ACTIVITY4~

Madeleine Schoderbek

2025-12-03

R Markdown

#QUESTION 1 A)

```
library(readxl)
library(openxlsx)

rip <- read_excel("SuicideRisk_Data-2.xlsx")

library(tidyr)
library(readxl)
library(broom)
library(car)

## Loading required package: carData

library(performance)
library(ggplot2)
library(kableExtra)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:kableExtra':
##
##     group_rows

## The following object is masked from 'package:car':
##
##     recode

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

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

library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-6   2023-06-27
```

```
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
sapply(rip[, c("GENDER", "RACE", "ETHNICITY", "INCOME")], function(x) length(
unique(x)))
```

```
##    GENDER      RACE ETHNICITY    INCOME
##         2         3         2         5
```

```
rip <- rip %>%
  mutate(
    GENDER = factor(GENDER, levels = c("Female", "Male")),
    RACE = factor(RACE, levels = c("White/Caucasian", "Other")),
    ETHNICITY = factor(ETHNICITY, levels = c("Not Hispanic/Latino", "Hispanic
/Latino")),
    INCOME = factor(INCOME),

    across(starts_with("ACES___"), ~factor(.x, levels = c(0, 1), labels = c("
No", "Yes"))),

    HX_SUICIDE = factor(HX_SUICIDE, levels = c(0,1), labels = c("No", "Yes"))
  )



   str(rip)
```

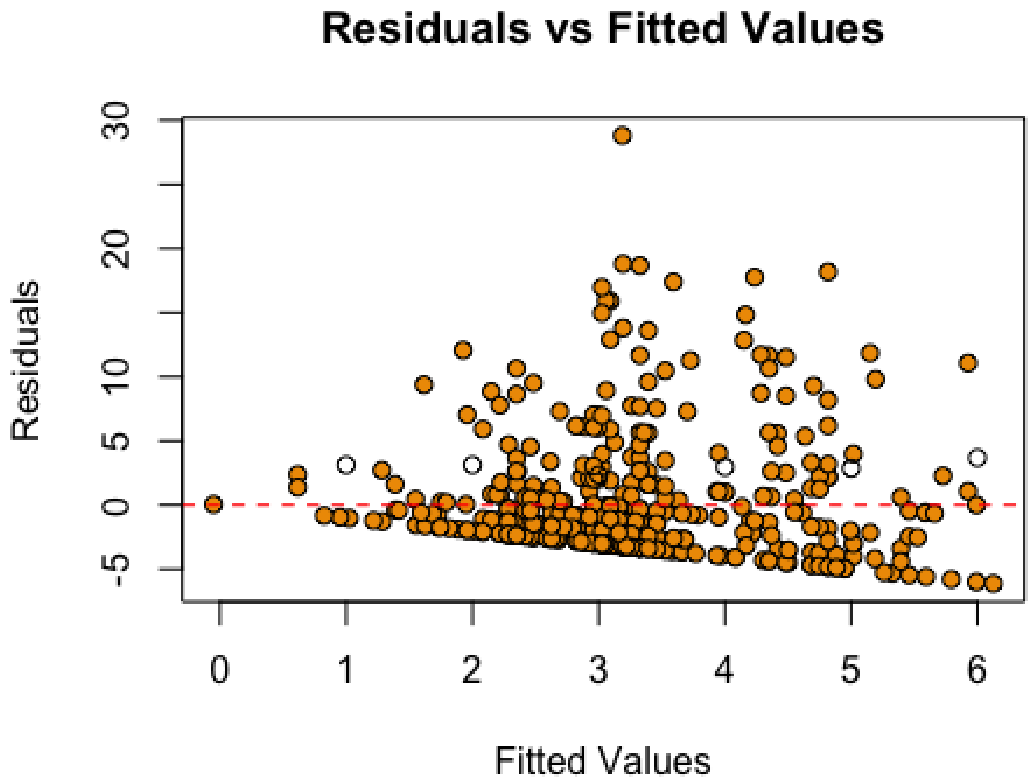
```
## tibble [546 × 19] (S3: tbl_df/tbl/data.frame)
##  $ RECORD_ID      : num [1:546] 1 4 7 10 12 14 15 17 19 20 ...
##  $ AGE            : num [1:546] 24 25 21 23 21 22 28 36 24 27 ...
##  $ GENDER         : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 2 1
1 ...
##  $ RACE           : Factor w/ 2 levels "White/Caucasian",..: 1 1 1 NA NA 1
1 1 1 NA ...
##  $ ETHNICITY      : Factor w/ 2 levels "Not Hispanic/Latino",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ INCOME         : Factor w/ 5 levels "< $30,000",">$100,000",..: 2 3 4 4
3 5 2 1 5 4 ...
```

```
##  $ ACES___1       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACES___2       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ ACES___3       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ ACES___4       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACES___5       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACES___6       : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 1 1 ...
##  $ ACES___7       : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...
##  $ ACES___8       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ ACES___9       : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 2 1 2 1 1 ...
##  $ ACES___10      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ CESDR_TOTAL_SUM: num [1:546] 8 15 2 20 9 2 18 0 2 9 ...
##  $ HX_SUICIDE     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ SABCS_TOTAL_SUM: num [1:546] 0 0 0 0 1 0 5 1 0 2 ...
```

```
model_demo <- lm(
  SABCS_TOTAL_SUM ~ AGE + GENDER + RACE + ETHNICITY + INCOME,
                 data = rip
  )

sabcs_model <- lm(SABCS_TOTAL_SUM ~ AGE + GENDER + RACE + ETHNICITY + INCOME,
data = rip)
summary(sabcs_model)
```

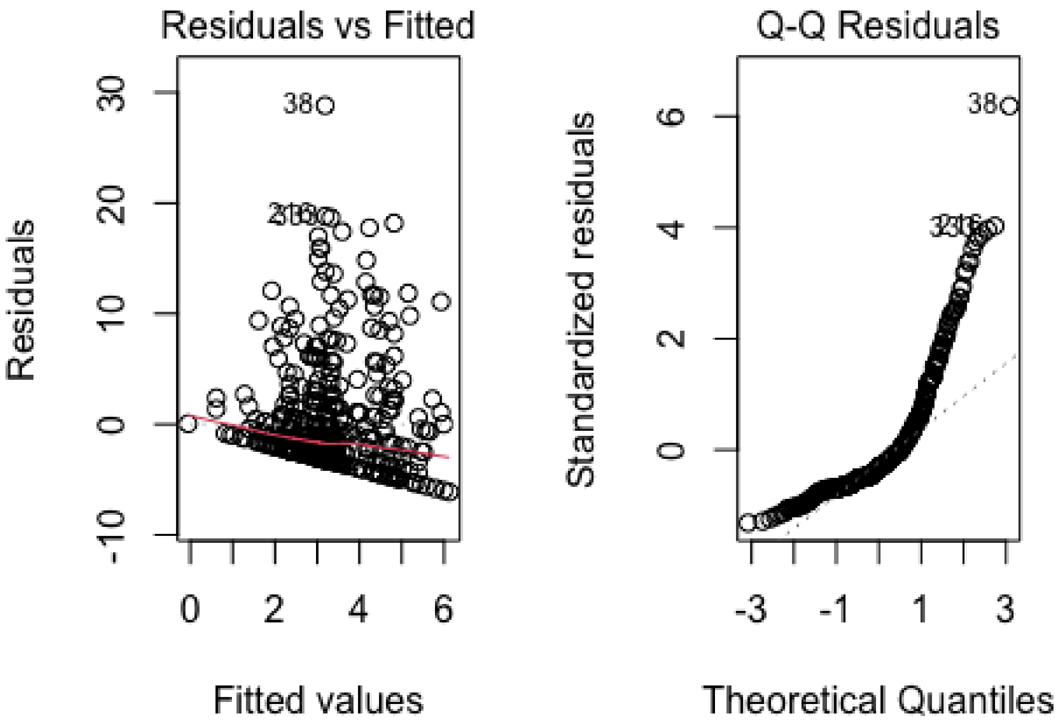
```
##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ AGE + GENDER + RACE + ETHNICITY +
##     INCOME, data = rip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1265 -2.9264 -1.5258  0.8287 28.8121
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.21777    1.01836   6.106 2.13e-09 ***
## AGE                        -0.06669    0.03276  -2.035  0.04236 *
## GENDERMale                 -0.13858    0.76471  -0.181  0.85627
## RACEOther                   0.26626    0.56837   0.468  0.63966
## ETHNICITYHispanic/Latino    0.90949    0.65573   1.387  0.16610
## INCOME>$100,000            -1.49091    0.68380  -2.180  0.02972 *
## INCOME$30,000 - $50,000    -1.42420    0.72772  -1.957  0.05093 .
## INCOME$51,000 - $75,000    -2.46747    0.69709  -3.540  0.00044 ***
## INCOME$76,000 - $100,000 -1.79142    0.69590  -2.574  0.01035 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.73 on 474 degrees of freedom
##   (63 observations deleted due to missingness)
## Multiple R-squared:  0.04189,    Adjusted R-squared:  0.02572
## F-statistic:  2.59 on 8 and 474 DF,  p-value: 0.008892
```

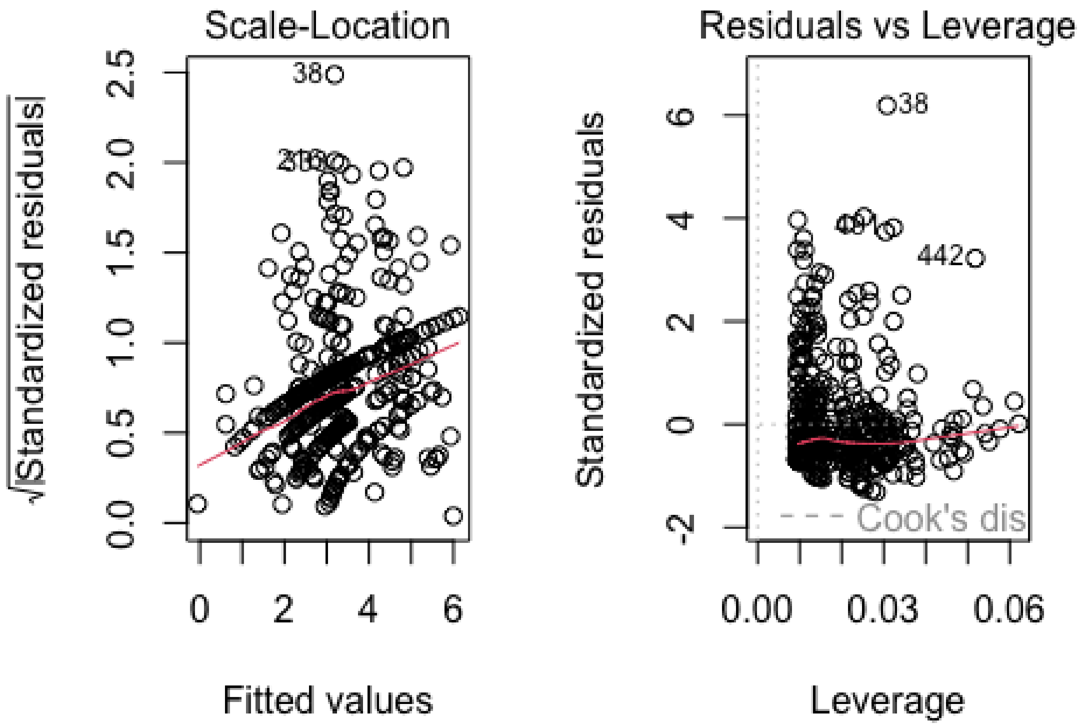
```r
plot(sabcs_model$fitted.values, resid(sabcs_model),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 21,
     bg = "orange2",
     col = "black")

points(x = sabcs_model$fitted.values)

abline(h = 0, col = "red", lty = 2)
```



```r
library(car)
vif(sabcs_model)

##                 GVIF Df GVIF^(1/(2*Df))
## AGE       1.068228  1        1.033551
## GENDER    1.023973  1        1.011916
## RACE      1.066518  1        1.032723
## ETHNICITY 1.095241  1        1.046538
## INCOME    1.128244  4        1.015197
```

```r
par(mfrow = c(1, 2))
plot(sabcs_model)
```

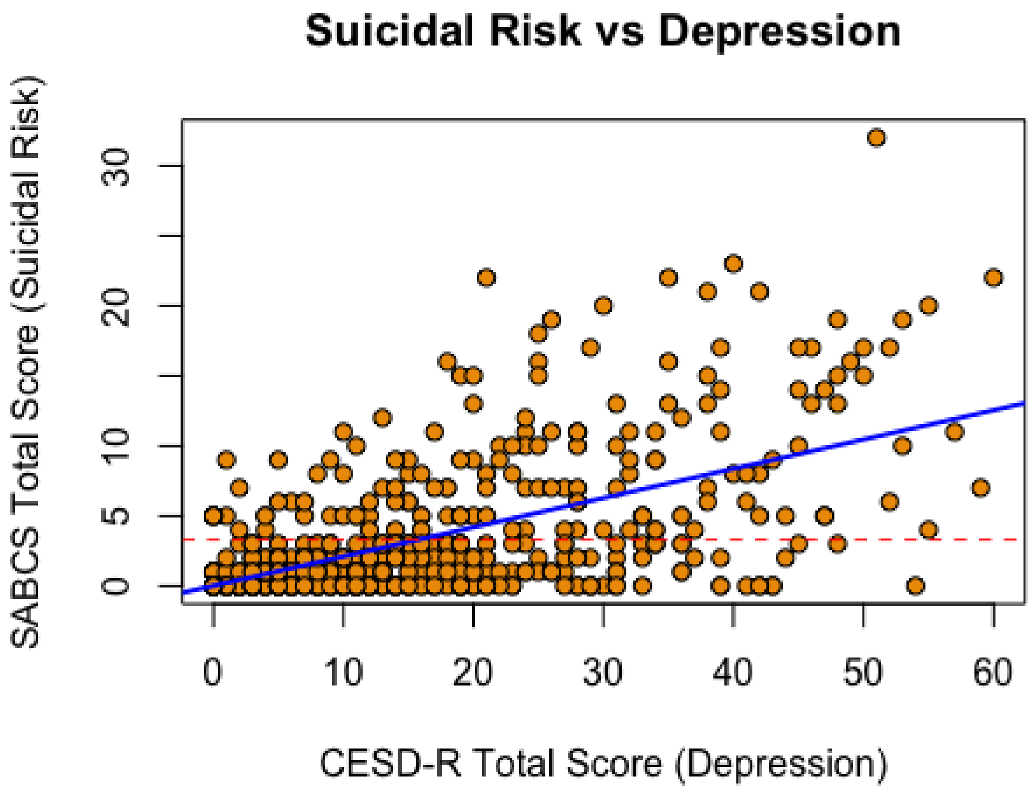

Residuals vs Fitted
Residuals
30
20
10
0
-10
38
0
2
4
6
Fitted values
Q-Q Residuals
Standardized residuals
6
4
2
0
38
-3
-1
1
3
Theoretical Quantiles



