Final\_project\_sds230

2024-07-25

# Introduction

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

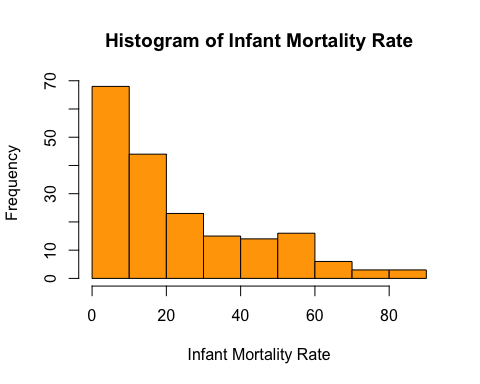
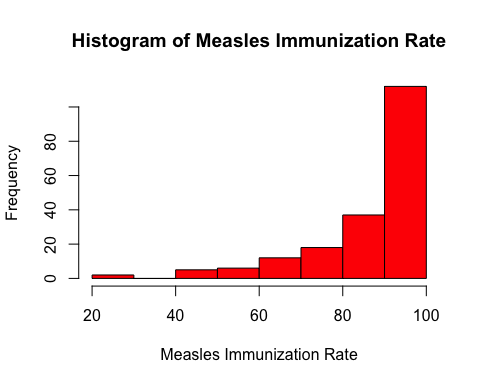
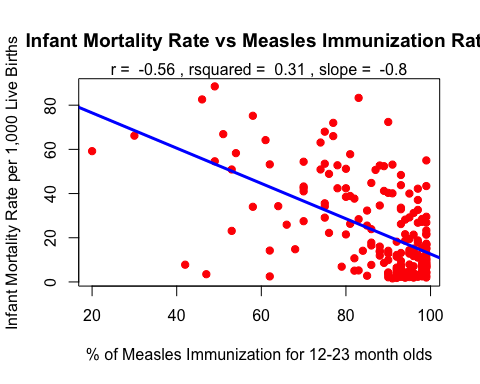
## data cleaning

## descriptive plots, summary information

# Analysis

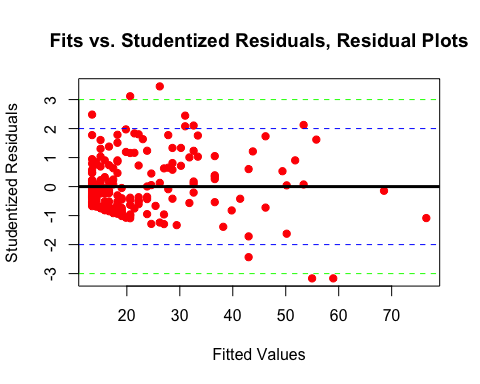
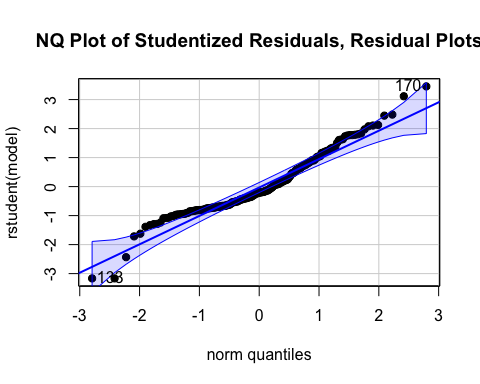
## T Test

## Correlation

We are performing a simple linear regression to model a linear relationship between the response variable, infant mortality rate per 1,000 live births, and the predictor variable, measles immunization rate. The assumptions for correlation and linear regression is that we have random, normally distributed errors centered at zero with constant variance. There should be homoskedasticity, with constant variance across groups. In addition, the variables must be linearly related. First, we want to investigate whether there is a relationship between infant mortality and measles vaccination rates. We check the distribution of the variables and see that Measles vaccination rates are heavily left skewed and infant mortality is heavily right skewed. We fit an inital linear model to these variables and calculate correlation and rsquared value.  

