Estimation Of Maternal Mortality Rate And The Analysis Of Its Indicators Using Regression Techniques

Project Report

Megha Das, Sudha Sahithi Murikipudi

December 8, 2023

Introduction

Background

WHO defines Maternal Mortality rate at any given time period as the number of deaths occurred to a woman, aggravated by pregnancy or its management, per 100,000 live births. The ratio is usually measured after gathering information on pregnancy status, timing of death, and cause of death as follows:

$MMR = \frac{Number\ of\ maternal\ deaths}{Number\ of\ live\ births} * 100000$

Based on projections from various United Nations agencies, the global maternal mortality ratio (MMR) decreased by 34% between 2000 and 2020, dropping from 339 to 223 deaths per 100,000 live births. This translates to an annual decline rate of 2.1 percent. Although this reduction is notable, it accounts for only approximately one-third of the 6.4 percent annual reduction necessary to achieve the Sustainable Development Goal (SDG) target of 70 maternal deaths per 100,000 live births by the year 2030. While there was significant success in decreasing the global maternal mortality ratio (MMR) from 2000 to 2015, the statistics have shown a plateau in progress with relatively stable rates of reduction from 2016 to 2022, in most regions (Cresswell and World Health Organization 2023).

Motivation

A major barrier to gauging progress toward maternal mortality goals in low- and middle-income countries, lies in the absence of reliable data. Due to the lack of functioning civil registration systems, a large number of developing countries rely on national surveys like the Multiple Indicator Cluster Survey (MICS) and the Demographic and Health Survey (DHS) to collect data on maternal mortality. Maternal Mortality Ratio (MMR) estimation based on models is becoming more and more popular as a way to fill in information shortages. Additionally, identifying the significant factors affecting the MMR, will aid in providing a better understanding and interpretation of what factors drive MMR. (Gebremedhin 2018).

Research Question

This project aims to estimate the Maternal Mortality ratio, by constructing a regression model that explores the underlying relationship between the MMR and its potential indicators/predictors. The aim is to identify statistically significant predictors, estimating their regression coefficients, and subsequently, conduct hypothesis tests to evaluate the significance of the model and its predictors. This will ultimately contribute to a comprehensive understanding of the factors that influence MMR, help reduce its risk, and achieve the SDG Target.

Data Description

The data being used for this project, with the response variable as MMR, and all the other variables was collected by various sources. Since 1990, 248 nationwide surveys on health and demographics, as well as other relevant studies, have been carried out in 80 low- and middle-income nations. The majority of it were conducted as standard Demographic and Health Surveys (DHS), while the rest encompassed continuous DHS, National Family Health Surveys, or specialized surveys focused on maternal mortality.

The author obtained and reviewed 248 survey reports from the Measure DHS website. They extracted key information, including survey details (country, year, and type), reported Maternal Mortality Ratios (MMR, if available), and data for the nine predictor variables, and organized it in an Excel spreadsheet (Gebremedhin 2018).

The data contains 248 observations (from the 248 surveys) and 9 potential predictors. The nine variables were considered as potential predictors of Maternal Mortality Ratio (MMR). These variables include the modern contraceptive prevalence rate (CPR) among married women of reproductive age, the proportion of mothers who had 4 or more antenatal (ANC) visits, utilized health institution delivery, and received postnatal care (PNC) within the first two days after delivery for recent births within the last 5 years.

Additionally, the proportion of women who underwent Cesarean Section (C-section) for recent births within the last 5 years, the prevalence of anemia during pregnancy, prevalence of HIV, as well as factors such as thinness (Body Mass Index (BMI) $<$ 18.5 kg/m2) and short stature (height less than 145 cm) among women of reproductive age were also considered. The data dictionary is provided below:

Data Dictionary

- `MMRo` : Maternal mortality rate. This is the response variable.
- `CPR_modern` : Contraceptive Prevalence coverage (%)
- `ANC_4` : Antenatal care coverage
- `HID` : Health Institution delivery rate
- `CS` : Cesarean section rate
- `PNC` : Postnatal care coverage (%)
- `Animia_preg` : Percentage of people having maternal anemia
- `Thin` : Prevalence of thinness (%) (BMI $<$ 18.5 kg/m2)
- `HIV_final` : HIV prevalence among women of reproductive age
- `Ht_145` : Prevalence of maternal stunting (%)

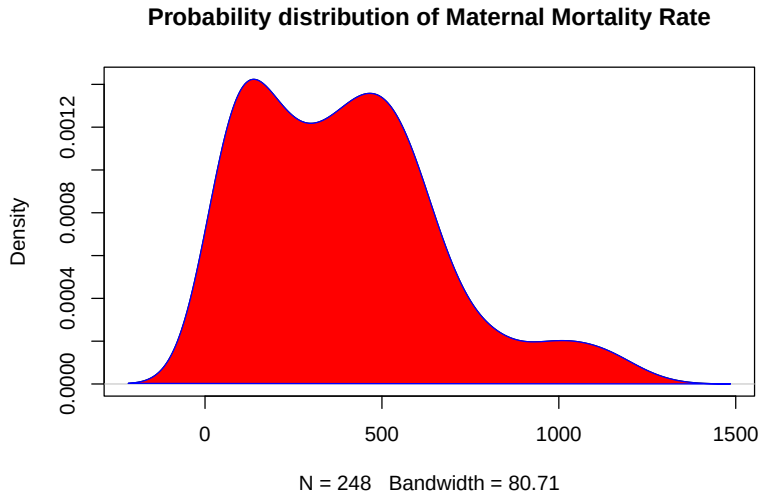
Exploratory Data Analysis

Data cleaning

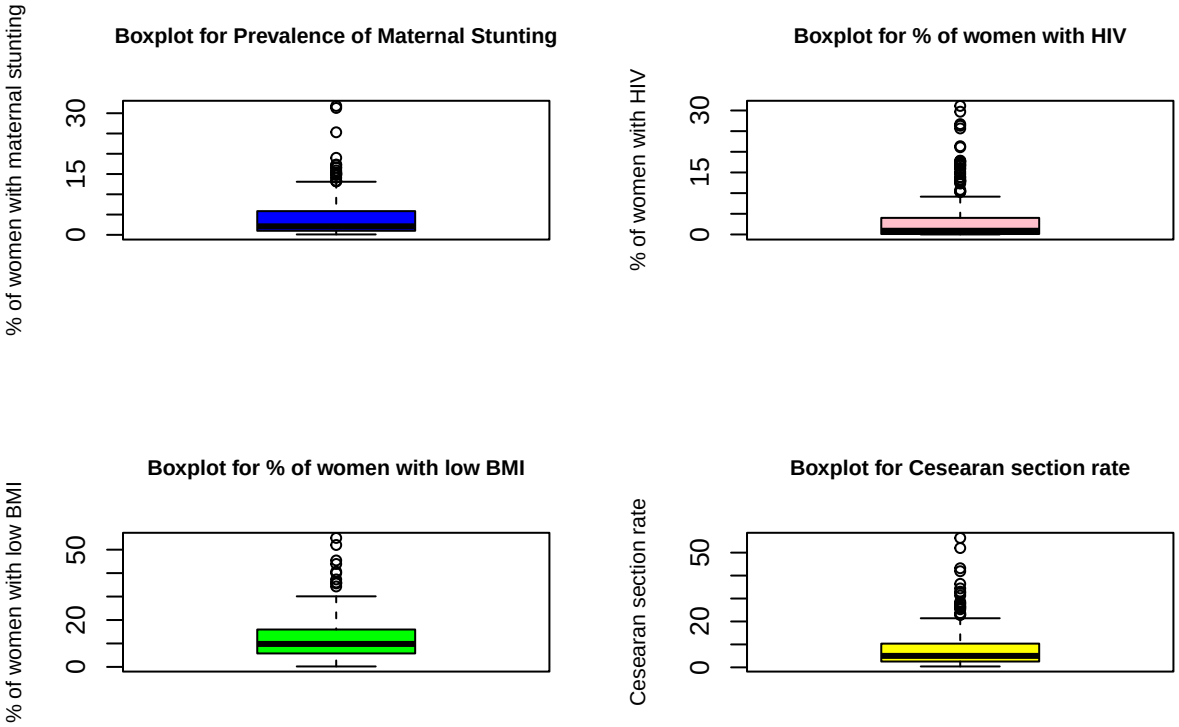
Initial data cleaning and wrangling was performed to bring the data to the required format. It was observed that `Ht_145` contains a few missing values, which was treated by imputing it with the median. There are no duplicates in the data.

Univariate Analysis

Univariate analysis was carried out to understand the distribution, spread, and presence of outliers in each variable. For the response, `MMRo` , the distribution is bi-modal, with a slightly left skew. This is because of the presence of few outliers in the data.



As part of missing value treatment, we observed that `Ht_145`, which is the percentage of women with short stature, contained outliers and is a left skewed distribution.



