ANCOVA

Zander Bonnet

2024-05-15

```r
miniPONS <- read.csv("~/Desktop/GCU/DSC_520/Data/Database MiniPONS.csv", sep=";", string
sAsFactors=TRUE)
sum(is.na(miniPONS))
```
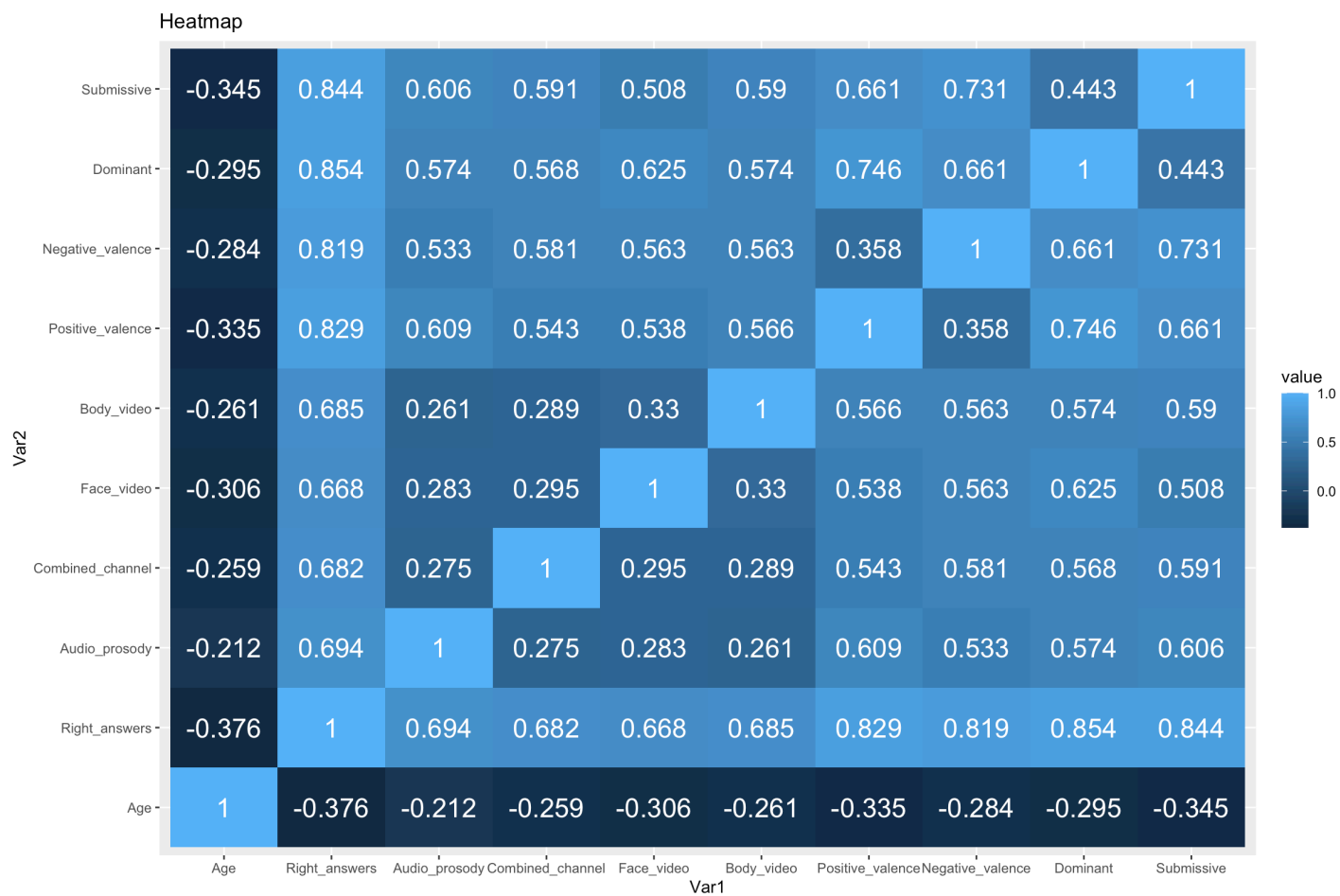
```
## [1] 0
```

```r
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
```

```r
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
```

