Assignment 5
Mixed effect models

Jan Izquierdo & Ainhoa Lopez

2023-11-14

Libraries

```
options(warn=-1)
library(nlme)
library(lmtest)
library(lme4)
library(ggplot2)
```

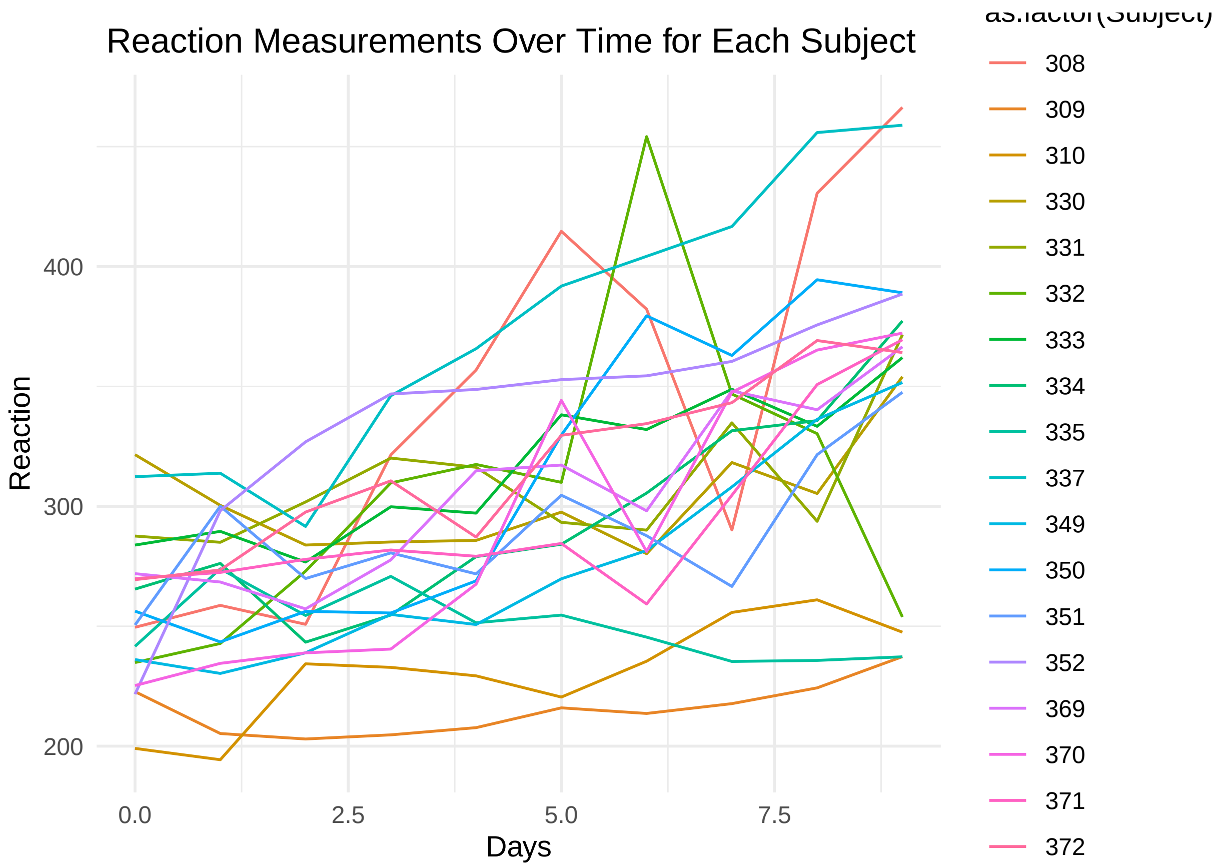
1.a) Read the dataset into the R environment. .

```
mixed_effect_data <-readxl::read_excel("./Assignment-data/sleepstudy_fixed.xlsx")
sapply(mixed_effect_data, class)
```

```
##          Order       Reaction           Days        Subject EducationLevel
##      "numeric"      "numeric"      "numeric"      "numeric"      "numeric"
```

1.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? .

```
ggplot(mixed_effect_data, aes(x = Days, y = Reaction, group = Subject, color = as.factor(Subject))) +
  geom_line() +
  labs(title = " Reaction Measurements Over Time for Each Subject",
       x = "Days", y = "Reaction") +
  theme_minimal()
```



```
## The data in mixed_effect_data is nested and structured in a data frame where variables vary at one
## or more levels. Therefore, the data follows a mixed-effects model, also known as a hierarchical
## model, due to the influence of different levels on the data. The reaction of each subject is
## independent of the reaction of the other subjects. However, if we focus on a specific subject, their
## reaction measurements will be interrelated. Since there are different observations for the same
## subject, we can assume that there is grouping. The fact that there are repeated measurements of
## individuals over time gives us reason to expect that the reaction measurements cannot be considered
## independent.
```

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

```
X <-groupedData(Reaction~Days|Subject,data=mixed_effect_data)
reaction <- as.numeric(X$Reaction)
```

***1.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? .**

```
number_subj <- length(unique(X$Subject))

table(X$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
```

```
ES<-table(X$EducationLevel, X$Subject)
Ed1<-table(ES[1,])
Ed2<-table(ES[2,])
Ed3<-table(ES[3,])
```

```
## The total number of subjects is  18

## We can see that the data is not balanced because:

## Education level 1:

## 
##  0 10 
## 11  7

## Education level 2:

## 
##  0 10 
## 12  6

## Education level 3:

## 
##  0 10 
## 13  5

## For education level 1 10 measurements were only done on 7 subjects, while on level 2 10
##     measurements were done on 6 subjects and on level 3 10 measurements were done on only 5 subjects

## From the tables we can see that the maximum mesaurements on a single subject is 10
```

1.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?** .

```
model<-lm(Reaction~Days, data=X)
summary(model)
```

