Assignment 2 - Language Development in ASD - Part 1 - Explaining development

Klara, Pernille, Søren, Clement & Julia

16/09-2020

Assignment 2

In this assignment you will have to discuss a few important questions (given the data you have). More details below. The assignment submitted to the teachers consists of: - a report answering and discussing the questions (so we can assess your conceptual understanding and ability to explain and critically reflect) - a link to a git repository with all the code (so we can assess your code)

Part 1 - Basic description of language development - Describe your sample (n, age, gender, clinical and cognitive features of the two groups) and critically assess whether the groups (ASD and TD) are balanced - Describe linguistic development (in terms of MLU over time) in TD and ASD children (as a function of group). - Describe how parental use of language (in terms of MLU) changes over time. What do you think is going on? - Include individual differences in your model of language development (in children). Identify the best model.

Part 2 - Model comparison - Discuss the differences in performance of your model in training and testing data - Which individual differences should be included in a model that maximizes your ability to explain/predict new data? - Predict a new kid's performance (Bernie) and discuss it against expected performance of the two groups

Part 3 - Simulations to plan a new study - Report and discuss a power analyses identifying how many new kids you would need to replicate the results

The following involves only Part 1.

Learning objectives

- Summarize and report data and models
- Critically apply mixed effects (or multilevel) models
- Explore the issues involved in feature selection

Quick recap

Autism Spectrum Disorder is often related to language impairment. However, this phenomenon has not been empirically traced in detail: i) relying on actual naturalistic language production, ii) over extended periods of time.

We therefore videotaped circa 30 kids with ASD and circa 30 comparison kids (matched by linguistic performance at visit 1) for ca. 30 minutes of naturalistic interactions with a parent. We repeated the data collection 6 times per kid, with 4 months between each visit. We transcribed the data and counted: i) the amount of words that each kid uses in each video. Same for the parent. ii) the amount of unique words that each kid uses in each video. Same for the parent. iii) the amount of morphemes per utterance (Mean Length of Utterance) displayed by each child in each video. Same for the parent.

This data is in the file you prepared in the previous class.

NB. A few children have been excluded from your datasets. We will be using them next week to evaluate how good your models are in assessing the linguistic development in new participants.

This RMarkdown file includes 1) questions (see above). Questions have to be answered/discussed in a separate document that you have to directly submit on Blackboard. 2) A break down of the questions into a guided template full of hints for writing the code to solve the exercises. Fill in the code and the paragraphs as required. Then report your results in the doc for the teachers.

REMEMBER that you will have to have a github repository for the code and submit the answers to Blackboard without code (but a link to your github/gitlab repository). This way we can check your code, but you are also forced to figure out how to report your analyses :-)

Before we get going, here is a reminder of the issues you will have to discuss in your report:

1- Describe your sample (n, age, gender, clinical and cognitive features of the two groups) and critically assess whether the groups (ASD and TD) are balanced 2- Describe linguistic development (in terms of MLU over time) in TD and ASD children (as a function of group). 3- Describe how parental use of language (in terms of MLU) changes over time. What do you think is going on? 4- Include individual differences in your model of language development (in children). Identify the best model.

Let's go

Loading the relevant libraries

Load necessary libraries : what will you need? - e.g. something to deal with the data - e.g. mixed effects models - e.g. something to plot with

```r
library(pacman)
p_load(tidyverse,MuMIn,sjPlot,lme4,reshape2, MuMIn,nlme, broom, ggplot2)
```

Define your working directory and load the data

If you created a project for this class and opened this Rmd file from within that project, your working directory is your project directory.

If you opened this Rmd file outside of a project, you will need some code to find the data: - Create a new variable called locpath (localpath) - Set it to be equal to your working directory - Move to that directory (setwd(locpath)) - Load the data you saved last time (use read_csv(fileName))

```r
alldata <- read_csv("cleanedData.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   Diagnosis = col_character(),
##   Ethnicity = col_character(),
##   Gender = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

Characterize the participants (Exercise 1)

Identify relevant variables: participants demographic characteristics, diagnosis, ADOS, Verbal IQ (expressive lang raw), Non Verbal (mullenraw) IQ, Socialization, Visit, Number of words used, Number of unique words used, mean length of utterance in both child and parents.

Make sure the variables are in the right format.

Describe the characteristics of the two groups of participants and whether the two groups are well matched.

```
alldata %>% group_by(Diagnosis) %>% filter(VISIT == 1) %>% summarize(mean(Age, na.rm = T))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   Diagnosis `mean(Age, na.rm = T)`
##   <chr>                      <dbl>
## 1 ASD                         33.0
## 2 TD                          20.4
```

```
alldata %>% filter(VISIT == 1) %>% count(Diagnosis)
```

```
## # A tibble: 2 x 2
##   Diagnosis     n
##   <chr>     <int>
## 1 ASD          31
## 2 TD           35
```

```
alldata %>% count(Diagnosis)
```

```
## # A tibble: 2 x 2
##   Diagnosis     n
##   <chr>     <int>
## 1 ASD         176
## 2 TD          196
```

```
176/31
```

```
## [1] 5.677419
```

```
196/35
```

```
## [1] 5.6
```

```
alldata %>% filter(VISIT == 1) %>% count(Gender)
```

```
## # A tibble: 2 x 2
##   Gender     n
##   <chr>  <int>
## 1 F         11
## 2 M         55
```

```r
alldata %>% filter(VISIT == 1) %>% count(Ethnicity)
```

```
## # A tibble: 8 x 2
##   Ethnicity            n
##   <chr>            <int>
## 1 African American     2
## 2 Asian                1
## 3 Bangladeshi          1
## 4 Latino               1
## 5 Lebanese             1
## 6 White               57
## 7 White/Asian          1
## 8 White/Latino         2
```

```r
alldata %>% group_by(Diagnosis) %>% filter(VISIT == 1) %>% count(Ethnicity)
```

