Verifying Assumptions

Zander Bonnet

2024-05-08

```
miniPONS <- read.csv("~/Desktop/GCU/DSC_520/Data/Database MiniPONS.csv", sep=";", string
sAsFactors=TRUE)
sum(is.na(miniPONS))
```

```
## [1] 0
```

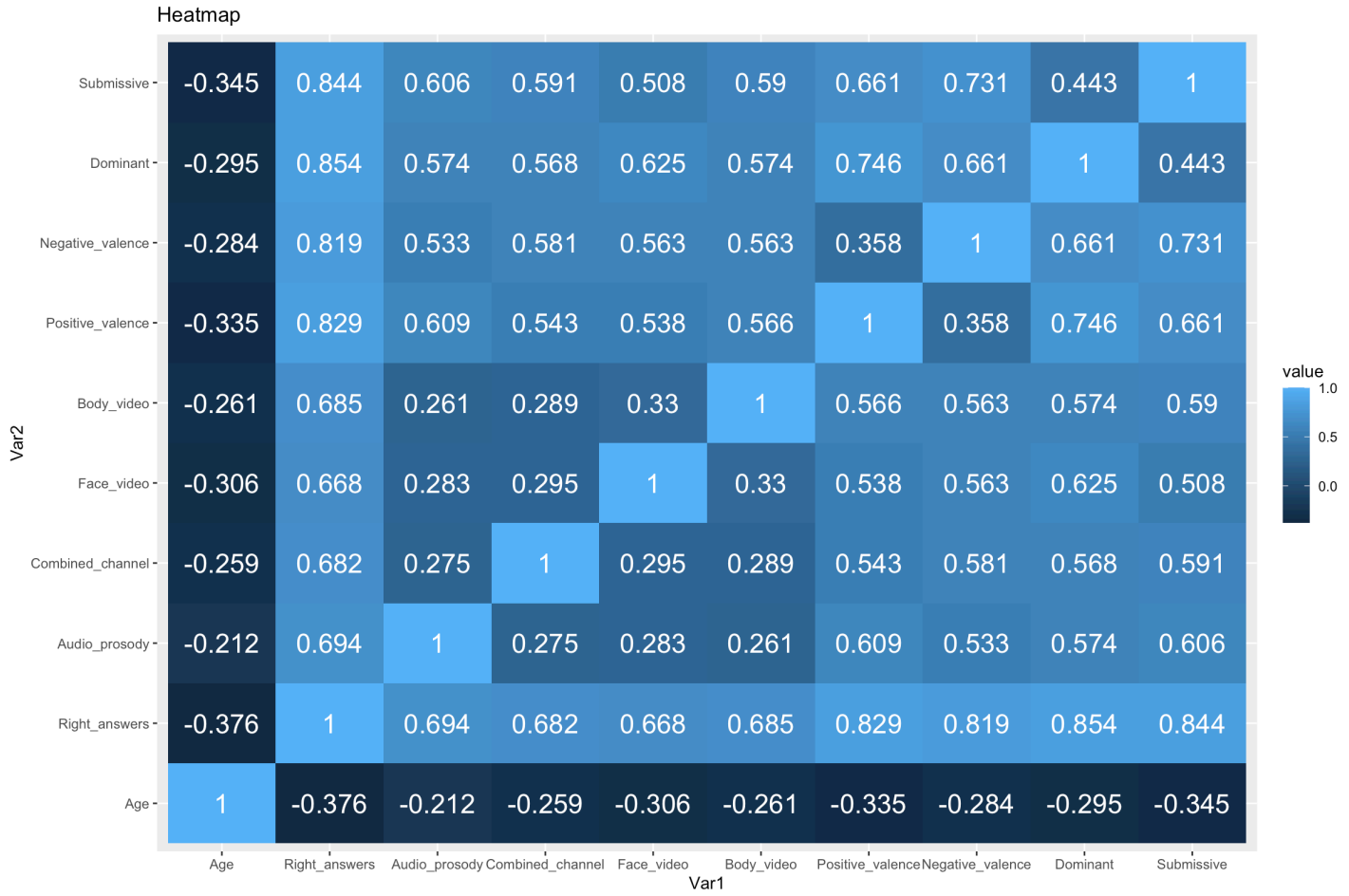
```
head(miniPONS)
```

```
##     Group  Type Age Right_answers Audio_prosody Combined_channel Face_video
## 1 Bipolar  BD I  47            40             9               11          9
## 2 Bipolar  BD I  49            49            13               13         11
## 3 Bipolar  BD I  45            43             9               11         13
## 4 Bipolar  BD I  53            44            10               10         12
## 5 Bipolar BD II  50            50            14               13         11
## 6 Bipolar  BD I  31            54            13               14         14
##   Body_video Positive_valence Negative_valence Dominant Submissive
## 1         11               18               22       23         17
## 2         12               24               25       24         25
## 3         10               21               22       24         19
## 4         12               25               19       24         20
## 5         12               23               27       23         27
## 6         13               28               26       26         28
```

