**PUBH 7485/8485**

**Homework 1**

The data for this assignment come from the obstetrics and perdiodontal therapy (OPT) study. The full description of these data are available from the following citation:

Meredith Hyun, James S. Hodges and Ann M. Brearley, “Obstetrics and Periodontal Therapy Dataset”, *TSHS Resources Portal (2019)*. Available at [https://www.causeweb.org/tshs/obstetrics-and-periodontal-therapy/](http://www.causeweb.org/tshs/obsetrics-and-periodontal-therapy).

The main findings of this trial can be found in the following citation:

Michalowicz BS, Hodges JS, DiAngelis AJ, Lupo VR, Novak MJ, Ferguson JE, Buchanan W, Bofill J, Papapanou PN, Mitchell DA, Matseoane S, Tschida PA. “Treatment of periodontal disease and the risk of preterm birth.” *New England Journal of Medicine*. 2006 Nov 2. 355 (18): 1885-94.

Briefly, this randomized controlled trial tested whether or not treatment of maternal periodontal disease could reduce pre-term birth and improve other birth outcomes. A total of 823 participants who had periodontal disease and were 13-16 weeks pregnant were randomized to receive periodontal therapy during pregnancy (treatment condition) or after pregnancy (control condition). Although this was a randomized trial we are interested in studying only those from the treatment group who completed their periodontal therapy. This is, of course, a nonrandom subset of the treatment group in the population so there are important prognostic differences among the control group and this subset of treatment. The outcomes of interest are whether or not the pregnancy ended before 37 weeks (*Preg.ended...37.wk*, binary variable) and the birth weight (*Birthweight* in grams among those with live birth, continuous variable). Participants who were lost-to-follow-up prior to the end of pregnancy were not included in this analysis.

The data for the participants who met the inclusion/exclusion criteria for this study have been posted to the course website as an R workspace. These datafiles also include variables to consider as potential confounders (a subset of all the variables available) which have been modestly cleaned. Description of the variables can be found at

<https://www.causeweb.org/tshs/datasets/OPT_Data_Dictionary.pdf> .

Additionally, the R code to reproduce the subset of the data used in this assignment from the complete data on the website above is included. The analytic dataset for this study contains 590 participants and 43 variables. The data will require some cleaning, particularly with missing data.

In the intention-to-treat analysis, the percentage of pregnancies ending before 37 weeks (13.1% vs.12.3% in the control and treatment groups, respectively, p = 0.81) and mean birthweight (3181 g vs. 3217 g, p = 0.46) were not significantly different among the groups. The goal of this assignment is to estimate the average causal treatment effect (i.e., the effect that we would have observed had those in the treatment group fully complied) in this study.

1. Summarize the differences between the treatment and control for the covariates included in the dataset using the tableone package in R (or similar output using a different statistical language). Please make sure that the variable names and levels for the categorical variables are informative and clear and that the table includes the standardized mean difference.
2. Find the unadjusted average treatment effect, standard error, and 95% confidence intervals for the outcomes of interest.
3. Consider estimating the causal treatment effect using regression adjustment. Please report the estimated average (causal) treatment effect, standard error, and 95% confidence intervals for each outcome. For the purposes of this assignment, please impute the median for continuous variables and the most common category for categorical and adjust for the following variables. The regression model for birthweight should include main effects and interactions and the regression model for pregnancies ending before 37 weeks should include main effects for the following variables:

*Race\_ethnicity, Public.Asstce, Use.Tob, N.prev.preg, Live.PTB, BL.GE, BL..BOP, BL..PD.4, BL..CAL.3*

PUBH 8485 Questions

Suppose that we have conducted a clinical trial and have collected data on a covariate *X* which we believe to be highly prognostic of outcome. Without loss of generality, assume that *E(X) = 0* and *V(X) = 1* and that *P(A=1)=0.5*, where *A* is the binary treatment indicator. Note that because this is a randomized study *X* and *A* are independent.

1. Suppose that and we estimate these regression parameters using least squares or equivalently by solving the following estimating equations

Thus, the regression adjustment estimator for the average treatment effect is just  . Note you may also assume that the conditional variance of *Y* is constant and does not depend on *X* and *A*. Show that this estimator has limiting variance less than the estimator for the average treatment effect which is the difference in sample means of the outcome between the two groups.

1. Now suppose that we fit the same model as above but with . Thus, we have omitted the interaction between A and X. Show that is still a consistent estimator for the average treatment effect.