Psych 205
Assignment 8
Due April 14th at 9am

In this dataset, we are going to use logistic regression on a dataset of about kids number learning in the US vs. in an indigenous group in Bolivia, the Tsimane'. This is some data that we (and others) have gathered on early math learning (please do not share this data or post it online). The main outcome measure we will use is when children become "CP-knowers", which roughly means that they can pass a simple counting task. There is a column in the data called "KL" for "Knower-Level" and we will be using whether KL is "CP-knower" as a binary outcome. First, load the data and check it. Make a new, binary (0/1) column for whether kids are CP-knowers or not.

Q1. [5pts, SOLO] To visualize this data we'll bin age. The binning on age is only for visualization and will allow you to put a single point with error bars for each bin. List advantages and disadvantages of binning by (a) year, (b) month, (c) quantiles (e.g. 10% quantiles).

Q2. [15pts, SOLO] (a) Choose the best option for binning and plot the results. Make this a "publication quality" plot: include error bars, points for the mean, axis labels, and faceting by language. (b) Are the best error bars from mean_se or mean_cl_boot? (c) Finally, include in this plot a stat_smooth using a glm (NOT a lm). You may need to read online to see how to do this. Ensure that the plot goes from 0 to 12 and that the stat summary line extends through this range. (d) What does the plot show? What basic patterns should you expect to see in a logistic regression based on this plot?

Q3. [15pts, HELP] (a) Run a logistic regression Language*Age (don't use binned age in the regression) and write up the results as you would for a paper. (b) When you do this, write a short description of the figure like you might find in a paper ("Figure 1 shows…") and then present/interpret the regression results. (c) Is there a different age slope in the two languages? (d) Is there a different intercept? What might those mean?

Q4. [10pts, HELP] (a) Write 2-3 sentence as you might in a paper explaining the size of the Age coefficient (b) Using the regression coefficients, figure out the point at which Tsimane kids are 75% CP-knowers vs. US kids and report this age. (c) What percent of Tsimane newborns are expected to be CP-knowers according to the regression model, and does this number make sense? (d) At what age will exactly 100% of Tsimane kids be CP-knowers and does this number make sense?

Q5. [20pts, HELP] Next, we are going to use a new coding scheme. We talked about "dummy coding" in class where the regression coefficients for a discrete factor (like Language) are primarily in one "baseline" level. This means, for example, that the Age slope in the previous regression is specifically in the baseline condition. However, we can also set up a regression where the Age slope is the average of both conditions. In this case, the Language effect will be added to one level of language and subtracted from the other (same for the interaction). To do this, we can simply give the glm (or a lm) the additional argument *contrasts=list(Language=contr.sum)*. This tells it to use "sum coding" (contr.sum) for the Language condition instead of dummy coding. (a) Print a summary of both this and the dummy coded regression. (b) When you do this coding, are the coefficients added or subtracted for English (you should be able to tell by comparing the this to the previous regression)? (c) When you run this new regression, what happens to the age slope relative to the previous regression and why? (d) What happens to the language effect and interaction and why? (e) What happens to the intercept and why?

Q6. [10pts, SOLO] (a) Make the same plot as Q2 for Tsimane only but, now with grouping and color by Gender. (b) What do you see? What would you expect to see in a logistic regression? (c) Run the regression and report the results as you would in a paper, also talking through the figure "Figure 2 shows…"

Q7. [10pts, HELP] Your collaborator keeps asking why you aren't running a linear regression, and what the relationship is between what you're doing an linear regression – after all, they use the same model formula in the glm call. Write a friendly email explaining logistic regression to them, and include argument on how it is different, and why it is more appropriate for this dataset.