Final\_project\_sds230

2024-07-25

# Introduction

## data explanation

The data in this project includes

## data cleaning

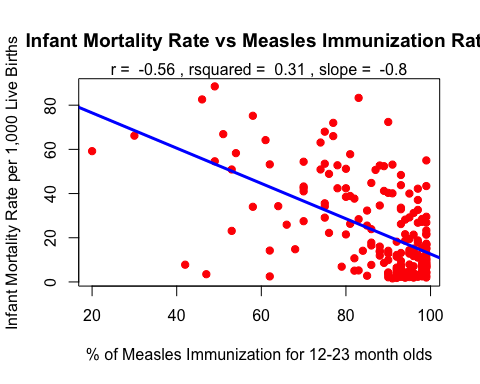
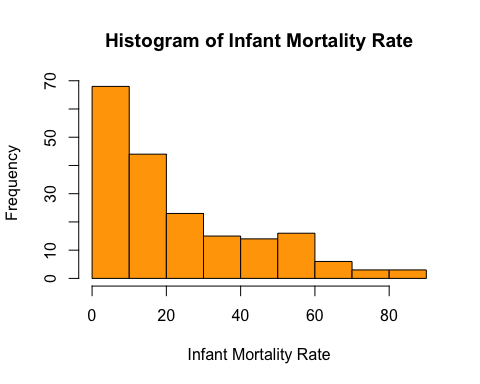
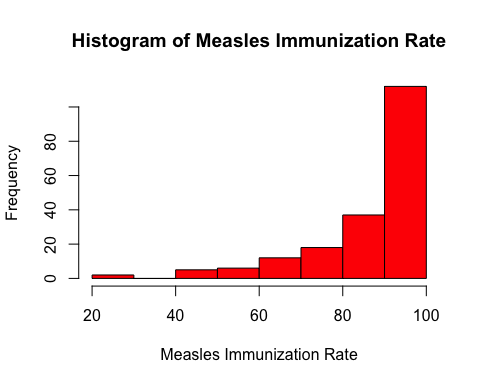
## descriptive plots, summary information

# Analysis

## T Test

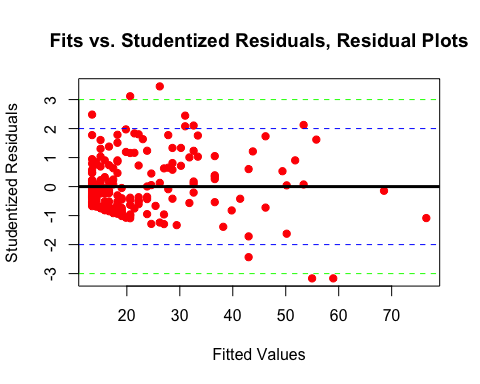
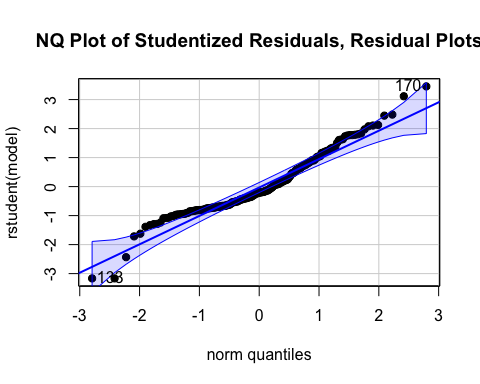
## Correlation

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 rat. 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. 

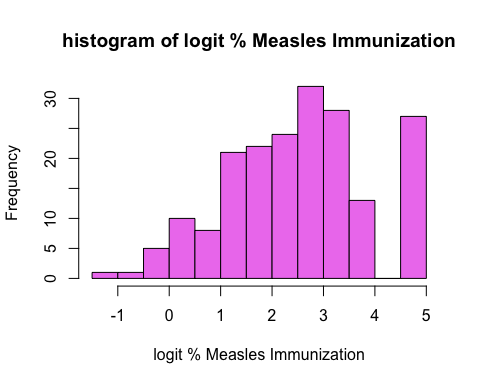
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 20.00 82.00 93.00 87.21 97.00 99.00

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.



