**MWSUG 2018 - Paper AA-047**

# Propensity Scores and Causal Inference for (and by) a Beginner

Bruce Lund, OneMagnify

## ABSTRACT

In an observational study the subjects are assigned to treatments through a non-randomized process. In the simplest and most typical case there are two treatments, one being deemed as “control”. Associated with the subjects is an “outcome” which is of interest to the researcher. The outcome could be discrete, very often binary, or have continuous numeric values. The researcher wants to know the effect of the treatment on the outcome. This is called the average treatment effect and abbreviated by ATE. Due to the non-random assignment of treatments, a simple comparison of outcomes, such as an average per treatment group, would be biased. One solution to removing the bias rests on finding covariates for the subjects such that the treatment assignment can be regarded as random for subjects having essentially equal covariate values. This is called “covariate balance” for the treatments. Although covariate balance is virtually a necessary condition when computing ATE, it is not also sufficient. Specifically, the covariate collection must also include any covariate which is strongly related to both treatment and outcome. Finding such covariates is a central challenge when performing a causal analysis. Once accomplished, then an analysis of outcomes can be undertaken. The SAS® procedure PROC PSMATCH provides methods to adjust the sample so that the treatment can be regarded as random. An output data set from PROC PSMATCH is then used by other SAS procedures to estimate the causal effect. This paper gives hands-on experience regarding the assumptions that enable the analysis of causal effects in an observational study and gives simple examples of usage of PSMATCH.

## INTRODUCTION

In an observational study the subjects are assigned to treatments through a non-randomized process. In this paper the treatment T will consist of two levels where one will be deemed as “control”. Associated with the subjects is an outcome Y which is of interest to the researcher. The outcome could be discrete, very often binary, or have continuous numeric values. The researcher wants to know the effect of the treatment on the outcome. An analysis of the treatment effect cannot be performed for individual subjects since a subject experiences either the treatment or the control but not both. Therefore, this analysis is done at the group level. The researcher wants to estimate the “Average Treatment Effect” (ATE) which is the difference in the average outcome between the treatment group and the control group. Due to the non-random assignment of treatments, a simple comparison of average outcomes would be biased.

One solution to removing the bias rests on finding covariates for the subjects such that the treatment assignment can be regarded as random for subjects having essentially equal covariate values. This is called “covariate balance” for the treatments. Although covariate balance is virtually a necessary condition when computing ATE, it is not also sufficient. Specifically, the covariate collection must also include any covariate which is strongly related to both treatment and outcome. Finding such covariates is a central challenge when performing a causal analysis. Once found, an analysis of outcomes can be undertaken.

PROC PSMATCH was introduced in SAS/STAT 14.2. PSMATCH carries out the balancing of covariates so that the treatment can be viewed as random. Additional SAS procedures are then used to measure the causal effect. Also in SAS/STAT 14.2 PROC CAUSALTRT was introduced. PROC CAUSALTRT provides a “one-stop” process for balancing of covariates and for the measurement of the causal effect.

## PURPOSE OF THIS PAPER

One purpose of this paper is to give hands-on experience regarding the assumptions that enable the analysis of causal effects in an observational study. This is done by giving simple examples and simple proofs of several key theorems.[1] These proofs employ only elementary rules of probability. In particular, a theorem is discussed which gives the “inverse probability weighting” method for computing ATE

[1] All theorems in this paper have been known for roughly 35 years. See Lunceford and Davidian (2004) or Angrist, and Pischke (2009) or Lamm and Yung (2017) for a list of references and commentary.

(Theorem 3). The consequences of the theorem might seem mysterious or even magical. This paper examines the “inverse probability weighting” method both in terms of assumptions and conclusions.[2]

A second purpose is to connect the “inverse probability weighting” method to the operation of PROC PSMATCH and follow-on SAS procedures. Only a small subgroup, however, of the options of PSMATCH are illustrated by the examples in this paper. In particular, none of the PSMATCH “matching methods” are included. Additionally, PROC CAUSALTRT is briefly discussed.

This paper might serve as preliminary reading before studying the paper “Propensity Score Methods for Causal Inference with the PSMATCH Procedure” by Yuan, Yung, and Stokes (2017).

## CONDITIONAL INDEPENDENCE OF RANDOM VARIABLES

Let X, T, and E be random variables on a data set S (formally, a probability space S). The random variables X, T, E can be vector-valued. For example, X may consist of k random variables $X_1$, …, $X_k$ so that X=( $X_1$, …, $X_k$) maps into $R^k$. Even though X may consist of k component random variables, the notation X and X=x will be used. There will be no need to single out the individual components of X.

Definition: X and T are conditionally independent given E if the following condition (A) is true:

(A) P(X=x, T=t | E=e) = P(X=x | E=e) * P(T=t | E=e) for all x, t, and e from S [3]

Consider data set S in Table 1. Each row (each subject) has probability of 1/7 of random selection.

In Table 1 the random variables X and T are conditionally independent given the random variable E. Condition (A) is verified below for the example where X=1, T=1, and E=0.5:

P(X=1, T=1 | E=0.5) = ¼ = ½ * ½ = P(X=1| E=0.5) * P(T=1| E=0.5)

Of course, there are other combinations of values for X, T, E that must be tested.

| X | T | E |
|---|---|---|
| 1 | 0 | 0.5 |
| 1 | 1 | 0.5 |
| 2 | 0 | 0.667 |
| 2 | 1 | 0.667 |
| 2 | 1 | 0.667 |
| 3 | 0 | 0.5 |
| 3 | 1 | 0.5 |

**Table 1: Data set S with random variables X, T, and E**

### Equivalent Conditions for Conditional Independence:

Conditions (A), (B), and (C) are equivalent. In each case X and T are conditionally independent given E.

(A) P(X=x, T=t | E=e) = P(X=x | E=e) * P(T=t | E=e) for all x, t, and e
(B) P(X=x | T=t, E=e) = P(X=x | E=e) for all x, t, and e
(C) P(T=t | X=x, E=e) = P(T=t | E=e) for all x, t, and e

**Exhibit 1: Equivalent Conditions for Conditional Independence of X and T given E**

The steps to prove (A) from (C) are shown here (in short-hand notation). By definition of conditional probability (twice): P(X, T | E)= P(X, T, E) / P(E)= P(T | X, E) * P(X, E) / P(E) … It is assumed P(E) > 0

Using (C) the “X” is removed from P(T | X, E) as shown below:

P(T | X, E) * P(X, E) / P(E)= P(T | E) * P(X, E) / P(E)= P(T | E) * P(X | E)

Therefore: P(X, T | E)= P(T | E) * P(X | E), which is condition (A)

[2] Not discussed in this paper is a topic in causal analysis called the “average treatment effect for the treated” (ATT). See Yuan, Yung, and Stokes (2017, p. 1) to begin a discussion of this topic.
[3] The comma inside “P(X=x, T=t)” stands for “and”, that is, X=x AND T=t

Conditions (A), (B), (C) are given for points (x, z, e), but all these conditions may be restated in terms of sets. For example, $P(X\in\mathbf{X})$ indicates the probability statement applies to values of X that belong to a set **X**. Condition (A) becomes: $P(X\in\mathbf{X}, T\in\mathbf{T} \mid E\in \mathbf{E}) = P(X\in\mathbf{X} \mid E\in\mathbf{E}) * P(T\in\mathbf{T} \mid E\in\mathbf{E})$ for all **X, T,** and **E**.

## CONDITIONAL INDEPENDENCE AND PROPENSITY SCORES

In the following, T is a binary treatment variable where T=1 is "treated" and T=0 is "control".[4] Covariates X (where X may denote many variables) are used in predicting T. Such covariates must have values determined before the determination of the value of T. Let S be a data set (or probability space) where each subject in S has values for T and X and each subject has equal probability of random selection.