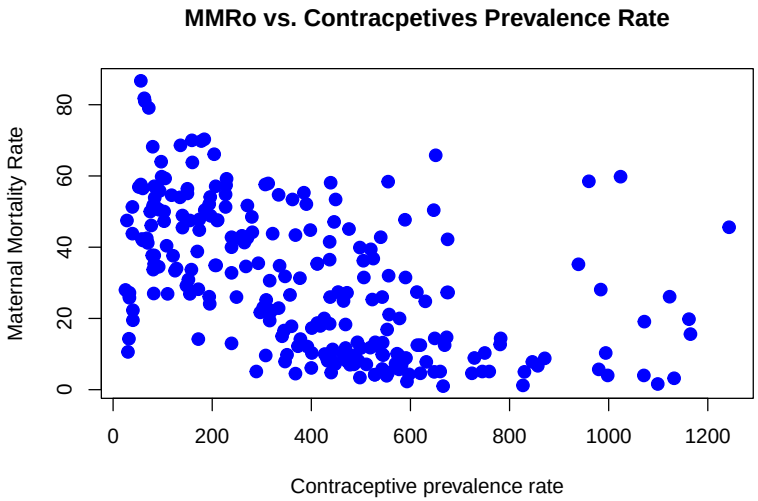
Similarly, it was observed that `HIV_final` (HIV Prevalence), Thinness, and Cesaeran section rate, is positively skewed as well, with multiple outliers. Outliers aren't treated at this stage as we suspect it might not negatively influence the data. Moreover, not all outliers are bad for a model, it could be that these are influential observations.

Bivariate Analysis

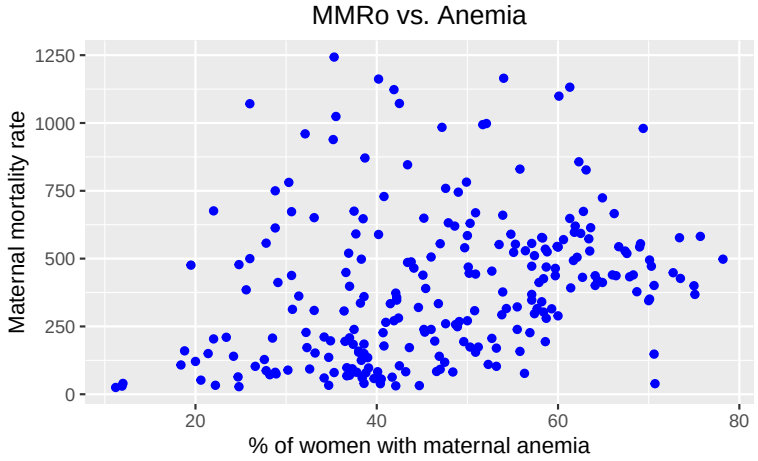
Next, the relationship between the response at the potential predictors was analysed to observe any patterns. The aim here is to identify variables, which could be used in the regression model, as well as indicate patterns which can be interpreted. Few variables that stood out are:

- `CPR_modern`: We suspect that as the usage of contraceptives increases, Maternal Mortality ratio will decrease. The scatter plot also confirms the hypothesis, as the linear relationship is relatively strong

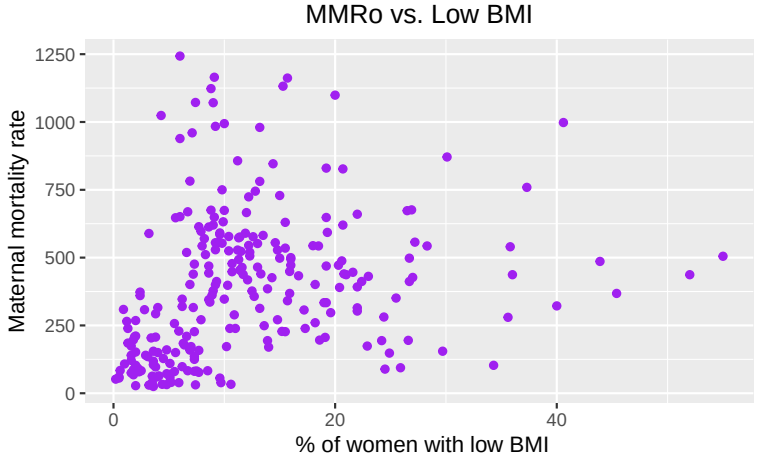
and negative. This could be a potential driver for MMR



- `Animia_preg`: The scatter plot of % of women with maternal anemia versus the Maternal Mortality rate also indicated a positive linear relationship

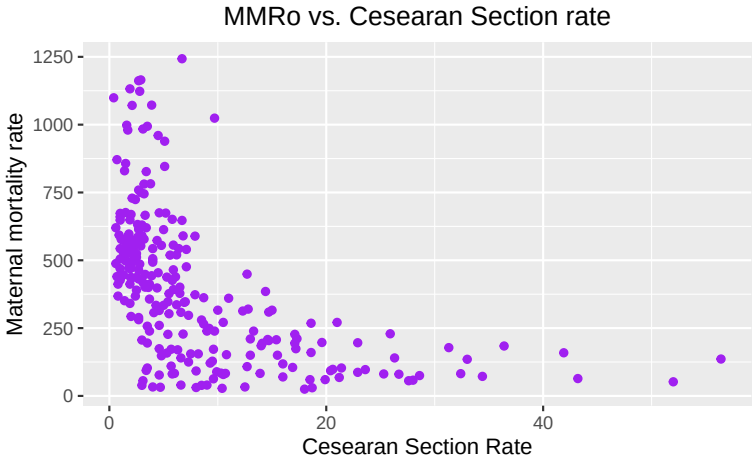


- `Thin`: One of the potential driving factors for high MMR could also be poor Body Mass Index, and this is reflected in the scatter plot as well. As the low BMI rate increases, the MMR increases as well

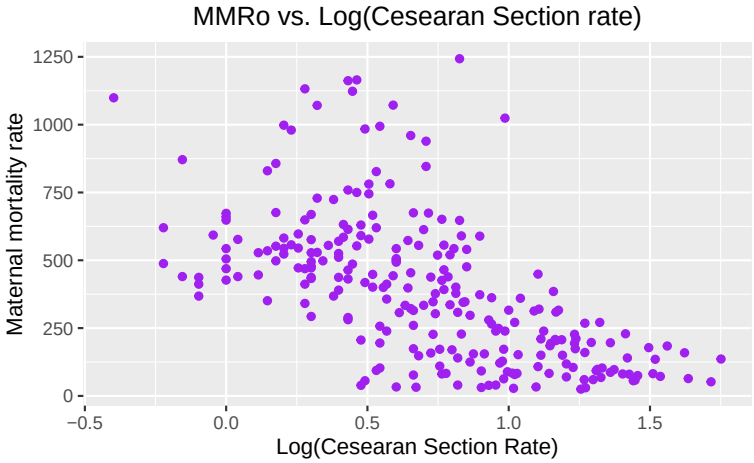


An interesting pattern was observed between MMR and the Cesearan section rate, wherein the linear

relationship was relatively weak as seen below



There is an exponential relationship which is observed instead. On performing log transformation of the variable `CS`, it is observed that the resulting relationship is relatively strongly linear with the response



Based on the above findings, it was determined that most of the variables do not have a clear linear relationship with the response, apart from `log(CS), Thin, CPR_Modern, and Anemia_preg`. In the upcoming sections, this is proven this statistically, through variable selection techniques.

Regression Analysis

Variable Selection

As the aim of this report is to not only build a robust model for predicting MMR, but also to identify the drivers for MMR using variables which can be easily measured, our goal was to do proper variable selection, followed with any variable transformations, and finally training various regression algorithms to arrive at the model with the best predictive power.

Selection process invalidates any inferences we make about the parameters including standard t-tests, F-tests and confidence intervals. To address this, data is split into two non-overlapping training and testing sets. Variable selection will be done on the training and inferences from the testing set.

- Training data - for fitting various regression algorithms
- Test data - for evaluating the candidate models based on a quality criterion

We used four different selection procedures to identify the variables which perform well in terms of Adjusted R-Squared. Our approach was as follows:

- Identify the best coefficients selected from each of the procedures - *Forward Selection AIC*, *Forward Selection BIC*, *Backward Selection AIC*, *Backward Selection BIC*, *Step-wise AIC*, *Step-wise BIC*, *Best subset selection AIC*, and *Best subset selection BIC*
- Refit OLS model using the best set of variables obtained from each method
- Evaluate each unique model based on $RMSE_{LOOCV}$

After performing the iterations as mentioned above, we obtained three unique set of models for which the selection procedures yielded the best quality criterion.