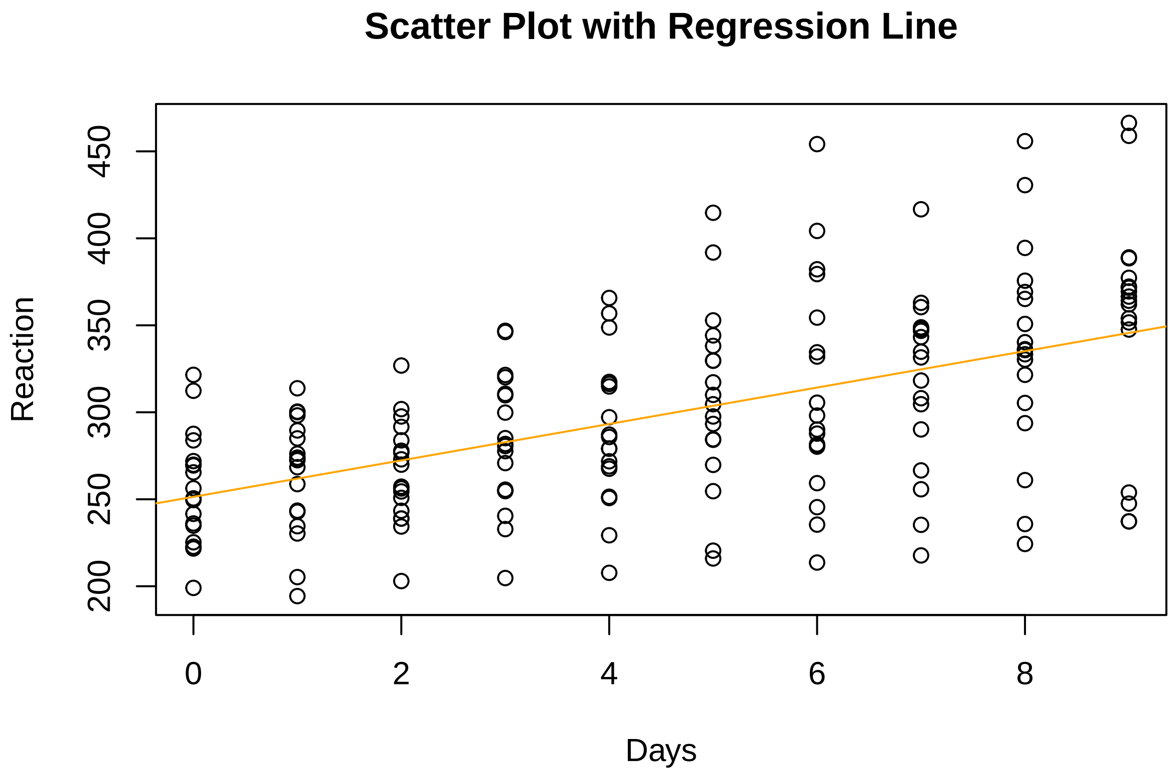
```
## 
## Call:
## lm(formula = Reaction ~ Days, data = X)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -110.848  -27.483    1.546   26.142  139.953 
## 
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  251.405      6.610  38.033  < 2e-16 ***
## Days          10.467      1.238   8.454 9.89e-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: 9.894e-15
```

```
## There is a significant relationship, as the p-value is inferior to the default critical value 0.05
```

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

```
plot(X$Days, X$Reaction, main = "Scatter Plot with Regression Line",
     xlab = "Days", ylab = "Reaction")
abline(model, col = "orange")
```



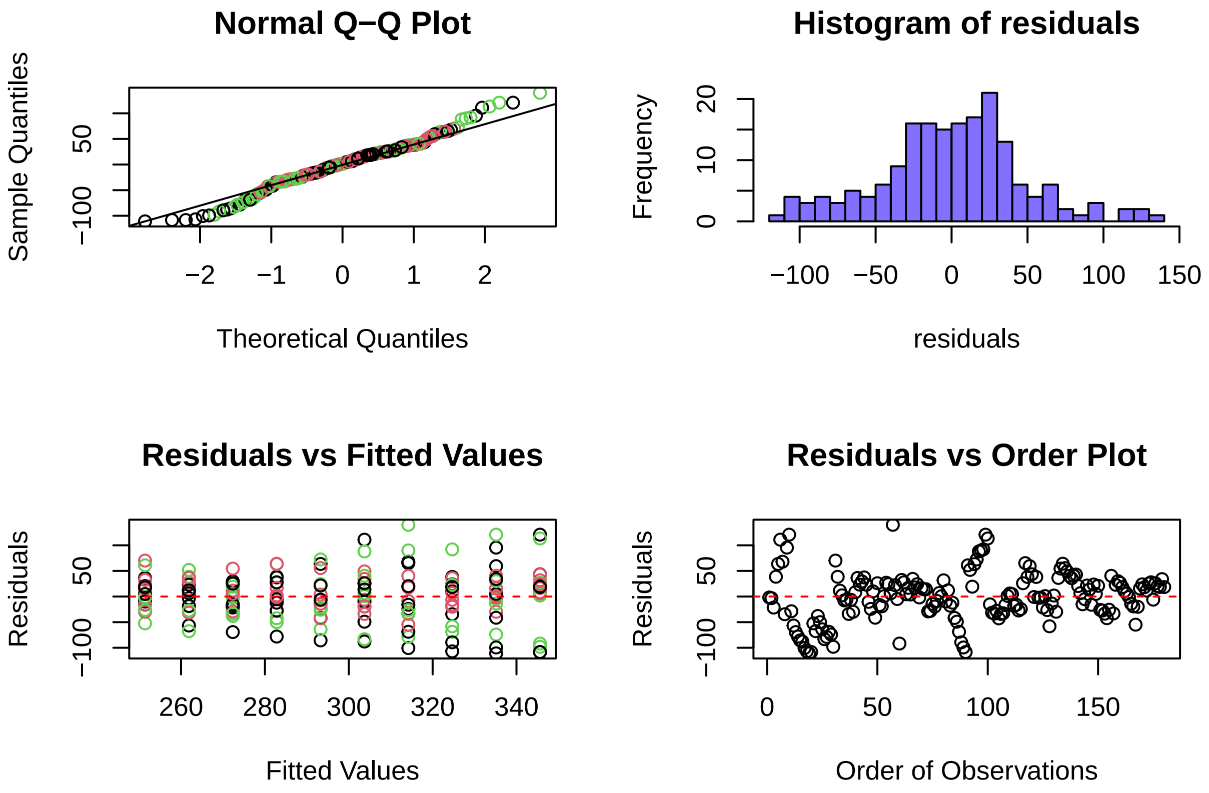
1.g) **Make the standard plots for the residuals of this regression (histogram, residuals versus fitted values, residuals versus order, normal probability plot) and indicate whether you believe if the standard regression assumptions hold or not.** .

```
par(mfrow = c(2, 2))
#Normal probability plot
residuals <- resid(model)
qqnorm(residuals, col = X$EducationLevel)
qqline(residuals)
yhat <- predict(model)


#Histogram
hist(residuals, breaks =25, col='lightslateblue')


#Residuals versus fitted values
plot(yhat, residuals, col = X$EducationLevel,
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)

plot(seq_along(residuals), residuals,
     main = "Residuals vs Order Plot",
     xlab = "Order of Observations",
     ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
```