There are no missing values.

```
library(ggplot2)
library(reshape2)

corr_mat <- round(cor(miniPONS[,-c(1:2)]),3)
melt_corr_mat <- melt(corr_mat)
plt <- ggplot(data = melt_corr_mat, aes(x = Var1, y = Var2, fill = value))
plt <- plt + geom_tile()
plt <- plt + geom_text(aes(Var2, Var1, label = value), color = "white", size = 6)
plt <- plt + labs(title = 'Heatmap')
plt
```



There are some moderatly correlated variables that could lead to possible coliniarity.

```
mod <- lm(Right_answers ~., data = miniPONS)
summary(mod)
```

```
##
## Call:
## lm(formula = Right_answers ~ ., data = miniPONS)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.331e-12 -7.100e-15  2.780e-15  1.476e-14  5.727e-14
##
## Coefficients: (4 not defined because of singularities)
##                     Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -1.693e-13  6.688e-14 -2.531e+00   0.0120 *
## GroupControl      5.269e-15  1.436e-14  3.670e-01   0.7140
## GroupUD           1.866e-14  1.895e-14  9.850e-01   0.3255
## TypeBD II         1.552e-14  1.601e-14  9.700e-01   0.3331
## TypeControl              NA         NA         NA       NA
## TypeUD                   NA         NA         NA       NA
## Age               1.276e-16  4.890e-16  2.610e-01   0.7944
## Audio_prosody     1.000e+00  3.376e-15  2.962e+14   <2e-16 ***
## Combined_channel  1.000e+00  3.531e-15  2.832e+14   <2e-16 ***
## Face_video        1.000e+00  4.216e-15  2.372e+14   <2e-16 ***
## Body_video        1.000e+00  3.591e-15  2.785e+14   <2e-16 ***
## Positive_valence  4.132e-15  2.845e-15  1.452e+00   0.1476
## Negative_valence         NA         NA         NA       NA
## Dominant         -7.257e-15  3.215e-15 -2.257e+00   0.0248 *
## Submissive               NA         NA         NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.365e-14 on 266 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.08e+29 on 10 and 266 DF,  p-value: < 2.2e-16
```