The *propensity score model* is defined as the probability that T=1 for covariate value X=x. The notation $e(x) = P(T=1 \mid X=x)$ will be used. By definition of e(x), the treated subjects with X=x will account for 100*e(x)% of the subpopulation of S where X=x. In this sense this theoretical propensity score is *perfect*.[5]

In the following sections, for given $x_0$ the set $E_{e(x0)}$ will denote the values x taken by X where:

$$E_{e(x0)} = \{x: e(x) = e(x_0)\}$$

### T AND X ARE CONDITIONALLY INDEPENDENT GIVEN $E_{e(x0)}$ … or $e(x_0)$

Given $E_{e(x0)}$, then T and X are conditionally independent. A proof is given below Table 2. In the probability statements below, the short-hand notation $E_{e(x0)}$ is used to replace the formally correct $X\in E_{e(x0)}$. Here, $X\in E_{e(x0)}$ indicates that the value of X belongs to $E_{e(x0)}$. By condition (C) of conditional independence:

$$P(T=t \mid X=x, E_{e(x0)}) = P(T=t \mid E_{e(x0)}) \text{ for all values of T and X in data set S, given } E_{e(x0)}.$$

In shortened notation, $e(x_0)$ takes the place of $E_{e(x0)}$ and conditional independence is expressed by:

$$P(T=1 \mid X=x, e(x_0)) = P(T=1 \mid e(x_0))$$

or even more succinctly by:

$$T \_||\_ X \mid e(x)$$

In Table 2, the propensity score is e(x) and it is true that T _||_ X | e(x).[6]

| X | T | e(x) | $E_{e(x)}$ | Y: outcome<br>(for later reference) |
|---|---|---|---|---|
| 1 | 0 | 0.5 | {1, 3} | 2 |
| 1 | 1 | 0.5 | {1, 3} | 3 |
| 2 | 0 | 0.667 | {2} | 1 |
| 2 | 1 | 0.667 | {2} | 4 |
| 2 | 1 | 0.667 | {2} | 2 |
| 3 | 0 | 0.5 | {1, 3} | 5 |
| 3 | 1 | 0.5 | {1, 3} | 0 |

**Table 2: Data set S with random variables X, T, propensity scores e(x) and sets $E_{e(x)}$**

### PROOF OF T _||_ X | e(x) USING ELEMENTARY METHODS

Short proofs of T _||_ X | e(x) are given in books and papers that utilize the notation and ideas of expected values.[7] The proof below uses basic rules of probabilities and avoids expected values. (References to the underlying data set S will be omitted.)

[4] Treatments must satisfy common sense but important conditions labeled SUTVA. See Lamm and Yung (2017, p. 5)

[5] But in applications, e(x) is fit from the available analysis data set and would rarely be perfect.

[6] One verification using condition (A): $P(X=1, T=1 \mid e(x) = 0.5) = ¼ = ½ * ½ = P(X=1 \mid e(x) = 0.5) * P(T=1 \mid e(x) = 0.5)$
Note: Probability of a row is 1/7 and $P(X=1, T=1 \mid e(x) = 0.5) = P(X=1, T=1, e(x) = 0.5) / P(e(x) = 0.5) = (1/7)/(4/7) = ¼$.
Similarly, applying conditional probability and 1/7 probability for each row gives the right hand side of ½ * ½ = ¼.

[7] Here is the proof from Rosenbaum (2010, p. 73): Taking the case T=1, condition (C) is verified below:
$P(T=1 \mid e(x)) = E\{T \mid e(x)\} = E\{E\{T \mid x, e(x)\} \mid e(x)\} = E\{P\{T=1 \mid x, e(x)\} \mid e(x)\} = E\{e(x) \mid e(x)\} = e(x) = P(T=1 \mid x, e(x))$

**THEOREM 1**: If e(x) is the perfect propensity score for covariates X and treatment T, then T _||_ X | e(x)

In this proof the conditional independence condition (C), given below, will be verified.

$$P( T=1 \mid X=x', E_{e(x0)} ) = P( T=1 \mid E_{e(x0)} ) \text{ … where } E_{e(x0)} = \{x: e(x) = e(x_0)\}$$

It can be assumed that x' (from X=x') in the equation above is contained in $E_{e(x0)}$.[8] The verification of condition (C) is given by showing that both the right hand side and the left hand side reduce to $e(x_0)$.

Left side:

$P( T=1 \mid X=x', E_{e(x0)} ) = P( T=1 \mid X=x' )$ … since $E_{e(x0)} \cap \{X=x'\} = \{X=x'\}$

$P( T=1 \mid X=x' ) = e(x_0)$ … since x' belongs to $E_{e(x0)}$

Right side:

$P( T=1 \mid E_{e(x0)} ) = P( T=1, E_{e(x0)} )) / P( E_{e(x0)}) )$

$= \{ \sum_{x' \in Ee(x0)} P( T=1, X=x', E_{e(x0)}) \} / P( E_{e(x0)} )$ … sum of probabilities across disjoint subsets

$= \{ \sum_{x' \in Ee(x0)} P (T=1, X=x' ) \} / P( E_{e(x0)})$ … since $E_{e(x0)} \cap \{X=x'\} = \{X=x'\}$

$= \{ \sum_{x' \in Ee(x0)} P( T=1 \mid X=x' ) * P( X=x' ) \} / P( E_{e(x0)} )$

$= \{ \sum_{x' \in Ee(x0)} e(x_0) * P( X=x' ) \} / P( E_{e(x0)} )$ … since $x' \in E_{e(x0)}$

$= \{ e(x_0) / P( E_{e(x0)}) \} * \sum_{x' \in Ee(x0)} P( X=x' )$

$\bigcup_{x' \in Ee(x0)} \{x'\} = E_{e(x0)}$ and the formula for the sum of probabilities across disjoint subsets gives …

$= \{ e(x_0) / P( E_{e(x0)} ) \} * P( E_{e(x0)} ) = e(x_0)$

The "trick" was the divide in $E_{e(x0)}$ into disjoint subsets consisting of the {x} which belong to $E_{e(x0)}$ (as is done in the second line). A similar proof applies when T=0.

## IMPORTANCE AND LIMITATIONS OF THEOREM 1

By Theorem 1 the probability distributions of covariates X, given a value of e(x), are the same distributions for both T=0 and T=1.[9] The covariates are "balanced" for treated and control.

For example, in Table 2 for e(x)=0.50, the distribution of X for T=0 is P(X=1)=0.5, P(X=2)=0, P(X=3)=0.5. The same is true for T=1 when e(x)=0.5.

Perhaps a non-randomized experiment (observational study) may be regarded as randomized if the averages of the outcome variable for treated and for control are compared only for subsets of subjects which have the same value of e(x)? This allows treated and control outcomes to be compared for subsets of subjects having balanced covariates, an apples to apples comparison.

For example, using the outcome Y in Table 2 the case might be made that the effect of the treatment T on Y (the ATE) could be measured as follows:

| e(x) | Average Y for T = 1 | Average Y for T = 0 | Difference "D" | % of Sample "S" | Weighted = "D" * "S" |
|---|---|---|---|---|---|
| 0.5 | 1.5 | 3.5 | -2.0 | 0.571 | -1.142 |
| 0.667 | 3.0 | 1.0 | 2.0 | 0.429 | 0.858 |
| | | | | ATE = | -0.284 |

**Table 3: Calculation of ATE for subsets of treated and control with the same value of e(x)**

However, for the analysis of Table 3 to be appropriate, the outcome and the treatment must also be independent given the covariates. Although covariate balance is virtually a necessary condition when computing ATE, it is not also sufficient. Specifically, the covariate collection must include any X which is strongly related to both treatment and outcome. The apples to apples comparisons must use the right apples.

[8] If X=x' is not contained in $E_{e(x0)}$, then equivalent condition (A) is satisfied. Both left and right sides of (A) equal zero.