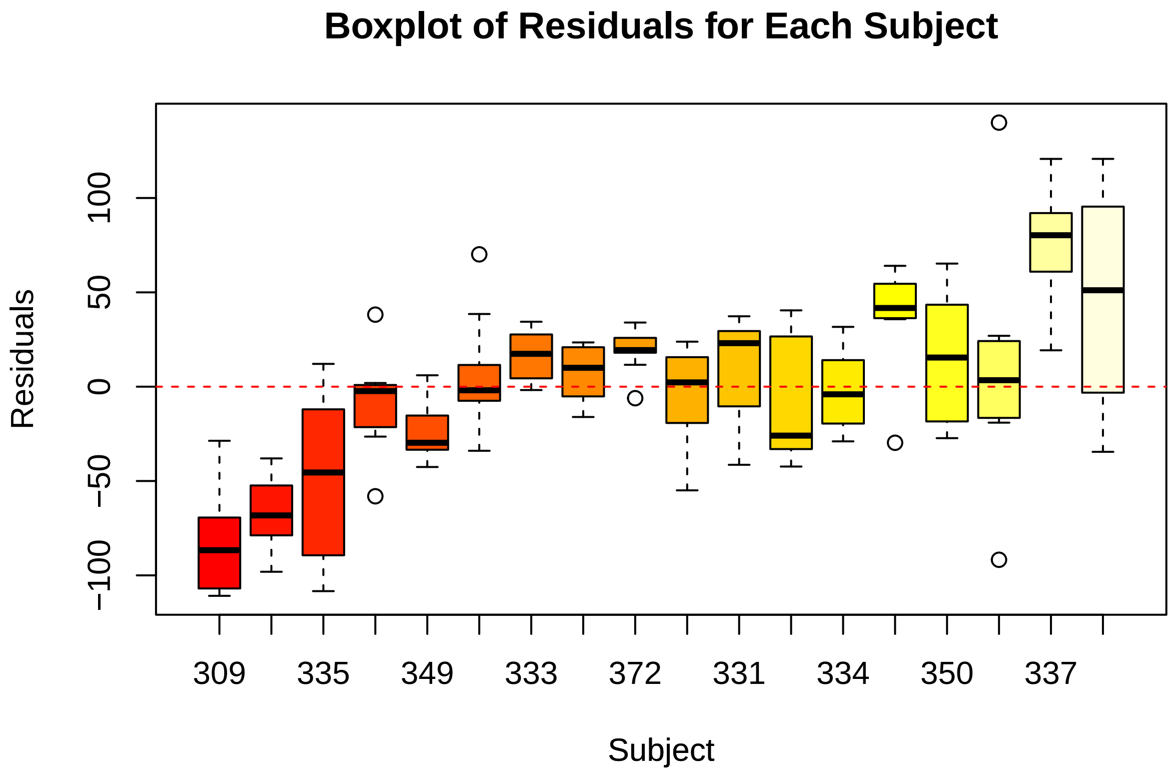
```
## The standard regression assumptions hold because the residuals mostly fit the distribution, for
## example in the QQ plot most of the residuals are located along the standard regression line, which
## means that the data does fit the distribution
```

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

```
boxplot(residuals ~ Subject, data = X, col = heat.colors(length(unique(X$Subject))), ylab="")

abline(h = 0, col = "red", lty = 2)

title(main = "Boxplot of Residuals for Each Subject", xlab = "Subject", ylab = "Residuals")
```