Heatmap

Var2	Age	Right_answers	Audio_prosody	Combined_channel	Face_video	Body_video	Positive_valence	Negative_valence	Dominant	Submissive
Submissive	-0.345	0.844	0.606	0.591	0.508	0.59	0.661	0.731	0.443	1
Dominant	-0.295	0.854	0.574	0.568	0.625	0.574	0.746	0.661	1	0.443
Negative_valence	-0.284	0.819	0.533	0.581	0.563	0.563	0.358	1	0.661	0.731
Positive_valence	-0.335	0.829	0.609	0.543	0.538	0.566	1	0.358	0.746	0.661
Body_video	-0.261	0.685	0.261	0.289	0.33	1	0.566	0.563	0.574	0.59
Face_video	-0.306	0.668	0.283	0.295	1	0.33	0.538	0.563	0.625	0.508
Combined_channel	-0.259	0.682	0.275	1	0.295	0.289	0.543	0.581	0.568	0.591
Audio_prosody	-0.212	0.694	1	0.275	0.283	0.261	0.609	0.533	0.574	0.606
Right_answers	-0.376	1	0.694	0.682	0.668	0.685	0.829	0.819	0.854	0.844
Age	1	-0.376	-0.212	-0.259	-0.306	-0.261	-0.335	-0.284	-0.295	-0.345

value

```
plt <- ggplot(data = miniPONS, aes(x = Age, y = Right_answers, colour = factor(Type)))
plt <- plt + geom_point() + geom_smooth(method = lm)
plt
```

```
## `geom_smooth()` using formula = 'y ~ x'
```
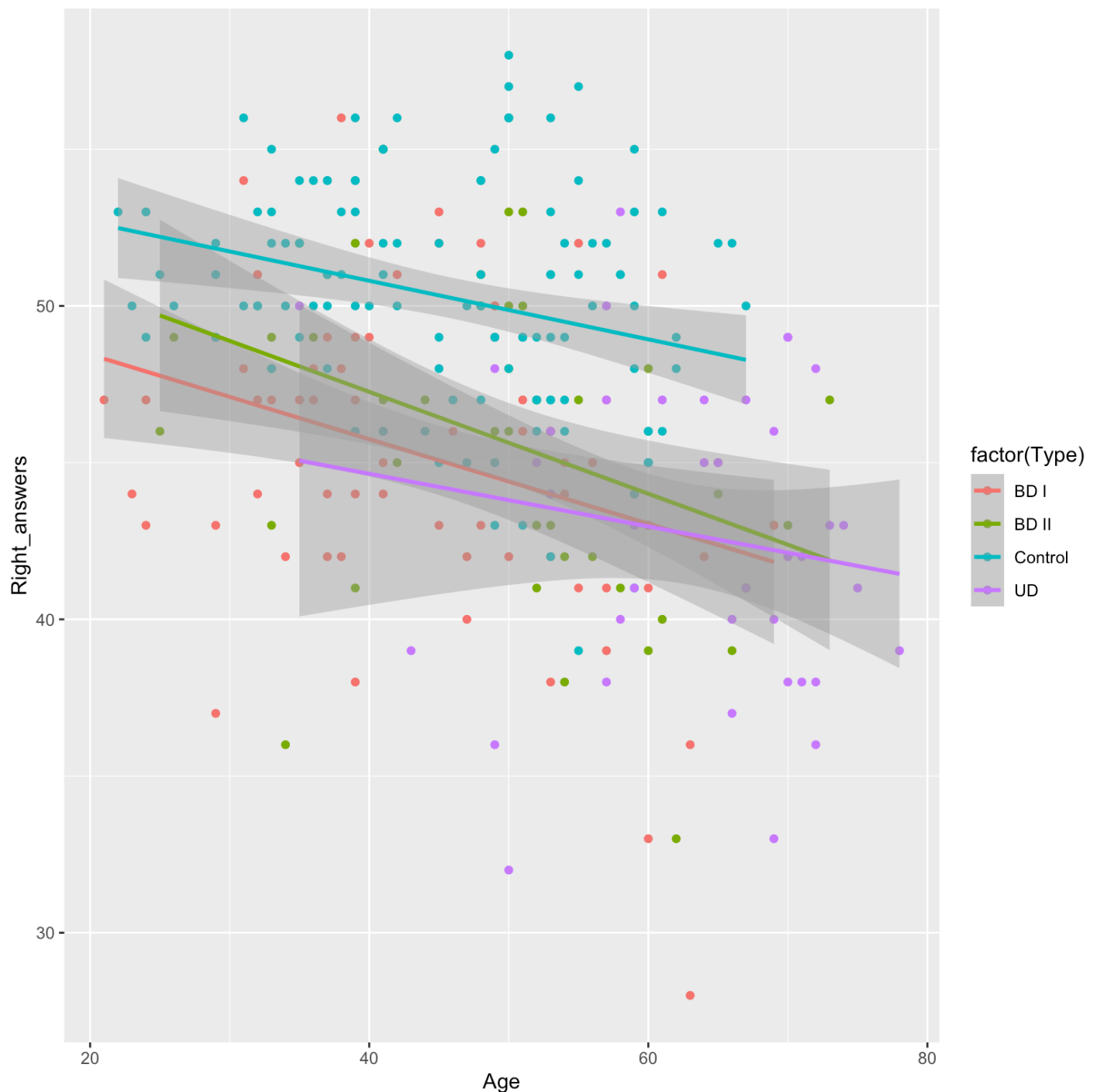
It is feasible that the types are parallel to each other as the BD1, BD2, and control groups are all roughly parallel and the UD group could possibly be parallel as the parallel line is in the error range.