[9] This is due to Condition (B) of the equivalent conditions for conditional independence.

Finding all such covariates is a central challenge when performing a causal analysis.

The earlier statement “the outcome and the treatment must also be independent given the covariates” needs quantification. To quantify this statement requires that the straightforward notion of “outcome” be refined by the concept of “potential” outcome. The concept of potential outcome is presented in the section below.

Employing the concept of potential outcomes and the “Strong Ignorability Assumption” (SIA), a setting is created where an observational study can be regarded as a randomized study. In this setting there are formulas for the computation of ATE.

The discussion of SIA follows the discussion of potential outcomes.

Stepping back from the theory of Theorem 1, which gives a result about probabilities, a modeler will, generally, have only a sample from some larger finite or conceptual population. Each subject in the sample will have a treatment and covariates. The modeler, using logistic regression or some other binary-target model, will fit an empirical propensity score model to the sample. How does the modeler assess the success of the propensity score in obtaining balance? This question is taken up when discussing, in a later section, the toolkit provided by PROC PSMATCH.

## POTENTIAL OUTCOMES AND POTENTIAL OUTCOME MEANS (POM)

Returning to theory, each subject in an observational study data set S will have an outcome Y, treatment T, and covariates X. If the $i^{th}$ subject is treated, then the outcome $Y_i$, due to being treated, is observed. But this subject had the *potential* of having a different outcome if this subject had been placed in control. Similar statements are made for subjects in control.

Of course, a potential outcome cannot be observed. It is imagined or missing. Here is the notation to be used for outcomes, both real and potential, by treatment:

- The outcome $Y_i$ (1) is the treated outcome for the $i^{th}$ subject, whether or not actually treated
- The outcome $Y_i$ (0) is the control outcome for the $i^{th}$ subject, whether or not actually in control

$Y_i$ (1) is imagined for any subject in control and $Y_i$ (0) is imagined for any treated subject.

The potential outcomes, the actual outcome, and the treatment status are connected by the equation:

$$Y_i = T_i * Y_i\ (1) + (1-T_i) * Y_i\ (0)$$

### POTENTIAL OUTCOME MEANS (POM’s)

The mean of the potential outcomes for the treated group is the average of $Y_i(1)$, or E[Y(1)]. This average is not directly computable due to imagined (or missing) values of $Y_i(1)$ for subjects in control.

The sections below will discuss assumptions which allow this average to be computed. Likewise, average of $Y_i(0)$, or E[Y(0)], is not computable without further assumptions.

## AVERAGE TREATMENT EFFECT

The difference between the average of the *actual* treated outcomes and the average of the *actual* control outcomes can be computed and forms the “naïve” ATE.

$$ATE_{naive} = (\textstyle\sum_{i:\,Ti=1} Y_i) / (\sum_{i:\,Ti=1} T_i) - (\sum_{i:\,Ti=0} Y_i) / (\sum_{i:\,Ti=0}\ (1 - T_i)$$

But, treated subjects may be very different than control subjects due to non-random assignment of treatment. The effect of the treatment on the outcome is obscured if using $ATE_{naive}$.

In an observational study the challenge to the researcher is to estimate ATE where ATE is the difference between the potential outcome means for treated and control:

$$ATE = E[Y(1)] - E[Y(0)]$$

ATE is an unknown parameter of data set S. ATE is not computable without further assumptions since it is the difference of two potential outcome means (POM’s). Such assumptions are given in sections below.

## CONDITIONAL INDEPENDENCE OF Y(t) AND T, GIVEN X … FOR t = 0, 1

Continuing with theoretical concepts, a data set S is considered with subjects having Y, T, X and the perfect propensity score model $e(x) = P(T=1 \mid X=x)$.

The Strong Ignorability Assumption (SIA) is given below:[10]

(a) Conditional independence of Y(1) and T given X. This is denoted Y(1) _||_ T | X

(b) Conditional independence of Y(0) and T given X. This is denoted Y(0) _||_ T | X

(c) $0 < P(T=1 \mid X=x) < 1$ for all x. This says that if X=x appears for a treated subject, then X=x also appears for some subjects in control, and vice versa.

As a result of assumption (a) of SIA, the version (B) of conditional independence (see Exhibit 1) says:

$$P(Y(1)=y \mid T=t, X=x) = P(Y(1)=y \mid X=x) \quad \text{… likewise for } Y(0) \text{ for assumption (b)}$$

That is, given X=x, the probability of the potential treated outcome does not depend on the assignment of the treatment. The observational study becomes a randomized study, conditional on X.

If there are important omitted covariates, then SIA fails. For an example of this failure, see Appendix B.

Strong Ignorability with respect to X implies Strong Ignorability with respect to e(x) as given by Theorem 2.

**THEOREM 2**: If Y(1), Y(0), T and X satisfy SIA and e(x) is the perfect propensity score, then

(a’) Y(1) _||_ T | e(x)
(b’) Y(0) _||_ T | e(x)

**Proof of THEOREM 2**

Condition (a’) will be proved for T=1 by verifying:

$P(\,Y(1)=y \mid T=1, E_{e(x0)}\,) = P(\,T=1 \mid E_{e(x0)}\,)$ … where $E_{e(x0)} = \{x: e(x) = e(x_0)\}$

Begin with:

$P(\,Y(1)=y \mid T=1, E_{e(x0)}\,) = P(\,Y(1)=y, T=1, E_{e(x0)}\,) / P(\,T=1, E_{e(x0)}\,)$

Consider the denominator: $P(\,T=1, E_{e(x0)}\,) = \sum_{x' \in \mathrm{Ee(x0)}} P(\,T=1, X=x'\,)$

$= \sum_{x' \in \mathrm{Ee(x0)}} P(\,T=1 \mid X=x'\,) * P(X=x') = e(x_0) * \sum_{x' \in \mathrm{Ee(x0)}} P(\,X=x'\,) = e(x_0) * P(\,E_{e(x0)}\,)$

Consider the numerator:

$P(\,Y(1)=y, T=1, E_{e(x0)}\,) = \sum_{x' \in \mathrm{Ee(x0)}} P(\,Y(1)=y \mid T=1, X=x'\,) * P(\,T=1, X=x')$

$= \sum_{x' \in \mathrm{Ee(x0)}} P(\,Y(1)=y \mid X=x'\,) * P(\,T=1, X=x')$ … by conditional independent of Y and T given X

$= \sum_{x' \in \mathrm{Ee(x0)}} P(\,Y(1)=y \mid X=x'\,) * P(\,T=1 \mid X=x') * P(\,X=x'\,)$

$= \sum_{x' \in \mathrm{Ee(x0)}} P(\,Y(1)=y \mid X=x'\,) * e(x_0) * P(\,X=x'\,) = e(x_0) * \sum_{x' \in \mathrm{Ee(x0)}} P(\,Y(1)=y, X=x'\,)$

$= e(x_0) * P(\,Y(1)=y, E_{e(x0)}\,)$

Now putting numerator over denominator gives: $\{e(x_0) * P(\,Y(1)=y, E_{e(x0)}\,)\} / \{e(x_0) * P(\,E_{e(x0)}\,)\}$

$= P(\,Y(1)=y, E_{e(x0)}\,) / P(\,E_{e(x0)}\,) = P(\,Y(1)=y \mid E_{e(x0)}\,)$

Similar proofs apply to T=0 and condition (b’).

Next, the important Inverse Probability Weighting (IPW) method (Theorem 3) is discussed. It provides a practical tool for estimating ATE. The proof relies on the assumption of SIA.

As shown below, the inverse probability weight for treated subjects is $w_1(x) = 1/e(x)$ and for control $w_0(x) = 1 / (1-e(x))$.

