EPsy 8282—Statistical Analysis of Longitudinal Data I
Lab #2

GRAPHING LONGITUDINAL DATA WITH ggplot2

You will use the `sleeplab.Rdata` data set you created in the first lab. Make a note of where this file resides on your hard drive. An Rdata file is loaded into `R` using the `load()` function. If you do not have this file, you must remake by following the steps in Lab 1. See Lab 1 for a complete description of the data set. The response variable is `Reaction`, which is the average reaction time for a number of cognitive tests. The time metric is `Days` $(0, 1, \ldots, 9)$, and the static predictors are `gpa` and two versions of sex, `female` (0 = male, 1 = female) and `female.c` ("male", "female"). In addition, the syntax presented below is available in the text file `Lab-2.R`.

Saving Graphs and Creating Image Files: This lab requires you to make multiple graphs. Similar to Lab 1, as you make a graph you can use RStudio to export the plot and paste it into your word processor. This is the recommended method for those new to `R`.

Another method is to include syntax in your script file that saves the graph for later insertion in your word processor. The syntax below shows an example of how to directly save a graph in a particular image format, in this case a PNG file for a Mac. In all cases, the pathname you enter for the `filename=` argument will need to be modified for your system. (Windows users will have to use the modified DOS pathname with two backslases, for example `C:\\Mine\\Courses\\8282\\Labs\\plot1.png`.)

```
> png(filename = "/Users/zief0002/Desktop/plot1.png")
> g1 <- ggplot(data = sleep.long3, aes(x=Days, y=Reaction)) +
            geom_point()
> print(g1)
> dev.off()
```

You can make other types of image files by replacing `png` with `jpeg`, `pdf`, or `tiff`. You must use the `print()` command as shown above in order to have the graph print to the file. In addition, you must execute the `dev.off()` function which causes the creation of the file. The disadvantage of this approach is that the graph will not be created in a pop-up window. Rather, it will be written directly to the image file. To create the graph in a pop-up window, run the `print()` line above by itself before the rest of the syntax. I leave to you to decide which approach to take.

Read in the Data: To load the Rdata file, type the following syntax in your script file and run it.

```
> # The following line must be tailored to your system:
> load(file = "/.../sleeplab.Rdata")
> ls()
```

`/.../` indicates you replace ... with the location of the file. For example,, on my computer the entire syntax would be `/Users/zief0002/Documents/EPsy8282/sleeplab.Rdata`. For Windows users, replace the pathname with `C:\\...\\`. Be sure to use double backslashes.

An indication you successfully loaded the file 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 file, then check your syntax and try again.

The `ls()` function shows the objects in the Rdata file, which are data frames. If you followed the directions in Lab 1, the long format data frame is named `sleep.long3`. If you renamed this data frame, then be sure to use the new name in the syntax below. It is a good idea to run `head(sleep.long3)` to see the top of the data set.

Assignment Guidelines: 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. There is a total of 39 points possible. There is no external R script for this lab. The R commands you will need are embedded in the lab itself.

Superimposed Individual Curves

Individual curves can be superimposed in the same graph. Run the syntax below and insert the graphs in your word processed document.

```
> library(ggplot2)
> theme_set(theme_bw())
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      geom_line() +
      opts(title = "Reaction Time by Day B1")

> ##
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      stat_smooth(method = "lm", se = FALSE) +
      opts(title = "Reaction Time by Day B2")
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Superimposed Plots Questions" in your write-up. Please keep the answers short.

1. What does this syntax do: `theme_set(theme_bw())`? Be specific.
2. What does this syntax do: `groups=SubNum`? `geom_point()`? `geom_line()`? Be specific.
3. What does this syntax do: `stat_smooth(method="lm", se=FALSE)`? Be specific.
4. Based on the first graph, how would you characterize the growth of the group as a whole? Briefly explain.
5. Based on the second graph, are there subgroups of subjects with different growth patterns? If so, what are these subgroups? Briefly explain.

Facet Plot of Individuals

It is often useful to create facet plots of individual curves. Run the syntax below to create two graphs and insert them in your word processed document.

```
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      geom_line() +
      facet_wrap(~SubNum) + opts(title = "Reaction Time by Day A1")
> ##
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      stat_smooth(method = "lm", se = FALSE) +
      facet_wrap(~SubNum, as.table = FALSE) +
      opts(title = "Reaction Time by Day A2")
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Facet Plots Questions" in your write-up. Please keep the answers short.