BD1, BD2, and the UD groups have similar trend lines so they may not significantly different.

1. Factor A - The main treatment effect: a) The adjusted population group means are all equal. b) The effect of each group equals zero.

   This is the categorical variable type.

   This means that all the groups means will be equal. In this case the various types will all have a similar mean Right answers. If any one group does not have a similar mean the null hypothesis will be rejected. The adjusted group mean would be the mean of the population minus the mean of each individual group. This would show if one group deviated from the population mean. If the effect of each group was zero then that would mean there was no difference in right answers between groups.

2. Factor B - The covariate effect: a) The pooled population slope equals zero.

   This is true for all covariates. In this analysis they are Age and submisive.

   The null states that the effect of the covariate is zero. This would be true if the covariate has no impact on the amount of right answers. This will be violated if the covariate has an effect size of nonzero and is statistically significant.

In this ANCOVA I will use Age, Type, and submisive to predict the number of right answers.

```
mod <- lm(Right_answers ~ Age + Type + Submissive, data = miniPONS)
summary(mod)
```

```
##
## Call:
## lm(formula = Right_answers ~ Age + Type + Submissive, data = miniPONS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7634 -1.6430  0.2824  1.7313  8.0953
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.02649    1.81336   9.941  < 2e-16 ***
## Age         -0.03742    0.01555  -2.407  0.01674 *
## TypeBD II   -0.16383    0.51655  -0.317  0.75137
## TypeControl  1.34698    0.45645   2.951  0.00344 **
## TypeUD      -0.19966    0.60899  -0.328  0.74327
## Submissive   1.28487    0.06575  19.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.715 on 271 degrees of freedom
## Multiple R-squared:  0.7355, Adjusted R-squared:  0.7306
## F-statistic: 150.7 on 5 and 271 DF,  p-value: < 2.2e-16
```

Age, Submissive, and Type all prove to be significant factors. Type is significant because the control groups mean is significantly different from BD1. This means that the null hypothesis the adjusted group means is zero is rejected, and they must be different. Part b of the hypothesis states that the effect of the constant is 0, but we can see that the effect size is 1.34 for the control group. and that is not 0, so that part of the null is also rejected.

For the covariates the null states that the pulled population slope is equal to zero. This is rejected for both of our covariates because they both have nonzero effect sizes and statistically significant.

1. Age and Submissive are coviariates and Type is a categorical variable.
2. The standard error in the response is 2.715
3. F ratios
   1. BD2 - -.317 F ratio. This is not significantly different in BD1, so it is not significant in predicting Right answers
   2. Control - 2.951 F ratio. This is significantly different then BD1, so it is significant in predicting Right answers.

