LMFinal_Cherkashina

Elizaveta Cherkashina

2025-04-09

Dataset and Variable Choice, RQ Formulation

For this project, I chose to study the "addhealth" dataset, specifically, I am interested in seeing whether or not depression in teenagers is correlated with relationships with friends and family. The reason why I am specifically interested in studying depression is because with the destigmatization of this diagnosis, the outcomes of the research could allow us to pay more attention to the groups that are more vulnerable to it. Moreover, the article "Depression in teenagers" (Martin, 1996) briefly mentions that family dynamics are related to depression in younger teenagers, which I want to check using the data of the provided dataset.

It is important to mention that in the context of this dataset, the "depress" variable does not signify being diagnosed with clinical depression but rather captures various negative emotions associated with depression (loneliness, sadness, etc.) and their strength.

My research question is as follows: How is the quality of relationships with friends and family correlated with depression, if at all? The hypotheses are the following:

- H0: There is no correlation between family and friends relationships and depression;

- H1: There is a significant linear relationship between depression and predictor variables.

Dependent variable: depress;

Independent variables: sex, age, frndscare, prntscare, famundrst, depress, momcare, dadcare, momrshp, dadrshp, esteem, intlgnce and race.

Although the main focus of the study is the relationship-related variables, the table also shows variables such as age, sex, and race. I also assume that esteem and intelligence (as perceived by a respondent) can have a strong effect on depression. These variables are known to have some relationship with depression and, thus, are taken as control variables. Studies have shown that women (Girgus & Yang, 2015) and teenage girls (Shokrgozar et al., 2019) are more likely to develop depression, similar to teenagers of specific ages (Shokrgozar et al., 2019), and racial/ethnic minorities (Patil et al., 2018).

My dependent variable is numeric, and independent variables are either numeric or factors, thus the best approach for modeling would be using a multiple linear regression.

Before modeling, the data should be cleaned:

```
# Reading the data
health_dat <- read.dta("addhealth.dta")

# Removing NAs in chosen rows
health_dat <- health_dat %>%
  drop_na(c(sex,age,frndscare,prntscare,
            famundrst,depress,momcare,dadcare,
            momrshp,dadrshp,hispanic,white,black,
            asian,othrace,esteem, intlgnce))
```

```r
# Converting age to integer (considering only years of age, not months)
health_dat$age <- as.integer(health_dat$age)

# Converting sex into numerals (1 = male, 2 = female)
health_dat$sex <- gsub("\\D", "", health_dat$sex)
health_dat$sex <- factor(health_dat$sex, levels = c(1, 2),
                         labels = c("Male", "Female"))

# Leaving only numeric values in frndscare, prntscare, famundrst,
# momcare, dadcare variables
health_dat$frndscare <- gsub("\\D", "", health_dat$frndscare)
health_dat$prntscare <- gsub("\\D", "", health_dat$prntscare)
health_dat$famundrst <- gsub("\\D", "", health_dat$famundrst)
health_dat$momcare <- gsub("\\D", "", health_dat$momcare)
health_dat$dadcare <- gsub("\\D", "", health_dat$dadcare)
health_dat$momrshp <- gsub("\\D", "", health_dat$momrshp)
health_dat$dadrshp <- gsub("\\D", "", health_dat$dadrshp)
health_dat$intlgnce <- gsub("\\D", "", health_dat$intlgnce)


# Making sure numeric variables are saved as numeric
health_dat <- health_dat %>%
  mutate(across(c(frndscare,prntscare,famundrst,momcare,dadcare,
                  momrshp,dadrshp,intlgnce,esteem), as.numeric))

# Race variables clean-up and factoring
health_dat$hispanic <- gsub("\\D", "", health_dat$hispanic)
health_dat$white <- gsub("\\D", "", health_dat$white)
health_dat$asian <- gsub("\\D", "", health_dat$asian)
health_dat$black <- gsub("\\D", "", health_dat$black)
health_dat$othrace <- gsub("\\D", "", health_dat$othrace)

health_dat$race <- ifelse(health_dat$hispanic == 1, "Hispanic",
                          ifelse(health_dat$white == 1, "White",
                                 ifelse(health_dat$asian == 1, "Asian",
                                        ifelse(health_dat$black == 1, "Black",
                                               ifelse(health_dat$othrace == 1,
                                                      "Other", "Unknown")))))
health_dat$race <- factor(health_dat$race,
                          levels = c("White",
                                     "Hispanic",
                                     "Asian",
                                     "Black",
                                     "Other"
                                     ),
                          )
```

Modeling

In this section I will build a model, which will also include testing some assumptions for better understanding why the model was changed. All assumption together will be included at the end of this document for better readability.

```r
# Linear model
linmod <- lm(depress ~ sex + age + race + frndscare + prntscare + famundrst +
               momcare + dadcare + momrshp + dadrshp + esteem + intlgnce,
             data = health_dat)
summary(linmod)
```

```
##
## Call:
## lm(formula = depress ~ sex + age + race + frndscare + prntscare +
##     famundrst + momcare + dadcare + momrshp + dadrshp + esteem +
##     intlgnce, data = health_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9734  -3.9891  -0.7349   2.9917  30.2591
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.69896    2.54543  14.418  < 2e-16 ***
## sexFemale     0.24316    0.32528   0.748 0.454865
## age           0.34069    0.08968   3.799 0.000151 ***
## raceHispanic  1.37270    0.52256   2.627 0.008710 **
## raceAsian     0.66236    0.77885   0.850 0.395228
## raceBlack     0.74885    0.45076   1.661 0.096874 .
## raceOther     1.49799    0.70362   2.129 0.033428 *
## frndscare    -0.23620    0.22541  -1.048 0.294867
## prntscare    -0.21402    0.37489  -0.571 0.568174
## famundrst    -0.71904    0.18820  -3.821 0.000139 ***
## momcare      -0.69193    0.39866  -1.736 0.082847 .
## dadcare      -0.59992    0.31049  -1.932 0.053533 .
## momrshp      -0.18036    0.22387  -0.806 0.420581
## dadrshp      -0.45223    0.20176  -2.241 0.025150 *
## esteem       -0.69114    0.05508 -12.548  < 2e-16 ***
## intlgnce     -0.37886    0.14825  -2.556 0.010703 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.865 on 1415 degrees of freedom
## Multiple R-squared:  0.278,  Adjusted R-squared:  0.2704
## F-statistic: 36.33 on 15 and 1415 DF,  p-value: < 2.2e-16
```

```r
linmod1 <- lm(depress ~ sex + age + race + famundrst + momcare + dadcare +
                momrshp + dadrshp + esteem + intlgnce,
              data = health_dat)
summary(linmod1)
```

```
##
## Call:
## lm(formula = depress ~ sex + age + race + famundrst + momcare +
##     dadcare + momrshp + dadrshp + esteem + intlgnce, data = health_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9633  -4.0035  -0.7408   2.9894  29.8569
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.81596    2.42253  14.785  < 2e-16 ***
## sexFemale     0.15049    0.31684   0.475  0.63488
## age           0.33927    0.08965   3.785  0.00016 ***
## raceHispanic  1.39900    0.52206   2.680  0.00745 **
## raceAsian     0.74506    0.77449   0.962  0.33622
## raceBlack     0.80401    0.44739   1.797  0.07253 .
## raceOther     1.52314    0.70327   2.166  0.03049 *
## famundrst    -0.75669    0.18554  -4.078 4.79e-05 ***
## momcare      -0.74897    0.38668  -1.937  0.05295 .
## dadcare      -0.68265    0.29660  -2.302  0.02150 *
## momrshp      -0.19672    0.22342  -0.881  0.37873
## dadrshp      -0.44978    0.20154  -2.232  0.02579 *
## esteem       -0.69795    0.05462 -12.779  < 2e-16 ***
## intlgnce     -0.39913    0.14739  -2.708  0.00685 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.864 on 1417 degrees of freedom
## Multiple R-squared:  0.2772, Adjusted R-squared:  0.2705
## F-statistic: 41.79 on 13 and 1417 DF,  p-value: < 2.2e-16
```

frndscare and prntscare variables do not have a significant relation, but reduce the significance of sex variable. esteem is significant and reduces the significance of sex variable. Based on the results above and studies that show a relation between being female and being depressed, we need to check for interactions between sex and variables that change its significance when removed from the model

```
# Checking for interaction effect
linmod2 <- lm(depress ~ sex + age + race + frndscare + prntscare + famundrst +
                momcare + dadcare + momrshp + dadrshp + sex*prntscare +
                sex*frndscare + esteem + intlgnce, data = health_dat)
# only the interaction between sex and prntscare is significant,
# momcare and dadcare are insignificant similar to previous models
summary(linmod2)
```

```
##
## Call:
## lm(formula = depress ~ sex + age + race + frndscare + prntscare +
##     famundrst + momcare + dadcare + momrshp + dadrshp + sex *
##     prntscare + sex * frndscare + esteem + intlgnce, data = health_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1398  -4.0267  -0.7285   3.0301  29.9791
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         31.56244    2.97769  10.600  < 2e-16 ***
## sexFemale           11.29795    3.34525   3.377 0.000752 ***
## age                  0.33974    0.08940   3.800 0.000151 ***
## raceHispanic         1.35435    0.52097   2.600 0.009428 **
## raceAsian            0.67048    0.77644   0.864 0.387996
## raceBlack            0.71555    0.44990   1.590 0.111956
```

```
## raceOther            1.52808    0.70197   2.177 0.029656 *
## frndscare           -0.03910    0.30062  -0.130 0.896523
## prntscare            0.30623    0.43149   0.710 0.478003
## famundrst           -0.71440    0.18796  -3.801 0.000150 ***
## momcare             -0.38746    0.40898  -0.947 0.343606
## dadcare             -0.53188    0.31033  -1.714 0.086768 .
## momrshp             -0.16947    0.22329  -0.759 0.448010
## dadrshp             -0.44463    0.20127  -2.209 0.027326 *
## esteem              -0.69139    0.05491 -12.592  < 2e-16 ***
## intlgnce            -0.39551    0.14786  -2.675 0.007562 **
## sexFemale:prntscare -1.81454    0.66698  -2.721 0.006598 **
## sexFemale:frndscare -0.51764    0.43888  -1.179 0.238412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.846 on 1413 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.275
## F-statistic: 32.91 on 17 and 1413 DF,  p-value: < 2.2e-16
```

```
linmod3 <- lm(depress ~ sex + age + race + famundrst + dadrshp + sex*prntscare +
                esteem + intlgnce, data = health_dat)
# variable sex itself is not significant, but its interaction
# with prntscare has a very high significance
summary(linmod3)
```

```
##
## Call:
## lm(formula = depress ~ sex + age + race + famundrst + dadrshp +
##     sex * prntscare + esteem + intlgnce, data = health_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5336  -4.0170  -0.7102   3.0990  29.3700
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         29.11821    2.62512  11.092  < 2e-16 ***
## sexFemale           11.66210    3.01265   3.871 0.000113 ***
## age                  0.34184    0.08908   3.837 0.000130 ***
## raceHispanic         1.38229    0.52117   2.652 0.008084 **
## raceAsian            0.76437    0.77679   0.984 0.325278
## raceBlack            0.81348    0.44663   1.821 0.068764 .
## raceOther            1.55524    0.70076   2.219 0.026620 *
## famundrst           -0.76175    0.18196  -4.186 3.01e-05 ***
## dadrshp             -0.59161    0.18377  -3.219 0.001314 **
## prntscare            0.02065    0.40010   0.052 0.958840
## esteem              -0.71865    0.05320 -13.508  < 2e-16 ***
## intlgnce            -0.41537    0.14715  -2.823 0.004828 **
## sexFemale:prntscare -2.36232    0.61844  -3.820 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.855 on 1418 degrees of freedom
## Multiple R-squared:  0.2789, Adjusted R-squared:  0.2728
## F-statistic: 45.71 on 12 and 1418 DF,  p-value: < 2.2e-16
```

```r
# Because of the interaction between variables we need to check VIF,
# it might suggest multicollinearity

vif(linmod3) # very high values of VIF for variables that interact
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## sex            94.708557  1        9.731832
## age             1.037145  1        1.018403
## race            1.071089  4        1.008621
## famundrst       1.297773  1        1.139199
## dadrshp         1.329331  1        1.152966
## prntscare       1.779553  1        1.333999
## esteem          1.361000  1        1.166619
## intlgnce        1.062511  1        1.030782
## sex:prntscare 95.874813  1        9.791568
```

```r
# Centering numeric variable to deal with high VIF values
health_dat$prntscare_centered <- scale(health_dat$prntscare,
                                       center = TRUE,
                                       scale = FALSE)

linmod_c <- lm(depress ~ sex + age + race + famundrst + dadrshp +
                 sex*prntscare_centered + esteem + intlgnce, data = health_dat)
summary(linmod_c)
```

```
##
## Call:
## lm(formula = depress ~ sex + age + race + famundrst + dadrshp +
##     sex * prntscare_centered + esteem + intlgnce, data = health_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5336  -4.0170  -0.7102   3.0990  29.3700
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    29.21803    2.03284  14.373  < 2e-16 ***
## sexFemale                       0.24504    0.31643   0.774 0.438831
## age                             0.34184    0.08908   3.837 0.000130 ***
## raceHispanic                    1.38229    0.52117   2.652 0.008084 **
## raceAsian                       0.76437    0.77679   0.984 0.325278
## raceBlack                       0.81348    0.44663   1.821 0.068764 .
## raceOther                       1.55524    0.70076   2.219 0.026620 *
## famundrst                      -0.76175    0.18196  -4.186 3.01e-05 ***
## dadrshp                        -0.59161    0.18377  -3.219 0.001314 **
## prntscare_centered              0.02065    0.40010   0.052 0.958840
## esteem                         -0.71865    0.05320 -13.508  < 2e-16 ***
## intlgnce                       -0.41537    0.14715  -2.823 0.004828 **
## sexFemale:prntscare_centered -2.36232    0.61844  -3.820 0.000139 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.855 on 1418 degrees of freedom
## Multiple R-squared:  0.2789, Adjusted R-squared:  0.2728
## F-statistic: 45.71 on 12 and 1418 DF,  p-value: < 2.2e-16
```

```
vif(linmod_c) # VIF values are good after centering
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## sex                     1.044818  1        1.022164
## age                     1.037145  1        1.018403
## race                    1.071089  4        1.008621
## famundrst               1.297773  1        1.139199
## dadrshp                 1.329331  1        1.152966
## prntscare_centered      1.779553  1        1.333999
## esteem                  1.361000  1        1.166619
## intlgnce                1.062511  1        1.030782
## sex:prntscare_centered 1.652718  1        1.285581
```

```
# Checking assumptions that can affect the final model
resid <- residuals(linmod_c)
# skewness is positive but less than 1, thus very low and acceptable
skewness(resid)
```

```
## [1] 0.9048995
```

