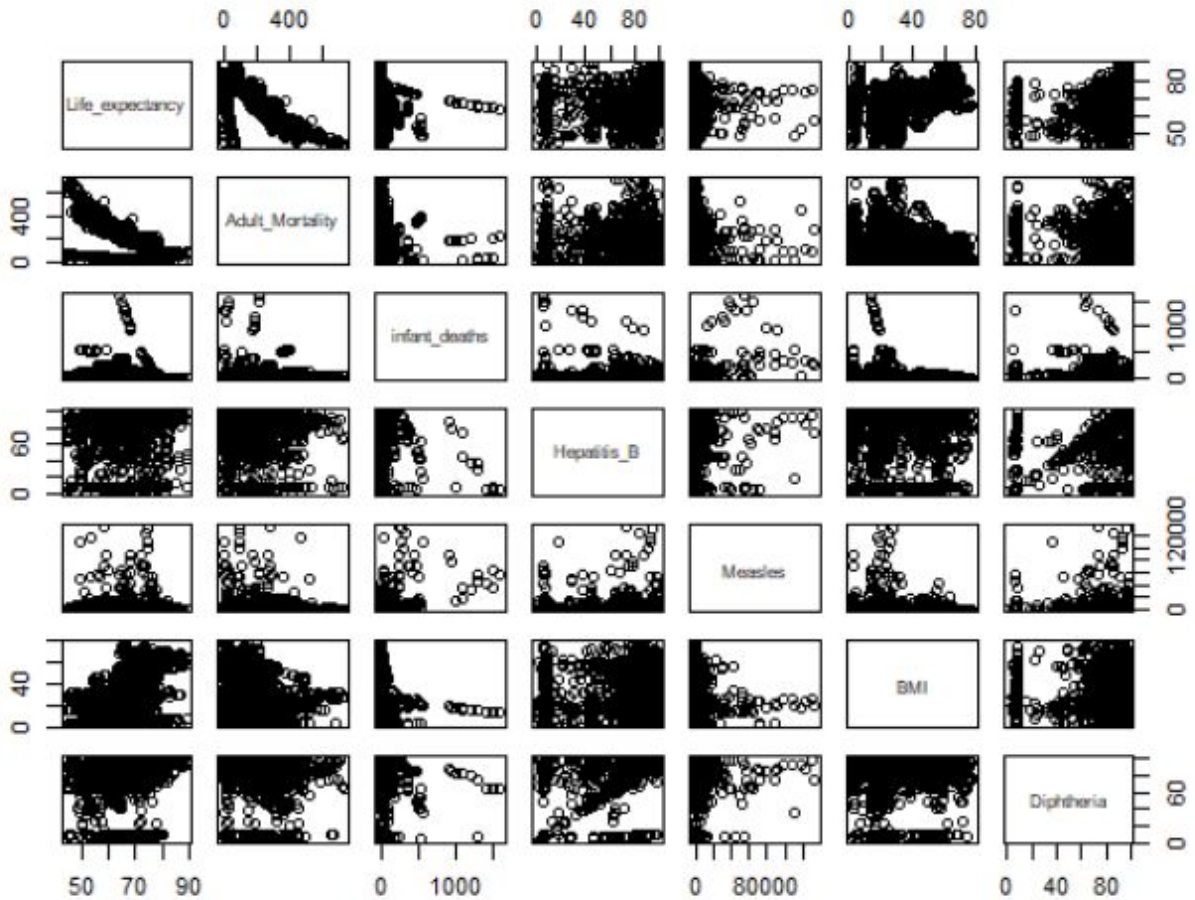
# MA 4710 Final

December 2020

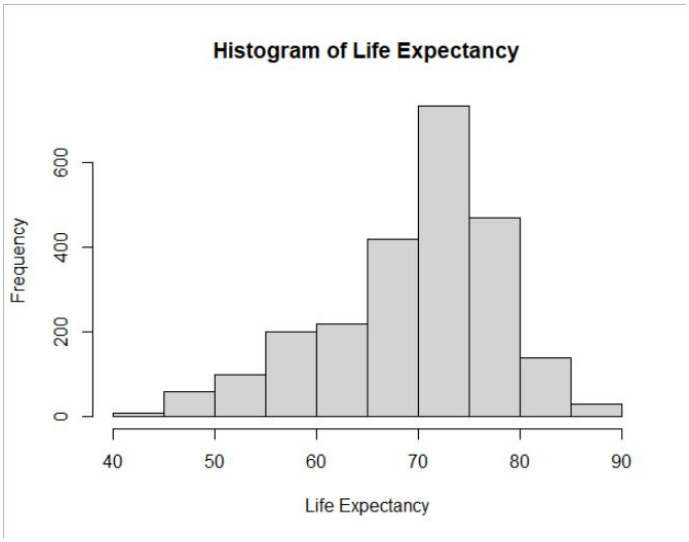
Marshal Gabala, Deanna Springgay, Sarah Wayward

