Lab assignment 2: linear models and mixed-effects models, sentence plausibility and selfies

Deadline: March 11

1 General lab information

This assignment consists of two parts. There are group assignments and individual assignments. Group assignments should be shown to your teacher and get graded directly *during the lab meeting*. Answers to individual questions should be submitted to Blackboard via a Blackboard quiz. If the question does not specify what type of question it is, it is a group question (there is only one individual question in this assignment).

There are 10 points in the assignment (2 points for question, 1 per question).

More details on the whole procedure:

You work in groups of 3 or 4 on these exercises, with help from a teacher/teaching assistant. We expect you to work on these exercises in class time so you can work with your group and teacher. It is not acceptable to miss these classes without agreement from your group, or to repeatedly miss classes. Then you will fail the assignment, which leads to failing the course. If your group members miss lab classes without agreement from your group, please inform your teacher.

We suggest all group members doing these exercises on their individual computers simultaneously: this improves (student) learning and also makes it easier to find mistakes. Don't rely on other group members' answers if you don't understand why they are correct: this is meant to be an interactive collaboration with your group, so ask your group members to explain. If your group gets stuck on a question or different group members can't agree on an answer, ask for help from your teacher. Please share your video if bandwidth and circumstances allow. This makes for a more personal conversation.

When your group is happy with your answer, work together to finalize your answer in a document shared with the whole group. Show these answers to your teacher as you work. You can share this document with the teacher too. Your teacher will grade you as you work to monitor your progress and address problems. But we need a record of all your answers, submitted at the end of the assignment (via Blackboard). At the end, you should submit one pdf file, which also includes your R calculations and code. You could either copy-paste your R code into the file, or, better, you could use an engine for dynamic report with R like knitr.[1] The assignment, including your answer to the individual question, has to be submitted on Blackboard by *Tuesday, 6pm, March 30*.

In group questions, it is generally best to start by asking every group member's opinion. Then work on a written answer together. Then explain your answer to your teacher. You can also ask your teacher to read what you wrote, but they will often ask questions. It is likely you will then have to update this answer after talking with your teacher. Please tell your teacher what changes you made next time you talk and show them what you wrote.

Many questions build on previous questions being completed correctly, so you should be confident of your answer before using it in further questions: ask for help if you are unsure. If you get stuck and the teacher can't get help immediately, you can move on to the next topic until your teacher can help.

[1] https://yihui.org/knitr/

2 Introduction

This assignment consists of two parts. The first part is based on the first experiment in Cunnings and Sturt (2018). Questions Q1 – Q4 belong to the first part. The second part (Q5 – Q9) is based on selfies data - a data collection from a study that investigated people's subjective reactions to selfies.

We recommend that you do the first part in the first lab meeting and the second part in the second lab meeting, since some of the concepts used in the second part are only explained in the next week lecture. It might be that you are done before the end of the lab the first or the second week. If so, it's good to use the extra time to think about your own experiment. Keep in mind that on March 10, you should submit a brief summary describing your idea of what you would like to run as an experimental study (in pairs).

3 What will you hand in?

You will hand in a pdf file with the analysis. The pdf file should include the code you used and the code should include all the steps, from loading the csv files up to the analysis required of you in questions. Aside from that, you have to respond to individual questions on Blackboard.

4 What can you use?

You can use R and any packages you find useful. Some packages are even recommended to use. You can (and should) reuse the code present in these assignment instructions. For an ease of reuse, we put the code separately into an R (knitr) file.

Part 1: Q1 – Q4

For the first part, you need this pdf, the pdf of the paper Cunnings and Sturt (2018) and the data from Exp1 from Cunnings and Sturt (2018), called cunnings_sturt_exp1.csv. You can also use the R code or the rnw file which includes all R snippets that we already created for you.

Q1: analyze the experiment (2 pts)

Go through the paper Cunnings and Sturt (2018). You can read the whole paper but since this assignment focuses only on the first experiment, it is enough if you only read pages 16–20.

The experiment (Experiment 1) described in the paper studies the effect of interference on memory retrieval. Imagine you are trying to replicate the experiment. For such a task, you need to gather all possible information about the setup of the experiment in the paper. Describe the experiment in enough details to make sure you would be able to replicate it as closely as possible. Find answers to at least these questions:

1. What is the design of the experiment? That is, how many factors are there? How many conditions? How was the data presented - in Latin square design? Were they randomized?
2. Are the stimuli available? If so, where?
3. How many participants took part and how many items were used?
4. What experimental method was used in this study (self-paced reading, acceptability study, fMRI...)? What kind of data (reaction times, ratings, EEG...) were gathered?
5. What statistical analyses were conducted? (Do not go into details here, name the statistical technique or tests used.)