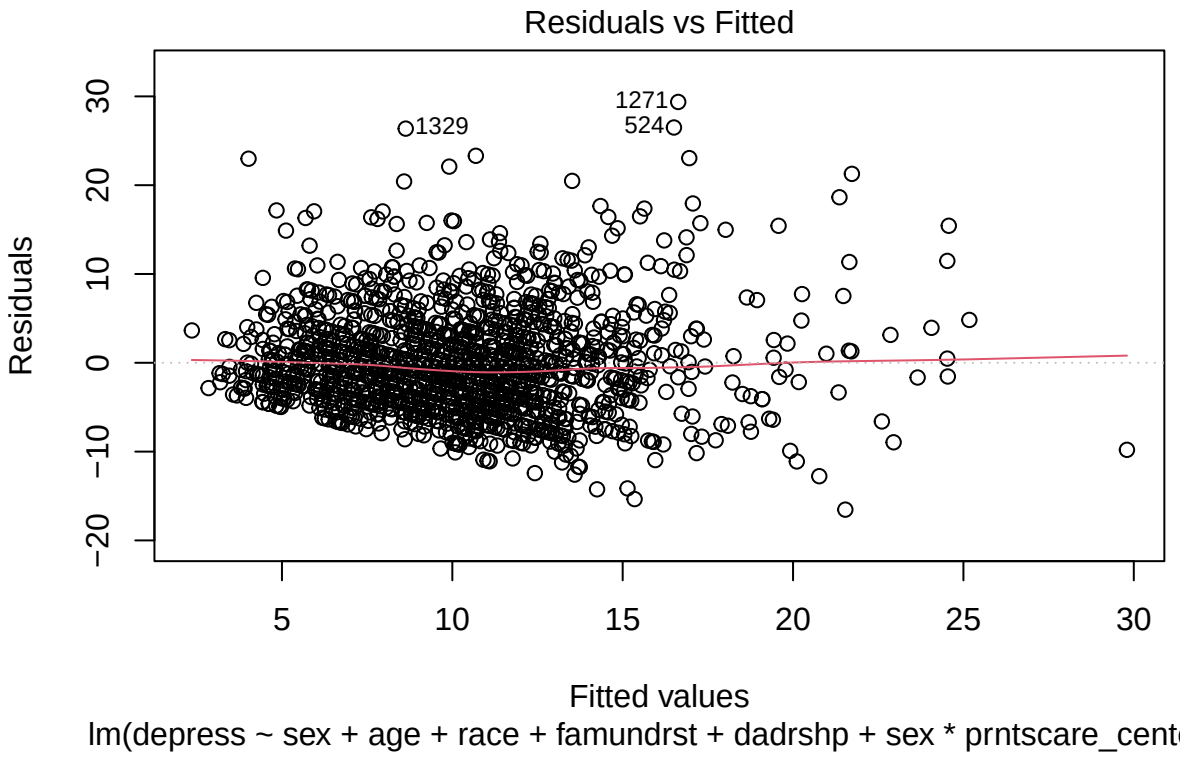
```
# value is very close to 0, thus the zero-mean assumption is satisfied
mean(resid)
```

```
## [1] -2.626985e-16
```

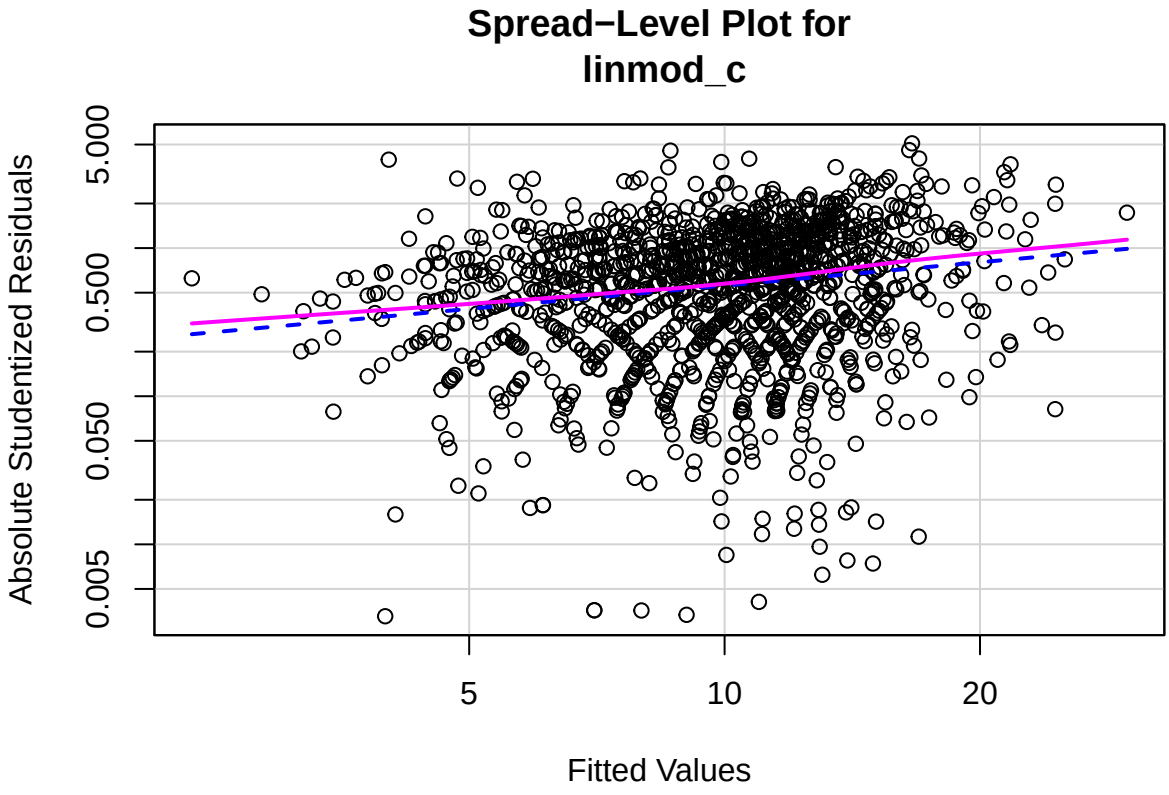
```
# DW test result is 2.0017, meaning that residuals are likely independent and
# p-value greater than 0.05 supports it
dwtest(linmod_c)
```

```
##
##  Durbin-Watson test
##
## data:  linmod_c
## DW = 2.0017, p-value = 0.5119
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# residuals distribution has a distinct cone shape, transformation is needed
plot(linmod_c, which = 1)
```




```
# suggested power of transformation is approximately 0.4770316
# (might change slightly because of how the function works)
spreadLevelPlot(linmod_c)
```