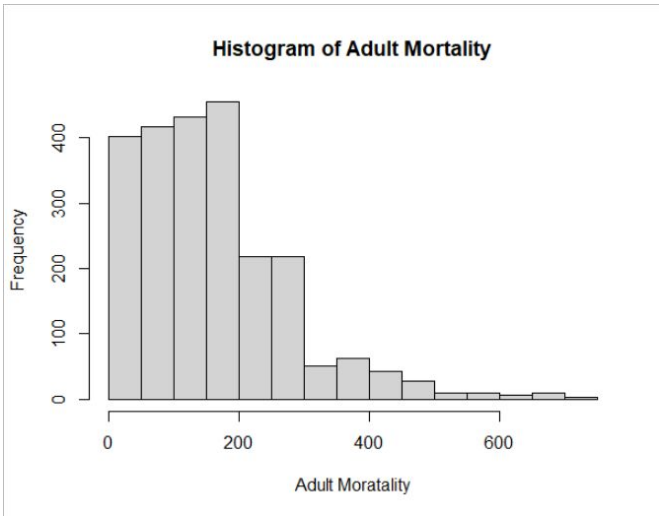
## Introduction

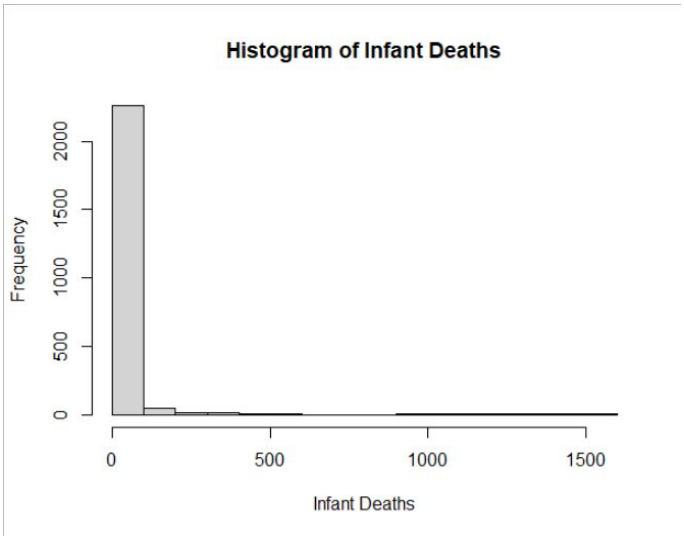
We chose Kaggle's Life Expectancy (WHO) dataset for our project. We chose this dataset because it had 20 predictor variables and almost 3000 rows of observations. The dataset from WHO includes the life expectancy of an average person by country and year, and other measured predictors that potentially affect life expectancy. In the end we chose the following predictors: Adult Mortality (AM) - the rate of adult death between the ages of 15 and 60 per 1000 people, Infant deaths (ID) - the rate of deaths of infants per 1000 people, Hepatitis B (HB) - the percentage of people greater than 1 year old that have been immunized against Hepatitis B, Measles (M) - the number of reported cases of measles per 1000 people, BMI (BMI) - the average Body Mass Index of the entire population, Diphtheria (D) - the percent of 1-year-olds who have been immunized against Diphtheria. We chose not to select any qualitative predictors since the qualitative predictors in the dataset contained more than two classes.

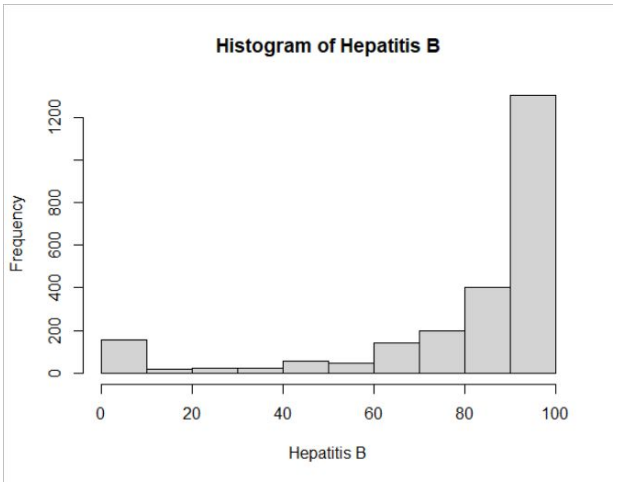


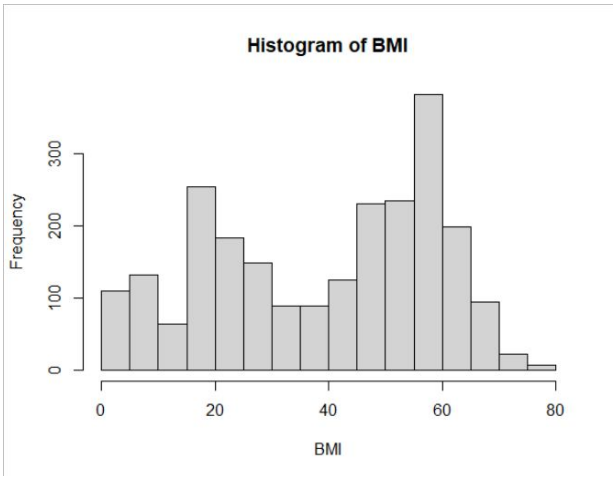
The above scatter plot shows there appears to be some slight correlation between a few of the variables. For example: life expectancy and adult mortality have a slight trend and pattern. However, when looking at Hepatitis B and BMI there is no pattern and the observations are all over the place. The same can be said about Hepatitis B and Adult Mortality and a few other plots.

