Problem Set 3: Inference and Multilevel Modeling

ANONYMOUS

Overview

The data in this assignment are drawn from Making Caring Common, a project run by HGSE professors Rick Weissbourd and Stephanie Jones and aimed at bringing attention to issues of caring and morality in schools. 59 schools participated in this survey, but we only have data from some of these schools.

In this project, researchers gave surveys to students from 5th grade onwards in the selected schools.[1] Different schools in this project served different ranges of students. Some were pure high schools, some were middle and below, and some had the full range of grades surveyed.

This is a group assignment.

[1] A copy of the survey given to students that generated these data is available on the course Canvas site.

Variables

The variables we will be working with are below. The two primary measures of interest will be *esafe* and *disc*.

var	label
esafe	a scale measuring students' sense of emotional safety at school (safety from teasing and ostracism)
grade	a measure of students' grade level
gender	a variable indicating students' gender identity (you will restrict this to female and male due to the very low number of students indicating another gender identity)
disc	a measure of students sense of discrimination at the school

Substantive Goals

Our aim is to focus on students' sense of emotional safety, and to see to what degree a sense of discrimination predicts lack of safety, and finally to explore whether this relationship differs substantially by school.

What to turn in

Turn in a report as a pdf file made from, e.g., Microsoft Word or R Markdown.[2]

Your final pdf report should have any code and output from R along with your prose discussion of the results. The TFs will focus on the results and discussion. The code is there as a reference to help the TFs figure out where things may have gone wrong so they can give you useful feedback. Make this

[2] We have proved a markdown template for this assignment.

code neat and commented.[3]

Unless you submit a knitted R Markdown file, you should also turn in an R script that has all the code to load the data and run all your analyses and make all your plots, in a single file.

[3] See this class's style guide.

Before you start

This assignment is partially to help you get a sense of what data prep and cleaning looks like. To accomplish this, you will be walked through a series of tasks to prepare for tackling the actual work. In particular, please do the following in a separate R script (and *do not* turn in any results from doing the following):

1. Read through this assignment before starting so you have a sense of where you are going.
2. Remind yourself of the assignment guidelines.[4]
3. Load your data. You may notice a lot of variables that you do not necessarily need. We recommend code such as:[5]

[4] Guidelines posted on Canvas.

[5] See the R for DS textbook for more on this under chapter 5.

```
dat = dplyr::select( dat, "var1", "var2" )
```

4. Clean up your gender variable. Try tabulating the gender variable. You can then drop students with rare[6] or missing gender with the `filter()` command:

```
dat = filter( dat, !is.na( gender ) )
```

and/or

```
dat = filter( dat, gender != "something" )
```

and/or

```
dat = filter( dat, gender %in% c( "something", "else" ) )
```

[6] There is some real tension as to what to do with very small subgroups in datasets. Nonbinary gender or Native American are two common such instances of such small subgroups in most datasets. Some advocate that for visability, all groups should be kept in a final analysis; this is generally a reasonable option, even if any estimates for such groups will tend to be too noisy to draw direct inference. In this assignment, you are welcome to maintain these other categories, but please focus on testing for self-identified male vs. self-identified female in the gender gap questions below. (You can try to test other pairwise comparisons; I would be curious to see if you find anything.)

5. Convert the *grade* variable from character to numeric. You can do this using the code below. Currently, the values of the *grade* variable are stored as "5th", "6th",..., "12th." Instead we want them to be stored as 5, 6, ..., 12. We'll do this in two steps: first, we remove the "th" character from the values of the *grade* variable and second, we convert the variable from character to numeric. The first step is necessary because in order to convert a character to numeric, we have to remove all non-numeric values from the variable. This is a taste of *data wrangling*, a sadly necessary component of any analysis.