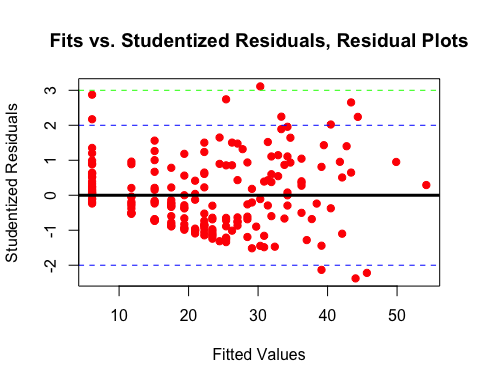
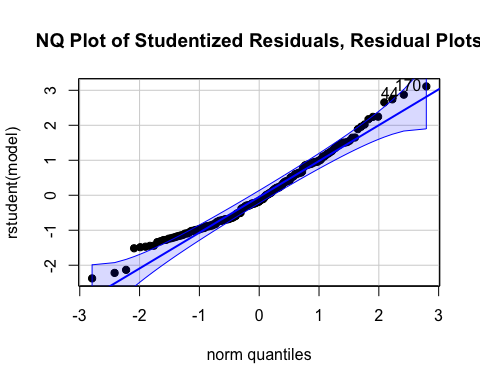
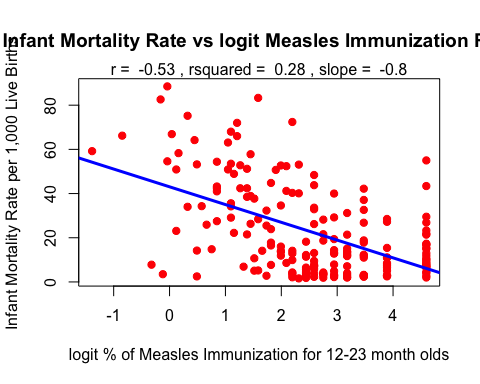
There looks to be some hetereoskedasticity in the fit vs studentized residuals, likely due to the extreme right skew of measles vacciination. This makes sense, as the median measles % vaccination is 93 and the mean is 87.21; most of the data is centered on the right. Since measles vaccination is a percentage,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



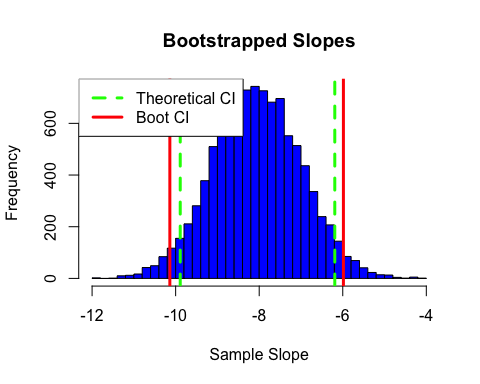
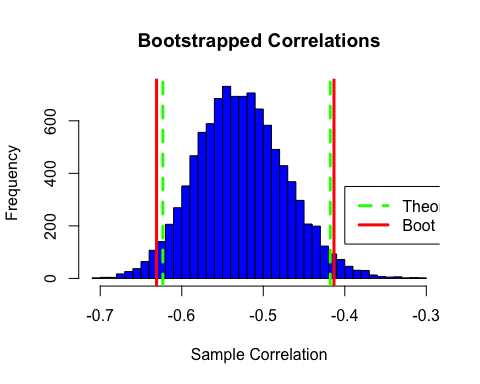
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.386 1.516 2.587 2.513 3.476 4.595

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.

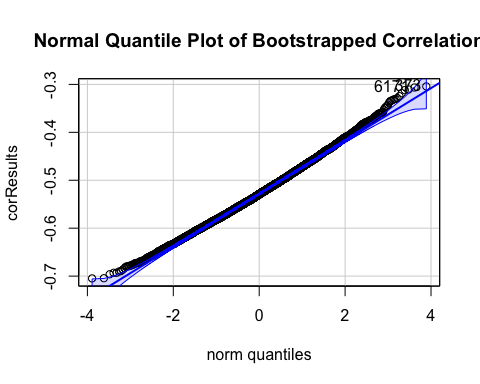
### Bootstrap CI for Correlation and Slope

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.

