Final\_project\_sds230

2024-07-25

# Introduction

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

## data cleaning

## descriptive plots, summary information

## 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

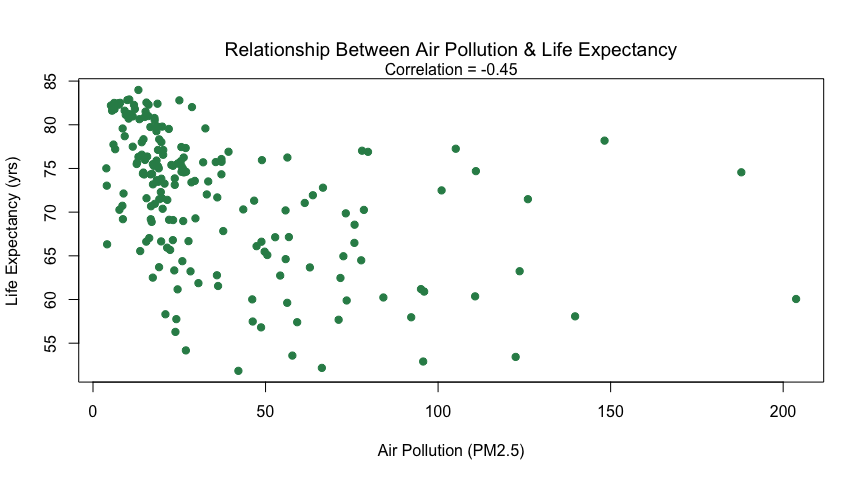
*For our permutation test, we want to see if air pollution significantly impacts life expectancy across different countries. Using the World Bank 2016 data variables “PM2.5” (mean annual exposure to air pollution) and “LifeExp”, our hypothesis is that higher levels of air pollution are associated with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

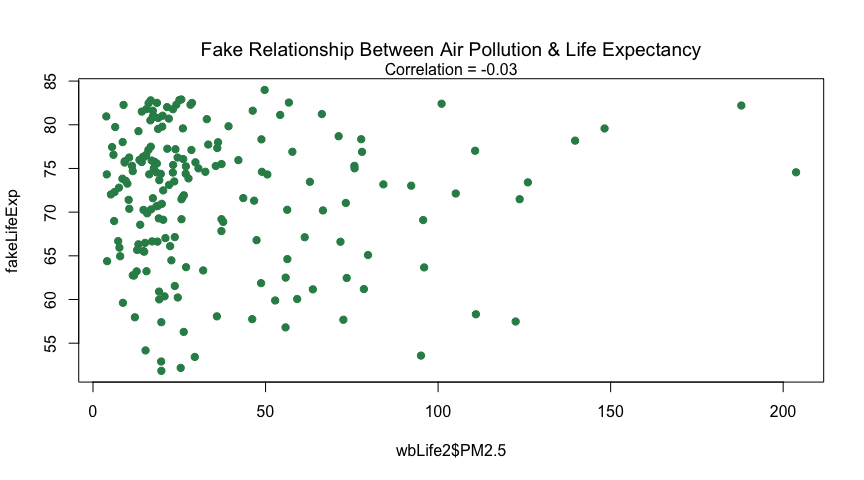
## 'data.frame': 188 obs. of 2 variables:  
## $ PM2.5 : num 62.9 14.6 37.2 36.2 15.7 ...  
## $ LifeExp: num 63.7 78.3 76.1 61.5 76.4 ...  
## - attr(\*, "na.action")= 'omit' Named int [1:29] 4 5 10 28 37 40 51 56 65 69 ...  
## ..- attr(\*, "names")= chr [1:29] "4" "5" "10" "28" ...

## 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

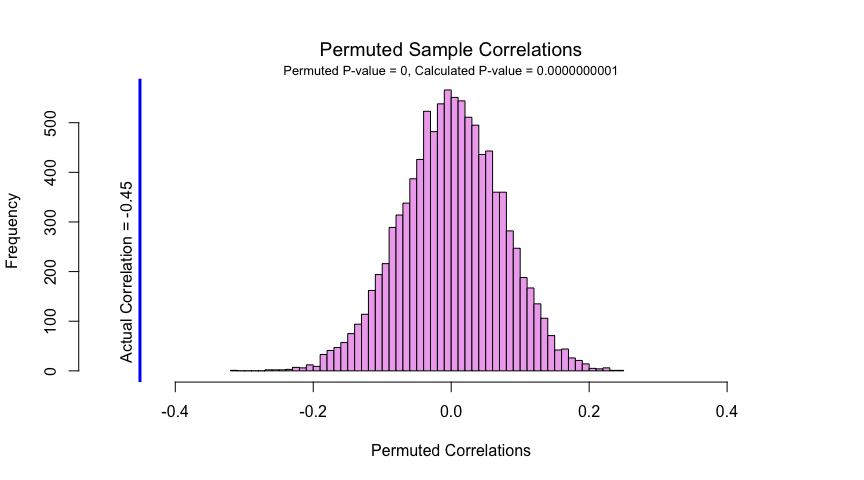
*The correlation is significant at alpha = .05 and .01, so there is evidence that there is statistically significant non-zero correlation between air pollution and life expectancy. So, let’s see what happens when we use fake data.*



