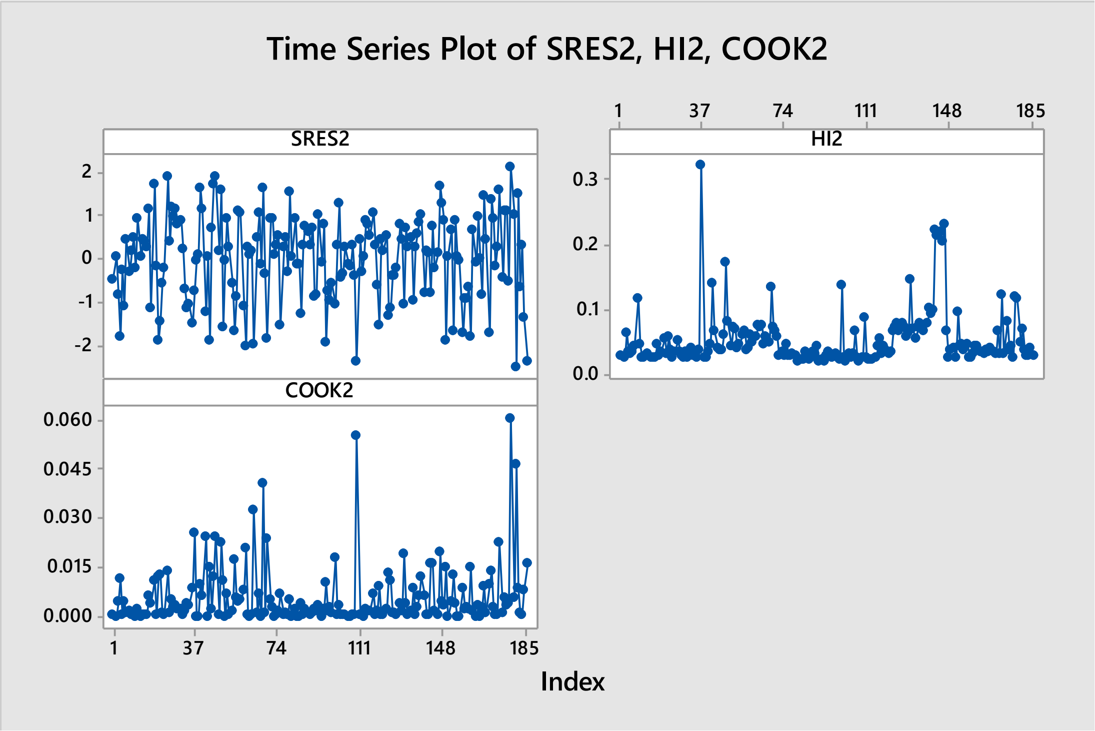
Regression Analysis: Life Expectancy versus Fertility, GDP, Population, Child Mortality

```
Method

Categorical predictor coding  (1, 0)


Analysis of Variance

Source                  DF   Adj SS   Adj MS  F-Value  P-Value
Regression               9  12655.3  1406.15   128.18    0.000
  Fertility              1     29.6    29.65     2.70    0.102
  GDP                    1    304.6   304.60    27.77    0.000
  Population             1      0.0     0.01     0.00    0.976
  ChildMortality         1    978.3   978.34    89.18    0.000
  Geographical Region    5    168.1    33.62     3.06    0.011
Error                  176   1930.7    10.97
Total                  185  14586.0


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
3.31208  86.76%     86.09%      85.21%


Coefficients

Term                             Coef   SE Coef  T-Value  P-Value   VIF
Constant                        76.41      1.01    75.54    0.000
Fertility                      -0.566     0.344    -1.64    0.102  5.92
GDP                          0.000102  0.000019     5.27    0.000  1.53
Population                   0.000000  0.000000     0.03    0.976  1.07
ChildMortality                -0.1045    0.0111    -9.44    0.000  6.49
Geographical Region
  East Asia & Pacific          -2.389     0.804    -2.97    0.003  1.52
  Europe & Central Asia        -1.775     0.767    -2.31    0.022  1.99
  Middle East & North Africa   -0.370     0.972    -0.38    0.704  1.47
  South Asia                    -0.26      1.64    -0.16    0.876  1.20
  Sub-Saharan Africa            -3.00      1.18    -2.55    0.012  3.97
```

