# SDS348 Project 2

Seungchul Yeom

4/18/2021

##Introduction

*I used the data collected in project 1 so I joined and prepared the data to be ready to perform EDA, MANOVA, Randomization test, linear regression model, and logistic regression. The data has variables that shows Suicide rate and income inequality. Furthermore, I added some more variables that fits certain conditions with other variables such as suicide_con which tells whether the suicide rate is low or high.*

```
#Data collecting
library(readxl)
IncomeMed <- read_excel("IncomeMed.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
IncomeInequal <- read_excel("IncomeInequal.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
Suicide <- read_excel("Suicide.xlsx")
```

```
## New names:
## * `` -> ...2
```

```
#Tidy data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
IncomeMed = IncomeMed %>%
  select(-...2, -AIAN, -Asian, -Black, -Hispanic,           - White)

Suicide = Suicide %>%
  select(-...2, -AIAN, -Asian, -Black, -Hispanic,           - White, -`Crude Rate`) %>%
  select(-`Error Margin (Age-adjusted)`)

IncomeInequal = IncomeInequal %>%
  select(-...2, -`County Value`, -"Z-Score")

Income_Suicide = left_join(Suicide,
                           IncomeMed,
                           by = "County")

IncomeInequal_Suicide = left_join(Income_Suicide,
                                  IncomeInequal,
                                  by = "County")

library(tidyr)
IncomeInequal_Suicide = IncomeInequal_Suicide %>%
  rename(Suicide_Rate = "County Value.x",
         "Household_Income" = "County Value.y",
         Death_Num = "# Deaths")

mydata = IncomeInequal_Suicide %>%
  mutate(Inequality = `80th Percentile Income`           - `20th Percentile Income`) %>%
  arrange(Suicide_Rate) %>%
  select(-`Error Margin`) %>%
  filter(!is.na(Suicide_Rate)) %>%
  mutate(State = "Texas")

mydata$suicide_con = ifelse(mydata$Suicide_Rate > 15,
                                          c("high"),
                                          c("low"))
```

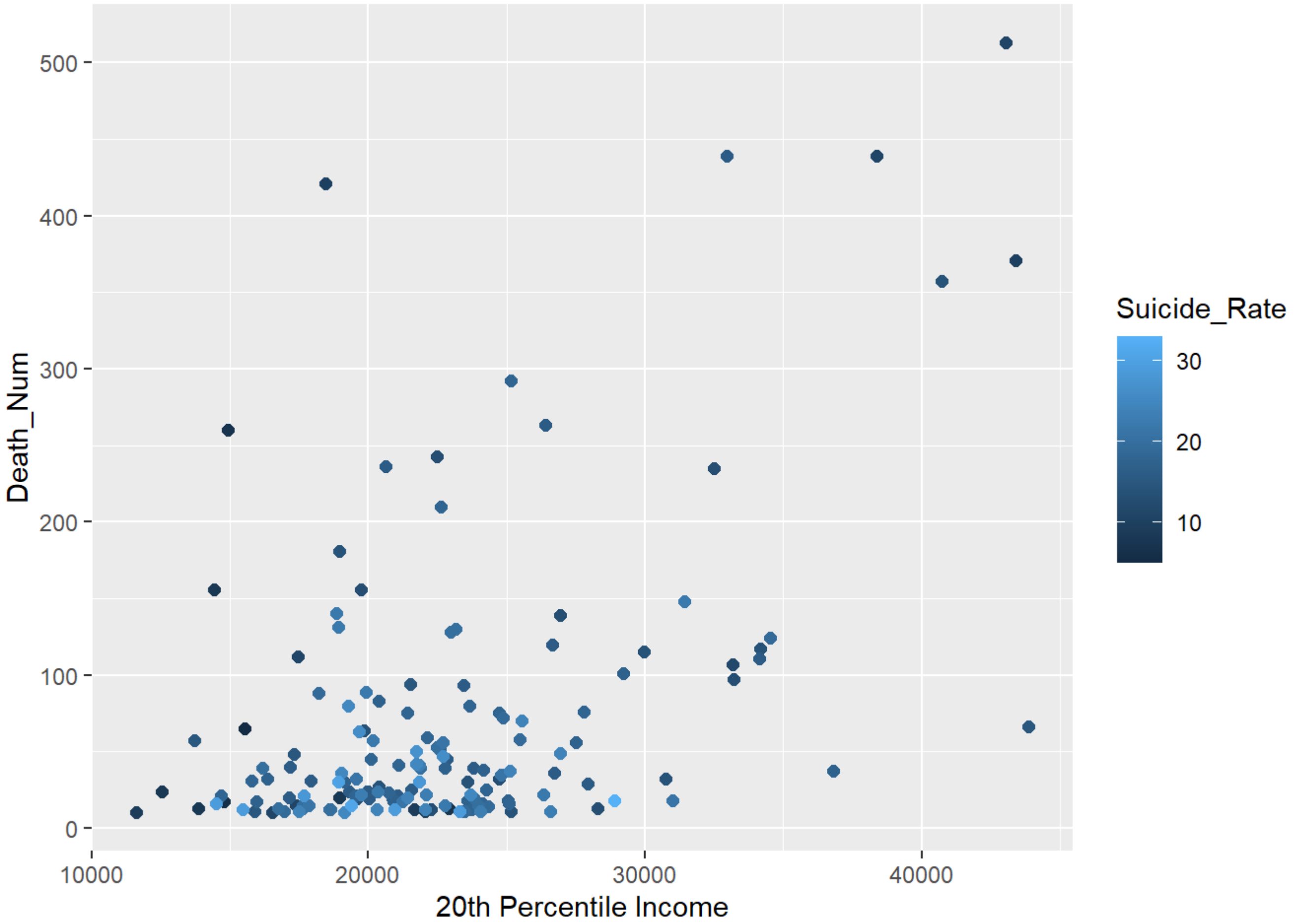
# Exploratory Data Analysis

```
stat = mydata %>%
  summarize(Mean_Suicide =
              mean(mydata$Suicide_Rate),
            Sd_Suicide =
              sd(mydata$Suicide_Rate),
            Min_Suicide =
              min(mydata$Suicide_Rate),
            Max_Suicide =
              max(mydata$Suicide_Rate),
            Med_Suicide =
              median(mydata$Suicide_Rate),
            Num_County =
              n(),
            Var_Suicide =
              var(mydata$Suicide_Rate),
            Mad_Suicide =
              mad(mydata$Suicide_Rate))

stat
```

```
## # A tibble: 1 x 8
##   Mean_Suicide Sd_Suicide Min_Suicide Max_Suicide Med_Suicide Num_County
##          <dbl>      <dbl>       <dbl>       <dbl>       <dbl>      <int>
## 1         17.7       5.40           5          33          17        159
## # ... with 2 more variables: Var_Suicide <dbl>, Mad_Suicide <dbl>
```