*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

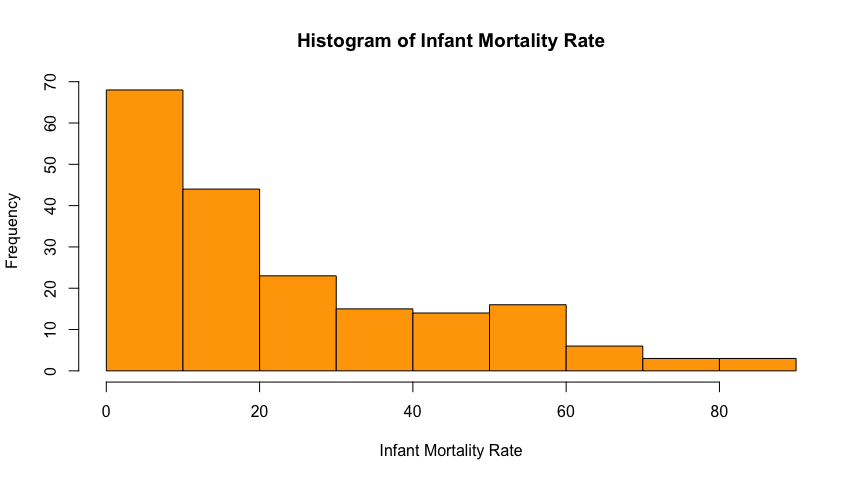
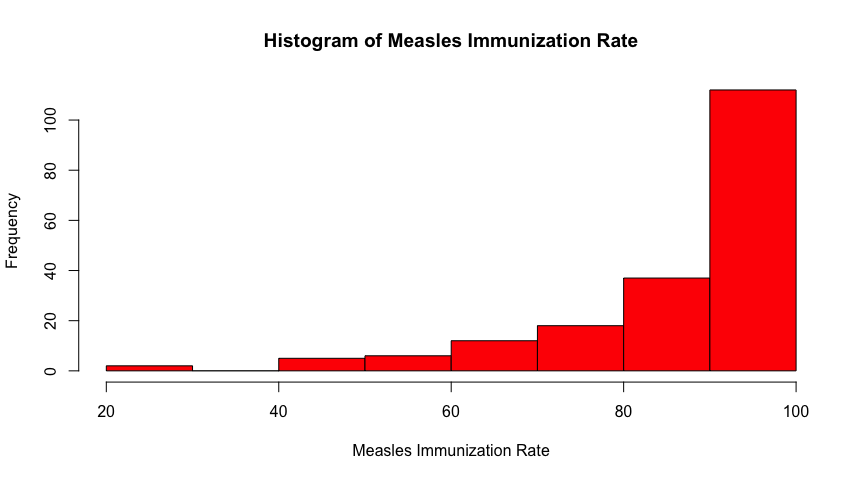
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.3131813 -0.0474443 0.0010860 0.0003454 0.0504363 0.2431204