Q2: t-test and linear regression

From now on, we will start working with the data. First, load useful packages and data and load relevant data. (We already did some basic data cleaning and subsetting for you.)

```
library(dplyr)
library(ggplot2)

exp1 <- read.table("cunnings_sturt_exp1.csv", col.names = c("subject", "item",
    "region", "measure", "condition", "rt"), sep = " ") %>% mutate(region = ifelse(region ==
    2, "verb", "spillover"), condition = case_when(condition == 1 ~ "a", condition ==
    2 ~ "b", condition == 3 ~ "c", condition == 4 ~ "d")) %>% mutate(plausibility = ifelse(condition %i
    c("a", "b"), "plausible", "implausible"), interference = ifelse(condition %in%
    c("a", "c"), "interf", "nointerf"))

glimpse(exp1)

## Rows: 9,210
## Columns: 8
## $ subject      <fct> exp1_1, exp1_1, exp1_1, exp1_1, exp1_1, exp1_1, ...
## $ item         <int> 27, 1, 22, 31, 12, 18, 16, 5, 13, 28, 19, 2, 9, ...
## $ region       <chr> "verb", "verb", "verb", "verb", "verb", "verb", ...
## $ measure      <fct> fp, fp, fp, fp, fp, fp, fp, fp, fp, fp, fp, fp, ...
## $ condition    <chr> "c", "a", "d", "c", "b", "d", "b", "a", "a", "b"...
## $ rt           <int> 388, 345, 318, 182, 319, 655, 179, 175, 140, 211...
## $ plausibility <chr> "implausible", "plausible", "implausible", "impl...
## $ interference <chr> "interf", "interf", "nointerf", "interf", "noint...
```

The data are in a so-called "long format" (each row shows one data point per subject) with 6 columns/variables and 9210 rows. The variables are described below.

- subject: a string denoting a participant
- item: an integer indicating a number of an item.
- region: either "verb" or "spillover". Identifies the region.
- measure: the measure of the reading time. "fp" for first-pass duration, "rp" for regression-path duration, "tt" for total viewing time
- condition: a letter ("a", "b", "c", or "d") indicating the condition based on the example 6 from the paper (p. 18). Copied below.

  The manor house was always very busy.

  (a) Plausible Sentence, Plausible Distractor
  Sue remembered the plate that the butler with the cup accidently shattered today in the dining room.

  (b) Plausible Sentence, Implausible Distractor
  Sue remembered the plate that the butler with the tie accidently shattered today in the dining room.

  (c) Implausible Sentence, Plausible Distractor
  Sue remembered the letter that the butler with the cup accidently shattered today in the dining room.

  (d) Implausible Sentence, Implausible Distractor
  Sue remembered the letter that the butler with the tie accidently shattered today in the dining room.

The owner of the house was not impressed.

- rt: reading times in ms.
- plausibility: plausible/implausible indicating whether we are in one of the plausible conditions (condition a or condition b) or implausible ones (condition c or condition d).
- interference: interf/nointerf indicating whether we are in one of the interference conditions (condition a or condition c) or no interference conditions (condition b or condition d)?

Use the dataset to address the following question: *is it so that early on (in first-pass duration) we can observe that readers are sensitive to the effect of plausibility?* Check this for two regions: the verb and the spillover. You will check it by two methods: using t-test and using linear models.

For the t-test: Make sure to develop the correct t-test. Keep in mind you should do some aggregations on the data and decide whether the test is paired or unpaired (it might also help to think about how many degrees of freedom the test should have).

For the linear model: Develop a linear model (Note: not a mixed-effects model yet). Try two versions of the linear model:

- In version 1, the dependent variable is reading times (so, no aggregation, no averaging, just plain reading times).
- In version 2, the dependent variable is subject-aggregated reading times.

Is the linear model justified for version 1 or version 2 or both? Why? Why not?

Finally, is there any worry about the fact that you test the same question twice: on the verb region and the spillover region? To answer this, it helps to check what Type 1 and Type 2 errors are and what Bonferroni correction is (you can check Wikipedia or other sources on these topics).