*The unit of Suicide rate is (%), and the mean, std, min, max, and median are 17.7, 5.4, 5, 33, and 17 %. None of the relationship was actually observed yet.*

```
library(ggplot2)
mydata %>%
  filter(Death_Num <= 700) %>%
  ggplot(aes(x = `20th Percentile Income`,
             y = Death_Num,
             color = Suicide_Rate)) +
  geom_point(size = 2)
```



*This is the scattor plot visualization between Death number and 20th percentile income corresponding to Suicide rate.*

##MANOVA

```
manova1 = manova(cbind(Death_Num,
                       Household_Income,
                       Inequality) ~ Suicide_Rate,
                 data = mydata)

summary(manova1)
```

```
##                Df  Pillai approx F num Df den Df   Pr(>F)
## Suicide_Rate    1 0.08995   5.1068      3    155 0.002135 **
## Residuals     157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Univariate ANOVA
summary.aov(manova1)
```

```
##  Response Death_Num :
##               Df   Sum Sq Mean Sq F value    Pr(>F)    
## Suicide_Rate   1   923589  923589  14.326 0.0002184 ***
## Residuals    157 10121864   64470                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##  Response Household_Income :
##               Df     Sum Sq   Mean Sq F value  Pr(>F)  
## Suicide_Rate   1 5.0019e+08 500185806  3.0587 0.08226 .
## Residuals    157 2.5674e+10 163528076                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##  Response Inequality :
##               Df     Sum Sq    Mean Sq F value  Pr(>F)  
## Suicide_Rate   1 1.0083e+09 1008253552  4.4862 0.03574 *
## Residuals    157 3.5285e+10  224745387                  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mydata %>%
  summarize(mean(Death_Num),
            mean(Household_Income),
            mean(Inequality))
```

```
## # A tibble: 1 x 3
##   `mean(Death_Num)` `mean(Household_Income)` `mean(Inequality)`
##               <dbl>                    <dbl>              <dbl>
## 1              110.                   53406.             81765.
```

```
pairwise.t.test(mydata$Death_Num,
                mydata$suicide_con,
                p.adj="none")
```

```
## 
##  Pairwise comparisons using t tests with pooled SD 
## 
## data:  mydata$Death_Num and mydata$suicide_con 
## 
##     high   
## low 1.8e-05
## 
## P value adjustment method: none
```

```
pairwise.t.test(mydata$Household_Income,
                mydata$suicide_con,
                p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mydata$Household_Income and mydata$suicide_con
##
##     high
## low 0.047
##
## P value adjustment method: none
```

```
pairwise.t.test(mydata$Inequality,
                mydata$suicide_con,
                p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  mydata$Inequality and mydata$suicide_con
##
##     high
## low 0.037
##
## P value adjustment method: none
```

```
1-.95^(7)
```

```
## [1] 0.3016627
```