3. UD - -.328 F ratio. This is not significantly different from BD1, so it is not significant in predicting Right answers
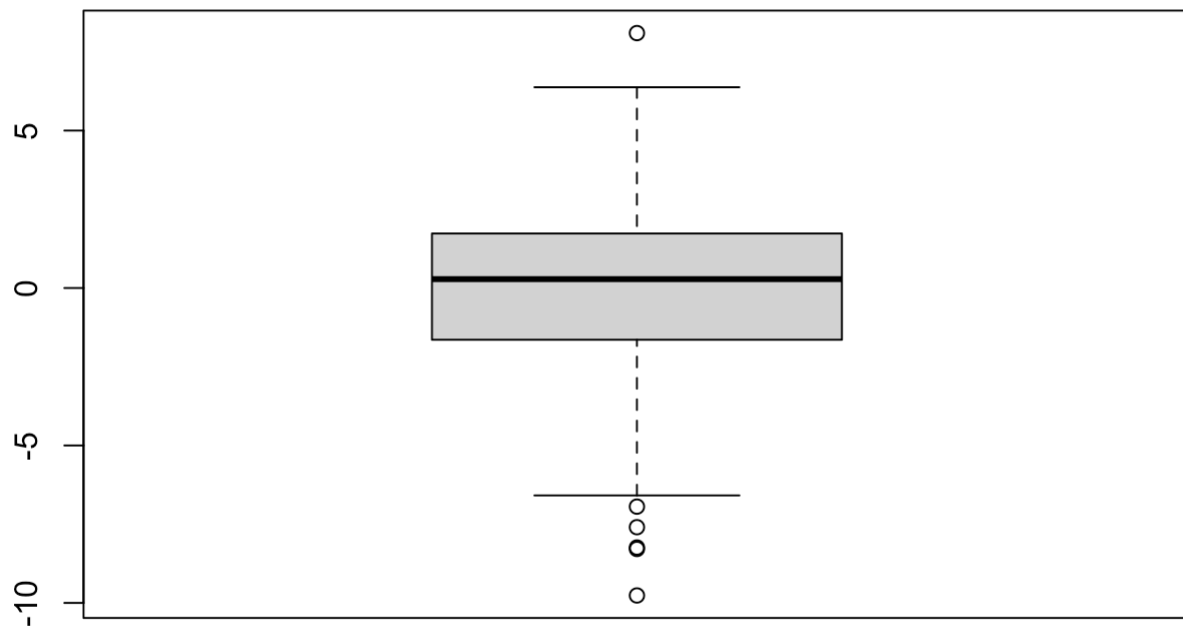
4. Covariate levels