Q3: linear models and mixed-effects models

We will now proceed with linear models and mixed-effects models. As you can see in the paper, the goal is not to find out whether there is a plausibility effect but the plausibility * interference interaction, which would show the illusion of plausibility. One way to interpret this interaction: interference of a plausible element helps reading, but only in the implausible condition; in the plausible condition, the effect flips or is missing. You might want to consult the paper for more details on the interpretation of the interaction.

Now, create a mixed-effects models that tries to see whether there is a significant effect of plausibility and a significant effect of plausibility * interference interaction. Keep in mind that reading times are the dependent variable (no aggregating or averaging needed) but you might want to transform this reading measure beforehand (using log). If you run into problems with log-transformation because some values are 0, change those values from 0 to 1 (since log(1)=0; we will come back to this issue in the next question).

Create this mixed-effects model on (i) first pass reading times on the region Verb and the region Spillover, (ii) for total viewing times on the same two regions (so you should have 4 mixed-effects model in total). Make sure you include correct fixed effects and your model also includes the random effect structures for subjects and items (it might help to check slides from Monday). What results did you get? Do you find differences for the effect of plausibility when you compare the now-created mixed-effects models and the linear model from Q2?

Q4: null responses

There is one strange measure in reading times (both first pass and total viewing times that we did not discuss yet but maybe you already noticed it: some responses are zero. The zero is encoded when the subject skipped the region or when the eye tracker lost track of the eyes in the region (this happens, for example, when the person blinks while fixating a word).

Address this question: is it correct to code such a case as zero? Or should we rather treat it as missing data? Which option is more adequate? Why?

After answering this, try the following: recode all zeroes as missing data (you can use the special element NA in R for this). Then, re-run linear models and mixed-effects models on total viewing times, checking whether the effect of plausibility changes. To answer this question, you should use the linear model from Q2 that did not aggregate any data (so, reading times were the dependent variable).

Did the estimates of plausibility interaction change when you compare the models in this question to the models in Q2 and Q3? Was the change bigger for mixed-effects models or for linear models?

Part 2: Q5 – Q9

For the second part, you need this pdf and the selfies data selfies.csv and novel_data.csv. You can also use the R code or the rnw file which includes all R snippets that we already created for you.