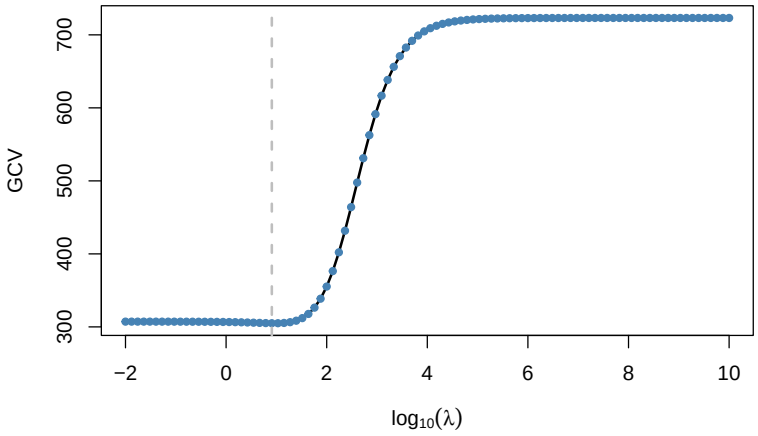
- Model 1: `MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145`
- Model 2: `MMRo ~ CPR_modern + HID + HIV_final + Ht_145`
- Model 3: `MMRo ~ CPR_modern + log(CS) + HIV_final + HID`

On fitting OLS models for each of the variable set, we calculate the $RMSE_{LOOCV}$ as well as the Adjusted R-squared, to choose our final set of variables. Based on the results as given in the table below, we choose Model 1, with a total of 5 variables `CPR_modern, HID, log(CS), HIV_final, and Ht_145`, as it has both the lowest RMSE as well as the highest Adjusted R-squared.

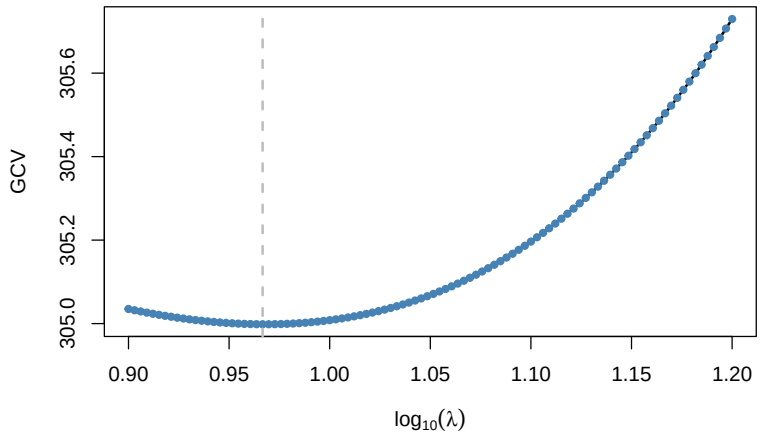
Model	$RMSE_{LOOCV}$	Adjusted R-squared
Model 1	191.976	56.33%
Model 2	193.17	55.71%
Model 3	194.598	55.19%

Ridge Regression for variable selection

After identifying the subset of variables from stage 1, we attempted to refine our predictors list by implementing Ridge Regression. This will enable us to be more confident of our predictors and provide a better and more robust explanation of the model. We began by specifying a range of λ values to try.



Based on this range of values, the value that minimizes the GCV error is 10.723. To refine our search, we tried a denser grid of 100 evenly spaced λ values between 0.5 and 1.



According to the more finer grid, the best value of λ is 10.501. Using this value we implement a Ridge Regression model to arrive at the significant predictors. The significant predictors can be obtained from the model summary.

```
##
## Call:
## lmridge.default(formula = MMRo ~ CPR_modern + HID + log(CS) +
##     HIV_final + Ht_145, data = mmr_train, K = k_best, scaling = "scaled")
##
##
## Coefficients: for Ridge parameter K= 9.261187
##             Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
## Intercept   676.8961     5940.6154   1295.6215       4.5851   <2e-16 ***
## CPR_modern   -2.4810      -47.9551     19.5942      -2.4474   0.0160 *
## HID          -3.0403      -79.6532     21.0161      -3.7901   0.0003 ***
## log(CS)     -54.9951      -53.1709     22.2170      -2.3932   0.0185 *
## HIV_final    19.7533      125.6235     16.4389       7.6419   <2e-16 ***
## Ht_145       -7.1267      -39.9831     17.7203      -2.2564   0.0261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge Summary
##         R2     adj-R2   DF ridge          F        AIC        BIC
##    0.55040    0.53340    4.31456   33.37944 1157.53970 1691.98798
## Ridge minimum MSE= 2202.596 at K= 9.261187
## P-value for F-test ( 4.31456 , 106.1464 ) = 8.772693e-19
## ------------------------------------------------------------------
```

Based on the results of the model summary, we found that all the variables, `CPR_modern, HID, log(CS), HIV_final, and Ht_145`. We simultaneously also check for the predictors that are significant using the Lasso model.

Lasso Regression for variable selection

Similar, to Ridge, a grid of values was specified for λ. On further deep dive, the best value of λ_{min}, was found to be 0.1353, while $\lambda_{1.se}$ was found to be 48.36555

Using λ_{min}, none of the predictors were deemed to be non-significant and did not have zero coefficients. However, on considering $\lambda_{1.se}$, it was observed that the coefficients for one predictor shrunk to zero. The final set of variables that are important according to Lasso `CPR_modern, HID, log(CS), and HIV_final`.

To further evaluate the two methods, using the final model chosen, we predict on the test data with the

quality criterion as R-squared.

Model	R-squared on test	RMSE on test
Ridge	62.446%	158.0957
Lasso (min)	62.1668%	158.7063
Lasso (1.se)	62.507%	165.8849

As observed from the table above, based on the predictions on test-data, we chose the Lasso (1.se) model, as the final choice for variable selection. This is because, with more number of variables, as seen in Ridge and Lasso(min), the R-squared does not improve significantly on the test-data

Hence, the variable selection yielded 4 variables, `CPR_modern, HID, log(CS), and HIV_final` with which we proceed to fit an OLS model, and perform model diagnostics.

Fitting the model

Upon fetching the best variables, an OLS model was fit on the test-data and diagnosed for any LINE violations and collinearity. Below, is a snapshot of the summary of the OLS model.

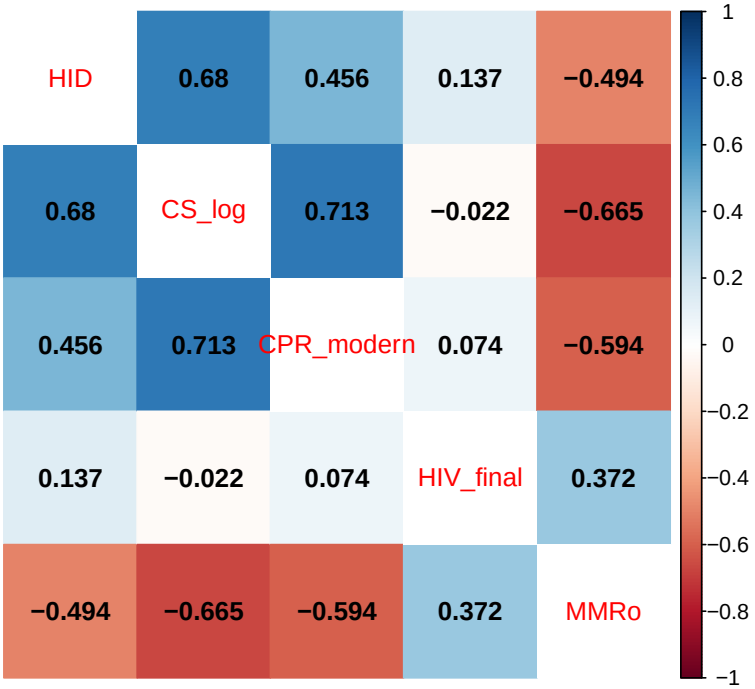
```
##
## Call:
## lm(formula = MMRo ~ CPR_modern + HID + log(CS) + HIV_final, data = mmr_test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -310.22  -75.92  -16.61   65.61  472.46
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 689.3329    34.8436  19.784  < 2e-16 ***
## CPR_modern   -4.2010     0.9962  -4.217 4.85e-05 ***
## HID          -2.0923     0.7476  -2.799  0.00599 **
## log(CS)     -67.0641    24.0888  -2.784  0.00625 **
## HIV_final    19.3145     2.6051   7.414 1.96e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.3 on 119 degrees of freedom
## Multiple R-squared:  0.6413, Adjusted R-squared:  0.6293
## F-statistic:  53.2 on 4 and 119 DF,  p-value: < 2.2e-16
```

We observe that the all the variables, are significant in terms of their β estimates, at the α=0.05 level. The adjusted R-squared is also decent with a value of 62.93%. The F-test also also yielded a low value of p. Thus, at-least, one of the predictors has a significant linear relationship with the response given that there are other predictors in the model.