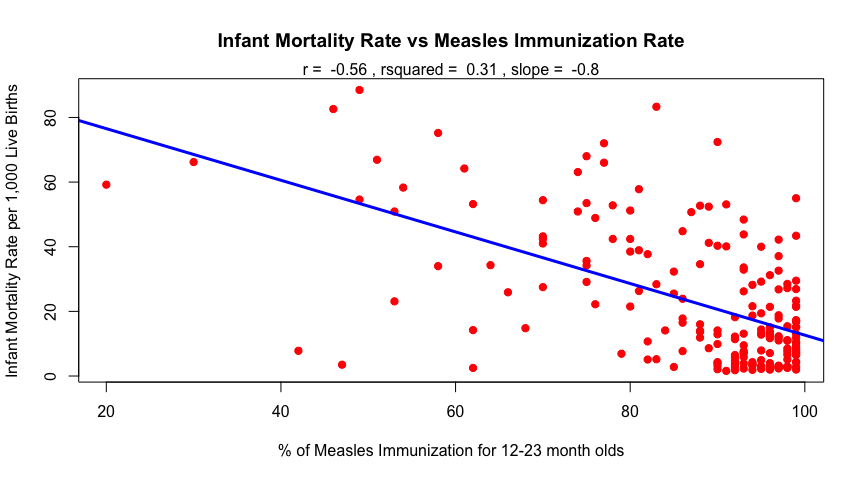


*The results from our permutation test reveal that the distribution of permuted correlations is normal and centered around zero. Since the actual observed correlation falls outside this distribution, this suggests that it is significantly different from what would be expected by random chance. This finding 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

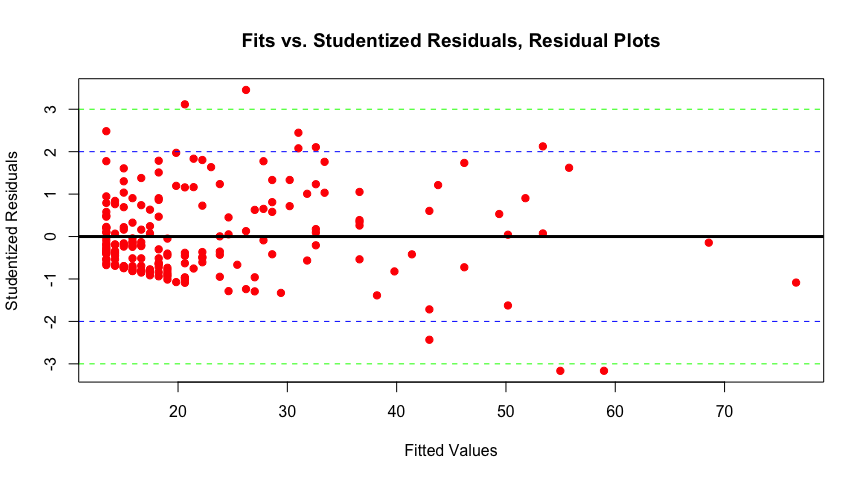
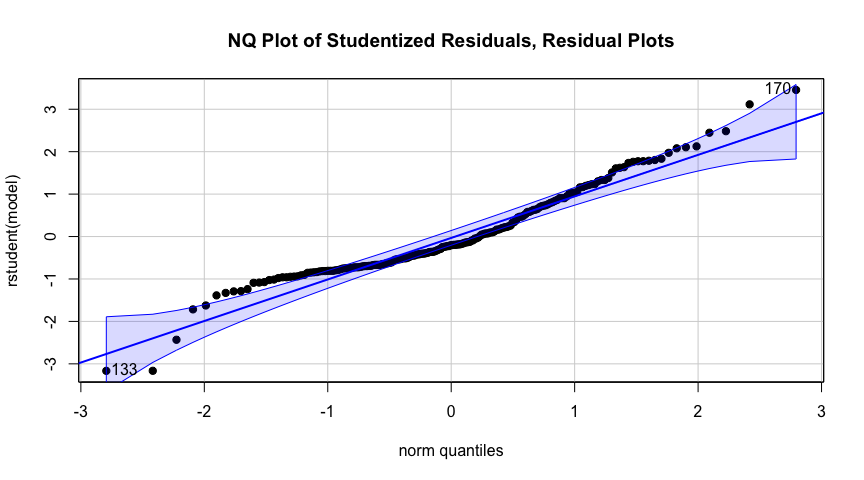
We are performing a simple linear regression to model the relationship between infant mortality rate (per 1,000 live births) and measles immunization rate for infants 12 - 23 months old. The assumptions are random, normally distributed errors centered at zero with constant variance (homoskedasticity) 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.



