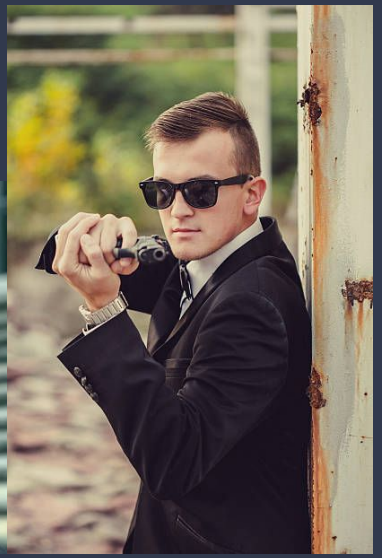
CIA Factbook

A Look into Maternal Mortality Rate

Carlos Aguilar & Harrison Plate



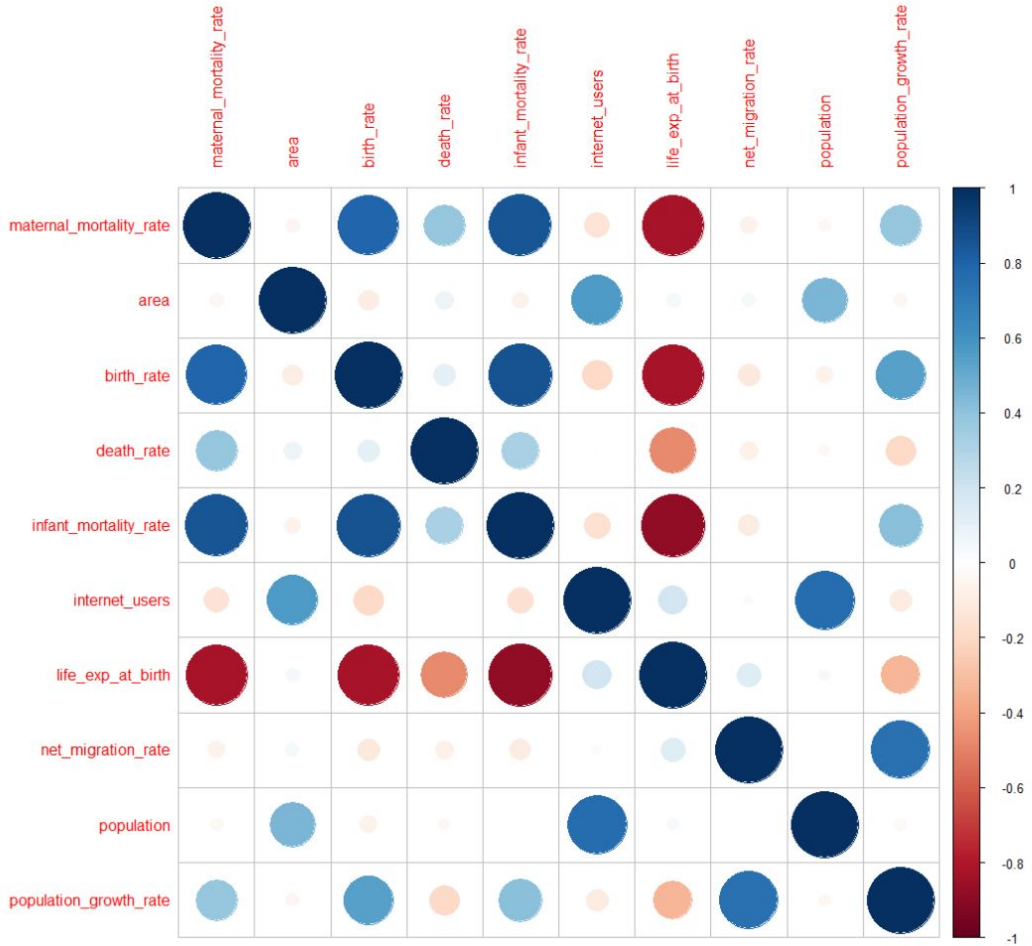
Data Description

- Collected in 2017
- Country-level statistics from the US Central Intelligence Agency (CIA)
- Data frame with 259 observations on 11 different variables