```
## # A tibble: 9 x 3
## # Groups:   Diagnosis [2]
##   Diagnosis Ethnicity            n
##   <chr>     <chr>            <int>
## 1 ASD       African American     2
## 2 ASD       Bangladeshi          1
## 3 ASD       Latino               1
## 4 ASD       Lebanese             1
## 5 ASD       White               23
## 6 ASD       White/Asian          1
## 7 ASD       White/Latino         2
## 8 TD        Asian                1
## 9 TD        White               34
```

Mean age: ASD = 33.03903 months and TD = 20.38294 months

The study included 31 children with ASD and 35 children with TD - both groups showing up to approximately 5.6 meetings.

There are 11 girls and 55 boys in the study.

The ethnicity of the children are not balanced…

The sample included mostly young (<20) white males …

[REPORT THE RESULTS]

Let's test hypothesis 1: Children with ASD display a language impairment (Exercise 2)

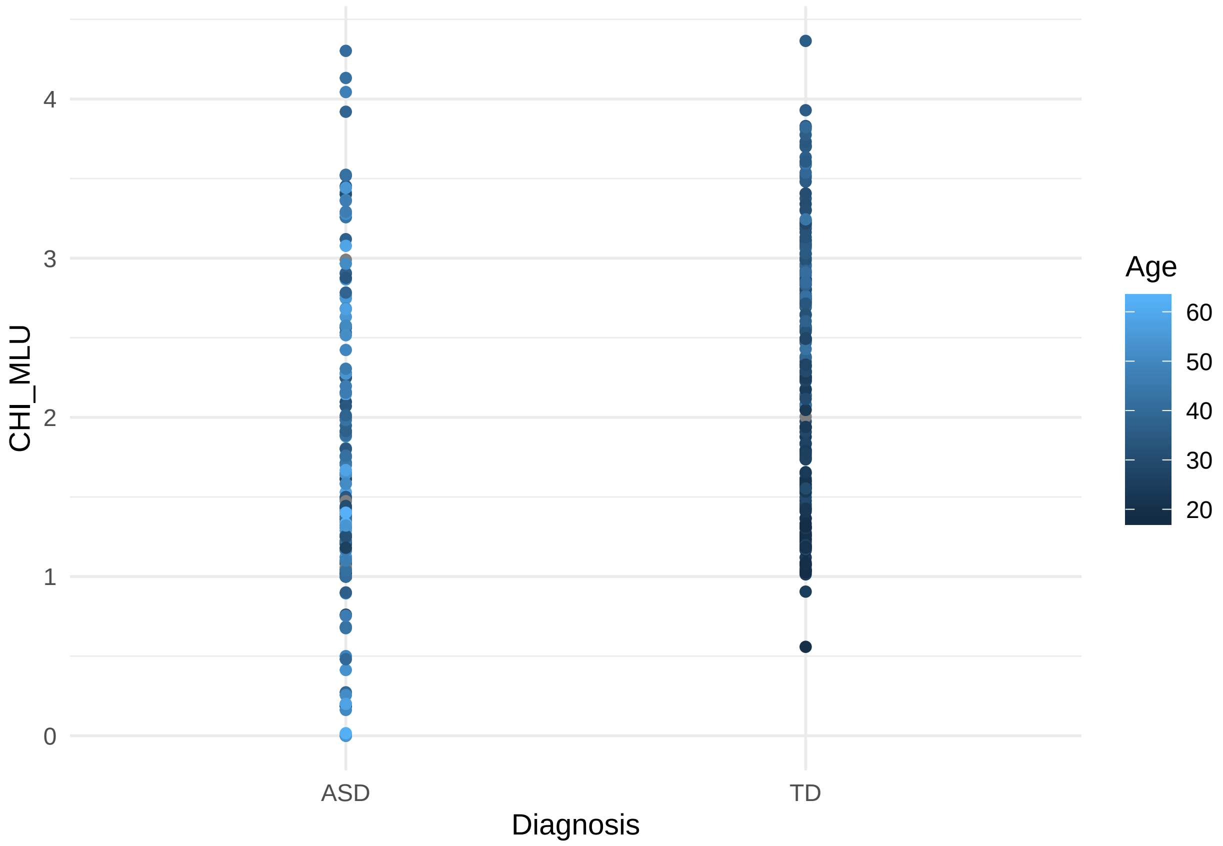
Hypothesis: The child's MLU changes: i) over time, ii) according to diagnosis

Let's start with a simple mixed effects linear model

Remember to plot the data first and then to run a statistical test. - Which variable(s) should be included as fixed factors? - Which variable(s) should be included as random factors?

```
# Plotting the data to see what is going on with the different variables

# Child MLU by diagnosis
ggplot(alldata, aes(x = Diagnosis, y = CHI_MLU, color = Age)) +
  geom_point() +
  theme_minimal()
```

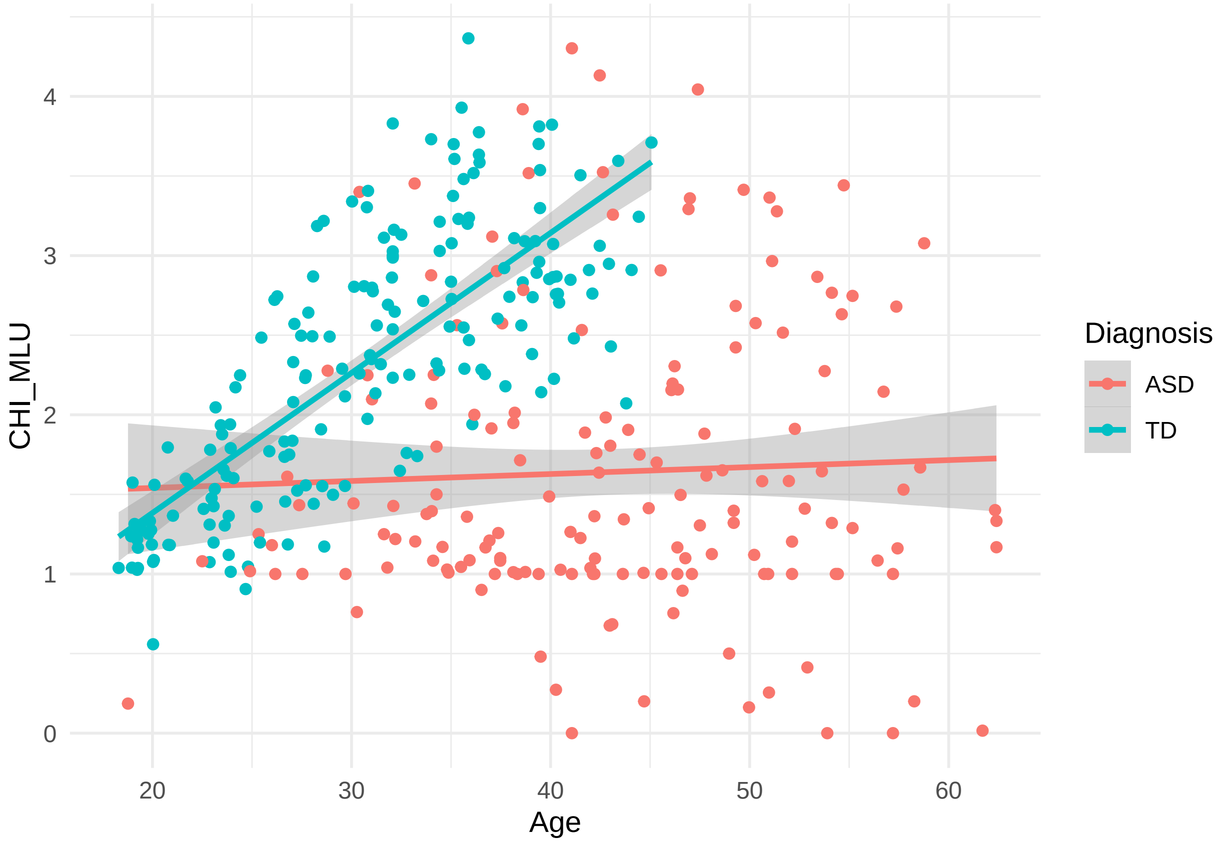
```
## Warning: Removed 20 rows containing missing values (geom_point).
```