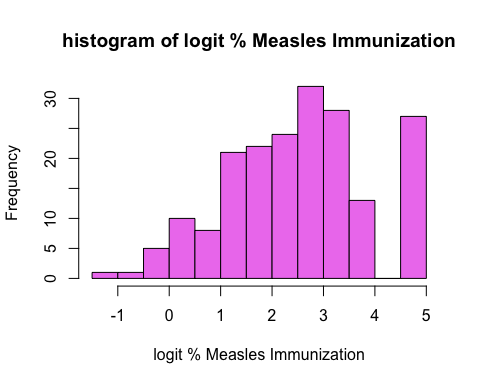
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 82.00 93.00 87.21 97.00 99.00

There is a moderate - strong negative relationship between % Measles immunizations and infant mortality, and the rsquared of .31 represents that 31% of the variability in infant mortality rate can be explained by this model. However, since measles immunization is piled towards 100%, this casues probelms of spread and variance. We can look at the residual plots to investiagte this issue further.



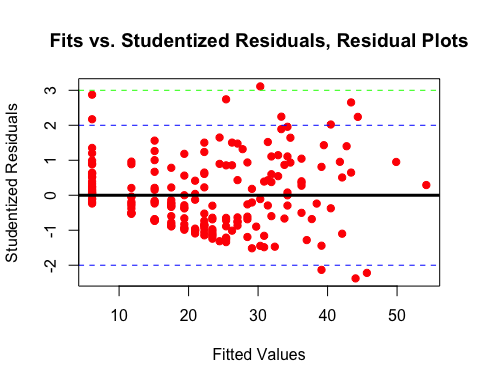
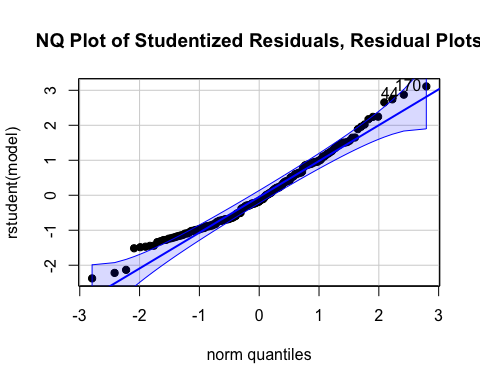
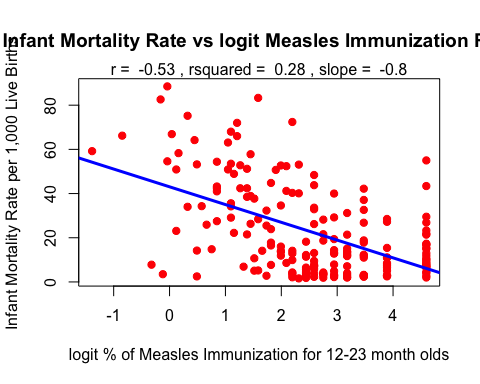
There looks to be some hetereoskedasticity in the fit vs studentized residuals, likely due to the extreme right skew of measles vacciination. This makes sense, as the median measles % vaccination is 93 and the mean is 87.21; most of the data is centered on the right. Since measles vaccination is a percentage,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



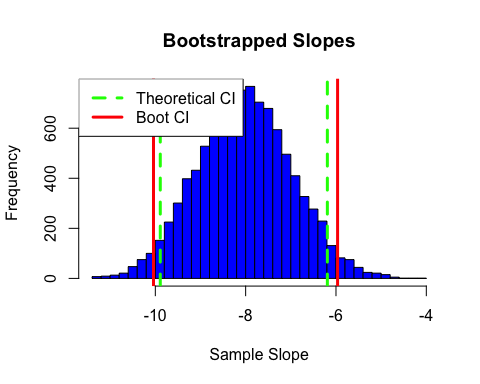
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.386 1.516 2.587 2.513 3.476 4.595

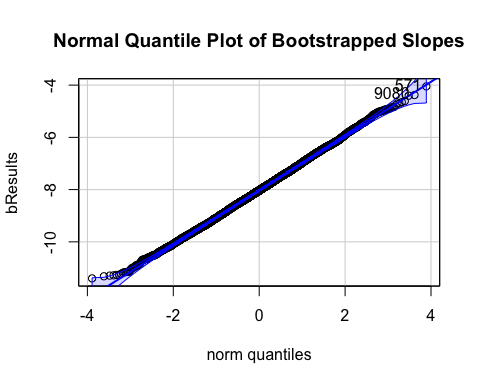
After taking logit of % Measles vaccination, we can see in the histogram is is more evenly distributed. There is still a left skew, but this is to be expected since countries still have high measles vaccination rates. With our transformed predictor, we make another scatterplot and fit a new regression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

 In the new scatterplot 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. With an rsquared value of .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 (rsquared = .31), the current model with logit measles vaccination is a better fit since the underlying assumptions are better met. The rsquared value is only one way of interpreting good model fit, and we don’t always necessarily want the model with the highest rsquared.