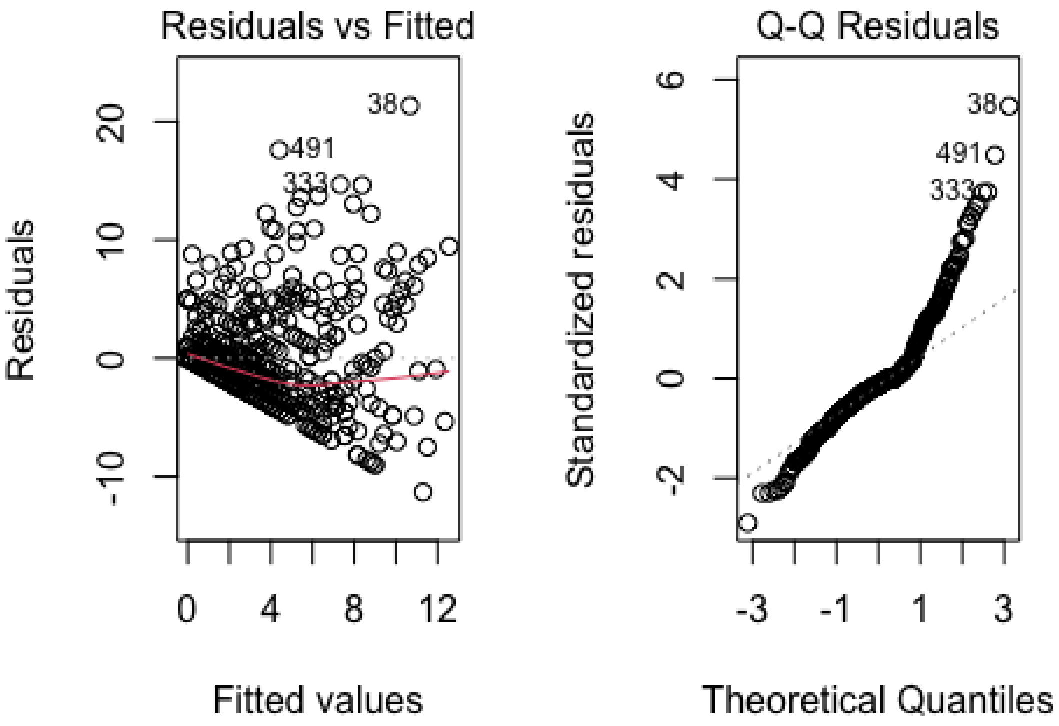
From the results comparing SABCS to demographic variables, the model is statistically significant but only relying on demographics to determine suicide risk would not be reliable. A few of the individuals have a much higher risk, but most individuals have a low SABCS score. The model represents that age and income are the more significant than the other factors when predicting suicide (higher age, higher income, lower risk of suicide). In sum, other factors would be a better predictor for suicide risk such as depression.

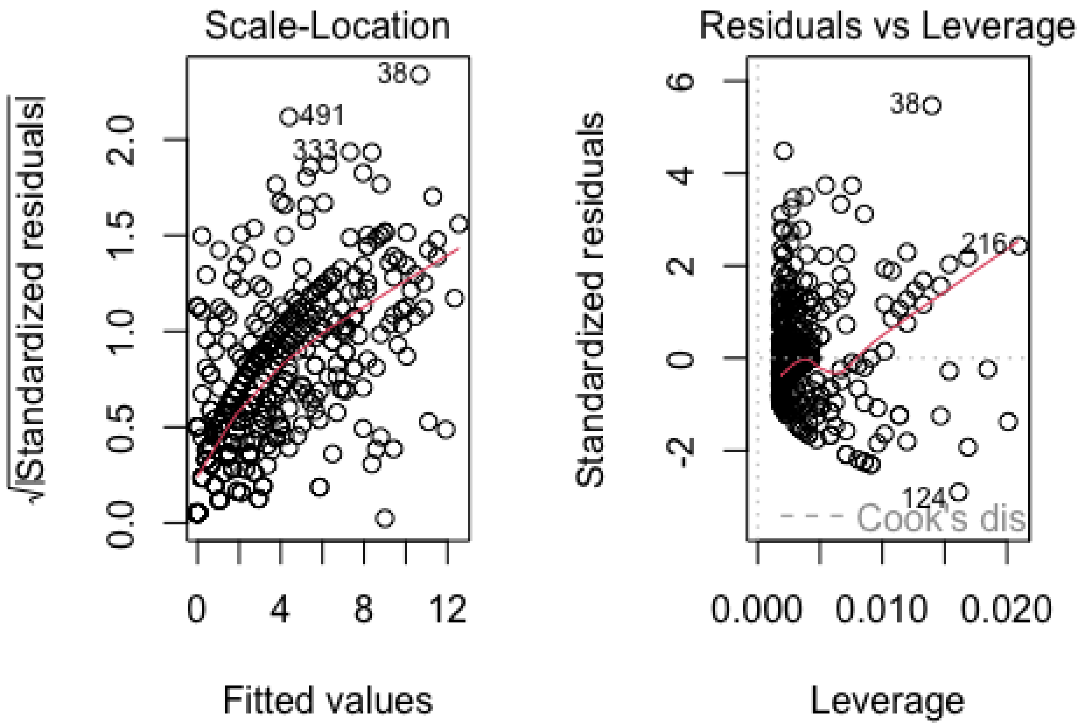
```
model_depression <- lm(SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = rip)


summary(model_depression)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = rip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2968  -2.1009  -0.4739   0.9891  21.3302
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.01088    0.25650   0.042    0.966    
## CESDR_TOTAL_SUM  0.20900    0.01228  17.013   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.93 on 544 degrees of freedom
## Multiple R-squared:  0.3473, Adjusted R-squared:  0.3461
## F-statistic: 289.4 on 1 and 544 DF,  p-value: < 2.2e-16
```

```r
plot(rip$CESDR_TOTAL_SUM, rip$SABCS_TOTAL_SUM,
     xlab = "CESD-R Total Score (Depression)",
     ylab = "SABCS Total Score (Suicidal Risk)",
     main = "Suicidal Risk vs Depression",
     pch = 21,
     bg = "orange2",
     col = "black")


abline(model_depression, col = "blue", lwd = 2)


abline(h = mean(rip$SABCS_TOTAL_SUM, na.rm = TRUE), col = "red", lty = 2)
```



```r
par(mfrow=c(1,2))
plot(model_depression)
```


Residuals vs Fitted
Residuals
Fitted values
20
10
0
-10
0
4
8
12
38
491
333
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
6
4
2
0
-2
-3
-1
1
3
38
491
333



This model shows that higher depression scores are associated with higher suicide rates. The relationship between the two is statistically significant, being that 35% of the variation in suicide risk scores is caused by the depression scores.

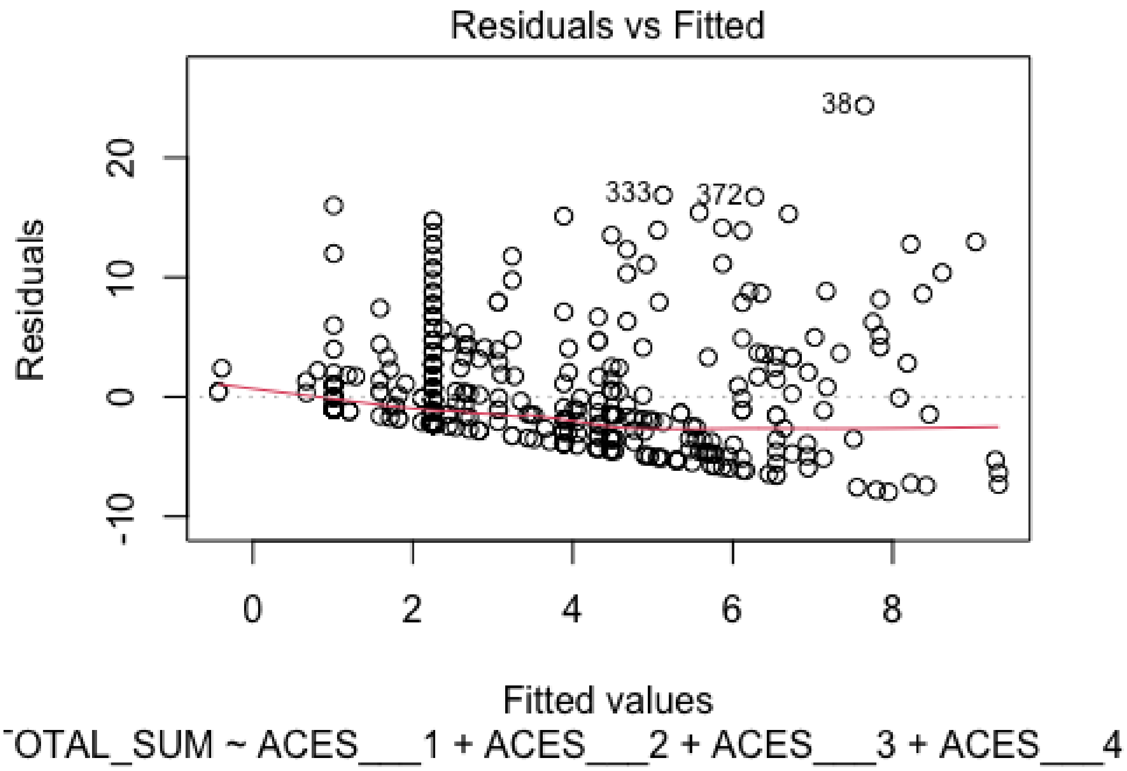
```
sapply(rip[, paste0("ACES___", 1:10)], function(x) table(x))

##     ACES___1 ACES___2 ACES___3 ACES___4 ACES___5 ACES___6 ACES___7 ACES___
8
## No       480      451      388      528      482      460      470      42
7
## Yes       66       95      158       18       64       86       76      11
9
##     ACES___9 ACES___10
## No       390       519
## Yes      156        27

aces_vars <- paste0("ACES___", 1:10)
rip[, aces_vars] <- lapply(rip[, aces_vars], function(x) {
  as.numeric(factor (x, levels = c("No", "Yes"), labels = c (0,1)))
})

sapply(rip[, aces_vars], table)
```

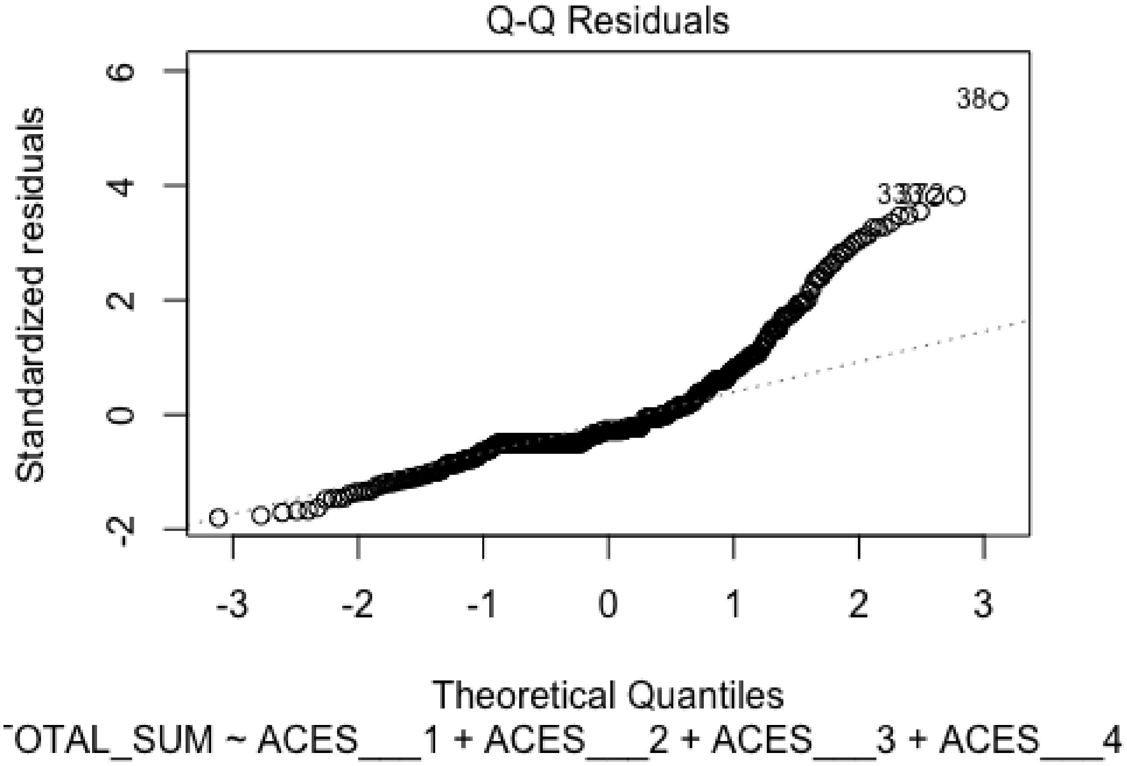
```
##   ACES___1 ACES___2 ACES___3 ACES___4 ACES___5 ACES___6 ACES___7 ACES___8
## 1      480      451      388      528      482      460      470      427
## 2       66       95      158       18       64       86       76      119
##   ACES___9 ACES___10
## 1      390       519
## 2      156        27
```

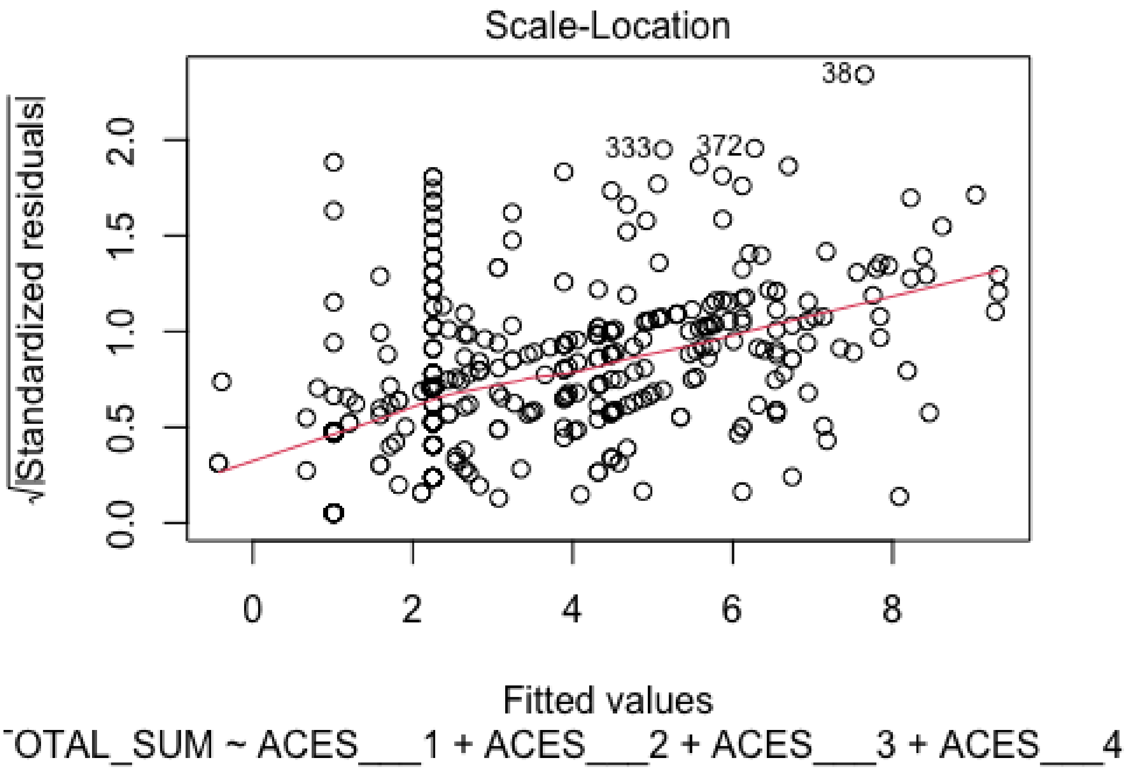
```
model_aces <- lm(SABCS_TOTAL_SUM ~ ACES___1 + ACES___2 + ACES___3 + ACES___4
+ ACES___5 + ACES___6 + ACES___7 + ACES___8 + ACES___9 + ACES___10, data = ri
p)


summary(model_aces)
```

