# Practical Mixed Effect Models

Ricard Garcia & Núria Cardona

2023-11-15

**a) Read the dataset into the R environment:**

```
library(readxl)

setwd("/home/nuria/Documents/Bioinformatics/Statistical_models_stochastic_processes/Assignment5_MixedMo

data <- read_excel("sleepstudy_fixed.xlsx")
```

**b) Does this data have a hierarchical structure? Do you have any a priori reasons to expect that Reaction measurements could not be independent? Do you have a longitudinal data case?**

To determine if our data has hierarchical structure we have to focus if it has a nested or structured organization. We can state that our data has hierarchical structure because of the repeated measurements taken, in each individual, over the time.

To consider if reaction measurements could not be independent, we have to focus in our data structure. Since the data is taken in each subject over the time, we can state that the data is correlated, dependent. We can assume it because we are talking about the same individual as the time goes by.

In our case, we have longitudinal data because we have multiple rows representing different time points, and we are specifically interested in observing changes over time. This is akin to saying that our data involves repeated measurements of the same subject over a specific period.

**c) Format the data for its use by the functions of the nlme package with the instruction: X <- groupedData(Reaction~Days|Subject,data=X)**

```
library(nlme)

formatted_data <- groupedData(Reaction ~ Days | Subject, data = data)
```

**d) What is the total number of subjects in the database? Is the data balanced? How many measurements were taken on each Subject at most?**

```
total_number_subjects <- length(unique(formatted_data$Subject))
cat("Total number of subjects:", total_number_subjects)
```

```
## Total number of subjects: 18
```

```
#Data is balanced if all subjects have the same number of measurements.
measurements_each_subject <- table(formatted_data$Subject)
print(measurements_each_subject)
```

```
##
## 309 310 335 351 349 330 333 369 372 371 331 370 334 352 350 332 337 308
##  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10  10
```

We can observe that all the subjects have the same number of measurements: 10. Therefore, the data is balanced. We can manually observe, since the number of individuals is low, that the maximum number of

measurements is 10.

If we want to compute it with R we can do it as follows:

```
max_measurements_individuals <- max(measurements_each_subject)
cat("Maximum number of measurements on a single subject:", max_measurements_individuals)
```

```
## Maximum number of measurements on a single subject: 10
```

**e) Fit an ordinary linear regression model, with Days as the predictor and Reaction as the response. Is there a significant relationship between these two variables? How much variance of Reaction can be explained by Days?**

```
#First we set the values as numeric.
formatted_data$Reaction <- as.numeric(as.character(formatted_data$Reaction))

#Then, we can apply the command:
OLR_model <- lm(Reaction ~ Days, data = formatted_data)

summary(OLR_model)
```

```
##
## Call:
## lm(formula = Reaction ~ Days, data = formatted_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.849  -27.484    1.547   26.141  139.956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  251.401      6.610  38.032  < 2e-16 ***
## Days          10.467      1.238   8.454  9.9e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.71 on 178 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2825
## F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.896e-15
```

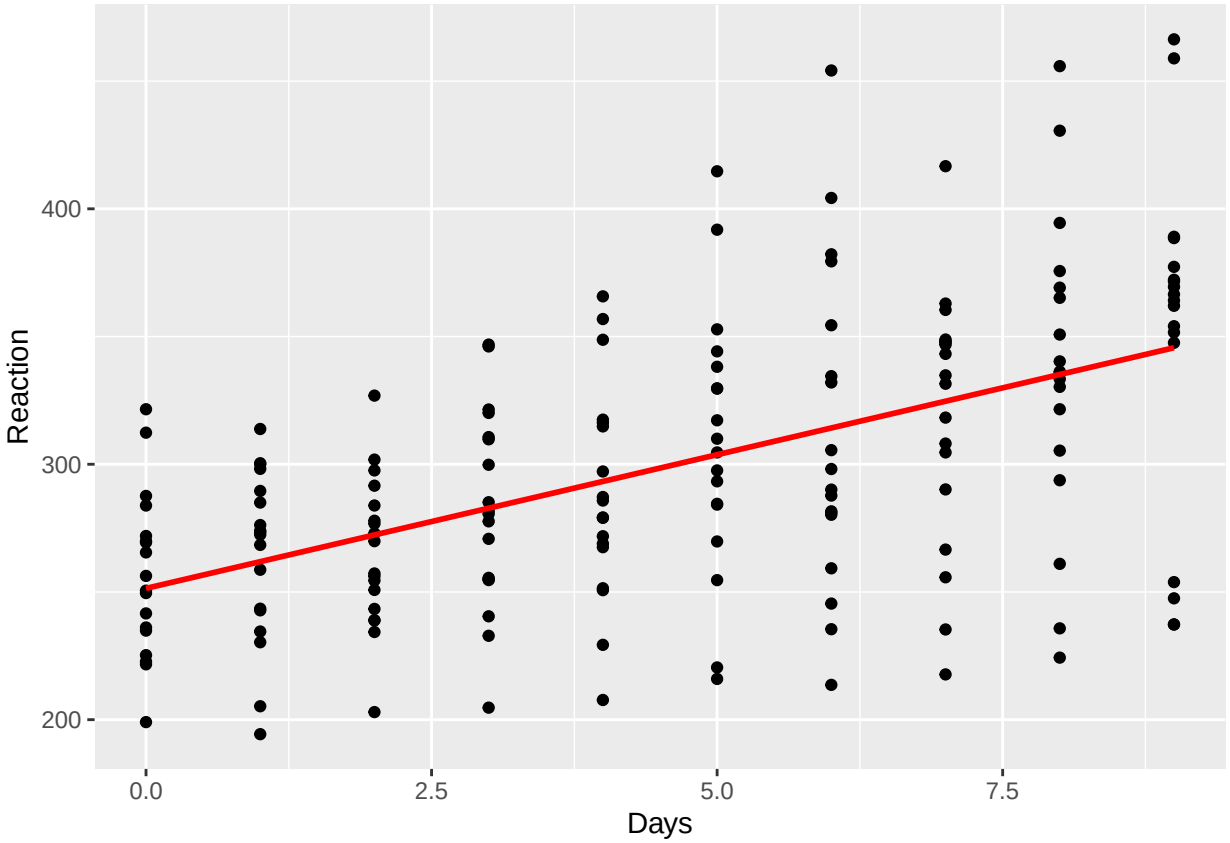
Since the p-value for days is 9.9e-15, less than 0.05, we can state that the relationship between DAYS and REACTION is highly statistically significant.

