# CIA Factbook: A Look into Maternal Mortality Rate

## Multiple Linear Regression

Carlos Aguilar & Harrison Plate

## Introduction:

When we first looked at the data, the most interesting column was maternal mortality rate; we wanted to investigate something related to biology since we're both interested in the subject. We were curious to see the maternal mortality rates around the world, and we quickly decided to create a multiple linear regression model with maternal mortality as the response variable. A linear regression model with a maternal mortality rate as the response helps us determine which predictors have a significant effect on the amount of mothers passing during childbirth in order to reduce maternal mortality rate.

## Data Description:

The data source we used was collected in 2017; the data contains country-level statistics from the US Central Intelligence Agency (CIA), with a data frame of 259 observations on 11 different variables. The response variable in our study was maternal mortality rate, and potential predictor variables were the following: area, birth rate, death rate, infant mortality rate, life expectancy at birth, net migration rate, population, population growth. We wanted to analyze the potential predictors from this data set of countries around the world that have an effect on maternal mortality rate. From a first look at the data, the paired scatter plot (Figure 1. ) gave us a general image of how the other variables were related to each other, along with how they related to maternal mortality rate (located in the top left of paired scatter plot).

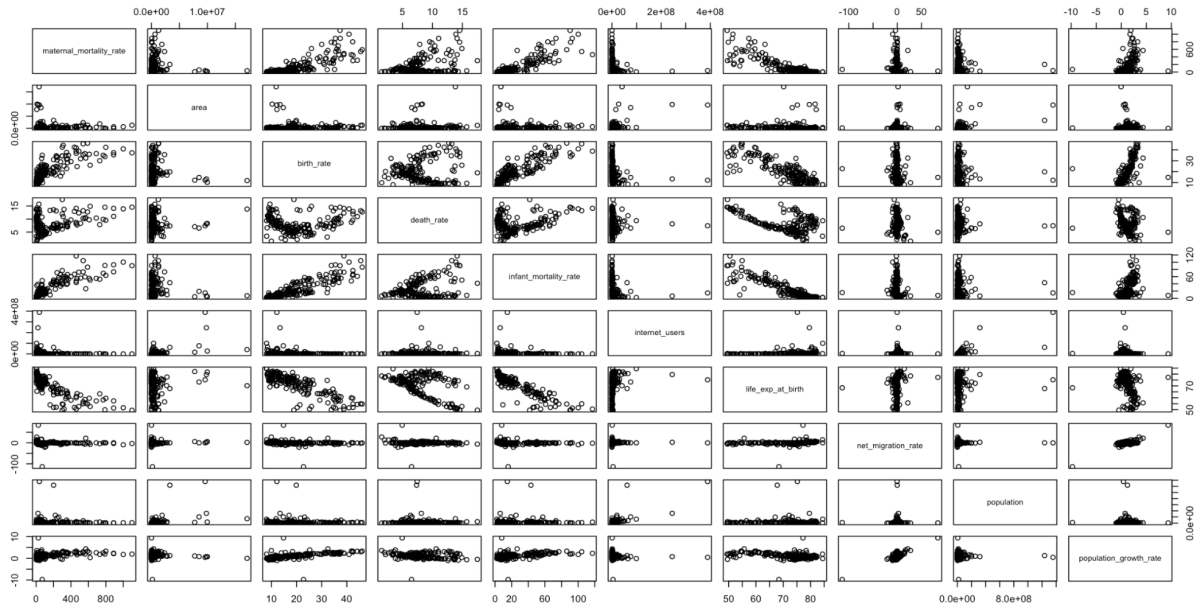


Figure 1. Paired Scatter Plot: Maternal Mortality Rate as response variable with nine other predictive variables

The correlation matrix (Figure 2. ) illustrated that some of the predictor variables were correlated with maternal mortality rate. Predictors like life expectancy at birth and birth rate, as well as life expectancy at birth and infant mortality rate had some correlation between themselves.