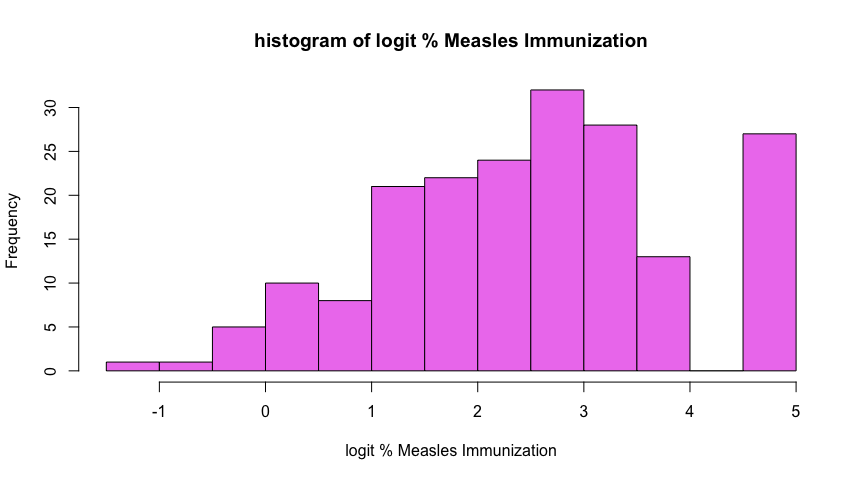
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 82.00 93.00 87.21 97.00 99.00

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.



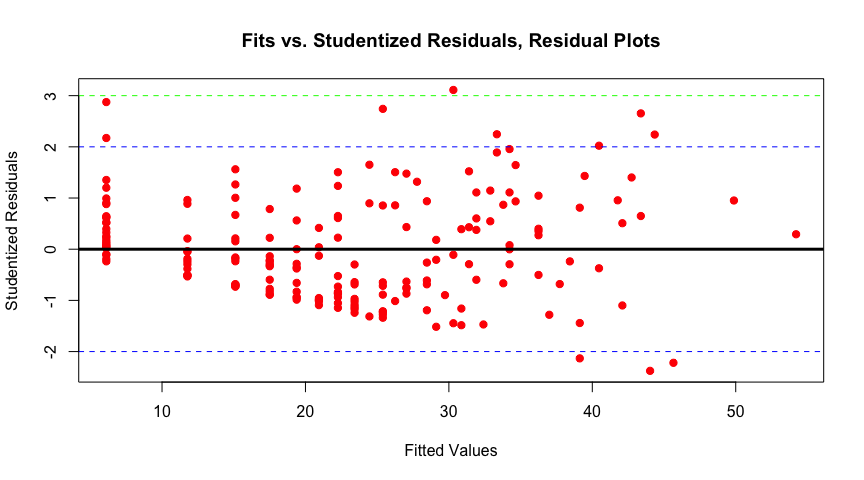
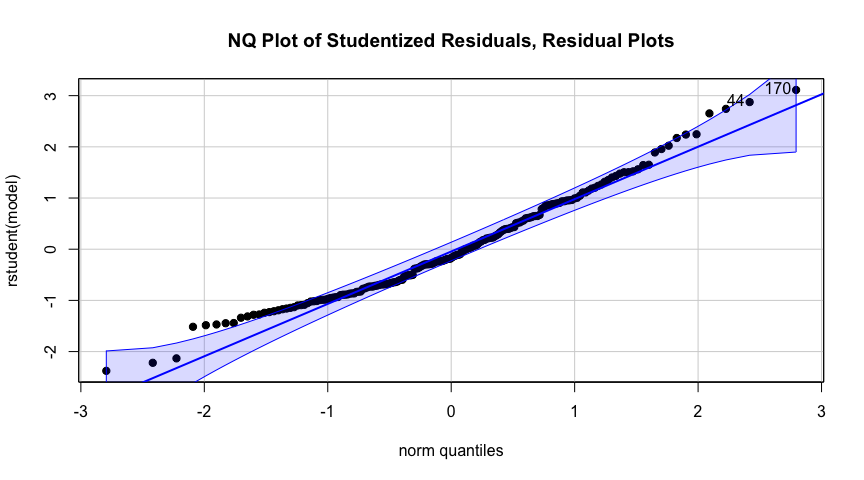
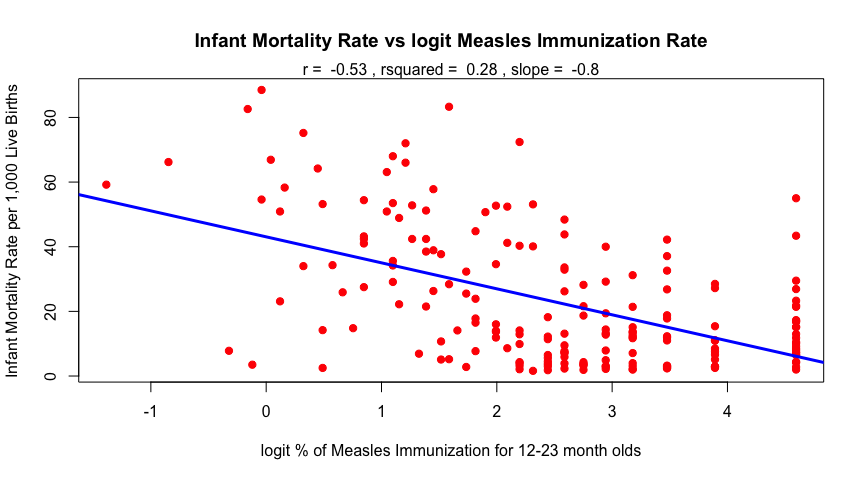
There looks to be some hetereoskedasticity in the fit vs studentized residuals, likely due to the extreme right skew of measles vacciination. This makes sense, as the median measles % vaccination is 93 and the mean is 87.21; most of the data is centered on the right. Since measles vaccination is a percentage,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