Since the multiple R-squared represents the proportion of variance in the response variable explained by the model, we can state that the model explains, approximately, 28.65% of the variance in the response variable, REACTION.

**f) Show the data adequately in a scatter plot by adding the relationship obtained by the regression model.**

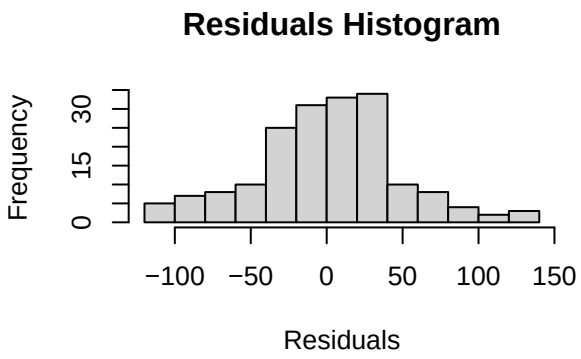
```
library(ggplot2)
scatter_plot <- ggplot(formatted_data, aes(x = Days, y = Reaction)) +
  geom_point() +  # Add points
  geom_smooth(method = "lm", se = FALSE, col = "red")  # Add regression line

# Print the scatter plot
print(scatter_plot)
```

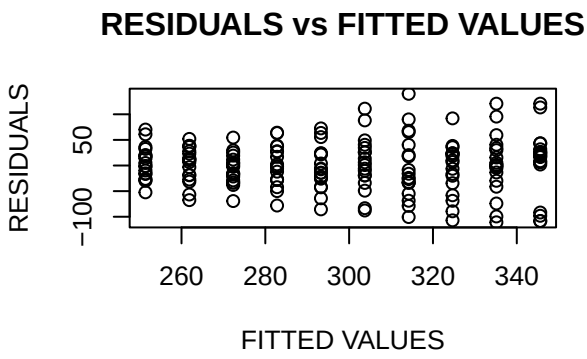
```
## `geom_smooth()` using formula = 'y ~ x'
```

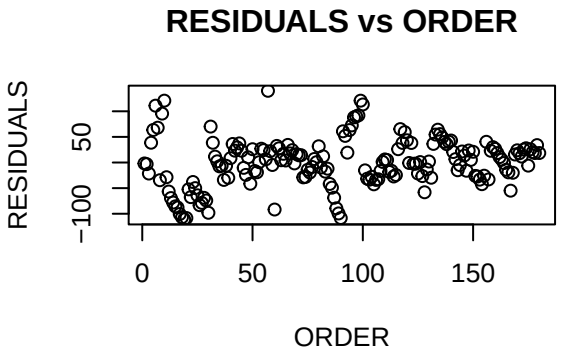


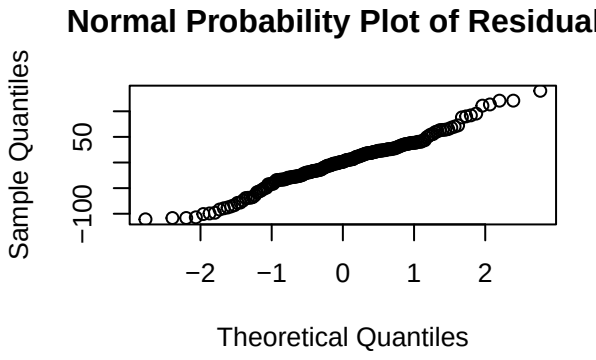

**g) Make the standard plots for the residuals of this regression (histogram, residuals versus fitted values, residuals versus order, normal probability plot) and indicate whether you believe if the standard regression assumptions hold or not.**

```
residuals_model <- resid(OLR_model)

#Way to cluster all the plots together
par(mfrow = c(2, 2))

#HISTOGRAM:
hist(residuals_model, main ="Residuals Histogram", xlab="Residuals" )


#RESIDUALS vs FITTED VALUES:
plot(fitted(OLR_model), residuals_model, main = "RESIDUALS vs FITTED VALUES",
     xlab = "FITTED VALUES", ylab = "RESIDUALS")


#RESIDUALS vs ORDER:
plot(1:length(residuals_model), residuals_model, main = "RESIDUALS vs ORDER",
     xlab = "ORDER", ylab = "RESIDUALS")


#NORMAL PROBABILITY PLOT:
qqnorm(residuals_model, main = "Normal Probability Plot of Residuals")
```