4.1 Data preparation

We start the second part by loading useful packages (dplyr for data manipulation and ggplot2 for graphics) and loading data as data frames and checking the structure of the data frames.

```
library(dplyr)
library(ggplot2)

selfies <- read.csv("selfies.csv")
str(selfies)
```

```
## 'data.frame': 2154 obs. of  16 variables:
##  $ ResponseId  : Factor w/ 135 levels "R_00OlMTXsIar3zHj",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Dur         : int  1732 1732 1732 1732 1732 1732 1732 1732 1732 1732 ...
##  $ Age         : int  34 34 34 34 34 34 34 34 34 34 ...
##  $ Country     : Factor w/ 17 levels "America","ENGLAN",..: 15 15 15 15 15 15 15 15 15 15 ...
##  $ Gender      : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Socialmedia : Factor w/ 31 levels "Facebook","Facebook,Instagram",..: NA NA NA NA NA NA NA NA NA N
##  $ Selfietaking: Factor w/ 4 levels "At least once a day",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ StimGender  : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Tilt        : Factor w/ 2 levels "Level","Tilted": 2 2 1 1 2 2 1 1 2 2 ...
##  $ Distance    : Factor w/ 2 levels "Far","Near": 2 2 2 2 1 1 1 1 2 2 ...
##  $ Eyes        : Factor w/ 2 levels "Direct","Side": 1 2 2 1 1 2 1 2 1 2 ...
##  $ Boring      : int  4 3 4 4 4 3 1 4 1 2 ...
##  $ Funny       : int  2 4 3 1 2 2 4 1 5 4 ...
##  $ Ironic      : int  4 4 5 4 2 3 2 2 2 4 ...
##  $ Serious     : int  4 3 2 4 3 3 2 3 1 1 ...
##  $ Ugly        : int  3 2 2 2 1 2 1 1 1 1 ...
```

```
selfies %>% group_by(ResponseId) %>% summarise(n = n(), m = mean(Boring))
```

```
## # A tibble: 135 x 3
##    ResponseId            n     m
##  * <fct>             <int> <dbl>
##  1 R_00OlMTXsIar3zHj    16  2.88
##  2 R_0IJR3pgtfcdxe2B    14  4.29
##  3 R_10BoO25kkjsfeLo    16  3.44
##  4 R_123YVnrORavW4At    16  1
##  5 R_12M70a3TN80PMDU    15  3.2
##  6 R_1C3aEpbLrFrYrFm    16 NA
##  7 R_1CguUOhL7Dy9bYp    16  2.44
##  8 R_1dNdwKIhC7Fxjmw    16  2.56
```

```
##  9 R_1E1ex0ct9qgnoKc    13  3.15
## 10 R_1EcHZSIxAECyCbs    16  3.62
## # ... with 125 more rows

glimpse(selfies)

## Rows: 2,154
## Columns: 16
## $ ResponseId   <fct> R_00OlMTXsIar3zHj, R_00OlMTXsIar3zHj, R_00OlMTXs...
## $ Dur          <int> 1732, 1732, 1732, 1732, 1732, 1732, 1732, 1732, ...
## $ Age          <int> 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, 34, ...
## $ Country      <fct> usa, usa, usa, usa, usa, usa, usa, usa, usa, usa...
## $ Gender       <fct> Male, Male, Male, Male, Male, Male, Male, Male, ...
## $ Socialmedia  <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ Selfietaking <fct> Less than once a week, Less than once a week, Le...
## $ StimGender   <fct> Female, Female, Female, Female, Female, Female, ...
## $ Tilt         <fct> Tilted, Tilted, Level, Level, Tilted, Tilted, Le...
## $ Distance     <fct> Near, Near, Near, Near, Far, Far, Far, Far, Near...
## $ Eyes         <fct> Direct, Side, Side, Direct, Direct, Side, Direct...
## $ Boring       <int> 4, 3, 4, 4, 4, 3, 1, 4, 1, 2, 3, 1, 4, 2, 4, 2, ...
## $ Funny        <int> 2, 4, 3, 1, 2, 2, 4, 1, 5, 4, 2, 2, 1, 4, 3, 4, ...
## $ Ironic       <int> 4, 4, 5, 4, 2, 3, 2, 2, 2, 4, 4, 5, 3, 3, 4, 3, ...
## $ Serious      <int> 4, 3, 2, 4, 3, 3, 2, 3, 1, 1, 3, 4, 3, 1, 2, 2, ...
## $ Ugly         <int> 3, 2, 2, 2, 1, 2, 1, 1, 1, 1, 3, 2, 1, 2, 2, 1, ...

library(extraDistr)
```

The columns in those data:

- ResponseId: participant ID
- Dur: How long did the experiment take?
- Age, Country, Gender: Age, country and gender of each participant?
- Selfietaking: How often each participant participates takes selfies.
- StimGender: The gender of the person on the selfie.
- Tilt: Is the selfie tilted or not?
- Distance: Is the person on the selfie very close to the camera or not?
- Eyes: Does the person on the selfie stay close to or far away from the camera?
- Boring, Funny Ironic, Serious, Ugly: Is the selfie boring, funny, ironic, serious, ugly? A scale ranging from 1 to 5 (1-lowest, 5 highest)

Q5: Does the gender of the selfie-taker predict boringness responses?

We will work with Boring responses. Prepare your dataset and test using t-test whether male selfies are seen as less/more boring than female selfies. Make sure to develop the correct t-test (keep in mind you might have to do some transformations on the data and decide whether the test is paired/unpaired; it might also help to think about how many degrees of freedom the test should have).

Then, develop a mixed-effects model with logit link (i.e., a logistic mixed-effects model) to address the same question. First, since this model requires a yes-no outcome, create a new variable called BoringYesNo. Then, transform each *Boring* response as follows:

- Response in *Boring*: 1 or 2 ⇒ BoringYesNo: 0
- Response in *Boring*: 4 or 5 ⇒ BoringYesNo: 1
- Response in *Boring*: 3 ⇒ BoringYesNo: NA (not available, i.e., a missing data)

Then, build a model that tests whether StimGender is a significant predictor for BoringYesNo. Check also whether this model provides a better fit to data than the model without the StimGender predictor. Keep in mind that this should be a mixed-effects model. Try to use the maximal random-effects structure that converges but use only subjects as random factors.

Do you find similarities and differences in the model? Do both approaches give the same answer wrt the significance of StimGender?