```
# Child MLU by age
ggplot(alldata, aes(x = Age, y = CHI_MLU, color = Diagnosis)) +
  geom_smooth(method = lm) +
  geom_point() +
  theme_minimal()
```

```
## Warning: Removed 26 rows containing non-finite values (stat_smooth).
```

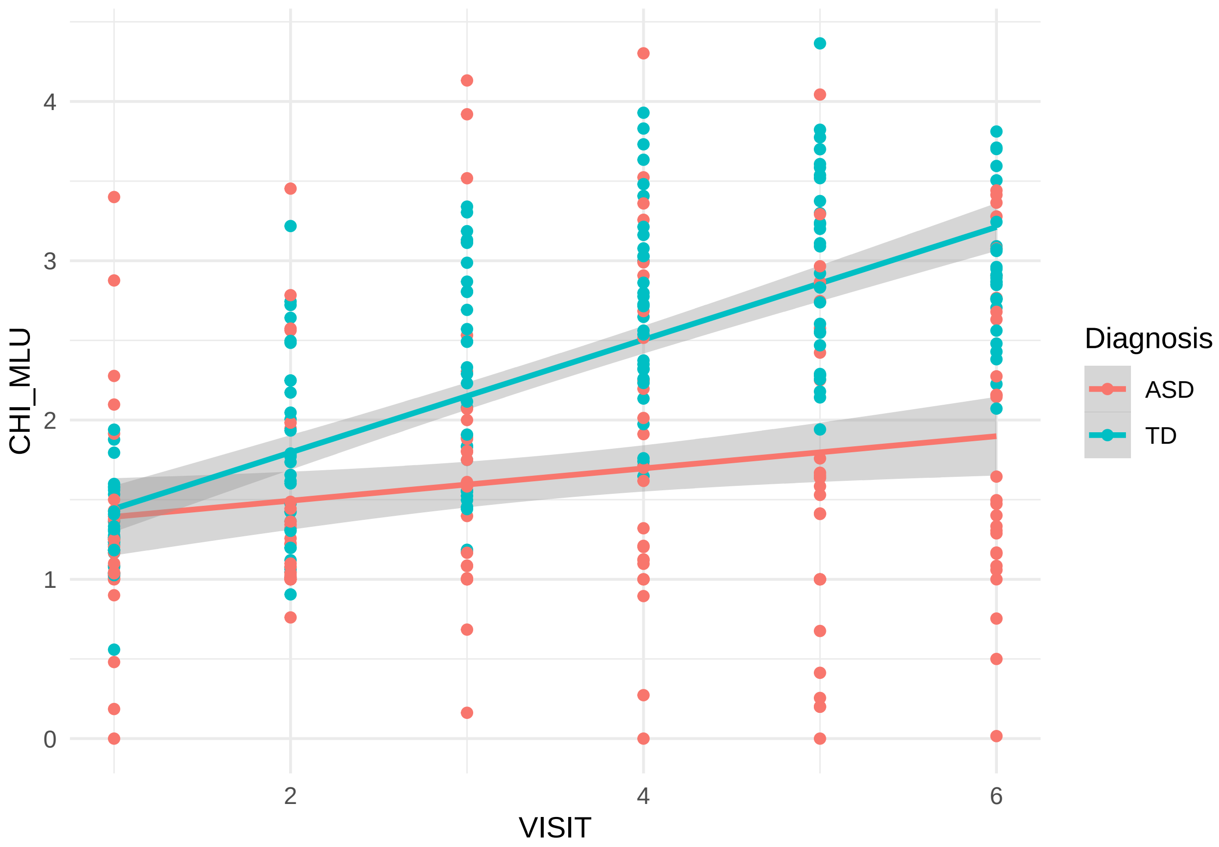
```
## Warning: Removed 26 rows containing missing values (geom_point).
```



```
# Child MLU by visit
ggplot(alldata, aes(x = VISIT, y = CHI_MLU, color = Diagnosis)) +
  geom_smooth(method = lm) +
  geom_point() +
  theme_minimal()
```

```
## Warning: Removed 20 rows containing non-finite values (stat_smooth).
```

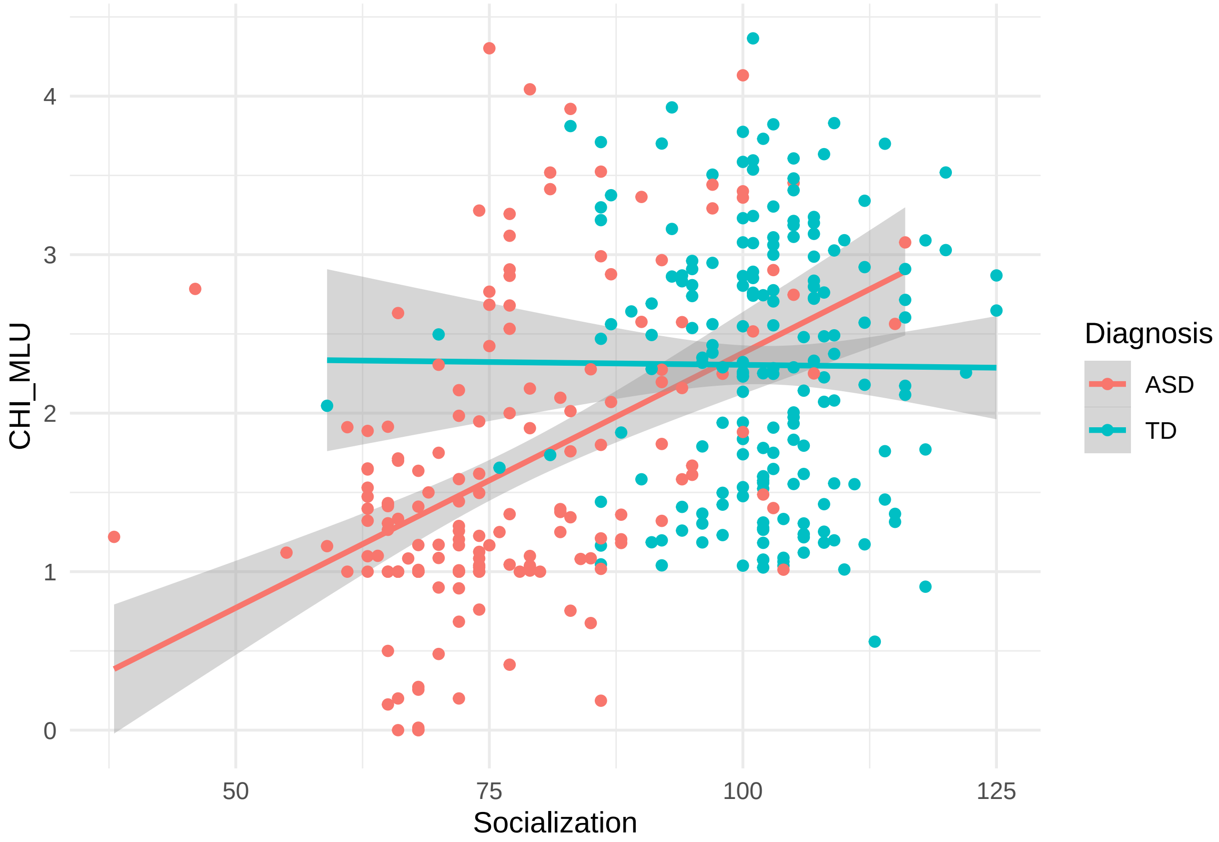
```
## Warning: Removed 20 rows containing missing values (geom_point).
```