Model Diagnostics

The first step toward diagnosing the OLS model, is identifying whether there is collinearity in the data. Any pairwise collinearity can also be assessed using the correlation plot

Collinearity



From the correlation plot, it can be observed that almost all the variables, have a decent linear relationship with the response. However, the variable pairs `log(CS)` and `HID`, and `log(CS)` and `CPR_modern` are also positively correlated. We checked for the overall collinearity in the model using condition number

```
##   Eigenvalue Condition Index
## 1     3.9837          1.0000
## 2     0.6782          2.4236
## 3     0.1836          4.6579
## 4     0.1062          6.1252
## 5     0.0482          9.0867
```

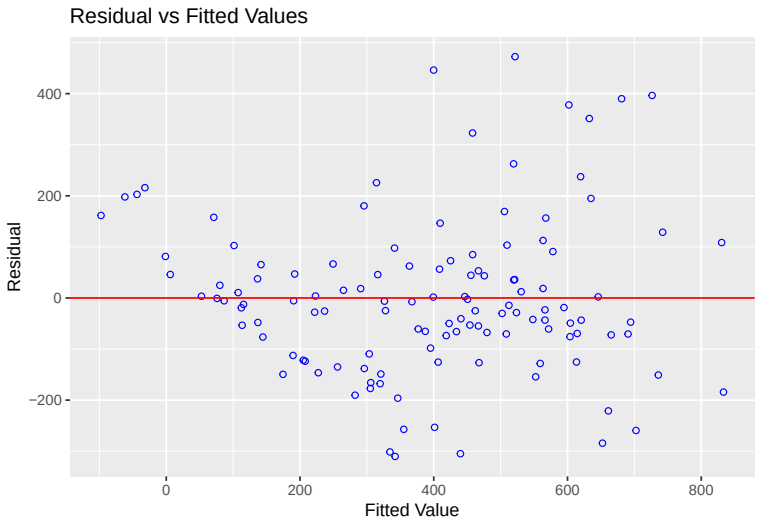
None of the condition indices have a value greater than 30. Hence, there does not seem to be any collinearity in the data. We verified the same using VIF or Variance inflation Factor.

```
## CPR_modern        HID    log(CS)  HIV_final
##   2.080563   1.958817   3.144150   1.067280
```

The VIF is also less than 5, thus suggesting that there is no multicollinearity in the model/data. Based on the two tests of collinearity, we have safely concluded that there is no collinearity in the data.

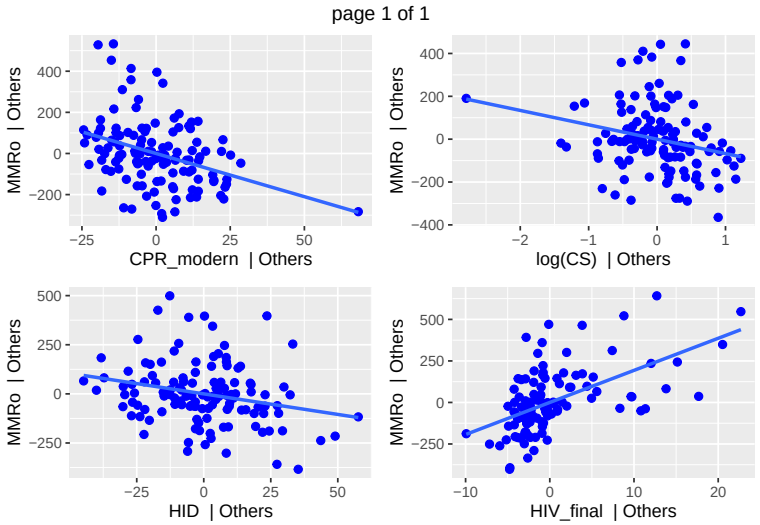
Model Violations

Our interpretations about the model might not hold true, if the *LINE* assumptions of OLS model are violated. As part of this project, we are aiming to establish the potential indicators of MMR as well. Therefore, we carried out a few tests of violation, beginning with *Linearity* and *Constant Variance* assumption.



On analysing, the fitted vs. residuals plot above, we observed that there was some patterns observed, where constant variance is concerned. It seems to have been violated, as the residuals on the right side of the graph seem to be more spread out. Also, the errors seem to show a pattern versus the fitted line.

Partial regression plots were analysed, to check for any linearity violation.



From these plots, it looks like there is a significant linear relationship between each predictor and the response, given the other predictor. However, there does seem to be influential observations, that skews the relationship between `MMRo` and `log(CS)`. This will be addressed in the upcoming sections.

To check for the constant variance violation, a statistical test - Breusch Pagan test was also performed.

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_model
## BP = 9.363, df = 4, p-value = 0.05264
```

The resulting p-value is 0.05264, which is more than 0.05. We reject the null hypothesis, and conclude that the errors are homoscedastic at $\alpha = 0.05$ level. The normality assumption was also evaluated using Q-Q plots and the Shapiro Wilk test.

As seen, from the Q-Q plot, the normality assumption is violated on the upper end of the plot, where the values do not lie along the line and deviate quite a bit. This is verified using a Shapiro-Wilk test.

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model)
## W = 0.95798, p-value = 0.0006928
```

The p-value of the model is 0.00064 which lesser than 0.05, and hence we reject the null hypothesis. The conclusion is that the errors are not normally distributed. As we have two potential model violations, we will attempt to detect any high leverage or high influential observations, which if removed can fix the violations.

```
## 119 129 210 227 245 246
##  57  62 108 116 123 124
```

There are 12 high leverage observations in the model. Next, the outliers were also checked, and there were no outliers detected.

```
## named integer(0)
```

```
##   8  52  64 132 139 140 193 209 210
##   2  27  34  63  66  67  97 107 108
```

On checking for high influential data-points, it was observed that there are 18 such observations. The next step was to remove these observations, refit the model and check the model diagnostics. Below is a summary of the new model.

```
##
## Call:
## lm(formula = MMRo ~ CPR_modern + HID + log(CS) + HIV_final, data = mmr_test,
##     subset = noninfluential_ids)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -281.43  -63.99  -10.02   66.86  418.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 643.6583    27.9248  23.050  < 2e-16 ***
## CPR_modern   -3.3360     0.7924  -4.210 5.24e-05 ***
```

```
## HID          -2.1525     0.6049  -3.558 0.000552 ***
## log(CS)     -61.2779    19.1124  -3.206 0.001761 **
## HIV_final    19.0030     2.2757   8.350 2.25e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 121.3 on 110 degrees of freedom
## Multiple R-squared:  0.7168, Adjusted R-squared:  0.7065
## F-statistic:  69.6 on 4 and 110 DF,  p-value: < 2.2e-16
```

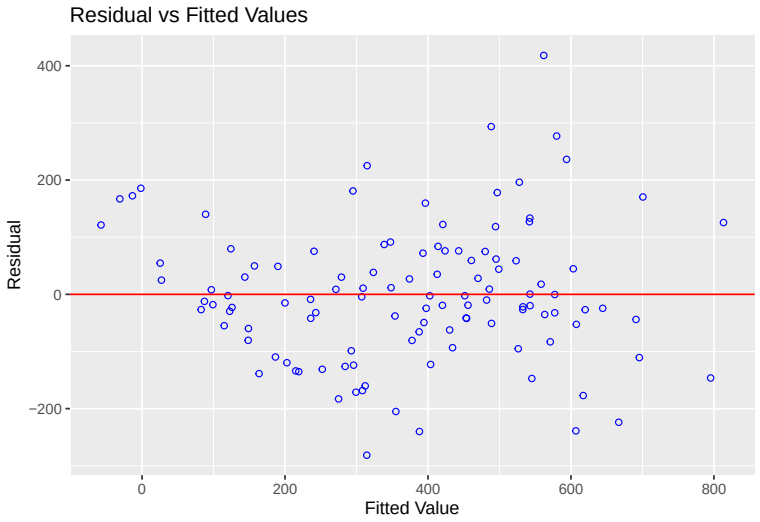
We then performed a both the Breusch Pagan test & Shapiro-Wilk test, to check whether the model is not violating the assumptions.