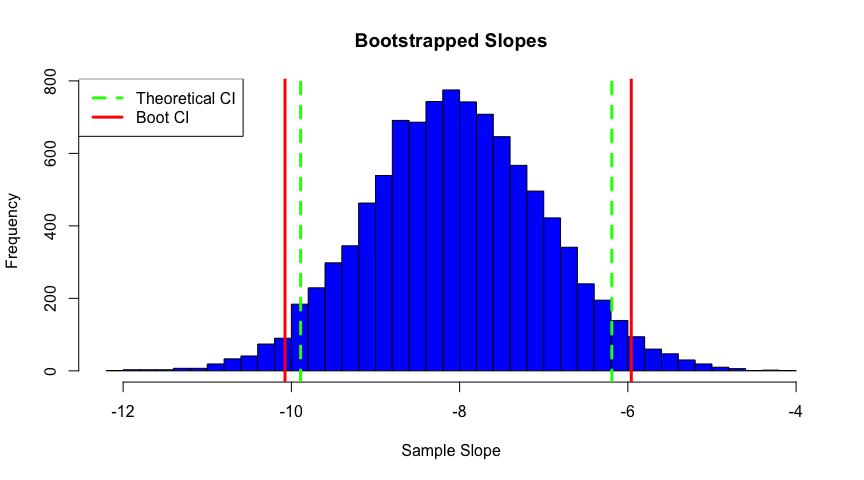
After taking logit of % Measles vaccination, we can see in the histogram is more evenly distributed. There is still a left skew, but this is to be expected since countries still have high measles vaccination rates. With our transformed predictor, we make another scatterplot and fit a new regression model.



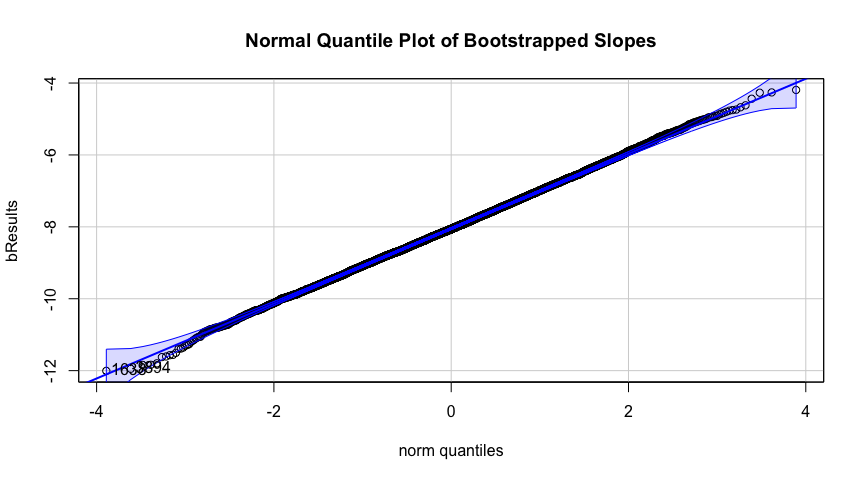
In the new scatter plot with logit measles, the data is more dispersed along the x axis. The spread and linear assumptions of correlation and linear models are better met with the transformed variable. The measles histogram is less skewed, the residual plot has less unequal variance, and the overall 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.