   1. Age - -2.4 t value. This is significant in predicting Right answers
   2. Submissive - 19.542 t value. This is very significant in predicting Right answers.
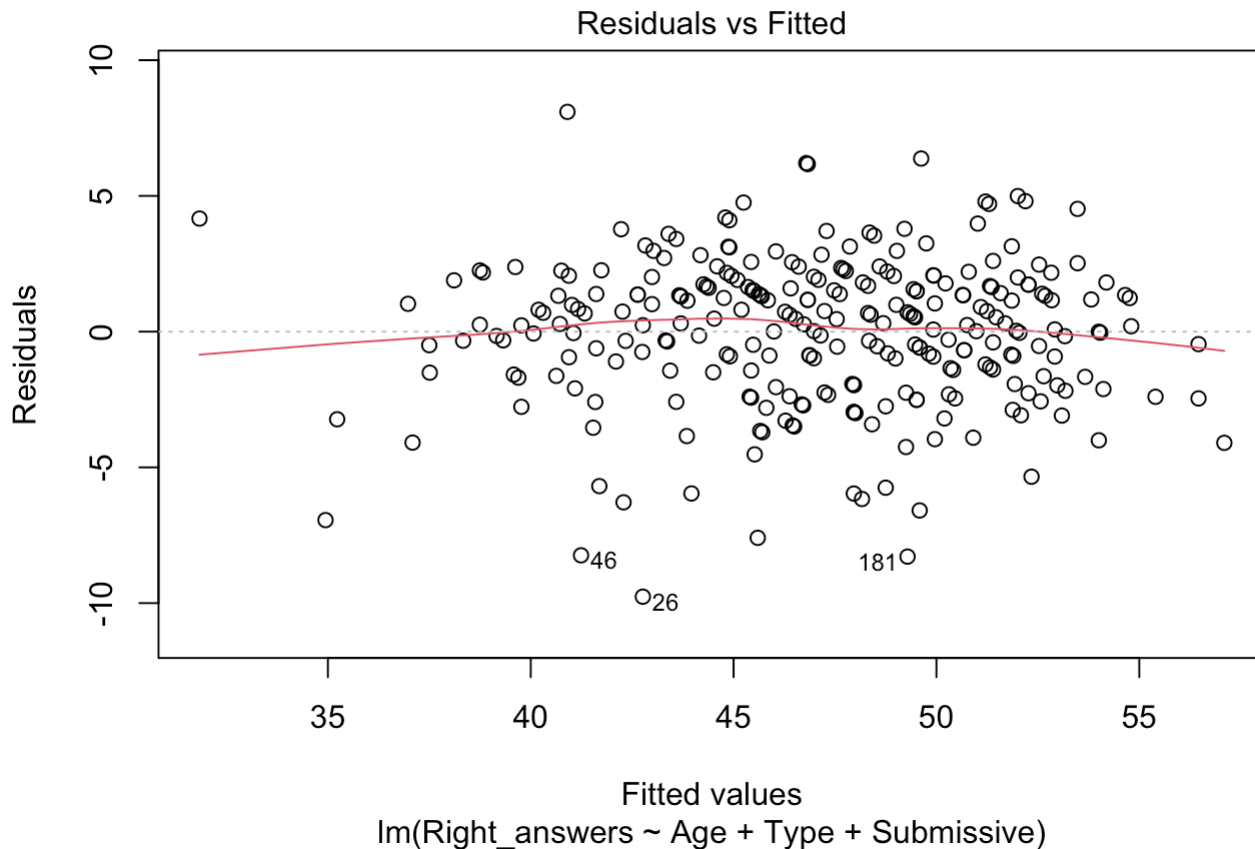
```
par(mfrow = c(2,2))
