Final\_project\_sds230

2024-07-25

# Introduction

…. We will introduce each subsection by describing the variables we are using in that section

## T Test - Claire

*We are investigating whether the global average life expectancy in 2016 significantly differs from 75 years using a one sample t-test. The dataset consists of the variables “Country” (the country name) and “LifeExp” (life expectancy in years) from the 2016 World Bank data.*

## [1] "Country" "LifeExp"

## 'data.frame': 217 obs. of 2 variables:  
## $ Country: chr "Afghanistan" "Albania" "Algeria" "American Samoa" ...  
## $ LifeExp: num 63.7 78.3 76.1 NA NA ...

## Country LifeExp   
## Length:199 Min. :51.84   
## Class :character 1st Qu.:66.67   
## Mode :character Median :74.31   
## Mean :72.20   
## 3rd Qu.:77.61   
## Max. :84.23

##   
## One Sample t-test  
##   
## data: wbLife$LifeExp  
## t = -5.0675, df = 198, p-value = 9.214e-07  
## alternative hypothesis: true mean is not equal to 75  
## 95 percent confidence interval:  
## 71.11137 73.29004  
## sample estimates:  
## mean of x   
## 72.20071

## Analysis - Claire

*The results from our one-sample t-test indicate a statistically significant difference from the hypothesized mean of 75 years for the global average life expectancy in 2016. The negative t-value shows that the sample mean is below the hypothesized mean. With an extremely low p-value, far below the alpha level of 0.05, we have strong evidence to reject the null hypothesis that the mean life expectancy is 75 years. This finding confirms that the mean life expectancy in 2016 is significantly lower than 75 years.*

## Permutation Test - Claire

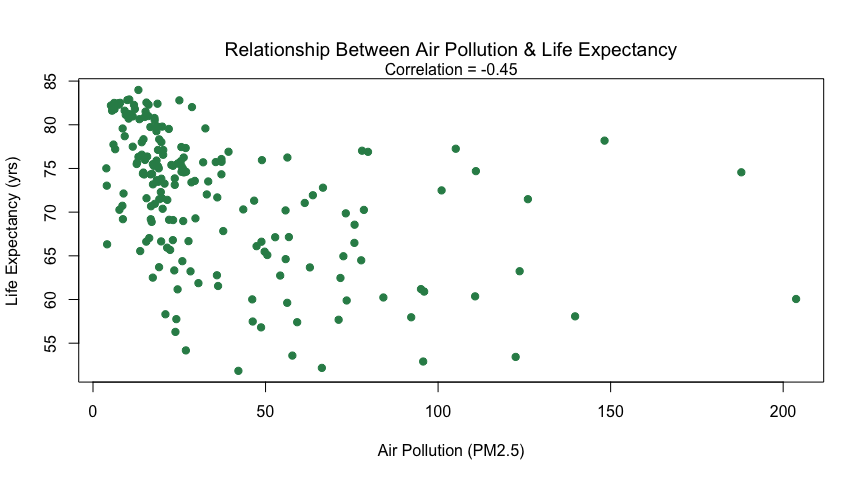
*For our permutation test, we want to see if air pollution significantly impacts life expectancy across different countries. Using the World Bank 2016 data variables “PM2.5” (mean annual exposure to air pollution) and “LifeExp”, our hypothesis is that higher levels of air pollution are associated with lower life expectancy.*

## [1] "Country" "PM2.5" "LifeExp"

## 'data.frame': 217 obs. of 3 variables:  
## $ Country: chr "Afghanistan" "Albania" "Algeria" "American Samoa" ...  
## $ PM2.5 : num 62.85 14.63 37.23 3.76 10.88 ...  
## $ LifeExp: num 63.7 78.3 76.1 NA NA ...

## 'data.frame': 188 obs. of 2 variables:  
## $ PM2.5 : num 62.9 14.6 37.2 36.2 15.7 ...  
## $ LifeExp: num 63.7 78.3 76.1 61.5 76.4 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:29] 4 5 10 28 37 40 51 56 65 69 ...  
## ..- attr(\*, "names")= chr [1:29] "4" "5" "10" "28" ...

## PM2.5 LifeExp   
## Min. : 3.857 Min. :51.84   
## 1st Qu.: 15.640 1st Qu.:66.58   
## Median : 23.155 Median :73.54   
## Mean : 35.758 Mean :71.78   
## 3rd Qu.: 46.867 3rd Qu.:77.22   
## Max. :203.744 Max. :83.98



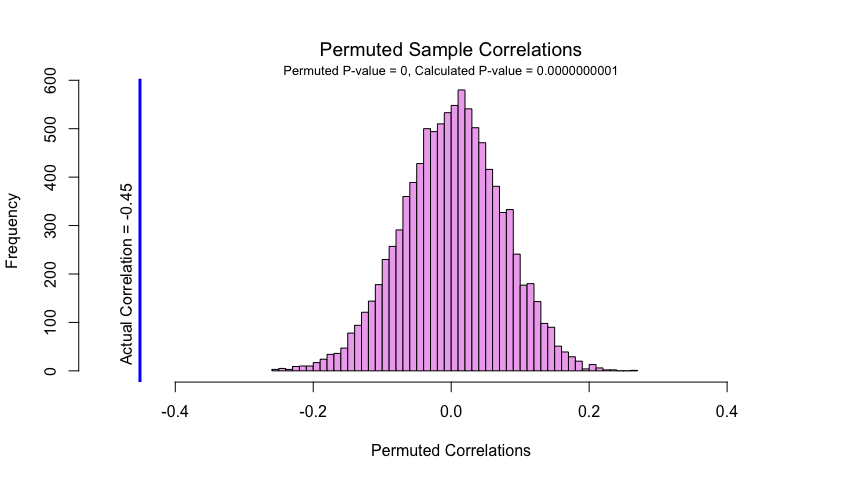
##   
## Pearson's product-moment correlation  
##   
## data: wbLife2$PM2.5 and wbLife2$LifeExp  
## t = -6.8954, df = 186, p-value = 8.122e-11  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.5582644 -0.3293592  
## sample estimates:  
## cor   
## -0.4512022

*The correlation is significant at alpha = .05 and .01, so there is evidence that there is statistically significant non-zero correlation between air pollution and life expectancy. So, let’s see what happens when we use fake data.*

*Now let’s create a LOT of fake data and run the permutation test. We’re going to get 10,000 fake correlations created on the assumption that there is no relationship between air pollution and life expectancy. Then, we’ll see how often we see a correlation close to our actual value just by chance.*

## [1] 0

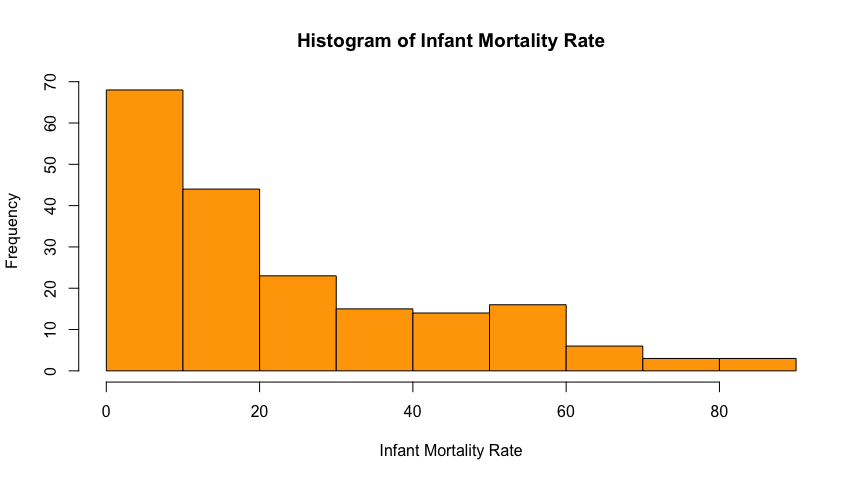
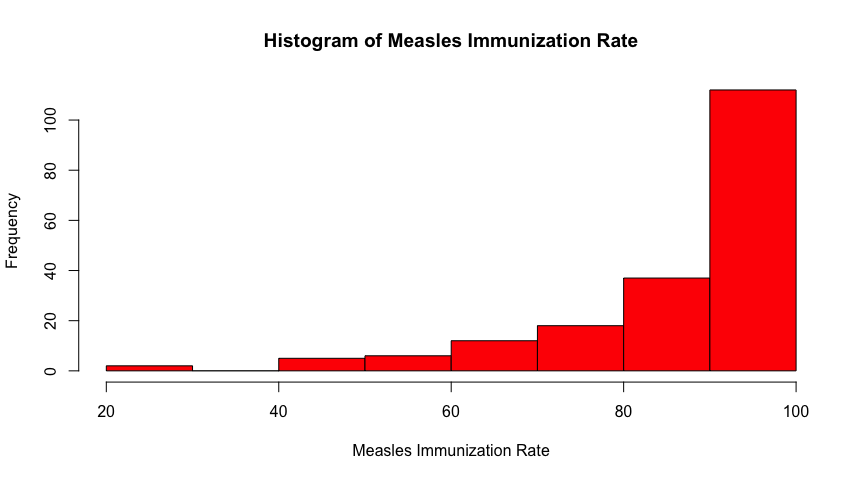
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.252915 -0.046399 0.003613 0.001935 0.051143 0.269793

