Data Analysis Assignment: Propensity Scores

For this assignment, the causal question of interest is: Does smoking cessation cause weight gain? The data we are going to use to try and answer this question is from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). A detailed description of the NHEFS, together with publicly available data sets and documentation, can be found at www.cdc.gov/nchs/nhanes/nhefs/nhefs.htm. For this assignment, we are going to use a subset of the NHEFS data used by Miguel Hernan and Jamie Robins in their online book "Causal Inference". This subset contains data on 1,566 cigarette smokers aged 25-74 years who, as part of the NHEFS, had a baseline visit (between 1971 and 1975) and a follow-up visit about 10 years later (1982). The dataset is further restricted to NHEFS individuals with non-missing sex, age, race, weight, height, education, alcohol use and intensity of smoking at the baseline (1971-75) and follow-up (1982) visits, and who answered the general medical history questionnaire at baseline.

Individuals are classified as treated if they reported having quit smoking before the follow-up visit, and as untreated if otherwise. Everyone's weight gain was measured (in kg) as the body weight at the follow-up visit minus the body weight at the baseline visit.

Here are some useful online tutorials for implementing a range of PS methods in SAS, STATA and R.

- https://www.ssc.wisc.edu/sscc/pubs/stata_psmatch.htm
- https://stanford.edu/~ejdemyr/r-tutorials-archive/tutorial8.html
- ftp://cran.r-project.org/pub/R/web/packages/ipw/ipw.pdf
- http://blog.stata.com/tag/inverse-probability-weighting/

The purpose of this assignment is to attempt to answer the question, "Does quitting smoking cause you to gain weight" using propensity score methods?

1. First construct a causal diagram for the question of interest and paste it below.
2. Before you get started analyzing the data, create a **studyid** variable for each participant and remove the outcome variable **wt82_71** as well as a copy of the study id variable from the dataset and set aside. The idea is for you to conduct your matching analysis using the matching data only (i.e., baseline covariates, treatment status, without the outcome variable). And then, after the treatment model and propensity score is finalized, merge the two data sets together to estimate the treatment effects of interest once.
3. Based on your causal diagram from question 1, choose a set of covariates that you think is an appropriate matching set and indicate why. If you think some of the covariates are more important than others, point that out. If there are variables that should not be in the treatment model, point that out as well and indicate why. Take some time to recode the data as appropriate and check for missing data and other problems. Before we build the propensity score model, make a balancing table comparing baseline (e.g., pre-treatment) covariates between those who quit (i.e., treated) and did not quit smoking (i.e., not treated). Explain the results. Do the statistics in the table tell you much about the likely value of matching in this application? Describe what statistical metric you are using to assess balance (e.g., t-tests, absolute standardized differences).
4. Next, estimate the propensity for an observation to be "treated" (or in this case to have quit smoking). Normally, a propensity score analysis would try many propensity score functional form assumptions to obtain good covariate balance between treatment groups. For this assignment, you should try out (say) 4-5 models and choose the preferred model for your analysis. Some of the more commons (from a coding perspective) methods of using propensity scores are a) inverse probability of treatment weighting, b) matching weights, c) stratification, d) pair matching. Select a method and use it to assess how well the propensity score "balances" the covariates in the data for several different candidate propensity score equations. To do so, make a similar table to the one you made in question 2, but this time add two more columns to compared characteristics between treatment groups after application of

the propensity score (e.g., weighting or matching). Be sure to play around with interaction terms and polynomials to see how well you can balance your data. Provide details regarding the steps you took, including how variables were selected into the PS model. Come up with a metric to compare your propensity score models to see which one balances the data best. Copy and paste from the linear predictor portion (i.e., the beta coefficients and the variable names) of the logistic regression used to adjust for your relation of interest.

5. Once you have your preferred propensity score equation in hand, you should assess the support condition (i.e., overlap, region of common support, or positivity). Make a histogram plot of the distribution of propensity scores in the treatment and untreated groups. Explain the idea of the common support assumption. Does it appear to hold in these data? Should you implement a trimming rule? If you do apply some trimming, explain what you are doing. If you are using a weighting approach, make a histogram of the weights. Should you implement a trimming or truncating rule for the weights?

6. With the propensity scores estimated, it is time to look at the outcome data. Merge the outcome data with the matching data and the added propensity scores. Use the propensity scores to estimate the average treatment effect of quitting smoking on weight gain (this can be a simple mean difference or you can use an ordinary least square linear regression model). Here again, you should try any way to use the propensity score in the model that you like (e.g., (a) inverse probability of treatment weighting, b) matching weights, and c) stratification, d) pair matching). Describe your results.

7. **Extra Credit.** Create a table like the one below comparing the effect of smoking cessation with weight gain using a variety of PS methods.

	No of unique participants in the final analysis	**Mean Difference in Weight Gain (The Treatment Effect)**	**95% CI**
Crude Model			
Multivariate adjusted model			
Matched on PS			
Greedy 1:1			
Optimal 1:1			
Greedy 1:many			
Optima 1: many			
Regression adjusted for PS			
Continuous			
Deciles			
Weighted Models			
Matching Weights			
IPW			
Overlap weights			