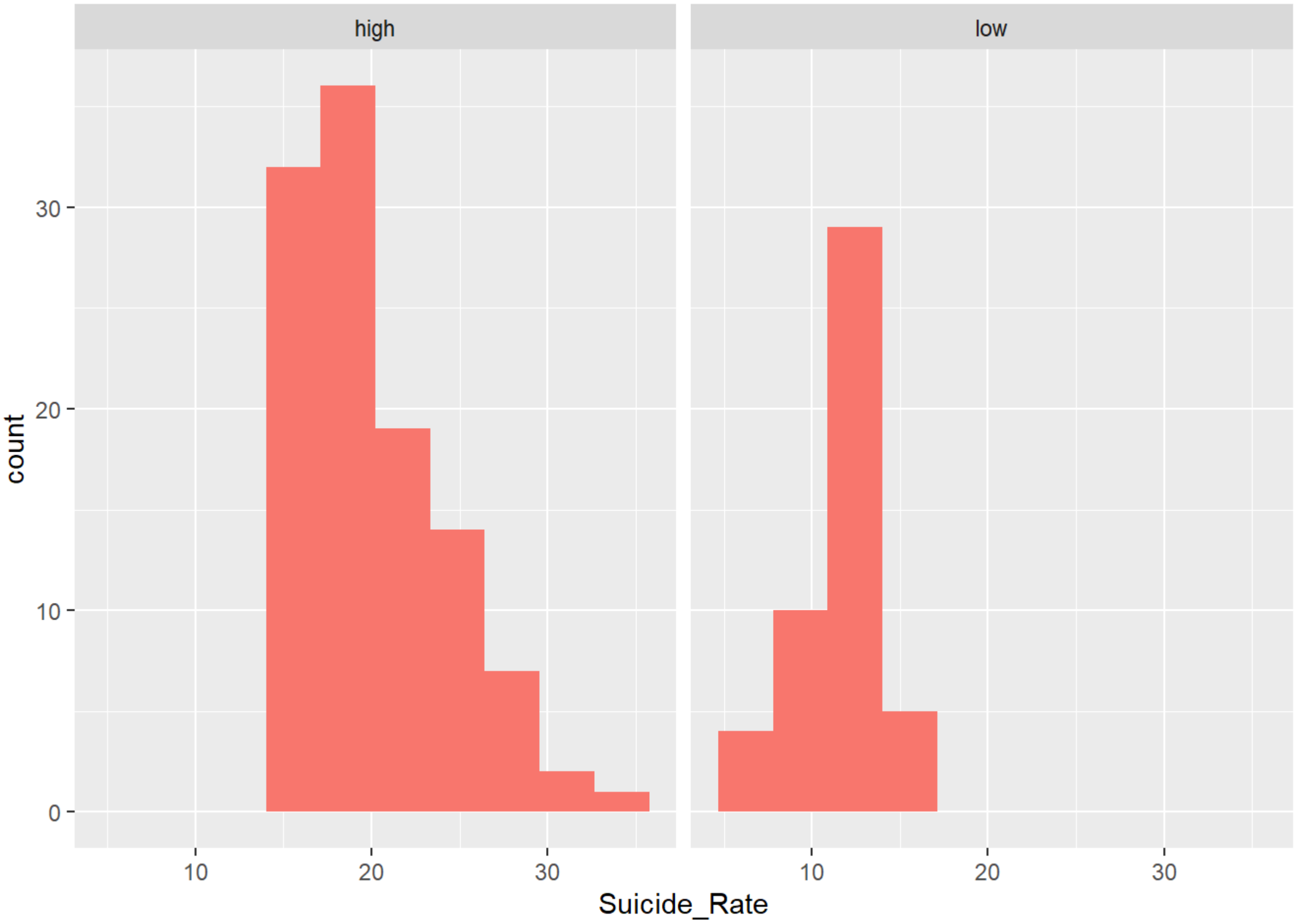
```
0.05/7
```

```
## [1] 0.007142857
```

*1 MANOVA, 3 ANOVA, and 3 t-tests, they are total of 7 tests. The probability of one type 1 error is 0.302. The adjusted significance level was calculated to be 0.05/7 = 0.007. I had to create new variable called suicide_con which represents whether the suicide rate is low or high. The standard point of determining suicide_con was 15. Random sample assumption is most likely not met since the samples were collected.*

##Randomization Test

```
library(ggplot2)
ggplot(mydata, aes(Suicide_Rate, fill = State)) +
  geom_histogram(bins=10) +
  facet_wrap(~suicide_con, ncol = 2) +
  theme(legend.position = "none")
```



```
perm = data.frame(suicide_con = mydata$suicide_con,
                  Suicide_rate = sample(mydata$Suicide_Rate))

perm %>%
  group_by(suicide_con) %>%
  summarize(means = mean(Suicide_rate)) %>%
  summarize(`mean_diff:` = diff(means))
```

```
## # A tibble: 1 x 1
##   `mean_diff:`
##          <dbl>
## 1      -0.0991
```

```
t.test(data = mydata, Inequality ~ suicide_con)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Inequality by suicide_con
## t = -1.8245, df = 67.255, p-value = 0.07251
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11400.4930    511.2857
## sample estimates:
## mean in group high  mean in group low 
##           80121.06           85565.67
```

*Null hypothesis: mean inequality of income is the same for low vs. high percentage of Suicide rate. Alternative hypothesis: mean inequality of income is different for low vs. high percentage of Suicide rate. We cannot reject the null hypothesis as the mean inequality of income is the same between low and high percentage of suicide rate.*

