Week 4 1-pager Subclassification and Matching

1) Takeaway
   a. Matching on observables is the common theme across each estimator
   b. Pro: we can control for selection on observable confounders
   c. Con: No way to control for selection on unobservable confounders; only way we know if ATE of matching is not biased is if we benchmark to RCT, but we don't need matching with an RCT
   d. Assumptions across subclassification and matching
      i. Conditional Independence Assumption $(Y^1, Y^0) \perp D|X$
      i. Common Support/Covariate Balance $0 < Pr(D = 1|X) < 1$
      ii. Stable Unit Treatment Value Assumption (SUTVA)
   e. Testable Assumptions: Common Support/Covariate Balance

2) Subclassification
   a. A method of satisfying the backdoor criterion by weighing differences in means by strata-specific weights
   b. Uses differences between treatment and control groups units and achieves covariate balance by using K probability weights to weight averages
   c. Smoking Death Rates Example: Divide into strata; calculate death rates; apply the weight ages of control to treatment group mortality rates; and sum to get weighted ATE
   d. $\hat{\delta}_{ATE} = \int (E[Y|D = 1] - E[Y|X, D = 0)dPr(X)$

3) Exact Matching
   a. Using covariates to find a match between a treatment observation and a control observation based upon confounding observables
   b. If there are more than one exact match? We can take M matches and calculate an average outcome as the counterfactual
      i. Where Yj(i) is the jth unit matched to the ith unit based on the jth being the "closest" to the ith unit for some X covariate
      ii. $\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left[Y_i - \left(\frac{1}{M}\sum_{m=1}^{M} Y_{j_m(1)}\right)\right]$
      iii. Where $2D_i - 1$ is switching equation: When D=1, the term is positive 1 and when D=0 the term is negative -1.
      iv. It is preferable to have M to be small, such as M=2

4) Propensity Score Matching
   a. Getting a single scalar, eliminates the problem of no matches among many covariates and eliminates the problem of lack of dimensionality
   b. The basis of PSM is to compare units based on observables have very similar propensities to be placed into the treatment group even though these units differ regarding treatment assignment
   c. Unit A and Unit B are assigned to treatment and control, respectively, and both have an estimated 60% chance of being assigned to treatment, but by random chance A receives treatment while B does not
   d. Under Conditional Independence Assumption: $(Y^1, Y^0) \perp D|X \rightarrow (Y^1, Y^0) \perp D|p(X)$
   e.

5) Inverse Propensity Weights (propensity scores)

a. Inverse Propensity Weights (IPW) use propensity scores but it is not matching
b. There is a two step process to calculate the ATE
    i. Estimate the propensity scores $p(X_i)$ with logit or probit
    ii. Estimate the ATE with the $\hat{p}(X_i)$
    iii. $\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\frac{D_i Y_i}{\hat{p}(X_i)} - \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i(1-D_i)}{(1-\hat{p}(X_i))}$
c. IPW method can be sensitive to extreme values of the propensity scores, so normalized weights are a method around this issue

6) K-Nearest Neighbor (KNN)
  a. Find the nearest k-neighbors within a particular radius of the treatment unit's propensity score to generate a pool of controls (where k is the number of matches)
  b. Similar to exact matching but using propensity scores to make the match
    i. $\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left[Y_i - \left(\frac{1}{M}\sum_{m=1}^{M}Y_{j_m(1)}\right)\right]$
  c. KNN needs a balanced sample, if not then forcing a balanced sample or excessive trimming can introduce bias