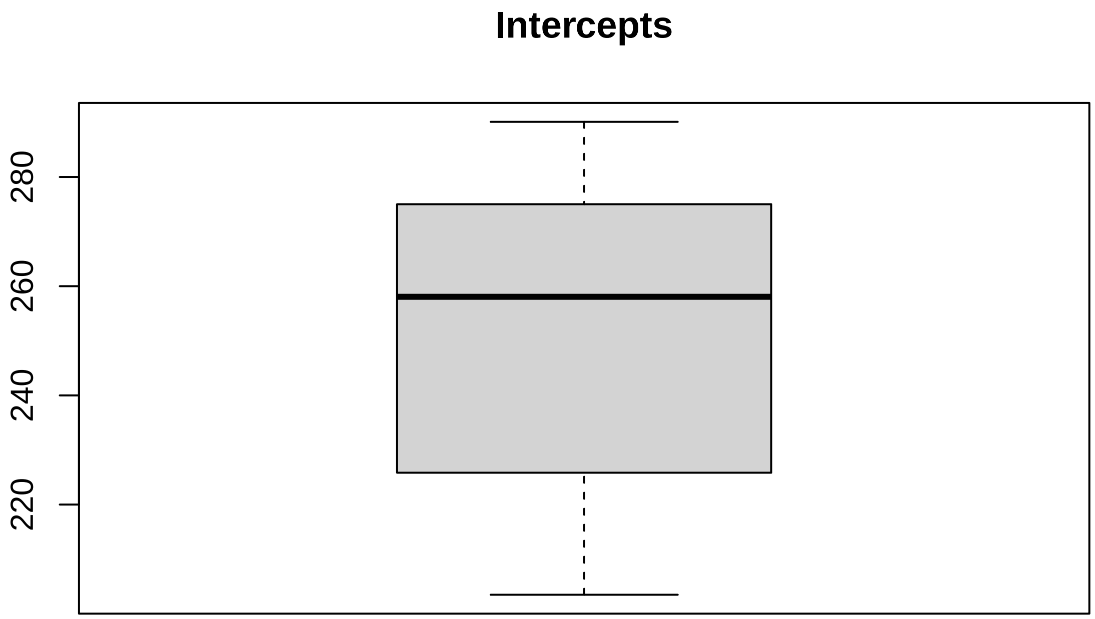


```
## The only problem I can observe is that the subject data is much more represented when the residuals
## are lower( this can be seen in the more intense coloration of the first 4 boxplots and it is caused
## by heat.color)
```

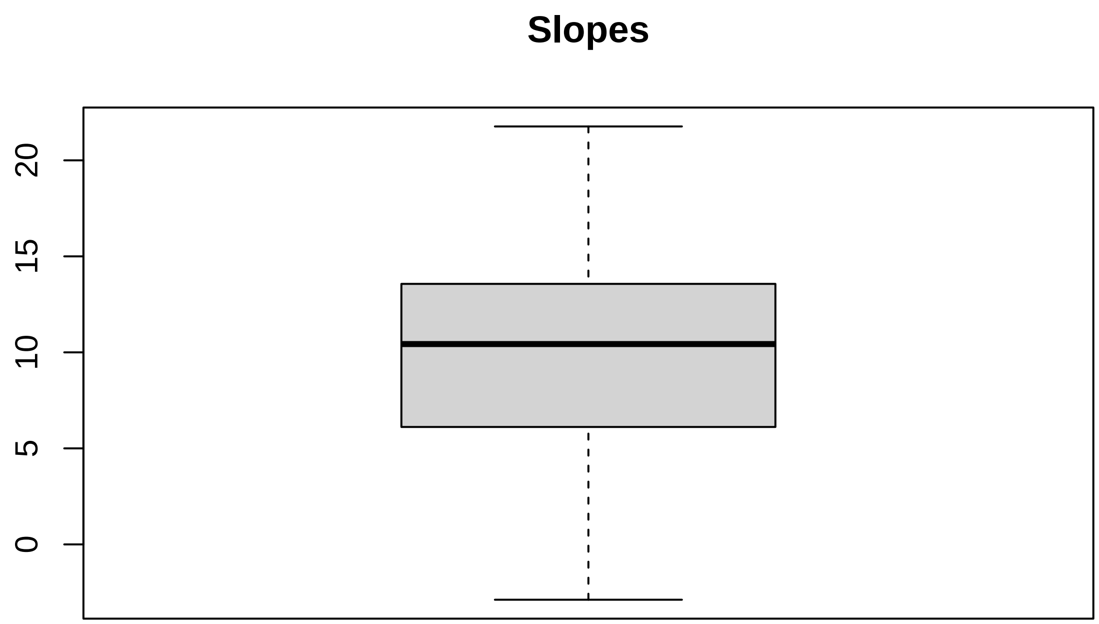
1.i) **Separate regressions for each Subject using the lmList instruction, and create all 95% confidence intervals for the intercepts and the slopes, using the intervals function. Display all intervals in a graph. Do you think intercepts and slopes vary significantly across Subjects? What model do the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

```
output <- lmList(Reaction ~ Days | Subject, data = X)
M <- coefficients(output)

boxplot(M[,1], main = "Intercepts")
```



```
boxplot(M[,2], main = "Slopes")
```