plot(mod)
```
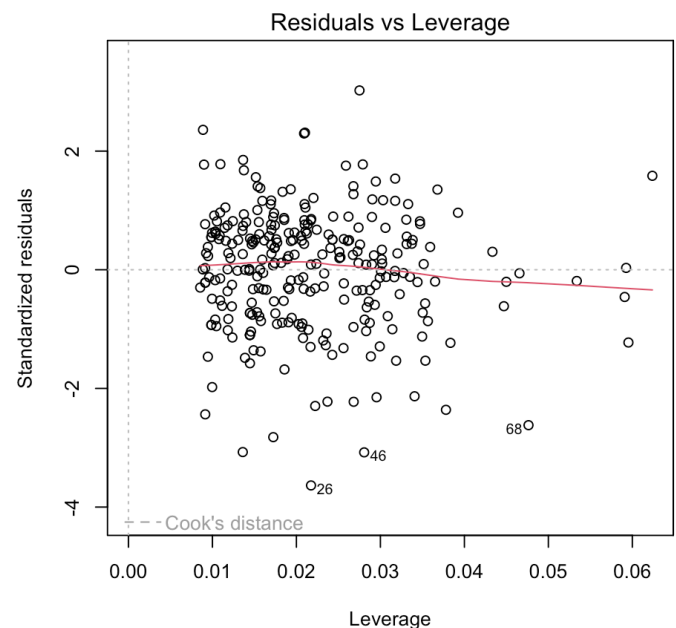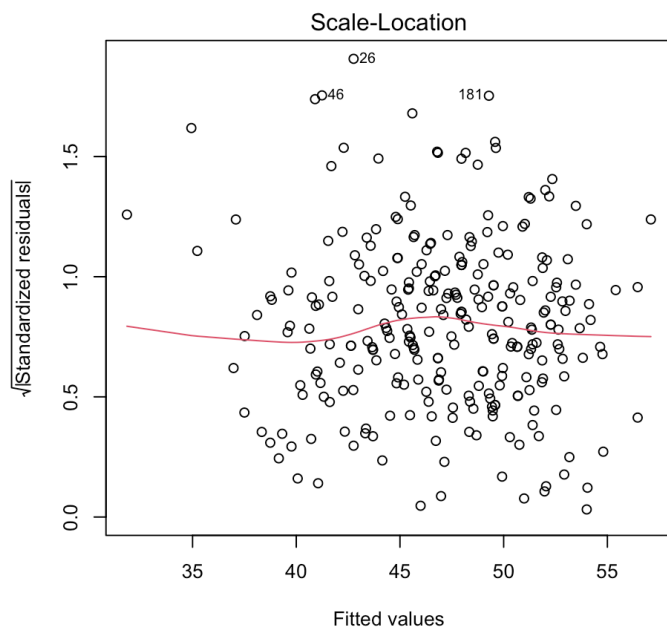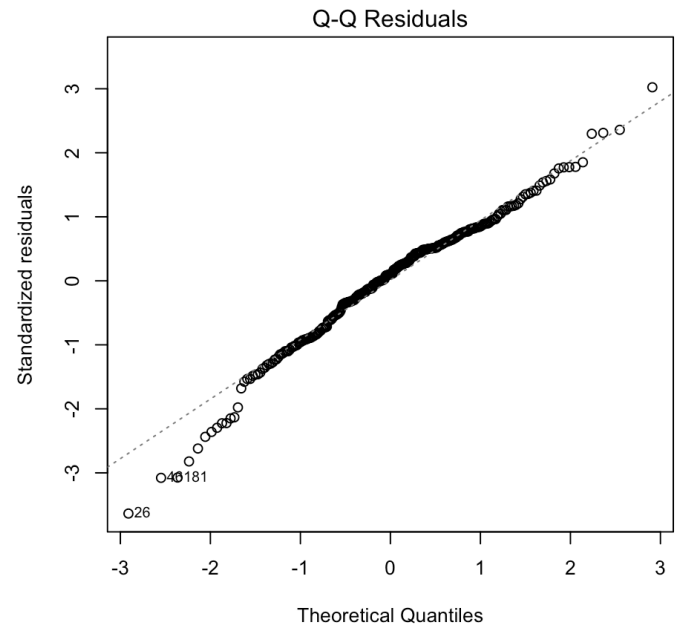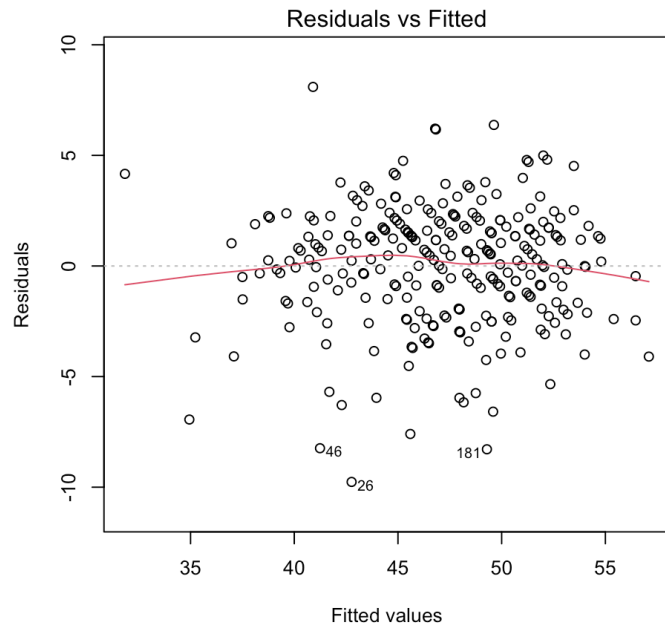


```
boxplot(mod$residuals)
```