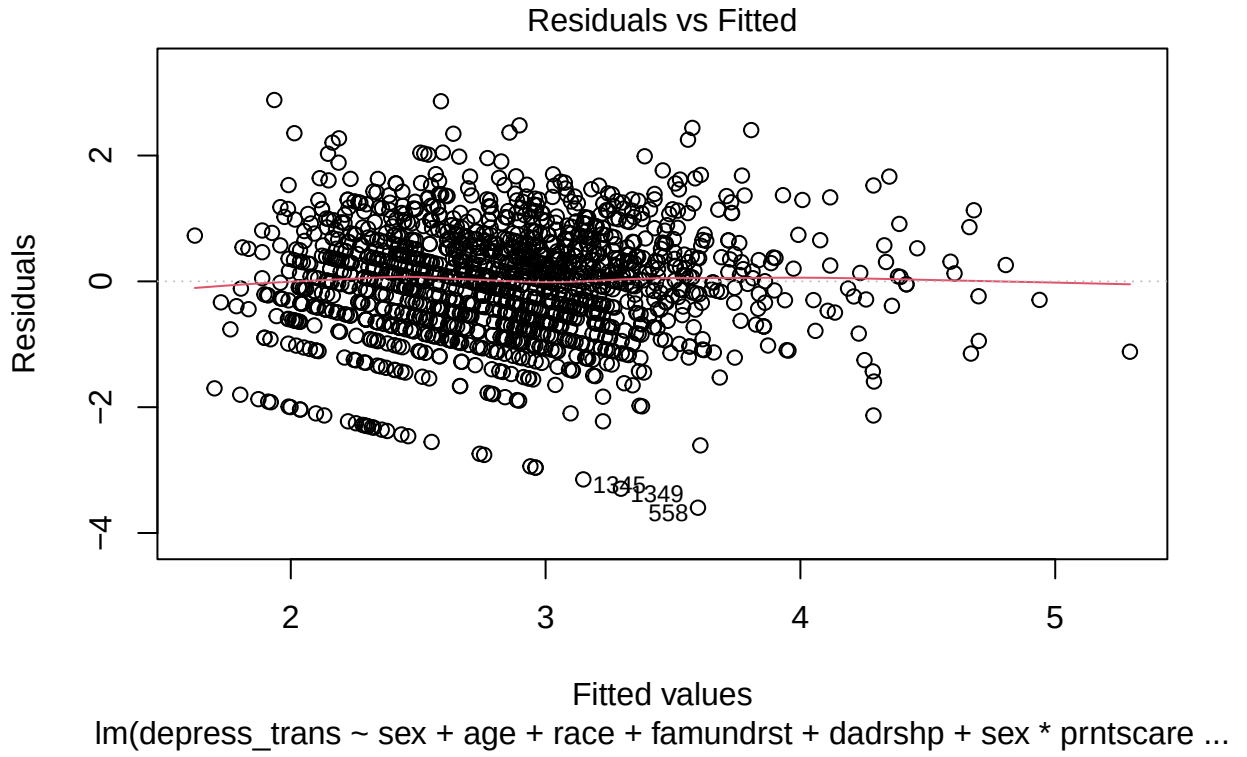

```
##
## Suggested power transformation:  0.4770316
```

```
# Model with transformation
depress_trans <- health_dat$depress^0.4770316
linmod_c <- lm(depress_trans ~ sex + age + race + famundrst + dadrshp +
                 sex*prntscare_centered + esteem + intlgnce, data = health_dat)
summary(linmod_c)
```

```
##
## Call:
## lm(formula = depress_trans ~ sex + age + race + famundrst + dadrshp +
##     sex * prntscare_centered + esteem + intlgnce, data = health_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5980 -0.5441  0.0301  0.5532  2.8821
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   5.46383    0.30148  18.123  < 2e-16 ***
## sexFemale                     0.02540    0.04693   0.541  0.58837
## age                           0.06235    0.01321   4.720 2.60e-06 ***
## raceHispanic                  0.19827    0.07729   2.565  0.01041 *
## raceAsian                     0.14343    0.11520   1.245  0.21332
## raceBlack                     0.10236    0.06624   1.545  0.12248
## raceOther                     0.23765    0.10393   2.287  0.02236 *
```

```
## famundrst                     -0.11839    0.02699  -4.387 1.23e-05 ***
## dadrshp                       -0.08763    0.02725  -3.215  0.00133 **
## prntscare_centered             0.03012    0.05934   0.508  0.61175
## esteem                        -0.10106    0.00789 -12.809  < 2e-16 ***
## intlgnce                      -0.08613    0.02182  -3.947 8.31e-05 ***
## sexFemale:prntscare_centered -0.24099    0.09172  -2.628  0.00869 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8683 on 1418 degrees of freedom
## Multiple R-squared:  0.2655, Adjusted R-squared:  0.2593
## F-statistic: 42.72 on 12 and 1418 DF,  p-value: < 2.2e-16
```

```
plot(linmod_c, which = 1) # much better spread residuals
```



Model Interpretation

F-statistic suggests that the model is overall significant, although there are some variables that are not statistically significant. Asian and Black levels of race are not significant, as well as sex and prntscare, but these variables are left in the model because their interaction is very statistically significant.

R-square is sufficient, but still low. This is explained by the fact that I have only concentrated on how relationships with people are connected to depression and added some control variables. Because causes of depression are very complex it is not surprising that this model only explains approximately 26% of the variance.

Variable coefficients:

sexFemale (0.02540): Being female (vs. male) has a very small positive effect on depression score, but the effect is not statistically significant ($p = 0.58837 > 0.05$).

age (0.06235): As age increases, depress_trans increases by 0.06235 units. This is statistically significant with a p-value of 2.60e-06, indicating a strong positive relationship between age and depression.

raceHispanic (0.19827): Being Hispanic is associated with an increase of 0.19827 units in depress_trans (compared to the reference group). This is statistically significant ($p = 0.01041$).

raceAsian (0.14343): Being Asian is associated with a small increase of 0.14343 units in depression, but this is not statistically significant ($p = 0.21332$).

raceBlack (0.10236): Being Black has a small positive effect on depression, but it is not statistically significant ($p = 0.12248$).

raceOther (0.23765): Being from another race is associated with an increase of 0.23765 units in depression, and this effect is statistically significant ($p = 0.02236$).

famundrst (-0.11839): Having a family background that is less supportive is associated with a decrease of 0.11839 units in depression. This effect is statistically significant ($p = 1.23e-05$).

dadrshp (-0.08763): Having a strong relationship with the father is associated with a decrease of 0.08763 units in depression. This effect is statistically significant ($p = 0.00133$).

prntscare_centered (0.03012): Parental care has a small positive effect on depression, but this effect is not statistically significant ($p = 0.61175$).

esteem (-0.10106): Higher self-esteem is associated with a decrease in depression by 0.10106 units. This effect is very strong and statistically significant ($p < 2e-16$).

intlgnce (-0.08613): Higher intelligence is associated with a decrease in depression by 0.08613 units. This effect is statistically significant ($p = 8.31e-05$).

sexFemale:prntscare_centered (-0.24099): There is an interaction effect between being female and parental care, with a negative effect of -0.24099. This suggests that for females, higher parental care is associated with a greater reduction in depression compared to males. This interaction is statistically significant ($p = 0.00869$).

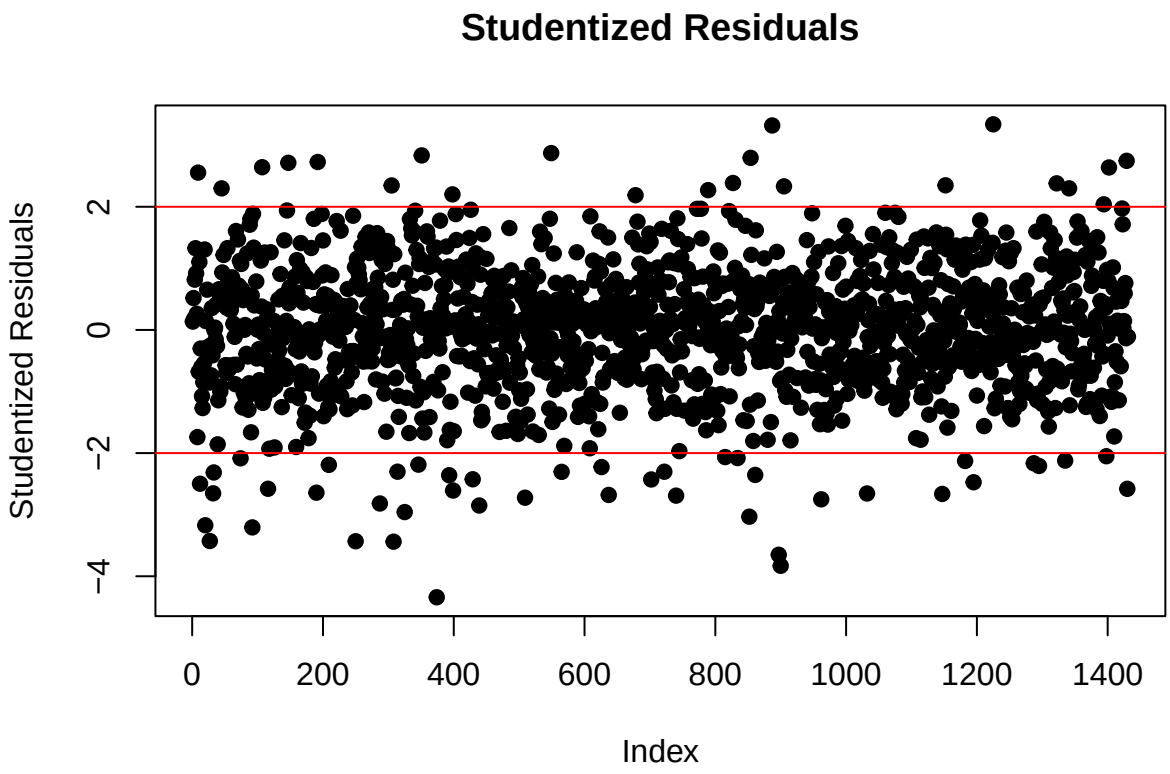
Influential Observations and Relative Importance Sets

Influential Observations

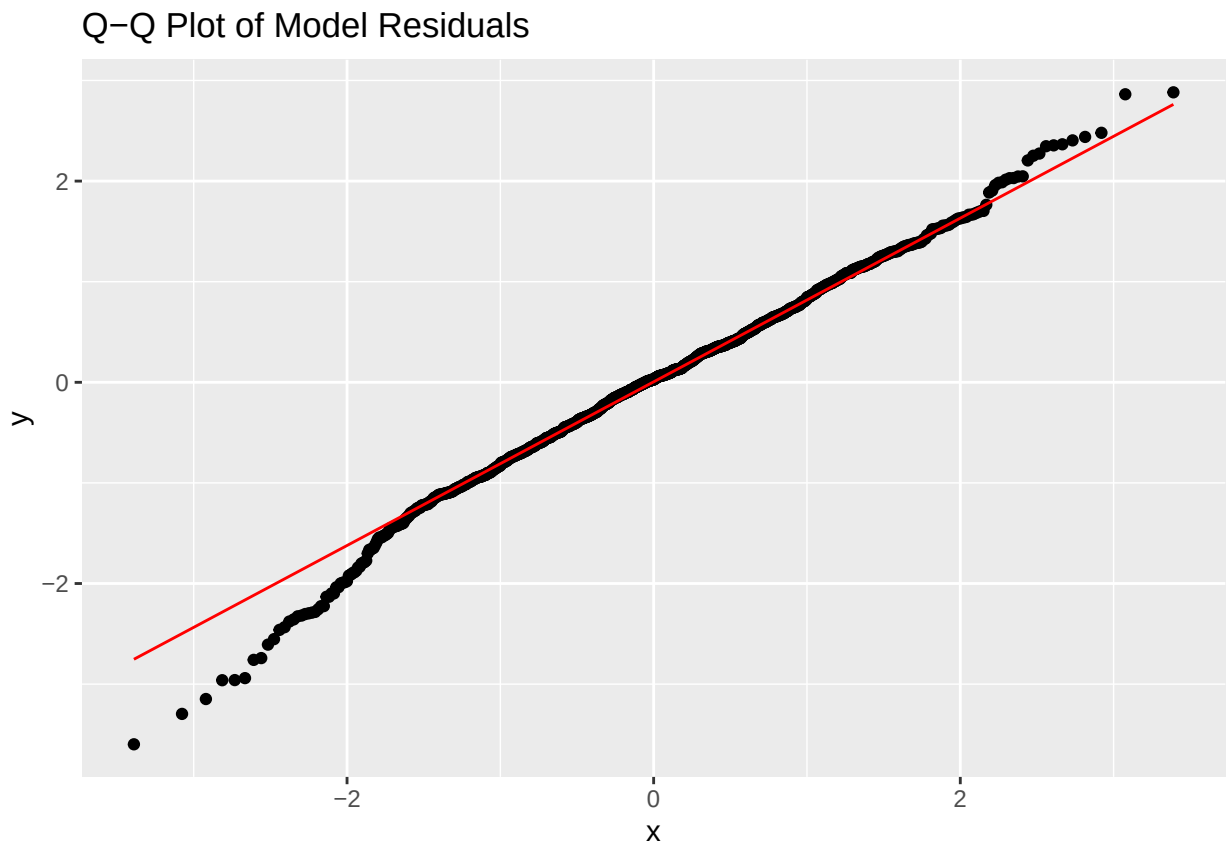
In this part of the project I will check for influential observations and comment on whether or not they should be removed. Relative importance sets will also be determined.