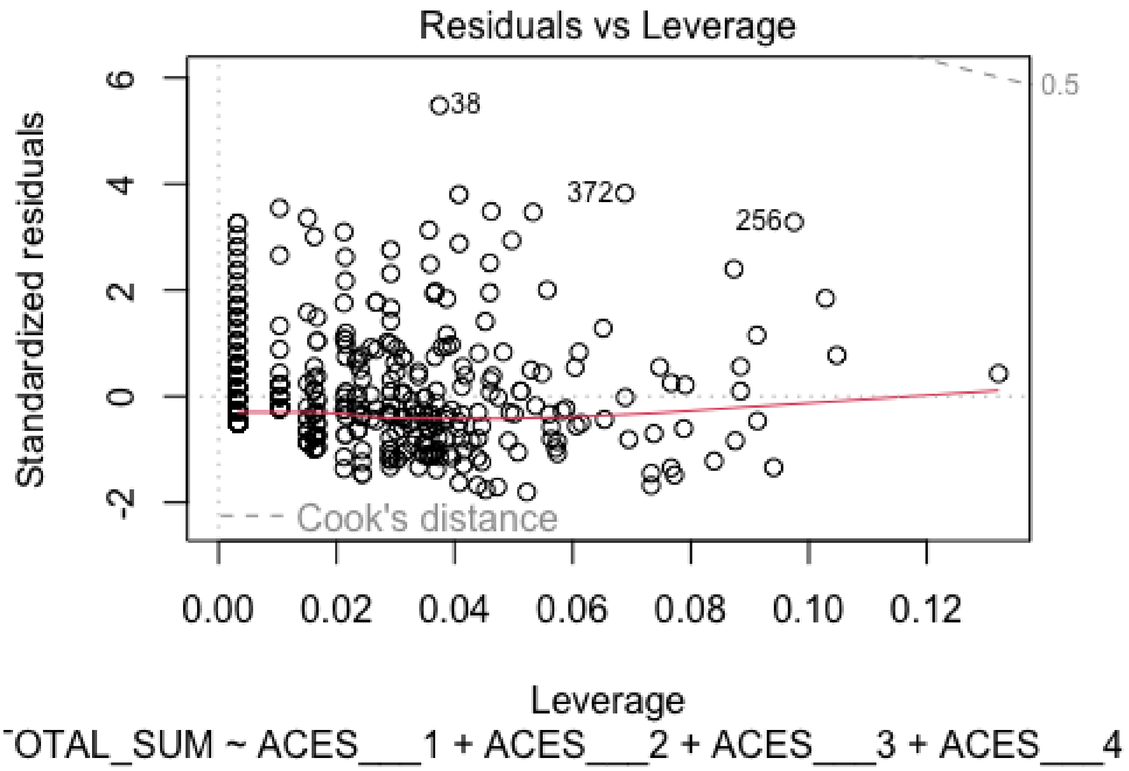
```
## 
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ ACES___1 + ACES___2 + ACES___3 + 
##     ACES___4 + ACES___5 + ACES___6 + ACES___7 + ACES___8 + ACES___9 + 
##     ACES___10, data = rip)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.9486 -2.2530 -1.2530  0.9881 24.3547 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -2.13521    1.41576  -1.508 0.132101    
## ACES___1     0.19534    0.72244   0.270 0.786969    
## ACES___2     2.06336    0.56740   3.637 0.000303 ***
## ACES___3     2.23049    0.56574   3.943 9.13e-05 ***
## ACES___4     0.09203    1.26894   0.073 0.942208    
## ACES___5     0.81234    0.70763   1.148 0.251491    
## ACES___6    -0.53680    0.65125  -0.824 0.410164    
## ACES___7     0.57879    0.66378   0.872 0.383619    
## ACES___8     1.63521    0.53899   3.034 0.002532 ** 
## ACES___9    -1.24110    0.47201  -2.629 0.008799 ** 
## ACES___10   -1.44145    0.99893  -1.443 0.149608    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.533 on 535 degrees of freedom
## Multiple R-squared:  0.146,  Adjusted R-squared:   0.13 
## F-statistic: 9.147 on 10 and 535 DF,  p-value: 5.074e-14
```

```
par(mfrow = c(1,1))
plot(model_aces)
```

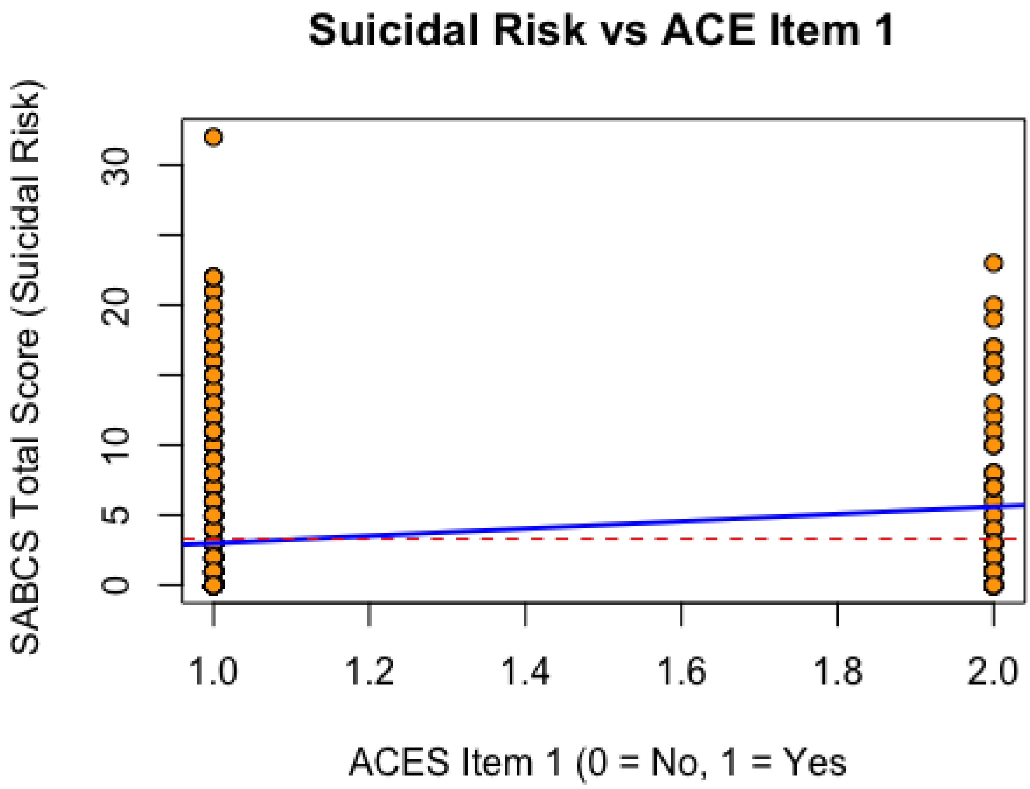

Residuals vs Fitted
Residuals
Fitted values
TOTAL_SUM ~ ACES___1 + ACES___2 + ACES___3 + ACES___4
-10
0
10
20
0
2
4
6
8
38
333
372


Q-Q Residuals
Standardized residuals
6
4
2
0
-2
-3
-2
-1
0
1
2
3
38
Theoretical Quantiles
TOTAL_SUM ~ ACES___1 + ACES___2 + ACES___3 + ACES___4


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
38
333
372
0
2
4
6
8
Fitted values
TOTAL_SUM ~ ACES___1 + ACES___2 + ACES___3 + ACES___4



```
library(car)
vif(model_aces)

##   ACES___1   ACES___2   ACES___3   ACES___4   ACES___5   ACES___6   ACES___7   ACE
S___8
##   1.474060   1.229727   1.749244   1.364332   1.377118   1.495848   1.403090   1.3
16016
##   ACES___9 ACES___10
##   1.208414   1.246614

plot(rip$ACES___1, rip$SABCS_TOTAL_SUM,
     xlab = "ACES Item 1 (0 = No, 1 = Yes ",
     ylab = "SABCS Total Score (Suicidal Risk) ",
     main = "Suicidal Risk vs ACE Item 1",
     pch = 21,
     bg = "orange1",
     col = "black")
abline(lm(SABCS_TOTAL_SUM ~ ACES___1, data = rip), col = "blue", lwd = 2)
abline(h = mean(rip$SABCS_TOTAL_SUM, na.rm = TRUE), col = "red", lty = 2)
```

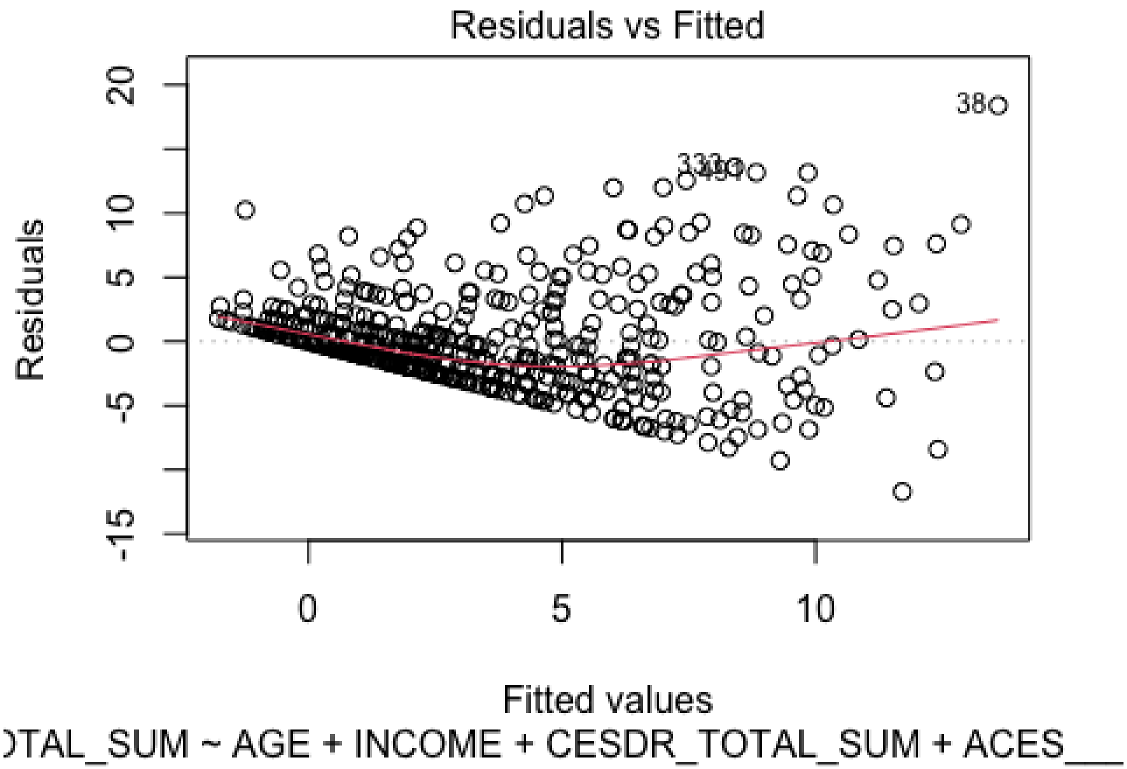


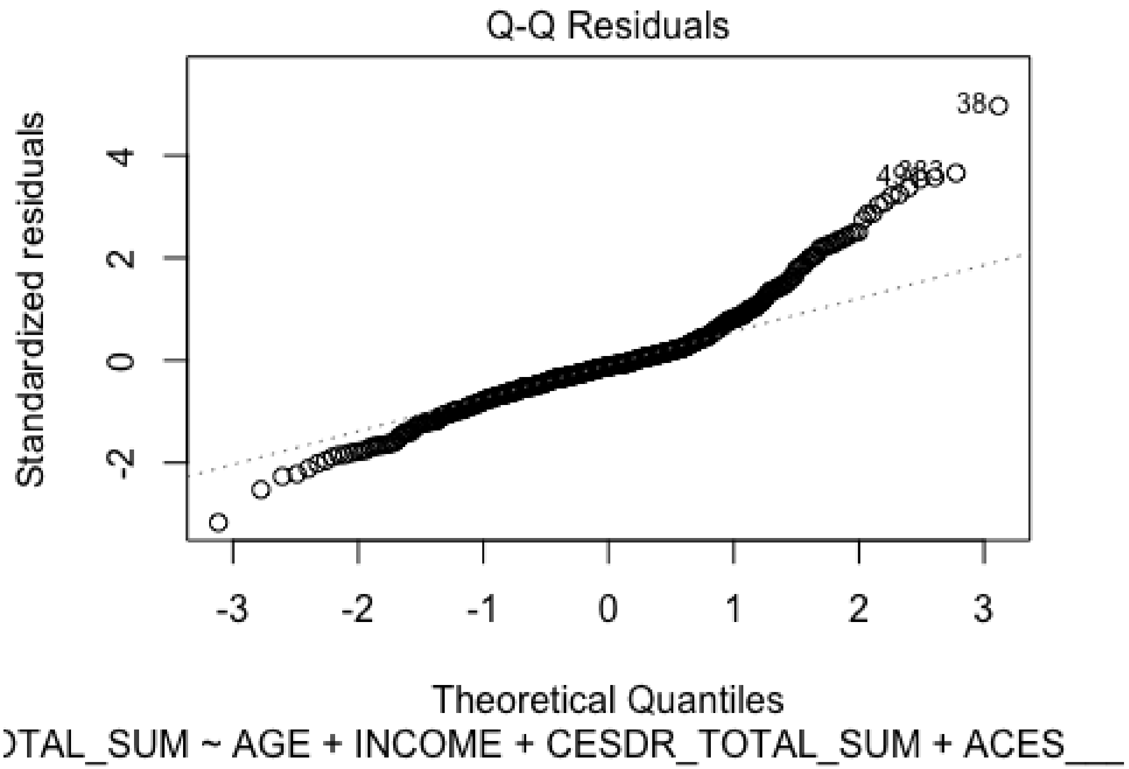
From our regressing model showing the effects of the ACES' variables on suicide, our model shows us that ACES 2, 3 and 8 were positively associated with chances of suicide risk and had the highest SABCS scores. ACES 9 had the most negative association with suicidal risk and had the lowest SABCS score. All 4 of these ACE variables were statistically significant and majority of the other ACES were not. Our adjusted R-squared variable shows us that there is only a 13% difference in suicide risk for the ACES variables.