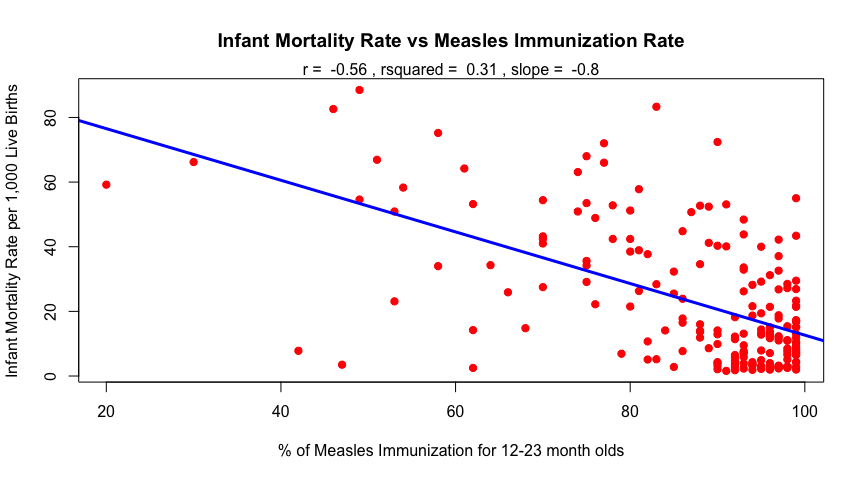


*The results from our permutation test reveal that the distribution of permuted correlations is normal and centered around zero. Since the actual observed correlation falls outside this distribution, this suggests that it is significantly different from what would be expected by random chance. This finding is further supported by a very low p-value, providing strong evidence that the observed correlation is statistically significant and suggesting a real association between air pollution and life expectancy.*

## Correlation

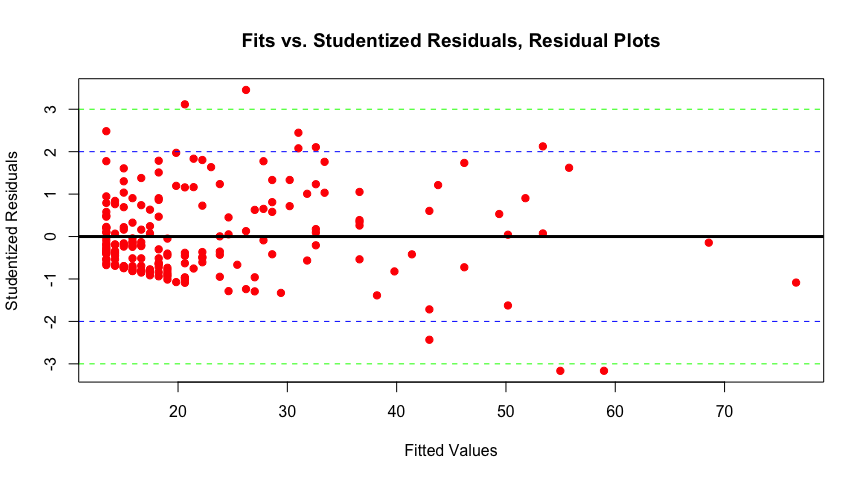
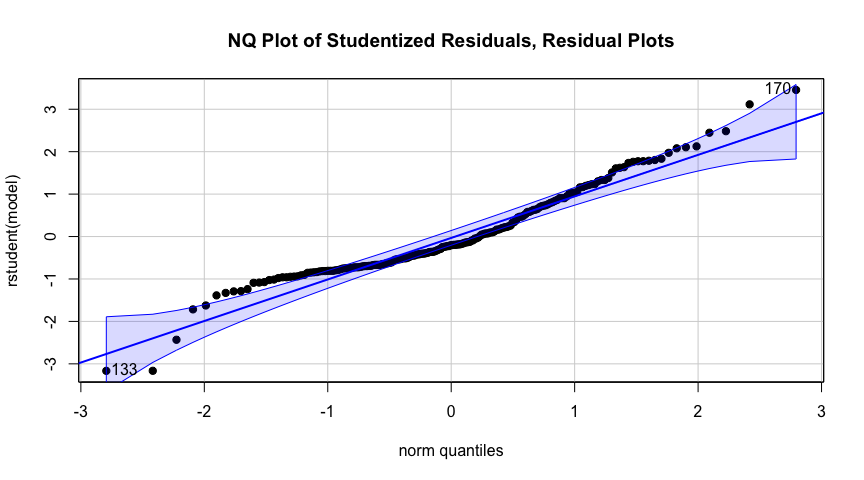
We are performing a simple linear regression to model the relationship between infant mortality rate (per 1,000 live births) and measles immunization rate for infants 12 - 23 months old. The assumptions are random, normally distributed errors centered at zero with constant variance (homoskedasticity) and linearity between variables. Histograms show that Measles vaccination rates are heavily left-skewed, and infant mortality is heavily right-skewed.

 We fit an initial linear model to these variables and calculate the correlation and R-squared value.



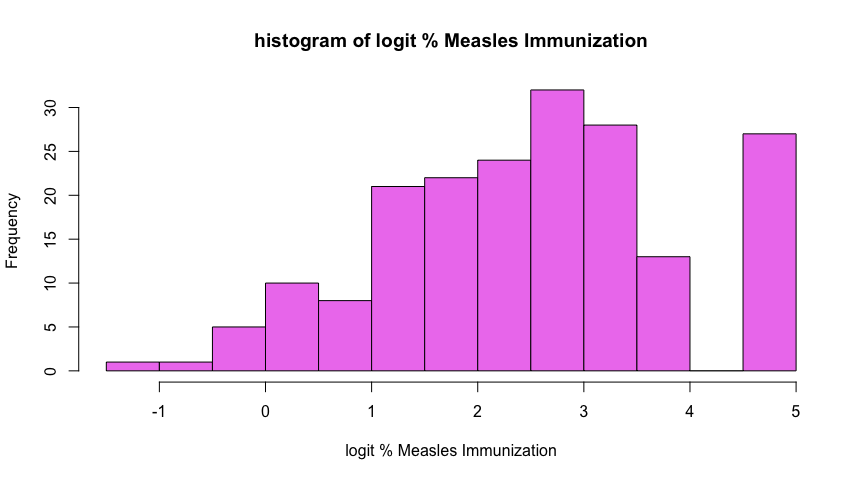
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 82.00 93.00 87.21 97.00 99.00

There is a moderate to strong negative relationship between % Measles vaccinations and infant mortality. An R-squared of 0.31 means 31% of the variability in infant mortality rate is explained by the model. However, since measles immunization rates cluster near 100%, this causes issues with spread and variance. Residual plots can help investigate this further.



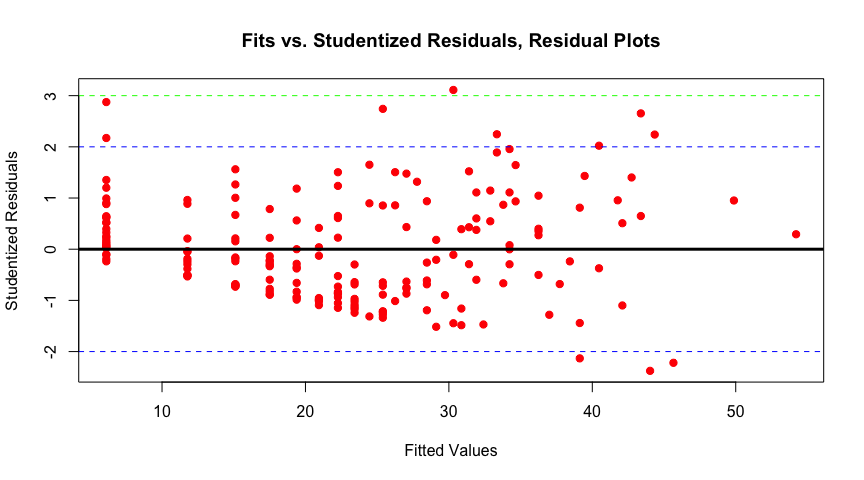
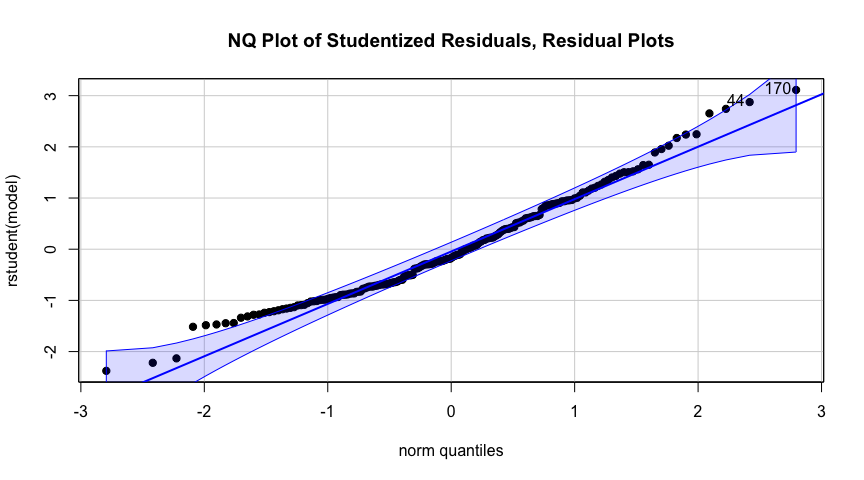
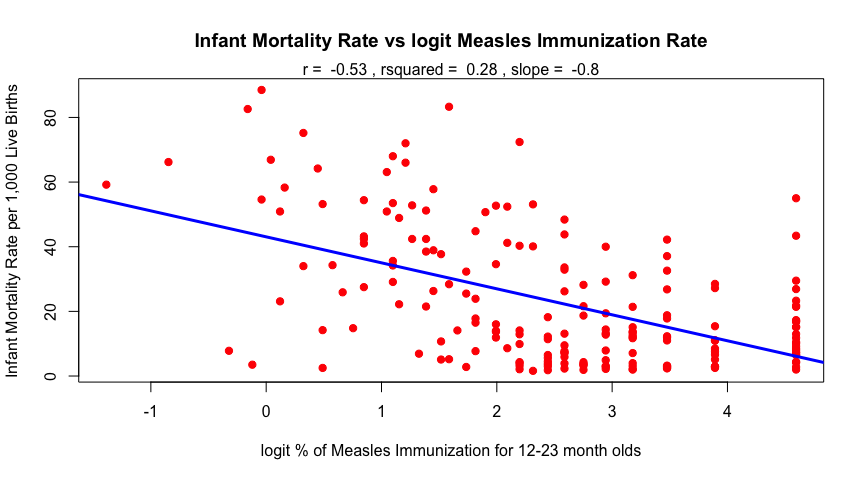
There looks to be some hetereoskedasticity in the fit vs studentized residuals, likely due to the extreme right skew of measles vacciination. This makes sense, as the median measles % vaccination is 93 and the mean is 87.21; most of the data is centered on the right. Since measles vaccination is a percentage, we can perform logit transformation and see if this improves the fit.