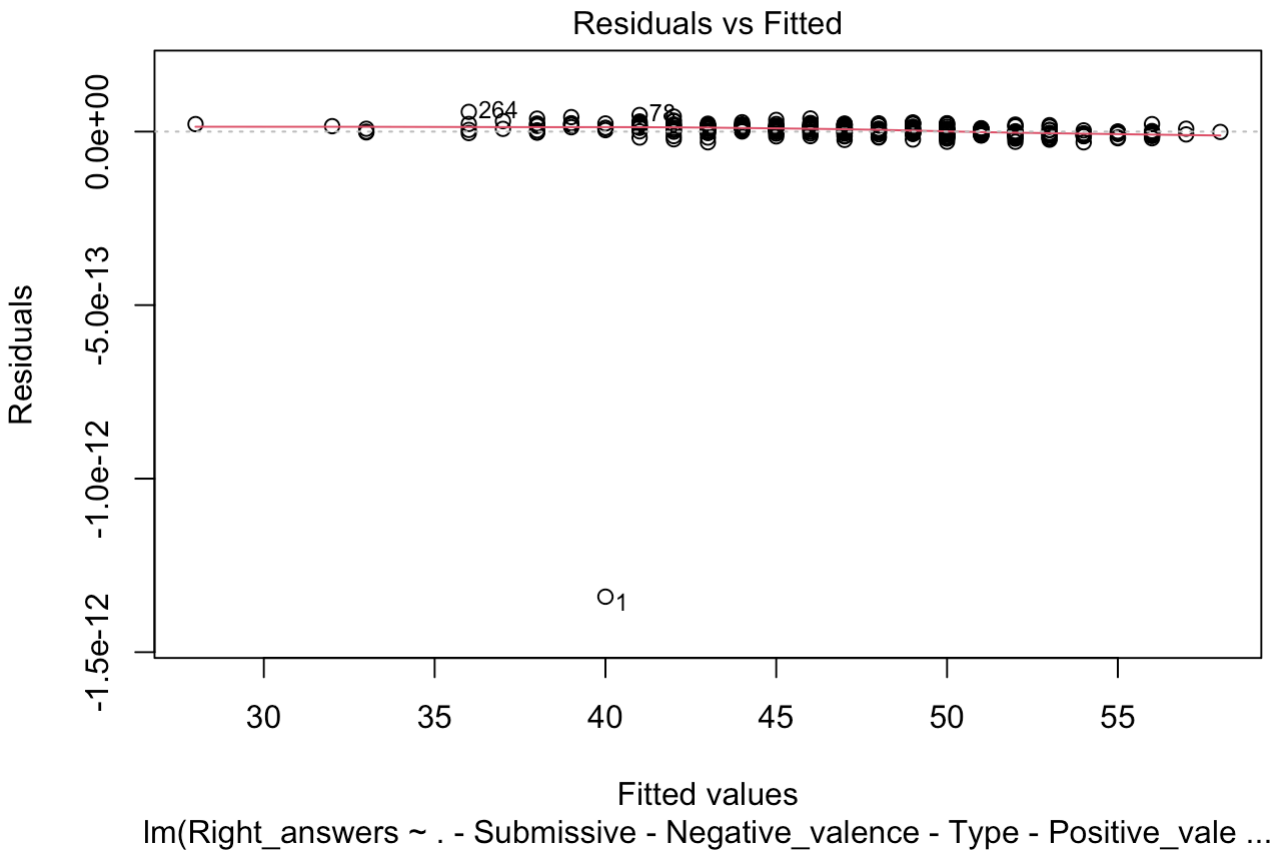
There are several insignificant factors.

```
mod2 <- lm(Right_answers ~. - Submissive - Negative_valence -
             Type - Positive_valence - Group - Age, data = miniPONS)
summary(mod2)
```

```
##
## Call:
## lm(formula = Right_answers ~ . - Submissive - Negative_valence -
##     Type - Positive_valence - Group - Age, data = miniPONS)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.340e-12 -4.530e-15  4.830e-15  1.496e-14  5.685e-14
##
## Coefficients:
##                    Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -1.622e-13  4.589e-14 -3.535e+00 0.000479 ***
## Audio_prosody     1.000e+00  2.917e-15  3.428e+14  < 2e-16 ***
## Combined_channel  1.000e+00  3.168e-15  3.157e+14  < 2e-16 ***
## Face_video        1.000e+00  3.827e-15  2.613e+14  < 2e-16 ***
## Body_video        1.000e+00  3.212e-15  3.113e+14  < 2e-16 ***
## Dominant         -6.667e-15  3.099e-15 -2.151e+00 0.032352 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.282e-14 on 271 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.203e+29 on 5 and 271 DF,  p-value: < 2.2e-16
```

The model is significant, and has a perfect R-Squared so it almost perfectly explains the variance in the data.

```
plot(mod2, which = 1)
```



The residual values are all very small, but there is one extreme outlier in the residuals.

```
library(car)
```