First, I am going to test for outliers with studentized residuals:

```
# Studentized residuals
studentized_residuals <- rstudent(linmod_c)

# Plot studentized residuals
plot(studentized_residuals, main = "Studentized Residuals",
     ylab = "Studentized Residuals", pch = 19)
abline(h = c(-2, 2), col = "red")  # thresholds for identifying large residuals
```




```
# Quantile comparison plot
ggplot(data = data.frame(resid = residuals(linmod_c)), aes(sample = resid)) +
  stat_qq() +
  stat_qq_line(col = "red") +
  ggtitle("Q-Q Plot of Model Residuals")
```



Studentized residuals plot suggests presence of outliers, most of them are around 2 and -2 thresholds, which means these are moderate, although there are some outliers that surpass -3 and even -4 thresholds. These outliers are potentially more influential on the model.

Quantile comparison plot suggests that generally the distribution is normal, but there is some deviation at both tales. This means there are either outliers or heavier tails in the residual distribution.

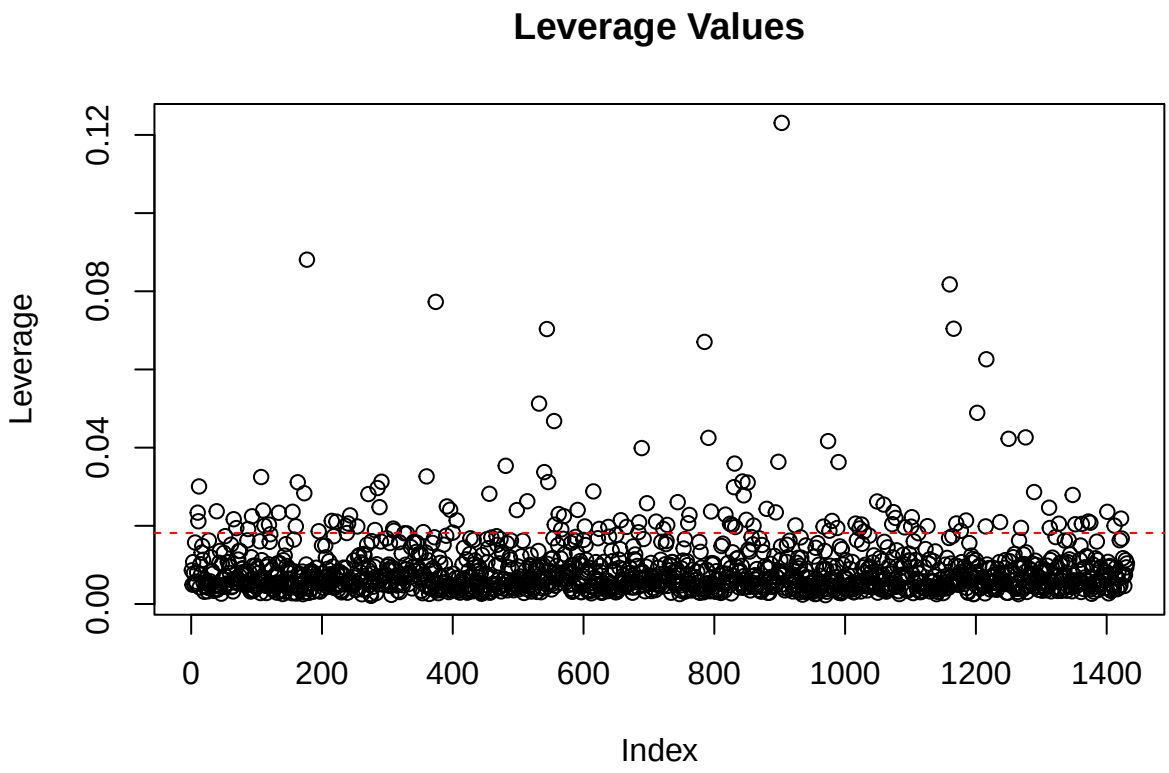
Although we have determined that there are some outliers, we need to see whether or not those are influential on the model:

```
# Influence through leverage
leverage <- hatvalues(linmod_c)
threshold <- 2 * (length(coef(linmod_c)) / nrow(health_dat))

# Find high leverage observations
which(leverage > threshold)
```

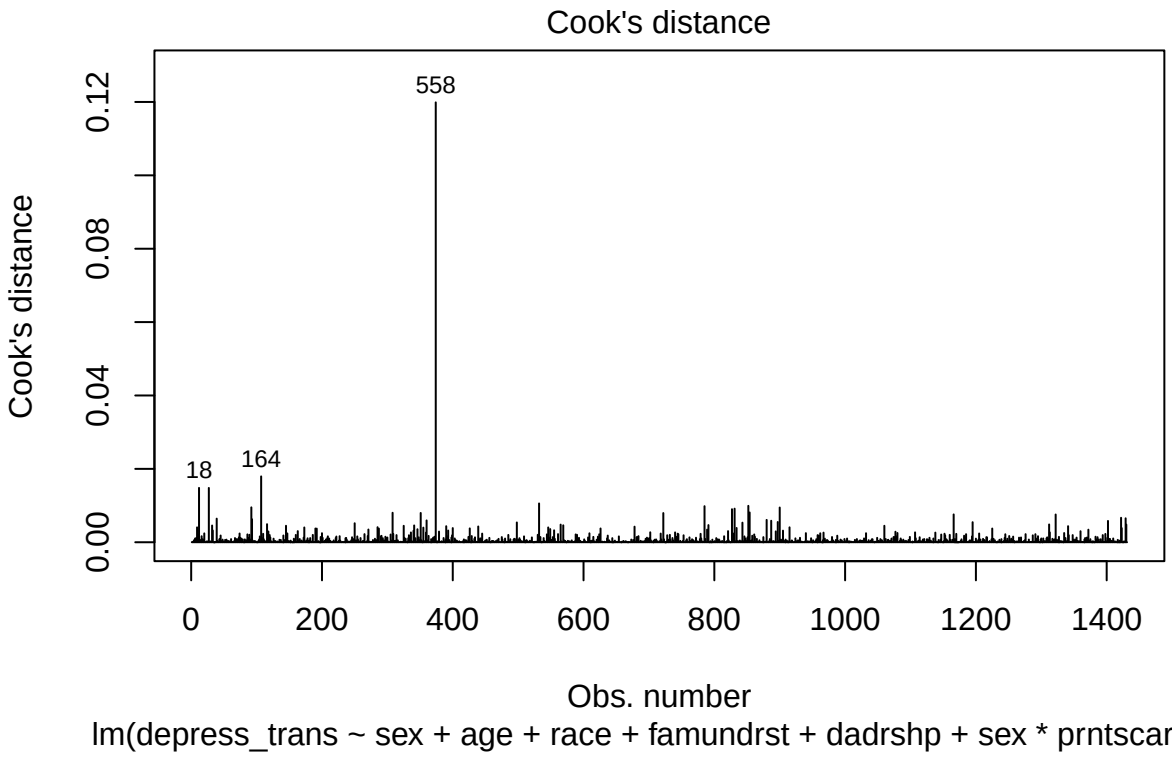
```
##   13   16   18   57   95  100  133  140  164  168  171  182  210  238  245  248
##   10   11   12   39   65   69   87   93  107  110  112  119  134  155  160  163
##  260  266  296  327  335  357  363  367  387  407  420  425  428  431  462  463
##  173  177  195  215  222  236  240  243  254  271  281  285  288  291  309  310
##  528  535  558  581  589  595  608  680  718  744  767  792  801  808  811  824
##  355  360  374  391  396  400  406  456  481  498  514  532  540  544  546  555
##  825  839  845  852  884  897  911  927  953  977  988 1012 1013 1019 1032 1053
##  556  562  566  570  591  602  615  624  638  657  666  684  685  689  697  711
## 1068 1075 1098 1121 1124 1167 1176 1181 1214 1228 1234 1239 1240 1241 1257 1260
##  722  728  744  760  762  785  791  795  817  824  827  830  831  832  843  845
## 1264 1266 1280 1320 1341 1347 1355 1383 1445 1454 1456 1463 1478 1480 1514 1524
```

```
##  849  851  860  880  894  898  903  924  967  974  976  980  988  990 1016 1022
## 1527 1534 1561 1575 1592 1594 1599 1618 1631 1632 1669 1721 1733 1741 1755 1773
## 1025 1029 1049 1059 1072 1074 1077 1091 1101 1102 1126 1160 1166 1170 1177 1185
## 1798 1816 1819 1852 1871 1899 1911 1931 1966 1967 1992 2026 2032 2043 2058 2063
## 1202 1215 1216 1237 1250 1269 1276 1289 1312 1313 1327 1348 1352 1362 1372 1375
## 2099 2119 2135
## 1401 1412 1422
```

```r
plot(leverage, main = "Leverage Values", ylab = "Leverage")
abline(h = threshold, col = "red", lty = 2)
```




```r
# Influence through Cook's distance
plot(linmod_c, which = 4)
```



Based on the leverage plot, we can see that there are quite a lot of values that have leverage above the threshold, which could be potentially problematic. To check the influence, Cook's distance was plotted.

Looking at Cook's distance plot, we see that there is only one truly problematic point with high influence on the model. Next, I need to decide whether it should be removed from the model (in case it is a mistake) or not (in case it is a valid data point).

```
# Identifying the observation
which.max(cooks.distance(linmod_c))
```

```
## 558
## 374
```

Observation named 558 has an index of 374. This observation has a 0 score for depression, which is rare, but not impossible, considering this observation does not have everything else at 0, thus it is unlikely an error in data, but rather a rare case. While it might distort the model to some extent, I would suggest not to remove this observation, since it is still a valid case.