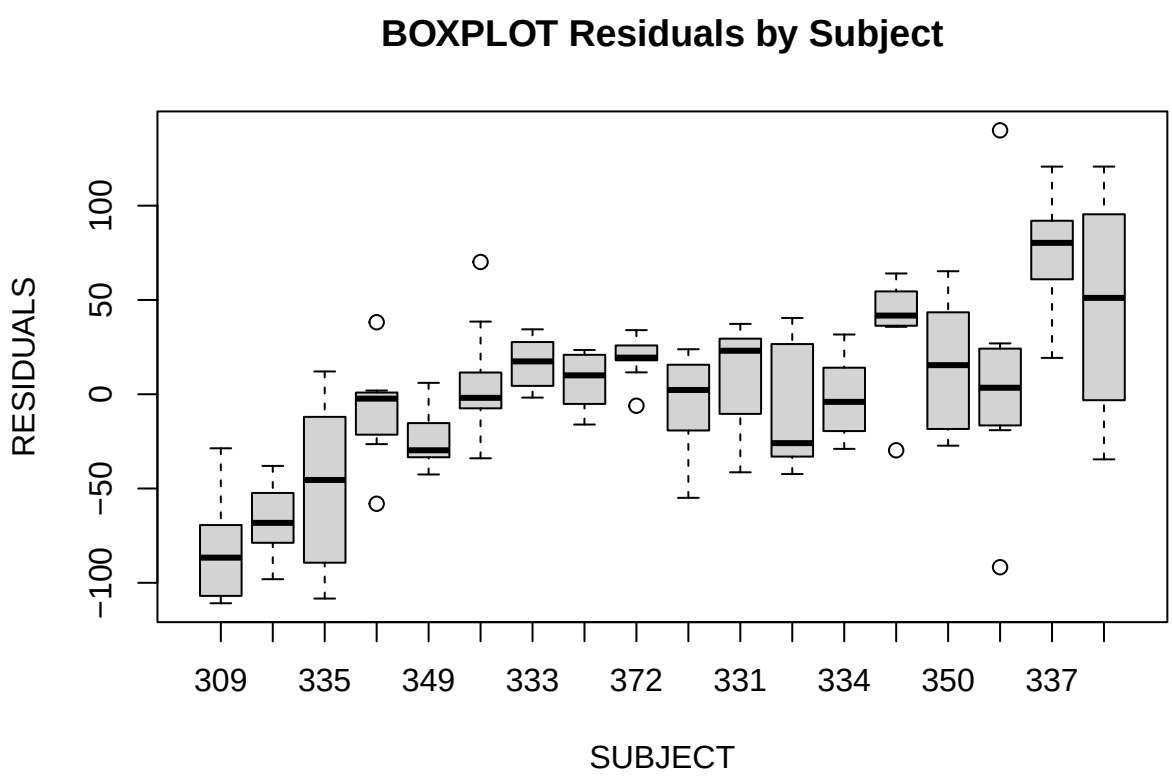

We can observe, in the histogram, that does not follow a normal distribution –> the assumption of normality does not hold.

We can observe that, in the residuals vs fitted values plot, does not follow homoscedasticity (constant variance) –> standard regression is not hold.

We can observe that, in normal probability plot, the point follow a straight line, meaning that standard regression assumptions is not violated.