```r
# Child MLU by socialization
ggplot(alldata, aes(x = Socialization, y = CHI_MLU, color = Diagnosis)) +
  geom_smooth(method = lm) +
  geom_point() +
  theme_minimal()
```

```
## Warning: Removed 22 rows containing non-finite values (stat_smooth).
```

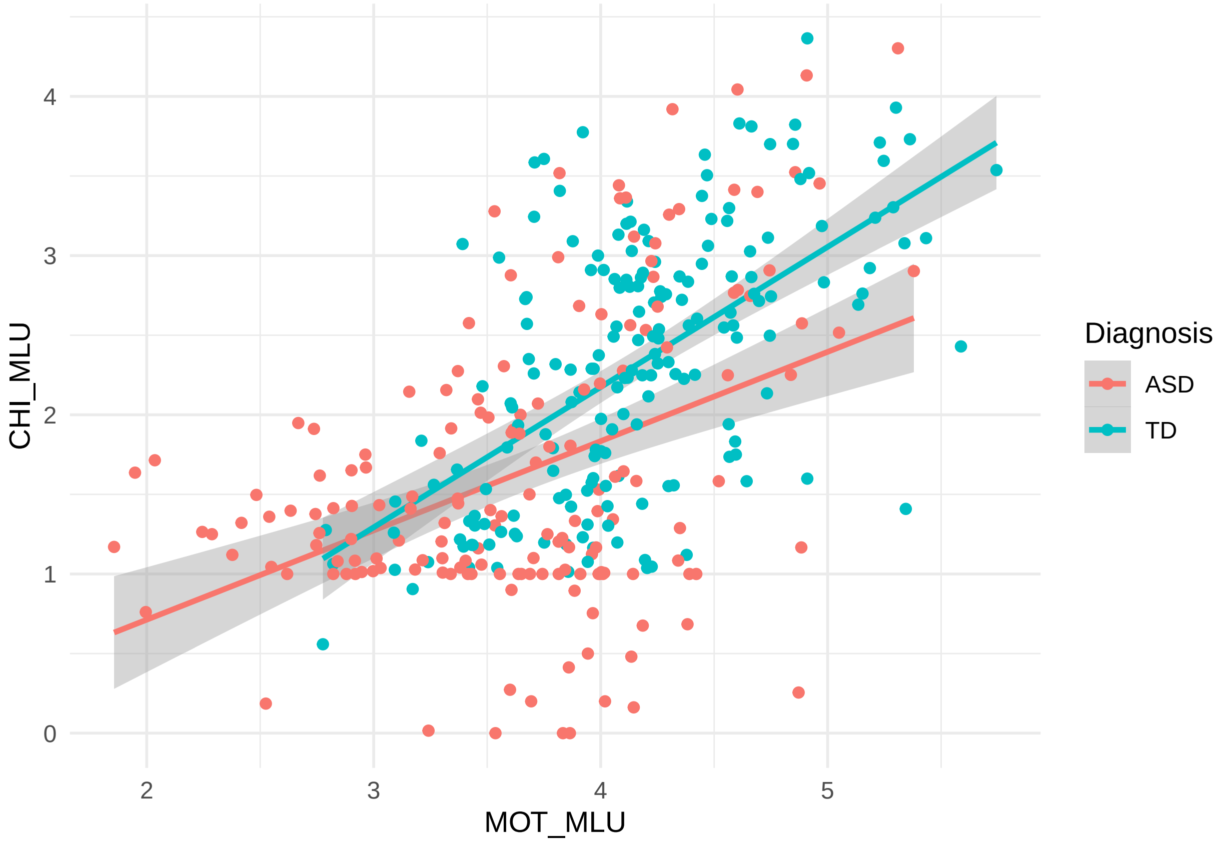
```
## Warning: Removed 22 rows containing missing values (geom_point).
```



```
# Child MLU by mother MLU
ggplot(alldata, aes(x = MOT_MLU, y = CHI_MLU, color = Diagnosis)) +
  geom_smooth(method = lm) +
  geom_point() +
  theme_minimal()
```

```
## Warning: Removed 20 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```