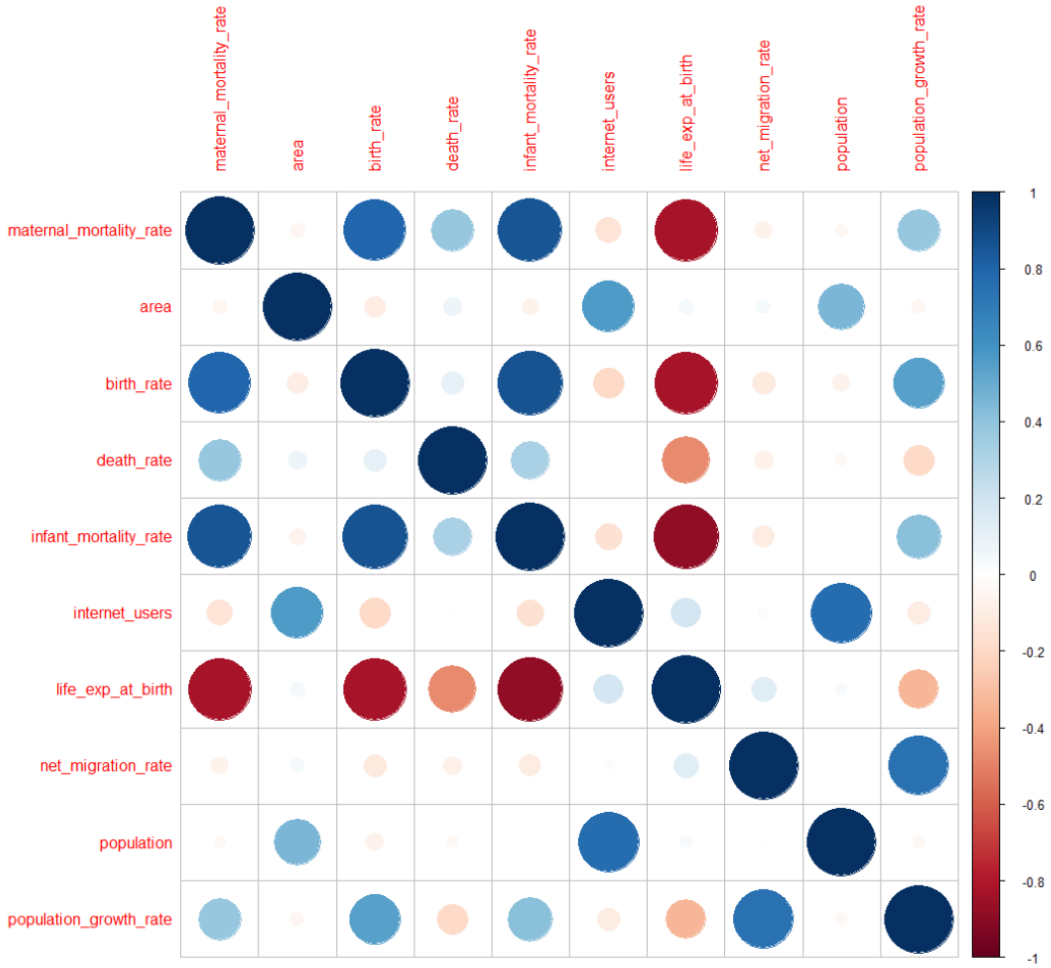


Figure 2. Correlation Matrix: Maternal Mortality Rate with Nine other Predictive Variables.

From observing our data through a paired scatter plot and correlation matrix, we decided to proceed with our data cleaning process in order to create the best model for our response variable, maternal mortality rate.

## Methods and Results:

The first thing we did was remove all the N/A values from the dataset. The countries that did have missing variables in the predictors were smaller countries which most people have never heard of and could not find on a map. We had no problem omitting them from our dataset and final model; one day with more data engineering practice we could create values, but that is something we would try in the future. After omission, we looked at a matrix scatterplot and we could see that there were several predictive variables that have a linear correlation with maternal mortality rate. We then decided to remove the column ‘country’ from the dataset, since we know

it won't be a significant predictor, and we move maternal mortality to the front so that we could read our correlation matrix easier.