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_model_fix
## BP = 2.8032, df = 4, p-value = 0.5913
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model_fix)
## W = 0.98355, p-value = 0.1721
```

On removing the high influential observations, the Shapiro test has a p-value of 0.1721. This is more than 0.05. Hence, at the $\alpha = 0.05$ significance level, we fail reject the null hypothesis and concluded that the errors of the refit model are normally distributed.

The constant variance assumption was also verified again, by performing the Breusch-Pagan test. For the new model, the p-value this time was 0.5913, which is much higher than 0.05. Therefore, we conclude that the errors are homoscedastic.



On analysing the residuals vs the Fitted values for this new model,it did not show any patterns in the data. The errors seem to randomly distributed along the 0 line, and there is no clear pattern which is observed for the residuals.

Since, the refit model is an OLS model, and the errors satisfy all the classical assumptions of an OLS model, we conclude that these are the best linear unbiased estimates, in other words, the OLS estimates will produce the smallest variance of all possible linear estimators.

Discussion

Our regression analysis consisted of quite a few stages, which began with variable selection, followed by training an OLS model, to finally fixing any violations in the data. We concluded that the OLS model will provide the best estimates as it did not violate any of **LINE** violations. The estimated regression equation is given below:

Estimated Regression equation

$$\hat{MMRo_i^{0.4571}} = 643.6583 - 3.336CPR\_modern_i - 2.1525HID_i + 19.003HIV\_final_i - 61.2779log(CS)_i$$

Interpreting the coefficients

The coefficient interpretations both in the context of the model, and our hyothesis at the beginning of this exercise were assessed.

- `CPR_modern` - This is the contraceptive prevalence rate. In the context of the model, with a 1% increase in the contraceptive prevalence rate, the mean or expected Maternal Mortality rate will decrease by 3.336%, given that there are other predictors in the model.

This is also confirms our hypothesis that the use of contraceptives methods, will significantly reduce the number of maternal deaths occurring from pregnancy, as it promotes maternal health by reducing the number of pregnancies for a woman. It is recommended, based on the model, that contraceptive prevalence rate be measured and used as a potential indicator of Maternal Mortality rate

- `HID` - This is the Health Institution delivery rate. In the context of the model, with a 1% increase in the Health Institution Delivery rate, the mean or expected Maternal Mortality rate will decrease by 2.153%, given that there are other predictors in the model.

Again in the context of the real world, it is expected that the maternal deaths will decrease if there is proper medical care and delivery. A hospital or health institution's medical conditions, quality of gyneacologists, and delivery rate, will determine the safety of the mother when it comes to her pregnancy.

- `log(CS)` - This is the Cesearan Section rate. In the context of the model, with a 1% increase in the logarithm of Cesearan Section rate, the mean or expected Maternal Mortality rate will increase by 61.28%, given that there are other predictors in the model.

Recall, that we considered the natural logarithm of `CS` as a potential predictor for MMR, as the scatterplot showed a clear relationship between the two. As a potential indicator, it confirms our hypothesis, that the higher will be the CS rate, the lower the number of natural deliveries, or mid-wife induced deliveries, which in turn will lead to lower maternal deaths.

- `HIV_final` - This is the % of women with HIV. In the context of the model, with a 1% increase in the HIV prevalence rate, the mean or expected Maternal Mortality rate will decrease by 19%, given that there are other predictors in the model.

Intuitively, HIV is life threatening disease, and is one of the main causes for maternal deaths worldwide. To prevent maternal related deaths, there should be frequent awareness program for HIV. If a women suffers from HIV, proper medical care should be provide to avoid any maternal complications and loss of life

R-squared and significance tests

The final OLS model with has an adjusted R^2 of 70.65%. This means, that about 70.65% of variance in the model, can be explained by the linear relationship with the predictors `CR_modern`, `HID`, `log(CS)`, and `HIV_final`.

The t-tests for each of the predictors are also significant. The p-values for all the predictors are less than 0.05. Therefore, we reject the null hypothesis that β values are 0 at the $\alpha = 0.05$ significance level.

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 643.658326  27.924819 23.049686 6.441607e-44
## CPR_modern   -3.335966   0.792440 -4.209739 5.244398e-05
## HID          -2.152541   0.604944 -3.558249 5.523370e-04
## log(CS)     -61.277876  19.112418 -3.206181 1.761311e-03
## HIV_final    19.002976   2.275737  8.350251 2.253458e-13
```

Therefore, each of the predictor has a significant linear relationship with the response, given that there are three other predictors in the model.

The F-statistic is 69.6 with a p-value of less than 0.05. We therefore conclude atleast one of the predictors from `CR_modern`, `HID`, `log(CS)`, and `HIV_final`, has a significant linear relationship with the response. This was anyway concluded from the t-tests at the 5% significance level.

Our regression model, therefore not only has the best linear unbiased estimates, but is also stable and robust to any changes to data. The interpretations will hold true because of the absence multi-collinearity, and we have also removed any potential over-fit issues by splitting the data into a training and test-data. However, there are a couple of limitations that is discussed in the next section.

Limitations

Although our model, is potentially free of any OLS violations, we have very less data to make real-world decisions based on this model. Ideally, if the data were huge and contained a lot more surveys, than the 248 available, it would have captured more variance, and fetched more robust estimates.

A real test of any model is it's performance on unseen data. However, due to lack of data-points, we did not create a validation data-set. Therefore measuring the true predictive power would require more data. Other algorithms such as GLS, WLS, and Huber, were not implemented since, we got the satisfactory results based on the data available. If there was a lot more samples, then our results may have been different, and the model would have not satified all the classical OLS assumptions. The more the data, the better our estimates of the population mean.

Secondly, the value of R^2 is not as high as we would expect from a linear regression model.Ideally, a score of more than 85% is what we wanted to accomplish. Perhaps, this could have been achieved with more variables in the model. However, this would have affected the interpretation of the coefficients, and the explain-ability of the model. Although the predictions would have been good, the interpretation would be useless. So it always a trade-off.

Conclusion

The aim of this regression analysis was to establish and estimate the indicators and drivers for Maternal Mortality Rate. As it is difficult to measure MMR, due to lack of proper methods and techniques, we attempted to build a regression model, that can accurately measure MMR, and also provide an understanding of the factors that affect it. The data used for this experiment, was 248 observations with 9 potential predictors. We observed from the initial EDA, that there are 4 variables that have a strong linear relationship with MMR. The variable, `CS` or cesearan rate was transformed using a natural logarithm, as the log of the variable showed a linear relationship pattern with the response.

Although, there were not many variables to begin with, we performed variable selection using Forward, Backward, Stepwise, and Best Subset selection methods. It fetched 3 unique models, out of which the model with least RMSE was chosen. The variable selection was refined further using Lasso and Ridge regression, to choose only the significant predictors. Post this analysis, and OLS model was fit using the test data, to make inferences on the model and perform model diagnostics.

The candidate model was diagnosed for any violations of OLS model assumptions, or **LINE** assumptions. It violated the normality assumption, which was fixed by removing 18 influential observations, based on the

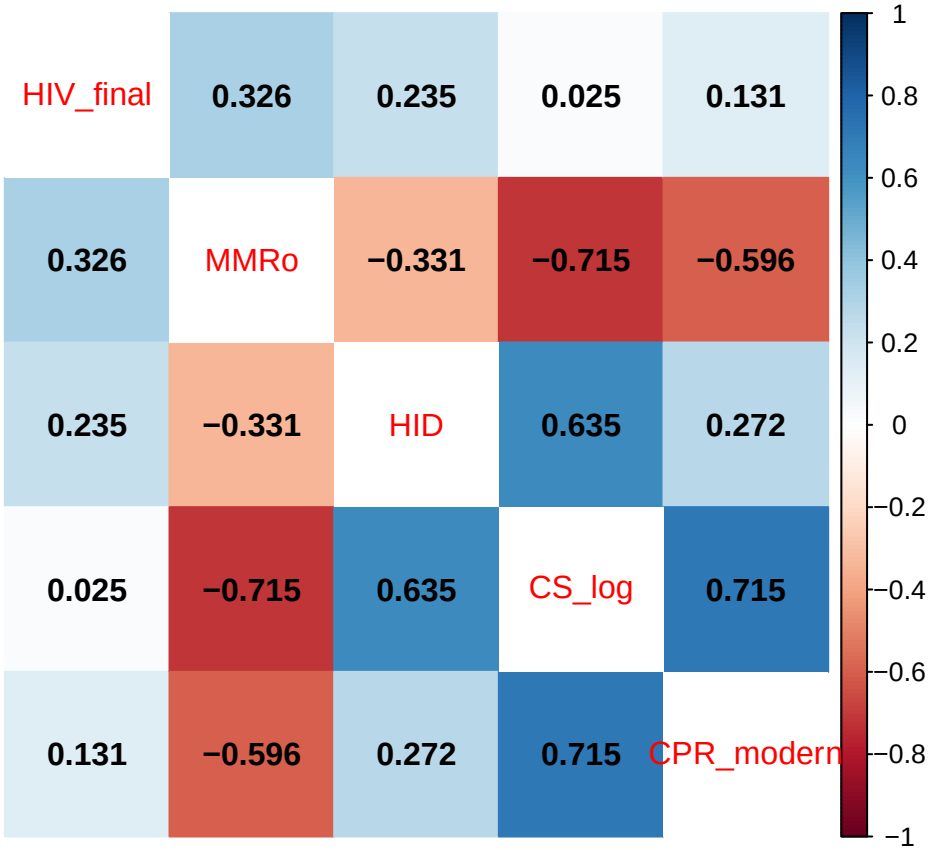
OLS model. The p-value from Shapiro Wilk test for the new model, was larger than 0.05. The resulting model, was free of any violations, with an R^2 of 71%. All of the predictors were significant as well, with a p-value less than 0.05.