[10] Called the “conditional independence assumption” (CIA) in Angrist and Pischke (2009)

## INVERSE PROBABILITY WEIGHTING (IPW) METHOD

**THEOREM 3**: Under SIA and where e(x) is the perfect propensity score, the ATE is given by:

$$\text{ATE} = (1/n) * \{ \sum_{i=1}^{n} y_i*t_i \,/\, e(x_i) - \sum_{i=1}^{n} y_i*(1-t_i) \,/\, (1 - e(x_i)) \}$$

where $y_i$ is the value of Y for subject "i", $t_i$ is the treatment T for subject "i", and n is sample size

This is a surprising result. After all, ATE involves E[Y(1)] and E[Y(0)] with missing values. Yet Theorem 3 gives a computable formula for ATE. The proof of Theorem 3 is given in Appendix A.

Importantly, Theorem 3 provides a formula for *estimating* ATE for a given sample from some larger or conceptual population, provided the modeler fits an empirical e(x) and is willing to assume SIA.

### WEIGHTS AND FILLING-IN MISSING VALUES IN THEOREM 3

Considering treated subjects with weight $w_1(x) = 1/e(x) = 3$, then (since e(x) is perfect [11]):

$$3 = (\text{total in sample with } w_1{=}3) \,/\, (\text{treated number in sample with } w_1{=}3).$$

Each treated subject with $w_1$=3 "stands for" 3 subjects from the total. The sum of the treated weights equals "n", the total sample count.[12] Likewise for $w_0(x) = 1 / (1-e(x))$. The treated group is viewed as a stratified sample from "n" treated subjects, and, similarly for control group … two conceptual data sets.

The weighted sum of treated outcomes $\sum_{i=1\ \&\ t=1}^{n} y_i*t_i * w_1(x_i)$ has n "weighted terms". In the sum, each weighted treated outcome $y_i*w_1(x_i)$ stands for the sum of "$w_1(x_i)$ actual and missing $y_i$" where $x=x_i$. In this sense, the weighting of the stratified sample "fills-in" the missing treated outcomes. Similarly for control.

### EXAMPLE: COMPUTING ATE USING THEOREM 3

A calculation of ATE using Theorem 3 is shown below. The reader may work through the calculations.

| Y | T | X | e(x) | Y*T | Y*T / e(x) | Y*(1-T) | Y*(1-T)/(1-e(x)) |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0.5 | | | 1 | 2 |
| 2 | 1 | 1 | 0.5 | 2 | 4 | | |
| 3 | 0 | 1 | 0.5 | | | 3 | 6 |
| 1 | 1 | 1 | 0.5 | 1 | 2 | | |
| 1 | 0 | 3 | 0.5 | | | 1 | 2 |
| 1 | 1 | 3 | 0.5 | 1 | 2 | | |
| 4 | 0 | 4 | 0.3333 | | | 4 | 6 |
| 2 | 0 | 4 | 0.3333 | | | 2 | 3 |
| 3 | 1 | 4 | 0.3333 | 3 | 9 | | |
| 2 | 0 | 4 | 0.3333 | | | 2 | 3 |
| 3 | 0 | 4 | 0.3333 | | | 3 | 4.5 |
| 2 | 1 | 4 | 0.3333 | 2 | 6 | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| n = | 12 | | | E[Y(1)] = | 1.917 | E[Y(0)] = | 2.208 |
| | | | | | ATE = E[Y(1)] - E[Y(0)] = | | -0.291 |

**Table 4: Calculation of ATE using Theorem 3**

### ALTERNATIVE, USING THEOREM 2

Since Strong Ignorability with respect to X implies Strong Ignorability with respect to e(x), the calculation of ATE for the data set of Table 4 may proceed as given in Table 5.

[11] If $e(x_1) = e_0$, then the treated observations account for $100*e_0$% of the subpopulation of S where $X=x_1$

[12] In Table 4 there are 3 treated subjects with $w_1 = 2$ and 2 treated subjects with $w_1 = 3$. So, 2 + 2 + 2 + 3 + 3 = 12
Generally, let weights be called $wgt_i$ where "i" indexes the distinct weight values.
Then $\sum_i$ (# treated with $wgt_i$) * $wgt_i$ = $\sum_i$ (# treated with $wgt_i$) * (# with $wgt_i$) / (# treated with $wgt_i$) = $\sum_i$ (# with $wgt_i$) = n

| e(x) | Average Y for T = 1 | Average Y for T = 0 | Difference "D" | % of Sample "S" | Weight = "D" * "S" |
|---|---|---|---|---|---|
| 0.5 | 1.333 | 1.667 | -0.333 | 0.5 | -0.166 |
| 0.333 | 2.500 | 2.750 | -0.250 | 0.5 | -0.125 |
| | | | | ATE = | -0.291 |

**Table 5: Alternative Calculation of ATE for the Data of Table 4**

The weights in Table 6 are the same as the weights of Theorem 3 (after dividing by n=12). The weights in Table 6 are implicitly used in the calculations of Table 5. The calculation in Table 5 gives the same results as in the formula of Theorem 3.

| e(x) | Divisor for T=1 | Divisor for T=0 | Fraction of Sample | Treated Weights ($w_1$ / n is the right side) | Control Weights ($w_0$ / n is the right side) |
|---|---|---|---|---|---|
| 0.5 | 1/3 | 1/3 | 1/2 | 1/3 * 1/2 = (1/0.5)*(1/12) | 1/3 * 1/2 = (1/(1-.05)) *(1/12) |
| 0.333 | 1/2 | 1/4 | 1/2 | 1/2 * 1/2 = (1/0.333)*(1/12) | 1/4 * 1/2 = (1/(1-0.333))*(1/12) |

**Table 6: The Implicit Weights Associated with Table 5**

## COMPUTE ATE BY TABLE 4 OR TABLE 5?

The method of Table 5 could be used to compute ATE when X is comprised of a very few discrete covariates so that there are only a few distinct values of e(x). But in actual practice there could be many distinct values of e(x) and there could be rows where there are subjects in treated and none in control, or conversely.[13] In this case the computations of Table 5 cannot be carried out, column D is "missing". (But then, unique values of e(x) could be replaced by strata. See Yuan, Yung, and Stokes (2017).)

## USING PROC TTEST

The weights for the IPW method imply a stratified sample is taken from the larger sample of treated subjects which consists of all actual and missing outcomes. Similarly for control. PROC TTEST can utilize these weights to estimate ATE (as the difference of two weighted means) and also provide a confidence interval. There is two-step logic for this confidence interval. First, the calculations of means and standard errors are computed using weights. Then the weighted sample is assumed to be drawn from an infinite population for the purpose of computing a 1 - α confidence interval:

$$\text{ATE} \;\; +/- \;\; t_{d.f.,\alpha/2} * \text{StdErr}$$

The d.f. (degrees of freedom) for the t-statistic are computed without weights.

## SUMMARY RELATED TO THEOREM 3

A disappointment with Theorem 3 is that the assumptions (a) and (b) are aspirational. Given a data sample, conditions (a) and (b) are not empirically verifiable. Probability P(Y(1)=y | X=x) cannot be computed due to missing values of Y(1). Similarly, probability P(Y(0)=y | X=x) cannot be computed.[14] So the properties, such as those below, of conditional independence cannot be verified.

- P(Y(1)=y | T=1, X=x) = P(Y(1)=y | X=x) for all y and x
- P(Y(0)=y | T=0, X=x) = P(Y(0)=y | X=x) for all y and x

[13] Normally, a propensity score model utilizes many covariates so that e(x) would have many distinct values.

[14] The conditions (a), (b) of SIA are satisfied in the following data set if hypothetical values, highlighted in yellow and italicized, are added to the table.

| X | T | Y | Y(1) | Y(0) |
|---|---|---|---|---|
| 1 | 0 | 2 | *7* | 2 |
| 1 | 0 | 4 | *3* | 4 |
| 1 | 1 | 3 | 3 | *2* |
| 1 | 1 | 7 | 7 | *4* |

To sum up, for the conclusions of Theorem 3 to be fully realized for an observational study, the modeler must be willing to believe that:

(1) The propensity score e(x) leads to covariate balance.