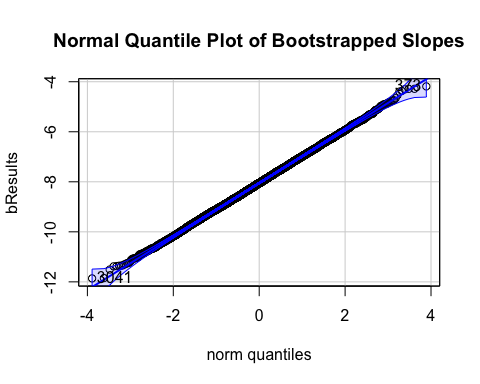


##   
## Call:  
## lm(formula = InfMort ~ Measles, data = mort)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.476 -11.761 -3.471 10.704 57.082   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 92.51959 7.55429 12.247 <2e-16 \*\*\*  
## Measles -0.79881 0.08546 -9.347 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 17.04 on 190 degrees of freedom  
## Multiple R-squared: 0.315, Adjusted R-squared: 0.3114   
## F-statistic: 87.36 on 1 and 190 DF, p-value: < 2.2e-16

The histograms above show the bootstrapped data both correlation and slope, the bootstrapped confidence intervals are a bit wider than the theoretical. In both cases, the bootstrapped confidence intervals are more conservative than the theoretical, which makes sense because the bootstrapping is non parametric therefore capturing more of the true variability in the data. There doesn’t appear to be a large difference between the theoretical and bootstrapped confidence intervals. This could mean that the linear model we fit approximated the assumptions of normality, homoskedasticity, and independence of errors. We can look at normal quantile plots of the bootstrapped data to further visualize the bootstrapped distribution.



## [1] 373 6171



## [1] 373 3041

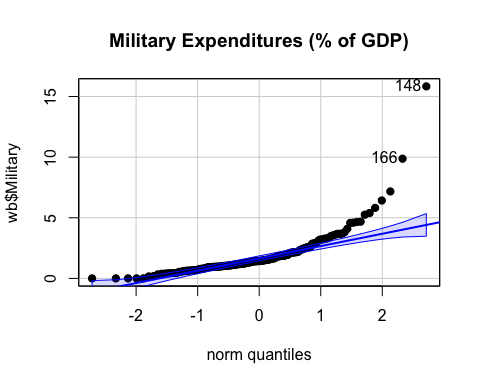
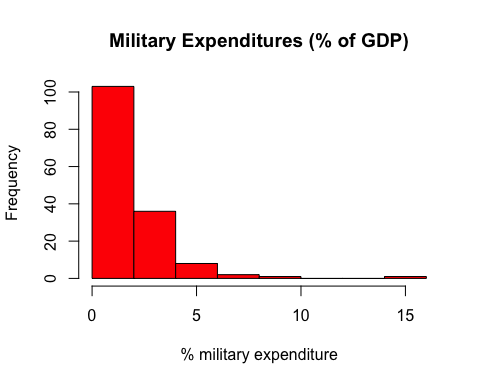
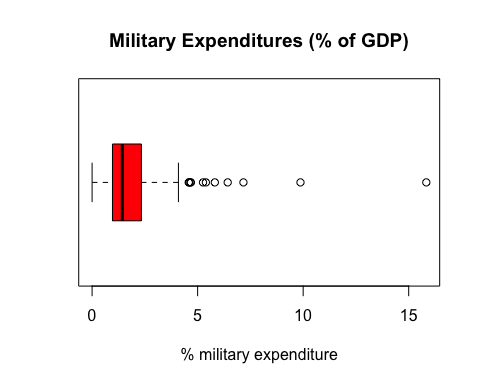
As expected, distributions for correlation and slope approximate reflect normality. There is a slight right skew in the normal quantile plot of bootstrapped correlation which is also reflected in the histogram. The histogram for slope looks very near normal, and the data falls almost entirely along the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

