**PUBH 7485**

**Homework 1**

The data for this assignment come from the following sources, both of which are included in the *causaldata* package in R.

*nhefs*

The National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS) was a longitudinal survey designed to investigate the relationships between clinical, nutritional, and behavioral factors and long-term health outcomes. The subset in this R package includes those who smoked in 1971 (i.e., the number of cigarettes smoked per day in 1971 was at least 1). The “treatment” of interest is whether or not the participant quit smoking between the first survey in 1971 and 1982 (variable *qsmk*). The outcomes of interest are the weight change between 1971 and 1982 (wt82\_71) and the systolic (*sbp*) and diastolic blood pressure (*dbp*) in 1982.

Please note, a description of each of the variables is included in the dataset *nhefs\_codebook*.

*black\_politicians*

These data come from a study by Broockman (2013, *American Journal of Political Science*) on whether state legislators responded to a letter requesting information about unemployment benefits. The design is somewhat complex and we will discuss more later. The key “treatment” variable is whether the state legislator was Black (*leg\_black*) and the response variable is whether they responded to the letter (*responded*). Description of the other variables included in the dataset is available from <https://cran.r-project.org/web/packages/causaldata/causaldata.pdf>.

1. For each dataset, summarize the differences between the treatment/exposure groups 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. For the *nhefs* data please include only those variables which were collected in 1971. Other follow-up variables (e.g., year of death) should not be included.
2. Find the unadjusted average treatment effect (i.e., use the simple estimators we would use if we were to assume there was no confounding), standard error, and 95% confidence intervals for the outcomes of interest for each dataset.