(2) Conditioning on X converts the study of the effect of T on Y to an experiment following the SIA.[15]

The modeler can come to believe (1) and (2) by assessing the balance achieved by the covariates between treated and control subjects and by using subject matter expertise.

### AN INAPPROPRIATE ASSUMPTION

Is there reason to consider the assumption Y _||_ T | X? There is no issue with missing data since Y is always observable. However, this assumption implies there is no treatment effect. That is, ATE equals zero. (The proof of this statement is not provided here.)

## PROC PSMATCH

The PSMATCH procedure provides tools for fitting a propensity score and then utilizing this propensity score to prepare an output data set for the estimation of the causal effect of the treatment using other SAS procedures. PSMATCH does not, itself, provide the estimation of the causal effect. The outcome variable is not even needed in the input data set for PSMATCH.

PSMATCH can prepare an output data set for further analysis by each of these three methods:

- Matching
- Stratification
- Weighting

Matching is heuristic and subjective. This is not to say that Matching is wrong. "Heuristic" is used since the matching of treated and control is based on "distances" between their propensity scores (or log-odds of the score) and there are many options for using the distances to implement the matching. It is subjective in that the success of the matching is judged by assessment of covariate balance. The range of propensity scores is, usually, restricted so that matching is done on treated and control subjects with propensity scores from a common range. Often, not all control subjects in the common range are used in matching.

The method of Stratification is now viewed as being superseded by Matching.

The method of Weighting is a formal statistical model-based approach that creates the inverse probability weights. (But PROC PSMATCH does not include the functionality to estimate ATE.) All treated and control observations can be used in the weighting method but the preferred approach is to remove subjects whose propensity scores are near 0 or 1.

In this paper the discussion of PSMATCH will limited to its implementation of the method of **weighting**. See Yuan, Yung, and Stokes (2017) for an introduction to propensity score methods and the PSMATCH procedure. Matching, stratification, and weighting are discussed in this paper.

In the example below PSMATCH is applied to the data of Table 4. The propensity score e(x), based on CLASS X or dummies X1 and X2, is perfect for this sample.[16]

```
DATA Table4;
input Y T X;
X1= (X=1); X2= (X=2);
datalines;
1 0 1
2 1 1
3 0 1
1 1 1
```

[15] See Rosenbaum (2010 pp. 86-87) for a roadmap to a discussion of this topic.

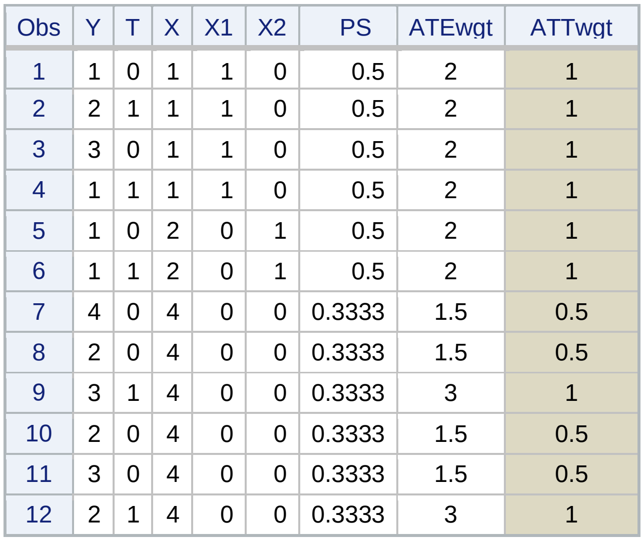
[16] The predictors X1, X2 (or the equivalent model CLASS X; PSMODEL T = X;) produce the perfect propensity score model. However, in practice, having a perfect propensity score model is very unlikely.

```
1 0 2
1 1 2
4 0 4
2 0 4
3 1 4
2 0 4
3 0 4
2 1 4
;
run;
```

The statement PSMODEL T = X1 X2 creates a propensity model with predictors X1 X2. In PSMODEL the statement (TREATED="1") specifies that T=1 is modeled. PSMATCH creates the weights for estimating ATE. These weights (ATEwgt) are included in OUTPUT OUT = OUTTable4. See Table 7.

```
PROC PSMATCH DATA= Table4;
CLASS T;
PSMODEL T(TREATED="1")= X1 X2;
OUTPUT OUT= OUTTable4 ATEwgt= ATEwgt;
PROC PRINT DATA= OUTTable4;
run;
```

If T = 1, then ATEwgt = 1 / _PS_; Else if T = 0, then ATEwgt = 1 / (1 - _PS_). The right-most column _ATTwgt_ is automatically added and it gives the weight for a different analysis called "average treatment effect for the treated" (ATT). The ATT is not discussed in this paper.



| Obs | Y | T | X | X1 | X2 | PS | ATEwgt | ATTwgt |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 1 | 0 | 0.5 | 2 | 1 |
| 2 | 2 | 1 | 1 | 1 | 0 | 0.5 | 2 | 1 |
| 3 | 3 | 0 | 1 | 1 | 0 | 0.5 | 2 | 1 |
| 4 | 1 | 1 | 1 | 1 | 0 | 0.5 | 2 | 1 |
| 5 | 1 | 0 | 2 | 0 | 1 | 0.5 | 2 | 1 |
| 6 | 1 | 1 | 2 | 0 | 1 | 0.5 | 2 | 1 |
| 7 | 4 | 0 | 4 | 0 | 0 | 0.3333 | 1.5 | 0.5 |
| 8 | 2 | 0 | 4 | 0 | 0 | 0.3333 | 1.5 | 0.5 |
| 9 | 3 | 1 | 4 | 0 | 0 | 0.3333 | 3 | 1 |
| 10 | 2 | 0 | 4 | 0 | 0 | 0.3333 | 1.5 | 0.5 |
| 11 | 3 | 0 | 4 | 0 | 0 | 0.3333 | 1.5 | 0.5 |
| 12 | 2 | 1 | 4 | 0 | 0 | 0.3333 | 3 | 1 |

**Table 7: Output Data Set from PSMATCH**

At this point the PSMATCH output is not displayed. This output will be discussed in a later section.

If it is assumed that Y(1) _||_ T | X and Y(0) _||_ T | X, then ATE can be estimated by computing the difference between the weighted treated mean and weighted control mean. PROC TTEST computes ATE by finding this difference with weights supplied by ATEwgt. The TC FORMAT puts the CLASS levels in the correct order.

```
PROC FORMAT; PICTURE TC
0 = "B_Control"
1 = "A_Treated";
PROC TTEST DATA= OUTTable4 ORDER= formatted;
CLASS T; VAR Y; WEIGHT ATEwgt;
FORMAT T TC.;
run;
```

Potential outcome means are 1.9167 for treated and 2.2083 for control. “Diff(1-2)” agrees with Table 4. The t-statistic two-sided tail probability for ATE is 62% (Table 8b).

| T | Method | Mean | 95% CL Mean | |
|---|---|---|---|---|
| A_Treated | | 1.917 | 0.863 | 2.971 |
| B_Control | | 2.208 | 1.169 | 3.247 |
| Diff (1-2) | Pooled | -0.292 | -1.575 | 0.991 |
| Diff (1-2) | Satterthwaite | -0.292 | -1.562 | 0.979 |

**Table 8a: PROC TTEST Reports**

| Method | Variances | DF | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Pooled | Equal | 10 | 0.51 | 0.6235 |
| Satterthwaite | Unequal | 9.9191 | 0.51 | 0.6198 |

**Table 8b: PROC TTEST Reports**

It is not true, however, that PROC PSMATCH followed by PROC TTEST replicates the formula of Theorem 3. The use of the formula of Theorem 3 would generally give a different estimate of ATE. Instead, the estimate of ATE that is provided by PROC PSMATCH followed by PROC TTEST is that same as that which would be provided by the METHOD = IPWR of PROC CAUSALTRT.[17] IPWR is “inverse probability weighting with ratio adjustment”. PROC CAUSALTRT is discussed in a later section.