**h) (1p) Make boxplots of the residuals for each Subject. Do you observe any problems?**

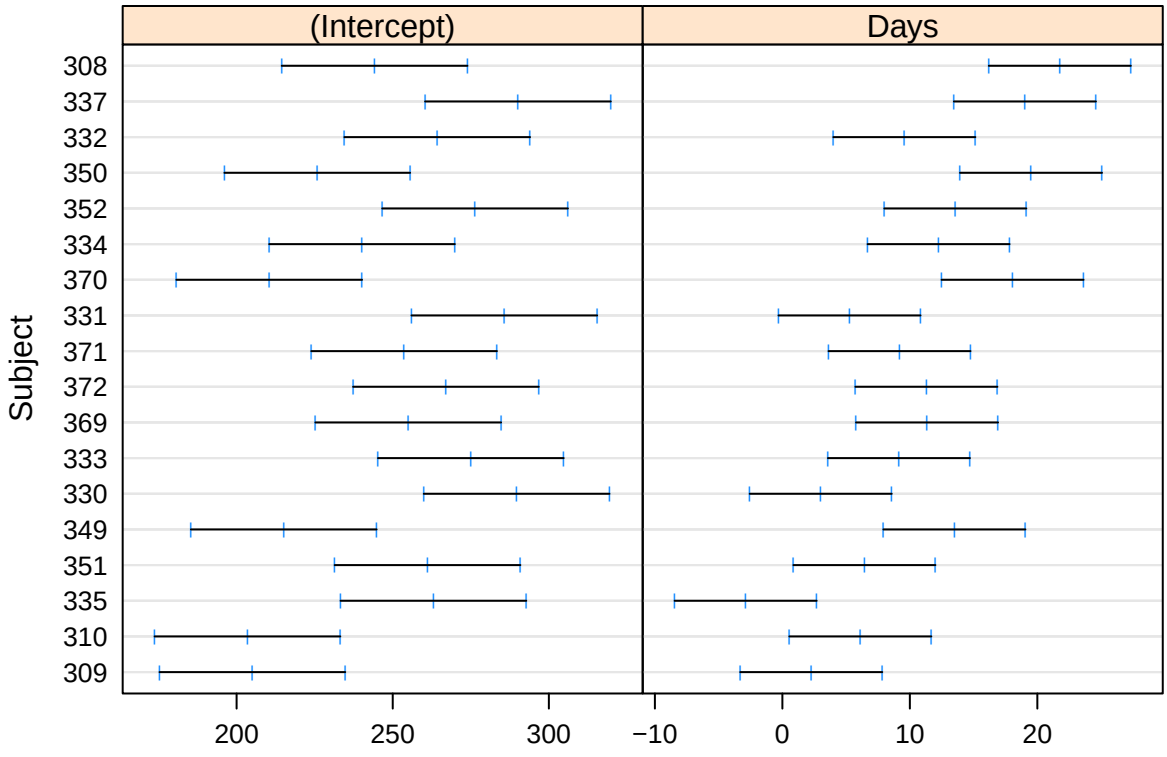
```
#We must combine residuals & subject into a data frame:
residuals_df <- data.frame(Subject = formatted_data$Subject, Residuals = residuals_model)

#Boxplots:
boxplot(Residuals ~ Subject, data = residuals_df,
        main = "BOXPLOT Residuals by Subject",
        xlab = "SUBJECT", ylab = "RESIDUALS")
```




The plot reveals that for some individuals the residuals are big and not centered at 0. If residuals are not centered at zero it means that the model doesn't contain all the information needed to explain the dependent variable (reaction) and thus it is not performing well on these individuals.

**i) Separate regressions for each Subject using the lmList instruction, and create all 95% confidence intervals for the intercepts and the slopes, using the intervals function. Display all intervals in a graph. Do you think intercepts and slopes vary significantly across Subjects? What model do the graphs suggest you?**

```
lm_list <- lmList(Reaction ~ Days | Subject, data = formatted_data)


# Obtain confidence intervals for intercepts and slopes
ci <- intervals(lm_list, level = 0.95)

# Plot confidence intervals for intercepts and for slopes
plot(ci)
```




Visually inspecting the graph, there seems to be a significant variation in the intercepts and slopes across subjects. Taking this characteristic of the graphs into account and knowing that all individuals have more or less a linear trend (i.e. there seems to be a linear relationship between "reaction" and "days") it seems that the model would be a mixed model.

**j) Fit a random intercept model to the data with lme. Use the output to obtain an estimate of the intraclass correlation coefficient. Do you think observations are independent?**

```
random_intercept_model <- lme(Reaction ~ Days, random = ~1 | Subject, data = formatted_data)

# Extract the variance components
var_components <- VarCorr(random_intercept_model)
residual_var <- var_components["Residual", "Variance"]
residual_var = as.numeric(residual_var)
re_var <- var_components["(Intercept)", "Variance"]
re_var = as.numeric(re_var)

# Calculate the intraclass correlation coefficient (ICC)
icc <- re_var/(re_var+residual_var)

print(paste("Estimated Intraclass Correlation Coefficient (ICC):", round(icc, 4)))
```

```
## [1] "Estimated Intraclass Correlation Coefficient (ICC): 0.5893"
```

Knowing that the ICC is 0.5893, it can be concluded that observations are probably not independent (because the observations inside the same group in the current grouping are more similar among them than with the rest of observations).

**k) Compare the ordinary regression model with the random intercept model using a likelihood ratio test (LRT). Which model fits the data better?**

```
anova(random_intercept_model, OLR_model)
```

```
##                         Model df      AIC      BIC    logLik   Test  L.Ratio
## random_intercept_model      1  4 1794.467 1807.194 -893.2334
## OLR_model                   2  3 1899.664 1909.210 -946.8322 1 vs 2 107.1975
##                         p-value
## random_intercept_model
## OLR_model                <.0001
```

As the p-value it's smaller than 0.0001, there are differences between the two models and the random intercept model (i.e. the model with random effect in the intercept) fits the data better.

**l) Give the value of the corresponding LR test statistic, its reference distribution and the p-value.**

As can be seen in the output of the likelihood ratio test (LRT) performed in the previous section (i.e., question k), the LR test statistic is 107.1975. Secondly, the reference distribution it's a chi-squared distribution (with 1 degrees of freedom). Regarding the p-value, it turns out to be smaller than 0.0001.

