PROPENSITY SCORE MATCHING IN STATA

Acknowledgement: This is adapted tutorial from the lecture series of Robert Greevy within Vanderbilt University's Center for Health Services Research

```
ssc install psmatch2
```

```
ssc install table1
```

To motivate the propensity score matching, let's use the `cattaneo2` dataset, a STATA example dataset. It can be loaded with the following command:

```
webuse cattaneo2
```

The data in cattaneo2 is a subset of data that was analysed in the following journal articles:

- Almond, D., Chay, K.Y., Lee, D.S., 2005. *The costs of low birth weight,* Quarterly Journal of Economics 120, 1031-1083.
- Cattaneo, M.D. 2010. *Efficient semiparametric estimation of multi-valued treatment effects under ignorability,* Journal of Econometrics, 155(2), 138-154.

The dataset included information about infant/mother/father characteristics from singleton births in Pennsylvania between 1989 and 1991. The original dataset included nearly 500,000 births. The STATA example dataset includes 4642 births.

The paper reference of this exercise is Elizabeth A. Stuart. 2010. *Matching Methods for Causal Inference: A Review and a Look Forward*, Statistical Science, Vol. 25, No. 1, 1–21.

RESEARCH QUESTION: What is the effect of maternal smoking during pregnancy on the infant's birthweight?

The dataset includes the following set of variables:

Infant	Mother	Father
birth weight (grams)	married/not married	
birth month	hispanic (yes/no)	hispanic (yes/no)

Infant	Mother	Father
	race (white/not white)	race (white/not white)
	age	age
	education	education
	foreign born (yes/no)	
	birth number	
	months since last birth	
	infant of previous births died (yes/no)	
	number of prenatal care visits	
	trimester of first prenatal care visit	
	alcohol during pregnancy (yes/no)	
	smoking during pregnancy (0, 1-5, 6-10, 11+ daily)	
	smoking during pregnancy (yes/no)	

You can access the complete codebook with the command `codebook` after loading the data.

BIG PICTURE: BACKGROUND

	Randomized Clinical Trial	Observational Study
Treatment Assignment:	Investigators generate a treatment schedule prior to patient enrollment. The schedule is constructed based on the design of the study, which includes randomization in some fashion. Physicians (who may be blind to treatment as well) assign treatments/exposures to study participants following the sequence in the schedule.	• Physicians assign treatment/exposure based on • characteristics of the patient • personal preference • regional preference • insurance restrictions • Study participants choose for themselves treatments or behaviors • Exposures are based on geographic location or cultural identity
CONSEQUENTLY		
Probability of Treatment:	Known	Unknown, may be 0 or 1
Covariate Balance:	Relationship between covariates and treatment assignment are known from study design. Usually the study is designed so that there is no relationship between treatment assignment and covariates.	Relationship between covariates and treatment assignment is unknown. There may be covariate imbalance.

BIG PICTURE: CHALLENGES

Differences in outcomes between the treated and untreated (or exposed and unexposed) may be the consequence of confounding variables and not the treatment (or exposure).

Dataset may include sub-groups for which a treatment effect should not be calculated because

- the sub-group may not be eligible for one treatment (a treatment effect should only be calculated in populations eligible for both treatments)
- though a sub-group may be eligible for both treatments, there may not be enough data for a comparison without extrapolation.

BIG PICTURE: SOLUTIONS

There are several methods for estimating a treatment effect with observational data. In this lecture series, you have been exposed (not randomly) to a family of methods which use the propensity score. The primary focus has been on propensity score matching.

STEPS (A SUMMARY OF STUART 2010)

1. Defining “closeness”: the distance measure used to determine whether an individual is a good match for another.
    0. Variables to include
        - Researchers should thus be liberal in terms of including variables that may be associated with treatment assignment and/or the outcomes.
        - One type of variable that should not be included in the matching process is any variable that may have been affected by the treatment of interest (Rosenbaum, 1984; Frangakis and Rubin, 2002; Greenland, 2003).
    1. Distance Measures
        - Exact
        - Mahalanobis
        - Propensity score / linear propensity score
            - i) With propensity score estimation, concern is not with the parameter estimates of the model, but rather with the resulting balance of the covariates (Augurzky and Schmidt, 2001).
2. Implementing a matching method, given that measure of closeness.
    - Methods:
        - i) k:1 Nearest Neighbor

          Estimates Average Treatment Effect in the Treated (ATT or ATET)

          May need a caliper
        - ii) Subclassification
        - iii) Full matching

iv) Weighting (IPTW)

- Assess Common Support
    - Examining the common support may indicate that it is not possible to reliably estimate the ATE.

3. Assessing the quality of the resulting matched samples, and perhaps iterating with steps 1 and 2 until well-matched samples result.
    - Perhaps the most important step in using matching methods is to diagnose the quality of the resulting matched samples
    - Tools:

        i. Histograms / Density Plots / ECDF

        ii. Standardized Bias / Standardized difference in means

4. Analysis of the outcome and estimation of the treatment effect, given the matching done in step 3.

    i) After k:1 matching

    - May not need to account for matched pair
    - Must use weights if matching was *with replacement*
    - Use regression adjustment

    ii) After subclassification

    - Aggregation weights determine estimation of ATT or ATE
    - Use regression adjustment