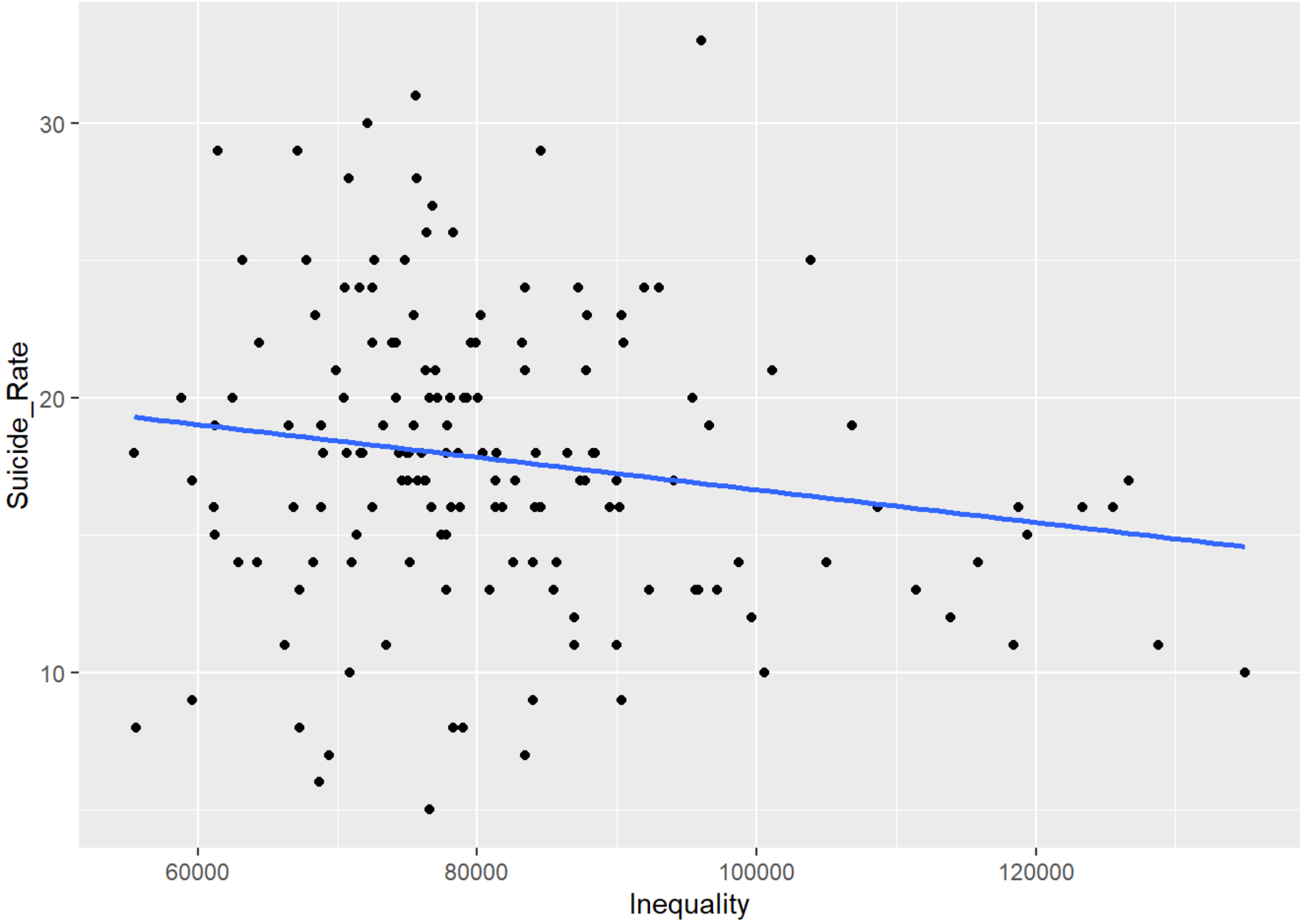
##Linear regression model

```
fit = lm(Suicide_Rate ~ Inequality + Household_Income, data = mydata)

summary(fit)
```

```
## 
## Call:
## lm(formula = Suicide_Rate ~ Inequality + Household_Income, data = mydata)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.9025  -3.2945  -0.2522   3.1692  16.0983 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       2.289e+01  2.458e+00   9.315   <2e-16 ***
## Inequality       -8.353e-05  6.709e-05  -1.245    0.215    
## Household_Income  3.134e-05  7.900e-05   0.397    0.692    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.353 on 156 degrees of freedom
## Multiple R-squared:  0.02876,    Adjusted R-squared:  0.01631 
## F-statistic:  2.31 on 2 and 156 DF,  p-value: 0.1027
```

```
qplot(x = Inequality, y = Suicide_Rate, data = mydata)+
  stat_smooth(method = "lm", se = FALSE,
              fullrange = TRUE)
```

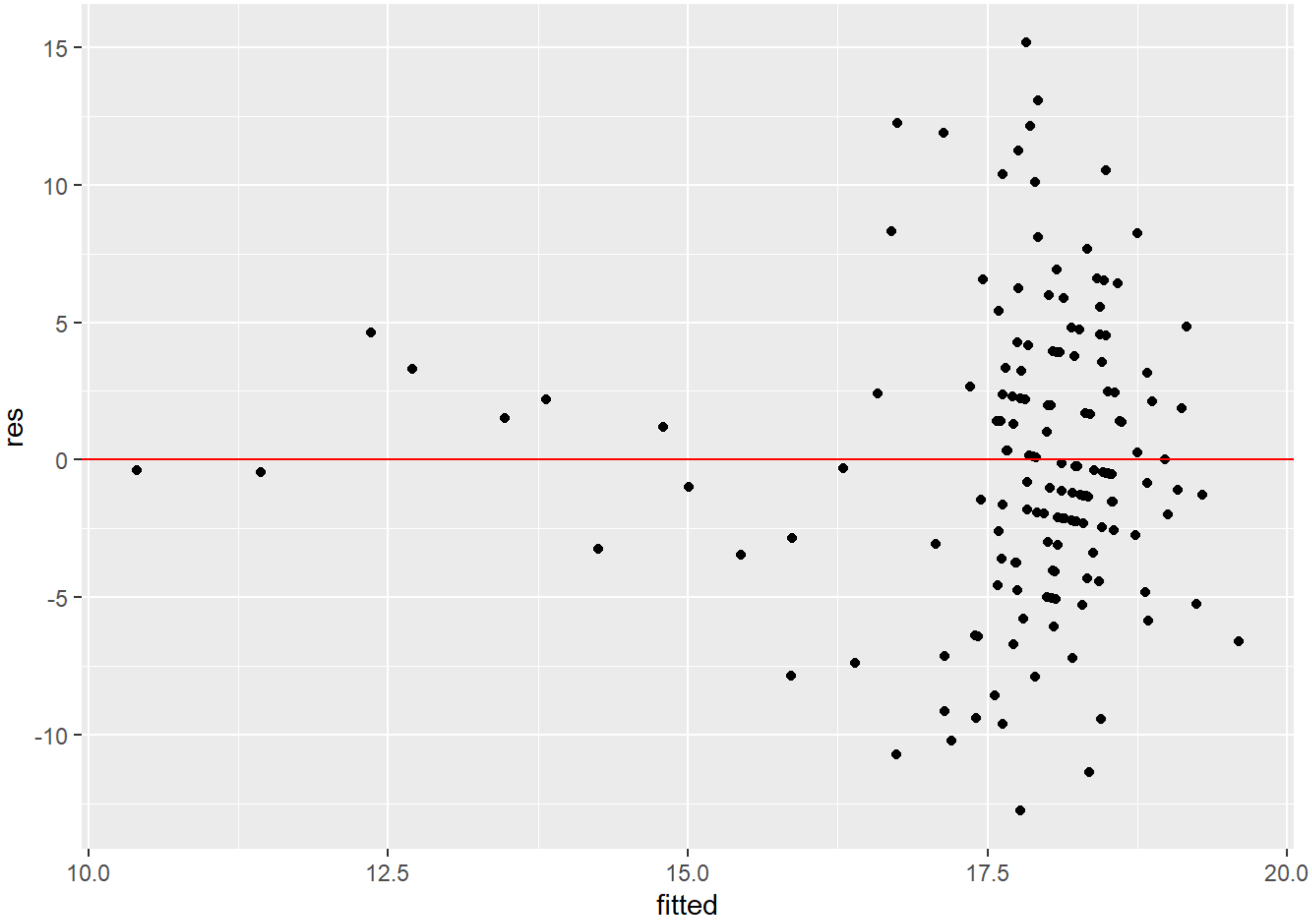
```
## `geom_smooth()` using formula 'y ~ x'
```