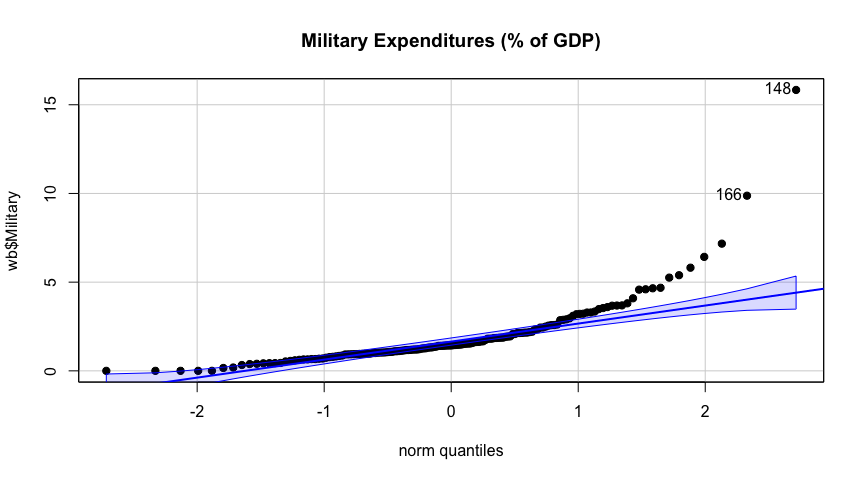
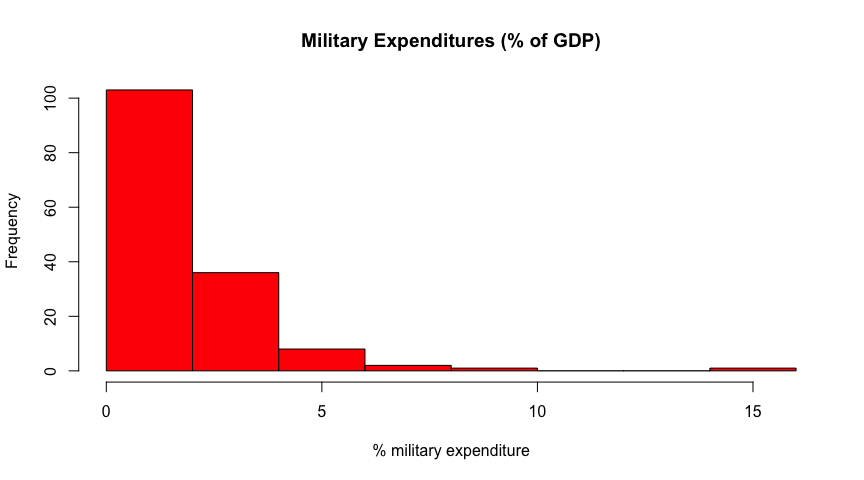
## [1] 1638 3894

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*

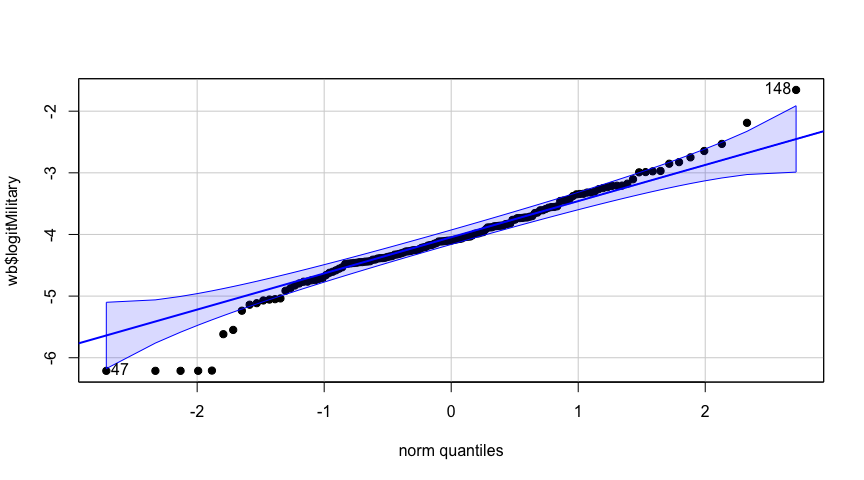
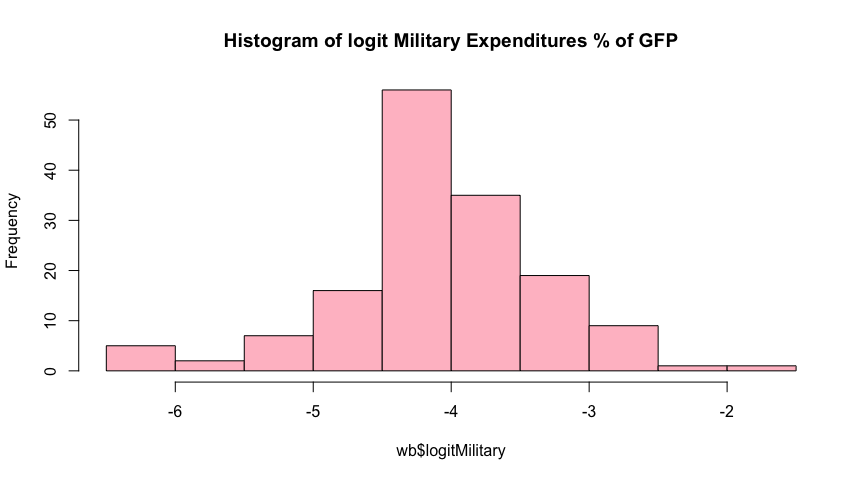
### look at response variable Military Expenditures