## Multiple regression

### Multiple Regression

#### 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



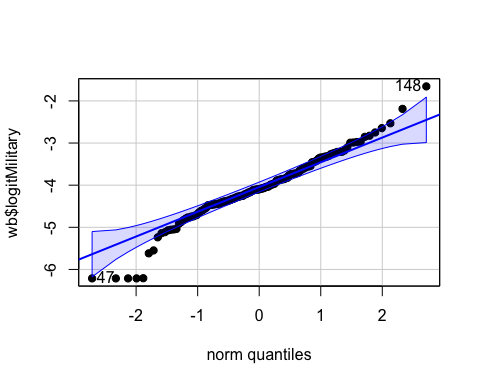
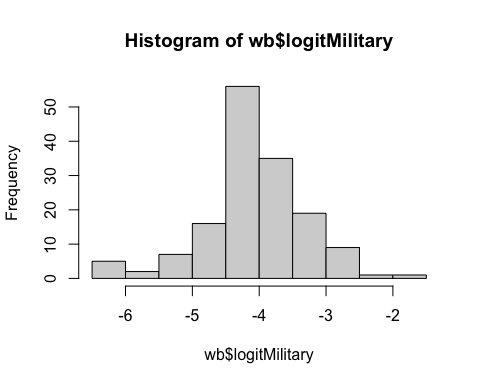
## [1] 148 166

*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 its %. logit function is log(p/(1-p)) or it is log(%/(100-%))*

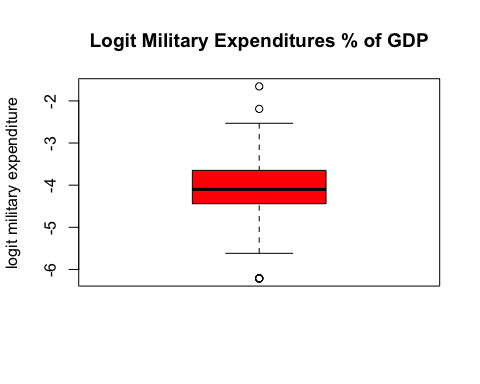
## Note: largest value of p > 1 so values of p interpreted as percents

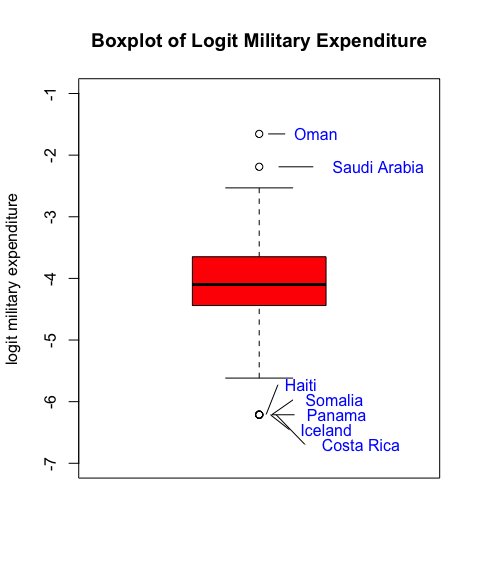
## [1] -6.212606 -6.212606 -6.212606 -6.212452 -6.208375 -5.617533 -5.548782  
## [8] -5.237265 -5.139962 -5.114686 -5.070914 -5.056188 -5.049748 -5.037522  
## [15] -4.913687 -4.872999 -4.827866 -4.794383 -4.767037 -4.765325 -4.743495  
## [22] -4.741470 -4.719833 -4.706733 -4.659060 -4.622170 -4.604560 -4.581933  
## [29] -4.555325 -4.532690 -4.477940 -4.473240 -4.468414 -4.462912 -4.458578  
## [36] -4.449765 -4.448399 -4.441805 -4.439004 -4.433982 -4.409882 -4.406643  
## [43] -4.390511 -4.384788 -4.383016 -4.377629 -4.365033 -4.356480 -4.350673  
## [50] -4.328546 -4.324704 -4.316073 -4.304345 -4.292412 -4.275911 -4.274303  
## [57] -4.268288 -4.257613 -4.253849 -4.252830 -4.239741 -4.217996 -4.209144  
## [64] -4.203064 -4.181094 -4.173386 -4.168690 -4.151195 -4.139714 -4.115001  
## [71] -4.113618 -4.111898 -4.106841 -4.101890 -4.100589 -4.099449 -4.089269  
## [78] -4.082789 -4.076405 -4.071863 -4.064889 -4.047097 -4.040242 -4.034812  
## [85] -4.033167 -4.014669 -3.989744 -3.983536 -3.981144 -3.967217 -3.962206  
## [92] -3.922975 -3.887239 -3.885577 -3.883430 -3.867828 -3.864535 -3.860454  
## [99] -3.857007 -3.854997 -3.833057 -3.826553 -3.821052 -3.763537 -3.759719  
## [106] -3.735724 -3.734464 -3.731882 -3.725047 -3.723474 -3.711693 -3.702063  
## [113] -3.652411 -3.644634 -3.608557 -3.606485 -3.586464 -3.566236 -3.555831  
## [120] -3.553516 -3.542986 -3.457119 -3.450313 -3.441052 -3.421707 -3.377344  
## [127] -3.349585 -3.343333 -3.340084 -3.320550 -3.319790 -3.303826 -3.264374  
## [134] -3.248138 -3.232867 -3.211914 -3.209400 -3.207478 -3.176256 -3.104951  
## [141] -2.992209 -2.987743 -2.975580 -2.969466 -2.852528 -2.826362 -2.749015  
## [148] -2.645858 -2.531159 -2.189169 -1.655668

