Final Exam – QPM 2024

Jacob M. Montgomery

Instructions

You have twelve (12) hours to answer these questions.

- You should attach a document that includes your answer including all figures or tables. Formatting is not particularly important, but be sure you provide enough information in the figures and/or text for us to be able to grade it clearly.
- You should also turn in an R script WITH COMMENTS that shows how your analysis was conducted.

Please note that this is an open book, open notes, open Internet exam. You can use any resource you want BUT DO NOT COPY CODE FROM OTHER SOURCES WITHOUT ATTRIBUTION. Stay far, far away from academic integrity violations. When in doubt – cite/give credit to the original source. Further, DO NOT DISCUSS THIS EXAM WITH ANYONE ELSE EITHER DURING OR AFTER THE EXAM.

Finally, keep an eye on the clock and make sure you get to all the parts parts before you start adding bells and whistles to your analysis.

Introduction: Dataset 1

We will begin by analyzing the data for the article "Are Mechanical Turk worker samples representative of health status and health behaviors in the U.S.?" published in *PlosOne*.

- You can read more about Mechanical Turk here: https://www.mturk.com/
- Here is an article that talks about how it is used in social science research: https://daily.jstor.org/amazons-mechanical-turk-has-reinvented-research/

Imagine we are interested in understanding how using "master" workers on Mechanical Turk affects our studies of health outcomes. For background, master workers are those who are viewed as doing high quality work and have a record of excellent performance.

Here is a short codebook of the variables in the Mechanical Turk dataset.

- *TurkMaster*: Coded 0 if they are not a master worker and 1 if they are a master worker.
- *bmi*: Body mass index (https://en.wikipedia.org/wiki/Body_mass_index)
- *exercise*: 1 = Excercise regularly; 0 = Does not exercise regulsalry
- *agecat*: 1= < 30; 2=30-40; 3=40-50; 4=50+
- *race_cat*: 1=White; 2=Black; 3=Hispanic/Latino; 4=Asian; 5=Other
- *income_cat*: 1= less than 35k; 2=35-50k; 3=50-75k; 4=75k+; 5=Other
- *female*: 0 = Female responden; 1 = Male respondent

1. First let's use BMI as our outcome variable of interest. We are interested in understanding if master workers are different than other workers when it comes to BMI. Fit a standard (classical) linear regression model to this data. You can go ahead and just control for all of the covariates (except `exercise`).
    a. Interpret the results. We are particularly interested in the results for the `TurkMaster`. Be sure to provide a correct interpretation for the p-values/stars. What conclusions do you draw about whether master workers are different when it comes to BMI?
    b. Thinking about this data, discuss two ways that the inferences you draw from this standard regression might be invalid. Focus on the kinds of assumptions we need to make to do inference in

a classical/frequentist setting.

2. In problem 1, you made use of the Null Hypothesis Significance Testing framework. Answer the questions below focusing again on the `TurkMaster` variable.
   a. Define the null and alternative hypothesis for this test.
   b. Explain what the p-value is and what it means in the context of this regression.
   c. What is the type-I error rate of the test you used to determine whether or not the variable was "significant."
   d. The significance test here is calculated simply by taking the coefficient and dividing by the standard error to derive a t-statistic with $n - p$ degrees of freedom. Assume that the standard errors are fixed and plot the "power curve" for different values of β. Is your test well powered? Why or why not?
3. Your co-author theorizes that the differences in BMI could be a result of differing lifestyles. Perhaps master workers are more (or less) likely to exercise.
   a. Fit an appropriate generalized linear model using `exercise` as the dependent variable.
   b. Write one paragraph interpreting these results to the readers of a hypothetical article. Explain what the "effect size" is. Provide a plot to help the reader understand what your model says.

Introduction: Dataset 2

For the questions below, you will be working on data from the paper, "Do Gladiators Fight for Their Masters?" This has something to do with voting on the United Nations Security Council, but you don't need to know much about the substance.

What you do need to know, is that the data covers the period 1946-1965, includes 46 countries, and (most importantly) that most countries appear in the dataset more than once.

Here is a short codebook of the variables in the UNSC dataset.

- *DependentVariable*: Percent of the time you vote against the us in the United Nations Security Council (UNSC).
- *Pro_Campaign* : Indicator for whether the US supported this member becoming a member of the UNSC.
- *l_pctaagreeus_unga*: Lagged percent ideological agreemenent with the USA.
- *wdicode*: Country code
- *year*: Year

4. Using the `lm_robust` function, fit a model that (a) includes fixed effects for each country *and* separate fixed effects for each year and (b) has clustered standard errors at the country level.[1] The only other covariate you need is `l_pctagreeus_unga`.

5. Now fit a Bayesian linear model with random intercepts (not slopes) for each country and year. Be sure to carry out appropriate diagnostics. Compare your results to those above. (Note: You may end up with more countries in this analysis than in problem 4. Countries that only appear in the dataset once may be dropped from the fixed effects model depending on how you implemented it.)

[1] Note that `lm_robust` has a nice syntax for putting in fixed effects easily. See an example on page 34 of the manual: https://cran.r-project.org/web/packages/estimatr/estimatr.pdf. If there is any confusion, ping me. But you are just adding a "dummy" for each respondent. You might try the `fastDummies` package: https://cran.r-project.org/web/packages/fastDummies/vignettes/making-dummy-variables.html