## Note: largest value of p > 1 so values of p interpreted as percents



## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.386 1.516 2.587 2.513 3.476 4.595

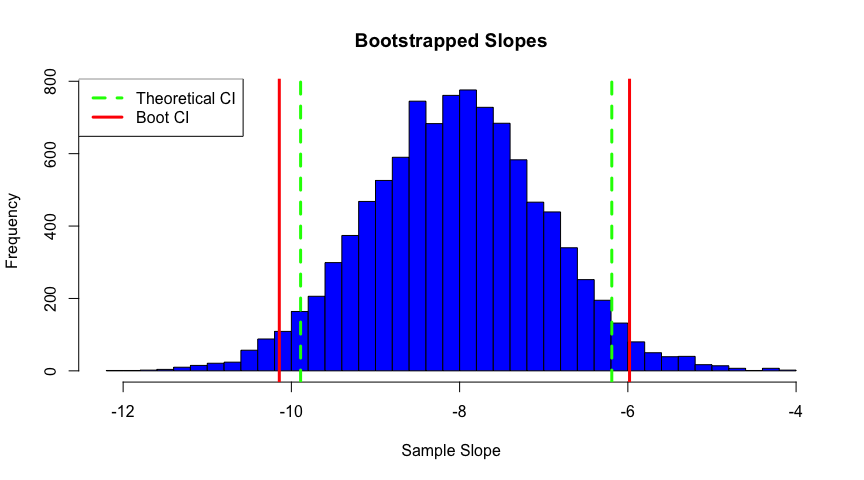
After taking logit of % Measles vaccination, we can see in the histogram is more evenly distributed. There is still a left skew, but this is to be expected since countries still have high measles vaccination rates. With our transformed predictor, we make another scatterplot and fit a new regression model.



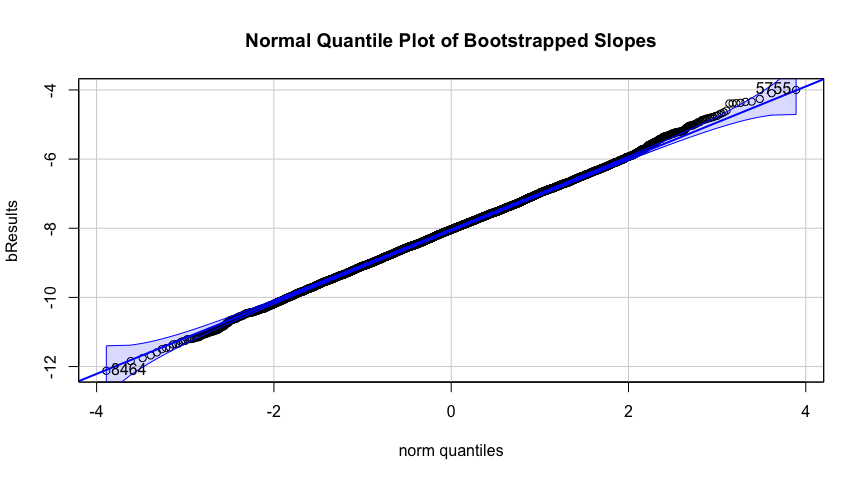
In the new scatter plot with logit measles, the data is more dispersed along the x axis. The spread and linear assumptions of correlation and linear models are better met with the transformed variable. The measles histogram is less skewed, the residual plot has less unequal variance, and the overall model fit is better with the transformed predictor variable. Since R-squared is .28, 28% of the variability in infant mortality can be explained by this model. While this is lower than the squared value of the previous model (R-squared = .31), the current model with logit measles vaccination is a better fit since the underlying assumptions are better met. The best model fit isn’t necessarily the model with the highest R-squared.

### Bootstrap CI for Correlation

In order to check the slope we calculated using parametric tests, we employ non parametric bootstrapping in order to calculate confidence intervals for the slope between logit Measles immunization rate and infant mortality rate.



The histograms above show the bootstrapped slopes. The bootstrapped confidence intervals are only a bit wider than the theoretical. This could mean that the linear model we fit approximated the assumptions of normality, homoskedasticity, and independence of errors. The bootstrapping is non parametric therefore capturing more of the true variability in the data. We can look at normal quantile plots of the bootstrapped data to further visualize the bootstrapped distribution.



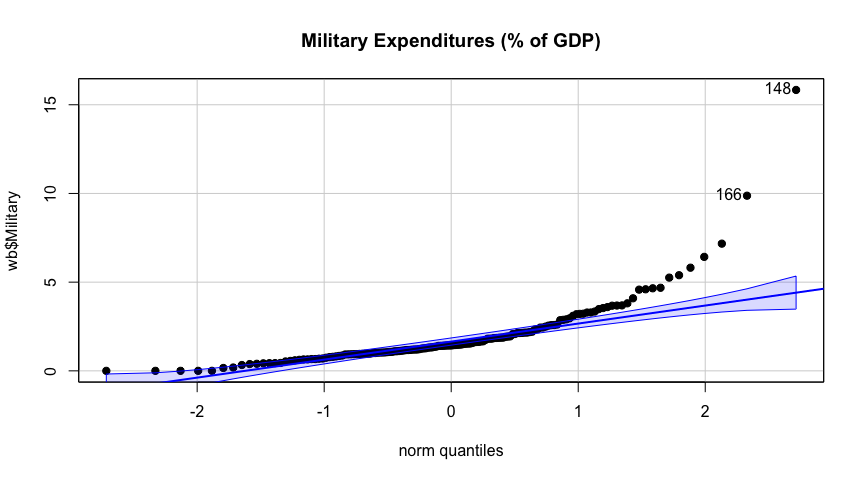
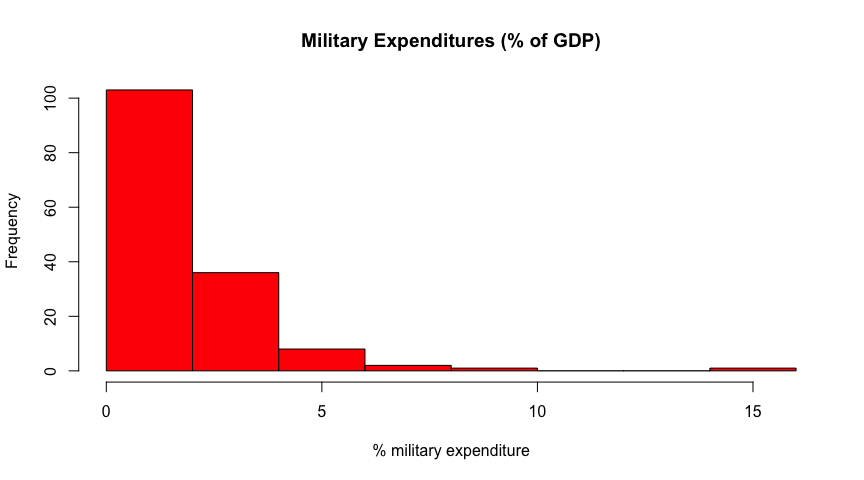
## [1] 8464 5755

As expected, distributions for slope approximate normality. There is a slight right skew in the normal quantile plot of bootstrapped correlation which is also reflected in the histogram.The histogram for slope looks very near normal, and the data falls almost entirely along the straight line in the normal quantile plot.