### Bootstrap CI for Correlation

In order to check the correlation and slope we calculated theoretically using parametric tests, we employ non parametric bootstrapping in order to calculate confidence intervals for the true correlation and slope of the relationship 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] 571 9080

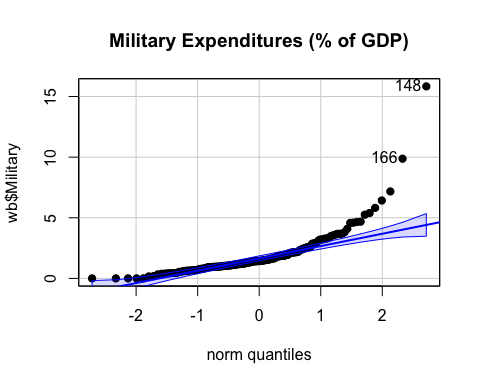
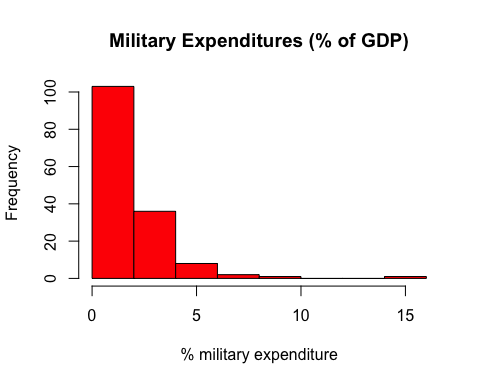
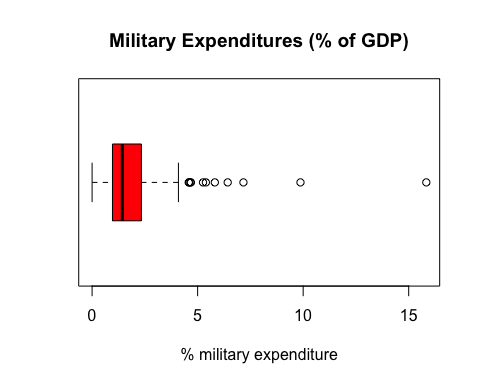
As expected, distributions for slope approximate normality. There is a slight right skew in the normal quantile plot of bootstrapped correlation which is also reflected in the histogram.The histogram for slope looks very near normal, and the data falls almost entirely along the straight line in the normal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

## Multiple regression

*introduction, data explanation, variables, etc*

### look at response variable Military Expenditures

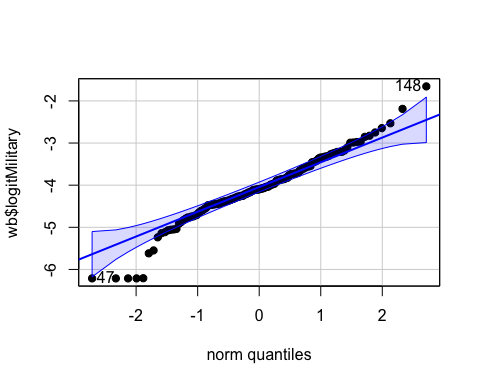
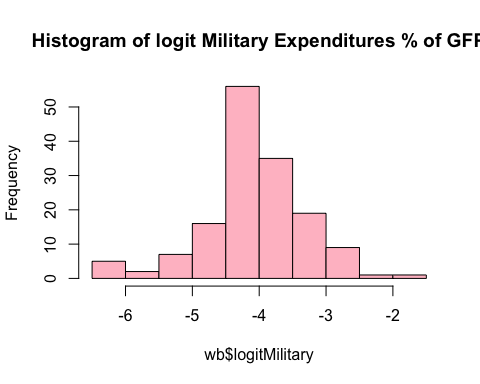
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.9654 1.4311 1.9383 2.3369 15.8344 66