**m) Fit an ordinary regression model (with lm) using EducationLevel as the sole predictor for Reaction. Is there evidence for an effect of EducationLevel on Reaction? Which EducationLevel/s has/d the highest levels of Reaction?**

```
# Fit the ordinary regression model:
ORM_EducationLevel <- lm(Reaction ~ EducationLevel, data = formatted_data)
summary(ORM_EducationLevel)
```

```
##
## Call:
## lm(formula = Reaction ~ EducationLevel, data = formatted_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.800  -41.114   -8.874   39.230  170.748
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     292.338     10.683  27.364   <2e-16 ***
## EducationLevel    3.264      5.199   0.628    0.531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.42 on 178 degrees of freedom
## Multiple R-squared:  0.00221,    Adjusted R-squared:  -0.003396
## F-statistic: 0.3942 on 1 and 178 DF,  p-value: 0.5309
```

No, there is no evidence for an effect of the predictor "EducationLevel" on the response variable "Reaction" because the associated p-value (0.531) is not smaller than 0.05.

To answer which EducationLevel/s has/d the highest levels of Reaction, the following approach can be used:

```
# Find the mean level of Reaction for each EducationLevel:
sum_level_1 = 0
sum_level_2 = 0
sum_level_3 = 0
```

```
for (i in 1:nrow(formatted_data)) {
  if (formatted_data[i, "EducationLevel"] == 1) {
    sum_level_1 = sum_level_1 + formatted_data[i, "Reaction"]
  } else if (formatted_data[i, "EducationLevel"] == 2) {
    sum_level_2 = sum_level_2 + formatted_data[i, "Reaction"]
  } else {
    sum_level_3 = sum_level_3 + formatted_data[i, "Reaction"]
  }
}

level_1 = sum_level_1/length(sum_level_1)
level_2 = sum_level_2/length(sum_level_2)
level_3 = sum_level_3/length(sum_level_3)

cat("Mean Reaction in level 1: ", level_1)
```

```
## Mean Reaction in level 1:  20543.65
```

```
cat("Mean Reaction in level 2: ", level_2)
```

```
## Mean Reaction in level 2:  18228.94
```

```
cat("Mean Reaction in level 3: ", level_3)
```

```
## Mean Reaction in level 3:  14958.03
```

According to the results, the Education Level with highest level of Reaction is level 1.

**n) Fit now a random intercept model with EducationLevel as the sole predictor. Does this fit the data better than a model with no random intercept?**

```
RIM_EducationLevel <- lme(Reaction ~ EducationLevel, data = formatted_data, random = ~1|Subject)

summary(RIM_EducationLevel)
```

```
## Linear mixed-effects model fit by REML
##   Data: formatted_data
##        AIC      BIC    logLik
##   1905.549 1918.276 -948.7746
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    36.91384 44.25914
##
## Fixed effects:  Reaction ~ EducationLevel
##                    Value Std.Error  DF  t-value p-value
## (Intercept)    292.33775  23.63714 162 12.36773  0.0000
## EducationLevel   3.26419  11.50335  16  0.28376  0.7802
##  Correlation:
##                (Intr)
## EducationLevel -0.919
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -2.5057336 -0.5552854 -0.1407963  0.5038269  3.3328853
##
```

```
## Number of Observations: 180
## Number of Groups: 18
```

In order to know if this model fits the data better than a model with no random intercept we can perform a LRT (likelihood ratio test) using RIM_EducationLevel model and the model defined in the previous section (i.e., question m):

```
anova(RIM_EducationLevel, ORM_EducationLevel)
```

```
##                     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## RIM_EducationLevel      1  4 1905.549 1918.276 -948.7746
## ORM_EducationLevel      2  3 1956.817 1966.362 -975.4083 1 vs 2 53.26751  <.0001
```

Considering that the p-value is smaller than 0.0001 we can affirm that there is a difference between the two models and that the model with the random intercept (RIM_EducationLevel) fits the data better.

**o) Fit a mixed model with random intercept for Reaction, using both predictors, Days and EducationLevel. How many parameters has this model? Are all terms significant?**

```
# Fit the model:
RIM_EduDays <- lme(Reaction ~ Days+EducationLevel, data = formatted_data, random = ~1|Subject)
summary(RIM_EduDays)
```

```
## Linear mixed-effects model fit by REML
##   Data: formatted_data
##        AIC     BIC    logLik
##   1789.689 1805.57 -889.8445
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    38.24228  30.9914
##
## Fixed effects:  Reaction ~ Days + EducationLevel
##                    Value Std.Error  DF  t-value p-value
## (Intercept)    245.23506 23.912583 161 10.25548  0.0000
## Days            10.46727  0.804226 161 13.01533  0.0000
## EducationLevel   3.26419 11.503351  16  0.28376  0.7802
##  Correlation:
##                (Intr) Days
## Days           -0.151
## EducationLevel -0.909  0.000
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.2339025 -0.5536702  0.0102115  0.5207122  4.2424231
##
## Number of Observations: 180
## Number of Groups: 18
```

This new model has 2 parameters: DAYS & EDUCATION LEVEL. We could also consider the (random) intercept. The p-value for "EducationLevel" is 0.7802, indicating that it is not significant in predicting "Reaction". To the contrary, the p-value for "Days" is very close to 0 (0.0000), meaning that it is significant for predicting the outcome (Reaction).

