Data Analysis Prelim Exam

Department of Statistics, University of Chicago

Assigned Wednesday Sept 15 2021 at 4pm; Due Friday Sept 17 2021 at 4pm (Chicago time)

1 Data

The data set for this exam is taken from the paper

> *The effect of publishing peer review reports on referee behavior in five scholarly journals*, Bravo et al, *Nature Communications* 10.1: 1–8 (2019)

The data set consists of data from the peer review process for papers submitted to academic journals. When a paper is submitted, the journal editor sends it to potential reviewers. Each reviewer can choose to accept or decline the invitation to review the paper. If the invitation is accepted, then they need to (1) write a review (which can be any length) and (2) choose a recommendation for what the journal should do with the paper (accept / request minor revisions / request major revisions / reject).

Typically, reviewers are anonymous in the peer review process (i.e., the author will not see the names of the reviewers). However, the journals in this study implemented an "open review" policy several years ago, meaning that reviewers can choose to attach their names to the review. The goal of this paper is to examine changes in reviewer behavior that resulted from this change in policy. The data set contains data from years before and after this change was implemented.

The data and analysis scripts from the paper were downloaded from `https://www.nature.com/articles/s41467-018-08250-2#Sec9`. The data set contains the following variables:

- `id` and `journal` are unique identifiers for the paper and for the journal it was submitted to.
- `invitation.date` and `year` indicate the date/year that the reviewer was invited to review the paper.
- `open.review` indicates whether the journal is offering an open review option at the time of this paper.
- `review.complete` indicates whether the reviewer submitted the review.
- `name.published` indicates whether the reviewer chose to publish their name.
- `recommendation` is what the reviewer recommended for the paper: `Accept`, `Minor revisions`, `Major revisions`, or `Reject`
- `accepted` indicates whether the reviewer accepted the invitation to review the paper (note: this does not mean that the reviewer recommends acceptance of the paper).
- `review.time` is the number of days between when the reviewer was invited to review, and when the review was submitted.
- `polarity` and `subjectivity` are variables computed via natural language processing. `polarity` takes values in $[-1, 1]$, where positive and negative values indicate positive and negative sentiments (e.g., "great" or "terrible"). `subjectivity` takes values in $[0, 1]$, with larger values indicating an opinion (subjective) while smaller values indicate factual information (objective).
- `nchar` is the length of the submitted review (# of characters).
- `reviewer.status` takes values `Professor`, `Dr`, and `other`, recording whether the reviewer is a professor/faculty, or they have their PhD but are not a professor/faculty, or they do not have a PhD.
- `gender` is the gender of the reviewer. This information is not provided by the reviewer, but was imputed based on the name of the reviewer.

The first section of the R script included with the paper, contains code to preprocess the data by converting numerical codes into these named factor levels.

2 Questions

Your report should include code for all plots and results that you present. The report should be *concise, well-organized, and easy to read.* For example, if you examine plots across a long list of different subsets of the data, you may choose to show only a small selection of representative plots.

1. Please read the original paper (including the supplementary information appendix) and briefly summarize the design and goals of this study. Include a high-level description of the data collection, the modeling approach used, and the main conclusions drawn by the authors. Highlight some of the main statistical issues, as you see them, that arose during the analysis.

2. Attempt to reproduce Figure 2 of the paper. (Note: it suffices to reproduce the content of the figure; it is not necessary to match it in all aspects of style etc). Based on visual inspection alone, comment on whether the degree of smoothing provided by the authors' Loess lines appears appropriate.

3. In Table 1 of the paper, the authors used a logistic regression model with interactions to examine the effects of the open review policy on the acceptance probability of review invitations. An alternative approach is to run a logistic regression on each of the 9 subgroups separately (3 status levels * 3 gender categories). For simplicity, in this question let's omit the Year variable and the random effects terms of journal and submission in both approaches.

   Can we find a regression model with interactions that has the same model assumptions as a set of simple logistic regression models for each of the 9 subgroups seperately? If yes, will the estimates and confidence intervals of the open review effect on each subgroup be different from the two approaches? Provide an analytical justification and also check your conclusions numerically.

4. Answer the same questions as in Question 3 for the cumulative-logit model in Table 2 of the paper. The response in Table 2 is recommendation, which is treated as an ordered categorical variable in the paper; here you are to compare the cumulative-logit model with interactions with the cumulative-logit model on each subgroup separately. As in Question 3, ignore the Year variable and the random effects terms of journal and submission in both approaches.

5. As the open review policy is not randomized, the open review effect is confounded with year/time. The paper adjusts for the confounding year effect by adding a linear fixed effect term of year in their regression models. Assuming that the year effect is linear can be a strong assumption. For instance, Figure 2 clearly suggests that the Year effect could be non-linear.

   In this question let's use only the data on 3 journals – Journals 1, 3, and 5 – from years 2010 - 2014 (before the pilot study starts for Journal 1/5). We focus on estimating the policy effect on review time (days) for Journal 3. Instead of assuming a shared linear effect of year as in Table 3, we assume that the Year effect (mean review time differences across years, after controlling for all other variables) is the same for all 3 journals. Perform an analysis to estimate the average effect (averaged across the reviewers who have accepted and completed the review) of the open review policy on the review time for Journal 3 after adjusting for Year and test whether the average effect is 0 or not.

6. In this question we will examine how the probability that a potential reviewer accepts the review invitation varies among papers in each journal. For simplicity we assume that the probability that an invited reviewer accepts to review paper j, denoted p_j, is a property of the paper (and the journal it was submitted to), but not dependent on reviewer characteristics. You may model the reviewer acceptance/non-acceptance data for each paper j using either a binomial or negative binomial model, with success probability p_j. (The negative binomial model may be more natural because most journals have a target number of reviews for each paper – say two or three – and keep inviting reviewers until the target number of reviews are accepted; however in practice the data collection procedure is always a bit messier than this, and you may alternatively choose to use a binomial model for simplicity, treating the number of reviewers invited to review each paper as fixed in advance.)

   For each journal, assess the variation in p_j across papers. For which journal(s) is there strong evidence that p_j is not constant across papers? Which journals appear to have greatest variability in p_j? Using an Empirical Bayes approach, or otherwise, obtain an approximate posterior mean and 90% credible interval for each p_j. Compare the posterior mean estimates with the maximum likelihood estimates.