Goals

- Investigate what are the possible predictive variable in our data set that are linearly related to Maternal Mortality Rate?
- What are some of the inferences we can make with our data: What are some aspects we can improve to decrease maternal mortality rate?
- Try something new for our data set

CORRELATION MATRIX

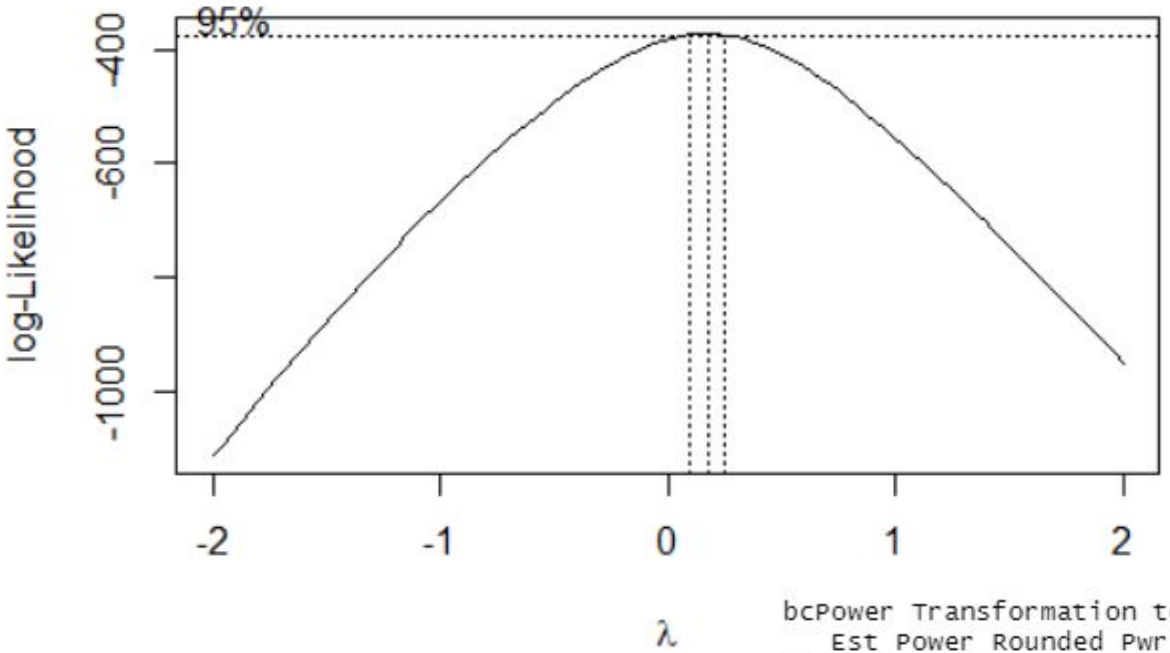


Predictive Variables for Full Model

Area
Birth Rate
Death Rate
Infant Mortality Rate
Life Expectancy at Birth
Net Migration Rate
Population
Population Growth Rate