**p) Fit a new model with random slope effects for both Days and EducationLevel. Does this fit the data better than a model without random slopes?**

```r
# Fit the model:
slope_model <- lme(Reaction ~ Days+EducationLevel, data = formatted_data,
                   random = ~Days+EducationLevel|Subject)
summary(slope_model)
```

```
## Linear mixed-effects model fit by REML
##   Data: formatted_data
##        AIC      BIC    logLik
##   1756.476 1788.237 -868.2379
##
## Random effects:
##  Formula: ~Days + EducationLevel | Subject
##  Structure: General positive-definite, Log-Cholesky parametrization
##                StdDev   Corr
## (Intercept)    40.52219 (Intr) Days
## Days            5.92208  0.026
## EducationLevel 22.67480 -0.892 -0.058
## Residual       25.59212
##
## Fixed effects:  Reaction ~ Days + EducationLevel
##                    Value Std.Error  DF   t-value p-value
## (Intercept)    249.47615 17.464827 161 14.284490  0.0000
## Days            10.46727  1.545781 161  6.771506  0.0000
## EducationLevel   1.81346  9.405206  16  0.192815  0.8495
##  Correlation:
##                (Intr) Days
## Days           -0.061
## EducationLevel -0.922 -0.030
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.9548448 -0.4610764  0.0232303  0.5213505  5.1736967
##
## Number of Observations: 180
## Number of Groups: 18
```

To see if this model fits the data better than a model without random slopes:

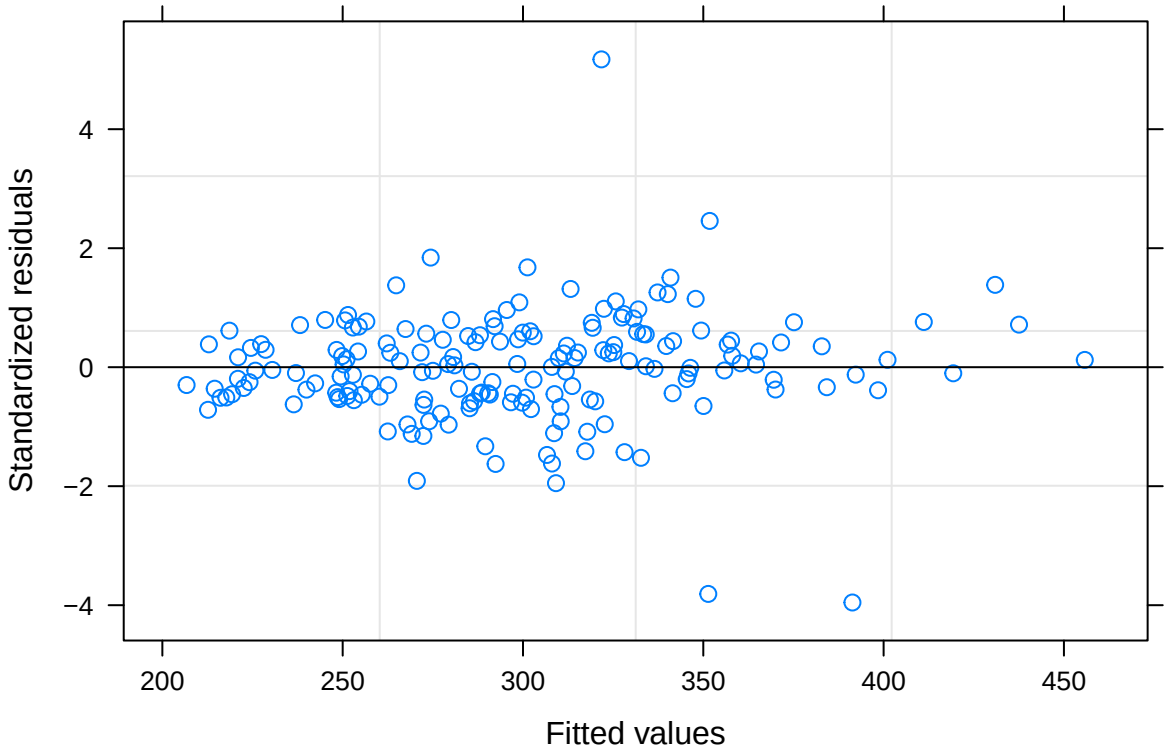
```r
anova(slope_model, RIM_EduDays)
```

```
##             Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## slope_model     1 10 1756.476 1788.237 -868.2379
## RIM_EduDays     2  5 1789.689 1805.570 -889.8445 1 vs 2 43.21314  <.0001
```