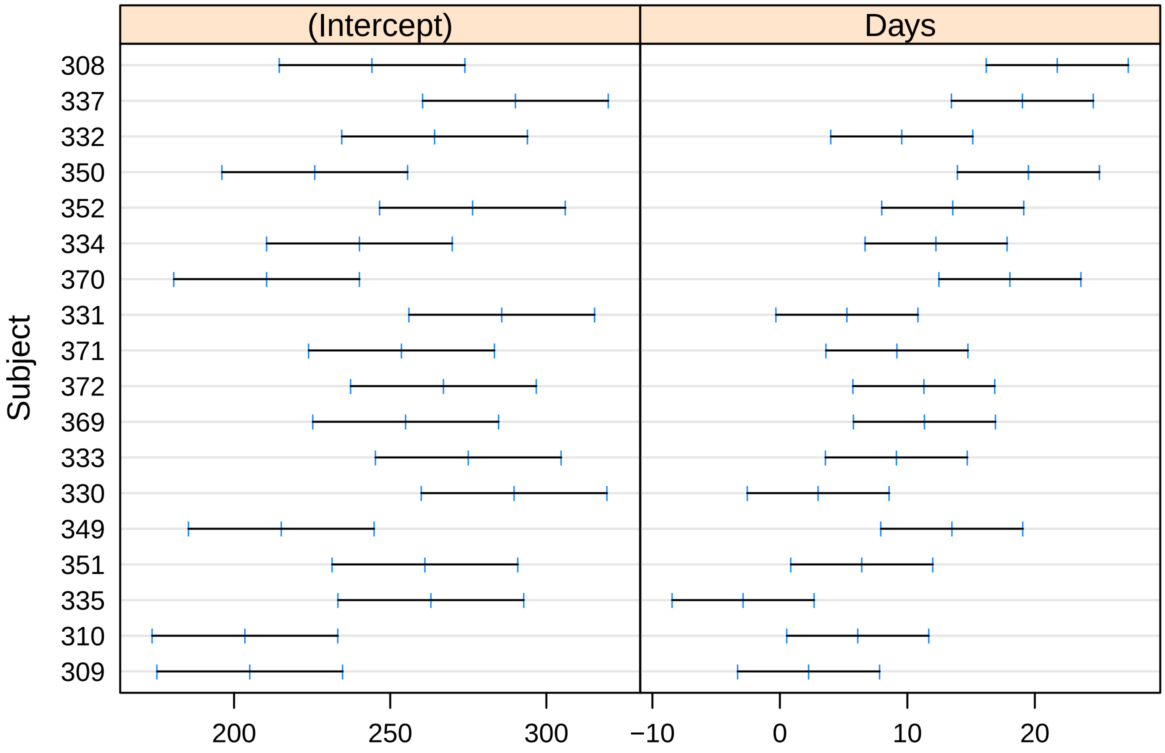


```
#95% confidence interval
inter <- intervals(output)
inter
```

```
## , , (Intercept)
##
##        lower     est.    upper
## 309 175.3239 205.0549 234.7860
## 310 173.7532 203.4842 233.2152
## 335 233.3037 263.0347 292.7657
## 351 231.4160 261.1470 290.8780
## 349 185.3807 215.1118 244.8428
## 330 259.9541 289.6851 319.4161
## 333 245.2881 275.0191 304.7501
## 369 225.2371 254.9681 284.6992
## 372 237.3138 267.0448 296.7758
## 371 223.9050 253.6360 283.3671
## 331 256.0079 285.7390 315.4700
## 370 180.7181 210.4491 240.1801
## 334 210.4319 240.1629 269.8939
## 352 246.6410 276.3721 306.1031
## 350 196.1036 225.8346 255.5656
## 332 234.5206 264.2516 293.9826
## 337 260.3731 290.1041 319.8352
## 308 214.4616 244.1927 273.9237
##
```

```
## , , Days
##
##          lower      est.     upper
## 309 -3.3073456  2.261785  7.830916
## 310  0.5457678  6.114899 11.684030
## 335 -8.4501650 -2.881034  2.688097
## 351  0.8643665  6.433498 12.002629
## 349  7.9248017 13.493933 19.063064
## 330 -2.5610583  3.008073  8.577204
## 333  3.5729144  9.142045 14.711176
## 369  5.7789781 11.348109 16.917240
## 372  5.7289423 11.298073 16.867204
## 371  3.6193138  9.188445 14.757576
## 331 -0.3031122  5.266019 10.835150
## 370 12.4870199 18.056151 23.625282
## 334  6.6840102 12.253141 17.822272
## 352  7.9974181 13.566549 19.135680
## 350 13.9348859 19.504017 25.073148
## 332  3.9976368  9.566768 15.135899
## 337 13.4568429 19.025974 24.595105
## 308 16.1955714 21.764702 27.333833
```

```
plot(inter)
```



```
## No, both models are near identical so it suggests mixed models
```

***1.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 = X)
summary(random_intercept_model)
```

```
## Linear mixed-effects model fit by REML
##   Data: X
##        AIC      BIC    logLik
##   1794.465 1807.192 -893.2325
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    37.12383 30.99123
##
## Fixed effects:  reaction ~ Days
##                 Value Std.Error  DF  t-value p-value
## (Intercept) 251.40510  9.746716 161 25.79383       0
## Days         10.46729  0.804221 161 13.01543       0
##  Correlation:
##      (Intr)
## Days -0.371
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.2256707 -0.5528788  0.0108521  0.5187971  4.2506162
##
## Number of Observations: 180
## Number of Groups: 18
```

```
variance<-  VarCorr(random_intercept_model)
sigma<- random_intercept_model$sigma
variance_total<- sigma^2  + var(fitted(random_intercept_model))
intraclasscor <- sigma^2/ variance_total
```

```
## We think there are  a significant relation because the p_values 4.605266e-59 6.412601e-27
## are smaller than the critical value 0.05 and the ICC is bigger than 0.5, this suggests substantial
## proportion of variability between groups
```

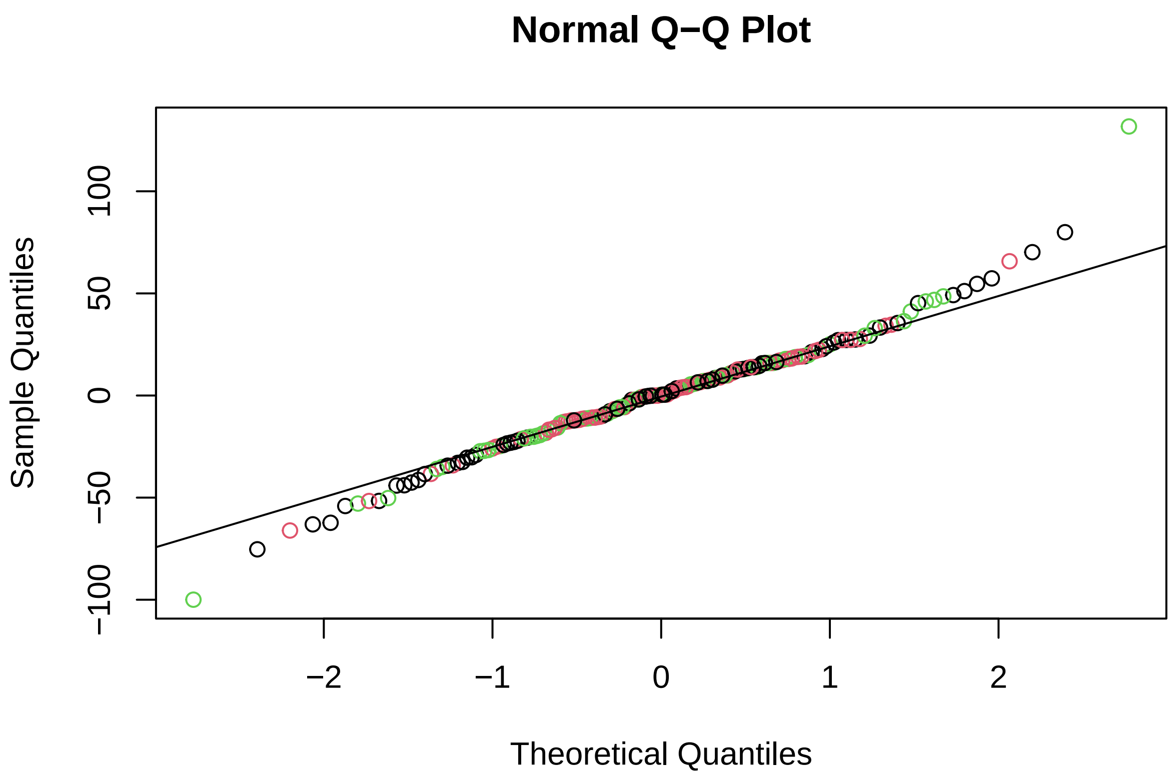
```
## Intraclass Correlation Coefficient (ICC): 0.3105238
```

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