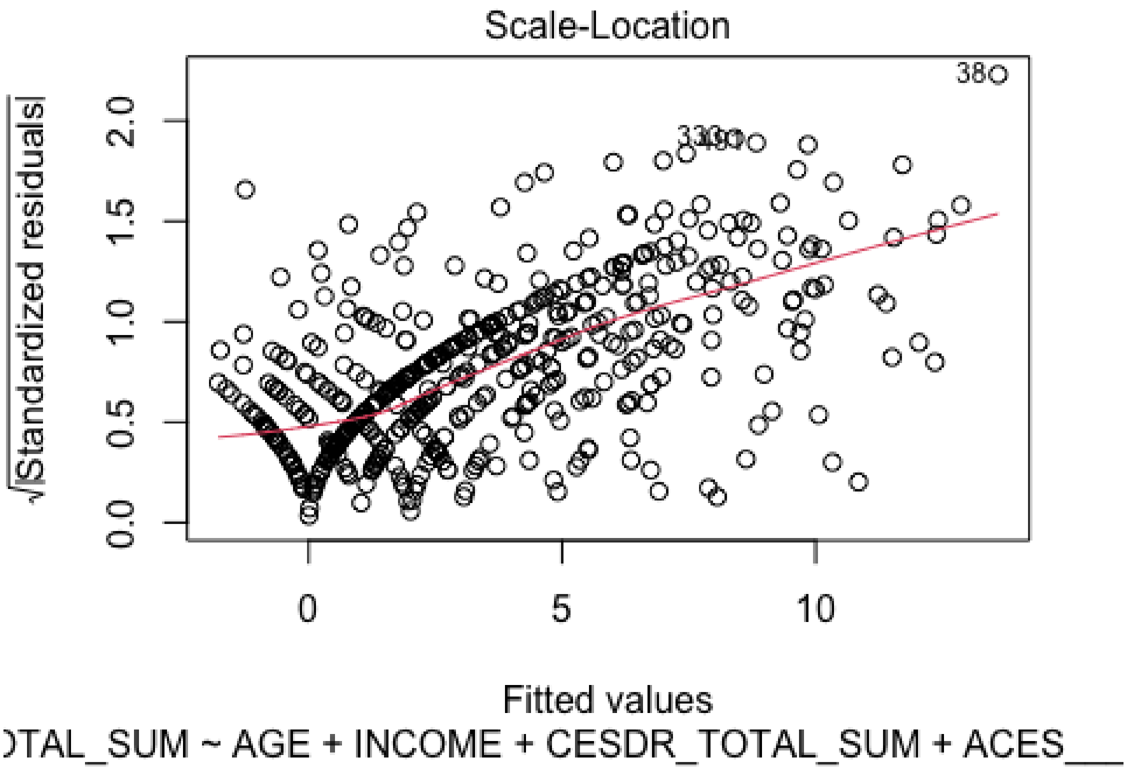
```
combined_model <- lm(SABCS_TOTAL_SUM ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES_
__2 + ACES___3 + ACES___8 + ACES___9 + ETHNICITY, data = rip)


summary(combined_model)

##
## Call:
## lm(formula = SABCS_TOTAL_SUM ~ AGE + INCOME + CESDR_TOTAL_SUM +
##     ACES___2 + ACES___3 + ACES___8 + ACES___9 + ETHNICITY, data = rip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7128  -1.9636  -0.4327   1.2970  18.3987
##
```

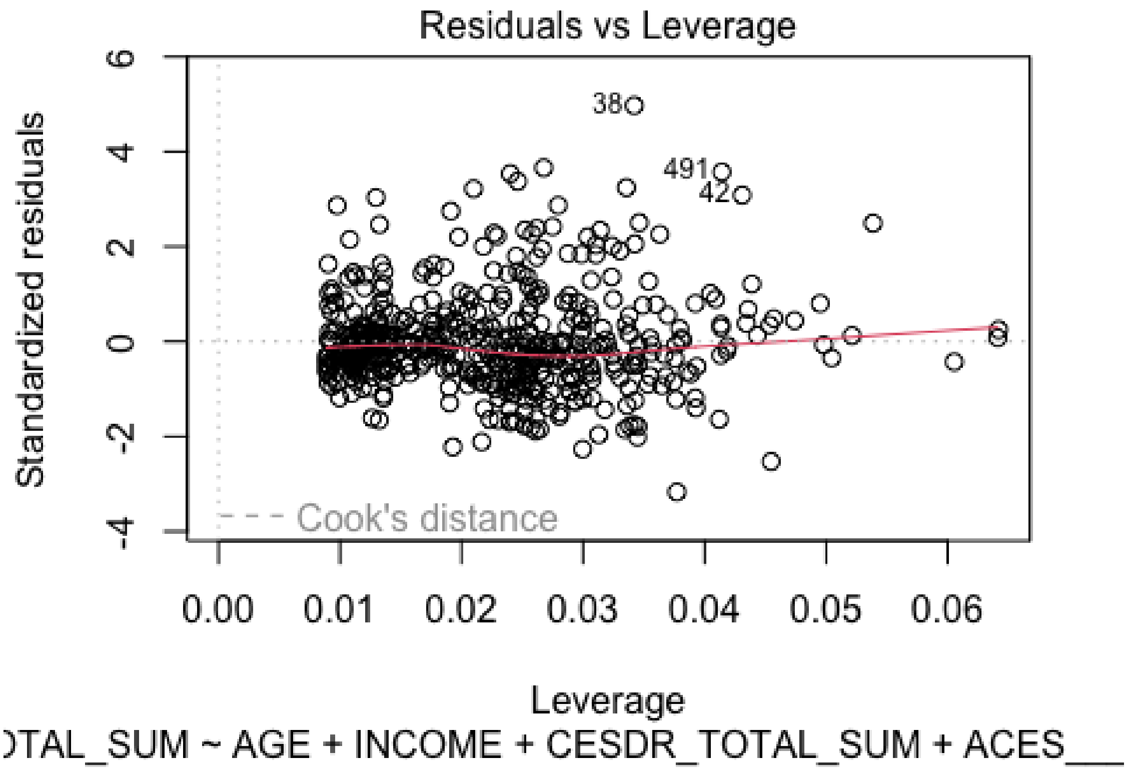
```
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 -0.85172    1.03792  -0.821  0.41224    
## AGE                         -0.05284    0.02608  -2.026  0.04324 *  
## INCOME>$100,000              0.21294    0.52198   0.408  0.68347    
## INCOME$30,000 - $50,000     -0.57805    0.52632  -1.098  0.27258    
## INCOME$51,000 - $75,000     -1.02281    0.52189  -1.960  0.05054 .  
## INCOME$76,000 - $100,000    -0.37808    0.53301  -0.709  0.47843    
## CESDR_TOTAL_SUM              0.18521    0.01262  14.672  < 2e-16 ***
## ACES___2                     1.36728    0.46374   2.948  0.00333 ** 
## ACES___3                     1.41619    0.42865   3.304  0.00102 ** 
## ACES___8                     0.81501    0.43466   1.875  0.06133 .  
## ACES___9                    -1.29322    0.38061  -3.398  0.00073 ***
## ETHNICITYHispanic/Latino     0.99645    0.49927   1.996  0.04646 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.766 on 534 degrees of freedom
## Multiple R-squared:  0.4116, Adjusted R-squared:  0.3995 
## F-statistic: 33.96 on 11 and 534 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,1))
plot(combined_model)
```

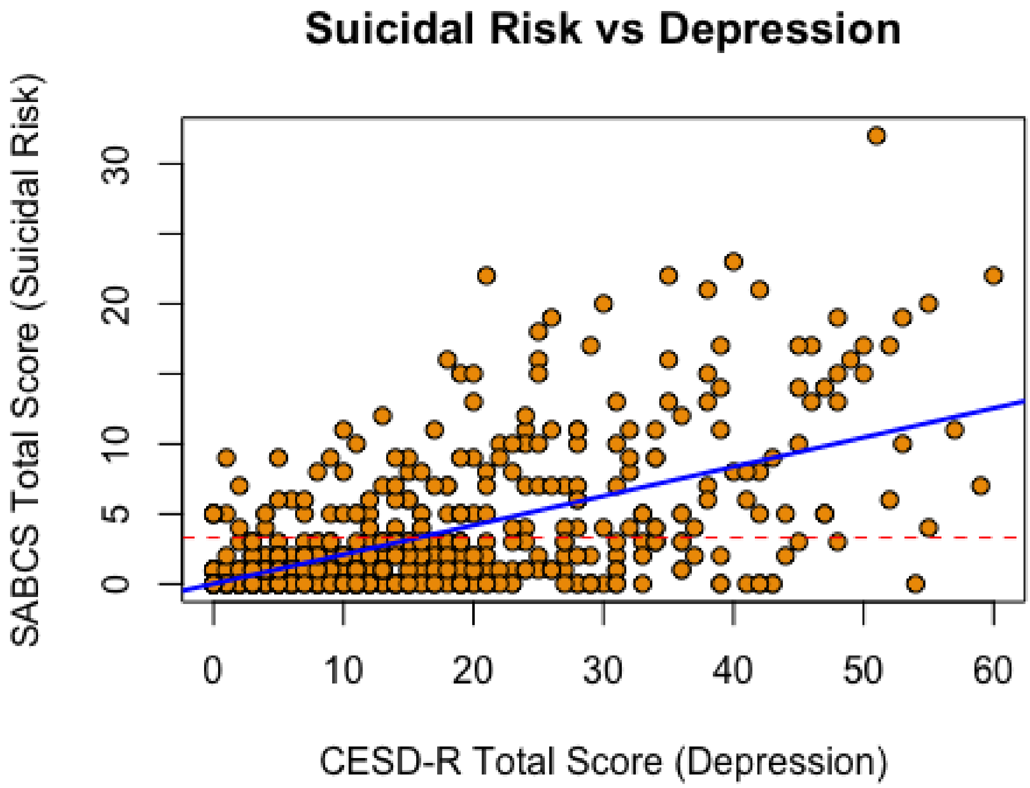

Residuals vs Fitted
Residuals
20
10
5
0
-5
-15
38
333
451
0
5
10
Fitted values
)TAL_SUM ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES___


Q-Q Residuals
Standardized residuals
4
2
0
-2
38
4983
-3
-2
-1
0
1
2
3
Theoretical Quantiles
)TAL_SUM ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES___


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
38
0
5
10
Fitted values
)TAL_SUM ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES___

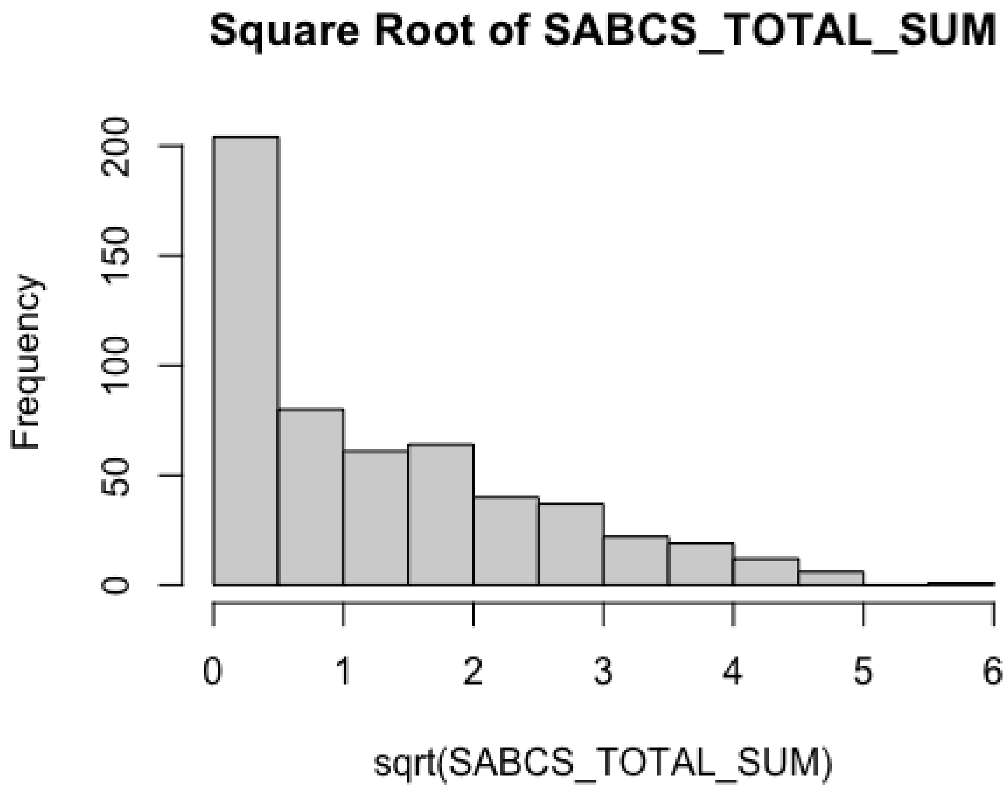


```
library(car)
vif(combined_model)

##                     GVIF Df GVIF^(1/(2*Df))
## AGE             1.105734  1        1.051539
## INCOME          1.169835  4        1.019801
## CESDR_TOTAL_SUM 1.149740  1        1.072259
## ACES___2        1.189974  1        1.090859
## ACES___3        1.454710  1        1.206114
## ACES___8        1.239853  1        1.113487
## ACES___9        1.138236  1        1.066881
## ETHNICITY       1.046398  1        1.022936

plot(rip$CESDR_TOTAL_SUM, rip$SABCS_TOTAL_SUM,
     xlab = "CESD-R Total Score (Depression)",
     ylab = "SABCS Total Score (Suicidal Risk)",
     main = "Suicidal Risk vs Depression",
     pch = 21,
     bg = "orange2",
     col = "black")

abline(lm(SABCS_TOTAL_SUM ~ CESDR_TOTAL_SUM, data = rip), col = "blue", lwd =
2)
abline(h = mean(rip$SABCS_TOTAL_SUM, na.rm = TRUE), col = "red", lty = 2)
```