## [1] 148 166

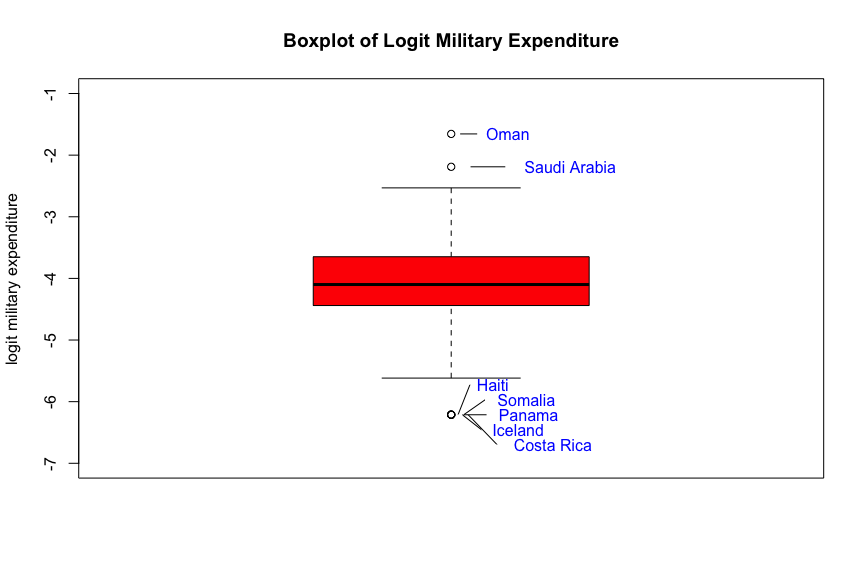
The data is heavily right-skewed and not normally distributed. These plots suggest using a logit transformation, which helps with probabilities or percentages. Due to zeros in the data, we add a small amount to each value to avoid issues with the logit function.

## 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.



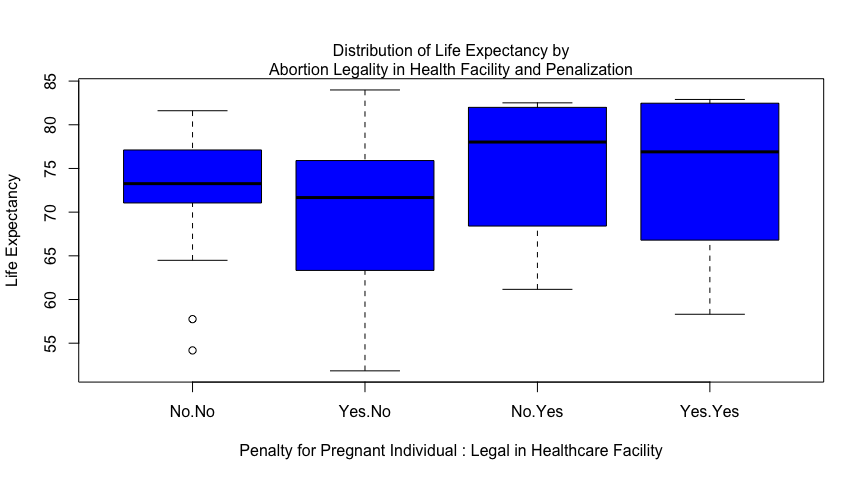
In this box plot, Haiti, Somalia, Panama, Iceland, and Costa Rica all have approximately 0% military expenditure. Oman and Saudi Arabia have relatively higher military expenditure compared to other countries in the 2016 World Bank dataset.