```
lr_test <- lrtest(model, random_intercept_model)
print(lr_test)
```

```
## Likelihood ratio test
##
## Model 1: Reaction ~ Days
## Model 2: reaction ~ Days
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -950.15
## 2   4 -893.23  1 113.83  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Normal QQ plot to test the fittnes to the data
residuals <- resid(random_intercept_model)
qqnorm(residuals, col = X$EducationLevel)
qqline(residuals)
```



```
yhat <- predict(random_intercept_model)
```

```
## The model that fits the data better is the random intercept model, as if we compare the
## normal QQ plot of this model we can see that the data fits the regression line much better
```

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

```
lr_test
```

```
## Likelihood ratio test
##
```

```
## Model 1: Reaction ~ Days
## Model 2: reaction ~ Days
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   3 -950.15
## 2   4 -893.23  1 113.83  < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
t_stat<-lr_test$Chisq
p_value<-lr_test$`Pr(>Chisq)`
```

```
## The value corresponding to the LR test statistic is  NA 113.828
```

```
## The reference distribution is a chi-square distribution with 1 degree of freedom
```

```
## The p-value is NA 1.421185e-26
```

1.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?** .

```
model2<- lm(reaction~EducationLevel, data = X)
summary(model2)
```

```
##
## Call:
## lm(formula = reaction ~ EducationLevel, data = X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -107.802  -41.117   -8.875   39.232  170.747
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      292.342     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:  0.002209,   Adjusted R-squared:  -0.003396
## F-statistic: 0.3941 on 1 and 178 DF,  p-value: 0.5309
```

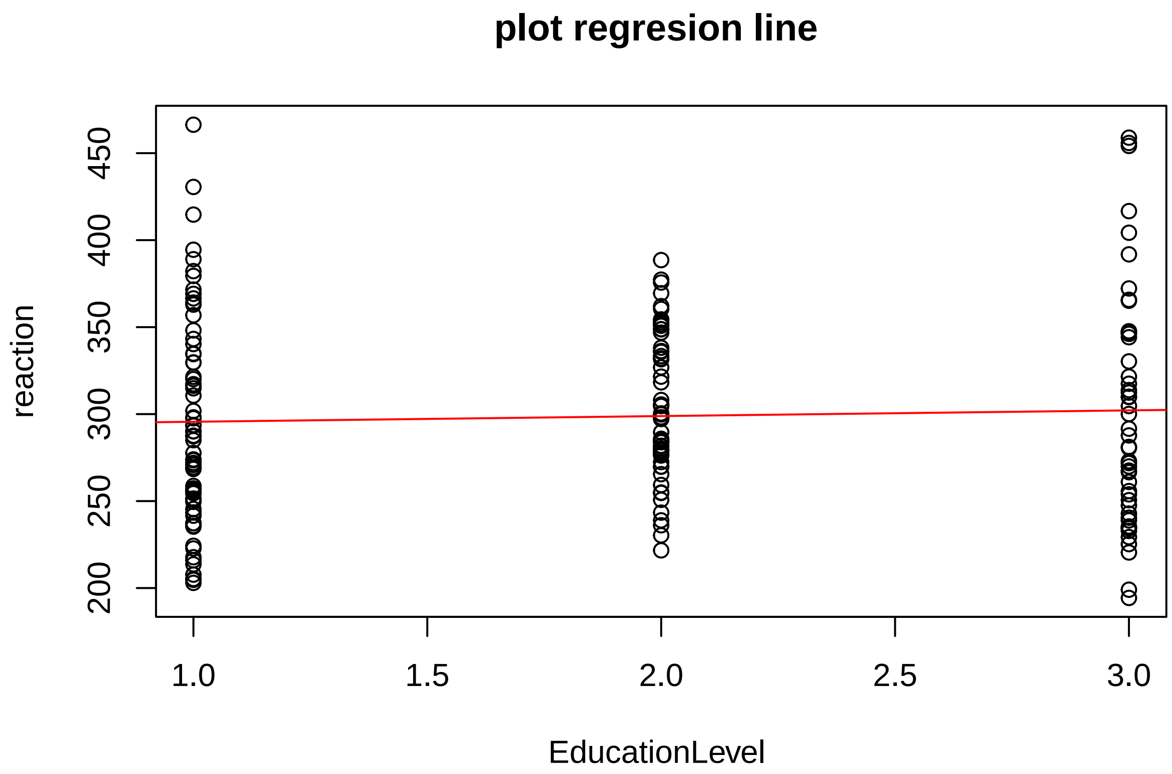
```
coe_model2<- coef(model2)
coe_model2
```