## [1] 148 166

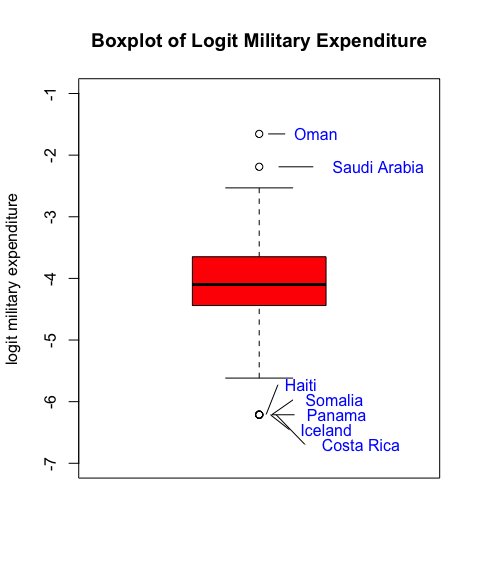
\*The data is heavily right skewed and not normally distributed. These plots all suggest we might want to look at things on the logit because logit can usually help when dealing with probabilities or percentages. Since there are zeros in the data, we must additionally add a small amount to every single value to avoid a function approaching negative infinity.

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



## [1] 148 47

Now the data seems more reasonably normally distributed and we begin to look at the predictor variables. We notice that there are a few potential outliers spending more and less than what we might expect from a normal distribution. We can see the countries that these outliers belong to below.



In this box plot, Haiti, Somalia, Panama, Iceland and Costa Rica all have approximately zero % military expenditure, so they are pointing to the same value. Oman and Saudi Arabia have relatively higher % military expenditure than the other countries listed in the world bank 2016 data set.

*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

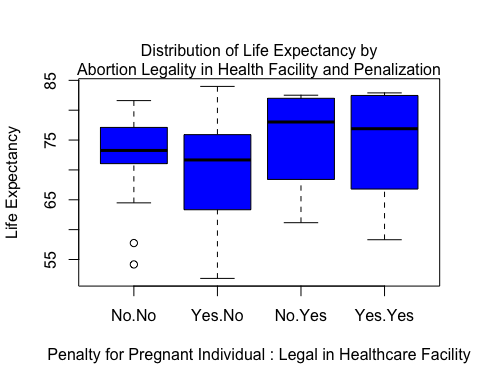
### Introduction

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 government 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.