## Multiple regression

*introduction, data explanation, variables, etc*

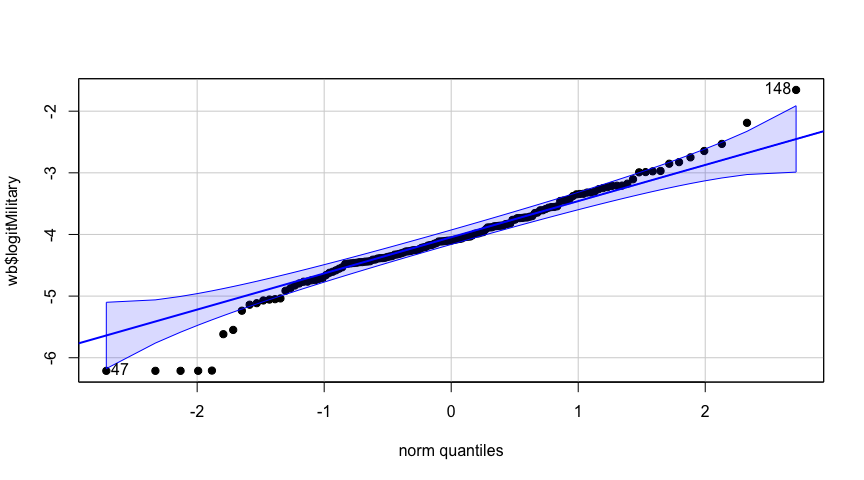
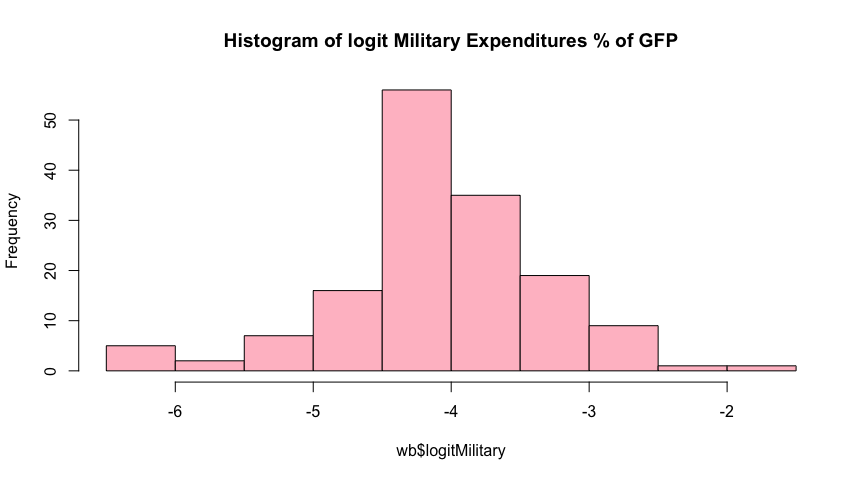
### look at response variable Military Expenditures



## [1] 148 166

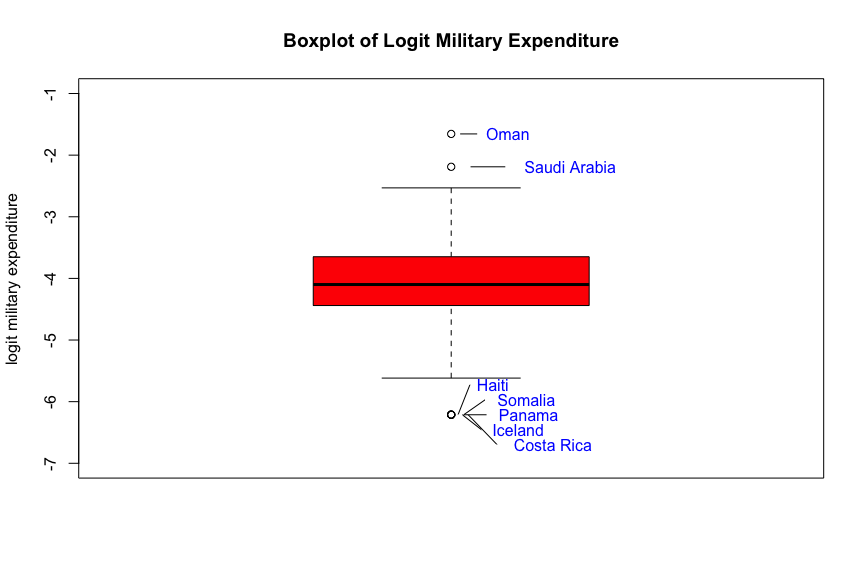
The data is heavily right-skewed and not normally distributed. These plots suggest using a logit transformation, which helps with probabilities or percentages. Due to zeros in the data, we add a small amount to each value to avoid issues with the logit function.