Relative Importance Sets

Although previously I have determined significant variables for the model through iteration process, here I will use relative importance sets to see whether I was correct when iterating the model. I will start the analysis with the first model I did.

```
# Compute the relative importance of each predictor
importance <- calc.relimp(linmod, type = "lmg", rela = TRUE)

# Print the relative importance
print(importance)
```

```
## Response variable: depress
## Total response variance: 47.14279
## Analysis based on 1431 observations
##
## 15 Regressors:
## Some regressors combined in groups:
##         Group  race : raceHispanic raceAsian raceBlack raceOther
##
##  Relative importance of 12 (groups of) regressors assessed:
##  race sex age frndscare prntscare famundrst momcare dadcare momrshp dadrshp esteem intlgnce
##
## Proportion of variance explained by model: 27.8%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                   lmg
## race      0.028598529
## sex       0.007759342
## age       0.038760027
## frndscare 0.024691208
## prntscare 0.032907213
## famundrst 0.109884061
## momcare   0.036802807
## dadcare   0.062675571
## momrshp   0.069735181
## dadrshp   0.100106225
## esteem    0.452470804
## intlgnce  0.035609032
##
## Average coefficients for different model sizes:
##
##                  1group     2groups    3groups    4groups    5groups    6groups
## sex           0.9358837  0.82521474  0.7315079  0.6509585  0.5805011  0.5179080
## age           0.5569137  0.49288194  0.4487053  0.4175036  0.3950035  0.3785400
## raceHispanic  1.9894797  1.84312885  1.7376308  1.6573984  1.5937745  1.5418929
## raceAsian     2.1774367  1.70519509  1.3969885  1.1903046  1.0470246  0.9440913
## raceBlack    -0.1193676  0.03237221  0.1526708  0.2516249  0.3358960  0.4098479
## raceOther     1.5556201  1.56438675  1.5703023  1.5733510  1.5734689  1.5706342
## frndscare    -1.4425282 -1.16338790 -0.9611755 -0.8095310 -0.6914284 -0.5959766
## prntscare    -2.7972659 -2.16855753 -1.6884335 -1.3205093 -1.0371858 -0.8177831
## famundrst    -2.1194739 -1.82634026 -1.5964821 -1.4129801 -1.2639761 -1.1411857
## momcare      -2.8915288 -2.28043565 -1.8404521 -1.5218940 -1.2895851 -1.1187566
## dadcare      -2.6834259 -2.21696435 -1.8520186 -1.5654272 -1.3391303 -1.1591526
## momrshp      -2.1940983 -1.81695590 -1.5118926 -1.2609134 -1.0508398 -0.8721091
## dadrshp      -2.1587914 -1.84685492 -1.5935010 -1.3837222 -1.2067534 -1.0549298
## esteem       -0.9484493 -0.89823622 -0.8583150 -0.8260050 -0.7993513 -0.7769438
## intlgnce     -0.9371036 -0.82653473 -0.7445091 -0.6802201 -0.6271964 -0.5815847
##                  7groups    8groups    9groups   10groups   11groups   12groups
## sex           0.4616228  0.4105367  0.3638039  0.3207238  0.2806872  0.2431606
## age           0.3664134  0.3574926  0.3509834  0.3462984  0.3429864  0.3406940
## raceHispanic  1.4989168  1.4630999  1.4332959  1.4086944  1.3886716  1.3727042
```

```
## raceAsian     0.8674842  0.8085113  0.7616266  0.7231770  0.6906745  0.6623563
## raceBlack     0.4763584  0.5373724  0.5942554  0.6480022  0.6993522  0.7488535
## raceOther     1.5648914  1.5563469  1.5451603  1.5315369  1.5157214  1.4979899
## frndscare    -0.5162024 -0.4475760 -0.3870711 -0.3325970 -0.2826669 -0.2362023
## prntscare    -0.6469543 -0.5133767 -0.4087082 -0.3267843 -0.2630262 -0.2140156
## famundrst    -1.0387936 -0.9526526 -0.8797184 -0.8176631 -0.7646224 -0.7190369
## momcare      -0.9919927 -0.8970009 -0.8250242 -0.7697403 -0.7265191 -0.6919253
## dadcare      -1.0147553 -0.8977390 -0.8018803 -0.7224846 -0.6560354 -0.5999229
## momrshp      -0.7178282 -0.5830421 -0.4641795 -0.3586439 -0.2645184 -0.1803581
## dadrshp      -0.9228046 -0.8064846 -0.7031430 -0.6106755 -0.5274661 -0.4522300
## esteem       -0.7577737 -0.7411206 -0.7264662 -0.7134308 -0.7017292 -0.6911400
## intlgnce     -0.5410636 -0.5041952 -0.4700572 -0.4380406 -0.4077354 -0.3788623
```

```
# Other model iterations
importance <- calc.relimp(linmod1, type = "lmg", rela = TRUE)
print(importance)
```

```
## Response variable: depress
## Total response variance: 47.14279
## Analysis based on 1431 observations
##
## 13 Regressors:
## Some regressors combined in groups:
##         Group  race : raceHispanic raceAsian raceBlack raceOther
##
##  Relative importance of 10 (groups of) regressors assessed:
##  race sex age famundrst momcare dadcare momrshp dadrshp esteem intlgnce
##
## Proportion of variance explained by model: 27.72%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                   lmg
## race      0.030424718
## sex       0.005849538
## age       0.039371308
## famundrst 0.120063392
## momcare   0.042921935
## dadcare   0.072941145
## momrshp   0.075323421
## dadrshp   0.106026168
## esteem    0.468280679
## intlgnce  0.038797696
##
## Average coefficients for different model sizes:
##
##                  1group     2groups    3groups    4groups    5groups    6groups
## sex           0.9358837  0.73421692  0.5862733  0.4749450  0.3888782  0.3206912
## age           0.5569137  0.48631457  0.4378594  0.4044318  0.3812736  0.3652446
## raceHispanic  1.9894797  1.84068205  1.7310929  1.6472329  1.5810459  1.5277558
## raceAsian     2.1774367  1.74808759  1.4590475  1.2587242  1.1143536  1.0057565
## raceBlack    -0.1193676  0.08169647  0.2349113  0.3564378  0.4566847  0.5422725
## raceOther     1.5556201  1.59704010  1.6176507  1.6244464  1.6215438  1.6113808
## famundrst    -2.1194739 -1.80813678 -1.5632790 -1.3683685 -1.2113141 -1.0833436
```

```
## momcare      -2.8915288 -2.27337794 -1.8273964 -1.5054497 -1.2718835 -1.1008746
## dadcare      -2.6834259 -2.19992208 -1.8233013 -1.5299305 -1.3006052 -1.1199709
## momrshp      -2.1940983 -1.78454796 -1.4483561 -1.1702783 -0.9380641 -0.7420590
## dadrshp      -2.1587914 -1.81871953 -1.5369870 -1.3013196 -1.1021137 -0.9319917
## esteem       -0.9484493 -0.89561827 -0.8526946 -0.8175353 -0.7884546 -0.7641490
## intlgnce     -0.9371036 -0.82675354 -0.7423017 -0.6744108 -0.6171320 -0.5667554
##                 7groups    8groups    9groups   10groups
## sex           0.2655781  0.2202952  0.1825106  0.1504875
## age           0.3542639  0.3469109  0.3421658  0.3392743
## raceHispanic  1.4844800  1.4493911  1.4212280  1.3990002
## raceAsian     0.9207093  0.8517537  0.7942038  0.7450576
## raceBlack     0.6174390  0.6849683  0.7467406  0.8040051
## raceOther     1.5954812  1.5749065  1.5505157  1.5231367
## famundrst    -0.9780889 -0.8908528 -0.8180347 -0.7566937
## momcare      -0.9741072 -0.8787735 -0.8058949 -0.7489670
## dadcare      -0.9759789 -0.8593820 -0.7632707 -0.6826525
## momrshp      -0.5748169 -0.4307299 -0.3056807 -0.1967222
## dadrshp      -0.7853779 -0.6581013 -0.5470354 -0.4497801
## esteem       -0.7436238 -0.7261218 -0.7110554 -0.6979453
## intlgnce     -0.5209802 -0.4783557 -0.4379415 -0.3991339
```

```
importance <- calc.relimp(linmod3, type = "lmg", rela = TRUE)
print(importance)
```

```
## Response variable: depress
## Total response variance: 47.14279
## Analysis based on 1431 observations
##
## 12 Regressors:
## Some regressors combined in groups:
##         Group  race : raceHispanic raceAsian raceBlack raceOther
##
##  Relative importance of 9 (groups of) regressors assessed:
##  race sex age famundrst dadrshp prntscare esteem intlgnce sex:prntscare
##
## Proportion of variance explained by model: 27.89%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                       lmg
## race           0.03007797
## sex            0.01012343
## age            0.04237419
## famundrst      0.13488682
## dadrshp        0.12854941
## prntscare      0.07586374
## esteem         0.50959570
## intlgnce       0.03907964
## sex:prntscare 0.02944910
##
## Average coefficients for different model sizes:
##
##                    1group    2groups    3groups    4groups    5groups    6groups
## sex            0.9358837  0.7763080  1.3350885  2.6013650  4.4263727  6.5248161
```

```
## age            0.5569137  0.4915702  0.4432522  0.4112054  0.3918800  0.3789318
## raceHispanic   1.9894797  1.8300122  1.7079375  1.6204328  1.5611229  1.5179568
## raceAsian      2.1774367  1.7492321  1.4425898  1.2207652  1.0605049  0.9473146
## raceBlack     -0.1193676  0.1024269  0.2802073  0.4125283  0.5046502  0.5725860
## raceOther      1.5556201  1.5365967  1.5213745  1.5115464  1.5106732  1.5207791
## famundrst     -2.1194739 -1.8320978 -1.5857714 -1.3859177 -1.2329428 -1.1142217
## dadrshp       -2.1587914 -1.8472577 -1.5707627 -1.3396285 -1.1595121 -1.0197205
## prntscare     -2.7972659 -2.2907879 -1.8743487 -1.5019776 -1.1442974 -0.7967238
## esteem        -0.9484493 -0.9030586 -0.8631800 -0.8297759 -0.8034757 -0.7828461
## intlgnce      -0.9371036 -0.8136932 -0.7137178 -0.6368831 -0.5821242 -0.5437347
## sex:prntscare        NaN        NaN -3.0677842 -2.8626377 -2.7026885 -2.5799437
##                   7groups    8groups     9groups
## sex            8.5707741 10.3196995 11.66210495
## age            0.3664319  0.3533815  0.34183835
## raceHispanic   1.4760280  1.4289970  1.38228978
## raceAsian      0.8692098  0.8123272  0.76437172
## raceBlack      0.6403193  0.7219687  0.81347867
## raceOther      1.5373305  1.5508340  1.55523642
## famundrst     -1.0038512 -0.8845311 -0.76174793
## dadrshp       -0.8896882 -0.7454789 -0.59160847
## prntscare     -0.4747938 -0.1986048  0.02065261
## esteem        -0.7636179 -0.7421387 -0.71864938
## intlgnce      -0.5091524 -0.4669750 -0.41537410
## sex:prntscare -2.4868875 -2.4165039 -2.36232263
```

```
importance <- calc.relimp(linmod_c, type = "lmg", rela = TRUE)
print(importance)
```

```
## Response variable: depress_trans
## Total response variance: 1.017937
## Analysis based on 1431 observations
##
## 12 Regressors:
## Some regressors combined in groups:
##         Group  race : raceHispanic raceAsian raceBlack raceOther
##
##  Relative importance of 9 (groups of) regressors assessed:
##  race sex age famundrst dadrshp prntscare_centered esteem intlgnce sex:prntscare_centered
##
## Proportion of variance explained by model: 26.55%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                                lmg
## race                   0.032594765
## sex                    0.008595136
## age                    0.062091493
## famundrst              0.139192739
## dadrshp                0.128811834
## prntscare_centered     0.051701481
## esteem                 0.500612362
## intlgnce               0.060946711
## sex:prntscare_centered 0.015453478
##
```

```
## Average coefficients for different model sizes:
##
##                              1group       2groups     3groups     4groups
## sex                       0.1246067  0.1018395542  0.08524912  0.07281332
## age                       0.0923397  0.0832323588  0.07651405  0.07208926
## raceHispanic              0.2933433  0.2693430365  0.25065325  0.23704929
## raceAsian                 0.3254744  0.2667202539  0.22553152  0.19691459
## raceBlack                -0.0319487 -0.0001725225  0.02535895  0.04435743
## raceOther                 0.2343369  0.2325678653  0.23137898  0.23084730
## famundrst                -0.3035021 -0.2633607694 -0.22919147 -0.20193727
## dadrshp                  -0.3075191 -0.2633589508 -0.22425341 -0.19191698
## prntscare_centered       -0.3503309 -0.2749127264 -0.21484440 -0.16361193
## esteem                   -0.1351673 -0.1285876767 -0.12275415 -0.11787857
## intlgnce                 -0.1578860 -0.1407871920 -0.12701794 -0.11654859
## sex:prntscare_centered          NaN           NaN -0.34027511 -0.31049893
##                              5groups     6groups     7groups      8groups
## sex                       0.06243933  0.05265707  0.04300816  0.033781741
## age                       0.06943710  0.06763630  0.06585421  0.063987387
## raceHispanic              0.22772583  0.22090161  0.21416492  0.206354969
## raceAsian                 0.17760144  0.16505786  0.15664590  0.149889022
## raceBlack                 0.05759129  0.06741243  0.07727855  0.089150561
## raceOther                 0.23127413  0.23280388  0.23495022  0.236768939
## famundrst                -0.18165763 -0.16630739 -0.15179849 -0.135504877
## dadrshp                  -0.16722914 -0.14840762 -0.13056602 -0.110017443
## prntscare_centered       -0.11667914 -0.07252915 -0.03223127  0.002337374
## esteem                   -0.11409213 -0.11114597 -0.10829146 -0.104906613
## intlgnce                 -0.10917355 -0.10399894 -0.09922242 -0.093318997
## sex:prntscare_centered -0.28757737 -0.27027346 -0.25743491 -0.247997536
##                              9groups
## sex                       0.02540278
## age                       0.06235385
## raceHispanic              0.19826965
## raceAsian                 0.14343192
## raceBlack                 0.10236102
## raceOther                 0.23764921
## famundrst                -0.11839387
## dadrshp                  -0.08763004
## prntscare_centered        0.03012451
## esteem                   -0.10106387
## intlgnce                 -0.08613324
## sex:prntscare_centered -0.24099331
```

Looking at the output we can see that the initial iteration process when building a model produced similar results in terms of determining significant and influential variables for the model. The importance and significance of the variables in different iterations of the model were discussed above.