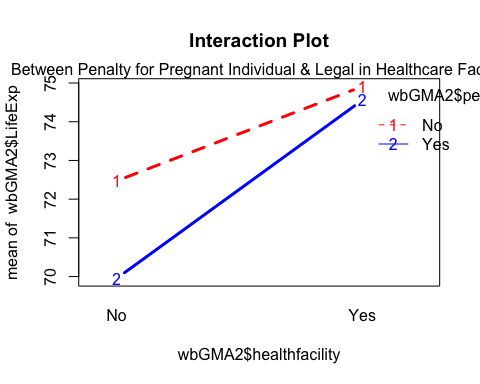
## tibble [220 × 52] (S3: tbl\_df/tbl/data.frame)  
## $ Jurisdictions : chr [1:220] "Aguascalientes" "Albania" "Algeria" "Andorra" ...  
## $ Effective Date : chr [1:220] "43094" "41275" "43435" "38353" ...  
## $ Valid Through Date : POSIXct[1:220], format: "2019-06-01" "2019-06-01" ...  
## $ glob\_regula : num [1:220] 1 1 1 1 1 1 1 1 1 1 ...  
## $ glob\_permis\_Economic reasons : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_Social reasons : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_Fetal impairment : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permisRape : chr [1:220] "1" "0" "0" "0" ...  
## $ glob\_permis\_Incest : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_Intellectual or cognitive disability of the pregnant person : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_Mental health : chr [1:220] "0" "0" "1" "0" ...  
## $ glob\_permis\_Physical health : chr [1:220] "0" "0" "1" "0" ...  
## $ glob\_permis\_Health : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permisLife : chr [1:220] "1" "0" "1" "0" ...  
## $ glob\_permis\_Age : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_Any grounds : chr [1:220] "0" "1" "0" "0" ...  
## $ glob\_permis\_Specific grounds not specified : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permis\_None : chr [1:220] "0" "0" "0" "1" ...  
## $ glob\_circum : chr [1:220] "1" "0" "1" "." ...  
## $ glob\_permit\_Medical practitioner : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permit\_Medical doctor : chr [1:220] "0" "0" "1" "0" ...  
## $ glob\_permit\_Specialist doctor : chr [1:220] "0" "1" "1" "0" ...  
## $ glob\_permit\_Midwife : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permit\_Nurse : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permit\_Pharmacist : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permit\_Lay health worker : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_permitPermitted health care professionals not specified : chr [1:220] "1" "0" "0" "0" ...  
## $ glob\_permit\_No health care professionals are legally permitted to provide an abortion: chr [1:220] "0" "0" "0" "1" ...  
## $ glob\_requir\_Ultrasound : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_requir\_Blood test : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_requir\_Urine test : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_requir\_Gynecological exam : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_requirRequired tests not specified : chr [1:220] "1" "0" "0" "0" ...  
## $ glob\_requir\_None : chr [1:220] "0" "1" "1" "1" ...  
## $ glob\_placea\_Hospital : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_placea\_Primary health care facility : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_placea\_Secondary health care facility : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_placea\_Health facility specifically designated to provide abortions : chr [1:220] "0" "1" "0" "0" ...  
## $ glob\_placea\_Government health facility : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_placeaRequired location not specified : chr [1:220] "1" "0" "1" "0" ...  
## $ glob\_placea\_Abortion is not permitted in any location : chr [1:220] "0" "0" "0" "1" ...  
## $ glob\_penaltAny individual who causes an unlawful abortion : chr [1:220] "1" "1" "1" "1" ...  
## $ glob\_penaltPregnant person : chr [1:220] "1" "0" "1" "1" ...  
## $ glob\_penaltHealth care professional who provides surgical abortion : chr [1:220] "1" "1" "1" "1" ...  
## $ glob\_penalt\_Individual who provides surgical abortion : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_penaltHealth care professional who provides abortion drugs : chr [1:220] "1" "1" "1" "1" ...  
## $ glob\_penalt\_Individual who provides abortion drugs : chr [1:220] "0" "1" "0" "0" ...  
## $ glob\_penalt\_Individual who provides substances : chr [1:220] "0" "0" "1" "0" ...  
## $ glob\_penalt\_Individual who provides instruments for abortion : chr [1:220] "0" "1" "0" "0" ...  
## $ glob\_penalt\_Individual who provides information on abortion : chr [1:220] "0" "1" "1" "0" ...  
## $ glob\_penalt\_Individual who assists in the abortion : chr [1:220] "0" "0" "0" "0" ...  
## $ glob\_penalt\_Individuals subject to penalties not specified : chr [1:220] "0" "0" "0" "0" ...

## penalpreg healthfacility LifeExp Country Fertility16  
## 2 No Yes 78.345 Albania 1.713  
## 3 Yes No 76.078 Algeria 2.776  
## 6 Yes No 61.547 Angola 5.694  
## 7 Yes No 76.364 Antigua and Barbuda 2.051  
## 8 Yes No 76.577 Argentina 2.293  
## 9 No No 74.618 Armenia 1.612

## [1] 143



Now we want to check interactions between the categorical variables and continuous variable. 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

The lines do not move in a parallel fashion, so there *may* be a significant interaction effect between whether a pregnant woman is penalized for having an abortion and whether an abortion is legally allowed to take place in a government health facility. For instances where the woman is penalized, countries permitting abortions in health facilities have a slightly higher mean life expectancy. In instances where the pregnant woman is not penalized, countries that do not permit abortions in health facilities look like they have a higher mean life expectancy. The plot itself however is not a statistical test and is merely a suggestion for what we may see when performing the statistical test.