Residuals appear to be normally distributed by looking at the boxplot. There is no visual skewness, the IQR is even and tighter then the two tails.

```
plot(mod,1)
```

Residuals vs Fitted

lm(Right_answers ~ Age + Type + Submissive)

The residuals are equally distributed and random.

```
library(car)
```

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

```
durbinWatsonTest(mod)
```

```
##   lag Autocorrelation D-W Statistic p-value
##    1     -0.02121021      2.039786   0.846
##  Alternative hypothesis: rho != 0
```

There is no evidence of autocorrelation so the residuals are independent.

```
library(patchwork)
p1 <- ggplot(miniPONS,aes(Age,Right_answers)) + geom_point() + geom_smooth(method = lm)
p2 <- ggplot(miniPONS,aes(Submissive,Right_answers)) + geom_point() + geom_smooth(method
= lm)

p1 + p2
```
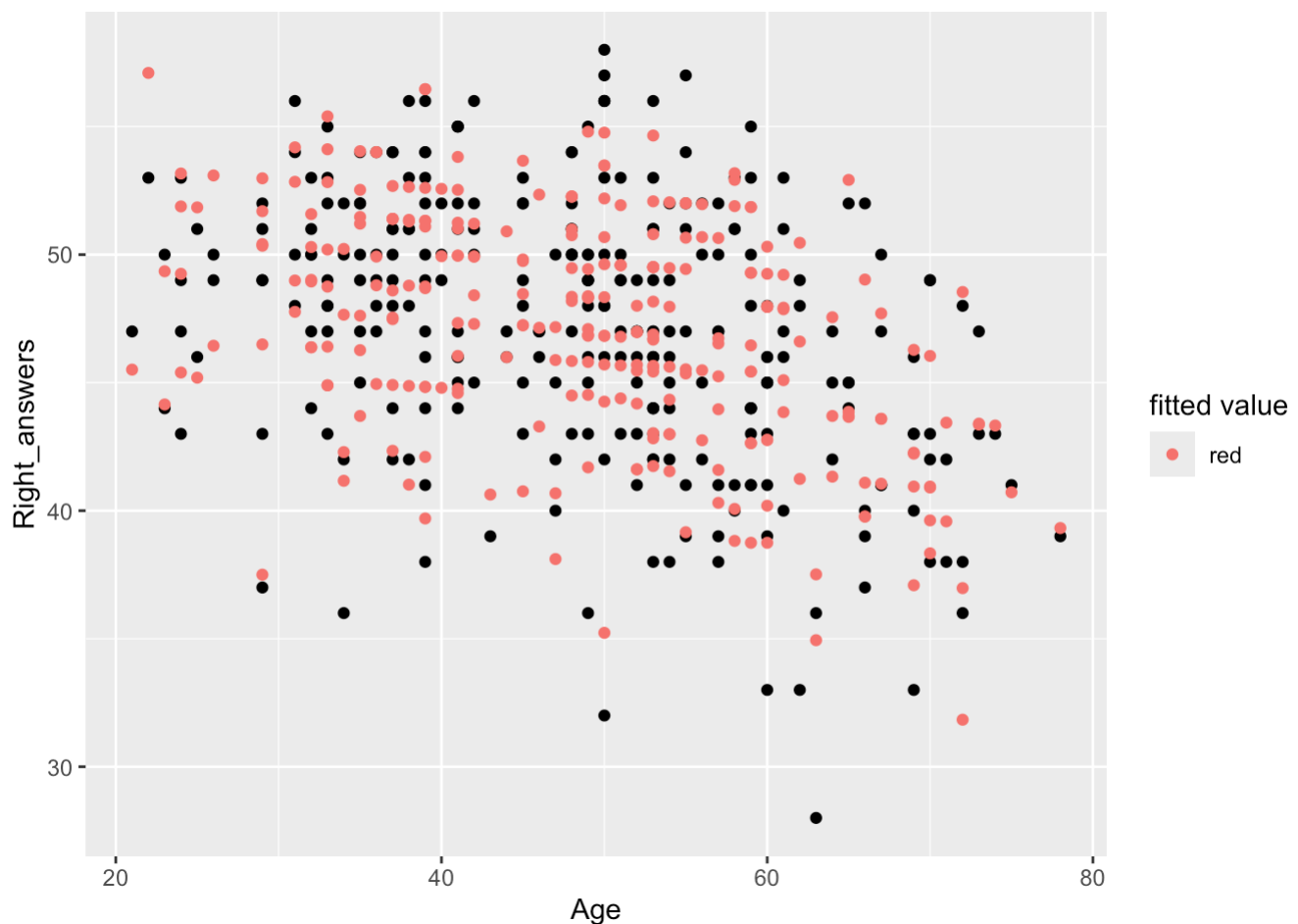
```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```
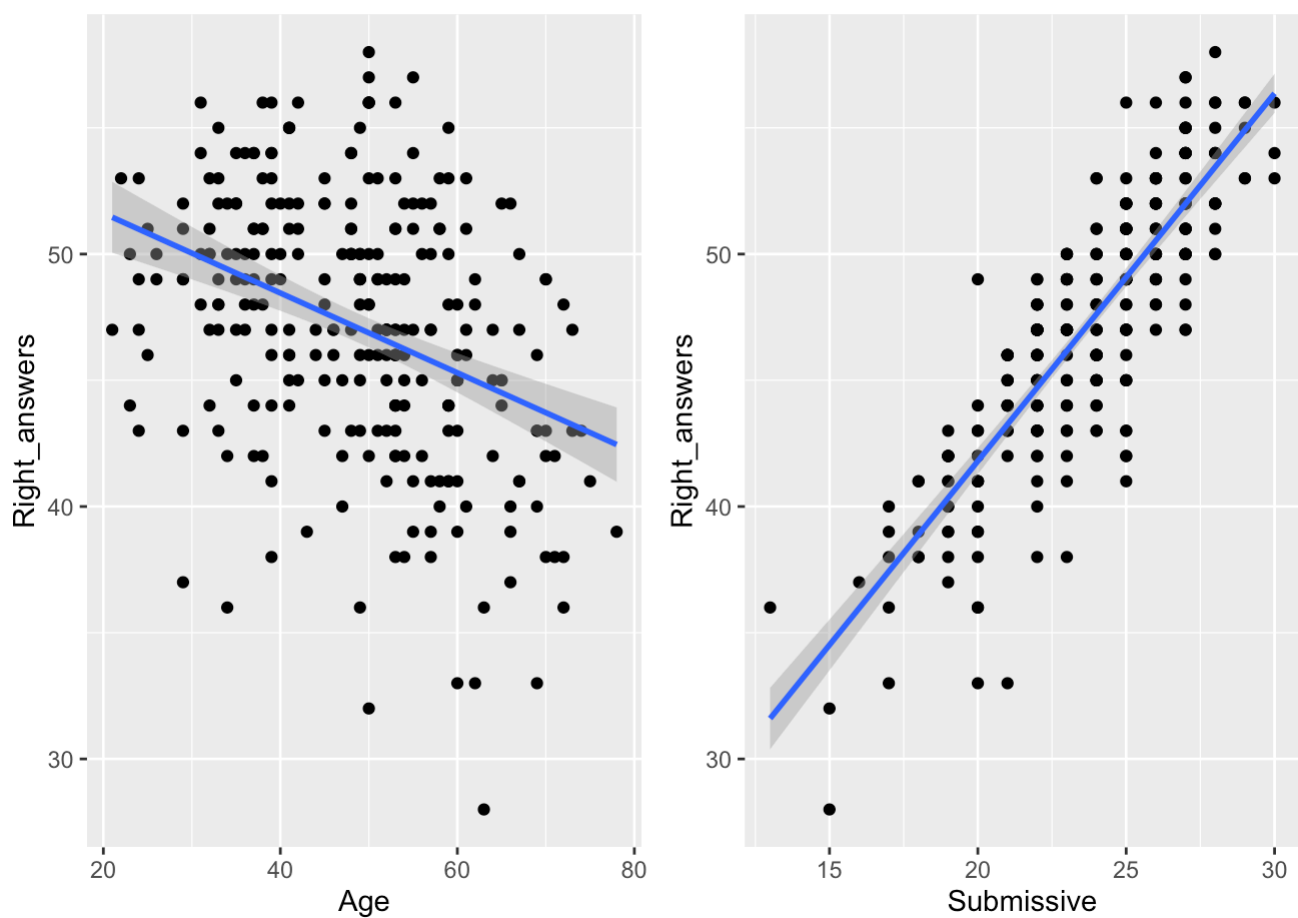

Each of the covariates appear to have roughly linear relationships with the response.

```
(ggplot(miniPONS) + geom_point(aes(Age, Right_answers)) + geom_point(aes(Age, mod$fitte
d.values,colour = 'red')) + labs(color = 'fitted value'))
```

We can see that the predicted values line up fairly well with the original values of the data.

All of the assumptions of the model are met, so we can use this model.

```
sum(mod$residuals<=5 & mod$residuals >= -5)/length(mod$residuals)
```

```
## [1] 0.9386282
```

We can see that the model does a fairly good job at predicting values based on the Age Vs.Right_answers plot when comparing the true value to the predicted. The model does meet all the assumptions of linear models and the model is significant with a very low p-value for both the overall model and the intercept. There is a decent r-squared value most the variation in the response is explained. I would say that the predictions from this model are good, but not excellent. One could use this model to generally estimate the number of right answers with the prediction being within 5 points of the true score 93.86% of the time based on this data.

I would be confident to use this model to make educated estimates based on the factors utilized above.

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)