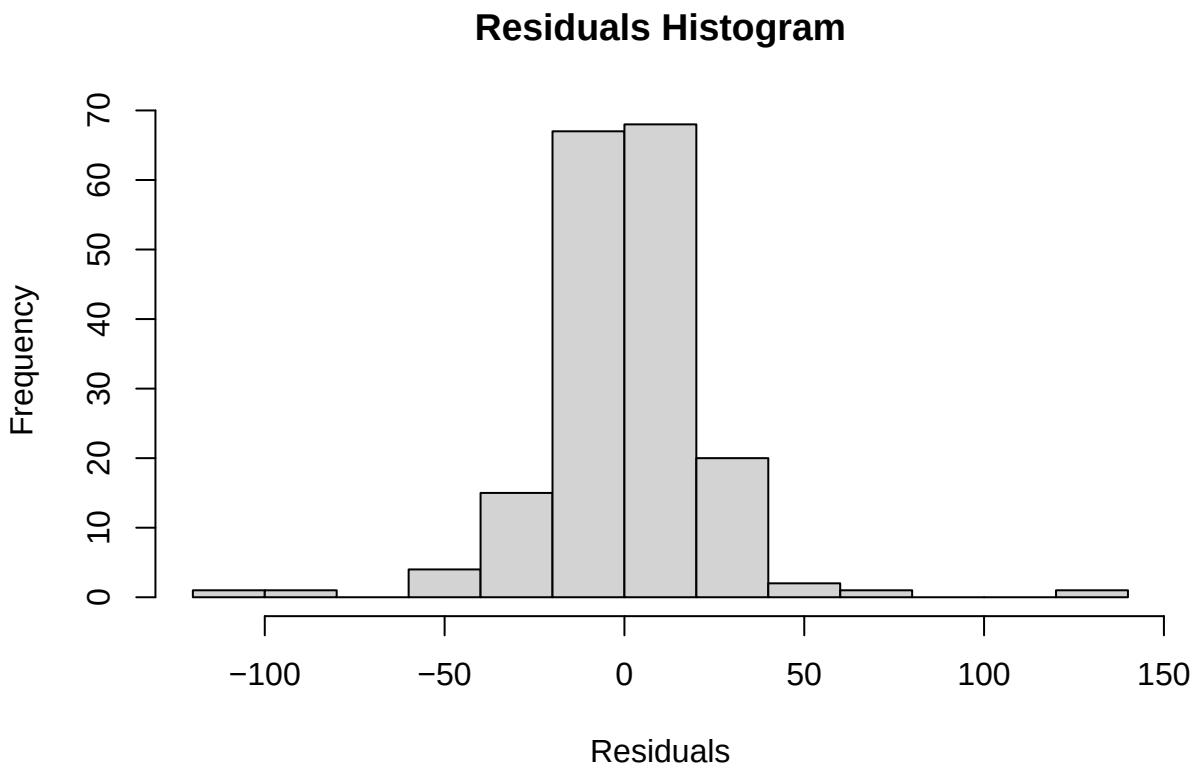
In the results it can be seen that the p-value is smaller than 0.0001. Consequently, it can be said that the model with the random slopes fits the data better.

**q) Investigate the residuals of this model of your by making some plots you consider adequate. Comment on your results.**

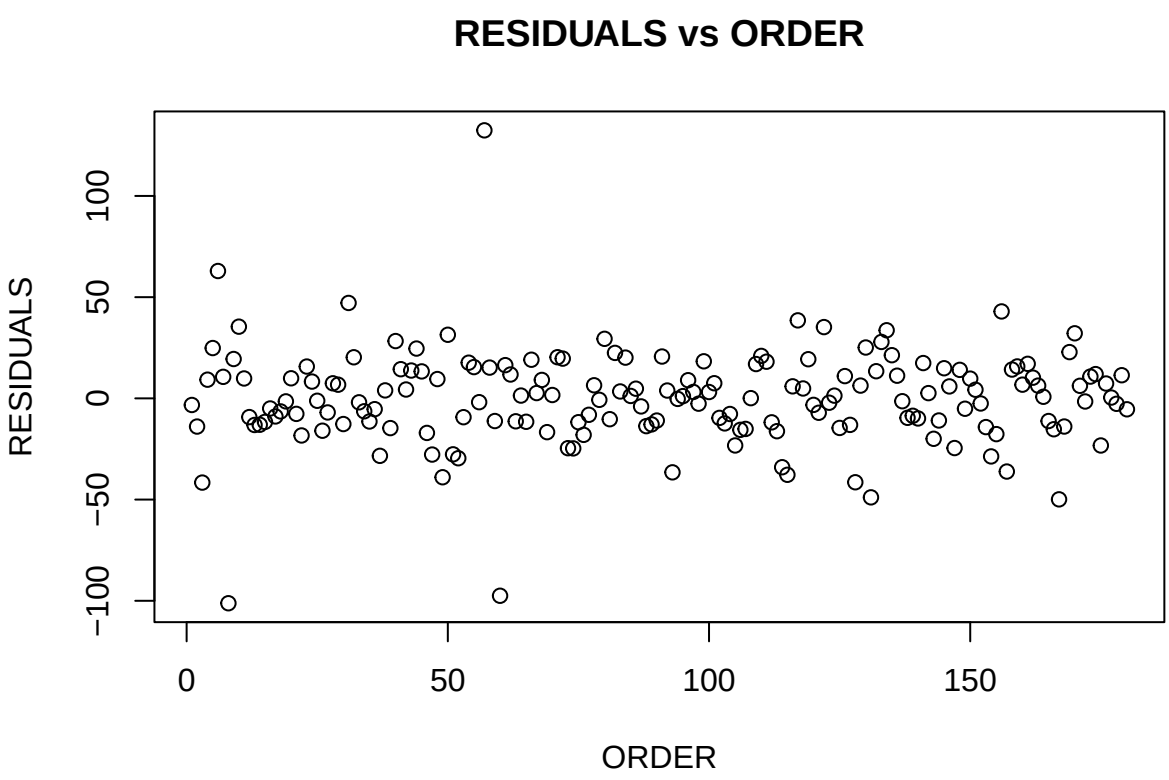
```r
# Plot 1
plot(slope_model)
```




```
# Plot 2:
hist(residuals(slope_model), main ="Residuals Histogram", xlab="Residuals" )
```




```
# Plot 3:
plot(1:length(residuals(slope_model)), residuals(slope_model), main = "RESIDUALS vs ORDER",
     xlab = "ORDER", ylab = "RESIDUALS")
```