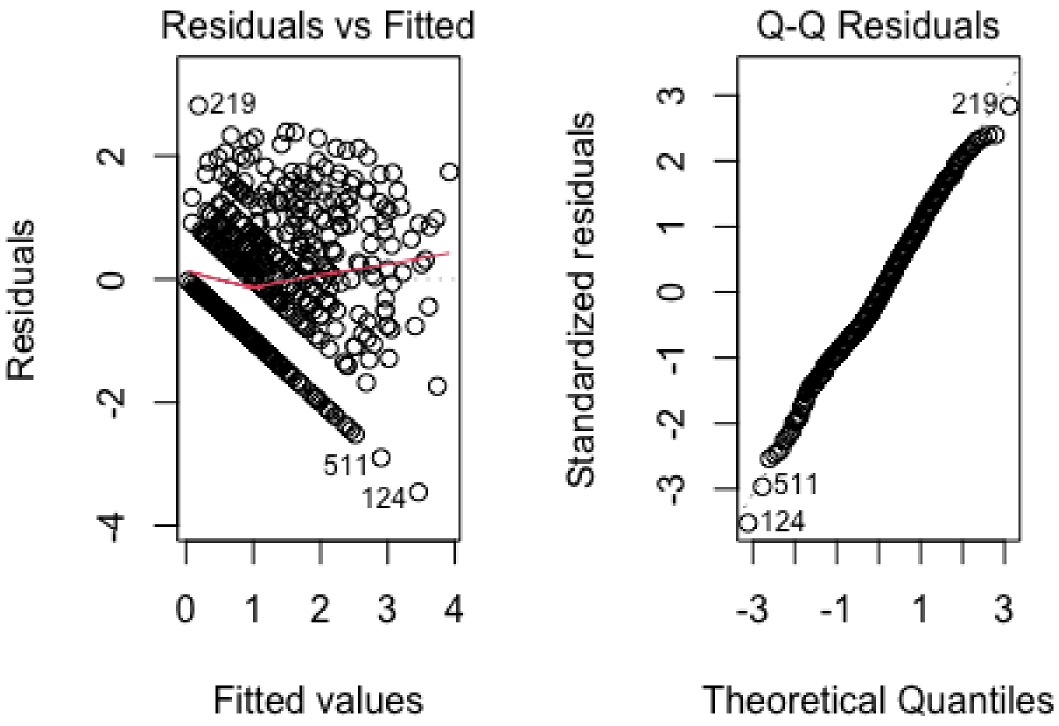
Based on the results, we can tell that the depression scores (CESDR_TOTAL_SUM) and ACES 2, 3, 8, 9 are significant indicators of suicide risk. Higher depression scores are associated with ACES 2, 3, and 8. ACES 9 is associated with lower suicide risk. Older age is linked to slightly lower risk, and some income categories (like $51k–$75k) show slight associations with risk. The model can explain about 40% of why suicidal risk is different between people, which is a lot better than just looking at ACEs' or demographics. The model looks okay overall, and the predictors don't get in each other's way.

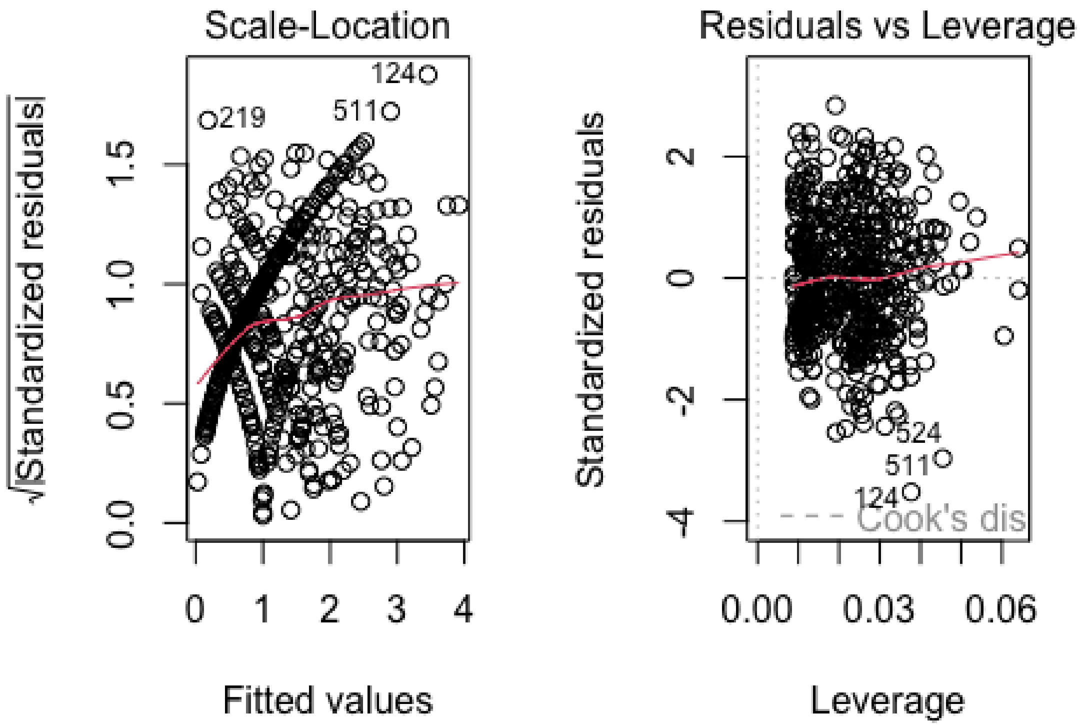
```
rip$SABCS_sqrt <- sqrt(rip$SABCS_TOTAL_SUM)


hist(rip$SABCS_sqrt, main="Square Root of SABCS_TOTAL_SUM", xlab="sqrt(SABCS_
TOTAL_SUM)")
```



```
model_sqrt <- lm(
  SABCS_sqrt ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES___2 + ACES___3 + ACES___
8 + ACES___9 + ETHNICITY,
  data = rip
)


summary(model_sqrt)

##
## Call:
## lm(formula = SABCS_sqrt ~ AGE + INCOME + CESDR_TOTAL_SUM + ACES___2 +
##     ACES___3 + ACES___8 + ACES___9 + ETHNICITY, data = rip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4568 -0.6813 -0.0796  0.6908  2.8157
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 0.072878   0.275999   0.264  0.79184
## AGE                        -0.011481   0.006934  -1.656  0.09835 .
## INCOME>$100,000             0.011714   0.138803   0.084  0.93278
## INCOME$30,000 - $50,000  -0.130373   0.139957  -0.932  0.35201    
## INCOME$51,000 - $75,000  -0.350764   0.138779  -2.528  0.01178 *  
## INCOME$76,000 - $100,000 -0.101121   0.141735  -0.713  0.47588    
## CESDR_TOTAL_SUM           0.047432   0.003357  14.130  < 2e-16 ***
## ACES___2                  0.368951   0.123316   2.992  0.00290 ** 
## ACES___3                  0.347332   0.113984   3.047  0.00242 ** 
## ACES___8                  0.243913   0.115584   2.110  0.03530 *  
## ACES___9                 -0.265520   0.101210  -2.623  0.00895 ** 
## ETHNICITYHispanic/Latino  0.255525   0.132764   1.925  0.05480 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.001 on 534 degrees of freedom
## Multiple R-squared:  0.3994, Adjusted R-squared:  0.3871 
## F-statistic: 32.29 on 11 and 534 DF,  p-value: < 2.2e-16
```

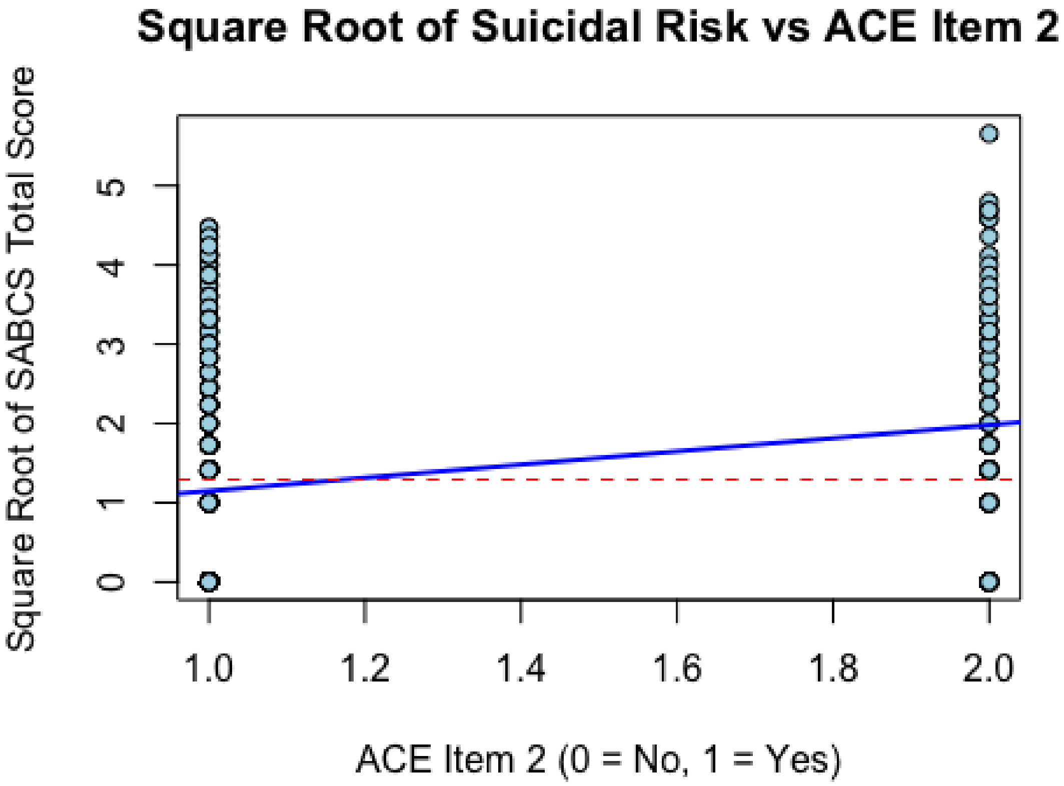
```
par(mfrow=c(1,2))
plot(model_sqrt)
```


Residuals vs Fitted
Residuals
Fitted values
219
511
124
2
0
-2
-4
0
1
2
3
4
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
219
511
124
3
2
1
0
-1
-2
-3
-3
-1
1
3



```
par(mfrow=c(1,1))


library(car)
vif(model_sqrt)

##                     GVIF Df GVIF^(1/(2*Df))
## AGE             1.105734  1        1.051539
## INCOME          1.169835  4        1.019801
## CESDR_TOTAL_SUM 1.149740  1        1.072259
## ACES___2        1.189974  1        1.090859
## ACES___3        1.454710  1        1.206114
## ACES___8        1.239853  1        1.113487
## ACES___9        1.138236  1        1.066881
## ETHNICITY       1.046398  1        1.022936

plot(rip$ACES___2, rip$SABCS_sqrt,
     xlab = "ACE Item 2 (0 = No, 1 = Yes)",
     ylab = "Square Root of SABCS Total Score",
     main = "Square Root of Suicidal Risk vs ACE Item 2",
     pch = 21,
     bg = "lightblue",
     col = "black")
```

```
abline(lm(SABCS_sqrt ~ ACES___2, data = rip),
       col = "blue", lwd = 2)


abline(h = mean(rip$SABCS_sqrt, na.rm = TRUE),
       col = "red", lty = 2)
```



Based on the results calculated using the square root of SABCS_TOTAL_SUM, we can see that CESDR (depression) as well as ACES 2,3,8 and 9 are still large indicators of suicide risk. The higher CESDR scores and ACES 2,3, and 8 are related to high suicide risk and 9 is related to low suicide risk. Model checks show that the leftover errors look pretty normal, the relationships are mostly straight, and the predictors aren't overlapping too much. Overall, this model shows about 39% of the differences in suicidal risk between participants and the transformation helped reduce the influence of extreme values without changing the key findings.

F) The model from part D shows the SABCS_TOTAL_SUM head on, while part E shows the square root of SABCS_TOTAL_SUM to reduce any influences of extreme values. I believe that the transformed model fit the assumptions of linear regression better because it contains a more even spread and more normal errors. The original model is a bit easier to understand though, being that the numbers match the risk scores directly. In my opinion, I would use the transformed model because it is more reliable from a statistics standpoint and it handles the outliers and skewed data better than the original.