$$H_0 : \beta_j = 0$$

$$H_A : \beta_j \neq 0 : \text{for at least 1 predictor}$$



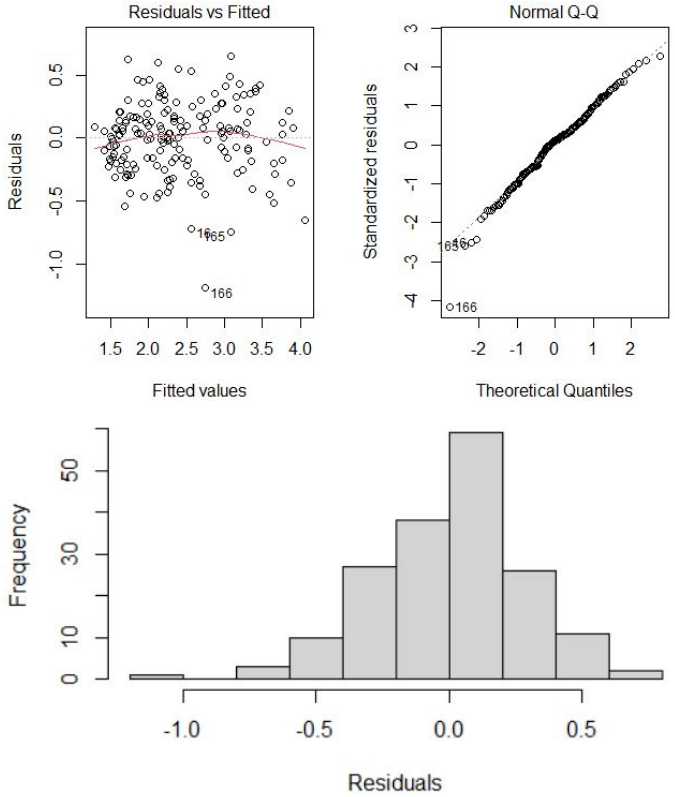
BOX-COX

```
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
```

Summary of Transformed Full Model

```
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
```

Significant!!!



Improving Our Model Step-by-Step

- Performed a step function in R to reduce our number of predictors being used
- Predictors are:
  - Birth Rate
  - Infant Mortality Rate
  - Death Rate
  - Life Expectancy at Birth

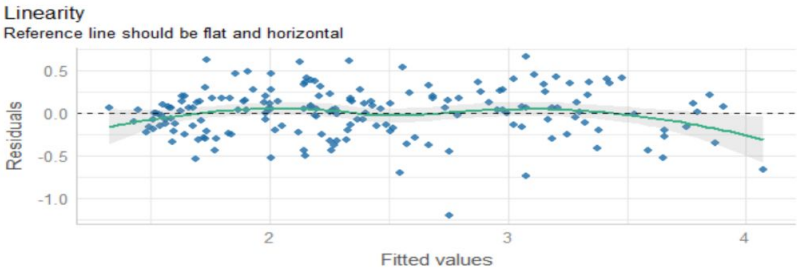
```
Step:  AIC=-435.54
(maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    life_exp_at_birth + birth_rate

                        Df Sum of Sq    RSS     AIC
<none>                              14.282 -435.54
- birth_rate             1    0.3830 14.665 -432.85
- infant_mortality_rate  1    1.2956 15.578 -422.17
- death_rate             1    1.6637 15.945 -418.03
- life_exp_at_birth      1    3.9079 18.190 -394.73
```

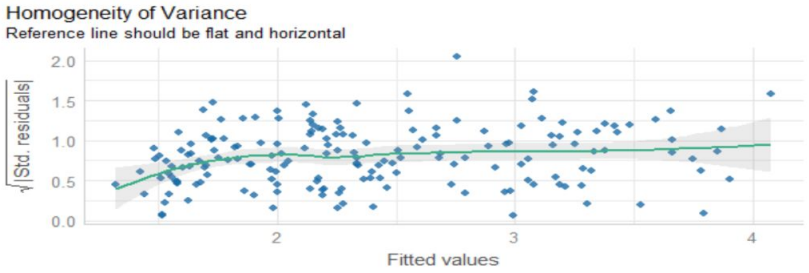
Done! Maybe..?