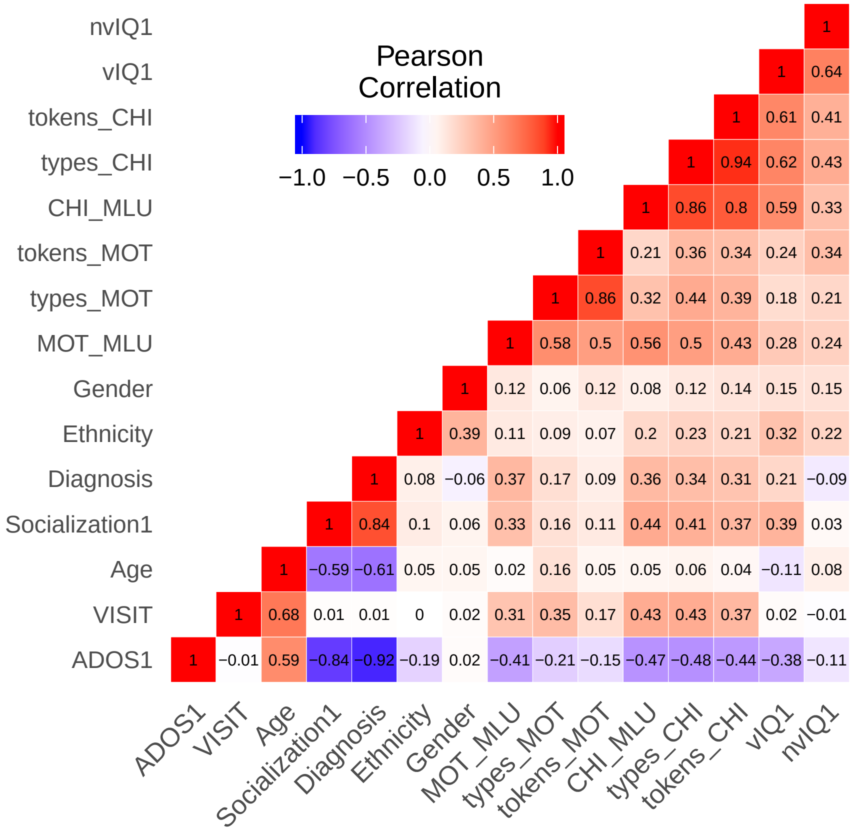
In order to avoid collinearity, we constructed a heatmap of the correlations between the variables.

```r
# Defining the functions needed for the heatmap
# Use correlation between variables as distance
reorder_cormat <- function(cormat) {
  dd <- as.dist((1 - cormat) / 2)
  hc <- hclust(dd)
  cormat <- cormat[hc$order, hc$order]
  }

# Get upper triangle of the correlation matrix
get_upper_tri <- function(cormat) {
  cormat[lower.tri(cormat)] <- NA
  return(cormat)
  }
```

```r
# Preparing a dataframe for heatmap without N/A values
sub_df <- alldata %>% select(-c(ID, ADOS, Socialization, nvIQ, vIQ))

sub_df <- na.omit(sub_df)

sub_df$Diagnosis <- as.factor(sub_df$Diagnosis)
sub_df$Diagnosis <- as.numeric(sub_df$Diagnosis)

sub_df$Ethnicity <- as.factor(sub_df$Ethnicity)
sub_df$Ethnicity <- as.numeric(sub_df$Ethnicity)
```

```r
sub_df$Gender <- as.factor(sub_df$Gender)
sub_df$Gender <- as.numeric(sub_df$Gender)
```

```r
#Building heatmap of correlations
cormat <- round(cor(sub_df),2)
cormat <- reorder_cormat(cormat)
upper_tri <- get_upper_tri(cormat)

melted_cormat <- melt(upper_tri, na.rm = TRUE)

ggplot(melted_cormat, aes(Var2, Var1, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(
    low = "blue",
    high = "red",
    mid = "white",
    midpoint = 0,
    limit = c(-1, 1),
    space = "Lab",
    name = "Pearson\nCorrelation"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(
    angle = 45,
    vjust = 1,
    size = 10,
    hjust = 1
  )) +
  coord_fixed() + geom_text(aes(Var2, Var1, label = value),
                            color = "black",
                            size = 2) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal"
  ) +
  guides(fill = guide_colorbar(
    barwidth = 7,
    barheight = 1,
    title.position = "top",
    title.hjust = 0.5
  ))
```



The correlations in this heatmap are a little different from the true correlations due to NA's omitted.

```r
cor.test(alldata$vIQ1, alldata$nvIQ1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  alldata$vIQ1 and alldata$nvIQ1
## t = 15.499, df = 369, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5620690 0.6859112
## sample estimates:
##       cor
## 0.6279491
```

How would you evaluate whether the model is a good model?

The model with the interaction, m, is significantly better than the simpler model and it has a higher marginal and conditional r^2 value of 62.0% and 81.1%, which is a pretty high value for a model in social science.

Let's check whether a growth curve model is better. Remember: a growth curve model assesses whether changes in time can be described by linear, or quadratic, or cubic (or… etc.) components. First build the different models, then compare them to see which one is better.

Exciting right? Let's check whether the model is doing an alright job at fitting the data. Plot the actual CHI_MLU data against the predictions of the model fitted(model).

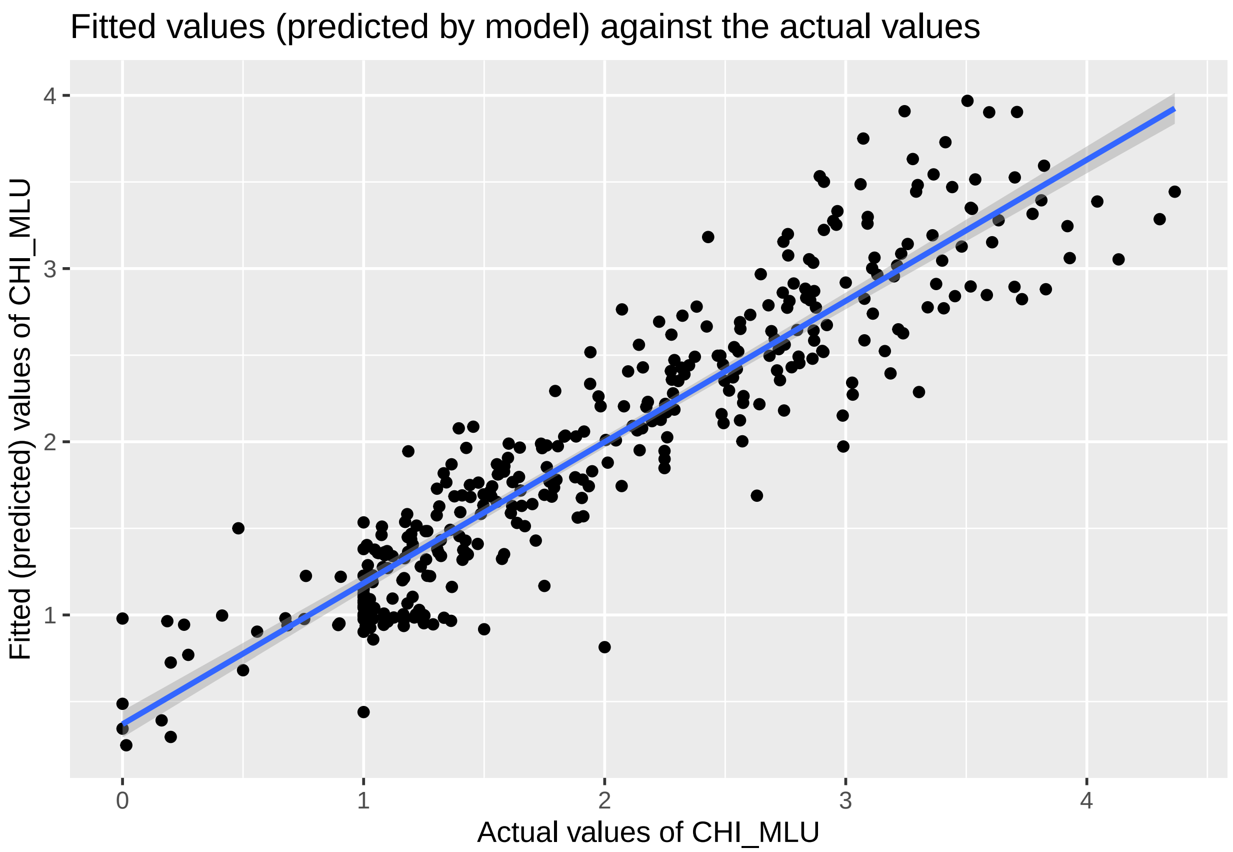
```
# Plot: Fitted values (predicted by model) against actual values

# Making df with selected variables
sub_df <- alldata %>% select(c(ID, CHI_MLU, VISIT, vIQ1, Diagnosis))

# Omit na's
sub_df <- na.omit(sub_df)

# Predicting the values using the
sub_df$pred_CHI_MLU <- predict(m) #same values

# Finding the difference in prediced and observed values
sub_df$diff <- sub_df$pred_CHI_MLU - sub_df$CHI_MLU

# Make model for fun - don't know if this is allowed? lol
model <- lm(CHI_MLU ~ pred_CHI_MLU, sub_df, REML = F)
```