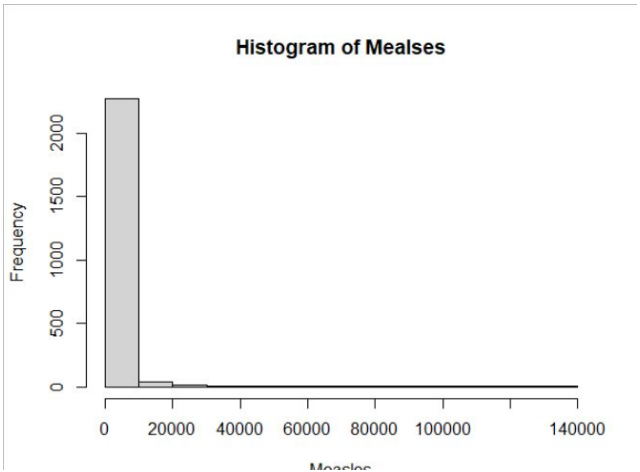


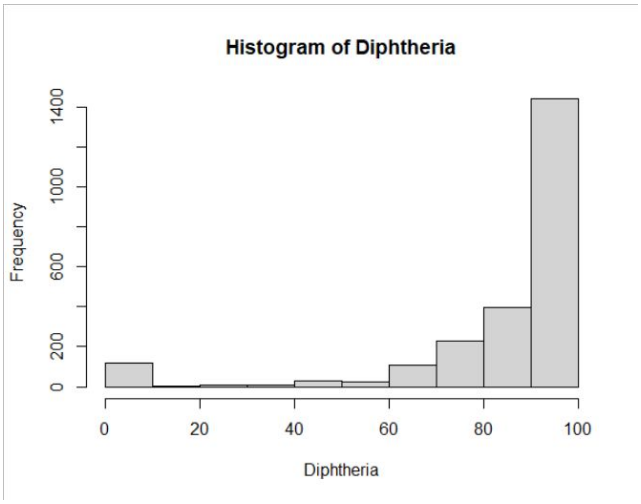






In this section you can see histograms of the outcome and predictors. An observation - the median life expectancy around the world is between 70 and 75 years old. For infant deaths there is a major right skew the same can be said about measles - an argument could be made that these two predictors potentially classify as near-zero-variance predictors. This means that around the world there are a lot of countries with very few infant deaths and only a couple that have a large number of infant deaths. With Measles it shows us that the majority of the world population do not contract measles while only a small percentage will end up with the disease at some point in their life. Both Hepatitis B and Diphtheria have a left skew which indicates that the majority of the world's population has been immunized against Hepatitis B and Diphtheria.

| | Life_expectancy | Adult_Mortality | infant_deaths | Hepatitis_B | Measles | BMI | Diphtheria |
|---|---|---|---|---|---|---|---|
| Life_expectancy | 1.00000000 | -0.71924932 | -0.19644276 | 0.2558713 | -0.09659744 | 0.5113353 | 0.36435595 |
| Adult_Mortality | -0.71924932 | 1.00000000 | 0.07949074 | -0.1631694 | 0.02989967 | -0.3559001 | -0.22839419 |
| infant_deaths | -0.19644276 | 0.07949074 | 1.00000000 | -0.2251359 | 0.53306240 | -0.2280307 | -0.16519070 |
| Hepatitis_B | 0.25587130 | -0.16316935 | -0.22513594 | 1.0000000 | -0.12242324 | 0.1551739 | 0.61170575 |
| Measles | -0.09659744 | 0.02989967 | 0.53306240 | -0.1224232 | 1.00000000 | -0.1571275 | -0.07249418 |
| BMI | 0.51133528 | -0.35590014 | -0.22803072 | 0.1551739 | -0.15712748 | 1.0000000 | 0.18057015 |
| Diphtheria | 0.36435595 | -0.22839419 | -0.16519070 | 0.6117058 | -0.07249418 | 0.1805701 | 1.00000000 |

We chose to center and scale the variables in order to make the predictors follow the same scale which could benefit the performance of a linear model though we will lose interpretability as a result.