Note: the `str_replace()` function is a useful function for modifying character variables. Basically, it's going through each observation's value for *grade*, searching for "th" and replacing it with nothing (hence the quote with no space).

```
dat$grade = str_replace( dat$grade, "th", "" ) # replace "th" with ""
dat$grade = as.numeric( dat$grade ) # convert from char to num
```

6. Explore the data. A good first step when working with a new dataset is to examine the individual variables to see what they look like. In particular, make tables and/or plots for *esafe*, *disc*, and *gender*.[7]
7. You might also plot the association between *esafe*, and *disc*, ignoring the clustering of students within school. Use `ggplot` to add smoothers.[8] Try showing the association separately for students identifying themselves as girls and boys.[9]
8. Generate a clean dataset. You will probably want to generate a dataset with no missing values. Use the filter command, listed above, to do this. You can also try the `na.omit()` command (but be sure to select only those columns you want in your data before you use this). How much data did you lose?

[7] Try using the `skimr` package, and try running `skim( dat )`. But be warned: R Markdown can crash when trying to render the output of this package, so use for exploration but not final reports!

[8] Use `geom_smooth()`

[9] In your `aes` (aesthetics) argument of `ggplot` you can say `col = gender` and it will color points by gender and also fit seperate smoothers. You can also use `facet_wrap( ~ gender )` to get two plots, one for each gender.

Again, do all these explorations in a simple R script rather than the main markdown report. At the end, save your cleaned data with

```
saveRDS( dat, file="mycleandata.rds" )
```

Then in your project rmd file you can load your cleaned data with

```
dat = readRDS( "mycleandata.rds" )
```

to directly load your cleaned and saved data.

A tip about writing results

We are asking you to do formal inference, i.e., conduct tests of significance and report confidence intervals as appropriate. Do not worry about model fitting issues unless explicitly asked. There is no need to try and report and/or discuss R^2 measures or similar measures of fit.

The Problems

1. Present a favorite plot or display from your work[10]

You can use a plot from your data exploration, above, or from some other question of interest you may have. Your presentation does not need to be in any particular style, but should be easy to read and understand. Make sure your plot has nice labels for the axes, a title, and a caption describing what we are seeing and why it is important. Your groups and legends should be clearly labeled as well.

Also provide a sentence or two justifying why you chose your plot and not another (e.g., why a scatterplot and not a histogram). You should *not* feel compelled to use strange or esoteric plots for this question!

[10] We highly recommend doing this question *after* the rest of the project, since then you will know by experience what was worth focusing on.

2. Does sense of emotional safety vary by gender and grade?

Plan and fit a model to assess whether sense of emotional safety varies by gender or grade. For this question, be sure to include the mathematical notation for your model as well the R code. Generate confidence intervals so you can assess whether your gender differences are statistically significant.

Are there gender differences and are they significant? For grade, think carefully about how to describe any changes, but do not worry about significance testing for this part (we will learn how to do this later).

What to turn in:

- Your model in mathematical form.
- The R code to fit the model.
- The summary of the model output.
- The confidence intervals for your parameters.
- Your discussion and interpretation.

3. Do those who feel discriminated against feel less safe?

Does sense of discrimination (`disc`) predict a student's sense of emotional safety (`esafe`)? Generate and write out a multilevel model (both in math and in R) to explore this question. Make sure this model allows for each school to have its own slope. Be sure to standardize both your discrimination covariate and your emotional safety covariate, if they are not already.

Given your model, interpret your results as you see fit. Be sure to test any identified relationships of interest for statistical significance, and report confidence intervals along with any point estimates.

What to turn in:

- Your model in mathematical form.
- The lmer command to fit the model.
- The summary of the model output.
- The code to do whatever statistical inference you need.
- Your discussion and interpretation, as outlined above.

4. Contextual effects for discrimination?

Update or augment your prior work as you see fit to allow for answering the following questions:

a) Is there a contextual effect for discrimination with regard to emotional safety?

b) Once you take out the contextual effect, if any, does the relationship at the student level of sense of discrimination and emotional safety change?

Once you have answered this, reflect on the following:

c) What, if anything, does it add substantively to do this within-school and contextual effect analysis?

What to turn in:

- The lmer command to fit the model (math not needed here).
- A printed summary of the model output.
- Your discussion and interpretation of (a), (b), and (c).