EPsy 8282—Statistical Analysis of Longitudinal Data I
Lab #1

DATA MANIPULATION AND RESHAPING WITH R

You will use the same data set for all the labs. The data set is available on the course website as the text file `sleeplab.txt`. You should download this file to a convenient location as you will read it into `R` in this lab. In addition, the syntax presented below is available in the text file `Lab-1.R`.

The data are from a sleep deprivation study described in Belenky et al. (2003) [*Journal of Sleep Research, 12*, 1-12]. Eighteen participants were studied over a 10 day period. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. Each day the participants were given a series of reaction time tests and the average was computed. The response variable is the average reaction time in ms. `sleeplab.txt` is in wide format, and as such, there are 10 columns of reaction time with column labels, `Reaction.0`, `Reaction.1`, ..., `Reaction.9`. The other variables are `SubNum` (subject number), `female` (1 = female, 0 = male) and gpa. Two participants, 105 and 116, dropped out after day 6 and their missing data are coded with −999. It should also be noted the variable names appear at the top of the columns. [Please note: this data set is different than the `sleepstudy` data set that comes with `lme4`!]

Assignment Guidelines: After downloading the data, you should invoke `R` and open `Lab-1.R` as a script file. You should also invoke your word processor or document processor as you will need to copy and paste output from `R`. Be sure to regularly save both your script file and the file in your word/document processor. In each section you are directed to produce output, which you should copy from `R` and paste into your word/document processor. Please label the sections as indicated below and use the question numbering as indicated. All questions are worth 1 point except for those in “Reshape the Data”, which are worth 2 points. There is a total of 30 points possible.

Read in the Data

To read in the data you will use `read.table()` with the appropriate options. You must supply the file directory (file location) and file name, an indication that the headers of the columns contain the variable names, and the missing data code. Type the following code in your script file and run it. Note you can continue a command on the next line.

```
# Reading in data:
> sleep.wide <- read.table(file=file.choose(),
                          header=TRUE, na.strings="-999")
```

An indication you successfully read in the data is the absence of any error messages. That is, success is indicated by a new prompt line in `R` (i.e., >). If you did not successfully read in the data, then check your syntax and try again. Having successfully read in the data, type and run the following code. You can execute it line by line or as a section. Copy the output to a word file and save it under the title "Reading in the Data". Now run the code below and save the output under the same section heading.

Examine the Data

Run the commands from the "Examine the Data" section of the `Lab-1.R`script file. Answer the following questions based on the output you produced. You can label this as "Examine the data" in your write-up. Please keep the answers short.

1. What does this code do: `sleep.wide[ , c(1, 9:13)]`? Be specific.

2. If you wanted to display just the data for `Reaction.0`, based on the syntax above, provide three ways you could do this (provide the code).

3. What is the reaction time score for subject 103 on day 2? (Recall day is counted as $0,1,\ldots,9$.)

4. For the `str()` output, why is `SubNum` and `female` denoted with "int" and the remaining variables denoted with "num"?

5. For the `summary()` output, why is there an extra row of output for the variables starting with `Reaction.7`?

6. What is the score that cuts off the lower 25% for `Reaction.3`?

7. What is the score that cuts off the lower 75% for gpa?

8. What do the median and mean indicate about the symmetry of the distribution of `Reaction.4`?

Reshape the Data

Recall that linear mixed model (LMM) analysis requires the data to be in long format. Having read in the wide format data, you now need to create a long format data frame using `reshape()`. You must supply six pieces of information: (1) the quoted list of columns names consisting of the repeated measures, (2) the quoted ID variable, (3) the quoted stacked response variable name (to be made), (4) the quoted stacked time metric variable name (to be made), (5) the quoted new time metric variable name, (6) and the quoted direction of reshaping. To make the first one much easier, we read in the last ten column names from `sleep.wide` into an object that can be used in the reshaping. Write the following code in your script file and run it.

```
> mynames <- colnames(sleep.wide[ , 4:13])
> mynames
> sleep.long <- reshape(sleep.wide, varying=mynames, idvar="SubNum",
     v.names="Reaction", timevar="Days", times=0:9, direction="long")
```

If you successfully reshaped the data there should be a prompt and no error messages. If there are error messages, check your code and try again. Having reformatted the data, Run the commands from the "Reshape the Data" section of the `Lab-1.R` script file. Copy the output to your word processor under the heading "Reshape the Data". Answer the following questions based on the output you produced immediate above. You can label this as "Reshape the Data" in your write-up. Please keep the answers short.

9. What does this code do: `sleep.long$female.c <- ifelse(sleep.long$female == 1, "female", "male")`?
10. How is `sleep.long2` different than `sleep.long`?
11. What does this code do: `sleep.long2[is.na(sleep.long2$Reaction)==TRUE, ]`?
12. How is `sleep.long3` different than `sleep.long2`?
13. Why might one use `sleep.long3` rather than `sleep.long2`? Explain.

Descriptive Statistics and Graphs in Long Format

Descriptive statistics should always be computed to understand the variables in the data frame and the nature of the missing data. Graphs are also essential. Here you will focus on basic graphs. In the next lab you will use the `lattice` package that has more extensive graphing capabilities. To save a graph made in *RStudio*, activate the graph window and click the "Save as PDF" button. Then active your word processing program and paste the graph.

Run the commands from the "Descriptive Statistics" section of the `Lab-1.R` script file. Copy the output and the graph to your word processor under the heading "Descriptive Statistics". Note we panel the plots so you only have to save two graphs. That is, you should make all graphs following a `par()` statement before pasting into your word processor. Answer the following questions based on the output you produced. You can label this as "Descriptive Statistics" in your write-up. Please keep the answers short.

14. What does this code do: `table(sleep.long3$female.c[sleep.long3$Days == 0])`? Be specific.
15. What does this code do: `tapply(sleep.long3$Reaction,sleep.long3$Days,summary)`? Be specific.
16. Based on the boxplot for GPA, is its distribution symmetric? Explain why or why not.
17. Using the descriptors of "symmetric", "positively skewed", or "negatively skewed", indicate which best characterizes the distribution of reaction time at each day. E.g., Day 0 = __, Day 1 = __, etc.
18. Based on the longitudinal boxplots of reaction time, what happens to the variation of the distributions over time?
19. As days elapse, what happens to mean reaction time?
20. As days elapse, what happens to the standard deviation (SD) of reaction time?
21. Based on the LOWESS curve, what type of model might you consider in the analysis?

Utilities

There are a number of useful utility functions for saving and loading data frames, and listing and removing objects. Use the syntax below but replace `C:\\...\\` with the pathname where you want to save the *.Rdata file. In future labs, you can use `load()` to load this file (you do not have to use reshape again).

```
> save(sleep.wide, sleep.long3, file = "C:\\...\\sleeplab.Rdata")
> ls()
> rm(list = ls())
> ls()
> load(file = "C:\\...\\sleeplab.Rdata")
> ls()
```

Answer the following questions based on the output you produced immediately above. You can label this as "Utilities" in your write-up. Please keep the answers short.

22. What does this code do: `save(sleep.wide, sleep.long3, file = "C: sleeplab.Rdata")`?

23. What would this code do: `save(sleep.wide, file = "C:\\...\\sleeplab.Rdata")`?

24. Can the `save()` function save three data frames? (Yes/No)

25. What does this code do: `rm(list = ls())`?