DUE: Wednesday, May 10, 11:59PM, BbLearn Groups (48 points scaled to 100).

You may discuss this assignment with whomever you wish, but please prepare and submit work in groups of ***ONE to THREE*** students, no more and no fewer. ***Each group*** will submit a single copy of their group's completed assignment, via BbLearn, including the ***names of all group members*** who participated on the assignment. All team members will receive the same score, which will be recorded in BbLearn. Again, please ensure that all team members' names are on your submitted assignment!

GROUP MEMBERS WHO DO NOT CONTRIBUTE SUBSTANTIALLY TO AN ASSIGNMENT MAY BE REQUIRED TO WORK IN THEIR OWN GROUP OF ONE FOR THE REMAINDER OF THE SEMESTER.

While you are permitted to discuss the assignment with other groups, please prepare your own group's code and written answers. GROUPS WHOSE CODE AND SOLUTIONS APPEAR SUBSTANTIALLY SIMILAR MAY BE SUBJECT TO A 20% PENALTY.

Please prepare solutions in a ***neat, organized fashion***. I prefer typeset presentations (e.g., cut and paste code/output into MS Word with added exposition when appropriate; knitr via EMACS and ESS, knitr or R Markdown via RStudio, etc.)—probably most appropriate for presenting (fixed width font) code/output, at least—but neatly handwritten presentations may also be appropriate for some problems. Sloppily prepared solutions will not receive full credit. IN ANY CASE, PLEASE SUBMIT ONLY ONE FILE with all material, including code, results, exposition, and any hand-written solutions. (You may submit multiple times in BbLearn, but only the last submitted file will be graded.)

To complete the items below, I expect you to find and use our lecture notes, including code/results, possibly after some modification. I don't intend that you consult resources beyond our notes, though I may at times ask questions that go a bit beyond our notes. Some exposition may be required beyond code and output.

For this homework, a generalized linear mixed model (GLMM) and GLM analysis of a subset of data from a longitudinal study of the health effects of air pollution, the so-called 'six-city study'. The data are in `faraway::ohio`. From `help(faraway::ohio)`, the data are comprised of $n = 2148$ measurements of obtained over four years from $m = 516$ children of ages 7 to 10. The response indicates whether a child wheezed (1=yes) or not (0=no). The variables are summarized below.

- `resp` 1 = wheezing, 0 = no wheezing
- `id` child identifier
- `age` age (7 years = -2, 8 yrs = -1, 9 yrs = 0, 10 yrs = 1)
- `smoke` indicator of maternal smoking at the first year of the study (1 = smoker, 0 = non-smoker)

1. Retrieve the data and use `base::factor` to convert the response to a factor with levels of 0 and 1 and labels "no" and "yes." Similarly, convert the smoking status variable to a factor with labels of "no" and "yes". Be sure to use the newly converted data for subsequent items when appropriate. (4 points)

2. Use `stats::ftable` to create a table of the number of children for the various combinations of age, maternal smoking status and wheezing status. No exposition required. (2 points)

3. Use `base::tapply` to create a table of the proportion of children wheezing by age and maternal smoking status. Round your proportions to 3 decimals, please. No exposition required. (3 points)

4. Create a plot of the proportion of children wheezing by age, with separate points/line segments for maternal smoking status. Be sure to annotate your plot appropriately, including a legend. It would be appropriate to indicate actual age rather than coded age on the plot, of course! No exposition required. (5 points)

5. Use `lme4::glmer` to fit an appropriate generalized linear mixed model accounting for the fixed effects of age, maternal smoking status and their interaction. Use a logit link. Include a random intercept for each child. Also, use `confint` with options `method=''boot''` and `parm=''theta_''` (note the underscore) to compute a 95% bootstrap confidence interval for the standard deviation of the random intercept. (This may take a few minutes, and you may get a few warnings, which you may ignore.) From the interval, does it appear that the random intercept is necessary? Report your code and output summarizing your fit, and clearly indicate the value of estimated standard deviation of the random intercept and its interval, with brief discussion. (See the very brief introduction to `lme4::glmer` in §8.2 of our notes. Be sure to specify the correct family!) (8 points)

6. Fit a richer model with both a random intercept and random slope for each child and use `stats::anova` to test if the richer model is necessary. Discuss briefly. (5 points)

7. Regardless of your conclusions in the previous item, continue with the model with random intercepts but without random slopes. Apply the `predict` function to your fitted model to obtain predicted (estimated) probabilities of wheezing at the eight combinations formed from the four age values in the data set, (-2, -1, 0, 1), and the two maternal smoking levels for eight predicted probabilities in total. You will need to create a new data frame with these eight new **x** values at which you would like to predict. (We've done this sort of thing in a previous homework.) We want to explore using only the fixed effects estimates for this prediction (so-called level 0 or population values when we talked of this sort of population prediction for linear mixed models); for this, use `re.form = NA` as one of the arguments to the `predict` function. Show your code and the eight model estimated probabilities of wheezing. No exposition necessary. (5 points)

8. Use `stats::glm` to fit a corresponding model without random effects—same fixed effects. Again, use `predict` to compute the eight model estimated probabilities of wheezing as in the previous item. Plot the eight observed proportions vs the corresponding eight model estimated probabilities from both your GLMM and your GLM—2 observed vs predicted plots, one for GLMM and one for GLM, on the same figure. You should have obtained the set of eight proportions and set of eight GLMM estimated probabilities already, above. Annotate your plot appropriately. Which model probabilities do you trust? Why or why not? Discuss briefly. (Hint: See §5.8 of our notes and similar discussion in [Wak13, Chapter 9]; comments about non-linear mixed models (NLMM) apply to generalized linear mixed models (GLMM). I don't believe I lectured on this material; it's not recorded.) (8 points)

9. Regardless of any conclusions drawn from the previous item, use the GLM model to infer about the effects of smoking on child probability of wheezing. Be sure to qualify your conclusions based on what you learned above about the covariance structure among measurements within each child. (8 points)

Bibliography

[Wak13] Jon Wakefield. *Bayesian and Frequentist Regression Methods*. Springer, New York, 2013.