## PSMATCH: COVARIATE ASSESSMENT REPORTS

The fit of the propensity model, as measured by a c-statistic or log-likelihood, is not reported by PSMATCH nor are the coefficients for the covariates. The purpose of the PSMATCH reports is to assess the balance achieved for the covariates after adjustment by weighting. Is the distribution of weighted X essentially the same for treated and control? Balance is assessed by analytical and graphical methods (graphical methods are not displayed here) when the ASSESS statement is included in PROC PSMATCH. [18]

In this example the PSMODEL statement has just one predictor, X. (Note here, that X is **not** in CLASS.)

```
PROC PSMATCH DATA=Table4;
CLASS T; /* Treatment T must be included in CLASS */
PSMODEL T(TREATED="1")= X;
ASSESS PS VAR=(X) / VARINFO WEIGHT=ATEwgt; /* PS = propensity score */
OUTPUT OUT=OUTTable4 ATEwgt=ATEwgt;
run;
```

[17] METHOD=IPW of PROC CAUSALTRT implements the formula of Theorem 3. METHOD=IPWR and METHOD=IPW both give approximately unbiased estimates of ATE (under SIA), assuming a good propensity score model. Their estimates of ATE would be similar for large samples. For details about METHOD=IPWR and METHOD=IPW of PROC CAUSALTRT, see SAS/STAT 14.3 User’s Guide, Chapter 34, p. 2166.

[18] If the modeler wants to fully manage the fitting of a propensity model, then the PSMODEL statement can be replaced by the PSDATA statement. Prior to running PSMATCH the modeler will have developed a propensity model using a modeling technique of the modeler’s choosing and saved the propensity scores as part of the input data set to PSMATCH.

In the PSDATA statement the modeler must specify these options: Treatment Variable, Treated Level for Modeling, and the name of the Propensity Score (or its Logits). The ASSESS statement still applies. The input data set must include any covariates that are specified in the VAR list of ASSESS.

Here is the code to replace PSMODEL with PSDATA. It is assumed that the propensity score has been added to the input data set “Table4” as the variable “e”.

```
PROC PSMATCH DATA= Table4;
CLASS T;
PSDATA TREATVAR=T(TREATED="1") PS= e;
ASSESS PS VAR=(X1 X2) / VARINFO WEIGHT=ATEwgt;
OUTPUT OUT=OUTTable4 ATEwgt=ATEwgt;
run;
```

Table 9 provides a summary report.

| Data Information | | |
|---|---|---|
| Data Set | WORK.TABLE4 | |
| Output Data Set | WORK.OUTTABLE4 | |
| Treatment Variable | T | |
| Treated Group | 1 | |
| All Obs (Treated) | 5 | |
| All Obs (Control) | 7 | |
| Support Region | All Obs | All observations for OUTPUT OUT= |
| Lower PS Support | 0.338401 | The minimum propensity score |
| Upper PS Support | 0.515201 | The maximum propensity score |
| Support Region Obs (Treated) | 5 | The support region is “All Obs” |
| Support Region Obs (Control) | 7 | The support region is “All Obs” |

In the reports below the rows for “Region” information have been removed since these are redundant with “All Obs”. The option VARINFO produces the table below (some columns were removed to fit the page).

| Variable Information | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Treated (T=1) | | | | Control (T=0) | | | | Treated - Control |
| Variable | Obs | N | Wgt | Mean | Std Dev | N | Wgt | Mean | Std Dev | Mean Difference |
| Prop Score | All | 5 | | 0.4323 | 0.0893 | 7 | | 0.4055 | 0.0861 | 0.0268 |
| | Wgted | 5 | 12 | 0.4169 | 0.0837 | 7 | 12 | 0.4169 | 0.0851 | 0 |
| X | All | 5 | | 2.4000 | 1.5166 | 7 | | 2.8571 | 1.4639 | -0.4571 |
| | Wgted | 5 | 12 | 2.6619 | 1.4233 | 7 | 12 | 2.6636 | 1.4464 | -0.0017 |

**Table 9: PSMATCH Reports** [19]

## HOW TO ASSESS COVARIATE BALANCE

The ASSESS statement for the example was:

```
ASSESS PS VAR=(X) /* PS = propensity score */
```

In Table 10 below there are two numeric measures for the assessment of balance. These measures will be discussed for the covariate X.[20] No discussion is given regarding PS in the following.

(1) Standardized Difference
(2) Variance Ratio

The definitions of Standardized Difference and Variance Ratio are given next:

### Standardized Difference of a Covariate

The Standardized Difference = Mean Difference / Standard Deviation … where:

- Mean Difference: Difference of the mean of treated subjects and the mean of control subjects
- Standard Deviation: The square root of { (Treated Variance + Control Variance) / 2 }

The standard deviation is computed using all observations in Table 4.

For covariate X the Standardized Difference is computed as shown:

For All Obs:

Mean Difference = 2.4000 - 2.8571 = -0.4571 … See Table 9 and 10
Standard Deviation = sqrt ( (1.5166**2 + 1.4639**2) / 2) ) = 1.4904 … See Table 9 and 10
Standardized Difference = -0.4571 / 1.4904 = -0.3067 … See Table 10

[19] The weighted means for X are given by: Treated subjects ∑ x*w1 / ∑ w1 and control ∑ x*w0 / ∑ w0
For discussion see: https://blogs.sas.com/content/iml/2016/01/06/weighted-mean-in-sas.html
For other formulas to compute weighted statistics, including the standard deviation, see:
http://support.sas.com/documentation/cdl/en/procstat/68142/HTML/default/viewer.htm#procstat_univariate_details03.htm

[20] Additionally, other covariates not used in fitting the propensity score could also be included in ASSESS.

For Weighted Obs:

Mean Difference = 2.6619 - 2.6636 = -0.0017 … See Table 9 and 10
Standardized Difference = -0.0017 / 1.4904 = -0.0012 … See Table 9 and 10
Percent Reduction = 100% * (| -0.3067 | - | -0.0012 |) / | -0.3067 | = 99.62%

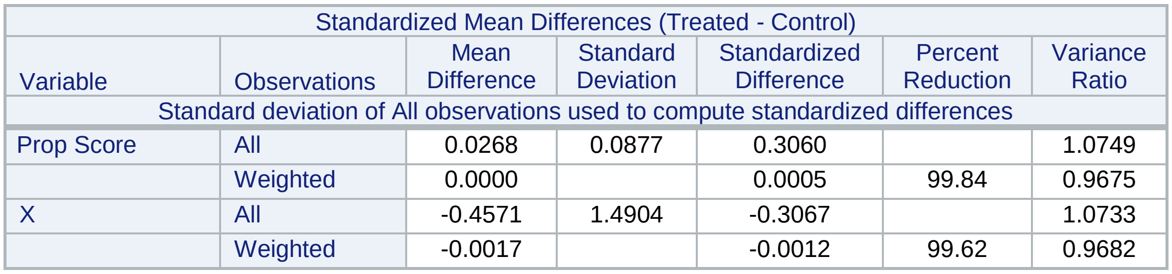
**Variance Ratio of a Covariate: {Standard Deviation (Treated) / Standard Deviation (Control) }**2**

For covariate X for All Obs … See Table 9 above

{Standard Deviation (Treated) / Standard Deviation (Control) }**2 = {1.5166 / 1.4639}**2 = 1.0733

For covariate X for Weighted Obs … See Table 9 above

{Standard Deviation (Treated) / Standard Deviation (Control) }**2 = {1.4233 / 1.4464}**2 = 0.9682



| Standardized Mean Differences (Treated - Control) | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Observations | Mean Difference | Standard Deviation | Standardized Difference | Percent Reduction | Variance Ratio |
| Standard deviation of All observations used to compute standardized differences | | | | | | |
| Prop Score | All | 0.0268 | 0.0877 | 0.3060 | | 1.0749 |
| | Weighted | 0.0000 | | 0.0005 | 99.84 | 0.9675 |
| X | All | -0.4571 | 1.4904 | -0.3067 | | 1.0733 |
| | Weighted | -0.0017 | | -0.0012 | 99.62 | 0.9682 |

**Table 10: PSMATCH Covariate Assessment Reports**

Documentation for PSMATCH version 14.3 gives guidelines for acceptable Standardized Mean Difference and Variance Ratio.[21] The guidelines are given below and are easily met by weighted X.