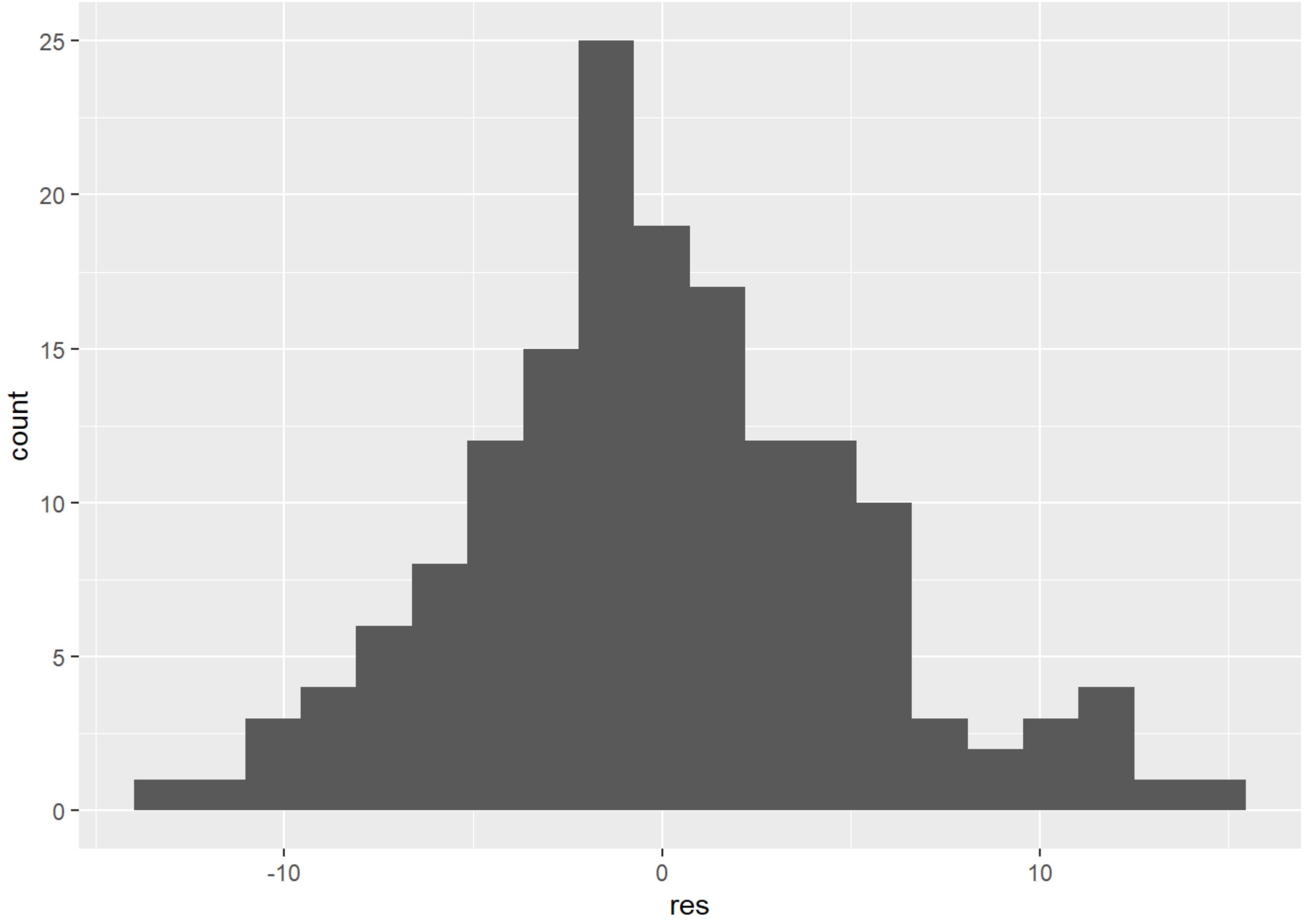
```
mydata1 = mydata %>%
  mutate(Inequality_m = Inequality - mean(Inequality),
         Household_Income_m = Household_Income -
           mean(Household_Income))

fit1 = lm(Suicide_Rate ~ Inequality_m * Household_Income_m, data = mydata1)

summary(fit1)
```

```
##
## Call:
## lm(formula = Suicide_Rate ~ Inequality_m * Household_Income_m,
##     data = mydata1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7739  -3.0756  -0.4635   3.1927  15.1728
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.840e+01  5.066e-01  36.316   <2e-16 ***
## Inequality_m                    -7.838e-05  6.620e-05  -1.184   0.2383
## Household_Income_m               1.017e-04  8.359e-05   1.216   0.2258
## Inequality_m:Household_Income_m -3.763e-09  1.620e-09  -2.322   0.0215 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.279 on 155 degrees of freedom
## Multiple R-squared:  0.06141,    Adjusted R-squared:  0.04324
## F-statistic:  3.38 on 3 and 155 DF,  p-value: 0.01984
```