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -6.213 -4.440 -4.099 -4.078 -3.649 -1.656 66



## [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 boxplot, 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 realtively higher % military expenditure than the other countries listed in the world bank 2016 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

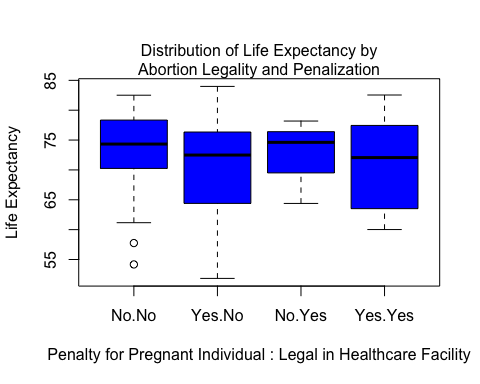
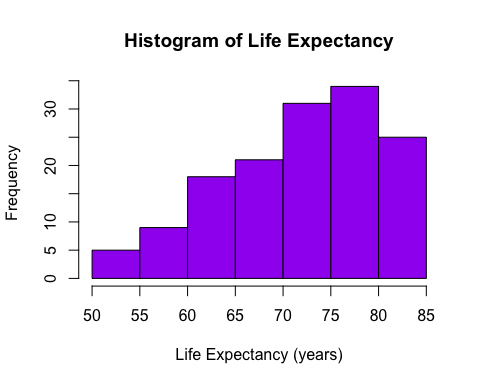
We are interested in predicting life expectancy based on some categorical variables a data set concerning global abortion laws concerning self-managed abortion. *need to talk about what variables we are using from GWB dataset.* Life expectancy is a variable from the world bank dataset that is slightly left skewed. However, this should not be a problem if the residuals are normally distirbuted and show homoskedasticity after fitting a model.

First, read in data, do some data cleaning by converting non 1 and 0 to NA, converting 1 to Yes and 0 to No. Then, we can join the GMA dataset by country name with the world bank dataset. The names in the GMA dataset are long, so we assign new names to them. THe indicator variable for pregnant women being penalized for participating in an unlawful abortion becomes “penalpreg” and the indicator variable indicating whether an abortion is legally premitted to take place in a government health facility becomes healthfacility. Additionally, in order to make sure the two way ANOVA is a balanced design, we removed rows with NAs in our columns of interest from the dataframe. In the end, to perform two way ANOVA, we have 143 unique countries with information from the world bank and GMA data set. In order to get a sense of the data, we can use boxplots to look at the distribution of life expectancy for each level of the two categorical variables. The variance between groups doesn’t seem hugely different, and there also doesn’t seem to be a real difference in life expectancy across the groups.