## Note: largest value of p > 1 so values of p interpreted as percents

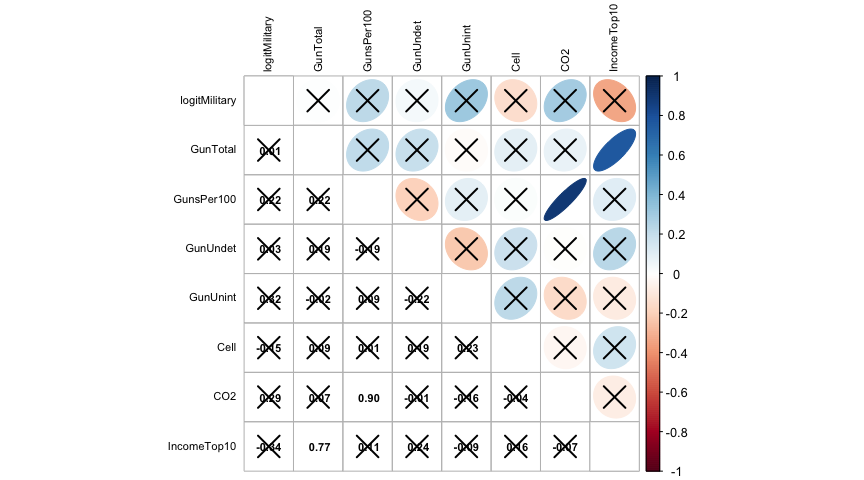


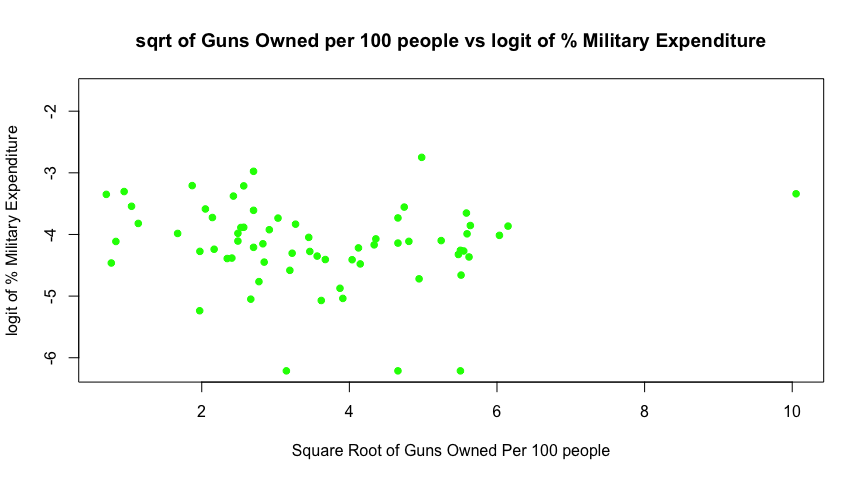
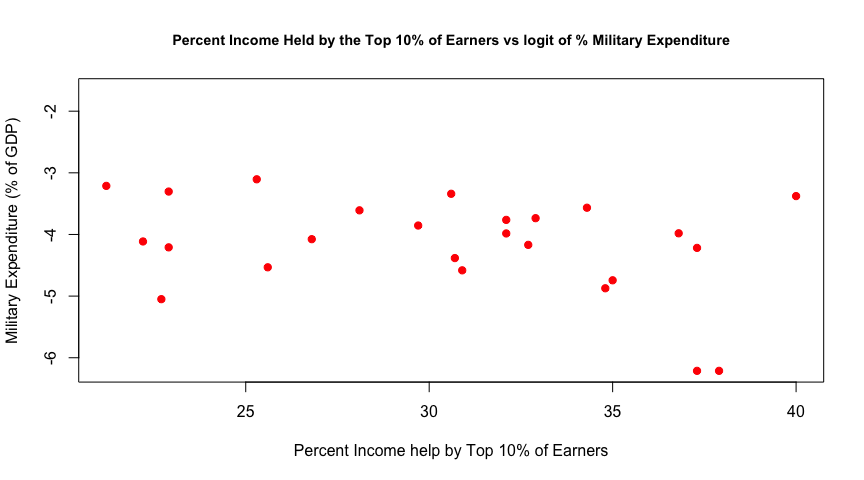
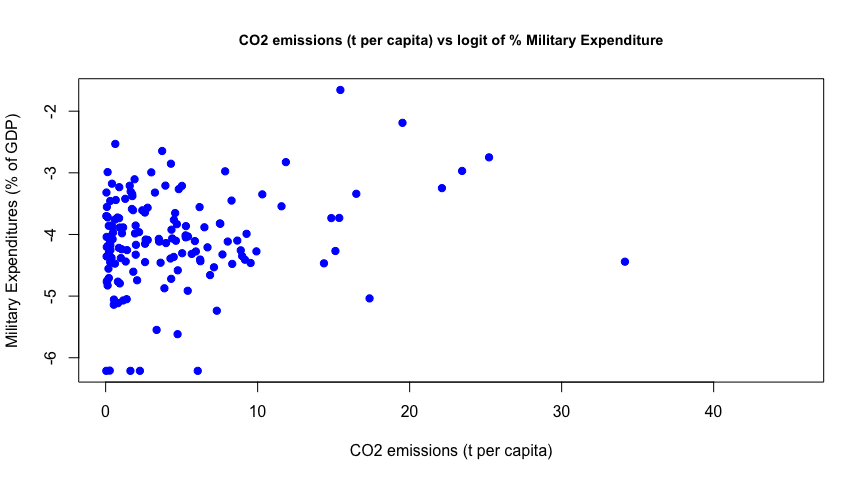
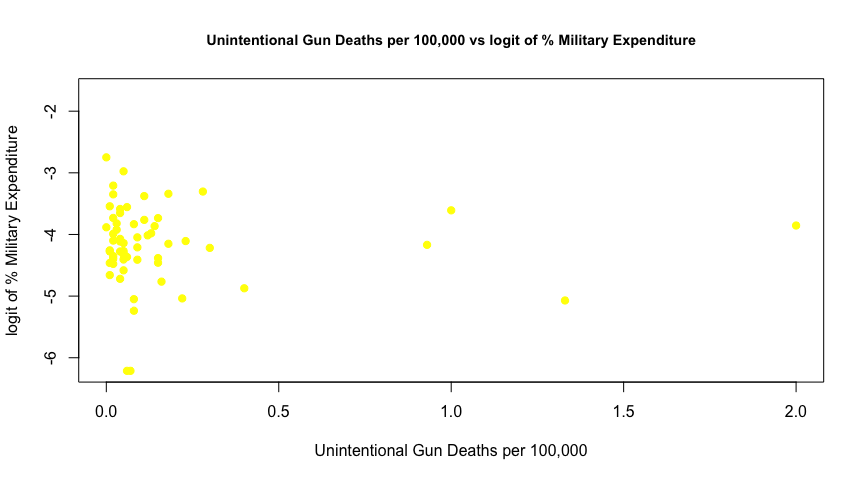
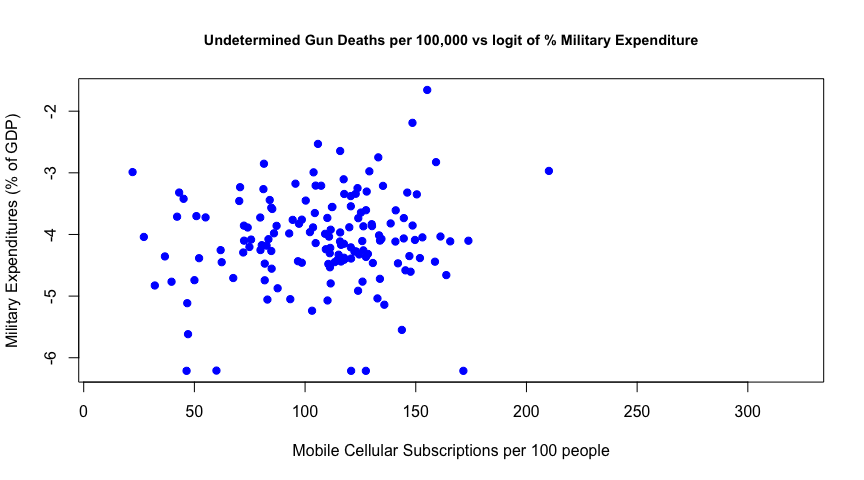
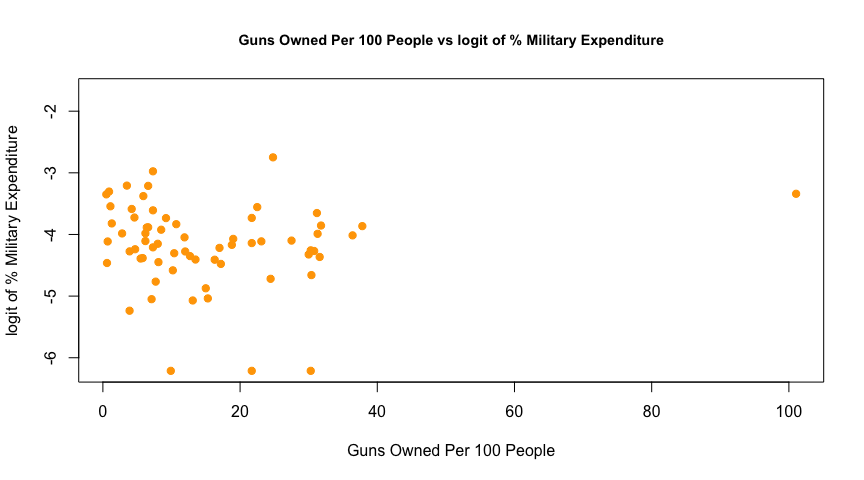
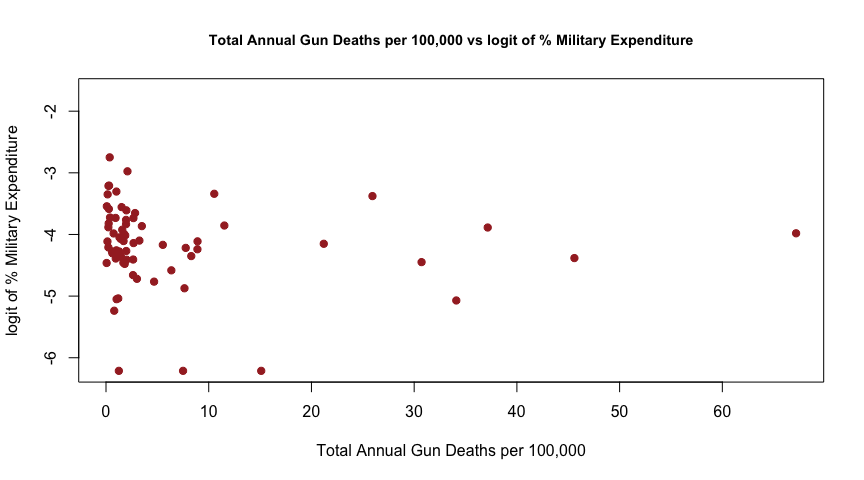
## [1] 148 47

Now the data is more normally distributed. There are a few potential outliers, spending more or less than expected on military. The box plot below shows these countries.

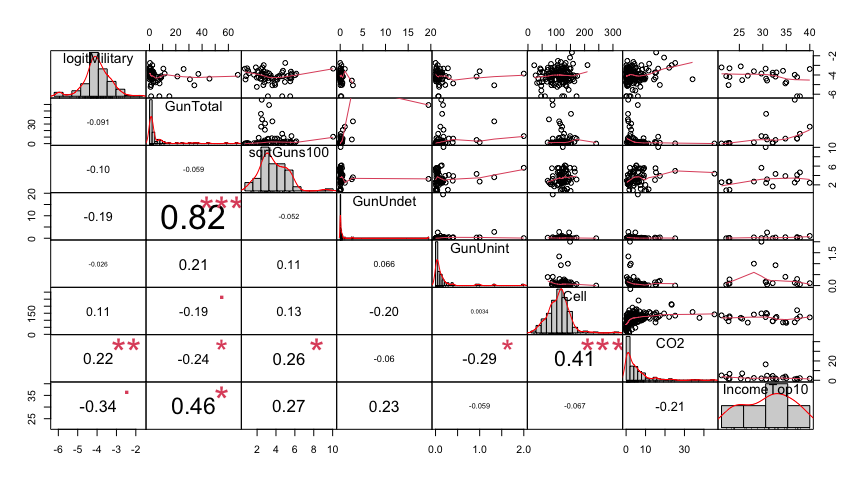


In this box plot, Haiti, Somalia, Panama, Iceland, and Costa Rica all have approximately 0% military expenditure. Oman and Saudi Arabia have relatively higher military expenditure compared to other countries in the 2016 World Bank dataset.

Now that our response variable, miltary expenditure, is transformed we begin to look at the relationships with this transformed variable and potential explanatory variables. First, we make correlation plot of all the possible predictors we want to include in our model. Here are the possible predictors we are including: total annual gun deaths per 100,000, the number of guns owned per 100 people, undetermined gun deaths per 100,000, Unintentional Gun Deaths per 100,000,percent income held by the top 10% of Earners, Mobile Cellular Subscriptions per 100 people, and CO2 emmisions in metric tons (t) per capita.  Through the correlation matrix, it appears that none of the predictor variable are signifant enough to have a clear relationship with logit of the military expenditure. However, it appears that some of the predictor variables including Guns per 100 people , Unintential Gun Deaths per 100,000, and CO2 emmissions have a somewhat positive correlation with Military Expenditure on the logit scale while Mobile Cell Subscriptions per 100 people, and Percent income held by the Top 10% of Earners have a somewhat negative linear relationship with Military Expenditure on the logit Scale.

Next I create, individual scatterplots between each of the predictor variables and response variable. It appears  Through these scatterplots of each predictor variable I observed that if the square root of Guns per 100 people is taken then the spread of data on the scatterplot vs Military Expenditure increases and their relationship can be better interpreted. I will now incorporate square root of Guns per 100 into the model instead of the raw version of the variable.

## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter  
## Warning in par(usr): argument 1 does not name a graphical parameter



Our relationships between the logit of Military Expenditure and each of the variables included appears to be either linear or in a blob shape. It appears that among our predictor variables, that there are some significant correlations between predictor variables including Total Gun Deaths per 100,000 vs Undetermined Gun Deaths per 100,000 people and Mobile Cellular Subscriptions vs Carbon Emissions in metric tons per capita. This will result in collinearity within our regression model. The most significant relationships between our response variable and predictor variables were the logit of Military Expenditure & CO2 Emssions per Capita and logit of Military Expenditure & Income in the top 10% of individuals. It appears that the correlation between logit of Military Expenditure and Income in the top 10% of individuals was negative and this was significant at a pvalue of 0.05 or less from a t distribution.

Now I am going to fit a regression model including all possible predictors for logit of Military Expenditure

##   
## Call:  
## lm(formula = logitMilitary ~ GunTotal + sqrtGuns100 + GunUndet +   
## GunUnint + Cell + CO2 + IncomeTop10, data = wbn2)  
##   
## Residuals:  
## 8 43 47 72 107 128 130 151 153 154   
## -0.2799 0.4145 -0.3206 0.2117 0.7837 -0.9181 -0.2377 -0.7204 1.2120 0.2381   
## 160 207 208   
## -0.3843 0.3199 -0.3191   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.53197 2.56026 -0.989 0.368  
## GunTotal 0.05329 0.06264 0.851 0.434  
## sqrtGuns100 -0.46129 0.32581 -1.416 0.216  
## GunUndet 0.13901 0.44130 0.315 0.765  
## GunUnint 1.43577 0.72997 1.967 0.106  
## Cell -0.01252 0.01295 -0.967 0.378  
## CO2 0.29421 0.17609 1.671 0.156  
## IncomeTop10 -0.02075 0.09678 -0.214 0.839  
##   
## Residual standard error: 0.9282 on 5 degrees of freedom  
## (204 observations deleted due to missingness)  
## Multiple R-squared: 0.632, Adjusted R-squared: 0.1168   
## F-statistic: 1.227 on 7 and 5 DF, p-value: 0.4253

From the summary of our model, it appears that none of the predictors are statistically significant and although there are 217 observations of logit Military expenditure in our original dataset, the total number of observations deleted among all predictors due to missingness was 204. I decided to perform backwards stepwise regression on the model to see if any of our predictors would become significant.

##   
## Call:  
## lm(formula = logitMilitary ~ GunTotal + sqrtGuns100 + GunUndet +   
## GunUnint + Cell + CO2, data = wbn2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5829 -0.2798 0.0245 0.3340 1.3713   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.919747 0.694783 -5.642 0.00000106 \*\*\*  
## GunTotal -0.013964 0.019248 -0.725 0.4719   
## sqrtGuns100 -0.045188 0.053881 -0.839 0.4061   
## GunUndet -0.092923 0.222588 -0.417 0.6783   
## GunUnint 0.382933 0.278843 1.373 0.1765   
## Cell -0.003265 0.005132 -0.636 0.5279   
## CO2 0.051395 0.019904 2.582 0.0131 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.619 on 45 degrees of freedom  
## (165 observations deleted due to missingness)  
## Multiple R-squared: 0.1914, Adjusted R-squared: 0.08362   
## F-statistic: 1.776 on 6 and 45 DF, p-value: 0.1258

##   
## Call:  
## lm(formula = logitMilitary ~ CO2, data = wbn2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.18423 -0.35924 0.01353 0.46643 2.09648   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.20675 0.07707 -54.584 < 0.0000000000000002 \*\*\*  
## CO2 0.02944 0.01086 2.712 0.00749 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.73 on 148 degrees of freedom  
## (67 observations deleted due to missingness)  
## Multiple R-squared: 0.04733, Adjusted R-squared: 0.04089   
## F-statistic: 7.353 on 1 and 148 DF, p-value: 0.007488

Through backwards regression of the model, one predictor variable was left to predict the logit of Military expenditures which was CO2 emissions. This variable is highly significant at a pvalue on the t distribtution of 0.001\*\*. This model however does not have a very high R squared value which is only 0.04 meaning that only around 4 percent of the variability in Military Expenditure on the logit scale can be explained by CO2 emissions. It appears that there is a positive correlation however and as carbon emission increase, there is expected to be a higher military expenditure in a country.

Now instead of using backwards stepwise regression, I will use best subsets regression on all of our original variables to see which model among all possible models could predict logit of military expenditure the best.

## [1] 217 8

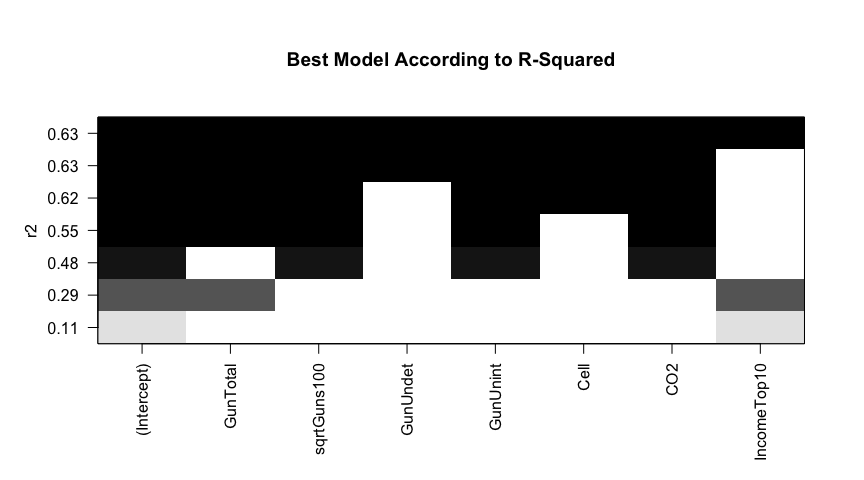
## [1] "np" "nrbar" "d" "rbar" "thetab" "first"   
## [7] "last" "vorder" "tol" "rss" "bound" "nvmax"   
## [13] "ress" "ir" "nbest" "lopt" "il" "ier"   
## [19] "xnames" "method" "force.in" "force.out" "sserr" "intercept"  
## [25] "lindep" "nullrss" "nn" "call"

## (Intercept) GunTotal sqrtGuns100 GunUndet GunUnint Cell CO2 IncomeTop10  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## 2 TRUE TRUE FALSE FALSE FALSE FALSE FALSE TRUE  
## 3 TRUE FALSE TRUE FALSE TRUE FALSE TRUE FALSE  
## 4 TRUE TRUE TRUE FALSE TRUE FALSE TRUE FALSE  
## 5 TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE  
## 6 TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  
## 7 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

##   
## Call:  
## lm(formula = logitMilitary ~ IncomeTop10, data = wbn2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7131 -0.3947 0.2622 0.4940 1.2539   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.67925 0.87863 -3.049 0.00569 \*\*  
## IncomeTop10 -0.04880 0.02834 -1.722 0.09853 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7758 on 23 degrees of freedom  
## (192 observations deleted due to missingness)  
## Multiple R-squared: 0.1142, Adjusted R-squared: 0.07566   
## F-statistic: 2.965 on 1 and 23 DF, p-value: 0.09853

It appears that after using best subsets regression on logitMilitary expenditure based on the predictor variables chosen (Total Gun Deaths, Square Root of Number of Gun Deaths, Undeterimned Gun Deaths, Unintentional Gun Deaths, Mobile Cellular Subscriptions per 100 people, Carbon emissions in metric tons per capita, and percent income held by the top 10% of Earners ) that of the models with a single predictor, the single best predictor for Military Expenditures was Percent Income Held by the Top 10% of Earners. It turns that that the best model with 3 predictors contains CO2 emissions, Unintentional Gun Deaths and the square root of the number of Guns per 100,000.

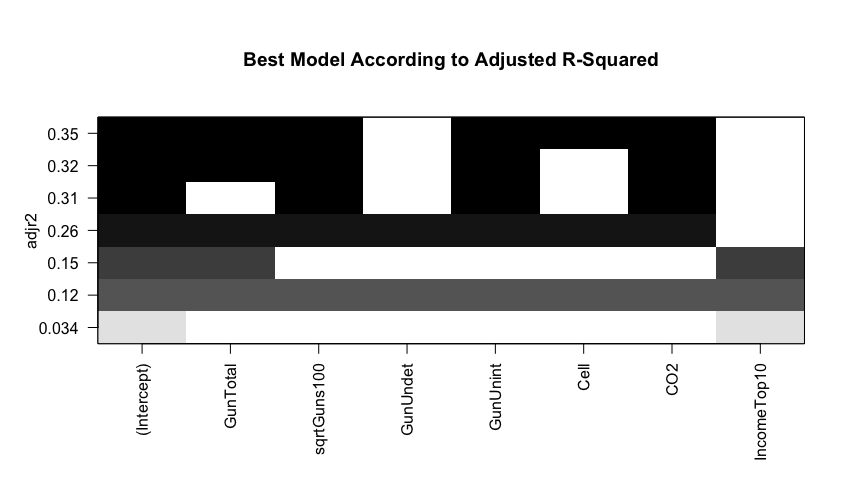
For every additional predictors added, an additional dimensional will also be added to the regression model,and r squared will increase as a result. Since this is the case, I want to find the model that will penalize additional predictor variables and this is done by determining the model according to adjusted r squared.