We run a two way ANOVA model including the main effects ‘penalpreg’ and ‘healthfacility’ and their interaction. However, none of the main effects nor interaction was significant. Therefore we run a second anova model with out the interaction effect. This model shows there is a significant main effect, ‘healthfacility.’ Whether or not abortions are allowed in a abortion specified government health facility or not has a statistical effect on life expectancy. When we run a linear model without the interaction effect included, the P value of the F statistic is .015, which is lower compared to the linear model with the interaction included (P = .03). This indicates that the model without the interaction may be a better fit. However, adjusted rsquared value could be is .04, meaning about 4% of the variance in life expectancy can be attributed to the model. Since there is a small sample size and only one significant main effect, the model may not explain the data very well.

## Anova Table (Type III tests)  
##   
## Response: wbGMA2$LifeExp  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 131304 1 2098.8459 <2e-16 \*\*\*  
## wbGMA2$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   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

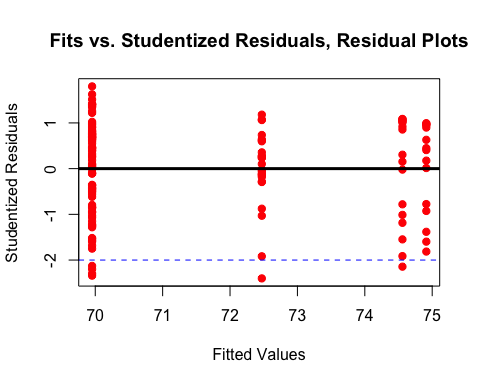
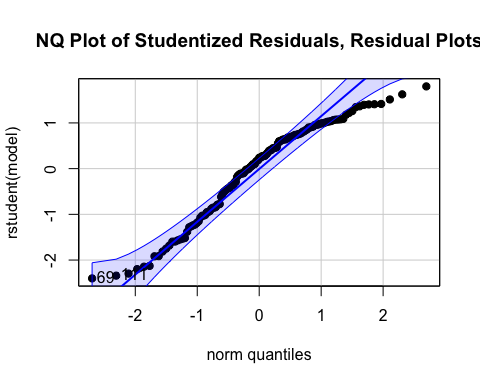
## Anova Table (Type III tests)  
##   
## Response: wbGMA2$LifeExp  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 167328 1 2685.7414 < 2e-16 \*\*\*  
## wbGMA2$penalpreg 97 1 1.5638 0.21319   
## wbGMA2$healthfacility 330 1 5.3042 0.02275 \*   
## Residuals 8722 140   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## lm(formula = wbGMA2$LifeExp ~ wbGMA2$penalpreg + wbGMA2$healthfacility +   
## wbGMA2$penalpreg \* wbGMA2$healthfacility)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.298 -6.125 1.850 6.071 14.032   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) 72.472 1.582 45.813  
## wbGMA2$penalpregYes -2.519 1.797 -1.402  
## wbGMA2$healthfacilityYes 2.441 2.583 0.945  
## wbGMA2$penalpregYes:wbGMA2$healthfacilityYes 2.165 3.329 0.650  
## Pr(>|t|)   
## (Intercept) <2e-16 \*\*\*  
## wbGMA2$penalpregYes 0.163   
## wbGMA2$healthfacilityYes 0.346   
## wbGMA2$penalpregYes:wbGMA2$healthfacilityYes 0.517   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.909 on 139 degrees of freedom  
## Multiple R-squared: 0.0608, Adjusted R-squared: 0.04053   
## F-statistic: 2.999 on 3 and 139 DF, p-value: 0.03281

##   
## Call:  
## lm(formula = wbGMA2$LifeExp ~ wbGMA2$penalpreg + wbGMA2$healthfacility)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.260 -6.351 1.891 6.018 13.890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 71.983 1.389 51.824 <2e-16 \*\*\*  
## wbGMA2$penalpregYes -1.888 1.510 -1.251 0.2132   
## wbGMA2$healthfacilityYes 3.744 1.626 2.303 0.0227 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.893 on 140 degrees of freedom  
## Multiple R-squared: 0.05794, Adjusted R-squared: 0.04448   
## F-statistic: 4.305 on 2 and 140 DF, p-value: 0.01533

Plotting fit vs studentized residuals shows that the residuals are approximately normally distributed and there isn’t a great violation in equal variances. However, there are some non conforming data on the right tail of the distribution. This is likely due to the fact that life expectancy is left skewed, with a maximum of about 84 years. In a normal distribution, we may expect the maximum life expectancy value to be higher to even out 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