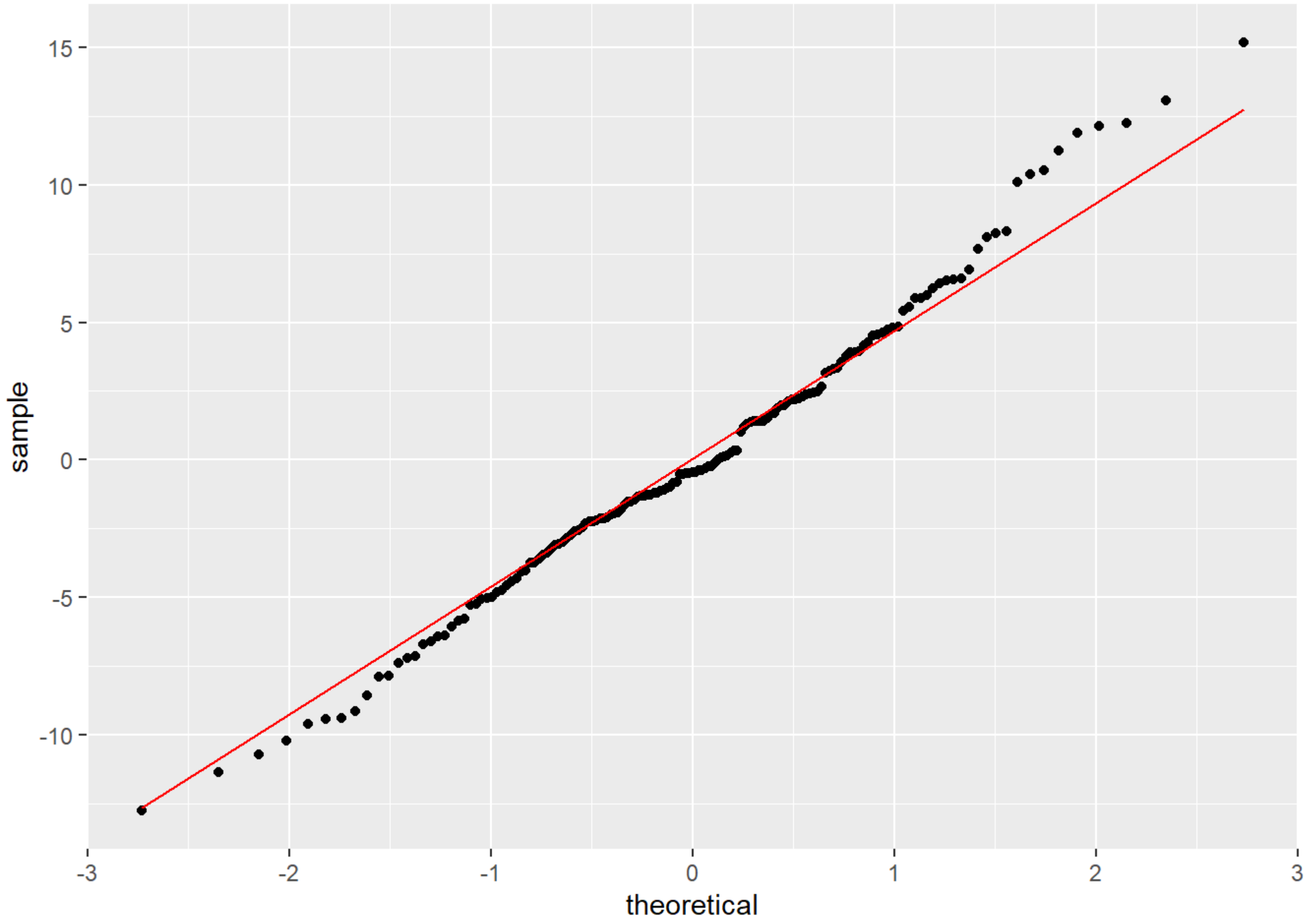
```
res = fit1$residuals
fitted = fit1$fitted.values

ggplot() +
  geom_point(aes(fitted, res)) +
  geom_hline(yintercept = 0,
             col = "red")
```



```
#Normally Distributed
ggplot() +
  geom_histogram(aes(res), bins = 20)
```



```
ggplot() +
  geom_qq(aes(sample=res)) +
  geom_qq_line(aes(sample = res),
               color = 'red')
```



```
#hypothesis test
ks.test(res, "pnorm", mean = 0, sd(res))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  res
## D = 0.065443, p-value = 0.5037
## alternative hypothesis: two-sided
```

```
#robust SEs
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```
summary(fit1)$coef[,1:2]
```

```
##                                   Estimate   Std. Error
## (Intercept)                   1.839816e+01 5.066159e-01
## Inequality_m                 -7.837854e-05 6.620172e-05
## Household_Income_m            1.016644e-04 8.359057e-05
## Inequality_m:Household_Income_m -3.762677e-09 1.620469e-09
```

```
summary(fit1)$r.sq
```

```
## [1] 0.06140884
```

*While controlling for Inequality, household income does not explain variation and vice versa. 2.289 is the predicted value of Suicide rate when inequality and household income are zero. The slope for inequality and household income are -8.353e-05 and 3.134e-05, respectively. After the robust SEs, intercept was 1.839. inequality_m was 0.507. Household income_m was 8.359e-5. Their ratio was 1.6205e-9.*