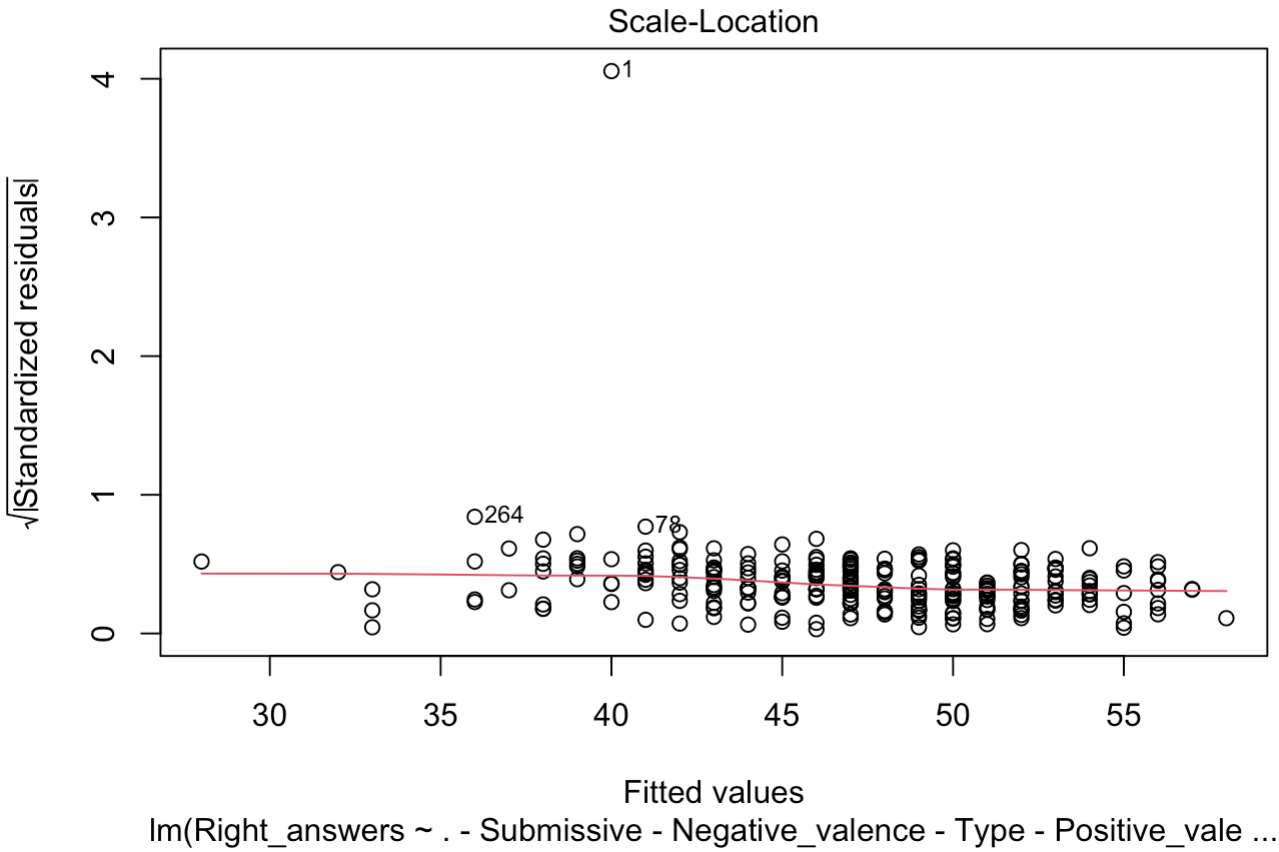
```
## Loading required package: carData
```

```
durbinWatsonTest(mod2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02417833     0.9852903       0
##  Alternative hypothesis: rho != 0
```

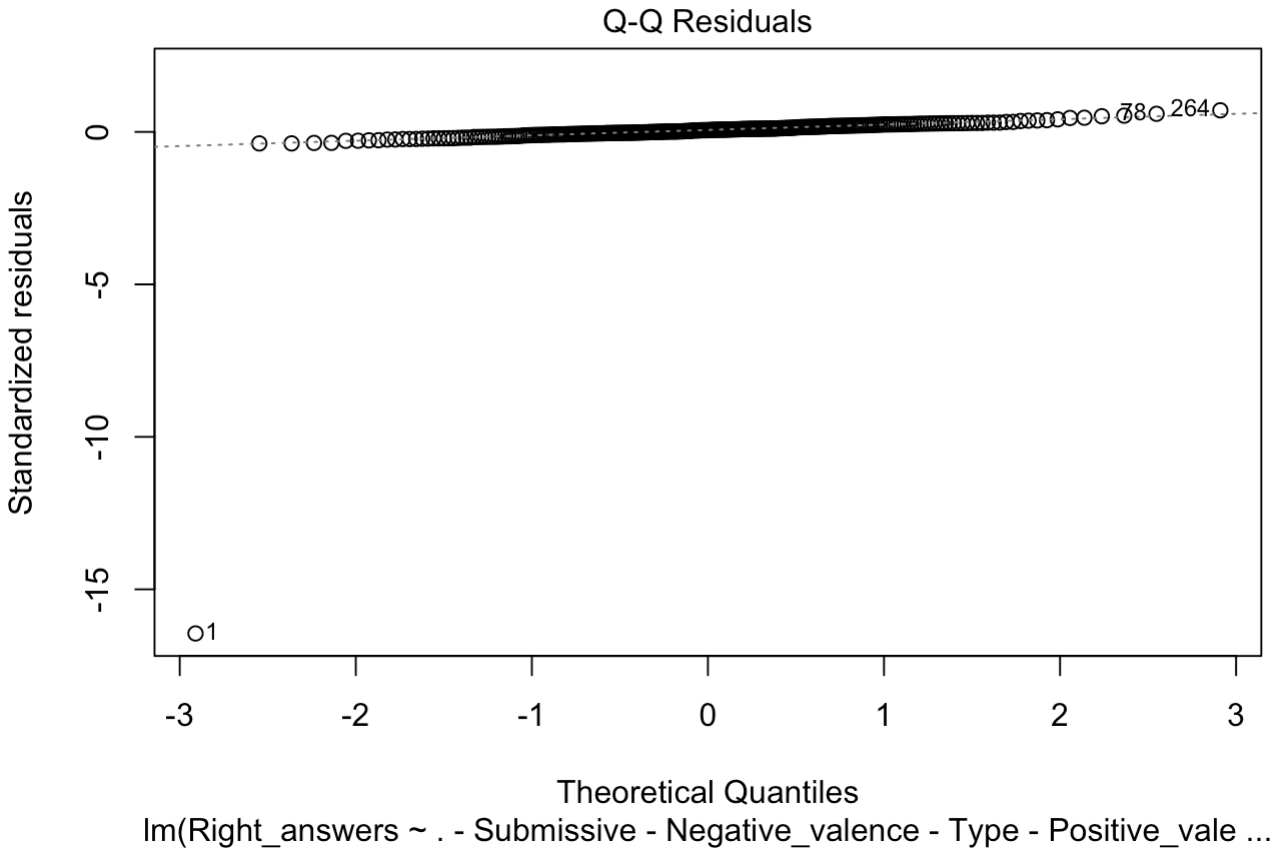
The Durbin Watson Test shows the possibility of a positive autocorrelation in the residuals. Meaning that they cannot be deemed independent. This could lead to some of the predictors to be falsely significant.

```
plot(mod2, which = 3)
```