**Models/Methods**

*Linear Model with Interaction Terms*

The first model we built was a linear model with interaction terms, which resulted in the following model:

```
Residuals:
     Min       1Q   Median       3Q      Max
-29.0090  -2.2741   0.3436   2.8077  21.1679

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.960e+01  1.489e-01 467.322  < 2e-16 ***
AM_center   -3.972e-02  1.116e-03 -35.583  < 2e-16 ***
ID_center   -2.653e-02  4.732e-03  -5.607 2.29e-08 ***
M_center     1.266e-05  2.736e-05   0.463 0.643468
BMI_center   1.008e-01  7.026e-03  14.342  < 2e-16 ***
D_center     7.845e-02  7.629e-03  10.284  < 2e-16 ***
HB_center    4.426e-03  5.442e-03   0.813 0.416121
AM_ID        2.471e-05  1.046e-05   2.363 0.018200 *
AM_HB        2.277e-05  4.337e-05   0.525 0.599634
AM_M        -6.080e-08  1.070e-07  -0.568 0.569943
AM_BMI       2.009e-04  5.895e-05   3.408 0.000666 ***
AM_D        -2.907e-04  5.290e-05  -5.496 4.31e-08 ***
ID_HB       -1.285e-04  5.245e-05  -2.450 0.014353 *
ID_M         1.705e-07  5.511e-08   3.094 0.002001 **
ID_BMI      -5.846e-04  2.190e-04  -2.669 0.007668 **
ID_D        -1.868e-05  6.539e-05  -0.286 0.775127
HB_M         1.864e-06  8.111e-07   2.298 0.021627 *
HB_BMI      -5.199e-04  2.667e-04  -1.950 0.051355 .
HB_D         4.994e-04  1.524e-04   3.277 0.001065 **
M_BMI        2.053e-07  1.405e-06   0.146 0.883831
M_D          2.442e-07  1.074e-06   0.227 0.820058
BMI_D       -1.809e-03  3.373e-04  -5.362 9.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.002 on 2341 degrees of freedom
Multiple R-squared:  0.6509,    Adjusted R-squared:  0.6478
F-statistic: 207.9 on 21 and 2341 DF,  p-value: < 2.2e-16
```

We then used stepwise regression with adjR2, Cp, AIC, and BIC in order to reduce the model using the most informative terms:

```
> stepwise(data=LE,y="Life_expectancy",select="adjRsq")
$process
   Step   EffectEntered EffectRemoved EffectNumber    Select
1     0       intercept                         1 0.0000000
2     1 Adult_Mortality                         2 0.5171151
3     2             BMI                         3 0.5916368
4     3      Diphtheria                         4 0.6227452
5     4           BMI_D                         5 0.6296720
6     5            AM_D                         6 0.6369629
7     6   infant_deaths                         7 0.6390063
8     7            HB_D                         8 0.6408037
9     8          ID_BMI                         9 0.6423742
10    9          AM_BMI                        10 0.6438600
11   10            HB_M                        11 0.6445632
12   11           ID_HB                        12 0.6452863
13   12         Measles                        13 0.6461398
14   13          HB_BMI                        14 0.6468042
15   14           AM_ID                        15 0.6468767

$variate
 [1] "intercept"       "Adult_Mortality" "BMI"             "Diphtheria"      "BMI_D"
 [6] "AM_D"            "infant_deaths"   "HB_D"            "ID_BMI"          "AM_BMI"
[11] "HB_M"            "ID_HB"           "Measles"         "HB_BMI"          "AM_ID"
```

```
> stepwise(data=LE,y="Life_expectancy",select="CP")
$process
   Step   EffectEntered EffectRemoved EffectNumber     Select
1     0       intercept                         1 4345.63423
2     1 Adult_Mortality                         2  878.16098
3     2             BMI                         3  379.42335
4     3      Diphtheria                         4  171.89664
5     4           BMI_D                         5  126.44866
6     5            AM_D                         6   78.60312
7     6   infant_deaths                         7   65.90279
8     7            HB_D                         8   54.85868
9     8          ID_BMI                         9   45.34217
10    9          AM_BMI                        10   36.39942
11   10            HB_M                        11   32.69264
12   11           ID_HB                        12   28.85614
13   12         Measles                        13   24.15377
14   13          HB_BMI                        14   20.71795

$variate
 [1] "intercept"       "Adult_Mortality" "BMI"             "Diphtheria"      "BMI_D"
 [6] "AM_D"            "infant_deaths"   "HB_D"            "ID_BMI"          "AM_BMI"
[11] "HB_M"            "ID_HB"           "Measles"         "HB_BMI"
```

```
> stepwise(data=LE,y="Life_expectancy",select="AIC")
$process
   Step   EffectEntered EffectRemoved EffectNumber    Select
1     0       intercept                         1 12440.091
2     1 Adult_Mortality                         2 10720.881
3     2             BMI                         3 10325.791
4     3     Diphtheria                          4 10139.554
5     4           BMI_D                         5 10096.762
6     5            AM_D                         6 10050.774
7     6   infant_deaths                         7 10038.433
8     7            HB_D                         8 10027.634
9     8          ID_BMI                         9 10018.277
10    9          AM_BMI                        10 10009.435
11   10            HB_M                        11 10005.760
12   11           ID_HB                        12 10001.943
13   12         Measles                        13  9997.245
14   13          HB_BMI                        14  9993.799

$variate
 [1] "intercept"       "Adult_Mortality" "BMI"             "Diphtheria"      "BMI_D"
 [6] "AM_D"            "infant_deaths"   "HB_D"            "ID_BMI"          "AM_BMI"
[11] "HB_M"            "ID_HB"           "Measles"         "HB_BMI"
```

```
> stepwise(data=LE,y="Life_expectancy",select="BIC")
$process
   Step   EffectEntered EffectRemoved EffectNumber    Select
1     0       intercept                         1 10074.957
2     1 Adult_Mortality                         2  8356.655
3     2             BMI                         3  7961.934
4     3     Diphtheria                          4  7776.027
5     4           BMI_D                         5  7733.288
6     5            AM_D                         6  7687.443
7     6   infant_deaths                         7  7675.131
8     7            HB_D                         8  7664.375
9     8          ID_BMI                         9  7655.071
10    9          AM_BMI                        10  7646.297
11   10            HB_M                        11  7642.661
12   11           ID_HB                        12  7638.894
13   12         Measles                        13  7634.265
14   13          HB_BMI                        14  7630.885

$variate
 [1] "intercept"       "Adult_Mortality" "BMI"             "Diphtheria"      "BMI_D"
 [6] "AM_D"            "infant_deaths"   "HB_D"            "ID_BMI"          "AM_BMI"
[11] "HB_M"            "ID_HB"           "Measles"         "HB_BMI"
```