QUESTION 2

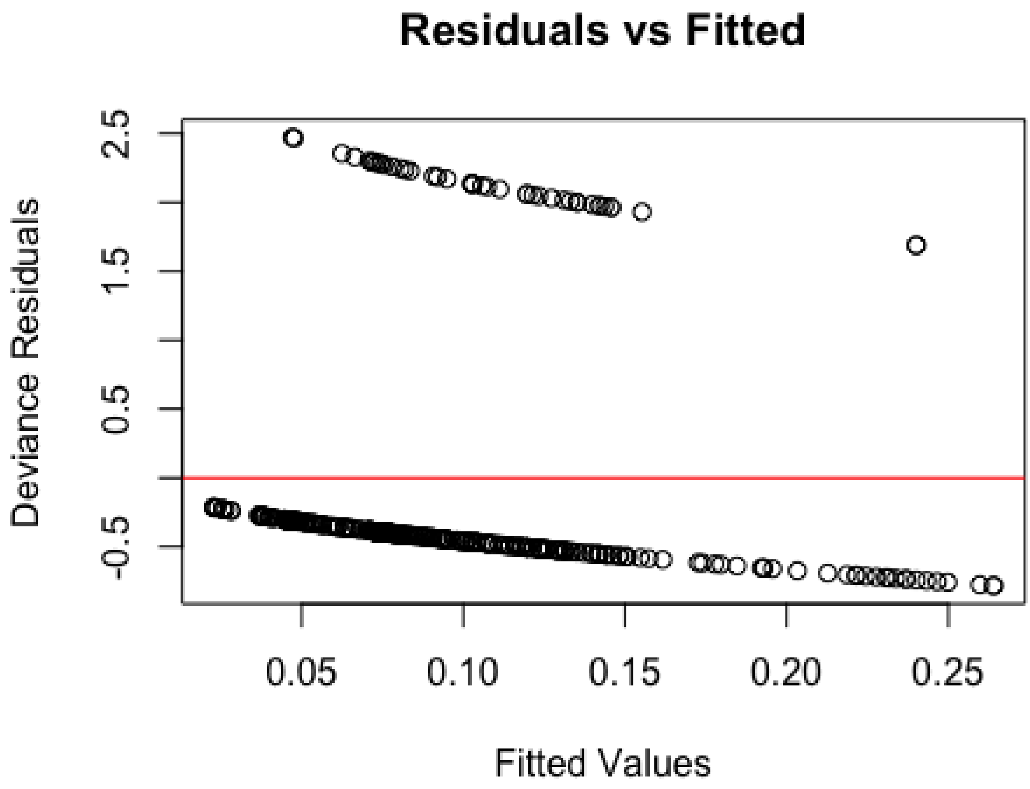
```
rip$HX_SUICIDE <- ifelse(rip$HX_SUICIDE == "Yes", 1, 0 )

hx_model <- glm(HX_SUICIDE ~ AGE + GENDER + RACE + ETHNICITY + INCOME,
                family = binomial, data = rip)


summary(hx_model)
```

```
##
## Call:
## glm(formula = HX_SUICIDE ~ AGE + GENDER + RACE + ETHNICITY +
##     INCOME, family = binomial, data = rip)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.36348    0.69648  -3.393  0.00069 ***
## AGE                       0.01751    0.02217   0.790  0.42969
## GENDERMale                0.21447    0.51692   0.415  0.67821
## RACEOther                -0.71707    0.47514  -1.509  0.13125
## ETHNICITYHispanic/Latino  0.75620    0.41806   1.809  0.07047 .
## INCOME>$100,000          -0.50810    0.48067  -1.057  0.29048
## INCOME$30,000 - $50,000  -0.17408    0.47013  -0.370  0.71117
## INCOME$51,000 - $75,000  -1.03244    0.55650  -1.855  0.06356 .
## INCOME$76,000 - $100,000 -0.54082    0.49226  -1.099  0.27193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 294.70  on 482  degrees of freedom
## Residual deviance: 284.11  on 474  degrees of freedom
##   (63 observations deleted due to missingness)
## AIC: 302.11
##
## Number of Fisher Scoring iterations: 5
```

```
install.packages("car")
```

```
##
## The downloaded binary packages are in
##  /var/folders/k1/cgnjqdxd59b4x6dmzj7dcw4m0000gn/T//RtmpSppoP0/downloaded_p
ackages
```

```
library(car)

plot(hx_model$fitted.values,
     residuals(hx_model, type = "deviance"),
     xlab = "Fitted Values", ylab = "Deviance Residuals",
     main = "Residuals vs Fitted")
abline(h = 0, col = "red")
```



```
vif(hx_model)

##                GVIF Df GVIF^(1/(2*Df))
## AGE       1.068811  1        1.033833
## GENDER    1.021839  1        1.010861
## RACE      1.066858  1        1.032888
## ETHNICITY 1.133027  1        1.064438
## INCOME    1.144678  4        1.017034

exp(coef(hx_model))

##              (Intercept)                      AGE               GENDERMale
##               0.09409208               1.01765998               1.23920870
##                RACEOther ETHNICITYHispanic/Latino          INCOME>$100,000
##               0.48818125               2.13016923               0.60163628
##  INCOME$30,000 - $50,000  INCOME$51,000 - $75,000 INCOME$76,000 - $100,000
##               0.84022718               0.35613550               0.58227044
```

```
exp(confint(hx_model))
```

```
## Waiting for profiling to be done...
```

```
##                               2.5 %    97.5 %
## (Intercept)              0.02422807 0.3791714
## AGE                      0.97105687 1.0602484
## GENDERMale               0.40084601 3.1641779
## RACEOther                0.17471990 1.1574115
## ETHNICITYHispanic/Latino 0.90348639 4.7195639
## INCOME>$100,000          0.22833520 1.5344328
## INCOME$30,000 - $50,000  0.32491353 2.0964144
## INCOME$51,000 - $75,000  0.10882499 1.0081083
## INCOME$76,000 - $100,000 0.21385764 1.5087338
```

```
with(hx_model, 1 - deviance/null.deviance)
```

```
## [1] 0.03591505
```

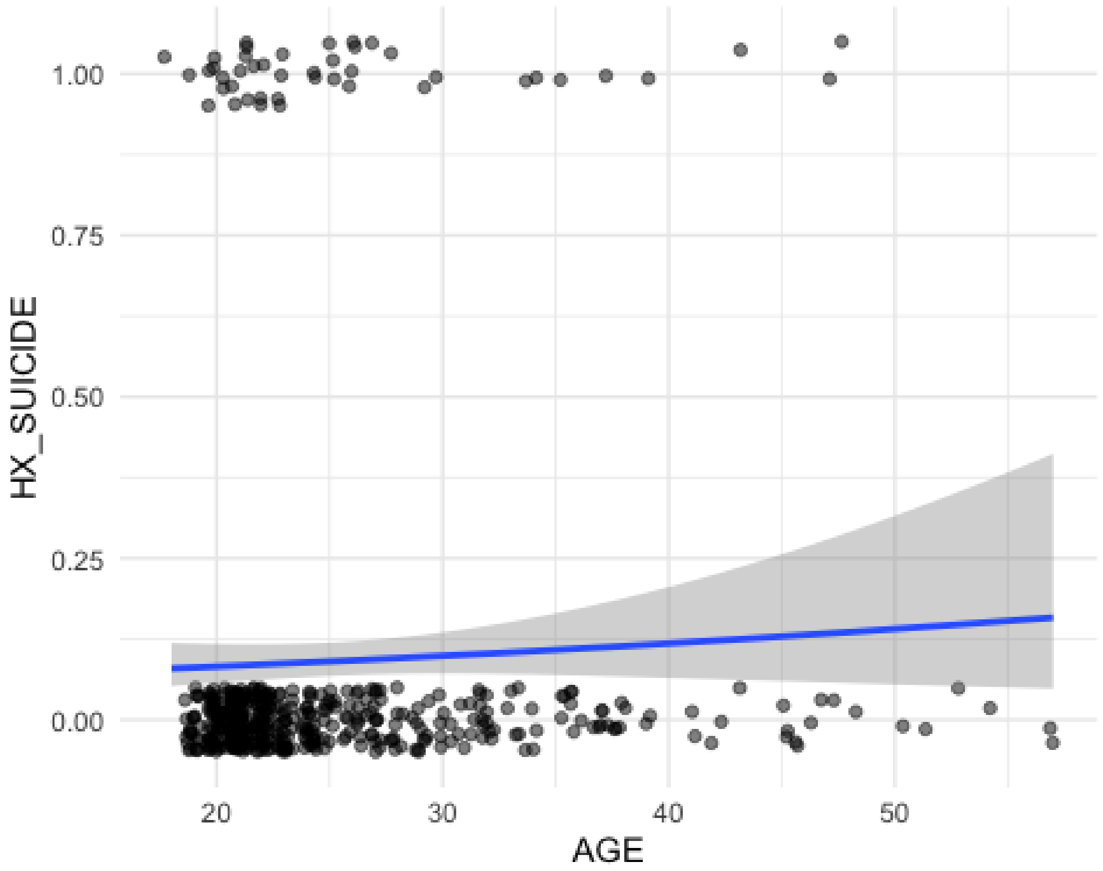
```
install.packages("ResourceSelection")
```

```
## 
## The downloaded binary packages are in
##  /var/folders/k1/cgnjqdxd59b4x6dmzj7dcw4m0000gn/T//RtmpSppoP0/downloaded_p
ackages
```

```
library(ResourceSelection)
```

```
ggplot(hx_model, aes(x = AGE, y = HX_SUICIDE)) +
  geom_jitter(height = 0.05, alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  theme_minimal()
```

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



When looking at HX_SUICIDE amongst demographic variables, none of the variables indicate a high prediction of history of suicide. Under the income category specifically at 51k- 75k shows weak numbers and may point toward lower history, but it is not statistically significant. For hispanic/latino participants, the scores indicated that they may have higher odds of a history, but are not considered statistically significant.

```
hx_depression_model <- glm(HX_SUICIDE ~ CESDR_TOTAL_SUM,
                           data = rip,
                           family = binomial)

summary(hx_depression_model)
```

```
##
## Call:
## glm(formula = HX_SUICIDE ~ CESDR_TOTAL_SUM, family = binomial,
##     data = rip)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.505651   0.295001 -11.884  < 2e-16 ***
## CESDR_TOTAL_SUM  0.057767   0.009868   5.854  4.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 329.72  on 545  degrees of freedom
## Residual deviance: 294.99  on 544  degrees of freedom
## AIC: 298.99
##
## Number of Fisher Scoring iterations: 5
```

```
library(broom)
tidy(hx_depression_model, exponentiate = TRUE, conf.int = TRUE)
```

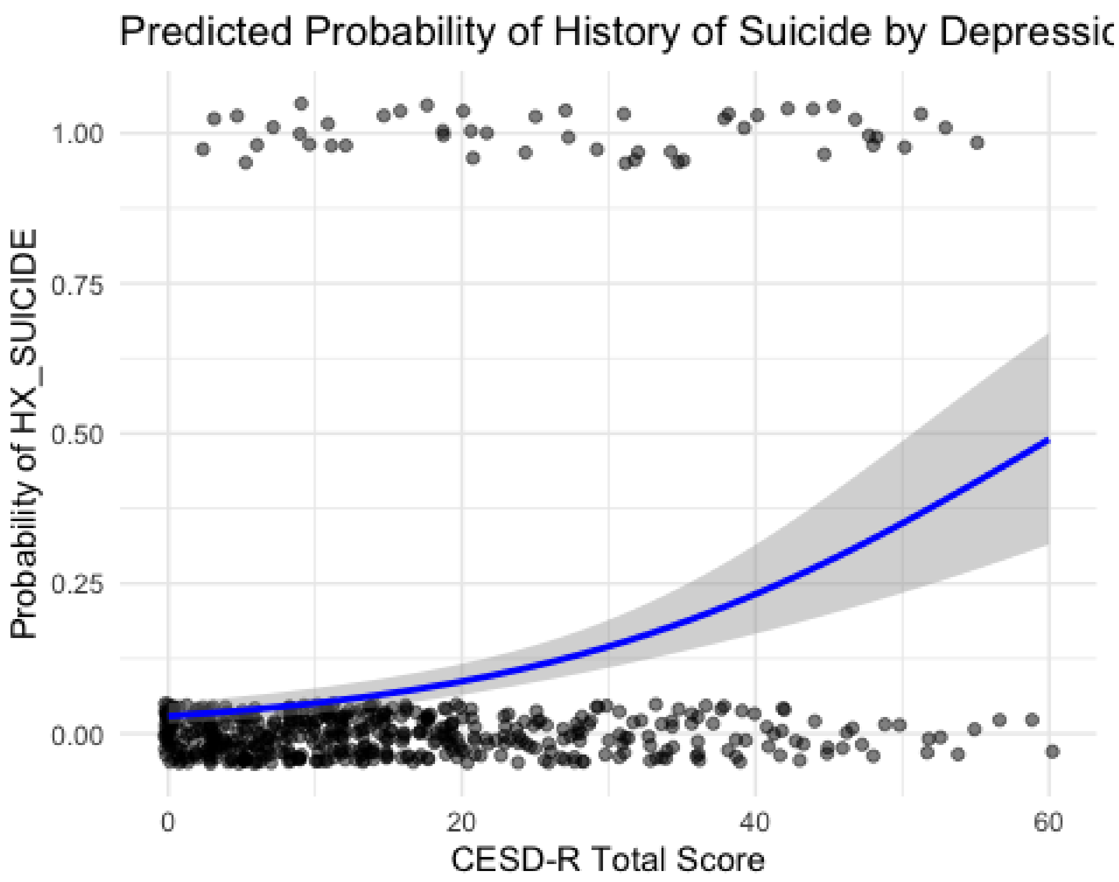
```
## # A tibble: 2 × 7
##   term            estimate std.error statistic  p.value conf.low conf.high
##   <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)       0.0300   0.295      -11.9  1.44e-32   0.0163    0.0519
## 2 CESDR_TOTAL_SUM   1.06     0.00987      5.85 4.80e- 9   1.04      1.08
```

```
library(ResourceSelection)
hoslem.test(rip$HX_SUICIDE, fitted(hx_depression_model), g=10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  rip$HX_SUICIDE, fitted(hx_depression_model)
## X-squared = 4.804, df = 8, p-value = 0.7783
```

```
library(ggplot2)

rip$predicted <- predict(hx_depression_model, type = "response")

ggplot(rip, aes(x = CESDR_TOTAL_SUM, y = HX_SUICIDE)) +
  geom_jitter(height = 0.05, alpha = 0.5) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), color
= "blue") +
  labs(title = "Predicted Probability of History of Suicide by Depression Sco
re",
       x = "CESD-R Total Score",
       y = "Probability of HX_SUICIDE") +
  theme_minimal()
```

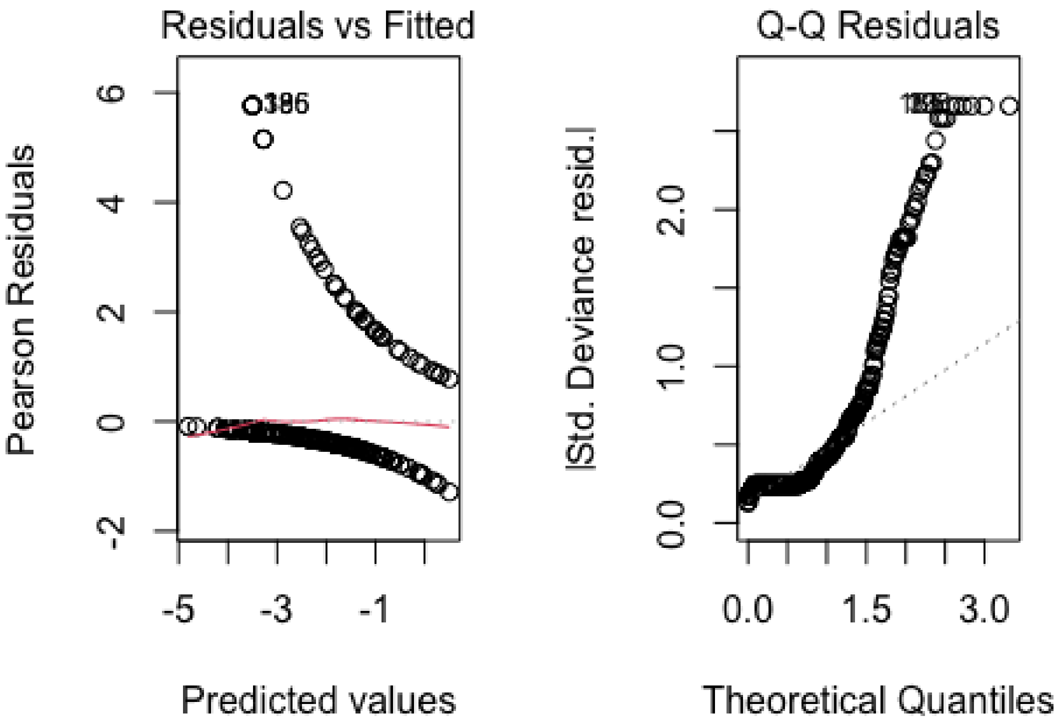
```
## `geom_smooth()` using formula = 'y ~ x'
```