As specified in the limitations, any future work on this would entail collection of more samples to have more robust estimates and a better R^2. Also, we must explore more variables which can be potential indicators of MMR, and data from various demographics to get a better understanding of the drivers for MMR across the world.

Additional work

Model with validation data as well to measure predictive power

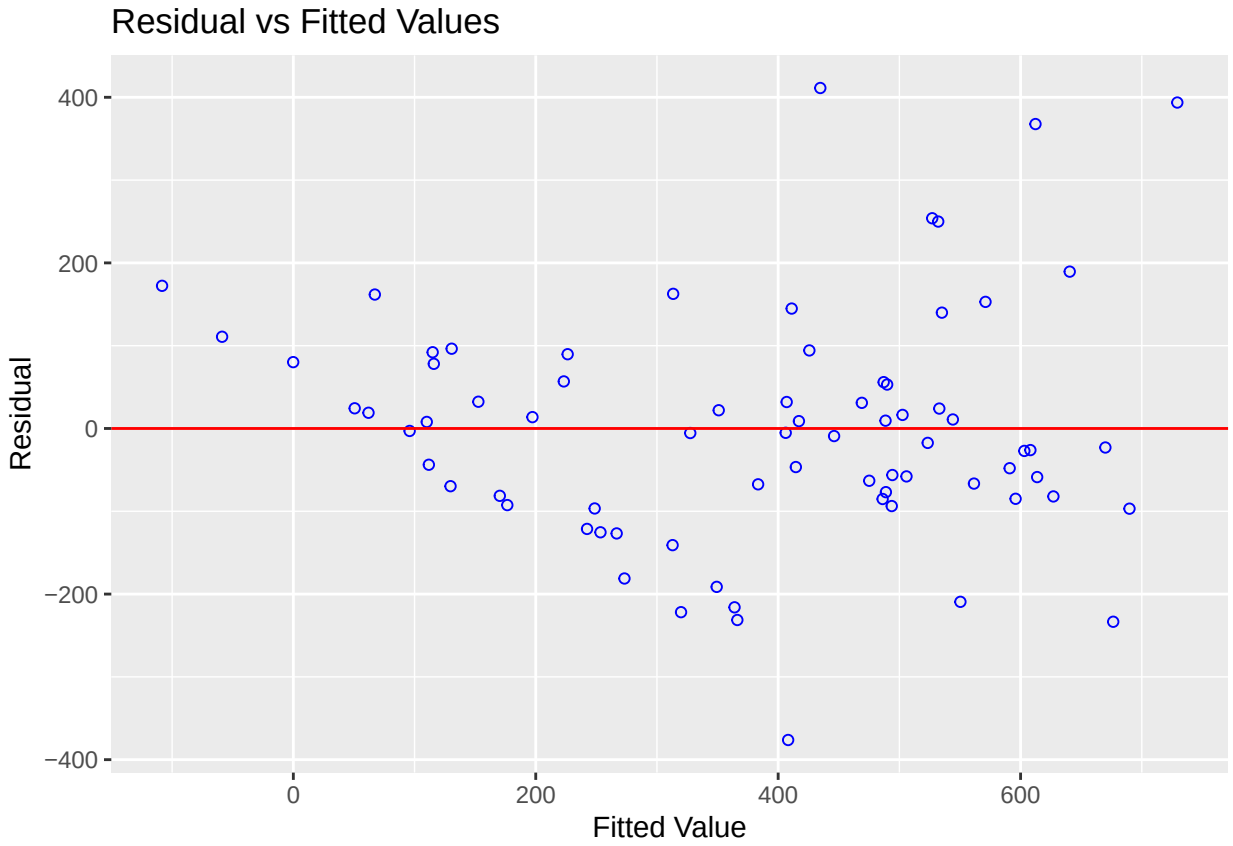
Data is split into validation data as well, to account for the predictive power of the model. We fit the new model based on the new test data-set, continuing from the predictors chosen by Lasso model.




```
##   Eigenvalue Condition Index
## 1     4.0434          1.0000
## 2     0.6415          2.5105
## 3     0.1877          4.6419
## 4     0.0981          6.4187
## 5     0.0293         11.7521
```

```
## CPR_modern        HID     log(CS)  HIV_final
##   2.510484   2.177163   3.983258   1.183961
```

There is no collinearity observed in the data as before. We check for model violations as before.

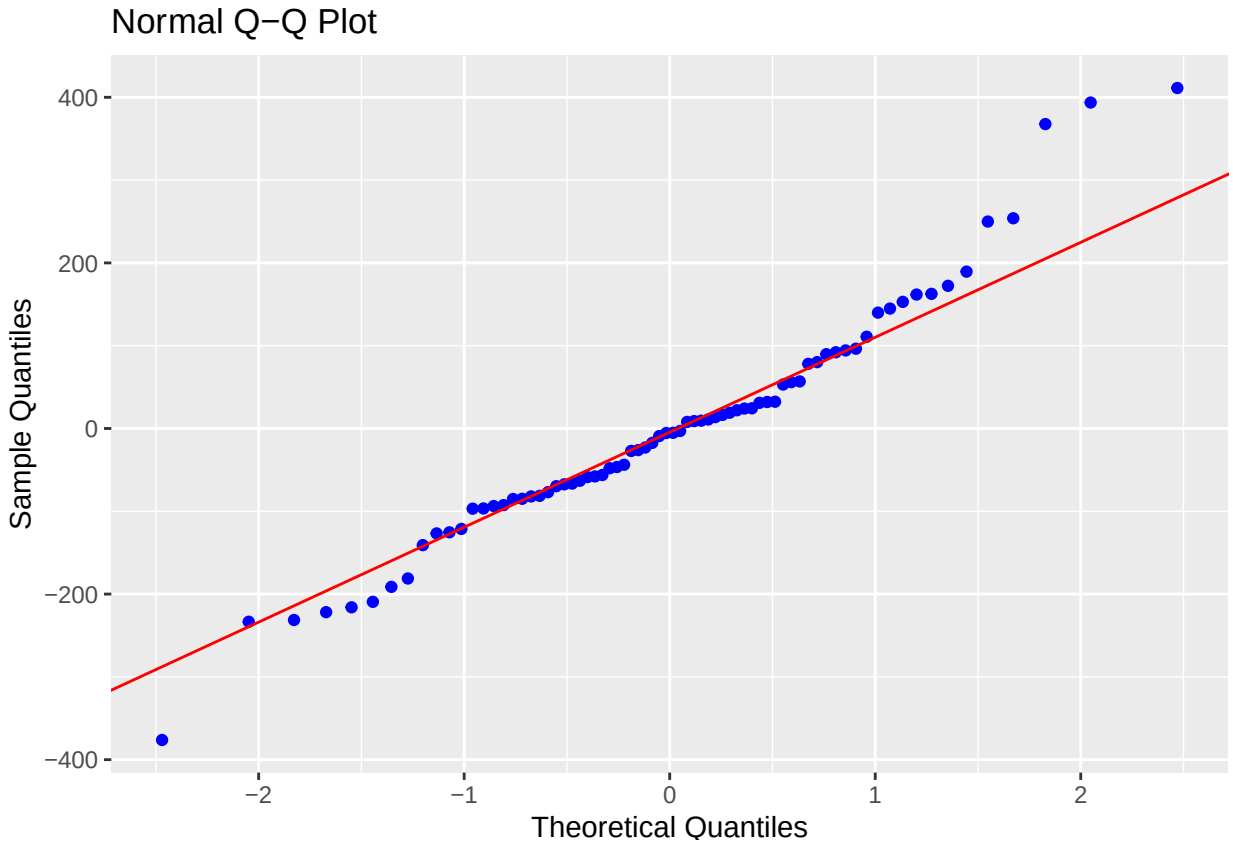


The residuals show some patterns in the data and might be violating the constant variance assumption and normality assumption.

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_model_new
## BP = 7.8916, df = 4, p-value = 0.09563
```

The residuals of the model do violate the constant variance assumption. This is because the p-value from the Breusch-Pagan test is 0.09, which is more than 0.05 and hence the errors are homoscedastic.

Next we check for the normality using q-q plot and shapiro test.




```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model_new)
## W = 0.96376, p-value = 0.03224
```

The p-value from the Shapiro Wilk test is 0.03, which is less than 0.05. Therefore, we reject the null hypothesis that the errors are normally distributed. The influential observations are removed to ensure the errors do not violate any assumption.