The standardized residuals show that the residuals are homoscedastic for the most part. They have consistint variance and are evenly spread. Again there is just the one extreme value.

```
plot(mod2, which = 2)
```



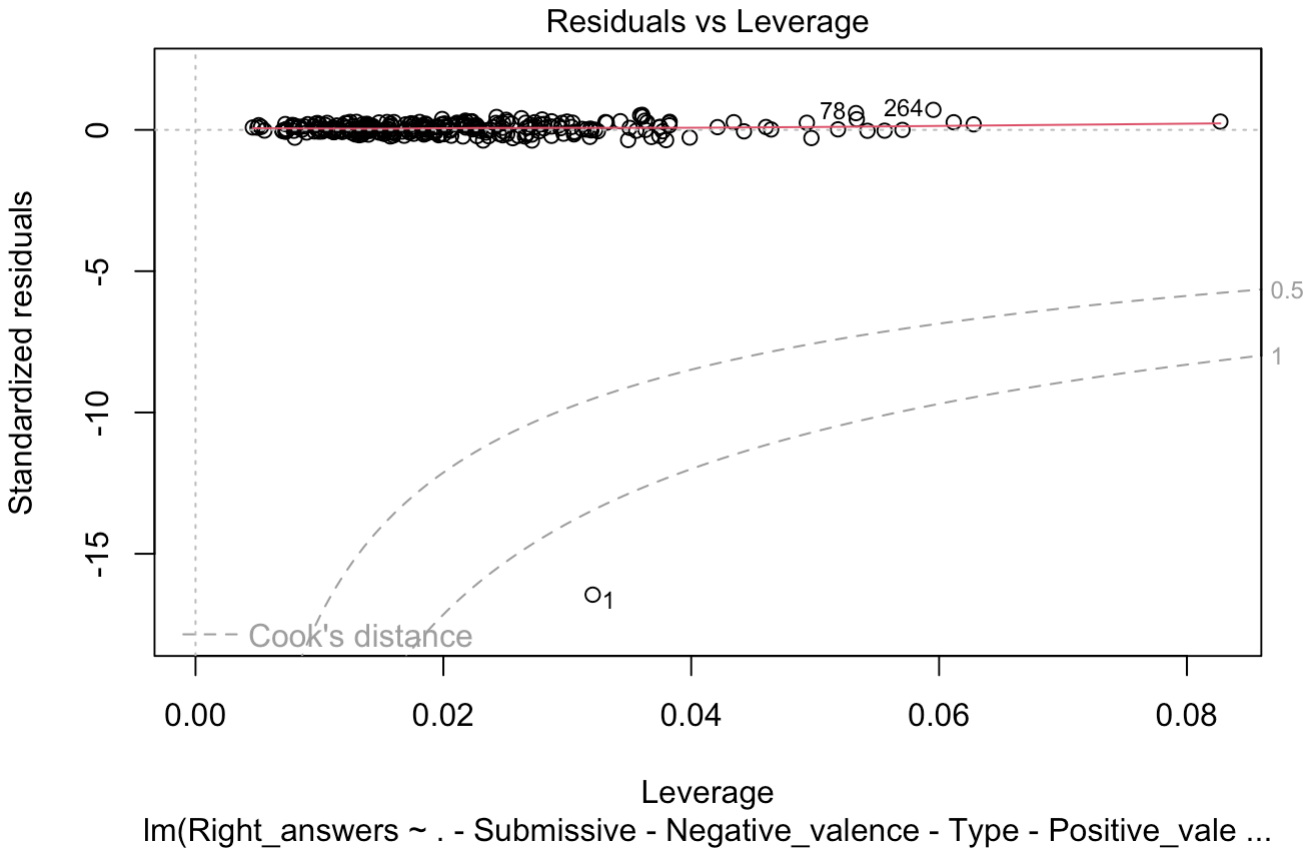
The residuals are normally distributed outside of the extreme value.

```
ncvTest(mod2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 242.7659, Df = 1, p = < 2.22e-16
```