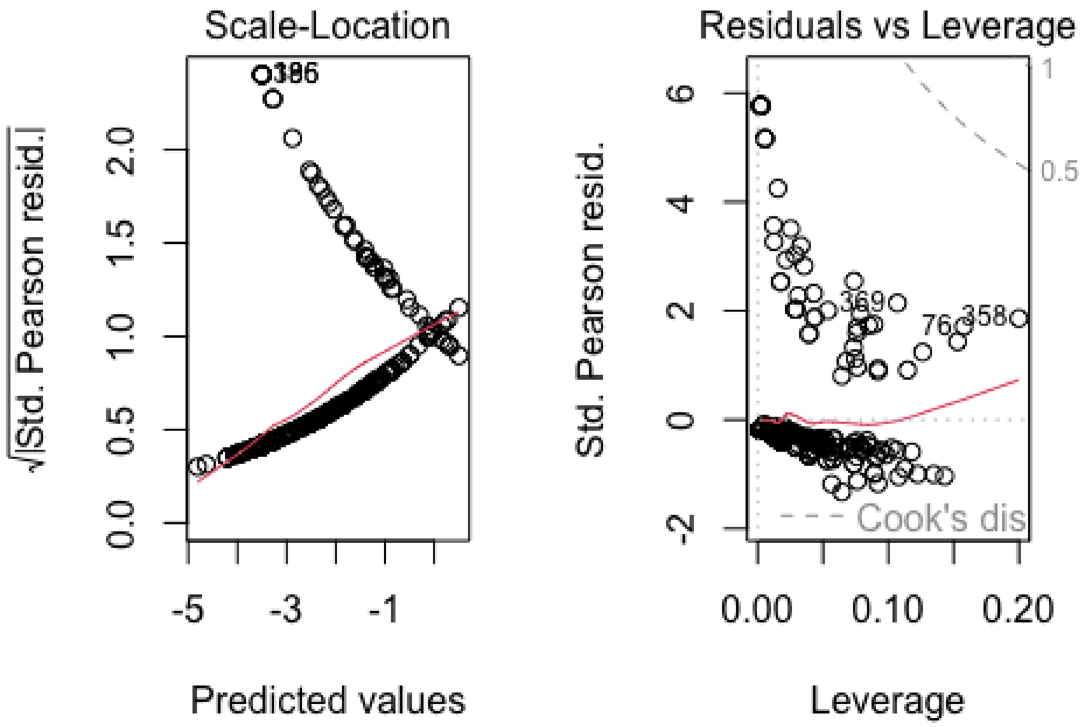


When comparing suicide on depression, the results show us that higher depression scores were correlated with having a history of committing suicide. For every 1 point increase of the CESDR scores, there is a 6% increase in the chances of having history of suicide. These results are statistically significant.

```
aces_model <- glm(
  HX_SUICIDE ~ ACES___1 + ACES___2 + ACES___3 + ACES___4 + ACES___5 +
               ACES___6 + ACES___7 + ACES___8 + ACES___9 + ACES___10,
  data = rip,
  family = binomial
)

summary(aces_model)

##
## Call:
## glm(formula = HX_SUICIDE ~ ACES___1 + ACES___2 + ACES___3 + ACES___4 +
##     ACES___5 + ACES___6 + ACES___7 + ACES___8 + ACES___9 + ACES___10,
##     family = binomial, data = rip)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8787     0.9650  -7.128 1.02e-12 ***
## ACES___1     -0.3472     0.4672  -0.743  0.45741    
## ACES___2      1.6642     0.3682   4.520 6.18e-06 ***
## ACES___3      0.9722     0.4101   2.371  0.01776 *  
## ACES___4      0.1678     0.7605   0.221  0.82538    
## ACES___5     -0.1769     0.5060  -0.350  0.72665    
## ACES___6     -0.4260     0.4449  -0.957  0.33833    
## ACES___7     -0.7034     0.5181  -1.358  0.17457    
## ACES___8      1.1463     0.3721   3.081  0.00207 ** 
## ACES___9      0.2246     0.3677   0.611  0.54136    
## ACES___10     0.8537     0.6117   1.396  0.16286    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 329.72  on 545  degrees of freedom
## Residual deviance: 269.50  on 535  degrees of freedom
## AIC: 291.5
## 
## Number of Fisher Scoring iterations: 6
```

```
par(mfrow=c(1,2))
plot(aces_model)
```


Residuals vs Fitted
Pearson Residuals
6
4
2
0
-2
386
-5
-3
-1
Predicted values
Q-Q Residuals
|Std. Deviance resid.|
2.0
1.0
0.0
0.0
1.5
3.0
Theoretical Quantiles

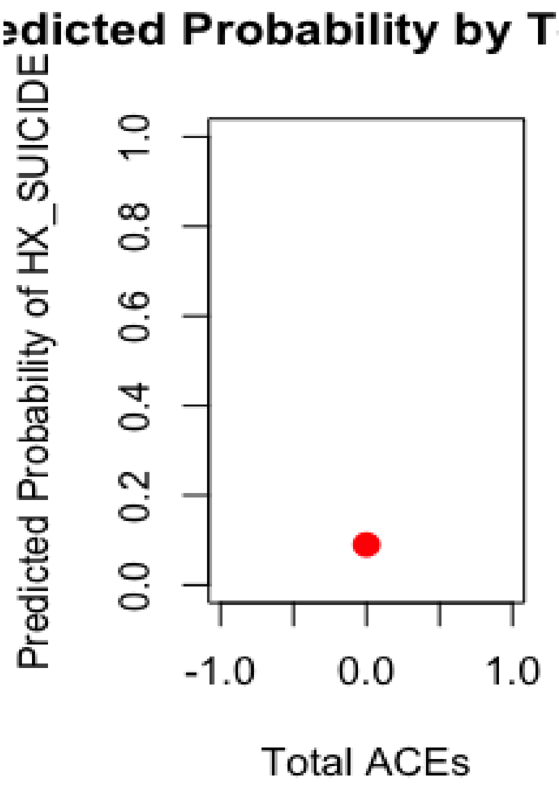


```
library(car)
vif(aces_model)

##   ACES___1   ACES___2   ACES___3   ACES___4   ACES___5   ACES___6   ACES___7   ACE
S___8
##   1.461075   1.266782   1.562255   1.516519   1.499994   1.457657   1.655530   1.2
93654
##   ACES___9 ACES___10
##   1.236776   1.458882

library(ResourceSelection)
hoslem.test(rip$HX_SUICIDE, fitted(aces_model), g = 10)

## Warning in hoslem.test(rip$HX_SUICIDE, fitted(aces_model), g = 10): The da
ta
## did not allow for the requested number of bins.

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  rip$HX_SUICIDE, fitted(aces_model)
## X-squared = 3.3026, df = 4, p-value = 0.5085
```

```r
rip$predicted <- predict(aces_model, type = "response")

rip$ACE_TOTAL <- rowSums(rip[aces_vars] == "Yes")

library(dplyr)
library(ggplot2)

plot_data <- rip %>%
  group_by(ACE_TOTAL) %>%
  summarise(mean_pred = mean(predicted, na.rm = TRUE))


plot_data
```

```
## # A tibble: 1 × 2
##   ACE_TOTAL mean_pred
##       <dbl>     <dbl>
## 1         0    0.0897
```

```r
rip$ACE_TOTAL <- as.numeric(rip$ACE_TOTAL)

plot_data <- aggregate(predicted ~ ACE_TOTAL, data = rip, FUN = mean)


plot(plot_data$ACE_TOTAL, plot_data$predicted,
     type = "b",
     pch = 19,
     col = "blue",
     xlab = "Total ACEs",
     ylab = "Predicted Probability of HX_SUICIDE",
     main = "Predicted Probability by Total ACEs",
     ylim = c(0, 1))
points(plot_data$ACE_TOTAL, plot_data$predicted, pch = 16, col = "red", cex =
1.5)
lines(plot_data$ACE_TOTAL, plot_data$predicted, col = "blue", lwd = 2)
```



When comparing HX Suicide to the ACES variables, our results show us that ACES 2 (sexual abuse), ACES 8 (household mental illness) and ACES 3 (emotional abuse) were all strong indicators of individuals having a higher risk of reporting suicide. ACES 2 had the highest significance with a p value of 6.18e-06. Majority of the other ACES were not great predictors of individuals having a higher risk of reporting suicide.

```
final_model <- glm(HX_SUICIDE ~ ETHNICITY + RACE + INCOME +
                   CESDR_TOTAL_SUM + ACES___2 + ACES___3 + ACES___8 + ACES___
9,
                   family = binomial, data = rip)

library(broom)
library(dplyr)


final_table <- tidy(final_model) %>%
  mutate(
    OR = exp(estimate),
    lower_CI = exp(estimate - 1.96 * std.error),
    upper_CI = exp(estimate + 1.96 * std.error),
    p_value = round(p.value, 3)
  ) %>%
  select(term, OR, lower_CI, upper_CI, p_value)

library(broom)
library(dplyr)


final_table <- tidy(final_model, conf.int = TRUE, exponentiate = TRUE) %>%
  select(Term = term, OR = estimate, `95% CI Lower` = conf.low, `95% CI Upper
` = conf.high, `p-value` = p.value)
library(broom)
library(dplyr)


final_table <- tidy(final_model, conf.int = TRUE, exponentiate = TRUE) %>%
  select(Term = term, OR = estimate, `95% CI Lower` = conf.low, `95% CI Upper
` = conf.high, `p-value` = p.value)

final_table <- final_table %>%
  mutate(Term = case_when(
    Term == "(Intercept)" ~ "(Intercept)",
    Term == "ETHNICITYHispanic/Latino" ~ "ETHNICITY: Hispanic/Latino",
    Term == "RACEOther" ~ "RACE: Other",
    Term == "INCOME30to50k" ~ "INCOME: 30-50k",
    Term == "INCOME51to75k" ~ "INCOME: 51-75k",
    Term == "INCOME76to100k" ~ "INCOME: 76-100k",
    Term == "INCOME>100k" ~ "INCOME: >100k",
    Term == "CESDR_TOTAL_SUM" ~ "CESDR Total Score",
    Term == "ACES___2" ~ "ACE 2",
    Term == "ACES___3" ~ "ACE 3",
    Term == "ACES___8" ~ "ACE 8",
    Term == "ACES___9" ~ "ACE 9",
    TRUE ~ Term
  ))



final_table <- final_table %>%
  mutate(across(c(OR, `95% CI Lower`, `95% CI Upper`), ~ round(.x, 2)),
         `p-value` = signif(`p-value`, 3))


final_table
```

```
## # A tibble: 12 × 5
##    Term                          OR `95% CI Lower` `95% CI Upper` `p-value
`
##    <chr>                      <dbl>          <dbl>          <dbl>     <dbl
>
##  1 (Intercept)                    0              0              0     6.51e-1
```

```
3
##  2 ETHNICITY: Hispanic/Latino  3.59          1.36          9.21  8.34e-
```

Term	OR	95% CI Lower	95% CI Upper	p-value
(Intercept)	0.00	0.00	0.00	0.000000000000651
ETHNICITY: Hispanic/Latino	3.59	1.36	9.21	0.008340000000000
RACE: Other	0.24	0.07	0.68	0.012500000000000
INCOME >$100,000	1.66	0.54	5.11	0.371000000000000
INCOME $30,000 - $50,000	1.05	0.36	2.96	0.928000000000000
INCOME $51,000 - $75,000	0.59	0.16	1.92	0.396000000000000
INCOME $76,000 - $100,000	1.38	0.46	4.09	0.559000000000000
CESDR Total Score	1.05	1.03	1.08	0.000058100000000
ACE 2	4.81	2.25	10.38	0.000052000000000
ACE 3	1.58	0.70	3.55	0.264000000000000

```
3
her
0.07
##  4 INCOME>$
1.66
5.11  3.71e- 1
0,000 - $50,00
0.36
##  6 INCOME$5
0     0.59
1.92  3.96e- 1
```

Term	OR	95% CI Lower	95% CI Upper	p-value
ACE 8	2.09	0.96	4.56	0.062200000000000
ACE 9	1.84	0.87	3.87	0.109000000000000

```
##  3 RACE: Ot
0.24
0.68  1.25e- 2
100,000
0.54
##  5 INCOME$3
0     1.05
2.96  9.28e- 1
1,000 - $75,00
0.16
##  7 INCOME$7
6,000 - $100,000    1.38          0.46          4.09  5.59e- 1
##  8 CESDR Total Score          1.05          1.03          1.08  5.81e-
5
##  9 ACE 2                      4.81          2.25         10.4   5.20e-
5
## 10 ACE 3                      1.58          0.7           3.55  2.64e-
1
## 11 ACE 8                      2.09          0.96          4.56  6.22e-
2
## 12 ACE 9                      1.84          0.87          3.87  1.09e-
1
```

```
library(flextable)
```

```
##
## Attaching package: 'flextable'
```

```
## The following objects are masked from 'package:kableExtra':
##
##     as_image, footnote
```

```
flextable(final_table)
```

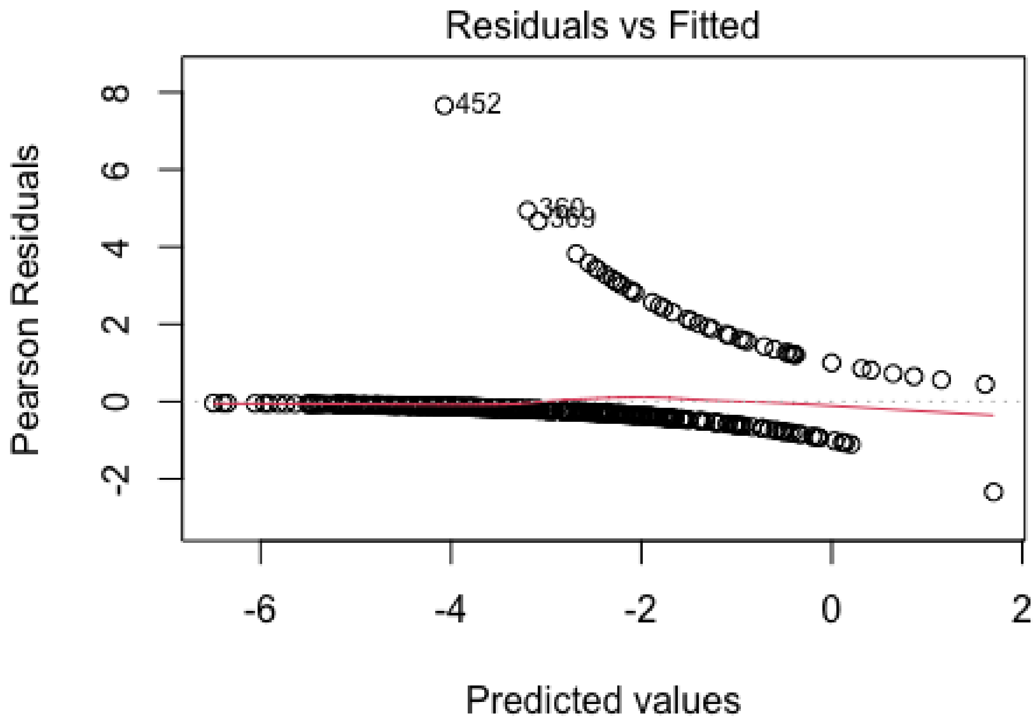
```
library(car)


vif(final_model)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## ETHNICITY       1.209744  1        1.099884
## RACE            1.202317  1        1.096502
## INCOME          1.324522  4        1.035756
## CESDR_TOTAL_SUM 1.120077  1        1.058337
## ACES___2        1.160257  1        1.077152
## ACES___3        1.312665  1        1.145716
## ACES___8        1.222636  1        1.105729
## ACES___9        1.123463  1        1.059935
```