From the correlation matrix we could see that there were many predictors that were correlated to maternal mortality. There were some predictors that were correlated within themselves, such as birth rate and infant mortality rate, and birth rate and life expectancy rate. Since we were happy with the correlation matrix, we decided to move forward and create our model followed by a Box-Cox transformation (Figure 3.).

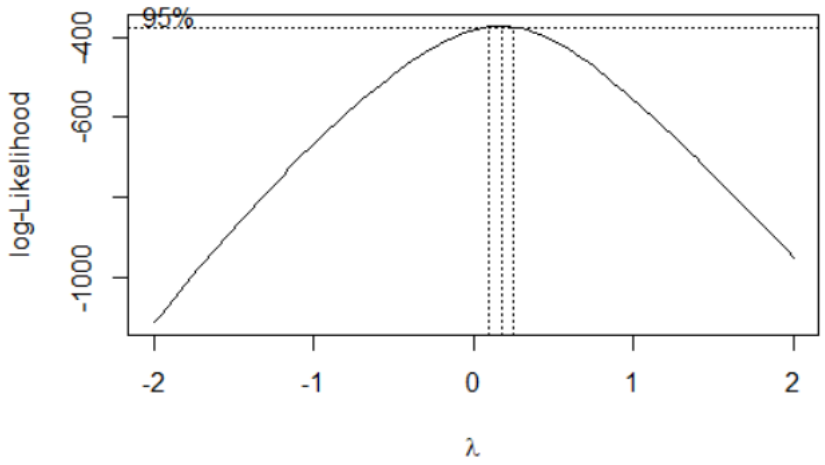


Figure 3. Box-Cox Transformation and Results from R

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.1738        0.17       0.0989       0.2487

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)

Likelihood ratio test that no transformation is needed
```

The Box-Cox returned a rounded power of 0.17, which indicates to us that we needed to do a 0.2 transformation, which we then applied to the response of our model. After the 0.2 transformation, we ran a summary of our model and saw that two of our nine predictors, infant mortality rate and life expectancy at birth, were significant. We also made sure that the four assumptions for multiple linear regression were satisfied by creating residuals vs fitted, QQ, and histogram of the residuals plots.

```
Call:
lm(formula = (maternal_mortality_rate)^(1/5) ~ area + internet_users +
    death_rate + infant_mortality_rate + life_exp_at_birth +
    birth_rate + net_migration_rate + population + population_growth_rate,
    data = df_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19200 -0.17924  0.02994  0.17049  0.65045