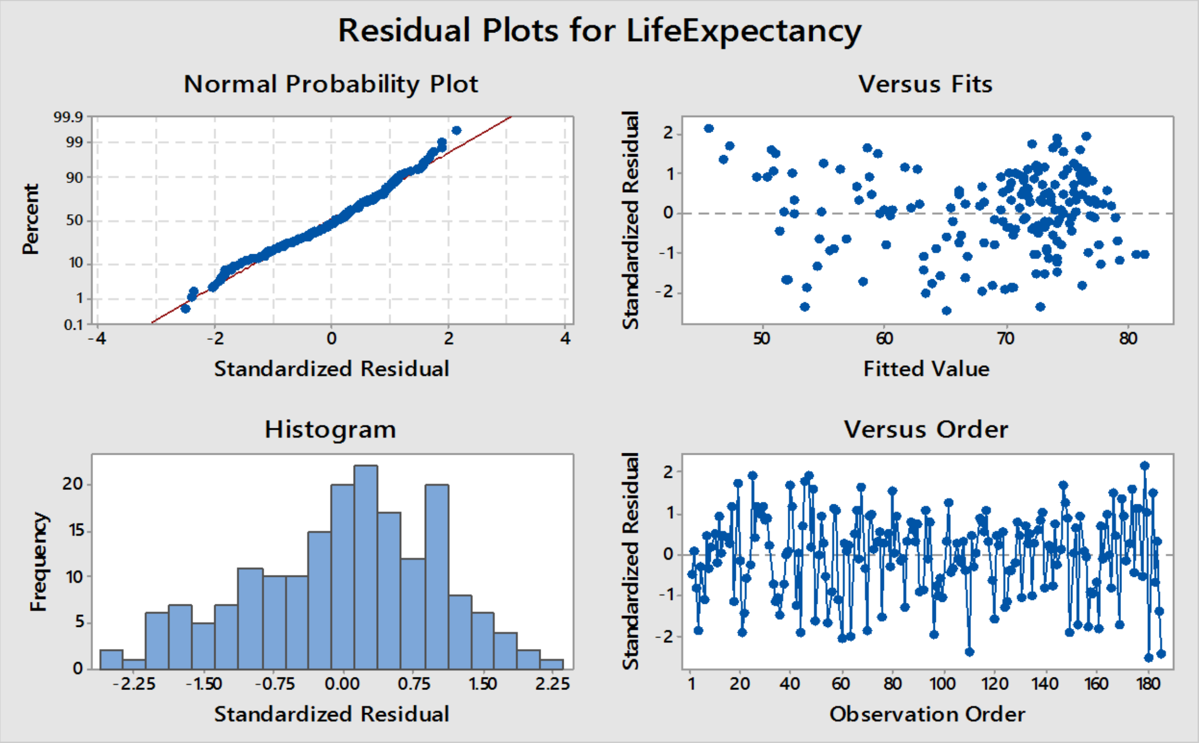


Regression Equation

```
Geographical Region
America
LifeExpectancy =
76.41 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality

East Asia & Pacific
LifeExpectancy =
74.02 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality

Europe & Central Asia
LifeExpectancy =
74.639 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality

Middle East & North Africa
LifeExpectancy =
76.04 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality

South Asia
LifeExpectancy =
76.16 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality

Sub-Saharan Africa
LifeExpectancy =
73.42 - 0.566 Fertility + 0.000102 GDP + 0.000000 Population -
 0.1045 ChildMortality
```



We also see some improvement in the residual plots after eliminating the above-mentioned unusual values. Variance of this data is close to constant but still it can be improved further. We will try to improve it by performing the transformations on the predictor values.



Dummy Variable:

We also analyzed the model using the dummy variables to compare the data amongst categorical variables. We set the 'Sub-Saharan Africa' as the base here to compare other regions with this region.

Result shows us that p values for all the other five regions are less than 0.05 (95% confidence interval) which proves that life expectancy differs significantly in all the five regions when compared to the base region that is Sub-Saharan Africa. Hypotheses that we tested here is

$\mathrm{H_0}$: There is no significant difference in the life expectancy of the region and Sub − Saharan region
$\mathrm{H_a}$: There is significant difference in the life expectancy of the region and Sub − Saharan region

Therefore, in all the five cases or regions we reject the null hypothesis based on the p values and accept the alternative hypothesis to prove there is a significant difference.

```
Method:

Categorical predictor coding  (1, 0)


Analysis of Variance

Source                  DF  Adj SS   Adj MS  F-Value  P-Value
Regression               5   10256  2051.26    55.79    0.000
  Geographical Region    5   10256  2051.26    55.79    0.000
Error                  195    7170    36.77
Total                  200   17426


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
6.06372  58.86%     57.80%      56.21%


Coefficients

Term                           Coef  SE Coef  T-Value  P-Value   VIF
Constant                     56.276    0.866    64.96    0.000
Geographical Region
  America                     17.02     1.29    13.17    0.000  1.45
  East Asia & Pacific         13.78     1.38    10.00    0.000  1.39
  Europe & Central Asia       17.62     1.21    14.52    0.000  1.52
  Middle East & North Africa  16.12     1.58    10.19    0.000  1.28
  South Asia                   8.01     2.31     3.46    0.001  1.12


Regression Equation

LifeExpectancy = 56.276 + 17.02 America + 13.78 East Asia & Pacific
+ 17.62 Europe &      Central Asia + 16.12 Middle East & North Africa
+ 8.01 South Asia
```