As you can see from the above figures of the stepwise function the model has been narrowed down to these 13 variables and the constant:

```
 [1] "intercept"       "Adult_Mortality" "BMI"             "Diphtheria"      "BMI_D"
 [6] "AM_D"            "infant_deaths"   "HB_D"            "ID_BMI"          "AM_BMI"
[11] "HB_M"            "ID_HB"           "Measles"         "HB_BMI"
```

```
Call:
lm(formula = LE$Life_expectancy ~ AM + BMI + D + BMI_D + AM_D + 
    ID + HB_D + ID_BMI + AM_BMI + HB_M + ID_HB + M + HB_BMI + 
    AM_ID)

Residuals:
     Min       1Q   Median       3Q      Max 
-28.8766  -2.2586   0.3662   2.7688  21.5767 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.567e+01  7.478e-01  87.819  < 2e-16 ***
AM          -4.000e-02  1.110e-03 -36.036  < 2e-16 ***
BMI          1.020e-01  6.971e-03  14.636  < 2e-16 ***
D            8.229e-02  7.077e-03  11.628  < 2e-16 ***
BMI_D       -1.834e-03  3.264e-04  -5.618 2.16e-08 ***
AM_D        -2.749e-04  4.265e-05  -6.444 1.40e-10 ***
ID          -2.365e-02  4.405e-03  -5.369 8.71e-08 ***
HB_D         5.039e-04  1.489e-04   3.383 0.000728 ***
ID_BMI      -6.366e-04  1.999e-04  -3.184 0.001470 ** 
AM_BMI       2.048e-04  5.895e-05   3.475 0.000521 ***
HB_M         1.765e-06  4.980e-07   3.544 0.000402 ***
ID_HB       -1.517e-04  4.517e-05  -3.359 0.000796 ***
M            3.812e-05  1.389e-05   2.743 0.006128 ** 
HB_BMI      -6.022e-04  2.575e-04  -2.339 0.019421 *  
AM_ID        1.120e-05  9.203e-06   1.217 0.223577    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.009 on 2348 degrees of freedom
Multiple R-squared:  0.649,     Adjusted R-squared:  0.6469 
F-statistic: 310.1 on 14 and 2348 DF,  p-value: < 2.2e-16
```

Based on the information from the regression above we get a final model of:

$$Life\ Expectancy = 6.567e^{1} - 4.00e(AM) + 1.02e^{-1}(BMI) + 8.23e^{-2}(D) - 1.83e^{-3}(BMI\_D) - 2.75e^{-5}(AM\_D) - 2.37e^{-2}(ID) + 5.04e^{-4}(HB\_D) - 6.37e^{-4}(ID\_BMI) + 2.05e^{-4}(AM\_BMI) + 1.77e^{-6}(HB\_M) - 1.52e^{-4}(ID\_HB) + 3.81e^{-5}(M) - 6.02e^{-4}(HB\_BMI) + 1.12e^{-5}(AM\_ID)$$

Looking at our model diagnostic next we will be looking at the adequacy of the final model. First we will look at multicollinearity using variance importance factors (VIFs):

```
> vif(reduced.lmfit)
      AM       BMI         D     BMI_D      AM_D        ID      HB_D    ID_BMI    AM_BMI      HB_M     ID_HB 
1.623085  1.807581  2.157929  1.684548  1.382604 20.183419  2.112809 21.772550  1.511300  2.172317  6.001458 
       M    HB_BMI     AM_ID 
1.667982  1.594136  1.179309 
```

There are two variables that show signs of severe multicollinearity: ID and ID_BMI of 20.183 and 21.773 respectively. All the other variables are well below the cut off of 10.