##Logistic Regression

```
class_diag<-function(probs,truth){
  tab<-table(factor(probs>.5,
                    levels=c("FALSE","TRUE")),truth)
  acc=sum(diag(tab))/sum(tab)
  sens=tab[2,2]/colSums(tab)[2]
  spec=tab[1,1]/colSums(tab)[1]
  ppv=tab[2,2]/rowSums(tab)[2]
  if(is.numeric(truth)==FALSE & is.logical(truth)==FALSE)
    truth<-as.numeric(truth)-1
  ord<-order(probs, decreasing=TRUE)
  probs <- probs[ord]; truth <- truth[ord]

             TPR=cumsum(truth)/max(1,sum(truth))

             FPR=cumsum(!truth)/max(1,sum(!truth))
             dup<-c(probs[-1]>=probs[-length(probs)],
                    FALSE)
             TPR<-c(0,TPR[!dup],1);
             FPR<-c(0,FPR[!dup],1)
             n <- length(TPR)
             auc<- sum( ((TPR[-1]+TPR[-n])/2) * (FPR[-1]-FPR[-n]) )
             data.frame(acc,sens,spec,ppv,auc)}
```

```
logit = function(x)log(odds(x))
mydata2 = mydata %>%
  mutate(bin = ifelse(suicide_con == "high", 1, 0))

logfit = glm(bin ~ Household_Income + Inequality,
             data = mydata2,
             family = binomial(link = "logit"))

coef(logfit)
```

```
##      (Intercept) Household_Income       Inequality
##     2.670357e+00    -5.662162e-06    -1.845940e-05
```

```
exp(coef(logfit))
```

```
##      (Intercept) Household_Income       Inequality
##       14.4451190        0.9999943        0.9999815
```

```
mydata2$prob = predict(logfit, type = "response")

mydata2$predicted = ifelse(mydata2$prob > .5, 1, 0)

table(truth = mydata2$bin,
      prediction = mydata2$predicted) %>%
  addmargins()
```

```
##      prediction
## truth   0   1 Sum
##   0     4  44  48
##   1     3 108 111
##   Sum   7 152 159
```

```
#Accuracy
(4+108) / 159
```

```
## [1] 0.7044025
```

```
#Sensitivity(TPR)
108/152
```

```
## [1] 0.7105263
```

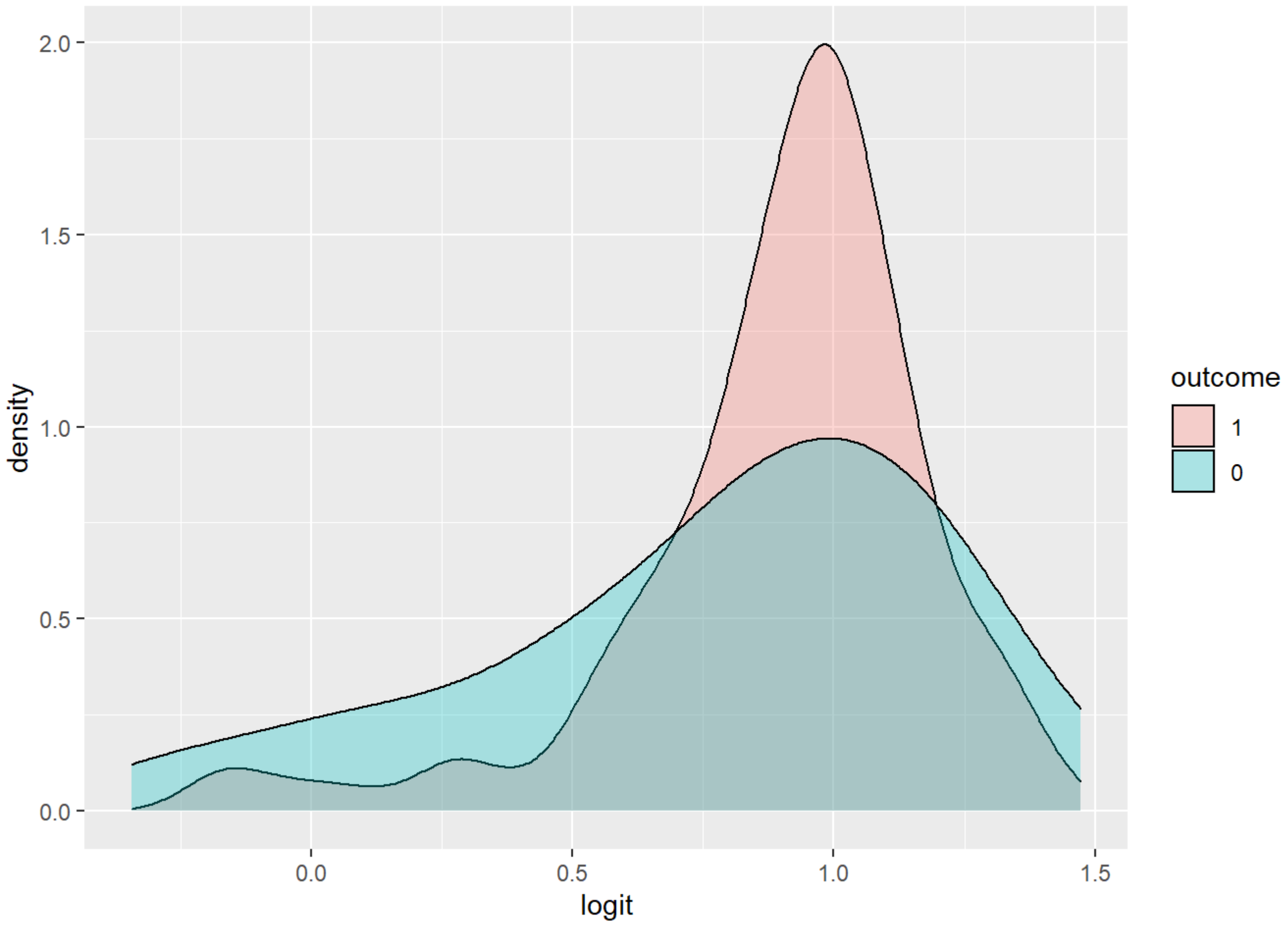
```
#Specificity(TNR)
4/7
```

```
## [1] 0.5714286
```

```
#Precision(PPV)
108/111
```

```
## [1] 0.972973
```