The test shows that there is a non constant variance in the fitted values. This is most likely due to the fact that there is the extreme value creating inconsistencies. This test differs from the Durbin Watson test because NCV tests if the values have a consistent variance, and the Durbin Watson test tests if the residuals have any autocorrelation. Meaning that the residuals depend on the previous residual.

```
plot(mod2, which = 5)
```



There is one value that is largely different from the rest and has a Cook's distance of over 1. This could cause the tests to give false results, and make the models significance be questioned.

```
which(abs(scale(mod2$residuals)) > 3)
```

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

There is one large outlier in the residuals.

Remove the outlier residual

```
#remove outlier
out <- which(abs(scale(mod2$residuals)) > 3)
no_out <- miniPONS[-out,]
```

```
mod3 <- lm(Right_answers ~. - Submissive - Negative_valence -
             Type - Positive_valence - Group - Age, data = no_out)
summary(mod3)
```

```
## Warning in summary.lm(mod3): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = Right_answers ~ . - Submissive - Negative_valence -
##     Type - Positive_valence - Group - Age, data = no_out)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.445e-13  1.120e-16  6.410e-16  9.710e-16  1.031e-14
##
## Coefficients:
##                    Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -6.843e-15  4.969e-15 -1.377e+00    0.170
## Audio_prosody     1.000e+00  3.159e-16  3.166e+15   <2e-16 ***
## Combined_channel  1.000e+00  3.420e-16  2.924e+15   <2e-16 ***
## Face_video        1.000e+00  4.163e-16  2.402e+15   <2e-16 ***
## Body_video        1.000e+00  3.463e-16  2.888e+15   <2e-16 ***
## Dominant          3.504e-17  3.367e-16  1.040e-01    0.917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.922e-15 on 270 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.886e+31 on 5 and 270 DF,  p-value: < 2.2e-16
```

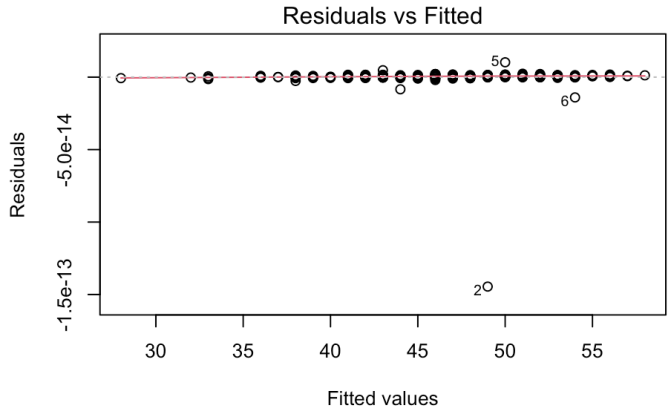
Dominant is no longer significant after the removal of the outlier.

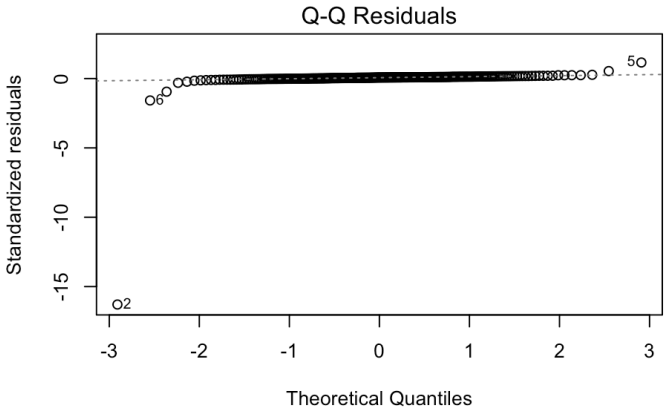
```
mod4 <- lm(Right_answers ~. - Submissive - Negative_valence -
             Type - Positive_valence - Group - Age -Dominant, data = no_out)
summary(mod4)
```