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'REML' will be disregarded
```

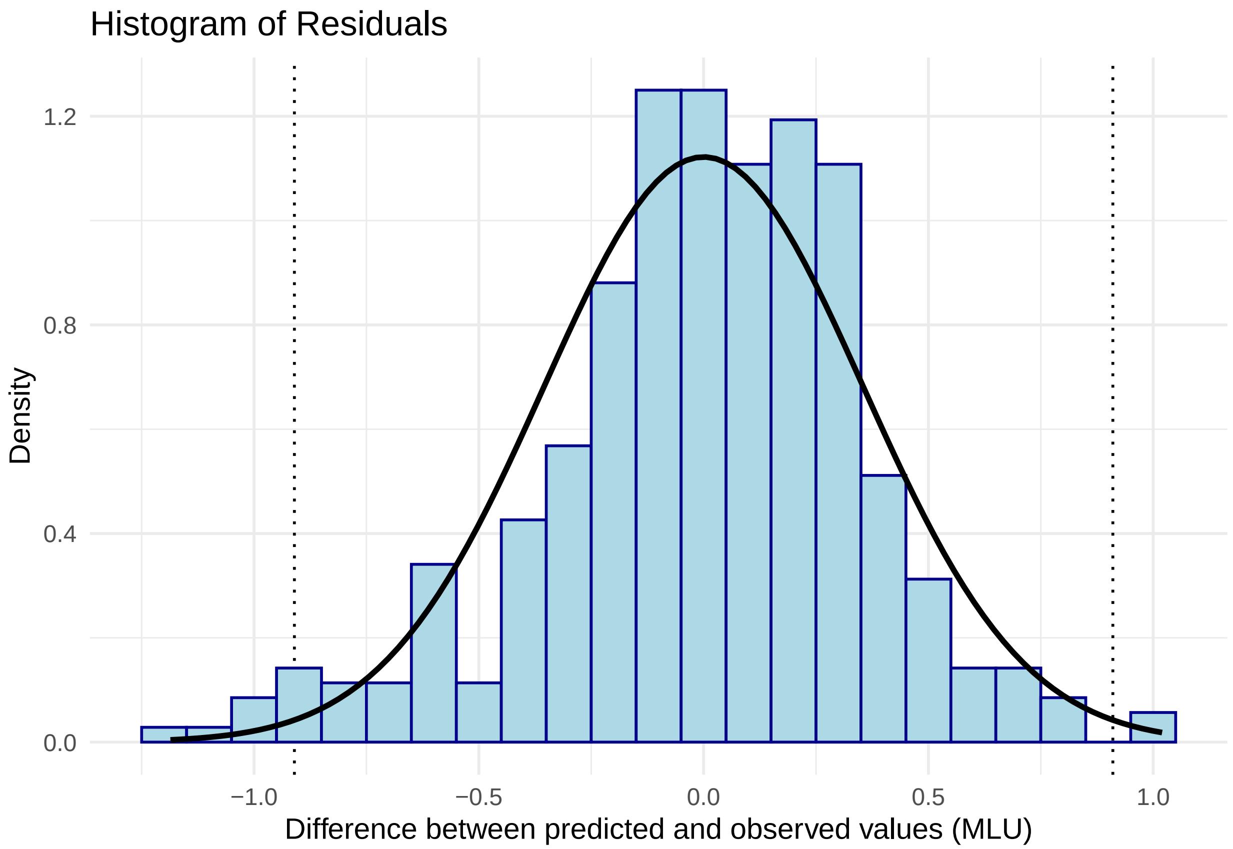
```
summary(model)
```