In the plots it can be seen that the residuals of the model are mainly located around 0. Specifically, the Q-Q plot reveals some issues in the right and left tails but the other parts seem to be correct.

**r) What would be your final model for the data? Justify your answer.**

In total, six models have been fitted:

1. OLR_model (ordinary linear regression model with Days as the predictor and Reaction as response)
2. random_intercept_model (random intercept model with Days as the predictor and Reaction as response)
3. ORM_EducationLevel (ordinary linear regression model with Days as the predictor and Reaction as response)
4. RIM_EducationLevel (random intercept model with EducationLevel as the predictor and Reaction as response)
5. RIM_EduDays (random intercept model with Days and EducationLevel as predictors and Reaction as response)
6. slope_model (model with random slope effects for both Days and EducationLevel)

To begin with, as discussed in previous sections ordinary linear regression models do not contain enough information to predict correctly the response variable "Reaction" and they perform worse than a model with random intercept containing the same predictor (this has been tested with LRT in previous sections). Consequently, OLR_model and ORM_EducationLevel would be discarded. Next, in order to compare different mixed effect models, AIC (Akaike information criteria) can be used because the values of likelihood are corrected or characterized by penalizing the fact of using a lot of factors. Usually, the smaller the AIC, the better the model is.

```
AIC_random_intercept_model = AIC(random_intercept_model)
AIC_RIM_EducationLevel = AIC(RIM_EducationLevel)
```

```r
AIC_RIM_EduDays = AIC(RIM_EduDays)
AIC_slope_model = AIC(slope_model)
cat("AIC of random_intercept_model: ", AIC_random_intercept_model)
```

```
## AIC of random_intercept_model:  1794.467
```

```r
cat("AIC of RIM_EducationLevel: ", AIC_RIM_EducationLevel)
```

```
## AIC of RIM_EducationLevel:  1905.549
```

```r
cat("AIC of RIM_EduDays: ", AIC_RIM_EduDays)
```

```
## AIC of RIM_EduDays:  1789.689
```

```r
cat("AIC of slope_model: ", AIC_slope_model)
```

```
## AIC of slope_model:  1756.476
```

The model with lowest AIC is slope_model (1756.476), so it would be the final model for the data (among the ones that have been fitted until now).

**s) What would be your next step in the analysis of this data set? Comment your suggestion.** After fitting several ordinary linear regression models and mixed effects models, comparing them and deciding which will be the final model, the next step could be to interpret the results, analyze in more depth the residuals of the model and prepare a report explaining the performance of the model.

**t) Give examples of outcomes that can be modelled using mixed models, longitudinal, hierarchical, cluster, and repeated measurements are possible options.**

**MIXED MODELS:**

- Growth of plants in different experimental conditions.
- Student test scores within classrooms.

**LONGITUDIONAL MODELS:**

- Blood pressure measurements over time for a group of patients.
- Academic Performance of Students Over Several Semesters.

**HIERARCHICAL MODELS:**

- Student academic performance within classrooms, within schools.
- Patient outcomes within hospitals, within regions.

**CLUSTER MODELS**

- Survey responses on job satisfaction from employees in different companies.
- Exam scores of students from different districts.

**REPEATED MEASUREMENTS MODELS**

- Body weight measurements of individuals over the course of a diet intervention.
- Reaction times of individuals in response to different stimuli measured in a psychology experiment.

**u) How can we extend the linear model regression to a generalized linear model? Which library and function in R can be used to fit a generalized linear model?**

Linear regression models make assumptions about the normal distribution of the response variable, but real data often don't follows normality. Generalized linear models (GLMs) address this by extending linear models to accommodate a wider range of response variable distributions and relationships.

To implement a generalized linear model in R, the '**glm function**' is utilized. This function enables the specification of a distribution family for the response variable and a link function connecting predictors to the response variable's mean.

This function is included in the primary R library in a package called **'stats'.**