Absolute Value of Standardized Mean Difference ≤ 0.25

0.5 ≤ Variance Ratio ≤ 2.0 … with 1.0 being the perfect variance ratio

## REMOVING SUBJECTS WITH HIGH OR LOW PROPENSITY SCORES

But there is a problem when an empirical propensity score model has values of e(x) near 0 or 1. In this case the weight w1 for treated or the weight w0 for control can be very large. In applications, the estimate of ATE could be distorted. This leads to the idea removing subjects where e(x) is near 0 or 1. Such removal can be implemented using PSMATCH.

```
DATA REMOVE;
DO I = 1 TO 500;
   X= rannor(1);
   XBETA= X + 0.5* rannor(4);
   PROB= EXP(XBETA)/(1 + EXP(XBETA));
   T= (PROB < 0.5);
   OUTPUT;
   END;
run;
```

The REGION statement with ALLOBS is qualified by PSMIN=0.05 and PSMAX=0.95. Only subjects where 0.05 ≤ Propensity Score ≤ 0.95 are output to OUTREMOVE. This in example, there were 500 subjects used in fitting the propensity score model but only 304 had scores within the range of 0.05 to 0.95.

```
PROC PSMATCH DATA=REMOVE REGION= ALLOBS (PSMIN=.05 PSMAX=.95);
CLASS T;
PSMODEL T(TREATED="1")= X;
OUTPUT OUT= OUTREMOVE ATEwgt= ATEwgt;
run;
```

[21] See https://support.sas.com/documentation/onlinedoc/stat/142/psmatch.pdf, page 7714

| Data Information | |
|---|---|
| Data Set | WORK.REMOVE |
| Output Data Set | WORK.OUTREMOVE |
| Treatment Variable | T |
| Treated Group | 1 |
| All Obs (Treated) | 246 |
| All Obs (Control) | 254 |
| Support Region | PS Bounded Obs |
| Lower PS Support | 0.05 |
| Upper PS Support | 0.95 |
| Support Region Obs (Treated) | 154 |
| Support Region Obs (Control) | 150 |

**Table 12: PSMATCH Report Showing Effect of Removal**

## PROC CAUSALTRT

PROC CAUSALTRT combines some of the functionality of PROC PSMATCH together with analytical methods to compute the causal effect for an observational study. This paper only very briefly discusses PROC CAUSALTRT. See Lamm and Yung (2017) for an introduction to PROC CAUSALTRT.

An example of PROC CAUSALTRT is given below. See the PROC CAUSALTRT code.

When METHOD=IPWR the “outcome model” (see MODEL statement) is not fit (even if predictors are provided in the MODEL statement). The statement MODEL Y is required in order to identify the outcome variable Y. The PSMODEL statement fits the logistic propensity score model CLASS X; PSMODEL T= X; Then ATE is computed by following a formula similar to that of Theorem 3.[22] An output data set is, optionally, created.

```
PROC CAUSALTRT DATA= Table4 METHOD= IPWR;
CLASS X;
PSMODEL T (descending)= X; /* PROPENSITY MODEL */
MODEL Y; /* OUTCOME MODEL */
BOOTSTRAP BOOTCI(NORMAL) SEED= 1234;
OUTPUT OUT= causalout IPW= IPW PS= PS;
run;
```

| Analysis of Causal Effect | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Para meter | Treat ment Level | Est- imate | Robust Std Err | Boot strap Std Err | Wald 95% Confidence Limits | | Bootstrap Wald 95% Confidence Limits | | Z | Pr > \|Z\| |
| NOTE: 889 out of 1000 bootstrap samples are used to calculate standard errors. | | | | | | | | | | |
| POM | 1 | 1.917 | 0.275 | 0.331 | 1.377 | 2.456 | 1.267 | 2.566 | 6.96 | <.0001 |
| POM | 0 | 2.208 | 0.364 | 0.379 | 1.496 | 2.921 | 1.466 | 2.951 | 6.07 | <.0001 |
| ATE | | -0.292 | 0.382 | 0.428 | -1.041 | 0.458 | -1.131 | 0.547 | -0.76 | 0.446 |

**Table 13: PROC CAUSALTRT Reports**

Z-stats are computed using the Robust Std Err’s. There are two confidence intervals computed by CAUSALTRT, one using Robust Std Err and the other using bootstraps.[23] In Table 14 these are compared

[22] For details about METHOD=IPWR of PROC CAUSALTRT, see SAS/STAT 14.3 User’s Guide, Chapter 34, p. 2166.
[23] See SAS/STAT 14.3 User’s Guide, Chapter 34, p. 2149.

with the confidence interval produced by PROC TTEST (taken from Table 8a).

For this small sample there is a noticeable difference in 95% confidence intervals for ATE between PROC TTEST and the two confidence intervals from CAUSALTRT. (Further study of this issue is merited.)

| METHOD | 95% confidence interval for ATE |
|---|---|
| Robust Std Err | -1.041 to 0.458 |
| Bootstrap Std Err | -1.131 to 0.547 |
| TTEST (see Table 8) | -1.562 to 0.979 |

**Table 14: Comparison of Confidence Intervals**

## SUMMARY

Conditional independence is illustrated by examples. Conditional independence of outcome and treatment, given covariates, is a central concept in analysis of observation studies. Simple proofs are given of key theorems including the IPW method. Calculations of ATE by inverse probability weighting are illustrated by applying PROC PSMATCH / PROC TTEST and PROC CAUSALTRT to a small data set.

*Indianapolis, IN, MWSUG 2018, version 26*

## RECOMMENDED REFERENCE RESOURCES

Angrist, J. D. and Pischke, Jörn-Steffen (2009), *Mostly Harmless Econometrics*, Princeton, NY, Princeton University Press. See: Chapters 1, 2, 3

Davidian, M. (2007), Double Robustness in Estimation of Causal Treatment Effects, lecture notes, https://www4.stat.ncsu.edu/~davidian/double.pdf (last tested 6/23/2018).

Lunceford, J.K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. https://www4.stat.ncsu.edu/~davidian/statinmed.pdf (last tested 7/3/2018). See pages 1-10. Published in *Statistics in Medicine* 23:2937–2960.

## REFERENCES

Lamm, M., and Yung, Y.-F. (2017). “Estimating Causal Effects from Observational Data with the CAUSALTRT Procedure.” In Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc. http: //support.sas.com/resources/papers/proceedings17/SAS0374-2017.pdf.

Rosenbaum, P. (2010), *Design of Observational Studies*, New York, NY, Springer.

Yuan, Y., Yung, Y.-F., and Stokes, M. (2017). “Propensity Score Methods for Causal Inference with the PSMATCH Procedure.” In Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc. http: //support.sas.com/resources/papers/proceedings17/SAS0332-2017.pdf.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bruce Lund
OneMagnify, Detroit, MI
blund_data@mi.rr.com, blund.data@gmail.com

All code in this paper is provided by OneMagnify "as is" without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose. Recipients acknowledge and agree that OneMagnify shall not be liable for any damages whatsoever arising out of their use of this material. In addition, OneMagnify will provide no support for the materials contained herein.