```
##
## Call:
## lm(formula = CHI_MLU ~ pred_CHI_MLU, data = sub_df, REML = F)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99446 -0.23378 -0.00612  0.19153  1.24665
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.10249    0.04957  -2.068   0.0394 *
## pred_CHI_MLU  1.05142    0.02300  45.707   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3536 on 350 degrees of freedom
## Multiple R-squared:  0.8565, Adjusted R-squared:  0.8561
## F-statistic:  2089 on 1 and 350 DF,  p-value: < 2.2e-16
```

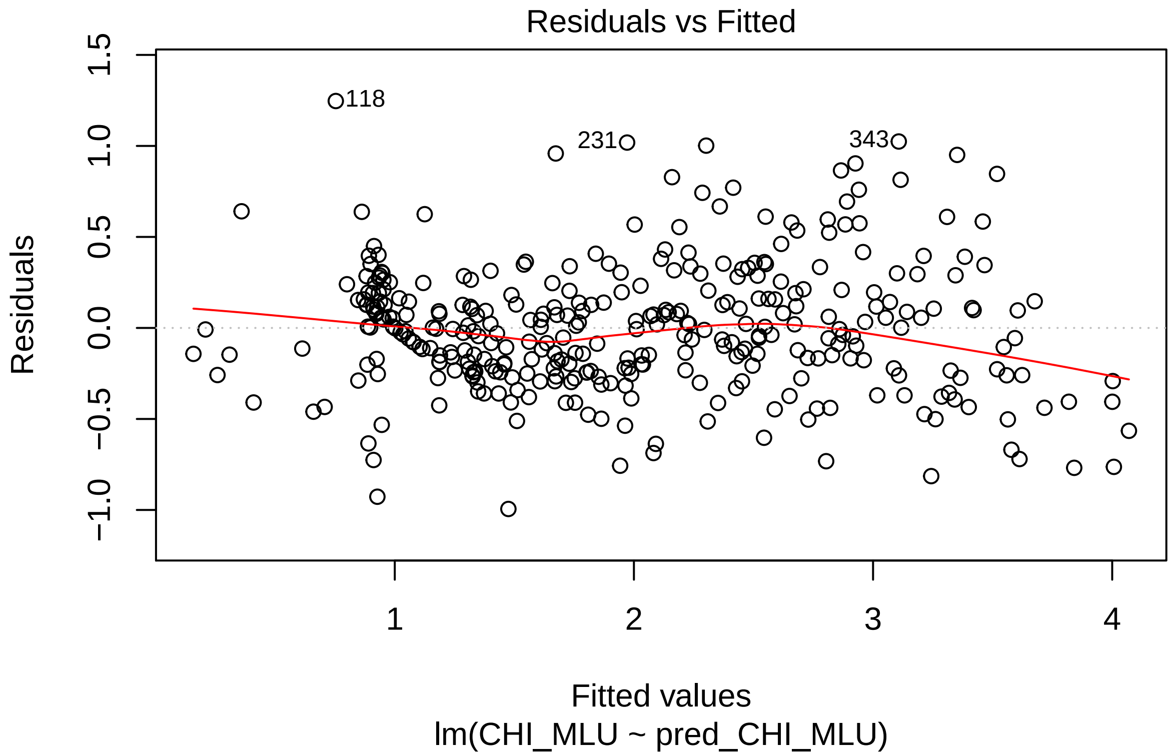
```
# Making a cool plot
ggplot(sub_df, aes(x = CHI_MLU, y = pred_CHI_MLU)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ggtitle("Fitted values (predicted by model) against the actual values") +
  ylab("Fitted (predicted) values of CHI_MLU") +
  xlab("Actual values of CHI_MLU")
```

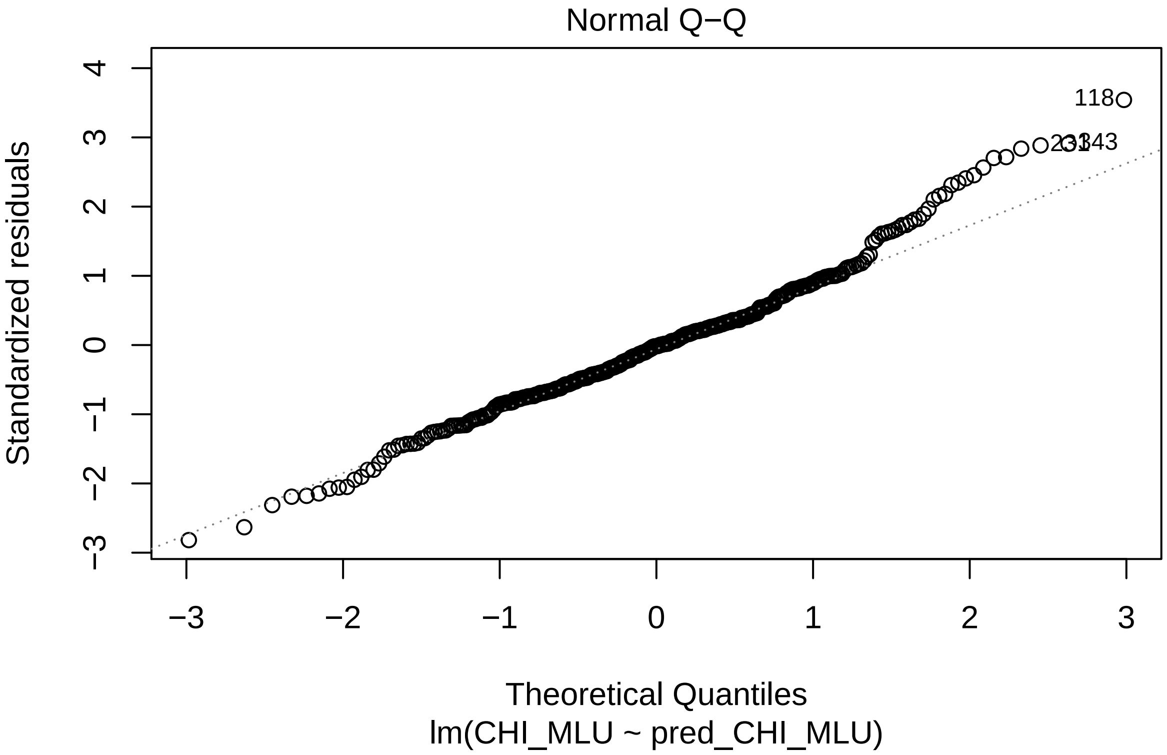


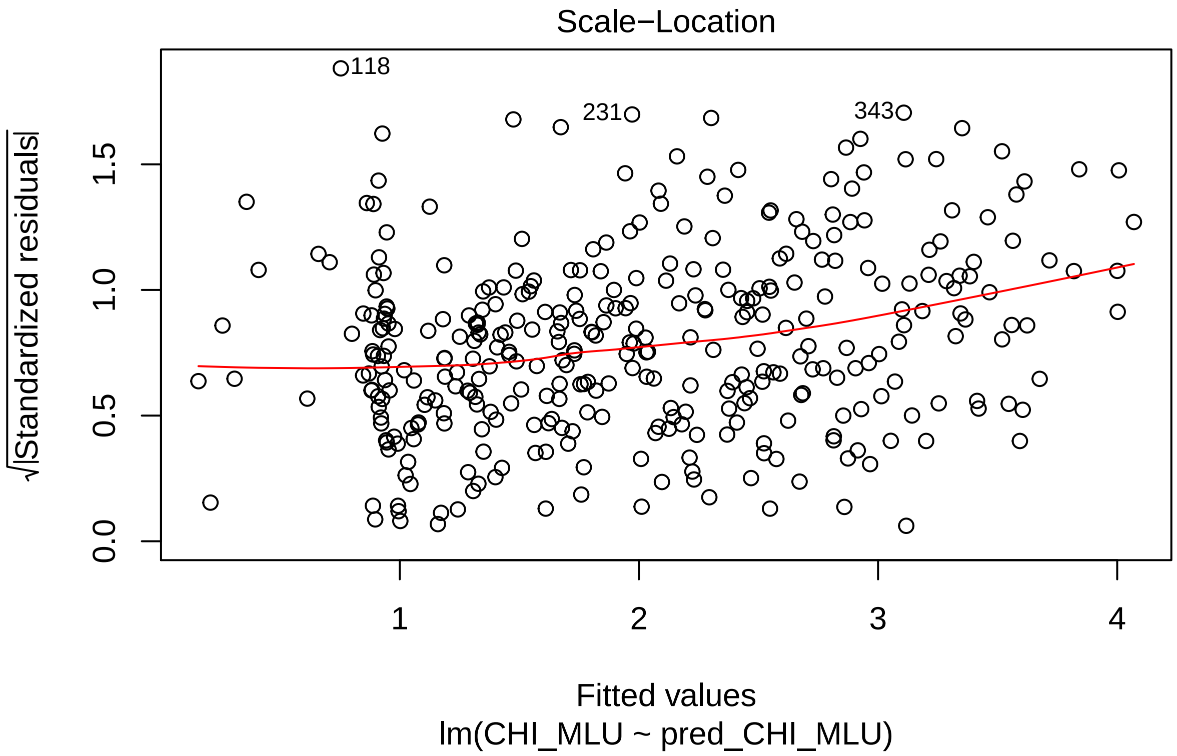
```
# Making a histogram plot over the residuals between predicted and actual
ggplot(sub_df, aes(x = diff)) +
  geom_vline(
    data = sub_df,
    aes(xintercept = mean(diff) + 2.56 * sd(diff)),
    colour = "black",
    linetype = "dotted"
  ) +
  geom_vline(
    data = sub_df,
    aes(xintercept = mean(diff) - 2.56 * sd(diff)),
    colour = "black",
    linetype = "dotted"
  ) +
  geom_histogram(
    aes(y = ..density..),
    color = "darkblue",
    fill = "lightblue",
    binwidth = .1
  ) +
  stat_function(
    fun = dnorm,
    args = list(
      mean = mean(sub_df$diff, na.rm = TRUE),
      sd = sd(sub_df$diff, na.rm = TRUE)
    ),
    colour = "black",
    size = 1
  ) +
  labs(title = "Histogram of Residuals",
       y = "Density",
       x = "Difference between predicted and observed values (MLU)") +
  theme_minimal()
```

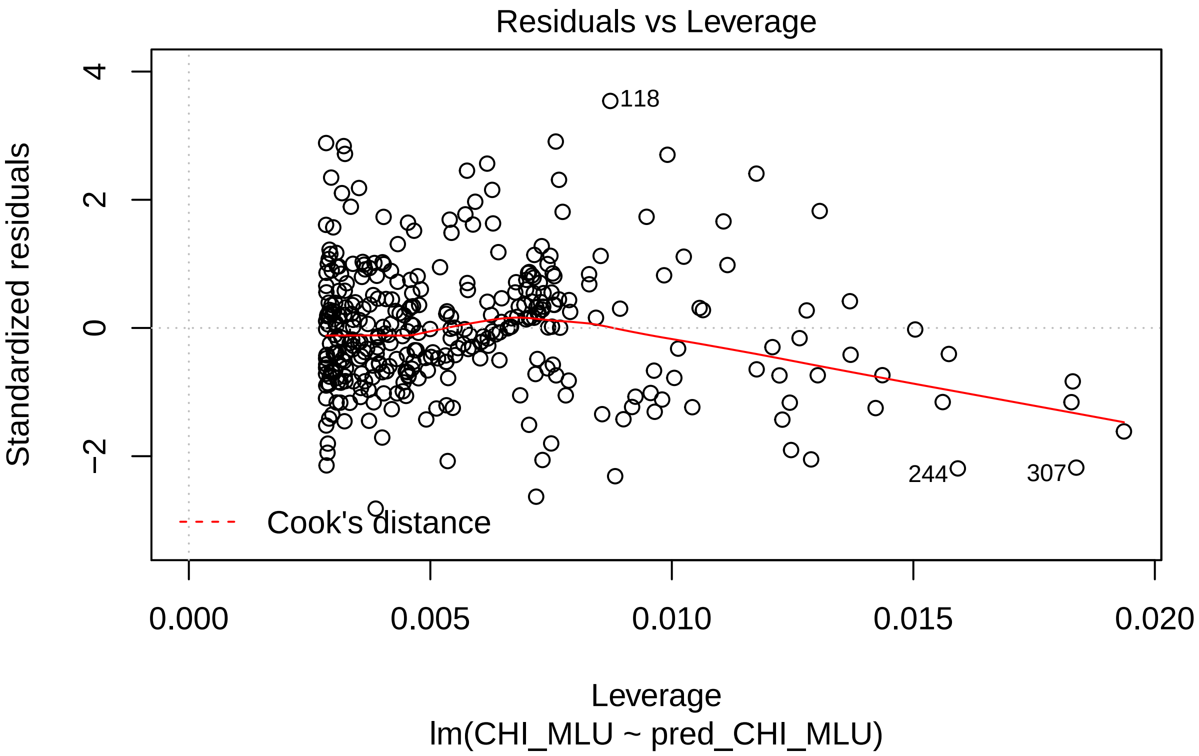