```
##    (Intercept) EducationLevel
##     292.342456       3.264054
```

```
hip_test<- anova(model2)
hip_test
```

```
## Analysis of Variance Table
##
## Response: reaction
##                 Df Sum Sq Mean Sq F value Pr(>F)
## EducationLevel    1   1255  1254.8  0.3941 0.5309
## Residuals       178 566699  3183.7
```

```
plot(X$EducationLevel, reaction, main = "plot regresion line ",
     xlab = "EducationLevel ", ylab = "reaction")
abline(model2, col = "red")
```



```
## The p-value of 0.531 suggests that the observed lack of statistical significance indicates
## weak support for an impact of the variable EducationLevel on Reaction. Despite the positive
## coefficient for EducationLevel, signifying that an increase in EducationLevel is associated
## with an expected increase in Reaction, the absence of compelling evidence at the given significance
## level makes it difficult to confidently assert that this relationship is not merely a result of
## random chance. In essence, the data does not provide sufficient support to confirm the existence
## of a meaningful effect attributable to EducationLevel on Reaction.
```

```
## The education levels that have the highest level of reaction should be 3.0
```

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?** .

```
no_random_intercept <- lm(reaction ~ EducationLevel, data = X)
random_intercept <- lmer(reaction ~ EducationLevel + (1|Subject), data = X)

lr_test2 <- anova(random_intercept,no_random_intercept)
```

```
## refitting model(s) with ML (instead of REML)
```

```
lr_test2
```

```
## Data: X
## Models:
## no_random_intercept: reaction ~ EducationLevel
## random_intercept: reaction ~ EducationLevel + (1 | Subject)
##                     npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## no_random_intercept    3 1966.7 1976.2 -980.33   1960.7
## random_intercept       4 1918.5 1931.2 -955.23   1910.5  50.2  1  1.388e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Looking at the BIC/AIC, we observe that the random intercept model has a better fit to the data
## (has the smallest value).
```

1.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?** .

```
mixed_model_random <- lmer(reaction ~ Days + EducationLevel +(1|Subject), data = X)
mixed_model_random
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: reaction ~ Days + EducationLevel + (1 | Subject)
##    Data: X
## REML criterion at convergence: 1779.687
## Random effects:
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 38.24
##  Residual             30.99
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
##    (Intercept)            Days  EducationLevel
##        245.240          10.467           3.264
```

```
summary_model <- summary(mixed_model_random)
p_values <- summary_model$coefficients[, "t value"]
p_values
```

```
##    (Intercept)           Days EducationLevel
##      10.255671      13.015428       0.283748
## We have 3 fixed effects (intercept, days, and EducationLevel) and 1 random effect (Subject).
```

```
## To determine if the fixed effects are significant, we look at the p-values. Our p-values for the
## intercept and days are <0.05, so they are considered statistically significant. However, our
## p-value for EducationLevel is > 0.05, indicating that it is not statistically significant.
```

1.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? .

```
random_model_slope <- lmer(reaction ~ Days + EducationLevel + (Days + EducationLevel | Subject), data =


no_random_model_slope <- lmer(reaction ~ Days + EducationLevel + (1 | Subject), data = X)


anova_results <- anova(no_random_model_slope, random_model_slope)
```

```
## refitting model(s) with ML (instead of REML)
```

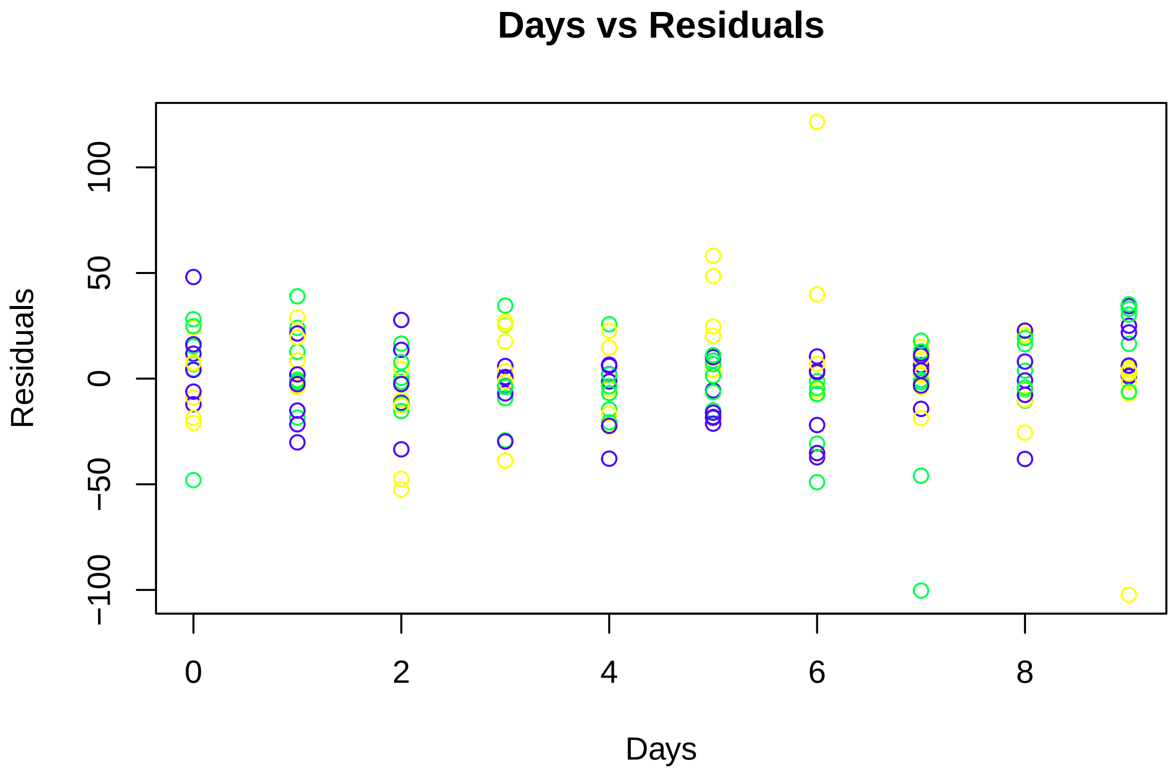
```
anova_results
```