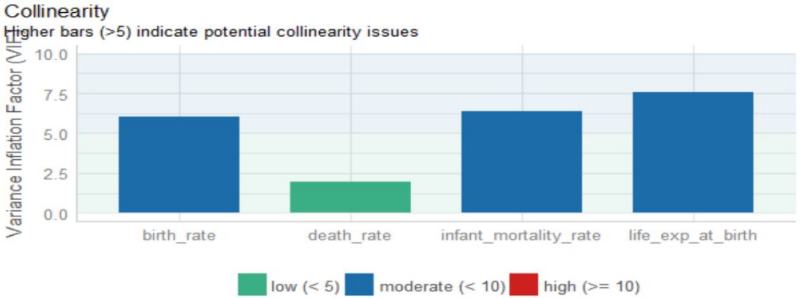
- 4 significant predictors
- Full model:
  - Adj. R^2: 0.841506
  - AIC: 75.99666
- Stepped model:
  - Adj. R^2: 0.8436858
  - AIC: 68.76710
- AOV p-value > 0.05 indicates that the 5 predictors can be dropped

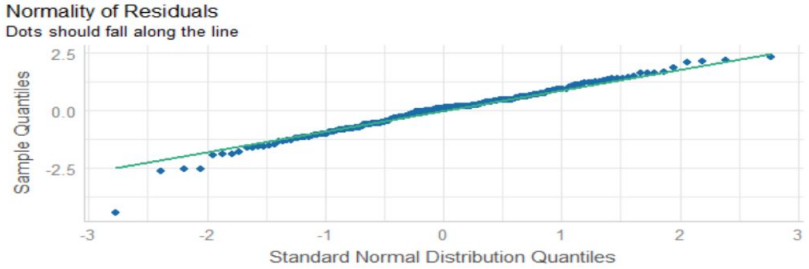
```
Call:
lm(formula = (maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate
+
    life_exp_at_birth + birth_rate, data = df_final)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2042 -0.1742  0.0360  0.1739  0.6626

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.547843   0.614250   9.032 3.36e-16 ***
death_rate          -0.042858   0.009575  -4.476 1.38e-05 ***
infant_mortality_rate  0.008784   0.002224   3.950 0.000114 ***
life_exp_at_birth   -0.046587   0.006791  -6.860 1.19e-10 ***
birth_rate           0.011681   0.005439   2.148 0.033144 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2882 on 172 degrees of freedom
Multiple R-squared:  0.8472,    Adjusted R-squared:  0.8437
F-statistic: 238.5 on 4 and 172 DF,  p-value: < 2.2e-16
```

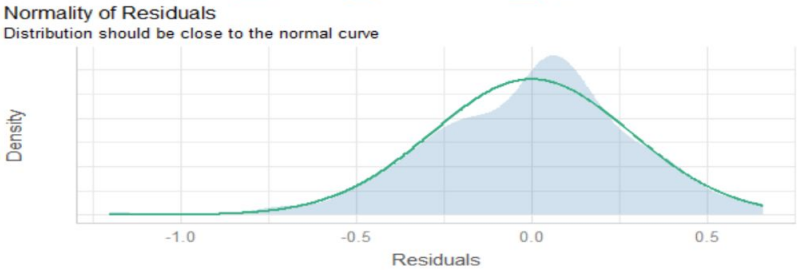
```
Analysis of Variance Table

Model 1: (maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    life_exp_at_birth + birth_rate
Model 2: (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
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    172 14.282
2    167 14.060  5    0.2218 0.5269 0.7557
```