```
result <- lm(CHI_MLU ~ pred_CHI_MLU, sub_df)

# Visual inspection of the assumptions
plot(result)
```


Residuals vs Fitted
Residuals
Fitted values
lm(CHI_MLU ~ pred_CHI_MLU)
1.5
1.0
0.5
0.0
−0.5
−1.0
1
2
3
4
118
231
343


Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(CHI_MLU ~ pred_CHI_MLU)
118
231
143


Scale−Location
118
231
343
√|Standardized residuals|
0.0
0.5
1.0
1.5
1
2
3
4
Fitted values
lm(CHI_MLU ~ pred_CHI_MLU)



Now it's time to report our results. Remember to report: - the estimates for each predictor (beta estimate, standard error, p-value) - A plain word description of the results - A plot of your model's predictions (and some comments on whether the predictions are sensible)

[REPORT THE RESULTS] Linguistic development of children MLU is affected by … [COMPLETE]

Let's test hypothesis 2: Parents speak equally to children with ASD and TD (Exercise 3)

Hypothesis: Parental MLU changes: i) over time, ii) according to diagnosis

Parent MLU is affected by … but probably not … [REPORT THE RESULTS]

Adding new variables (Exercise 4)

Your task now is to figure out how to best describe the children linguistic trajectory. The dataset contains a bunch of additional demographic, cognitive and clinical variables (e.g.verbal and non-verbal IQ). Try them out and identify the statistical models that best describes your data (that is, the children's MLU). Describe how you selected the best model and send the code to run the model to Victor and Byurakn.

In addition to …, the MLU of the children is also correlated with … Using AIC / nested F-tests as a criterium, we compared models of increasing complexity and found that …

[REPORT THE RESULTS]