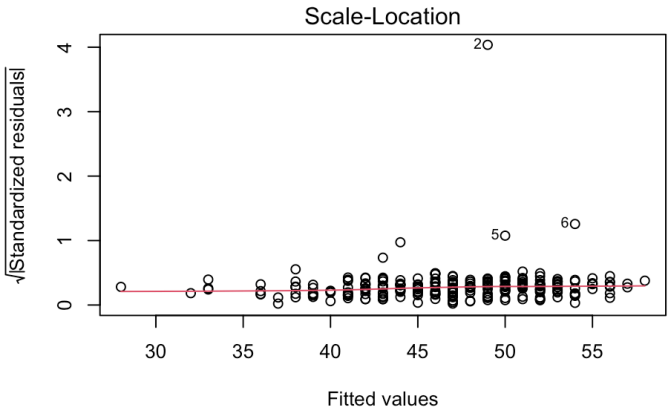
```
## Warning in summary.lm(mod4): essentially perfect fit: summary may be unreliable
```

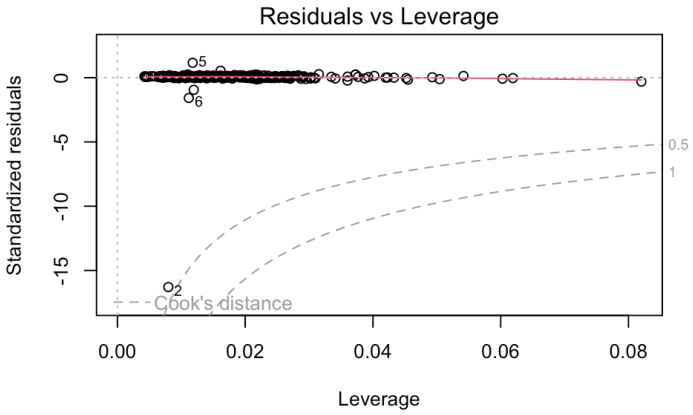
```
##
## Call:
## lm(formula = Right_answers ~ . - Submissive - Negative_valence -
##     Type - Positive_valence - Group - Age - Dominant, data = no_out)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.445e-13  1.090e-16  6.700e-16  9.800e-16  1.025e-14
##
## Coefficients:
##                    Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -6.843e-15  4.947e-15 -1.383e+00    0.168
## Audio_prosody     1.000e+00  2.724e-16  3.671e+15   <2e-16 ***
## Combined_channel  1.000e+00  3.020e-16  3.312e+15   <2e-16 ***
## Face_video        1.000e+00  3.505e-16  2.853e+15   <2e-16 ***
## Body_video        1.000e+00  3.074e-16  3.253e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.905e-15 on 271 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.366e+31 on 4 and 271 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(mod4)
```









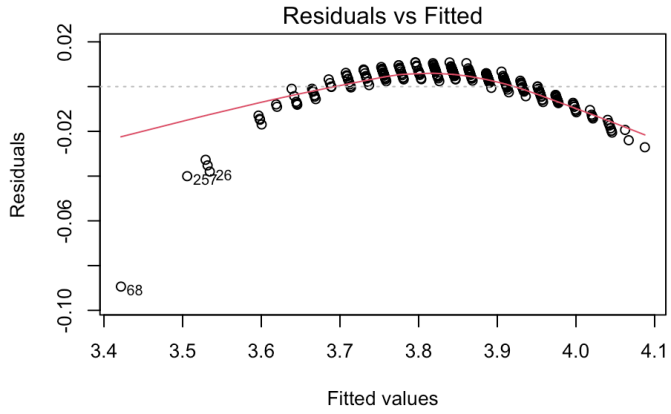
The intercept of the model is not significant.

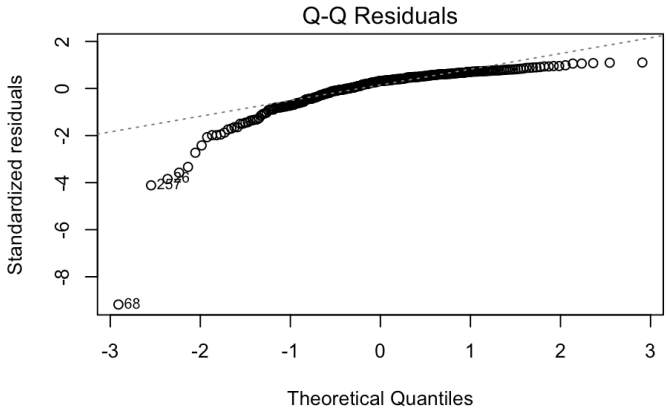
The outlier residual was replaced with the next value.