Q6: Reasoning about the model

Criticize the t-test and logistic mixed-effects models that you just created. While they are (hopefully) right from the statistical perspective, would you conclude from them that general population finds male selfies significantly more/less boring than female selfies? Think about the following issues: data aggregation/transformation, confounds, balancing in the data. All these issues might make it dubious that one can conclude that male selfies are generally found signficantly more/less boring than female selfies.

Q7: Predictions of the model

We will now look at deterministic predictions of logistic mixed-effects models.

Suppose you are still interested in boringness of selfies based on the sex of the selfie-taker but you add one confound into the picture: the sex of the person that judges the selfie. You want to see whether females judge male selfies as more boring than female selfies and whether male judgements differ. You furthermore add subjects random effects and have the maximal random effect structure that converges. So, you are thinking of this model:

```
# BoringYesNo ~ 1 + StimGender*Gender + (1 + StimGender * Gender |
# ResponseId )
```

Or some simpler version in the random structure, if this one does not converge.

Create the mixed-effects model.

Second, check predictions of your model. What does the model predict on unseen data? Suppose you got the dataset novel_data. You want to see what the model that you just created predicts for those data. With what probability will be each selfie (each row) considered as ugly by a median subject? (Saying that you make predictions for a median subject is just another way to say that you can assume random effects are zero, i.e., you can ignore them.)

```
novel_selfies <- read.csv("novel_data.csv")
```

First, calculate for each combination of the values in novel_data what probability your model estimates.

Once you are confident you can do this, calculate probabilities for all the rows (400 data points).

As a final check, load the following function:

```
drawprobabilities <- function(probs) {

    if (length(probs) != 400) {
        print("Wrong length of the vector of calculated probabilities. Should be 400 data points.")
    } else {

        matrixprobs <- matrix(ifelse(probs > 0.5, "X", ""), nrow = 20)
```

```
        x <- rep(NA, 400)
        y <- rep(NA, 400)

        k <- 1

        for (i in 1:20) {
            for (j in 1:20) {
                if (matrixprobs[i, j] == "X") {
                  y[k] <- i
                  x[k] <- j
                  k <- k + 1
                }
            }
        }

        plot(x, y, xlim = c(0, 40), ylim = c(0, 40), pch = 15)

    }

}
```

Now, run the function *drawprobabilities* with the argument the vector of predicted probabilities (e.g, if you stored your vector of predicted probabilites for novel_data as mypredictions, you would call it as drawprobabilities(mypredictions)). If everything was correct, you should see a picture as a result. What picture do you see?

Q8: Ordinal model

We will now inspect the original ordinal responses in the data (no transformation). You should use the package *ordinal* here.

Try to establish whether male selfies are considered more boring than female selfies. However, unlike in Q5 – Q7, use the original non-transformed response and model it using the ordered probit link function. Furthermore, try to control for various confounds that might obscure the effect of boringness of male selfies. Don't go beyond the data provided here (i.e., you might think of various other confounds that might be affecting the results but as long as they were not collected, you can ignore them). Discuss what you found. Can you conclude that male selfies generally significantly differ from female ones wrt boringness? Or is the difference more restricted and driven by specific factors?

```
library(ordinal)

# model investigation here
```

Q9: Inspecting ordinal model (individual component)

Check a simple ordinal model with only one condition (StimGender) and intercept-only random effects per subjects, i.e., use this formula on the ordinal model (commented out right now):

```
# Boring ~ StimGender + (1|ResponseId)
```

Check the output of the model and answer the following three questions:

1. Are 1-5 responses selected equally likely?
2. Which of the values 1-5 does the model estimate to be the most likely response for the male StimGender?

3. It is sometimes suggested that in Likert scale, the middle response (i.e., 3) should be removed and scales should be even because otherwise people will predominantly go for the middle, non-committal response, and the results will be useless. Based on your findings, is this justified?

Hint: You will need to look at thresholds and translate those into probabilities on standardized normal distribution (i.e., normal distribution with mean 0 and st.d. 1). You will probably want to use pnorm. When you consider a condtion, you will have to move the mean. If you are lost, go back into the last slides of the last lecture, or check discussions of ordinal models in the last video and on Wikipedia.

References

Cunnings, Ian, and Patrick Sturt. 2018. Retrieval interference and semantic interpretation. *Journal of Memory and Language* 102:16–27.