```
## 246 245 142 210
##   8  23  42  46
```

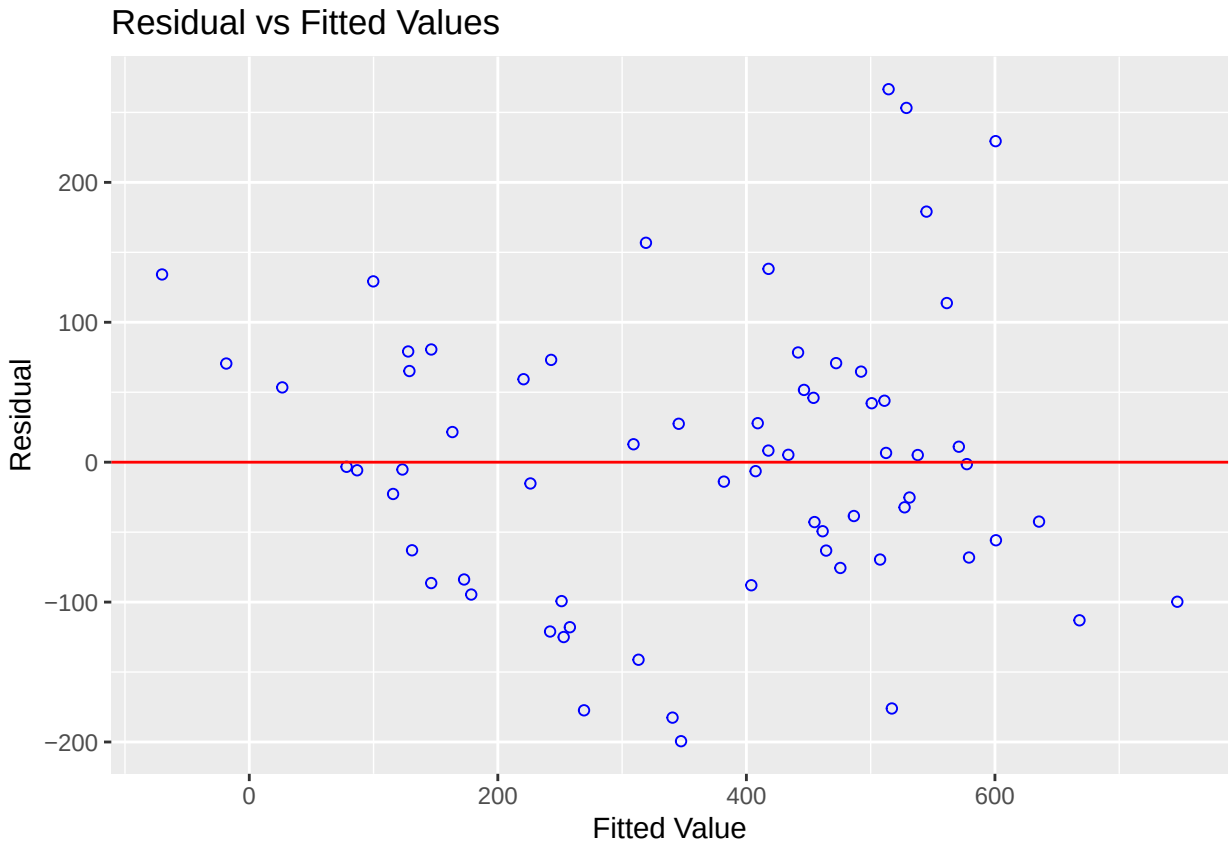
```
## named integer(0)
```

```
##  94  64 209 210   8 139 155
##   1  27  29  46  49  63  72
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(ols_model_fix_new)
## W = 0.9798, p-value = 0.346
```

```
##
##  studentized Breusch-Pagan test
##
## data:  ols_model_fix_new
## BP = 2.9591, df = 4, p-value = 0.5647
```

Both the shapiro wilk test and the BP Test, have p-values larger than 0.05. Hence, this is our final OLS model. We now chec whether the residuals show a random pattern from the fitted-vs-residuals plot.




```
##
## Call:
## lm(formula = MMRo ~ CPR_modern + HID + log(CS) + HIV_final, data = mmr_test_new,
##     subset = noninfluential_ids)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -199.428  -68.863   -3.289   62.017  266.578
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  621.11945   35.34300  17.574  < 2e-16 ***
## CPR_modern    -3.25738    1.02090  -3.191  0.00223 **
## HID           -0.09666    0.75903  -0.127  0.89908
## log(CS)     -113.40760   27.14447  -4.178 9.37e-05 ***
## HIV_final     19.98037    3.09326   6.459 1.84e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 104 on 62 degrees of freedom
## Multiple R-squared:  0.7728, Adjusted R-squared:  0.7582
## F-statistic: 52.74 on 4 and 62 DF,  p-value: < 2.2e-16
```

After checking, the summary of the model, we observed that the variable `HID` was not significant. That is the p-value from the t-test is 0.9 which is much higher than the $\alpha = 0.05$ significance level. This means that

given that there are other predictors in the model, `HID` does not have a significant linear relationship with the response.

We drop this variable, and fit a new model.

```
##
## Call:
## lm(formula = MMRo ~ CPR_modern + log(CS) + HIV_final, data = mmr_test_new,
##     subset = noninfluential_ids)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -197.542  -68.497   -3.891   62.995  263.860
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  618.6346    29.2380  21.159  < 2e-16 ***
## CPR_modern    -3.2050     0.9269  -3.458 0.000982 ***
## log(CS)     -115.8269    19.2368  -6.021 9.77e-08 ***
## HIV_final     19.8210     2.8066   7.062 1.56e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103.2 on 63 degrees of freedom
## Multiple R-squared:  0.7728, Adjusted R-squared:  0.762
## F-statistic: 71.42 on 3 and 63 DF,  p-value: < 2.2e-16
```

The new model is enhanced and has all predictors with significant β estimates. The adjusted R-squared value is strong as well at 76.2%. This is another candidate model, that was considered. We then tested its predictive power on the validation data.

```
## [1] 0.5721614
```

We get an R-squared value of 57.2% on the unseen data. This is much lower than the R^2 value from the training data. Hence, there is some over-fitting in the model, which needs to be improved. This can be done if we have more sample, which is a future scope.

Code Appendix

```
# EDA
## Data Cleaning

mmr = read.csv("mmr_data.csv")
library(dplyr)
library(tidyr)
library(ggplot2)

# Filled Density Plot
d <- density(mmr$MMRo)
plot(d, main="Probability distribution of Maternal Mortality Rate")
polygon(d, col="red", border="blue")

### Check for missing values
missing_values <- colSums(is.na(mmr))
```

```
#### One column has 16 missing values
#### to impute it we use either mean or median

boxplot(mmr$Ht_145,
        col=c("blue"),
        ylab='% of women with maternal stunting',
        main="Boxplot for Prevalence of Maternal Stunting")

#### it is observed that the data is skewed with a few outliers. hence, it would be appropriate to use

median_ht_145 <- median(mmr$Ht_145, na.rm = TRUE)
mmr$Ht_145[is.na(mmr$Ht_145)] <- median_ht_145

### Check for duplicates in the data
mmr[duplicated(mmr)]

### Check datatypes and verify all are numeric
str(mmr)

## Graphical Analysis

### Univariate analysis

#### Boxplot
for (col in colnames(mmr)) {
  boxplot(mmr[col],
  ylab = col,
  main = col)
}

### Bivariate Analysis

#### Scatter plot

library(ggplot2)

ggplot(mmr, aes(x = CPR_modern, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'CPR', y = "Mater

ggplot(mmr, aes(x = ANC_4, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'ANC_4', y = "Maternal

ggplot(mmr, aes(x = HID, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'HID', y = "Maternal mor

ggplot(mmr, aes(x = CS, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'CS', y = "Maternal morta

ggplot(mmr, aes(x = PNC, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'PNC', y = "Maternal mor

ggplot(mmr, aes(x = Animia_preg, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'Animia_preg', y

ggplot(mmr, aes(x = Ht_145, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'Short Stature', y =

ggplot(mmr, aes(x = Thin, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'Thin', y = "Maternal m

ggplot(mmr, aes(x = HIV_final, y = MMRo)) + geom_point(colour = "purple") + labs(x = 'HIV_Final', y = "
```

```r
ggplot(mmr, aes(x = log10(CS), y = MMRo)) + geom_point(colour = "purple") + labs(x = "Log(Cesearan Sect

# Regression Analysis
calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model)))^2))
}

# 50% train data as our sample is small
set.seed(42)
train = sample(1:nrow(mmr), round(0.5 * nrow(mmr)))
test = c(-train)

mmr_train = mmr[train,]
mmr_test = mmr[test,]


# Backward Selection AIC
# The model which backward search removes variables
# NOTE: It does not need to include all variables.

n = nrow(mmr_train)
mod_all_preds = lm(MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+Animia_preg, data = mm

# NOTE:  step defaults to using AIC
mod_back_aic = step(mod_all_preds, direction = 'backward',trace = 0)
coef(mod_back_aic)

# Backward Selection BIC
mod_back_bic = step(mod_all_preds, direction = 'backward',k=log(n), trace = 0,)
coef(mod_back_bic)