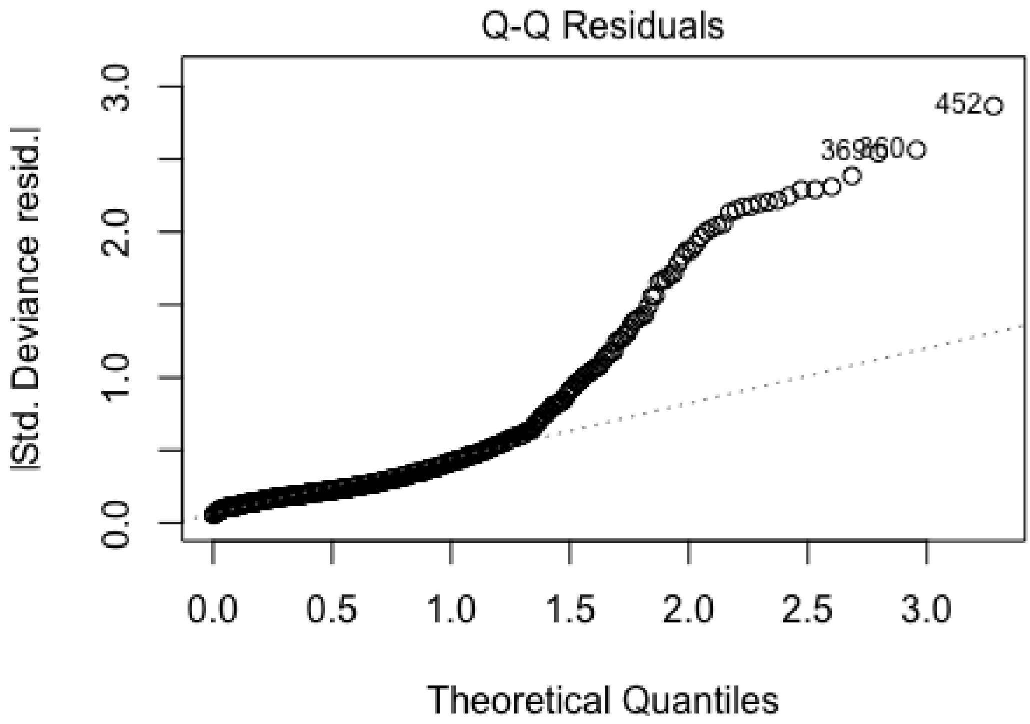
```
summary(final_model)
```

```
## 
## Call:
## glm(formula = HX_SUICIDE ~ ETHNICITY + RACE + INCOME + CESDR_TOTAL_SUM + 
##     ACES___2 + ACES___3 + ACES___8 + ACES___9, family = binomial, 
##     data = rip)
## 
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              -8.01598    1.11496  -7.189 6.51e-13 ***
## ETHNICITYHispanic/Latino  1.27802    0.48450   2.638  0.00834 ** 
## RACEOther                -1.41754    0.56749  -2.498  0.01249 *  
## INCOME>$100,000           0.50736    0.56687   0.895  0.37078    
## INCOME$30,000 - $50,000   0.04826    0.53047   0.091  0.92751    
## INCOME$51,000 - $75,000  -0.53549    0.63121  -0.848  0.39625    
## INCOME$76,000 - $100,000  0.32324    0.55320   0.584  0.55901    
## CESDR_TOTAL_SUM           0.04937    0.01228   4.021 5.81e-05 ***
## ACES___2                  1.56971    0.38794   4.046 5.20e-05 ***
## ACES___3                  0.45896    0.41081   1.117  0.26391    
## ACES___8                  0.73795    0.39570   1.865  0.06219 .  
## ACES___9                  0.60816    0.37922   1.604  0.10878    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 294.7  on 482  degrees of freedom
## Residual deviance: 220.6  on 471  degrees of freedom
##   (63 observations deleted due to missingness)
## AIC: 244.6
## 
## Number of Fisher Scoring iterations: 6
```

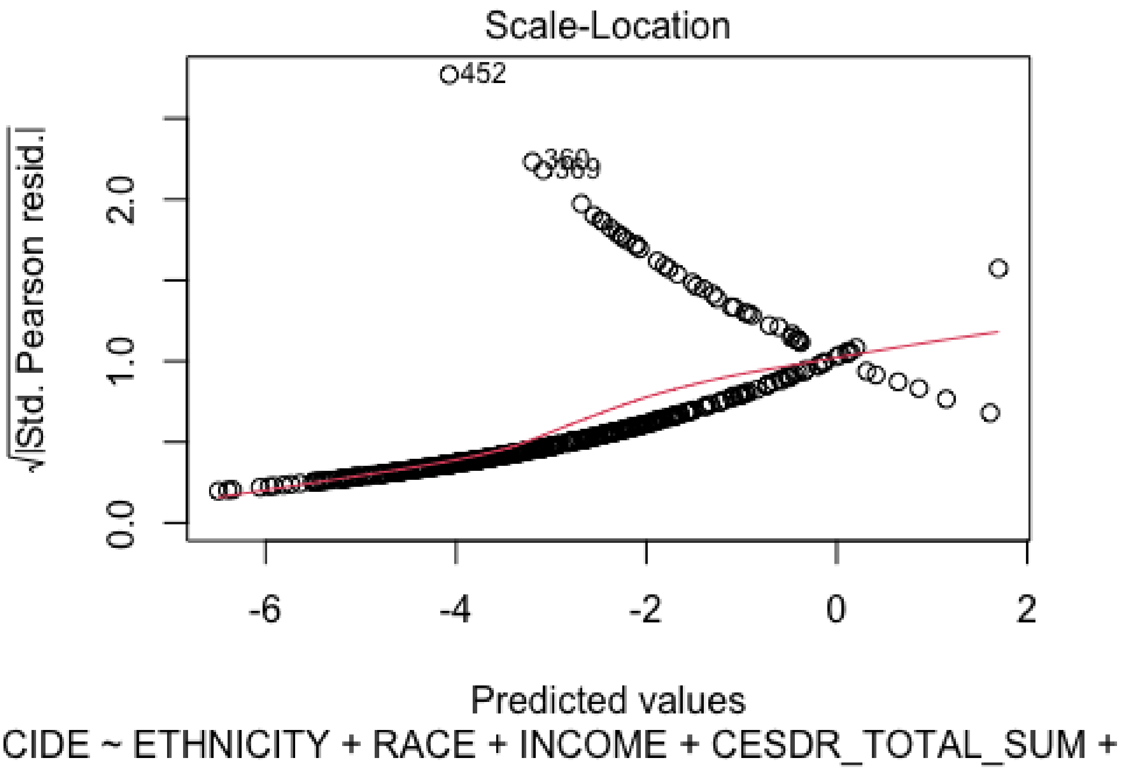
```
par(mfrow = c(1,1))
plot(final_model)
```

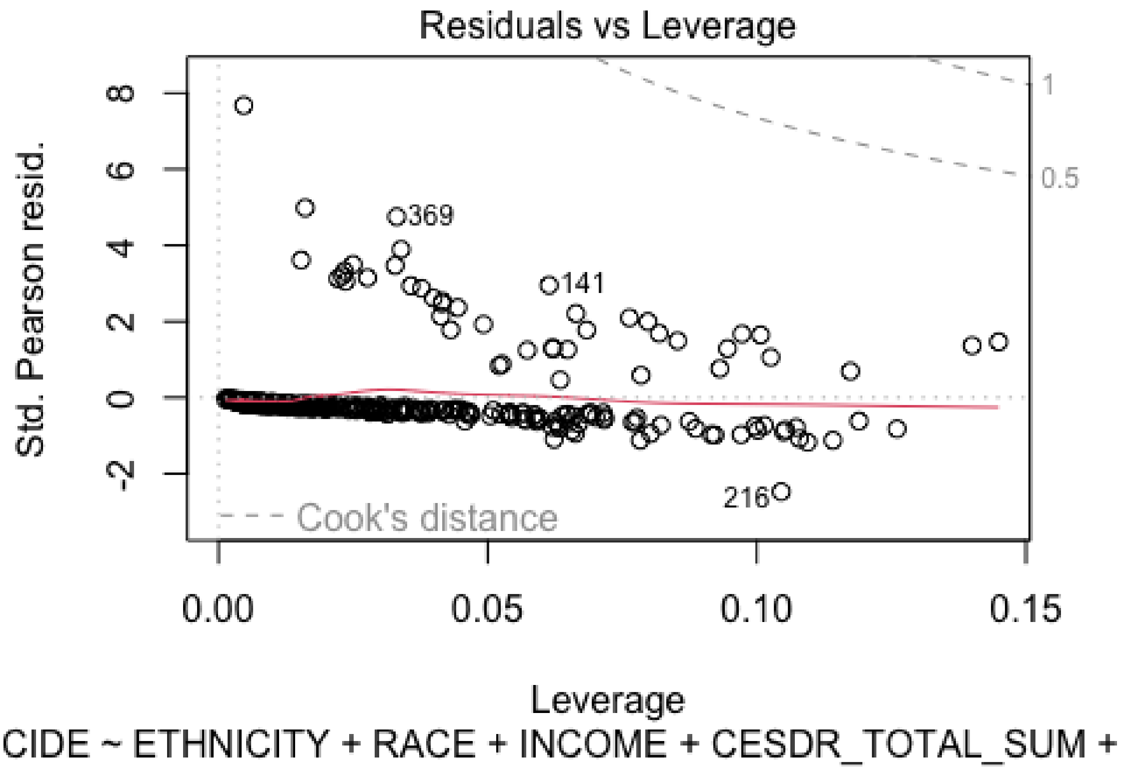


CIDE ~ ETHNICITY + RACE + INCOME + CESDR_TOTAL_SUM +



CIDE ~ ETHNICITY + RACE + INCOME + CESDR_TOTAL_SUM +


Scale-Location
452
360
369
√|Std. Pearson resid.|
0.0
1.0
2.0
-6
-4
-2
0
2
Predicted values
CIDE ~ ETHNICITY + RACE + INCOME + CESDR_TOTAL_SUM +



The final model I created shows that high CESDR scores (high depression), ACES 2(sexual abuse), 8 (household mental illness) and ACES 9 (Parental separation or divorce) are high indicators of history of suicide. Income did not represent a statistically significant prediction of suicide. The Hispanic/latino ethnicity also indicated a higher history of suicide, but was less significant than depression or the ACE variables.

#QUESTION 3

A) Significant independent variable from 1d: CESDR_TOTAL_SUM
   Significant independent variable from 2d: ACES 2
   Measure of effect:

```r
library(dplyr)
library(flextable)


summary_table <- tibble(
  Model = c("SABCS_TOTAL_SUM (Linear)", "HX_SUICIDE (Logistic)"),
  Variable = c("CESDR_TOTAL_SUM", "ACES 2: Yes"),
  Effect = c("β = 0.185", "OR = 2.17"),
  CI_or_SE = c("SE = 0.01266", "95% CI: 1.00–4.72"),
  p_value = c("< 0.001", "0.049"),
  Interpretation = c(
    "Higher depressive symptoms are associated with higher suicidal risk scor
es.",
    "Experiencing ACES 2 is linked to ~2.2 times higher odds of history of su
icide."
  )
)


ft <- flextable(summary_table) %>%
  autofit() %>%
  bold(j = 1, bold = TRUE) %>%
  bold(j = 2, bold = TRUE)


ft
```

Model	Variable	Effect	CI_or_SE	p_value	Interpretation
SABCS_TOTAL_SUM (Linear)	**CESDR_TOTAL_SUM**	β = 0.185	SE = 0.01266	< 0.001	Higher depressive symptoms are associated with higher suicidal risk scores.
HX_SUICIDE (Logistic)	**ACES 2: Yes**	OR = 2.17	95% CI: 1.00–4.72	0.049	Experiencing ACES 2 is linked to ~2.2 times higher odds of history of suicide.

In our interpretation for the respective corresponding measure of effects for each independent variable, we can conclude that depression symptoms are highly associated with suicide risk. For every 1 point increase in depression scores there is a 0.18 increase in suicidal risk. We can also see that exposure to ACES 2 (sexual abuse) also increases the chances in having a history of suicide because participants were inclined to have a 2.2 times higher odds.

B) Nonsignificant independent variable from 1d: Income >$100,000
   Nonsignificant independent variable from 2d: Income 51- 75k

```
library(dplyr)
library(flextable)


ns_table <- tibble(
  Model = c("SABCS_TOTAL_SUM (Linear)", "HX_SUICIDE (Logistic)"),
  Variable = c("INCOME > $100,000", "INCOME: 51–75k"),
  Effect = c("β = 0.054", "OR = 0.97"),
  CI_or_SE = c("SE = 0.517", "95% CI: 0.34–2.70"),
  p_value = c("0.917", "0.955"),
  Interpretation = c(
    "No significant association between high income and suicidal risk score."
,
    "No significant association between income $51–75k and history of suicide
."
  )
)


ft_ns <- flextable(ns_table) %>%
  autofit() %>%
  bold(j = 1, bold = TRUE) %>%
  bold(j = 2, bold = TRUE)

ft_ns
```

Model	Variable	Effect	CI_or_SE	p_value	Interpretation
SABCS_TOTAL_SUM (Linear)	**INCOME > $100,000**	β = 0.054	SE = 0.517	0.917	No significant association between high income and suicidal risk score.
HX_SUICIDE (Logistic)	**INCOME: 51–75k**	OR = 0.97	95% CI: 0.34–2.70	0.955	No significant association between income $51–75k and history of suicide.

From this model we can see that people with an income of over 100,000 did not have a large difference of suicide risk compared to the other groups. People with a mid income of $51,000- 75,000 did not have a significant link to suicide risk as well. The income category is not likely linked to high suicide risk.

#QUESTION 4

A) This analysis mostly resembles a cross sectional study. All of the variables measured (Suicide history, depression, ACES and demographics) are measured during a single point in time. The outcomes of either suicide history or depression are assessed simultaneously.

B) One limitation in this study is that because this is a cross sectional study, we are not able to prove cause and effect. This is because all variables are measured at the same time.

C) In order to minimize this limitation, we could rerun this study and design it as a prospective cohort study. The study subjects would be assessed for the same variables we used at the beginning, and the variables would then be tracked over time. Tracking them over time would allow us to determine the development process of these suicidal behaviors and what may change or cause the variables to vary.