## [1] 5

## [1] "GunTotal" "sqrtGuns100" "GunUnint" "Cell" "CO2"

##   
## Call:  
## lm(formula = logitMilitary ~ ., data = wbtemp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.73697 -0.24833 0.06399 0.36030 1.11129   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.784095 0.605064 -6.254 0.0000000763 \*\*\*  
## GunTotal -0.002161 0.007743 -0.279 0.781   
## sqrtGuns100 -0.050244 0.052187 -0.963 0.340   
## GunUnint 0.212278 0.265468 0.800 0.428   
## Cell -0.004288 0.004606 -0.931 0.356   
## CO2 0.048351 0.019978 2.420 0.019 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.634 on 52 degrees of freedom  
## (159 observations deleted due to missingness)  
## Multiple R-squared: 0.1224, Adjusted R-squared: 0.03799   
## F-statistic: 1.45 on 5 and 52 DF, p-value: 0.2222

 According to the model with the highest R-Squared all 7 predictors of Logit Military Expenditure % should be included, however, this is only because R squared increases as the dimensionality increases in our model and thus it is likely not the case that all 7 predictors explain 0.35 of the variability in logit of Military Expenditue as % of a counrty’s GDP.

Adjusted R squared penalizes extra predictor variables in our model which weren’t accounted for in our model where highest r squared value was taken. In this case, Total Gun Deaths per 100,000, the square root of the number of guns per 100 people, unintentional gun deaths per 100,000 people, mobile cellular subscriptions per 100 people, and carbon emissions in metric tons result in the model with the highest adjusted r squared value. In addition Carbon emissions in metric tons is a singificant statistic although 159 observations of the original 217 have been deleted and the model only has a multiple r squared value of 0.1224. Even with these predictors having a large adjusted r squared value not alot of the variability in Military Expenditure can be explained by the predictors.

Now, I will look at best model according to Bayesian Information Criteria

## [1] 3

## [1] "sqrtGuns100" "GunUnint" "CO2"

##   
## Call:  
## lm(formula = logitMilitary ~ ., data = wbtemp2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.82782 -0.33888 0.06642 0.32864 1.09237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.32107 0.21005 -20.571 <0.0000000000000002 \*\*\*  
## sqrtGuns100 -0.04891 0.05161 -0.948 0.3475   
## GunUnint 0.17203 0.25743 0.668 0.5068   
## CO2 0.04790 0.01892 2.532 0.0143 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6275 on 54 degrees of freedom  
## (159 observations deleted due to missingness)  
## Multiple R-squared: 0.1071, Adjusted R-squared: 0.05751   
## F-statistic: 2.159 on 3 and 54 DF, p-value: 0.1035

According to our Bayesian Information Criterion for our models for logit of military expenditure, the model with the lowest Bayesian Information criterion had 3 predictor variables which were the square root of Guns per 100 people, Unintentional Gun deaths per 100,000 people and carbon emissions per person in metric tons. This models criterion metric was set by measuring the likelihood of the model occuring, the number of parameters in the mdoel and the number of observations. It appears that our r-squared value is 0.10 which is around the rsquared value from our other models including the adjusted r squared model and the significance for Carbon Emissions in this model is still high (at a low pvalue). As carbon emissions increases in the model, we can expect the logit of military expenditure in % GDP of a country to increase.

Next I will look at the best model according to AIC

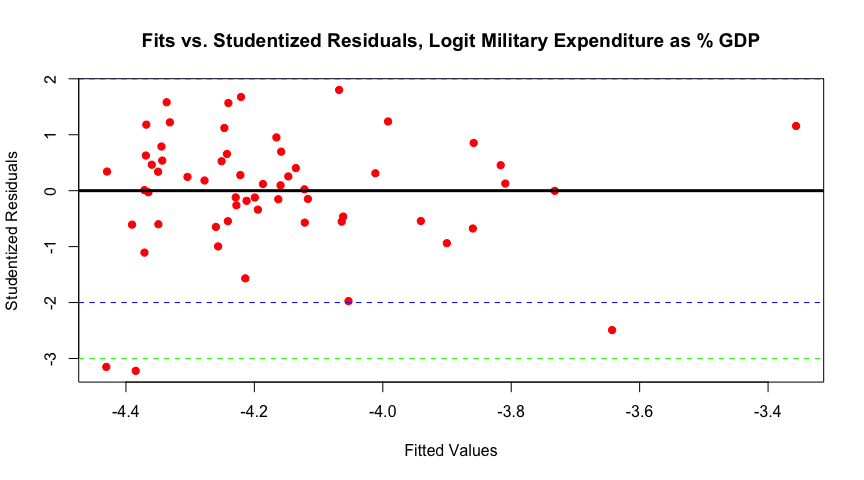
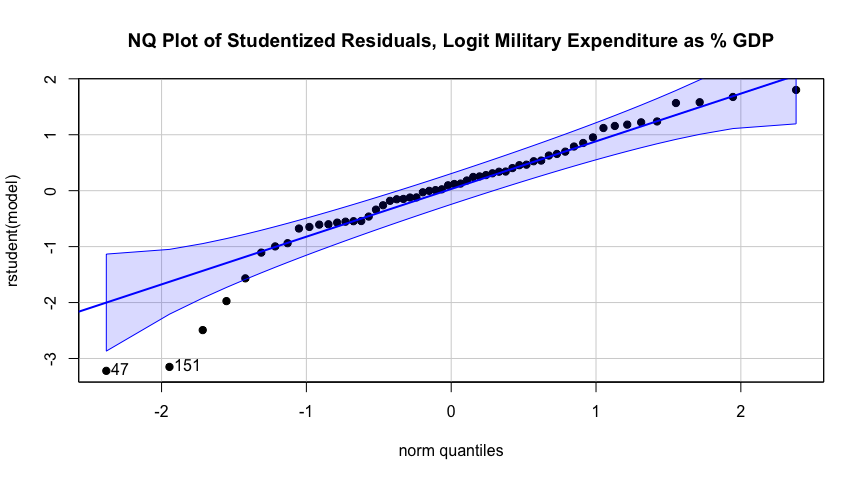
## [1] 62.16857 52.01907 116.39598 118.35477 119.39593 106.17296 40.53397

##   
## Call:  
## lm(formula = logitMilitary ~ ., data = wbtemp3)  
##   
## Residuals:  
## 8 43 47 72 107 128 130 151 153 154   
## -0.2799 0.4145 -0.3206 0.2117 0.7837 -0.9181 -0.2377 -0.7204 1.2120 0.2381   
## 160 207 208   
## -0.3843 0.3199 -0.3191   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.53197 2.56026 -0.989 0.368  
## GunTotal 0.05329 0.06264 0.851 0.434  
## sqrtGuns100 -0.46129 0.32581 -1.416 0.216  
## GunUndet 0.13901 0.44130 0.315 0.765  
## GunUnint 1.43577 0.72997 1.967 0.106  
## Cell -0.01252 0.01295 -0.967 0.378  
## CO2 0.29421 0.17609 1.671 0.156  
## IncomeTop10 -0.02075 0.09678 -0.214 0.839  
##   
## Residual standard error: 0.9282 on 5 degrees of freedom  
## (204 observations deleted due to missingness)  
## Multiple R-squared: 0.632, Adjusted R-squared: 0.1168   
## F-statistic: 1.227 on 7 and 5 DF, p-value: 0.4253

From the results of all of the best subsets regression functions including AIC, BIC , Adjusted R squared and Backwards Stepwise regression, I decided to include the model from the Bayesian Information Criterion since it’s most significant predictor variable had the lowest p-value for highest adjusted r squared value (0.05) with the most predictor variables kept in the model (3). This adjusted r squared value is still relatively low but it is the best of all models.

Now I will examine the model residuals based on this BIC model with 3 predictors

##   
## Call:  
## lm(formula = logitMilitary ~ ., data = wbtemp2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.82782 -0.33888 0.06642 0.32864 1.09237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.32107 0.21005 -20.571 <0.0000000000000002 \*\*\*  
## sqrtGuns100 -0.04891 0.05161 -0.948 0.3475   
## GunUnint 0.17203 0.25743 0.668 0.5068   
## CO2 0.04790 0.01892 2.532 0.0143 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6275 on 54 degrees of freedom  
## (159 observations deleted due to missingness)  
## Multiple R-squared: 0.1071, Adjusted R-squared: 0.05751   
## F-statistic: 2.159 on 3 and 54 DF, p-value: 0.1035

 According to our residuals it appears that for the most part our errors for logit of military expenditures are approximately normally distributed with some exceptions on the lower end of the normal quantile plot. We could infer that these countries had very low military expenditure as part of their overall GDP. In addition, there are 2 outliers that have a standard deviation of over 3 although the rest of the studentized residuals appear evenly scattered with no evidence of heteroskedasticity in the model.