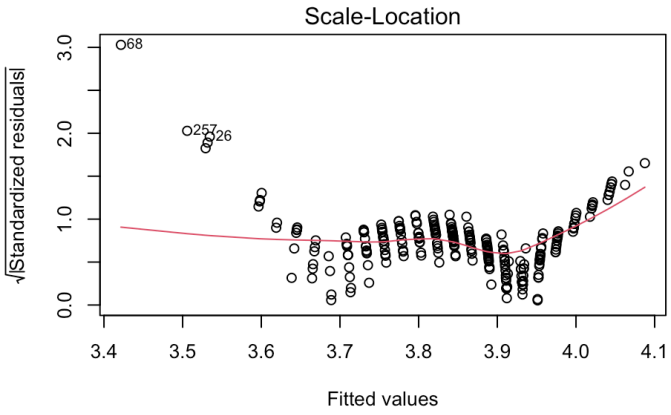
```
mod5 <- lm(log(Right_answers) ~. - Submissive - Negative_valence -
             Type - Positive_valence - Group - Age - Dominant, data = no_out)
summary(mod5)
```

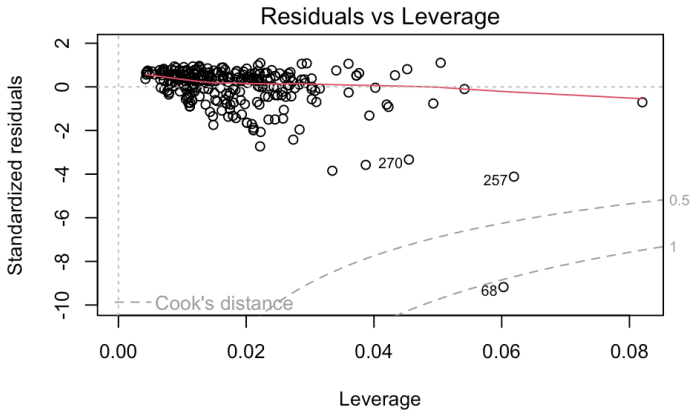
```
##
## Call:
## lm(formula = log(Right_answers) ~ . - Submissive - Negative_valence -
##     Type - Positive_valence - Group - Age - Dominant, data = no_out)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.089371 -0.002949  0.003209  0.005997  0.010863
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.8025657  0.0055798  502.27   <2e-16 ***
## Audio_prosody    0.0219946  0.0003073   71.58   <2e-16 ***
## Combined_channel 0.0226213  0.0003406   66.42   <2e-16 ***
## Face_video       0.0211435  0.0003954   53.47   <2e-16 ***
## Body_video       0.0228242  0.0003467   65.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01004 on 271 degrees of freedom
## Multiple R-squared:  0.9926, Adjusted R-squared:  0.9925
## F-statistic:  9148 on 4 and 271 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(mod5)
```









When trying to perform a log transformation on the response it causes the residuals to now follow a pattern and violate the conditions of the model.

When we try to remove the outliers in the data it is just replaced with another outlier, so we cannot go with that route to fix the model. When we try to apply a transformation to the data it takes away the linear relationship of the data and violates multiple conditions of the model.

```
ncvTest(mod5)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 259.1325, Df = 1, p = < 2.22e-16
```

Even after performing log transforamtion of the data there is non-constant variance in the fitted values, so the assumption of homoscedastity in the model are not met.

Looking at the leverage plot we can see that there are now multiple potential outliers that are dragging the predictions down as we can see by the red line in the plot.

```
#original model
vif(mod2)
```

```
##    Audio_prosody Combined_channel       Face_video       Body_video
##         1.542800         1.507051         1.689296         1.518299
##         Dominant
##         3.777094
```

```
#log transformed
vif(mod5)
```

```
##    Audio_prosody Combined_channel       Face_video       Body_video
##         1.158392         1.182251         1.213939         1.202092
```

There is no evidence of multicoliniarity in the both of the models as the VIF scores are all less than 4.

The log transformed model is not valid to be used. It does not meet the assumptions of the model as the relationship is no longer linear. Since the relationship is not linear it also fails the other assumptions of the model.

If we used the original model we can gather some data from it but it still does not follow the assumptions of the model. There is the one outlier, but even though the outlier is significant it is still very close to the true value. There is the possibility of auto correlation, but the residuals are so small that it might not be significant.

I would recommend using the original model for any predictions and analysis. The data is linear to begin with, so if we use a transformation it takes away the linearity of the data. The original model also does a very good job a predicting values based on the significant factors. There is a chance that the model is overly reported as significant, but the model is very strong.

Reference

Theory of mind in remitted bipolar disorder. (2019). Kaggle [Dataset]. https://www.kaggle.com/datasets/mercheovejero/theory-of-mind-in-remitted-bipolar-disorder/data (https://www.kaggle.com/datasets/mercheovejero/theory-of-mind-in-remitted-bipolar-disorder/data)