*Now that our response variable is transformed, we can begin to look at the relationships with this transformed variable and some potential explanatory variables. First, we make correlation plot of all the possible predictors we want to include in our model*

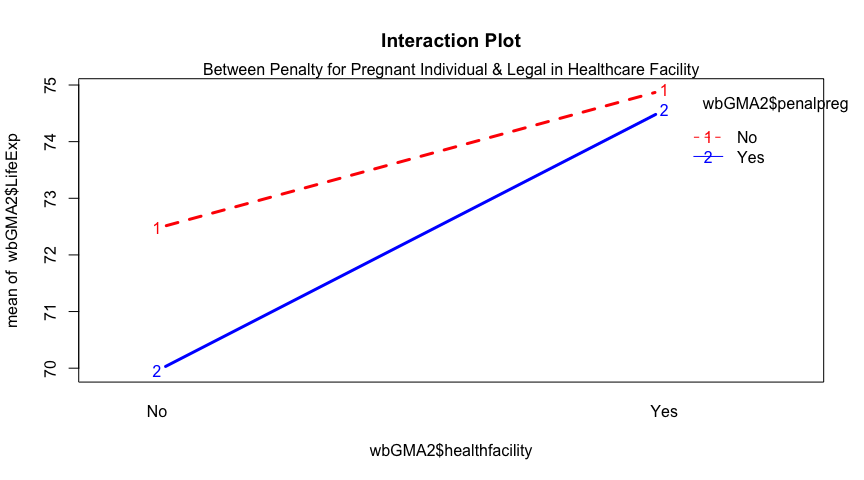
## 2-Way ANOVA

### 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.



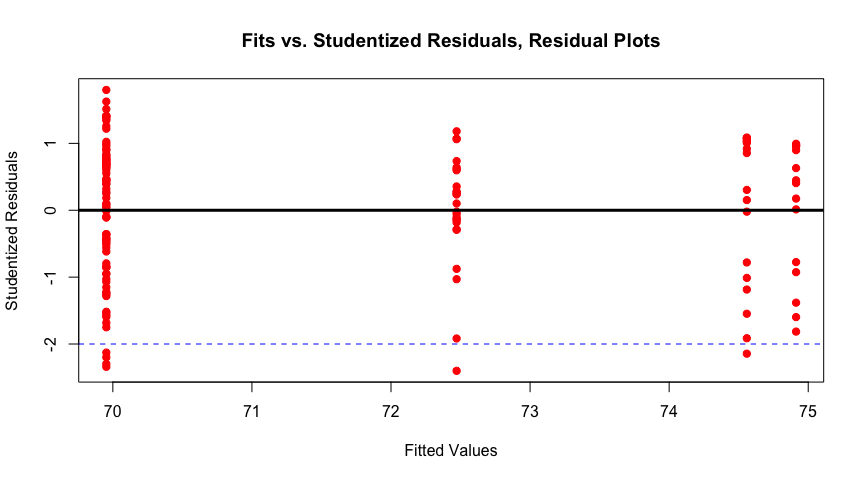
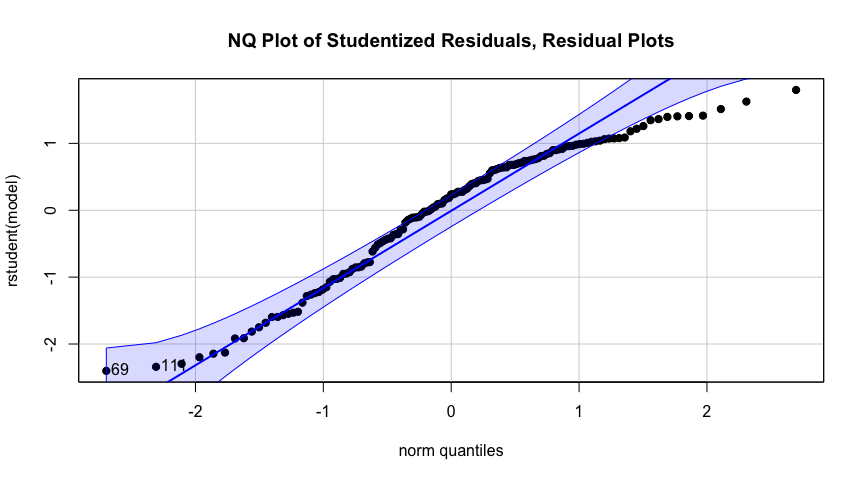
It is also important to note that we have a small sample size and an unbalanced design as seen in the table below, the number of observations in each group is not the same.

##   
## 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.

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