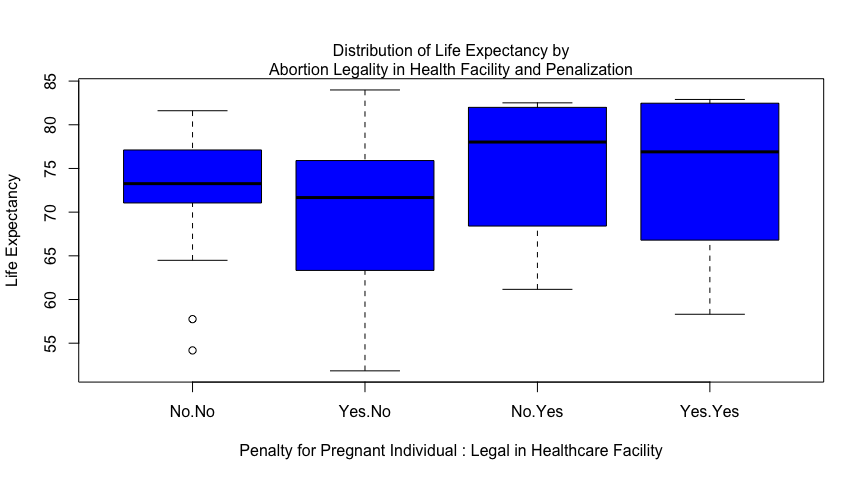
Results and Discussion:

After performing various types of best subsets regression including exhaustive search and forward selection, the criterion I decided to use for our model was Bayesian Information Criterion which included the predictor variables the square root of Guns per 100 people, the number of unintentional gun deaths per 100,000 people, and carbon emissions per person in metric tons which were used to predict military expenditure on the log scale. I chose this criterion in particular because when compared to the other criterion including backwards step wise regression, the adjusted r squared value and the exhaustive search method for best subsets regression, it had the largest r squared (0.1071) and adjusted r squared (0.05731) value with the most number of significant variables (although there was only 1). The most significant predictor variable was Carbon. emissions per person in metric tons which was significant at a pvalue of 0.0143. In the model, as carbon emissions per person of a country increases we can expect that the logit of Military expenditure to increase as well by 0.0479.

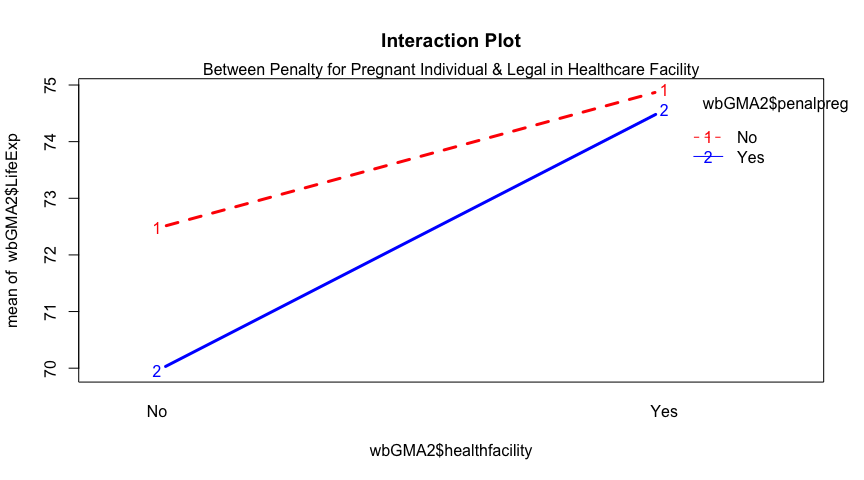
## 2-Way ANOVA

### Introduction

We are interested in predicting life expectancy based on some categorical variables a data set concerning global abortion laws concerning self-managed abortion. More information can be found [here](https://legacy.lawatlas.org/datasets/global-medication-abortion-laws). We want to understand how allowing abortions in abortion designated health facilities (‘healthfacility’) and penalizing individuals seeking abortions (‘penalpreg’) affects life expectancy. We clean the Global Medication Data (GMA) by converting non 1 and 0 values to NA and recoding 1 as Yes and 0 as No. Then, we join the GMA and World bank dataset by country, and remove rows with NAs. The resulting dataframe has 143 unique countries. We begin the analysis using boxplots to examine the life expectancy distributions for each level of ‘healthfacility’ and ‘penalpreg’. The variance and life expectancy differences across groups are minimal.



Now we want to check interactions between the categorical variables and continuous variable. The lines are not parallel, suggesting a potential interaction effect between penalizing a pregnant woman for abortion (‘penalpreg’) and allowing abortions in designated health facilities (‘healthfacility’). For penalized women, countries permitting abortions in health facilities have a slightly higher mean life expectancy. Conversely, for non-penalized women, countries not permitting abortions in health facilities have a higher mean life expectancy. However, this plot is not a statistical test.



It is also important to note that we have a small sample size and an unbalanced design as seen in the table below, the number of observations in each group is not the same.

##   
## No Yes  
## No 25 15  
## Yes 86 17

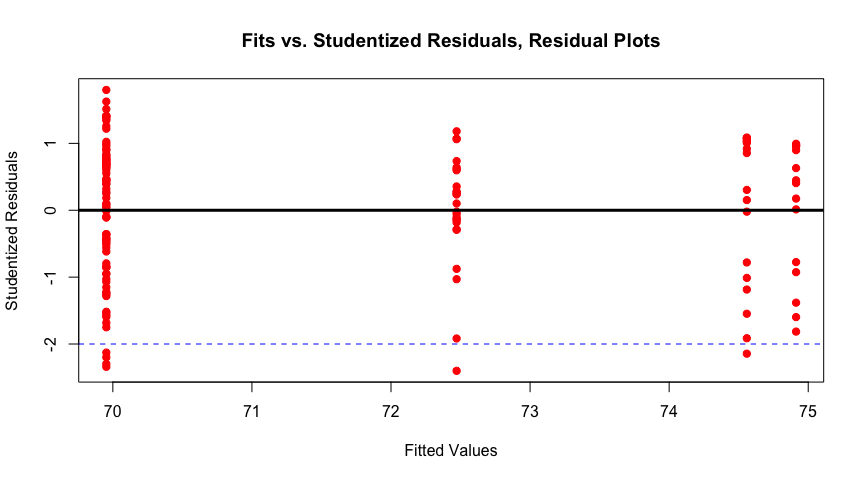
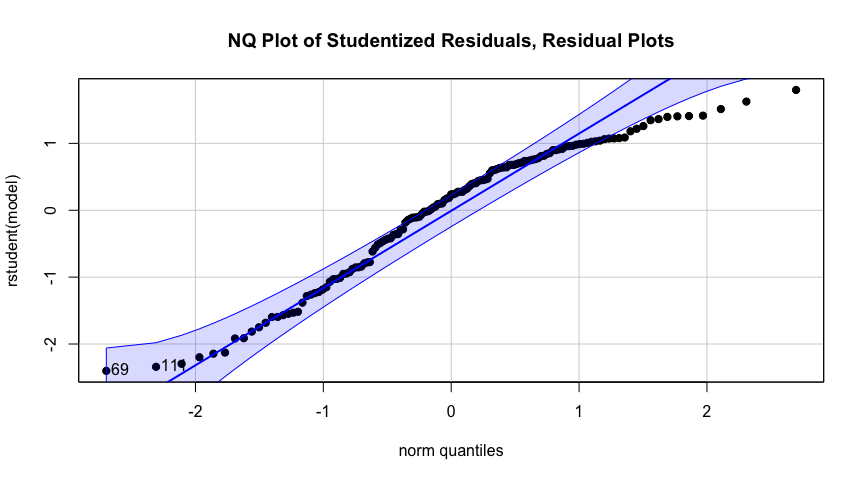
We ran a two-way ANOVA with ‘penalpreg,’ ‘healthfacility,’ and their interaction. None were significant. A second additive ANOVA without the interaction showed a significant main effect of ‘healthfacility’ only. Abortions in government health facilities significantly affect life expectancy. Running a linear model gave a significant F statistic P value of 0.015, lower than the model with the interaction (P = 0.03), indicating a better fit. However, the adjusted R-squared is 0.04, explaining only 4% of the variance in life expectancy. The model may not explain the data well due to the small sample size and only one significant effect, in addition to it being an unbalanced ANOVA.

## Anova Table (Type III tests)  
##   
## Response: wbGMA2$LifeExp  
## Sum Sq Df F value Pr(>F)  
## (Intercept) 131304 1 2098.8459 <0.0000000000000002  
## wbGMA2$penalpreg 123 1 1.9642 0.1633  
## wbGMA2$healthfacility 56 1 0.8929 0.3463  
## wbGMA2$penalpreg:wbGMA2$healthfacility 26 1 0.4228 0.5166  
## Residuals 8696 139   
##   
## (Intercept) \*\*\*  
## wbGMA2$penalpreg   
## wbGMA2$healthfacility   
## wbGMA2$penalpreg:wbGMA2$healthfacility   
## Residuals   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type III tests)  
##   
## Response: wbGMA2$LifeExp  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 167328 1 2685.7414 < 0.0000000000000002 \*\*\*  
## wbGMA2$penalpreg 97 1 1.5638 0.21319   
## wbGMA2$healthfacility 330 1 5.3042 0.02275 \*   
## Residuals 8722 140   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

In the graphs below, we see that the residuals are approximately normally distributed with no major violations of equal variances. However, some non-conforming data in the right tail likely result from life expectancy being left-skewed, with a maximum of about 84 years. In a normal distribution, the maximum value would be higher, balancing the upper right tail.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 51.84 65.80 73.26 71.46 77.19 83.98



## conclusion and summary