6. What does this syntax do: `facet_wrap(~SubNum)`? `as.table=FALSE`? Be specific.
7. Based on the first graph, what is the primary difference of subjects 105 and 116 from the remainder of the subjects?
8. Based on the first graph, is there variability in the observed change curves? Briefly explain.
9. Based on the first graph, is there *extensive* variability in initial levels? Briefly explain.
10. Based on the second graph, does a linear change curve seem to be adequate for all subjects? Briefly explain.
11. Based on the second graph, is there variability in the fitted change curves? Briefly explain.

Selecting and Graphing Subsets

When the number of subjects is large, it may be preferable to select a subset for graphing. Below you are asked to select the first half of the subjects based on their ID numbers, and select a random sample of four subjects. For grading purposes, you set the seed on the random number generator so that everyone will obtain the same random sample. In analyzing your own data, you would ordinarily not set the seed (more accurately, the seed is set to the system clock that constantly changes). Run the syntax below, and save the output and graphs to your word processor.

```
> # Obtain ID numbers.
> theIDs <- unique(sleep.long3$SubNum)
> theIDs
```

Based on the output immediately above, identify the ID number that cuts off the lower half of the sample using the `median()` function.

```
> # Find median.
> mymed <- median(theIDs)
> mymed
```

Now use the median value in the syntax below.

```
> # Select first half with index condition:
> ggplot(data = sleep.long3[sleep.long3$SubNum < mymed, ],
                aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      geom_line() +
      facet_wrap(~SubNum) +
      opts(title = "Reaction Time by Day, Half the Subjects")
```

Now you will select a subsample consisting of a random sample of size four.

```
> # Set seed of random number generator.
> set.seed(123)
> # Obtain random sample of size 4.
> mysub <- subset(sleep.long3, SubNum %in% sample(theIDs, 4))
```

Create a graph with the subsample.

```
> ggplot(data = mysub, aes(x = Days, y = Reaction, groups = SubNum)) +
      geom_point() +
      geom_line() +
      facet_wrap(~SubNum, nrow = 1) +
      opts(title = "Reaction Time by Day, Random Sample of Four Subjects")
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Subsets Questions" in your write-up. Please keep the answers short.

12. What does this sytax do: `subset(sleep.long3, SubNum %in% sample(theIDs, 4))`? Be specific.
13. Consider the first type of graph you produced. If you wanted to select the first two subjects and the last two subjects, what syntax for `subset=()` would you use?
14. Consider the second graph you produced. What is the advantage of this graph relative to the superimposed graph you made earlier with all the subjects?
15. Consider the second graph you produced. What is the disadvantage of this graph relative to the superimposed graph you made earlier with all the subjects?

Group-Level Graphs

Group-level plots are essential for an exploratory analysis as the form of the aggregate change curve is not known. Run the syntax below and save the graphs to your word processor.

```
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point() +
      stat_summary(fun.y = mean, geom = "line", lwd = 2)
> ##
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point() +
      stat_smooth(se = FALSE, lwd = 2)
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Group-Level Questions" in your write-up. Please keep the answers short.

16. What does this syntax do: `stat_summary(fun.y=mean, geom="line", lwd=2)`? Be specific.
17. What does this syntax do: `stat_smooth(se=FALSE, lwd=2)`? Be specific.
18. Based on the first graph you produced, describe the average trend.
19. Based on the second graph you produced, does the curve appear to be linear or nonlinear? Briefly explain.
20. Why is the curve for the second graph smoother than the curve for the first graph?

