# IPTW of MSM exercise

*January 30, 2020*

Throughout this exercise we will be addressing the question: what is the causal effect of quitting smoking on weight? We will be estimating this effect using data that was simulated using data from the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The dataset and codebook are available from the website for the Causal Inference book by Miguel Hernan and James Robins: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Begin by loading the dataset into R from my github

Q1. Our two variables of interest are the exposure *qsmk*, quit smoking, in 1982, and the outcome *wt82*, weight in 1982. For the purposes of these exercises, let's assume that the only confounders we are worried about are *sex*, *age*, *education* and *exercise*. Time-varying variables are followed by a 1 or 2 indicating the time they were measured. Run some descriptive statistics that you would normally include in a table 1 for *qsmk1* (the average value or proportion of confounders in each of the exposure groups). On the last page there are two tables to help you keep track of the of balance across covariates from different models and the estimates themselves from different models. You can either fill them in by hand or create a similar table in another program.

Q2. Use outcome regression to estimate the crude and adjusted effect of *qsmk1* on *wt82*.

Q3. Now we will start estimating the weights. Estimate a regression where the exposure, *qsmk1*, is the dependent variable in your model and the confounders are the independent variables. Use R to output a variable that is the probability of quitting smoking in each observation.

Q4. Create a variable that is the probability that each observation receives the exposure, *qsmk1*, that they were observed to receive. In other words, among those who quit smoking ($qsmk1 = 1$), the value should be the probability that they would quit smoking given confounders, $P(qsmk1 = 1|L = l)$. Among those who did not quit smoking ($qsmk1 = 0$), the value should be the probability that they would not quit smoking give confounders, $P(qsmk1 = 0|L = l)$ which is equivalent to $(1 - P(qsmk1 = 1|L = l))$. Taking the inverse of the this variable will give you the IPTW weights. Plot your weights on a histogram and check their mean and standard distribution.

Q5. These weights can now be used to create a pseudopopulation where the confounders have a different disitribution relative to exposure. Use the same functions you used in Q1 to create your table 1 but add the 'weight' option to use the weights you created in Q4. Has the balance between those who quit smoking and those who didn't quit smoking changed?

Q6. One of the advanatages of the propensity scores is that you can play with the model to get the balance right between the two groups without worrying about p-hacking (because you aren't looking at the outcome model). Add some interactions or non-linear effects to the model you used in Q3, recalculate the weights in Q4 and use the new weights in Q5. Have you improved the balance between the exposure groups?

Q7. Once you are satisfied with the balance between your exposure groups, estimate your outcome model using *qsmk* as the only independent variable but weight your model using the weights you created in Q5. How does the result compare to what you estimated in Q2?

Q8. Now lets calculate weights for *qsmk2*. Follow the same process to calculate weights for *qsmk2* being sure to include *qsmk1* and the confounders measured at time 2 in your propensity score model. Check for balance the same way you did with *qsmk1*. Once you're happy with the balance, multiply the weights for *qsmk1* by the weights for *qsmk2*. Check balance again with these new weights.

Q9. With the multiplied weights, estimate a marginal structural model of the effec of *qmsk1* and *qsmk2* on *wt82*.