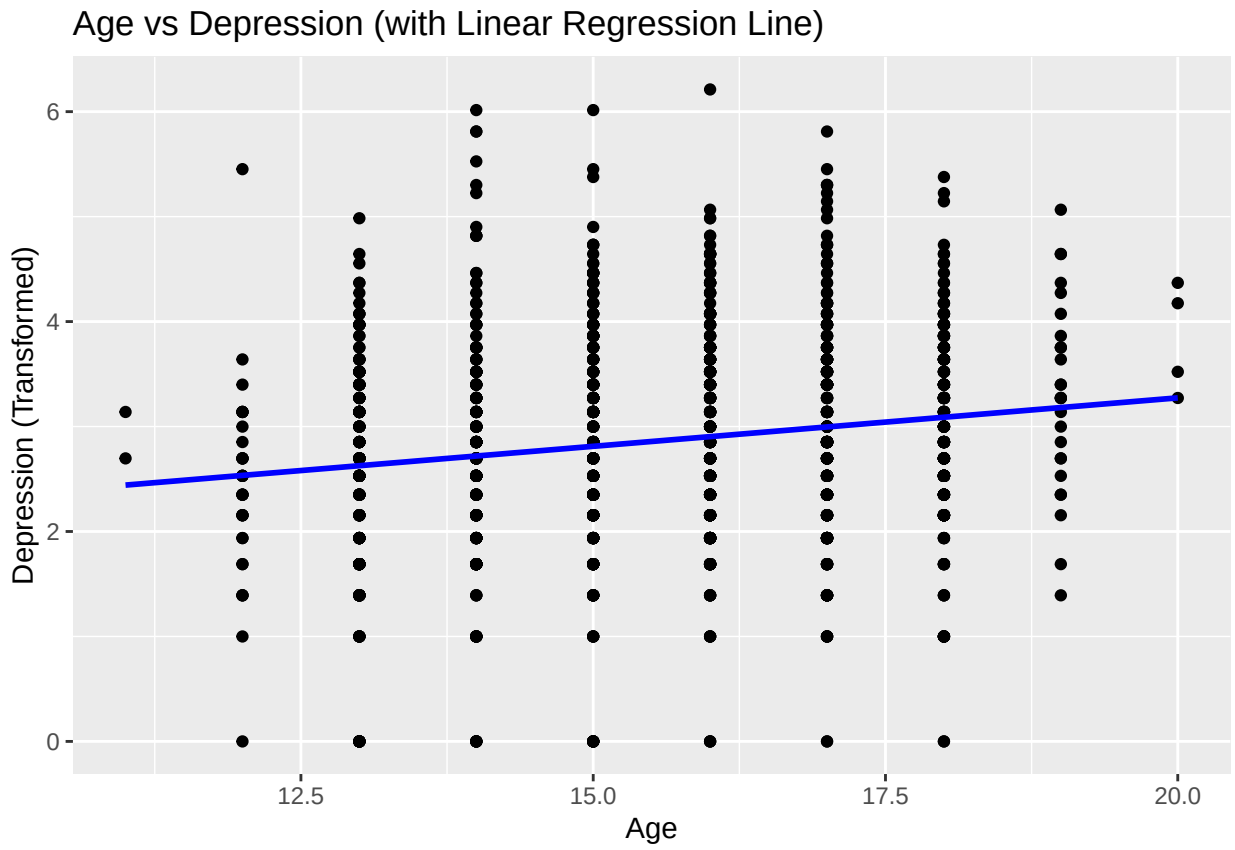
Model Plot

Because the final model has interaction, I cannot use crPlots function, thus it will plotted as follows:

```
# Scatter plot for continuous predictors vs. depress_trans (with regression line)
ggplot(health_dat, aes(x = age, y = depress_trans)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
```

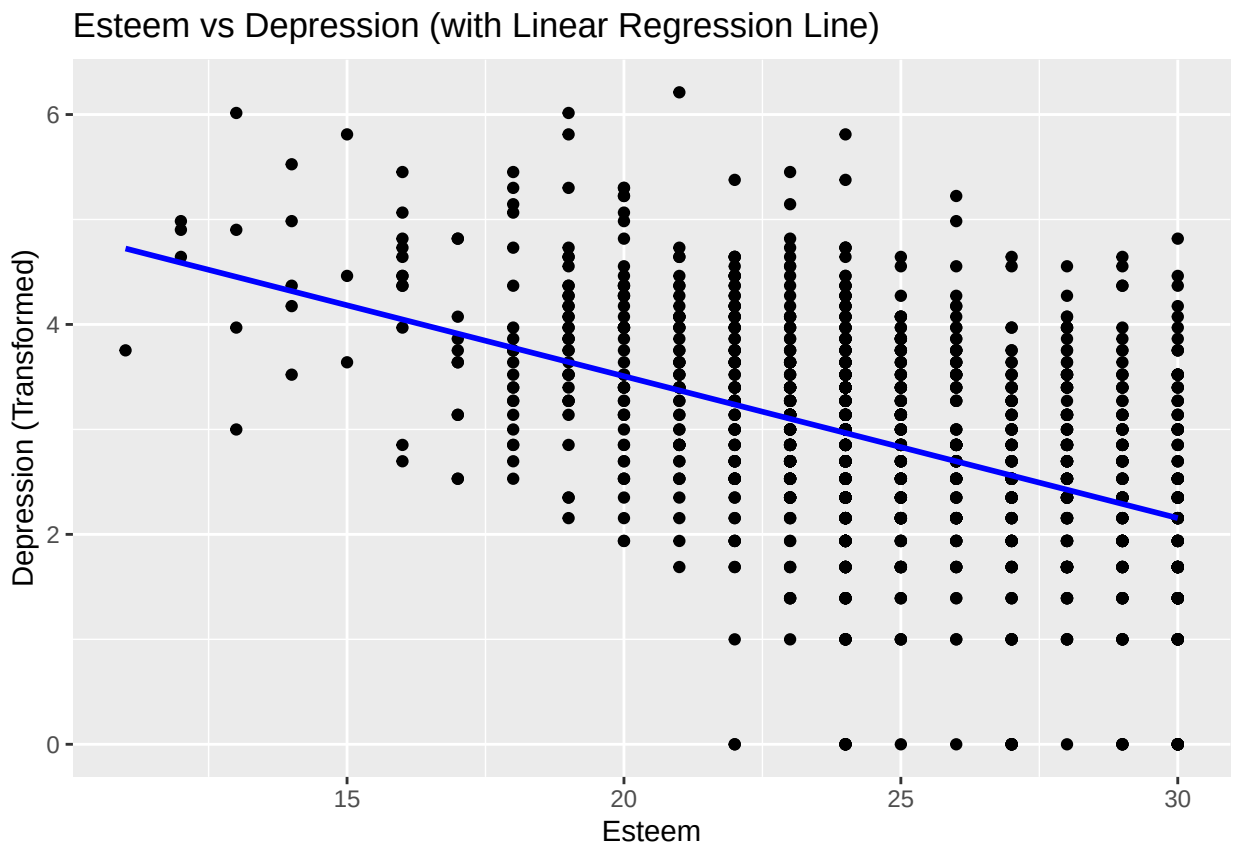
```
  labs(title = "Age vs Depression (with Linear Regression Line)", x = "Age",
       y = "Depression (Transformed)")
```

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



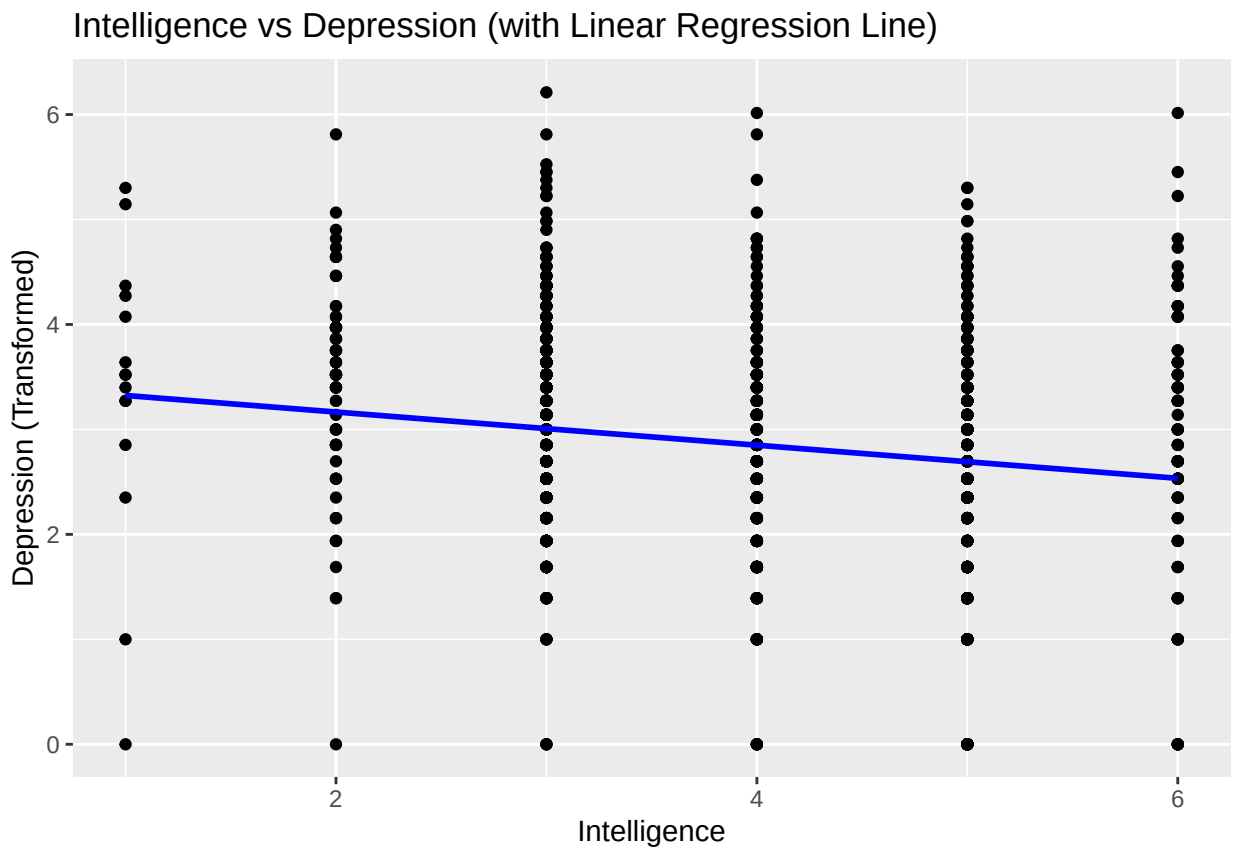

```
ggplot(health_dat, aes(x = esteem, y = depress_trans)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(title = "Esteem vs Depression (with Linear Regression Line)",
       x = "Esteem", y = "Depression (Transformed)")
```