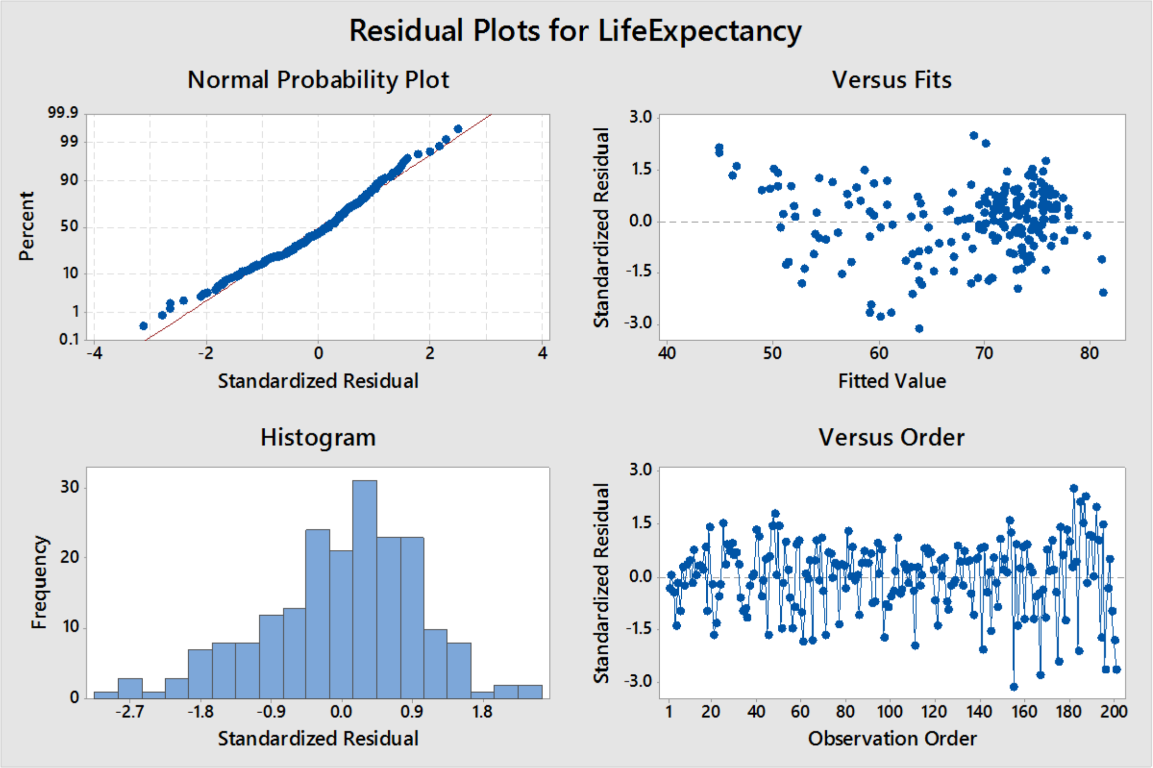
The constant intercept value 56.27 indicates that Life expectancy start at 56.27 years irrespective of country and continent. Here, Sub-Saharan Africa’s population serves as the baseline for rest of the continents. So, the coefficient of America indicates that life expectancy of Americans 17.02 years higher than Sub-Saharan Africans. Similarly, East Asian’s life expectancy 13.78 years higher than Sub-Saharan Africans but less than south Asians. Europe and Central Asia’s population have highest life expectancy with 73.89 years and South Asia’s populations have lowest life expectancy with 64.28 years.

Choosing the best model:

Earlier it was clear from the simple linear regressions that population is not contributing much towards deciding the life expectancy. It is further supported by the multiple linear regression and now we will decide the best model using the stepwise regression as a tool because of its advantages over other regression methods. We get the following results:

```
Regression Analysis: LifeExpectan versus Fertility, GDP, Population, ChildMortali,
...

Method

Categorical predictor coding (1, 0)


Stepwise Selection of Terms

Candidate terms: Fertility, GDP, Population, ChildMortality, Geographical Region

                     ------Step 1-----    ------Step 2-----    ------Step 3-----
                         Coef        P        Coef        P        Coef        P
Constant               76.587               74.610               75.183
ChildMortality       -0.14476    0.000    -0.13130    0.000    -0.11602    0.000
GDP                                       0.000092    0.000    0.000106    0.000
Geographical Region                                               -3.34    0.014
Fertility

S                             3.60921              3.40597              3.31942
R-sq                           83.57%               85.45%               86.55%
R-sq(adj)                      83.48%               85.29%               86.02%
R-sq(pred)                     83.07%               84.86%               85.37%
Mallows’ Cp                     36.50                13.52                 8.79

                     ------Step 4-----
                         Coef        P
Constant               76.421
ChildMortality        -0.1044    0.000
GDP                  0.000102    0.000
Geographical Region     -3.00    0.010
Fertility              -0.568    0.096

S                             3.30272
R-sq                           86.76%
R-sq(adj)                      86.17%
R-sq(pred)                     85.38%
Mallows’ Cp                      8.00

α to enter = 0.15, α to remove = 0.15
If a term has more than one coefficient, the largest in magnitude is shown.
```

```
Analysis of Variance

Source                 DF   Adj SS   Adj MS  F-Value  P-Value
Regression              8  12655.3  1581.91   145.02    0.000
  Fertility             1     30.6    30.60     2.81    0.096
  GDP                   1    305.8   305.78    28.03    0.000
  ChildMortality        1    998.9   998.93    91.58    0.000
  Geographical Region   5    169.5    33.90     3.11    0.010
Error                 177   1930.7    10.91
Total                 185  14586.0


Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
3.30272  86.76%     86.17%      85.38%

Coefficients

Term                              Coef   SE Coef  T-Value  P-Value   VIF
Constant                        76.421     0.983    77.78    0.000
Fertility                       -0.568     0.339    -1.67    0.096  5.77
GDP                           0.000102  0.000019     5.29    0.000  1.52
ChildMortality                 -0.1044    0.0109    -9.57    0.000  6.35
Geographical Region
  East Asia & Pacific           -2.389     0.801    -2.98    0.003  1.52
  Europe & Central Asia         -1.777     0.761    -2.34    0.021  1.96
  Middle East & North Africa    -0.370     0.969    -0.38    0.703  1.47
  South Asia                     -0.25      1.62    -0.15    0.877  1.17
  Sub-Saharan Africa             -3.00      1.17    -2.56    0.011  3.96

Regression Equation

Geographical Region
America
LifeExpectancy = 76.421 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality

East Asia & Pacific
LifeExpectancy = 74.03 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality
Europe & Central Asia
LifeExpectancy = 74.644 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality

Middle East & North Africa
LifeExpectancy = 76.05 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality

South Asia
LifeExpectancy = 76.17 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality

Sub-Saharan Africa
LifeExpectancy = 73.42 - 0.568 Fertility + 0.000102 GDP - 0.1044 ChildMortality
```



Validation of model assumption:

We get the residual plots as shown above from the Minitab after performing multiple linear regression, dummy variables analysis and selecting the best model. We can see from the plots that:

1. Residuals are following no particular pattern in the graph of residuals vs fitted value. This is indicative of constant variance which is one of the main assumption that need to be satisfied to get the good fit and results.
2. From Normal probability plot and histogram, we can see that Residuals are close to following a normal distribution. This is also confirmed using the Anderson-Darling test for normality.
3. Residuals are having a constant mean of zero here.
4. We also see no pattern in the residual Vs observation order (assumed the data order here) plot. This proves residuals are independent of each other.

4. Conclusions:

1. Excluded the population term from the predictor because it is not a significant contributor in deciding the life expectancy in that particular country or region.

2. Eliminating the unusual X, we get the improved R^2 and$\mathrm{R}^2(\mathrm{Adjusted})$. After selecting, the best model $\mathrm{R}^2(\mathrm{Adjusted})$ improves further.

	Original model (MR)	Eliminating the unusual X values (MR)	Selecting the best model (stepwise)
R^2	82.88%	86.76%	86.76%
$\mathrm{R}^2(\mathrm{Adjusted})$	82.07%	86.09%	86.17%

3. After removing the high influence and unusual parameters we are getting a good fit. This transformed model also satisfies all the basic assumptions of regression. This model can be used for predicting the values for the future. This will also give us an opportunity to test the prediction capability of the model.

Abbreviations:	
GDP	Gross Domestic Products
ANOVA	Analysis of Variance
MR	Multiple Regression

References:

Demographic Data - Multiple Geographical Areas on StatCrunch. (n.d.). Retrieved May 20, 2016, from https://www.statcrunch.com/app/index.php?dataid=1790796

COUNTRY COMPARISON :: LIFE EXPECTANCY AT BIRTH. (n.d.). Retrieved May 20, 2016, from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html

Gross Domestic Product (GDP) Definition | Investopedia. (2003). Retrieved May 20, 2016, from http://www.investopedia.com/terms/g/gdp.asp

DSS Data Subject Guides. (n.d.). Retrieved May 20, 2016, from http://dss.princeton.edu/cgi-bin/dataresources/guides.cgi