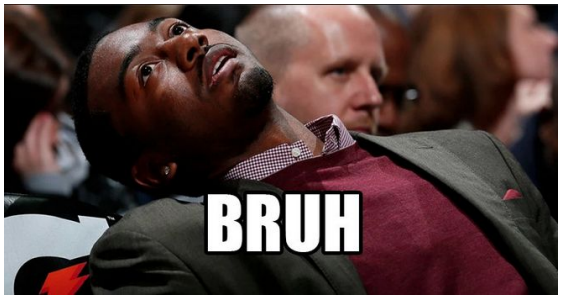












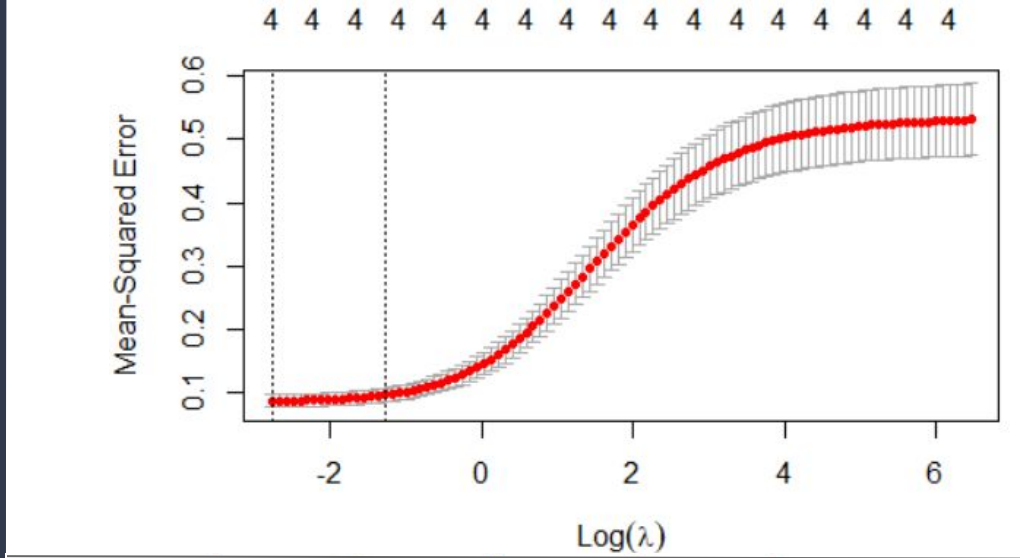
Collinearity Issue

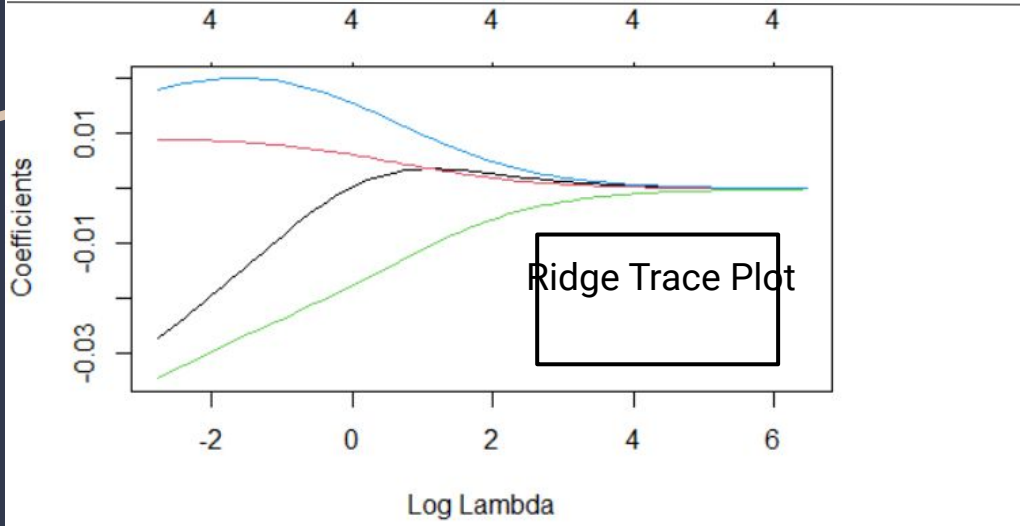
Ridge Regression

- We want to find the best lambda that produces the lowest MSE
  - The lowest MSE produces the best model
- Best Lambda: 0.06342977
- The ridge regression signifies that life_exp_at_birth is the least important predictor

```
5 x 1 sparse Matrix of class "dgCMatrix"
                                    s0
(Intercept)                4.431683368
death_rate                -0.027119913
infant_mortality_rate      0.008939495
life_exp_at_birth         -0.034548217
birth_rate                 0.018178334
```