# Forward Selection AIC
mod_start = lm(MMRo ~ 1, data = mmr_train)

mod_forwd_aic = step(mod_start, scope = MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+A

coef(mod_forwd_aic)

# Forward selection BIC

mod_forwd_bic = step(mod_start, scope = MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+A
coef(mod_forwd_bic)

# Stepwise Selection AIC

mod_start = lm(MMRo ~ 1, data = mmr_train)
mod_step_aic = step(mod_start,
scope = MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+Animia_preg,
direction = 'both', trace = 0)
coef(mod_step_aic)

## BIC stepwise selection
```

```r
mod_step_bic = step(mod_start, scope = MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+An
coef(mod_step_bic)

## run best subsets selection
library(leaps)
## run best subsets selection
mod_subsets = summary(regsubsets(MMRo ~ CPR_modern+HID+log(CS)+PNC+Thin+HIV_final+ANC_4+Ht_145+Animia_p
coef_names = colnames(mod_subsets$which)
## coeficients of the best model using adjusted R2
best_r2_ind = which.max(mod_subsets$adjr2)
coef_names[mod_subsets$which[best_r2_ind, ]]

## best model using AIC
p = ncol(mod_subsets$which)
mod_aic = n * log(mod_subsets$rss / n) + 2 * (2:p)
best_aic_ind = which.min(mod_aic)
coef_names[mod_subsets$which[best_aic_ind, ]]

## best model using BIC
mod_bic = n * log(mod_subsets$rss / n) + log(n) * (2:p)
best_bic_ind = which.min(mod_bic)
coef_names[mod_subsets$which[best_bic_ind, ]]

# fitting the best models

model1 = lm(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145,data = mmr_train)
model2 = lm(MMRo ~ CPR_modern + HID + HIV_final + Ht_145, data = mmr_train)
model3 = lm(MMRo ~ CPR_modern + log(CS) + HIV_final + HID, data = mmr_train)

# evaluation
calc_loocv_rmse(model1)
calc_loocv_rmse(model2)
calc_loocv_rmse(model3)

summary(model1)
summary(model2)
summary(model3)


# ridge and lasso for variable selection

library(lmridge)

# lambda values evenly spaced on the log-scale from 10^10 to 10^-2.
grid = 10 ^ seq(10, -2 , length = 100)

# estimate ridge regression on the training data
mod_ridge = lmridge(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145, data = mmr_train, scaling =

# extract the GCV errors and lambda that minimizes the GCV error
k_est = kest(mod_ridge)

# a plot of GCV vs. log10(lambda)
```

```r
plot(log10(mod_ridge$K), k_est$GCV, type = 'l', lwd = 2,
     xlab = expression(log[10](lambda)), ylab = 'GCV')

points(log10(mod_ridge$K), k_est$GCV,
       pch = 19, col = 'steelblue', cex = 0.75)

# horizontal line at log10(kGCV), i.e.,
# the base 10 logarithm of the best lambda value
abline(v=log10(k_est$kGCV), lty = 'dashed', col = 'grey',
       lwd = 2)

# lambda values evenly spaced on the log-scale from 10^1 to 10^2.5.
grid = 10 ^ seq(0.5,1 , length = 100)

# estimate ridge regression on the training data
mod_ridge = lmridge(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145, data = mmr_train, scaling =

# extract the GCV errors and lambda that minimizes the GCV error
k_est = kest(mod_ridge)

# a plot of GCV vs. log10(lambda)
plot(log10(mod_ridge$K), k_est$GCV, type = 'l', lwd = 2,
     xlab = expression(log[10](lambda)), ylab = 'GCV')

points(log10(mod_ridge$K), k_est$GCV,
       pch = 19, col = 'steelblue', cex = 0.75)

# horizontal line at log10(kGCV), i.e.,
# the base 10 logarithm of the best lambda value
abline(v=log10(k_est$kGCV), lty = 'dashed', col = 'grey',
       lwd = 2)

# best lambda value chosen by GCV
k_best = kest(mod_ridge)$kGCV

# re-fit the model using the best value of lambda according to GCV
mod_ridge_best = lmridge(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145, data = mmr_train,scali

summary(mod_ridge_best)

## Lasso Regression

# matrix of predictors with the first column of ones removed
# NOTE: This is quivalent to
#   x_train = model.matrix(Salary ~ . - 1, data = hitters_train)
x_train = model.matrix(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145, data = mmr_train)[, -1]

# the vector of responses
y_train = mmr_train$MMRo

library(glmnet)

# set the random seed for reproducibility
```

```r
set.seed(123)

# fit the lasso using 10-fold cross-validation to determine lambda
mod_lasso = cv.glmnet(x_train, y_train)
# the logarithm in this plot is the natural logarithm
plot(mod_lasso)

# set the random seed for reproducability
set.seed(123)

# lambda values evenly spaced on the natural log-scale from e^1 to e^6.
grid = exp(seq(-2, 4, length=100))

# fit the lasso using 10-fold cross-validation to determine lambda

mod_lasso = cv.glmnet(x_train, y_train, lambda = grid)

# lambda.min
mod_lasso$lambda.min
# lambda.1se
mod_lasso$lambda.1se

coef(mod_lasso, s = 'lambda.min')

coef(mod_lasso, s = 'lambda.1se')

# quick function to calculate RMSE
rmse = function(y_true, y_pred) {
    sqrt(mean((y_true - y_pred)^2))
}

x_test = model.matrix(MMRo ~ CPR_modern + HID + log(CS) + HIV_final + Ht_145, data = mmr_test)[,-1]
y_test = mmr_test$MMRo

# Model 1: ridge

# predict on the new test data. This is the same as lm
y_pred = predict(mod_ridge_best, newdata = mmr_test)

# calculate the RMSE
cor(y_test, y_pred)^2
rmse(y_test, y_pred)


# Model 2: lasso with lambda.min

# predictions using lambda.min
y_pred = predict(mod_lasso, newx = x_test, s = 'lambda.min')

# calculate the RMSE

cor(y_test, y_pred)^2
rmse(y_test,y_pred)
```

```r
# Model 3: lasso with lambda.1se

# predictions using lambda.1se
y_pred = predict(mod_lasso, newx = x_test, s = 'lambda.1se')

# calculate the RMSE
cor(y_test, y_pred)^2
rmse(y_test,y_pred)

#Fitting the model
ols_model = lm(MMRo~CPR_modern + HID + log(CS) + HIV_final, data = mmr_test)
print(summary(ols_model))

#Collinearity
library(corrplot)

mmr_test$CS_log <- log(mmr_test$CS)
data = mmr_test[c('CS_log','HID','HIV_final','CPR_modern','MMRo')]
# NOTE: we pass the output of cor() to corrplot()
corrplot(cor(data),
         method = 'color', order = 'hclust',  diag = FALSE,
         number.digits = 3, addCoef.col = 'black', tl.pos= 'd', cl.pos ='r')

library(olsrr)
round(ols_eigen_cindex(ols_model)[, 1:2], 4)

library(faraway)
vif(ols_model)

#Model violations
ols_plot_resid_fit(ols_model)

library(olsrr)
ols_plot_added_variable(ols_model)

library(lmtest)
bptest(ols_model)

ols_plot_resid_qq(ols_model)
shapiro.test(resid(ols_model))
which(hatvalues(ols_model) > 2 * mean(hatvalues(ols_model)))

outlier_test_cutoff = function(model, alpha = 0.05) {
    n = length(resid(model))
    qt(alpha/(2 * n), df = df.residual(model) - 1, lower.tail = FALSE)
}

# vector of indices for observations deemed outliers.
cutoff = outlier_test_cutoff(ols_model, alpha = 0.05)

which(abs(rstudent(ols_model)) > cutoff)

which(cooks.distance(ols_model) > 4 / length(cooks.distance(ols_model)))
```

```r
# ids for non-influential observations
noninfluential_ids = which(
    cooks.distance(ols_model) <= 4 / length(cooks.distance(ols_model)))

# fit the model on non-influential subset
ols_model_fix = lm(MMRo~CPR_modern + HID + log(CS) + HIV_final,
               data = mmr_test,
               subset = noninfluential_ids)

shapiro.test(resid(ols_model_fix))
bptest(ols_model_fix)

ols_plot_resid_fit(ols_model)

#R-squared and significance tests
summary(ols_model_fix)$coefficients
```

References

Cresswell, Jennifer, and World Health Organization. 2023. *Trends in Maternal Mortality 2000 to 2020: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division*. Genève, Switzerland: World Health Organization.

Gebremedhin, Samson. 2018. "Development of a New Model for Estimating Maternal Mortality Ratio at National and Sub-National Levels and Its Application for Describing Sub-National Variations of Maternal Death in Ethiopia." Edited by Tiziana Leone. *PLOS ONE* 13 (8): e0201990. https://doi.org/10.1371/journal.pone.0201990.