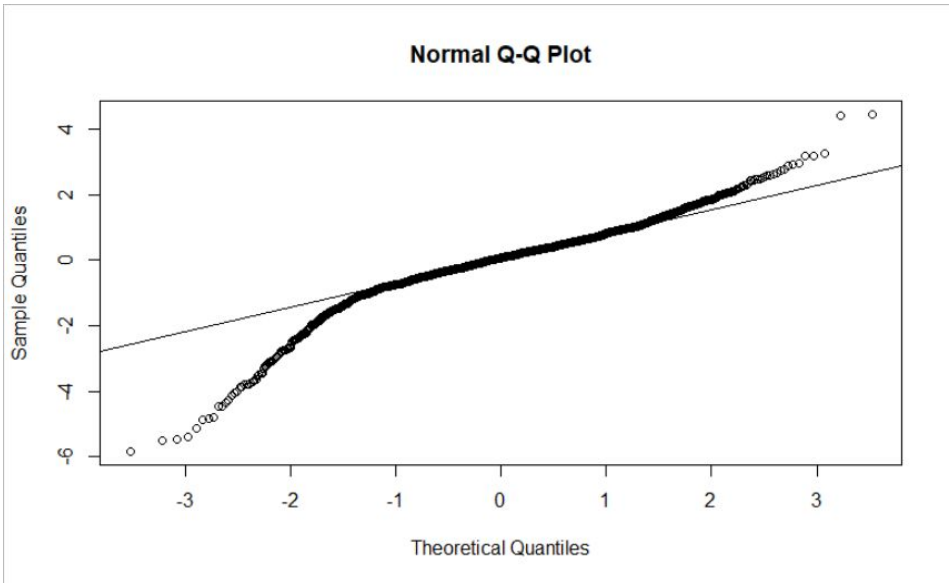
Following our check for multicollinearity we tested homoscedasticity using the Breusch-Pagan test and got the following results:

```
        studentized Breusch-Pagan test

data:  reduced.lmfit
BP = 371.59, df = 14, p-value < 2.2e-16
```

The Breusch-Pagan test gave us a p-value of less than 2.2e-16 which indicates that the error terms do not have a constant variance.

```
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.93027, p-value < 2.2e-16
```

Using the Shaprio-Wilk to test for normality we got a p value of less than 2.2e-16 which indicates that the residuals are not normally distributed.




We then checked for normality with a qq plot that indicates there is a left-skew. Factoring in all of these tests we conclude that the linear model with interaction terms is insufficient for the data.

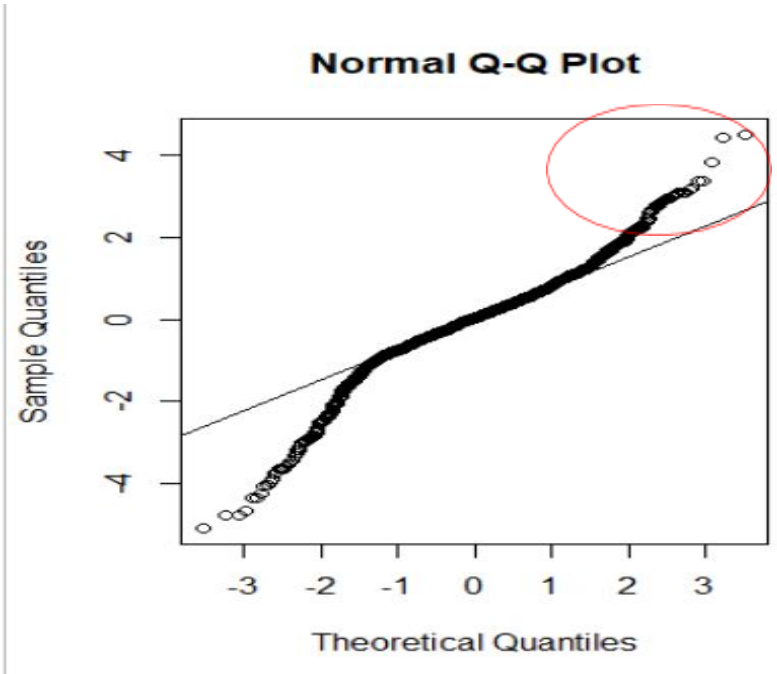
*Box-Cox Transformation*

We decided to try a BoxCox transformation in order to correct the issues found in assumption checking of the previous model. Running BoxCox gave us the following output:

```
> boxcox.summary
$lambda
[1] 2

$objective
[1] 0.9743273

$objective.name
[1] "PPCC"

$optimize
[1] TRUE

$optimize.bounds
lower upper
   -2     2

$eps
[1] 2.220446e-16
```

```
$lm.obj

Call:
lm(formula = LE$Life_expectancy ~ AM + BMI + D + BMI_D + AM_D +
    ID + HB_D + ID_BMI + AM_BMI + HB_M + ID_HB + M + HB_BMI +
    AM_ID, y = TRUE, qr = TRUE)

Coefficients:
(Intercept)           AM          BMI            D
  6.567e+01   -4.000e-02    1.020e-01    8.229e-02
      BMI_D         AM_D           ID         HB_D
 -1.834e-03   -2.749e-04   -2.365e-02    5.039e-04
     ID_BMI        AM_BMI         HB_M        ID_HB
 -6.366e-04    2.048e-04    1.765e-06   -1.517e-04
           M        HB_BMI        AM_ID
  3.812e-05   -6.022e-04    1.120e-05


$sample.size
[1] 2363

$data.name
[1] "reduced.lmfit"

attr(,"class")
[1] "boxcoxLm"
```

Transforming the data allowed us to attempt to correct the normality and homoscedasticity problems that the preliminary model indicated. Rechecking the QQ plot shows a slight change on the right side of the plot (the red circle), but this didn't correct the overall left-skew issue.