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              5.424e+00  6.300e-01   8.609 5.26e-15 ***
area                     8.479e-09  1.344e-08   0.631  0.52908
internet_users          -6.503e-10  1.046e-09  -0.622  0.53497
death_rate              -1.465e-01  7.584e-01  -0.193  0.84703
infant_mortality_rate    8.605e-03  2.293e-03   3.754  0.00024 ***
life_exp_at_birth       -4.534e-02  6.976e-03  -6.499 8.93e-10 ***
birth_rate               1.175e-01  7.576e-01   0.155  0.87697
net_migration_rate       1.031e-01  7.579e-01   0.136  0.89195
population               2.473e-10  2.471e-10   1.001  0.31836
population_growth_rate -1.048e+00  7.578e+00  -0.138  0.89019
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2902 on 167 degrees of freedom
Multiple R-squared:  0.8496,    Adjusted R-squared:  0.8415
F-statistic: 104.8 on 9 and 167 DF,  p-value: < 2.2e-16
```

Table 1. Summary of Full Model with transformation to the 0.2 Power

Since there were a lot of predictors with high p-values, indicating that they aren't significant, we decided to apply a step function to purge the model of insignificant predictors. The backwards stepwise selection removed five of the nine predictors, leaving birth rate, infant mortality rate, death rate, and life expectancy at birth. The four leftover predictors were deemed significant since their p-values were lower than 0.05. The stepped model produced higher adjusted R squared and lower AIC values compared to the original model, which suggested that the stepped model is better fitted. The Analysis of Variance (AOV) table produced a p-value of greater than 0.05 reinforced the fact that the stepwise model is the better model, and the five predictors can be removed. Next, we wanted to confirm that the four assumptions (linearity, independence of errors, normality of errors, and equal variances) were all satisfied. The performance check of the model confirmed that there were no violations of the assumptions. However, we noticed that there were some collinearity issues, which we wanted to address. In hindsight, the collinearity issues weren't very extreme and the model that we had was acceptable, but we decided to address them using a ridge regression since three out of four predictors had variance inflation factors exceeding five.

Instead of combining the predictors together, we decided to use a ridge regression to determine and drop the least significant of the four remaining predictors. We used the cv.glmnet function, which did a k-fold cross validation, to pinpoint a lambda value that generated the lowest Mean Squared Error (MSE), which builds the best model. We then looked at the coefficients of our ridge regression model, and we saw that life expectancy at birth had the furthest value away from 0, meaning it is the least significant, so we dropped it.

Having had only three predictors left, a summary of the model revealed that death rate was no longer a significant predictor based on its high p-value, so we decided to do another backwards stepwise selection. The higher adjusted R squared, lower AIC (107.8835), and high

AOV p-value confirmed that removing death rate as a predictor would create a superior model. Therefore, the only two predictors in our final model were infant mortality rate and birth rate, which we confirm were significant predictors based on the summary of the model. Before we could interpret the final model, we again used the performance check to confirm that the assumptions were satisfied; it also revealed that the collinearity issues had been fixed.

```
Call:
lm(formula = (maternal_mortality_rate)^(1/5) ~ infant_mortality_rate +
    birth_rate, data = df_final)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00088 -0.20711  0.02856  0.21028  0.81003

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.339008   0.066876  20.022  < 2e-16 ***
infant_mortality_rate 0.015183   0.001976   7.682 1.09e-12 ***
birth_rate           0.030970   0.004976   6.224 3.51e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3236 on 174 degrees of freedom
Multiple R-squared:  0.8051,    Adjusted R-squared:  0.8029
F-statistic: 359.4 on 2 and 174 DF,  p-value: < 2.2e-16
```

Table 2. Summary of Final Model:

The summary of our model resulted in p-values of both the model and predictors, infant mortality rate and birth rate, below 0.05, which indicates that they are both significant predictors in the model. The multiple R square indicates that 80.51% of the variability in maternal mortality rate can be predicted by the model. A one unit increase in infant_mortality_rate (1 more death per 1,000 live births), with the other predictor (birth_rate) held fixed, is associated with an increase in maternal_mortality_rate by (0.015183)^5 units, which equals 8.06841002E-10 units, which can be interpreted as 8.06841002E-10 more deaths (where the death is related to pregnancy or birth) per 100,000 live births. A one unit increase in birth_rate (1 birth per 1000 people), with the other predictor (infant_mortality_rate) held fixed, is associated with an increase in maternal_mortality_rate by (0.030970)^5 units, which equals 2.84908907E-8 units, which can be interpreted as 2.84908907E-8 more deaths (where the death is related to pregnancy or birth) per 100,000 live births. The outliers were Indonesia, Sudan, Iran, Mali, Afghanistan, Zimbabwe, Burkina Faso, Guyana, Lesotho, Sao Tome and Principe, & Kiribati.

## Conclusion:

Using data collected from the CIA factbook, we created a multiple linear regression model with maternal mortality rate as a response. After a 0.2 transformation and the removal of seven of nine predictors through two backwards stepwise selection and ridge regression, we determined that the two significant predictors were infant mortality rate and birth rate. We interpreted that as infant mortality rate and birth rate increases, when the opposite predictor is fixed, more mothers pass away during childbirth. We would want to further investigate our data by comparing our model with a model that takes into account interaction effects in order to see if any of the final predictors are correlated.

## Code Appendix:

https://github.com/HPL8/STAT-632-PROJECT