Black
Red
Green
Blue





Collinearity & Low p-value

```
Analysis of Variance Table

Model 1: (maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    birth_rate
Model 2: (maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    life_exp_at_birth + birth_rate
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    173 18.190
2    172 14.282  1    3.9079 47.063 1.189e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The low p-value indicates that life_exp_at_birth is a significant predictor
  - We chose to drop it due to the results of the ridge regression
- Alternative option: stack the collinear data into 1 column

Insignificant Predictors: Round 2

- After removing life expectancy at birth, we find that death rate could also be removed
- Step #2

- Before 2nd Step:
  - Adj. R^2: 0.8020654
  - AIC: 109.5773
- After 2nd Step:
  - Adj. R^2: 0.8028622
  - AIC: 107.8835

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.00872 -0.21522  0.02784  0.21176  0.83474

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.384471   0.106733  12.971  < 2e-16 ***
death_rate          -0.004806   0.008782  -0.547    0.585
infant_mortality_rate 0.015740   0.002227   7.068 3.71e-11 ***
birth_rate           0.029925   0.005339   5.605 8.08e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3243 on 173 degrees of freedom
Multiple R-squared:  0.8054,    Adjusted R-squared:  0.8021
F-statistic: 238.7 on 3 and 173 DF,  p-value: < 2.2e-16
```

```
Start:  AIC=-394.73
(maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    birth_rate

                      Df Sum of Sq    RSS     AIC
- death_rate           1    0.0315 18.221 -396.42
<none>                             18.190 -394.73
- birth_rate           1    3.3032 21.493 -367.19
- infant_mortality_rate 1    5.2521 23.442 -351.83

Step:  AIC=-396.42
(maternal_mortality_rate)^(1/5) ~ infant_mortality_rate + birth_rate

                      Df Sum of Sq    RSS     AIC
<none>                             18.221 -396.42
- birth_rate           1    4.0569 22.278 -362.84
- infant_mortality_rate 1    6.1802 24.401 -346.73
```

```
Analysis of Variance Table

Model 1: (maternal_mortality_rate)^(1/5) ~ infant_mortality_rate + birth_rate
Model 2: (maternal_mortality_rate)^(1/5) ~ death_rate + infant_mortality_rate +
    birth_rate
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    174 18.221
2    173 18.190  1   0.03149 0.2995 0.5849
```

Final Model : $\widehat{MaternalMortalityRate}^{.2} = \widehat{\beta}_0 + \widehat{InfantMortalityRate}x_1 + \widehat{BirthRate}x_2 = 1.339008 + 0.015183x_1 + 0.030970x_2$

```
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
```

AOV & Final Model

Final Model Interpretation

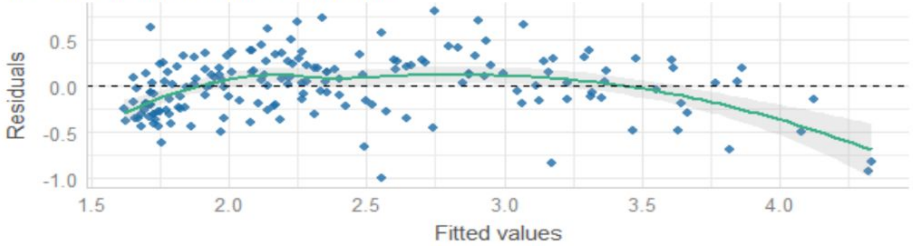
A one unit increase in infant_mortality_rate (1 more death per 1,000 live births), with the other predictor (birth_rate) held fixed, is associated with an increase in maternal_mortality_rate by (0.015183)^5 units, which equals 8.06841002E-10 units, which can be interpreted as 8.06841002E-10 more deaths (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.

```
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
```