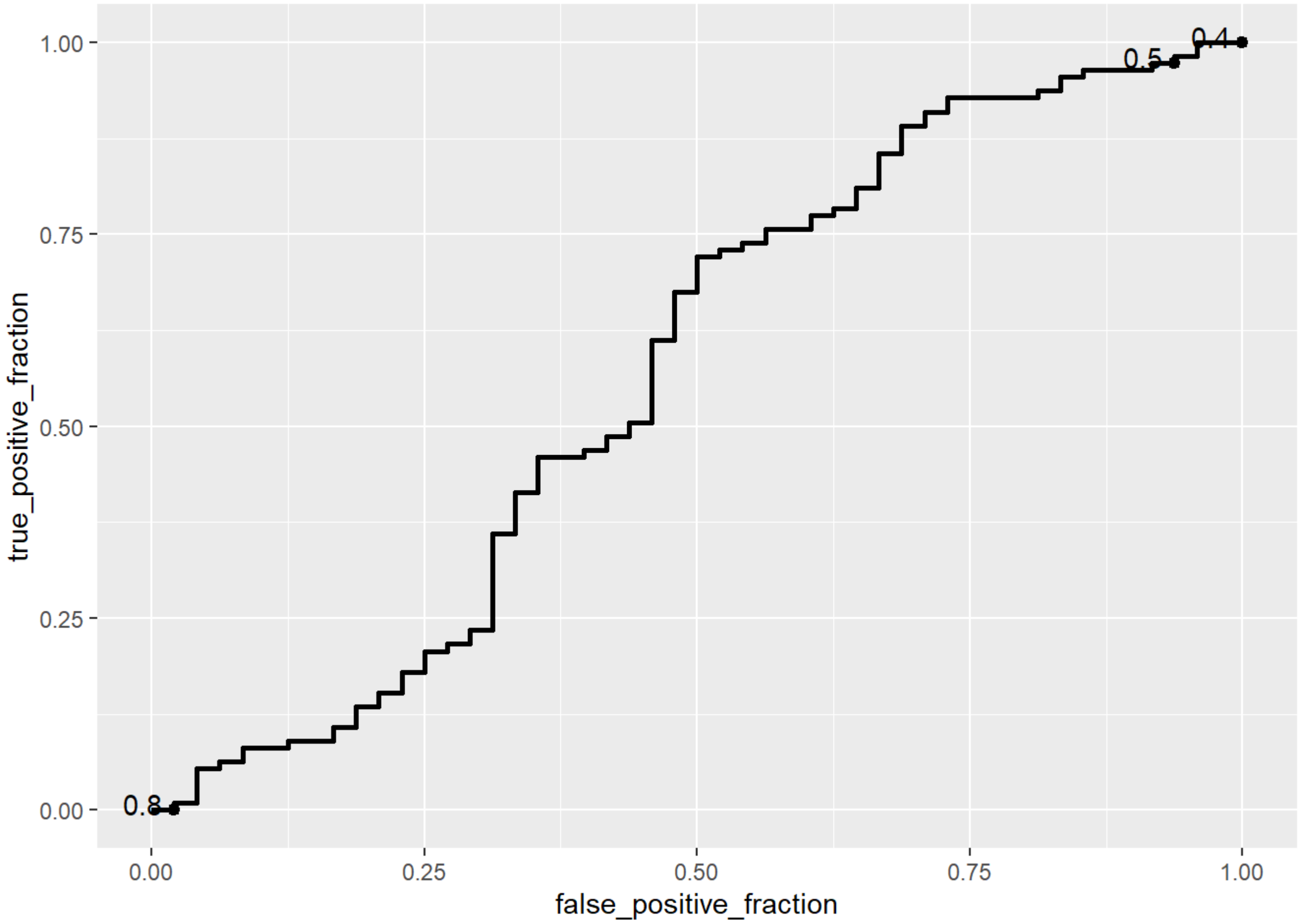
```
#Density plot of log-odds
mydata2$logit = predict(logfit)
mydata2$outcome = factor(mydata2$bin,
                         levels = c(1,0))
ggplot(mydata2, aes(logit, fill = outcome)) +
  geom_density(alpha=0.3)
```



```
#ROC
sens <- function(p, data = mydata2, y = y) mean(mydata2[mydata2$bin == 1, ]$prob > p)
spec <- function(p, mydata2 = data, y = y) mean(mydata2[mydata2$bin == 0, ]$prob <= p)

sensitivity <- sapply(seq(0,1,.01),sens,mydata2)
specificity<-sapply(seq(0,1,.01),spec,mydata2)
MyROC <- data.frame(sensitivity, specificity, cutoff = seq(0,1,.01))
MyROC$TPR <- sensitivity
MyROC$FPR <- 1-specificity


# Apply the functions to our data
sensitivity <- sapply(seq(0,1,.01),sens,mydata2)
specificity<-sapply(seq(0,1,.01),spec,mydata2)
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 4.0.5
```

```
ROCplot1 = ggplot(mydata2) +
  geom_roc(aes(d = bin, m = prob),
           cutoffs.at = list(0.1, 0.5, 0.9))

ROCplot1
```



```
#AUC
MyROC = MyROC %>%
  arrange(FPR)

widths = diff(MyROC$FPR)
heights = vector()
for(i in 1:100) heights[i] = MyROC$TPR[i] +
  MyROC$TPR[i+1]

AUC = sum(heights * widths / 2)

AUC
```

```
## [1] 0.5653153
```

```
calc_auc(ROCplot1)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5692568
```

*AUC is 0.569 (not good), which is really hard to predict whether a suicide rate is come from only household income and income inequality.*