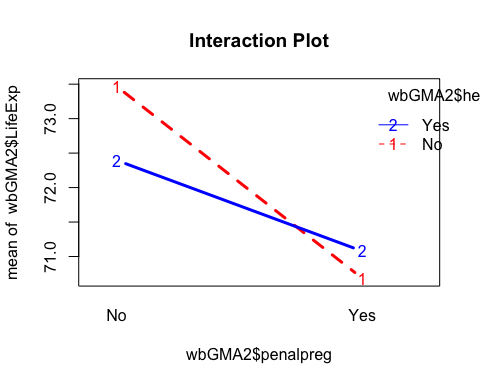
## tibble [220 × 52] (S3: tbl\_df/tbl/data.frame)  
## $ Jurisdictions : chr [1:220] "Aguascalientes" "Albania" "Algeria" "Andorra" ...  
## $ Effective Date : chr [1:220] "43094" "41275" "43435" "38353" ...  
## $ Valid Through Date : POSIXct[1:220], format: "2019-06-01" "2019-06-01" ...  
## $ glob\_regula : 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 No 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

 The lines do not move in a parallel fashion, so there might 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.

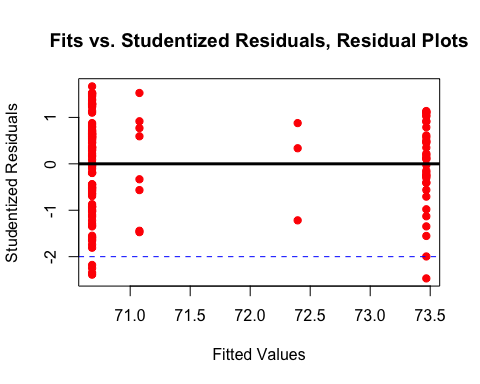
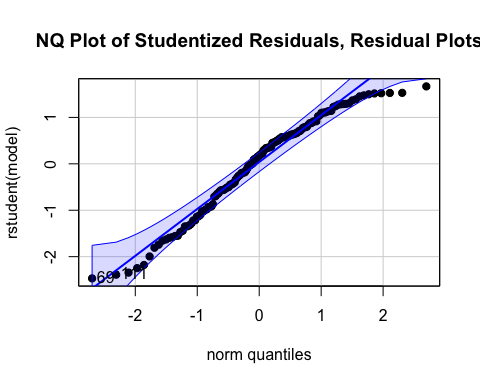
## Df Sum Sq Mean Sq F value Pr(>F)   
## wbGMA2$penalpreg 1 206 206.01 3.165 0.0774 .  
## wbGMA2$healthfacility 1 0 0.00 0.000 0.9981   
## wbGMA2$penalpreg:wbGMA2$healthfacility 1 4 4.34 0.067 0.7967   
## Residuals 139 9048 65.10   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## lm(formula = wbGMA2$LifeExp ~ wbGMA2$penalpreg + wbGMA2$healthfacility +   
## wbGMA2$penalpreg \* wbGMA2$healthfacility - 1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.294 -5.173 1.450 5.735 13.302   
##   
## Coefficients:  
## Estimate Std. Error t value  
## wbGMA2$penalpregNo 73.4675 1.3264 55.388  
## wbGMA2$penalpregYes 70.6825 0.8278 85.388  
## wbGMA2$healthfacilityYes -1.0720 4.8434 -0.221  
## wbGMA2$penalpregYes:wbGMA2$healthfacilityYes 1.4666 5.6816 0.258  
## Pr(>|t|)   
## wbGMA2$penalpregNo <2e-16 \*\*\*  
## wbGMA2$penalpregYes <2e-16 \*\*\*  
## wbGMA2$healthfacilityYes 0.825   
## wbGMA2$penalpregYes:wbGMA2$healthfacilityYes 0.797   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.068 on 139 degrees of freedom  
## Multiple R-squared: 0.9878, Adjusted R-squared: 0.9874   
## F-statistic: 2805 on 4 and 139 DF, p-value: < 2.2e-16

there is not a significant difference in life expectancy between groups, but the individual predictors are significant predictors of life expectancy. This could also be due to the fact that there are not very many observations (N = 143). The rsquared value could be too high due to over fitting the model, since again there are not many observations. Since there is a small sample size, the model may suit this data very well but may not be applicable if we were to introduce more new data.

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