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

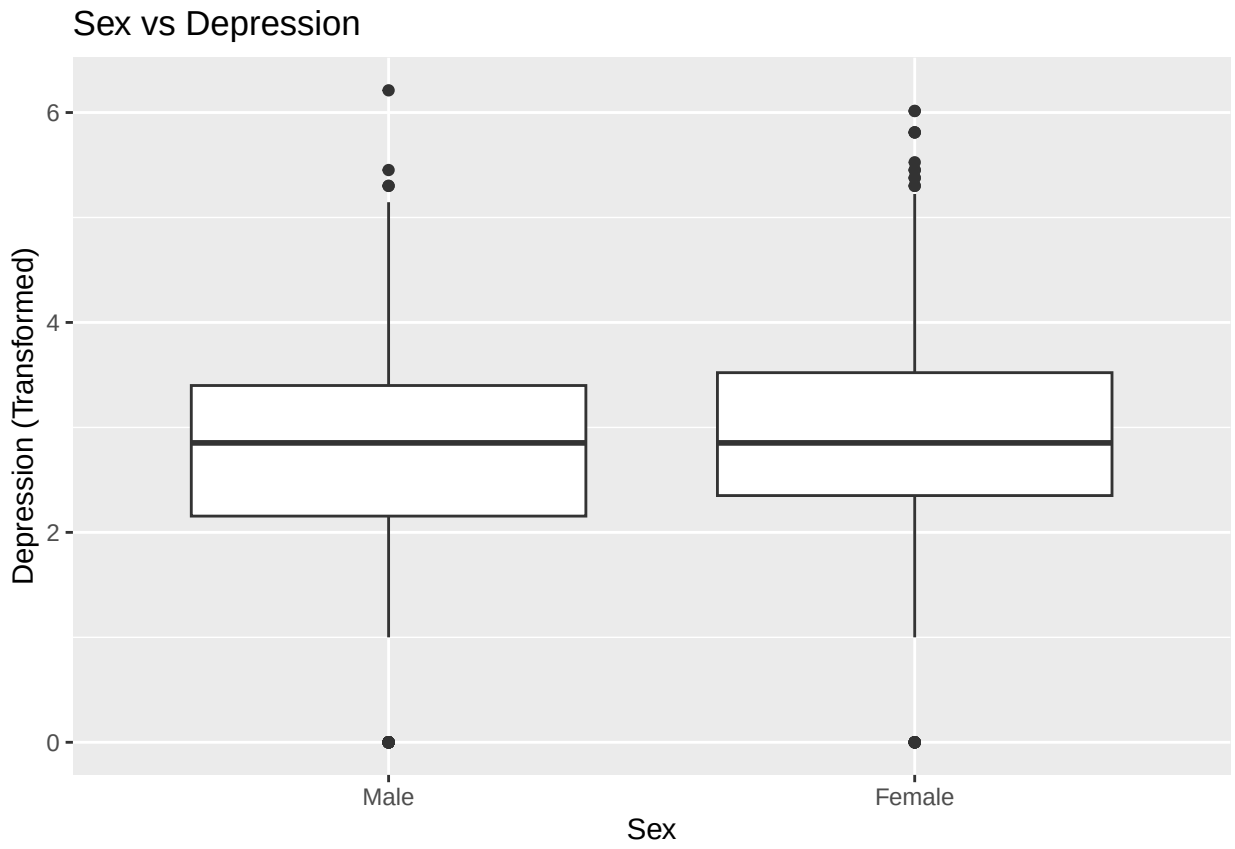



```
ggplot(health_dat, aes(x = intlgnce, y = depress_trans)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE) +
  labs(title = "Intelligence vs Depression (with Linear Regression Line)",
       x = "Intelligence", y = "Depression (Transformed)")
```

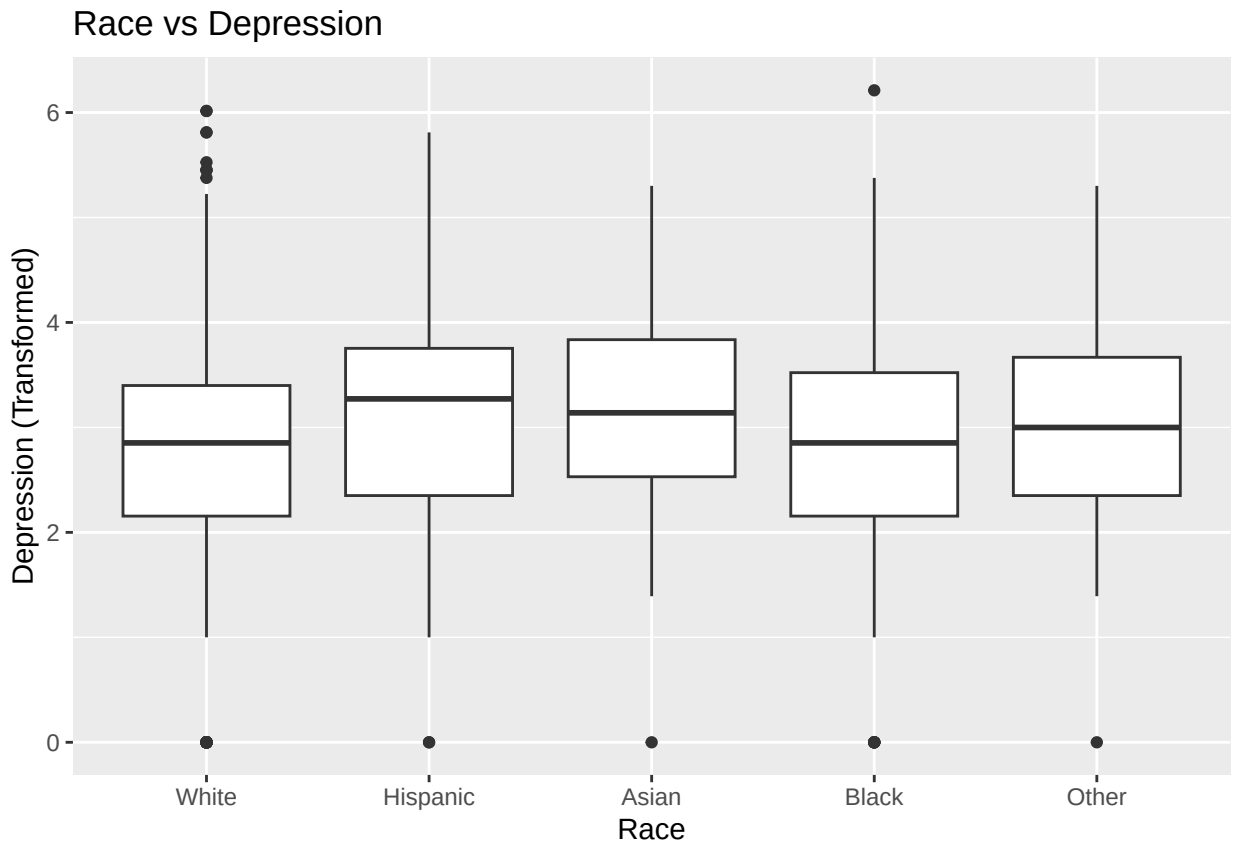
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Box plot for categorical predictors
ggplot(health_dat, aes(x = sex, y = depress_trans)) +
  geom_boxplot() +
  labs(title = "Sex vs Depression", x = "Sex", y = "Depression (Transformed)")
```

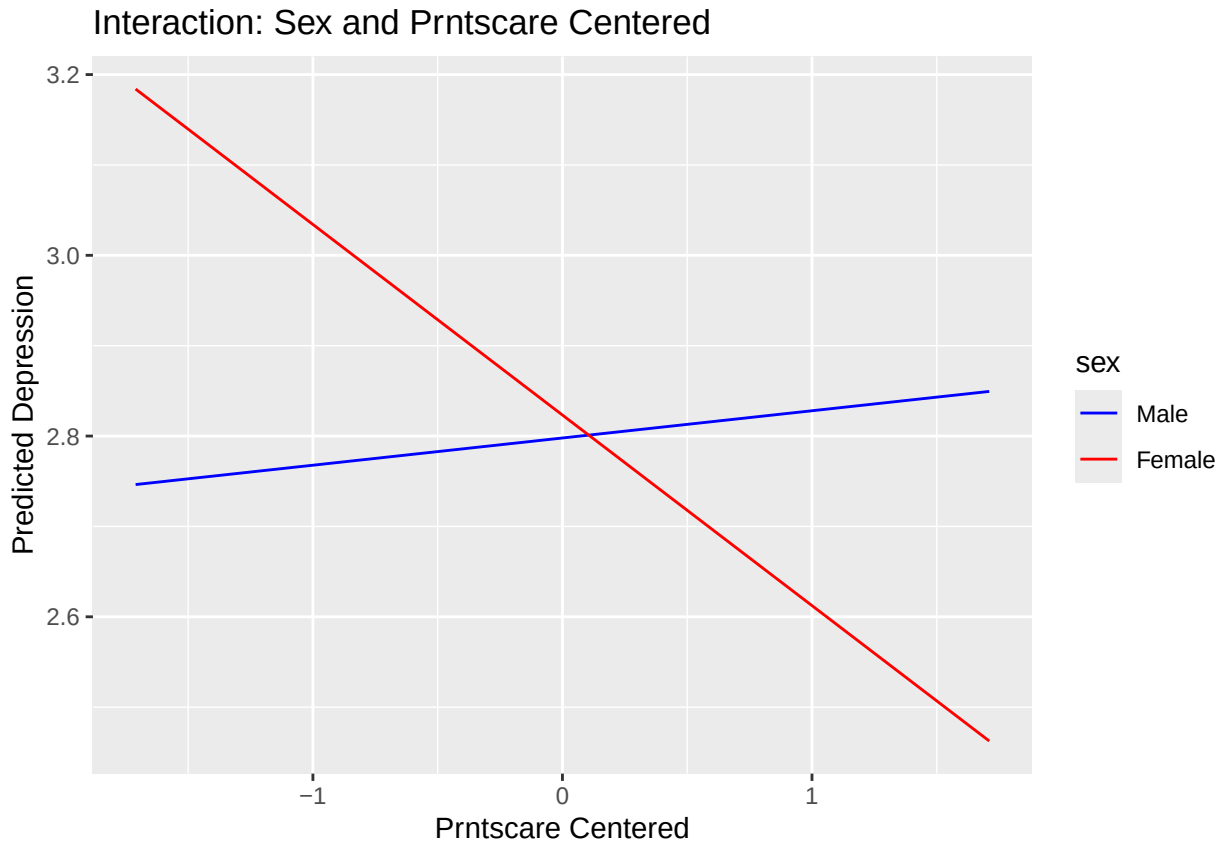



```
ggplot(health_dat, aes(x = race, y = depress_trans)) +
  geom_boxplot() +
  labs(title = "Race vs Depression", x = "Race", y = "Depression (Transformed)")
```




```r
# Visualize the interaction term between sex and prntscare_centered
# Generate predictions for different values of prntscare_centered
# (e.g., mean, -1 SD, +1 SD)
prntscare_vals <- seq(from = min(health_dat$prntscare_centered),
                      to = max(health_dat$prntscare_centered), length.out = 100)

# Create a new data frame with different values of prntscare_centered and sex
interaction_data <- expand.grid(prntscare_centered = prntscare_vals,
                                sex = unique(health_dat$sex))