Graphs with a Static Categorical Predictor

Graphs can be made conditional on values of static predictors. Static predictors can be categorical or quantitative (continuous). The former are easier to work with because their values define subgroups. For our data set, we consider graphs based on sex groups with the variable `female.c`. Consider the syntax below and save the graphs to your word processed document.

```
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point(colour = "grey80") +
      stat_summary(fun.y = mean, aes(line = female.c), geom = "line",
        lwd = 1) +
      opts(legend.position = c(0.3, 0.9 ), legend.background = theme_rect ())
> ##
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point(colour = "grey80") +
      facet_grid(female.c ~ .) + stat_smooth(se = FALSE, lwd = 2)
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Static Categorical Questions" in your write-up. Please keep the answers short.

21. What does this syntax do: `aes(line=female.c)`? Be specific.
22. What does this syntax do: `colour="grey80"`?
23. What does this syntax do: `opts(legend.position = c(0.3, 0.9 ), legend.b ackground = theme_rect ())`.
24. What is the advantage in using `female.c` rather than `female` in constructing the graphs?
25. Based on either graph, briefly comment on the extent of the difference between the growth curves at the initial value (Day 0).
26. Based on the smoothed curves graph, briefly comment on any differences in the change curves.

Graphing with a Static Quantitative Predictor

Graphs based on quantitative predictors are more difficult than with categorical predictors as subgroups are usually not readily defined. For graphing purposes (only), we can form subgroups by discretizing a quantitative predictor. We form equal-interval and equal-count groups for gpa with the `cut_interval()` function and `cut_number()` function, respectively. We treat gpa as continuous even though it is has the limits [0,4]). For illustration, we form three groups, but any number of groups can be made.

```
> # Create gpa3A with three equal-interval groups:
> sleep.long3$gpa3A <- cut_interval(sleep.long3$gpa, n = 3)
> table(sleep.long3$gpa3A)
> levels(sleep.long3$gpa3A) <- c("Low", "Med", "High")
> # Create gpa3B with three equal-count groups:
> sleep.long3$gpa3B <- cut_number(sleep.long3$gpa, n = 3)
> table(sleep.long3$gpa3B)
> levels(sleep.long3$gpa3B) <- c("Low", "Med", "High")
> # Graph 1
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point(colour = "grey80") +
      stat_summary(fun.y = mean, aes(line = gpa3A), geom = "line",
        lwd = 1) +
      opts(legend.position = c(0.3, 0.8 ),
        legend.background = theme_rect ()) +
      opts(title="Equal-Length")
> # Graph 2
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
      geom_point(colour = "grey80") +
      facet_grid(. ~ gpa3B, margins = TRUE) +
      stat_summary(fun.y = mean, geom = "line", lwd = 1) +
      opts(title = "Equal-Count")
```

GPA

GPA

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Static Quantitative Questions" in your write-up. Please keep the answers short.

27. What is the `levels()` syntax used to do above? Be specific.
28. What does this syntax do: `facet_grid(.~gpa3B, margins=TRUE)`? Be specific.
29. In comparison to the first graph, why is there no `aes()` argument in the `stat_summary()` component for the second graph? Be specific.
30. What is the length of the intervals for `gpa3A`? Show your work.
31. What is a reason for the unequal number of subjects in the groups of `gpa3B`? (You do not need to do any further analysis unless you want to.)

32. In the first graph, what are the differences among the mean curves?

33. In the second graph (with three facets), which panel displays the group with the lowest limits of reaction time?

34. In the second graph, explain any differences between the group curves and the marginal curve (all).

Graphs of Interactions Among Static Predictors

If interactions among the static predictors are of interest, then plots based on the combinations of levels of the predictors should be constructed. Suppose we anticipate a GPA by sex interaction. First we check to see if there is sufficient data to examine the combination of the two static predictors. We consider `gpa3B` as our GPA variable. Run the syntax below and save the output and graphs.

```
> # Make contingency table.
> with(sleep.long3, table(female.c, gpa3B))
> # Make the graph.
> ggplot(data = sleep.long3, aes(x = Days, y = Reaction)) +
    geom_point(colour = "grey80") +
    facet_grid(female.c ~ gpa3B) +
    stat_summary(fun.y = mean, geom = "line", lwd = 1)
```

Answer the following questions based on the two graphs you produced immediately above. You can label this as "Section: Static Interaction Questions" in your write-up. Please keep the answers short.

35. What does this syntax do: `facet_grid(female.c~gpa3B)`? Be specific.

36. How many subjects have the combination of female/GPA High? How many with male/GPA Low?

37. What combination of GPA and sex is not represented in the data set?

38. Based on the graph, is there an appreciable difference in the change curves of the GPA groups within a sex category? Explain.

39. Based on the graph, is there an appreciable difference in the change curves of the sex groups within the medium GPA category? Explain.