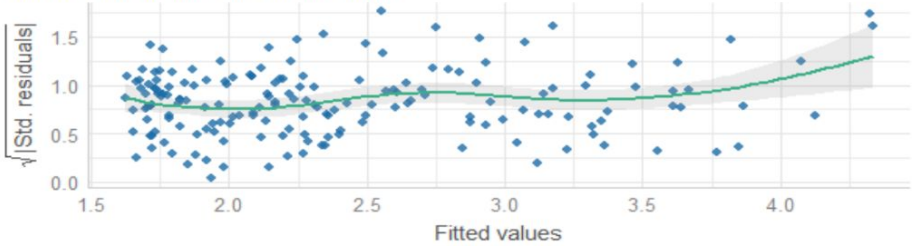
Linearity

Reference line should be flat and horizontal



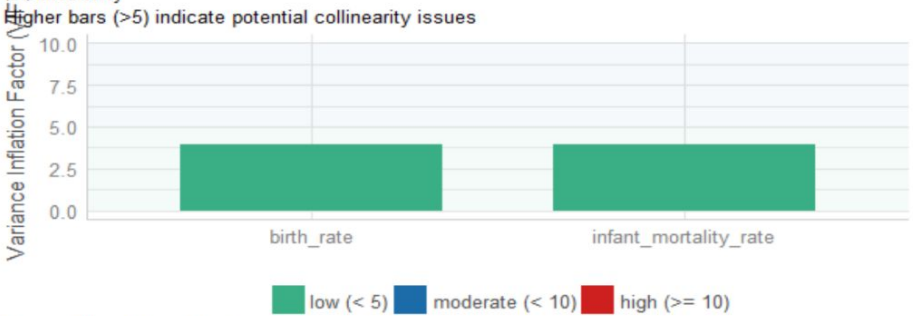
Homogeneity of Variance

Reference line should be flat and horizontal



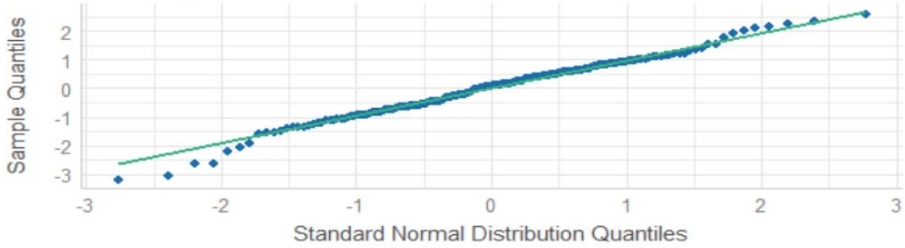
Collinearity

Higher bars (>5) indicate potential collinearity issues



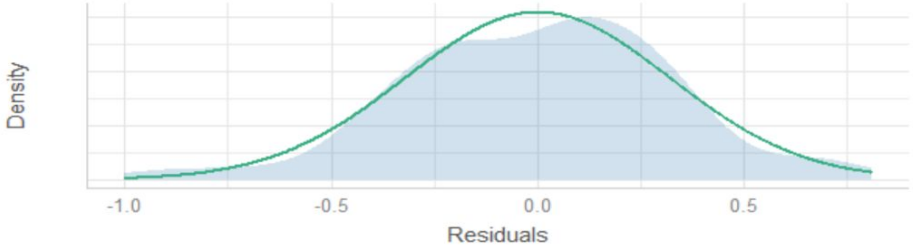
Normality of Residuals

Dots should fall along the line



Normality of Residuals

Distribution should be close to the normal curve




```
Shapiro-Wilk normality test

data:  resid(lmlogstep_no_col2)
W = 0.99015, p-value = 0.2621
```

Outliers and Leverage Points

- 11 outliers
- 15 leverage points
  - ~0.033
- Final Step: investigate outliers