```
## Data: X
## Models:
## no_random_model_slope: reaction ~ Days + EducationLevel + (1 | Subject)
## random_model_slope: reaction ~ Days + EducationLevel + (Days + EducationLevel | Subject)
##                       npar  AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## no_random_model_slope    5 1804 1820.0 -896.99     1794
## random_model_slope      10 1771 1802.9 -875.47     1751 43.041  5  3.625e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

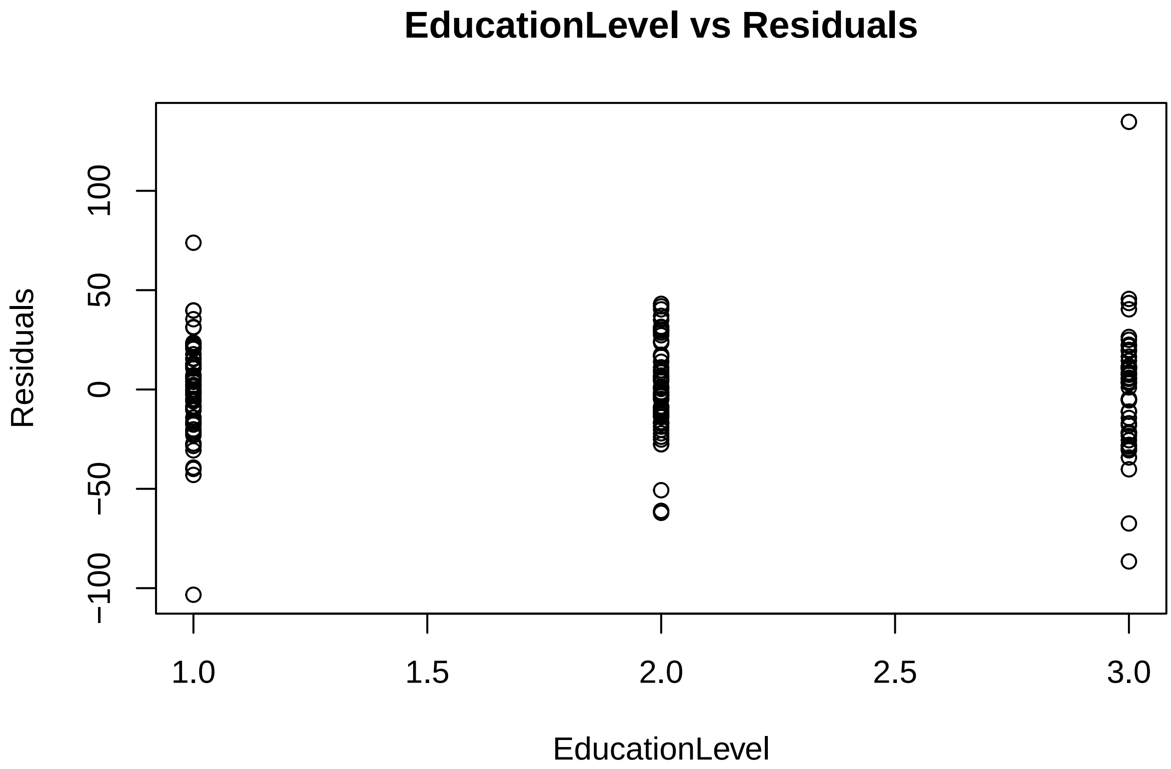
```
## Looking at the BIC/AIC, we observe that the random intercept model has a better fit to the data
## (has the smallest value).
```

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

```
res<-resid(random_model_slope)
random_effects <- ranef(random_model_slope)
plot(X$Days, res + random_effects$Subject[, "Days"], xlab = "Days", ylab = "Residuals",
     main = "Days vs Residuals", col=topo.colors(length(unique(X$EducationLevel))))
```



```
plot(X$EducationLevel, res + random_effects$Subject[, "EducationLevel"],
     xlab = "EducationLevel", ylab = "Residuals", main = "EducationLevel vs Residuals")
```



```
## The scatter plot of Days versus Residuals displays a random distribution of points, indicating that
## the model with random slopes for Days effectively addresses the variability associated with this
## predictor. Conversely, in the EducationLevel versus Residuals plot, the points are evenly scattered
## across the y-axis, providing confirmation that the model adeptly captures the overall variability
## present in the data.
```

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

```
## I would use the random intercept model with random slope, because as seen in the previous exercise
## it has a better fit to the data than the random intercept model with a non-random slope
```

1.s) **What would be your next step in the analysis of this data set? Comment your suggestion.**
.

```
## We would investigate the presence of unusual data and its impact on the model, in addition to
## considering additional models that were not initially contemplated. We would conduct comparisons
## between these models using the previously mentioned criteria as a guide. Furthermore, when
## communicating the results, we would present our main conclusions, justifying the choice of the
## model, and highlighting its practical implications. This comprehensive approach would ensure a
## thorough and well-founded evaluation of the findings, facilitating an informed interpretation and
## decision-making.
```

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

```
## Examples:
## Mixed models: Tracking performance of students in different subjects over several semesters in
## multiple schools
## Longitudinal model:Predict cardiovascular disease risk
## Hierarchical model:  Examination on factors that influence health trajectories for individuals
## across age
## Cluster model: Identify groups of households that are similar to each other in retail spending.
## Repeated measurements model: Analysis of drug efficacy in patients over multiple treatments
```

1.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?** .

```
## You can extend a linear regression model to a generalized linear model using the glm() function.
## The library is the base R library
```