# Add the missing predictor variables (e.g., age, race, famundrst, etc.)
interaction_data$age <- mean(health_dat$age, na.rm = TRUE)
interaction_data$race <- factor("White", levels = levels(health_dat$race))
interaction_data$famundrst <- mean(health_dat$famundrst, na.rm = TRUE)
interaction_data$dadrshp <- mean(health_dat$dadrshp, na.rm = TRUE)
interaction_data$esteem <- mean(health_dat$esteem, na.rm = TRUE)
interaction_data$intlgnce <- mean(health_dat$intlgnce, na.rm = TRUE)

# Scale the prntscare_centered variable for newdata (interaction_data)
interaction_data$prntscare_centered <-
  scale(interaction_data$prntscare_centered)

# Now make predictions (scaled prntscare_centered)
interaction_data$predicted_depress <-
  predict(linmod_c, newdata = interaction_data)
```

```r
# Plot the interaction between sex and prntscare_centered
ggplot(interaction_data, aes(x = prntscare_centered,
                             y = predicted_depress, color = sex)) +
  geom_line() + labs(title = "Interaction: Sex and Prntscare Centered",
                     x = "Prntscare Centered", y = "Predicted Depression") +
  scale_color_manual(values = c("blue", "red"))
```



Assumption Check on the Final Model

Model level assumptions

Completeness:

Because true models are never known, the model should be based on theory. The control variables and the variables of interest were chosen based either on existing articles or my assumptions, which are discussed above. Considering that I am studying depression from the perspective of human relationships, it is likely that the model will not be able to explain all the variance, as there are a lot more other factors that affect the severity of depression.

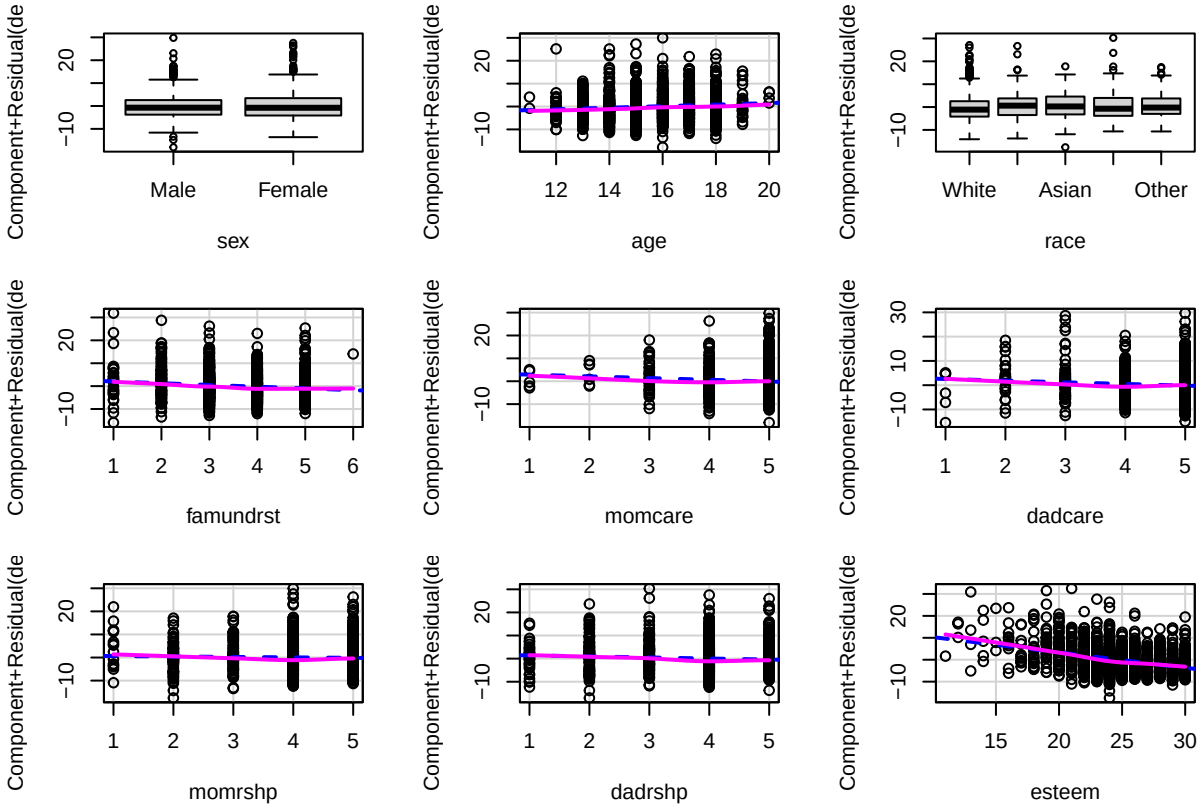
Additivity:

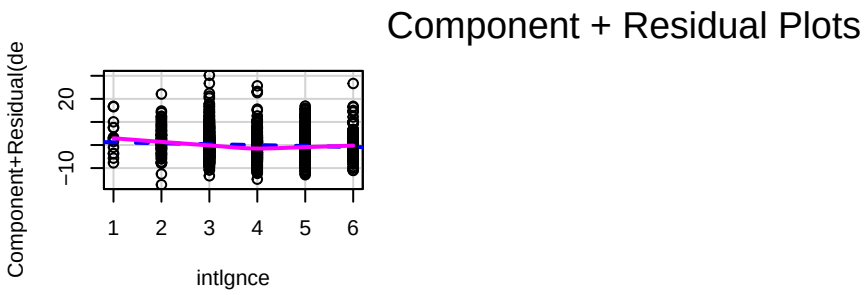
The model is additive, although through iteration one interaction effect was added to take into consideration the relationship between the sex variable and the prntscare variable (interaction effects do not affect this assumption).

Linearity:

Based on the graphs below, there are no clear linearity assumption violations. Although the purple lines are

not ideally linear, it is acceptable to model the relationship as linear for modeling purposes, as there are rarely pure linear relationships in real data. Categorical variables also have an expected behavior. Esteem seems to have the clearest linear relation to depress. (Assumption checked on the initial model's set of variables)





Variable level assumptions

Variables measured without error:

This assumption is difficult to access because the data was collected by other researchers. For the sake of the model, we have to assume that they have measured the variables without error or, more realistically, with small enough errors.

Variables measured at an interval or ratio scale: This assumption is met based on the dataset description and the type of model was chosen based on the fact that the dependent variable and independent variables are measured using either interval or ratio scales.

Error level assumptions

Normal distribution:

Skewness is negative, but less than -1, which is a small enough value to consider that the assumption is met

```
resid <- residuals(linmod_c)
skewness(resid)
```

```
## [1] -0.260847
```

Zero-mean assumption

The residuals' mean is very close to zero, thus, the assumption is satisfied.

```
mean(resid)
```

```
## [1] 2.155858e-17
```

Non-independence/autocorrelation:

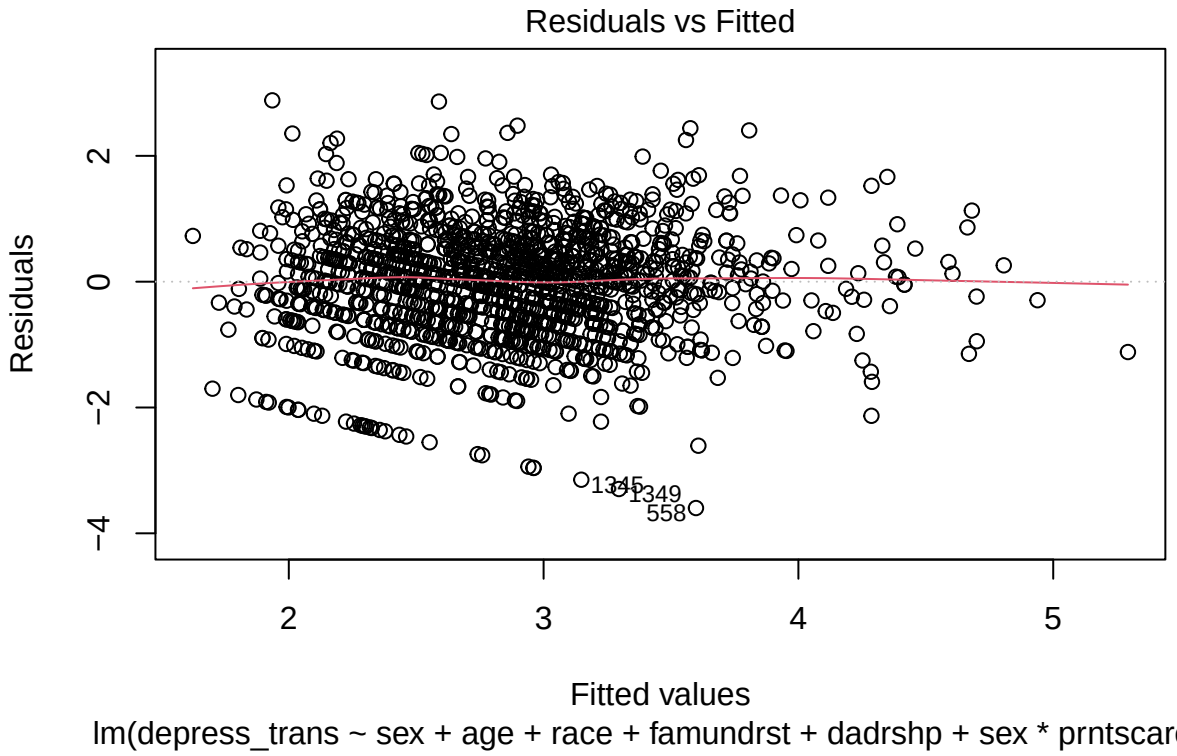
The Durbin-Watson test result is slightly above 2, with a p-value bigger than 0.05, suggesting that residuals are most likely independent and the assumption is satisfied.

```
dwtest(linmod_c)
```

```
##
##  Durbin-Watson test
##
## data:  linmod_c
## DW = 2.0388, p-value = 0.7679
## alternative hypothesis: true autocorrelation is greater than 0
```

Homoscedasticity: The initial model's residual vs fitted plot had a clear cone-like shape, suggesting heteroscedasticity. After using spreadlevelPlot, the response variable was transformed with a power of transformation around 0.477, which has resulted in a better plot, suggesting that the homoscedasticity assumption is satisfied at least to some extent.

```
plot(linmod_c, which = 1)
```



Predictors unrelated to errors:

From a theoretical standpoint, reciprocal causation is unlikely but still possible (for example, between depress and esteem). For the sake of the model building, we have to assume independence.

Errors unrelated to each other:

Because I am only dealing with one model equation, this assumption is irrelevant for this case.

Conclusion

I have built a model to see whether family and friends relationships are correlated with depression. Based on the results of the model, we reject H0 and accept H1, which states that at least one of the variables in the model has a significant linear relationship with depress variable.

Better relationships with parents (specifically dads) and understanding of a family are connected with decrease in depression, while relationships with friends do not seem to have any explanatory power over depression. Another interesting result is that parents' care matters more for females in terms of depression reduction.

References

Girgus, J. S., & Yang, K. (2015). Gender and depression. Current Opinion in Psychology, 4, 53–60. https://doi.org/10.1016/j.copsyc.2015.01.019

Martin, G. (1996). Depression in teenagers. Current Therapeutics, 37, 57-67.

Patil, P. A., Porche, M. V., Shippen, N. A., Dallenbach, N. T., & Fortuna, L. R. (2018). Which girls, which boys? the intersectional risk for depression by race and ethnicity, and gender in the U.S. Clinical Psychology Review, 66, 51–68. https://doi.org/10.1016/j.cpr.2017.12.003

Shokrgozar, S., Khesht-Masjedi, M., Abdollahi, E., Habibi, B., Asghari, T., Ofoghi, R., & Pazhooman, S. (2019). The relationship between gender, age, anxiety, depression, and academic achievement among teenagers. Journal of Family Medicine and Primary Care, 8(3), 799. https://doi.org/10.4103/jfmpc.jfmpc_103_18