Both the Shaprio-Wilk test and the Breusch-Pagan test return the same results as the previous model (again indicating that the error terms are not normally distributed or have constant variance):

```
> shapiro.test(boxcox.res)

        Shapiro-Wilk normality test

data:  boxcox.res
W = 0.94953, p-value < 2.2e-16

> bptest(boxcox.lmfit)

        studentized Breusch-Pagan test

data:  boxcox.lmfit
BP = 385.54, df = 14, p-value < 2.2e-16
```

Lastly, we checked for multicollinearity on the transformed model, however it returned the same results as the first model:

```
> vif(boxcox.lmfit) ### OK ###
       AM       BMI         D     BMI_D      AM_D        ID      HB_D    ID_BMI    AM_BMI      HB_M     ID_HB
 1.623085  1.807581  2.157929  1.684548  1.382604 20.183419  2.112809 21.772550  1.511300  2.172317  6.001458
        M    HB_BMI     AM_ID
 1.667982  1.594136  1.179309
```

**Results**

Our final model is as follows:

$$Life\ Expectancy = 4.39e^{3} - 5.38(AM) + 1.35e^{1}(BMI) + 1.11e^{1}(D) - 2.32e^{-1}(BMI\_D)$$
$$- 4.09e^{-2}(AM\_D) - 3.27(ID) + 6.25e^{-2}(HB\_D) - 8.69e^{-2}(ID\_BMI) + 1.30e^{-2}(AM\_BMI)$$
$$+ 2.21e^{-4}(HB\_M) - 2.06e^{-2}(ID\_HB) + 5.20e^{-3}(M) - 9.00e^{-2}(HB\_BMI) + 1.6e^{-3}(AM\_ID)$$

To determine whether each variable is relevant to our final model, an F-test is run as seen below.

```
Call:
lm(formula = trans.Y ~ AM + BMI + D + BMI_D + AM_D + ID + HB_D +
    ID_BMI + AM_BMI + HB_M + ID_HB + M + HB_BMI + AM_ID, data = trans.LE)

Residuals:
    Min      1Q  Median      3Q     Max
-3460.2  -328.1    26.6   362.1  2987.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.388e+03  1.025e+02  42.823  < 2e-16
AM          -5.380e+00  1.521e-01 -35.375  < 2e-16
BMI          1.347e+01  9.552e-01  14.107  < 2e-16
D            1.105e+01  9.696e-01  11.398  < 2e-16
BMI_D       -2.324e-01  4.472e-02  -5.196 2.21e-07
AM_D        -4.087e-02  5.844e-03  -6.994 3.46e-12
ID          -3.271e+00  6.035e-01  -5.420 6.57e-08
HB_D         6.246e-02  2.041e-02   3.061 0.002231
ID_BMI      -8.689e-02  2.739e-02  -3.172 0.001534
AM_BMI       1.297e-02  8.077e-03   1.606 0.108474
HB_M         2.214e-04  6.824e-05   3.245 0.001191
ID_HB       -2.055e-02  6.189e-03  -3.321 0.000911
M            5.195e-03  1.904e-03   2.729 0.006407
HB_BMI      -8.985e-02  3.528e-02  -2.547 0.010933
AM_ID        1.599e-03  1.261e-03   1.268 0.204800
```

F test

$H_0$: predictor variables are necessary

$H_a$: predictor variables are not necessary

FStat = 290.1

AM and BMI have F values greater than the F stat therefore we can reject the null of those two predictor variables. The others we fail to reject.

The adjR2 value is 0.6314 which is acceptable and indicates that the addition of the variables are helpful.

```
Analysis of Variance Table

Response: trans.Y
            Df     Sum Sq    Mean Sq   F value    Pr(>F)
AM           1 1506046979 1506046979 3197.6852 < 2.2e-16 ***
BMI          1  234336652  234336652  497.5508 < 2.2e-16 ***
D            1   95325479   95325479  202.3980 < 2.2e-16 ***
BMI_D        1   16453350   16453350   34.9343 3.907e-09 ***
AM_D         1   29021731   29021731   61.6198 6.279e-15 ***
ID           1    7629663    7629663   16.1995 5.881e-05 ***
HB_D         1    4885933    4885933   10.3740 0.0012955 **
ID_BMI       1    6345291    6345291   13.4725 0.0002475 ***
AM_BMI       1     922299     922299    1.9583 0.1618323
HB_M         1    2055049    2055049    4.3633 0.0368276 *
ID_HB        1    2619098    2619098    5.5609 0.0184470 *
M            1    3089601    3089601    6.5599 0.0104920 *
HB_BMI       1    3028657    3028657    6.4305 0.0112818 *
AM_ID        1     757666     757666    1.6087 0.2048004
Residuals 2348 1105861908     470980
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 686.3 on 2348 degrees of freedom
Multiple R-squared:  0.6336,    Adjusted R-squared:  0.6314
F-statistic: 290.1 on 14 and 2348 DF,  p-value: < 2.2e-16
```

The estimated coefficients indicate if they will increase or decrease the average age a person will live to be and to what extent. Most of the coefficients are relatively small, indicating that life expectancy is explained by many variables that don’t have much influence on an individual level.

