DRP Write-Up
Suh Young Choi
Mentors: Eric Morenz and Yiqun Chen

My DRP topic this quarter was “See What’s Not There” and dealt with the topic of identification and handling of missing data. We read from *Foundations of Agnostic Statistics* by Peter M. Aronow and Benjamin T. Miller (FAS) [1], with a few supplementary topics from Jeff Leek’s Advanced Data Science course [2].

Our first few meetings were spent talking about missing data problems in the wild. We talked about three kinds of missing data: missing completely at random (MCAR), in which the causes of missingness are unrelated to either the observed variables or patterns of missingness in the data; missing at random (MAR), in which the causes of missingness can be fully accounted for by observed variables; and missing not at random (MNAR), which is pretty much just every other possible kind of missing data. It was noted that we work under the MAR assumptions most of the time. MCAR is often unrealistic to assume in any applied setting, and MNAR data may not have sufficient information about missingness to be handled correctly.

We also discussed methods of deriving estimators when the data are missing. We discussed the methods in the context of the mean, but noted that they can be extended to other summary values of interest, such as the median. The simplest method is the naive estimator, which computes the mean from only the observed data and uses it to impute any missing values. Other estimators, such as the inverse-probability weighted (IPW), stabilized IPW, and doubly-robust estimators, require more calculation but give more nuanced results. For example, the IPW calculation includes weighting values according to their probabilities of observation, so that values less likely to be observed are accordingly downweighted in the estimation.

In the last four weeks, I developed and worked on a data project to demonstrate what I’ve learned about identifying and handling missing data. I found a dataset about books from GoodReads which had almost no missing values at all [3]. Therefore, I decided to simulate missing data according to the MAR assumptions on a variable of interest. I used the number of ratings on a book as my variable of interest and wanted to investigate its relationship with the number of pages each book has. I simulated the MAR data by leaving out books published before 1990. I then estimated the mean number of ratings on the MAR data according to the naive, regression-based, IPW, and doubly-robust estimators. Following these, I compared (i) the mean number of ratings; and (ii) the relationship between the number of ratings and number of pages, according to each estimator.

It turned out that introducing “missing” values of books published before 1990 removed many lower rating counts. The estimated means were more influenced by higher rating counts, particularly in the early 2000’s, which in turn led to overestimated means in the MAR data. For example, the original mean number of ratings was about 10,862 per book. The naive estimator produced a value of about 11,460, while the regression, IPW, and doubly-robust each gave about 11,462. In the examples we’ve seen from FAS, the latter three gave values that each became closer to the true mean. However, this was not the case for my data. One possible reason is that our simulated setting violates the positivity assumption; that is, the probability of observing the number of ratings for books published before 1990 should be strictly greater than zero. This helps us see that it’s important to consider the contexts of measurements and where missingness is coming from, in order to deliver the most accurate and meaningful interpretations of data with missingness.

References:

[1] Aronow, Peter M., and Benjamin T. Miller. *Foundations of Agnostic Statistics*. Cambridge University Press, 2019.

[2] http://jtleek.com/ads2020/

[3] https://www.kaggle.com/jealousleopard/goodreadsbooks