**Conclusion**

A linear model is not appropriate to model this data - a large number of the assumption tests failed even after attempting a transformation. A proper model would need to be a nonlinear regression model (such as neural networks, multivariate adaptive regression splines, k-nearest neighbors, or support vector machines). Such models are outside the scope of this class, and the final model we chose fit the data best out of the ones we attempted.

**R Code**

```
library(readr)
library(tidyverse)
library(ggpubr)
library(car)
library(olsrr)
library(lmtest)

Life_Expectancy_Data <-
read_csv("C:/Users/20539/OneDrive/Desktop/Regression/Regression-Analysis-WHO-Life-Expectancy-master/Regression-Analysis-WHO-Life-Expectancy-master/Life Expectancy Data.csv")

LE <- Life_Expectancy_Data %>%
select(c(Life_expectancy,Adult_Mortality,infant_deaths,Hepatitis_B, Measles,BMI, Diphtheria))%>%drop_na()

AM= LE$Adult_Mortality
ID= LE$infant_deaths
HB = LE$Hepatitis_B
M = LE$Measles
BMI = LE$BMI
D= LE$Diphtheria

###Center the predictor variables###
AM_center = (LE$Adult_Mortality - mean(LE$Adult_Mortality))
ID_center = (LE$infant_deaths - mean(LE$infant_deaths))
HB_center = (LE$Hepatitis_B - mean(LE$Hepatitis_B))
M_center = (LE$Measles - mean(LE$Measles))
BMI_center = (LE$BMI - mean(LE$BMI))
D_center = (LE$Diphtheria - mean(LE$Diphtheria))

AM_ID = AM_center*ID_center
AM_HB = AM_center*HB_center
AM_M  = AM_center*M_center
AM_BMI= AM_center*BMI_center
AM_D  = AM_center*D_center
ID_HB = ID_center*HB_center
ID_M  = ID_center*M_center
```

```
ID_BMI= ID_center*BMI_center
ID_D  = ID_center*D_center
HB_M  = HB_center*M_center
HB_BMI= HB_center*BMI_center
HB_D  = HB_center*D_center
M_BMI = M_center*BMI_center
M_D   = M_center*D_center
BMI_D = BMI_center*D_center

LE = cbind(LE,AM_ID,AM_HB,AM_M,AM_BMI,AM_D,
          ID_HB,ID_M,ID_BMI,ID_D,HB_M
          ,HB_BMI,HB_D,M_BMI,M_D,BMI_D)

view(LE)
full.lmfit = lm(LE$Life_expectancy~ AM + ID + HB + M + D +
                  AM_ID + AM_HB + AM_M + AM_BMI + AM_D +
                  ID_HB + ID_M + ID_BMI + ID_D + HB_M +
                  HB_BMI + HB_D + M_BMI + M_D + BMI_D,data = Life_Expectancy_Data)

library(leaps)
library(HH)
library(StepReg)

###adjRsQ###

stepwise(data=LE,y="Life_expectancy",select="adjRsq")

###CP###

stepwise(data=LE,y="Life_expectancy",select="CP")

#### AIC ####

stepwise(data=LE,y="Life_expectancy",select="AIC")

#### BIC ####

stepwise(data=LE,y="Life_expectancy",select="BIC")
```

```
reduced.lmfit = lm(LE$Life_expectancy~AM + BMI + D + BMI_D + AM_D + ID +
                   HB_D + ID_BMI + AM_BMI + HB_M + ID_HB + M + HB_BMI + AM_ID)
summary(reduced.lmfit )
###Diagnostics###
res <- rstudent(reduced.lmfit)
fitted.y <- fitted(reduced.lmfit)

###Multicollinearirty###
vif(reduced.lmfit)

### Constancy of Error Variances ###
bptest(reduced.lmfit)


######### Normality ############

qqnorm(res);qqline(res)
shapiro.test(res)

#### Transformation ####

library(EnvStats)

boxcox.summary <- boxcox(reduced.lmfit, optimize=TRUE)
lambda <- boxcox.summary$lambda

trans.Y <- LE$Life_expectancy^lambda

trans.LE <- cbind(LE,trans.Y)

### Re-fitting a model using the transformed response variable. ###
boxcox.lmfit <- lm(trans.Y ~AM + BMI + D + BMI_D + AM_D + ID +
                   HB_D + ID_BMI + AM_BMI + HB_M + ID_HB + M + HB_BMI + AM_ID ,
data=trans.LE)
summary(boxcox.lmfit)

boxcox.res <- rstudent(boxcox.lmfit)

boxcox.fitted.y <- fitted(boxcox.lmfit)
```

```
########## Multicollinearity ###########

vif(boxcox.lmfit) ### OK ###


########## Constancy of Error Variances ##########

bptest(boxcox.lmfit)

########## Normality ############

qqnorm(boxcox.res);qqline(boxcox.res)
shapiro.test(boxcox.res)

########## Final Model ###########

final.lmfit <- boxcox.lmfit
summary(final.lmfit)


anova(final.lmfit)
```