## APPENDIX A

**THEOREM 3**: Under SIA and where e(x) is the perfect propensity score, the ATE is given by:

$$(1/n) * \{ \sum_{i=1}^{n} y_i*t_i / e(x_i) - \sum_{i=1}^{n} y_i*(1-t_i) / (1 - e(x_i)) \}$$

where $y_i$ is the value of Y for subject "i", $t_i$ is the treatment T for subject "i", and n is sample size

**Proof of THEOREM 3** [24]

Begin with $(1/n) * \{ \sum_{i=1}^{n} y_i*t_i / e(x_i) \}$

Note that: $y*t = \{ y(1)*t + y(0)*(1-t) \} * t = y(1)*t$ … since $t^2 = t$ and $t*(1-t) = 0$ for values of t = 0, 1

This gives: $(1/n) * \{ \sum_{i=1}^{n} y_i*t_i / e(x_i) \} = (1/n) * \{ \sum_{i=1}^{n} y_i(1)*t_i / e(x_i) \}$

The probability of the ith subject ($i^{th}$ term in sum) is $1/n = P( Y(1)=y_i, T= t_i )$. This gives:

$(1/n) * \{ \sum_{i=1}^{n} y_i(1)*t_i / e(x_i) \} = \sum_{i=1}^{n} y_i(1)*t_i * P( Y(1)=y_i(1), T= t_i ) / e(x_i)$

Let a unique value of e(x) be denoted by $e(x_j)$ and let $E_j = \{x: e(x) = e(x_j)\}$. The $\{E_j\}$ partition the values of X into disjoint subsets. It does not matter which "$x_j$" is chosen to represent a partition member. In the example data set there are 2 unique $e(x_j)$ … 0.5 and 0.333 … these will be called $E_1$ and $E_2$.

$= \sum_{Ej} \sum_{i:xi \in Ej} y_i(1) * t_i * P( Y(1)=y_i(1), T= t_i ) / e(x_j)$

| Y | T | Y(1) | X | e(x) |
|---|---|---|---|---|
| 1 | 0 | | 1 | 0.5 |
| 2 | 1 | 2 | 1 | 0.5 |
| 3 | 0 | | 1 | 0.5 |
| 1 | 1 | 1 | 1 | 0.5 |
| 1 | 0 | | 3 | 0.5 |
| 1 | 1 | 1 | 3 | 0.5 |
| 4 | 0 | | 4 | 0.333 |
| 2 | 0 | | 4 | 0.333 |
| 3 | 1 | 3 | 4 | 0.333 |
| 2 | 0 | | 4 | 0.333 |
| 3 | 0 | | 4 | 0.333 |
| 2 | 1 | 2 | 4 | 0.333 |

**Table 15: Example Data Set for the Proof of Theorem 3**

In the example, the outside sum has two terms. Within the first term of the outside sum, the inside sum has 3 non-zero terms (blue colors in Table 15). For the second outside term there are 2 non-zero terms in the inside sum (pink colors in Table 15).

$= \sum_{Ej} \sum_{i:xi \in Ej} y_i(1) * P( Y(1)=y_i(1), T= 1 ) / e(x_j)$ + Zero-terms (those where t = 0)

The expansion is shown below:

= 2*(1/12)*(1/0.5) + 1*(1/12)*(1/0.5) + 1*(1/12)*(1/0.5) + 3*(1/12)*(1/0.333) + 2*(1/12)*(1/0.333) + 0's

Going forward, the zero terms are ignored.[25]

Condition (c): $0 < P(T=1 \mid X=x) < 1$ insures that the same $E_j$ and $x_i \in E_j$ appear in the summation where T=1 (below) as in the sum with all T values. That is, $\sum_{Ej} \sum_{i:xi \in Ej}$ is the same index of summation.

Next, group together common y(1) within each $E_j$.

The index "$y(1){:}\,Ej$" is read as "values of y(1) in the rows which occur in $E_j$"

$= \sum_{Ej} \sum_{y(1):Ej} y(1) * P( Y(1)=y(1), T=1 ) / e(x_j)$

[24] Here is a proof from Angrist and Pischke (2009, p. 82) of $(1/n) * \{ \sum_{i=1}^{n} y_i*t_i / e(x_i) \} = E[Y(1)]$. [Alternatively, see proof by Lunceford and Davidian (2004, on-line PDF version, top of p. 7).]

First: E[Y*T/e(x)] = E{ E[Y*T/e(x) | X=x] }.
Taking inside expectation of the right-side: E[Y*T/e(x) | X=x] = E[Y | T=1, X] * e(x) / e(x) = E[Y(1) | T=1, X] = E[Y(1) | X]
Finally, E{ E[Y(1) | X] } = E[Y(1)]

[25] The apparent loss of information about y(1) where t=0 is compensated by the use of 1/e(x) as a weight when t=1.

The probabilities in $E_1$, after grouping, are P( Y(1)=1, T=1 ) = 2/12 and P( Y(1)=2, T=1 ) = 1/12

Including $E_j$ into the Probability adds no new restriction

= $\sum_{\mathrm{Ej}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1), T=1, $E_j$ ) / e($x_j$) … $E_j$ means x∈$E_j$ where x is a value of X.

Applying the formula for conditional probability: [P(A,B,C] = P(A|B,C)*P(B,C)]

= $\sum_{\mathrm{Ej}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1) | T=1, $E_j$ ) * P(T=1, $E_j$ ) / e($x_j$)

Now the formula for the sum of probabilities of disjoint sets is applied to P( T=1, $E_j$ )

P( T=1, $E_j$ ) = $\sum_{\mathrm{y(1):Ej}}$ P( T=1, X=$x_i$ ) = $\sum_{\mathrm{y(1):Ej}}$ P( T=1 | X=$x_i$ ) * P( X=$x_i$ ) = e($x_j$) * P( $E_j$ )

Substituting into the entire expression and canceling e($x_j$) from top and bottom:

= $\sum_{\mathrm{Ej}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1) | T=1, $E_j$ ) * P( $E_j$ )

Finally, conditional independence is applied to remove T=1 from P( Y(1)=y(1) | T=1, $E_j$ )

= $\sum_{\mathrm{Ej}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1) | $E_j$ ) * P($E_j$ ) … all references to T are removed

= $\sum_{\mathrm{Ej}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1), $E_j$ )

The summation is over all rows of the original data set. Rows where t=0 contribute zeros.

Table 16 shows 4 non-zero terms in y(1) * P( Y(1)=y(1), $E_j$ ) that comprise the double summation.

| Y | T | Y(1) | X | e(x) | P(e(X)) | P(Y=y \| e(X)) | P(Y=y, e(X)) | Y * P(Y=y, e(X)) |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 |  | 1 | 0.5 |  |  |  |  |
| 2 | 1 | 2 | 1 | 0.5 | 0.5 | 0.333 | 0.167 | 0.333 |
| 3 | 0 |  | 1 | 0.5 |  |  |  |  |
| 1 | 1 | 1 | 1 | 0.5 | 0.5 | 0.667 | 0.333 | 0.333 |
| 1 | 0 |  | 3 | 0.5 |  |  |  |  |
| 1 | 1 | 1 | 3 | 0.5 | 0.5 |  |  |  |
| 4 | 0 |  | 4 | 0.333 |  |  |  |  |
| 2 | 0 |  | 4 | 0.333 |  |  |  |  |
| 3 | 1 | 3 | 4 | 0.333 | 0.5 | 0.5 | 0.250 | 0.750 |
| 2 | 0 |  | 4 | 0.333 |  |  |  |  |
| 3 | 0 |  | 4 | 0.333 |  |  |  |  |
| 2 | 1 | 2 | 4 | 0.333 | 0.5 | 0.5 | 0.250 | 0.500 |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | Total = |  | 1.917 |

**Table 16: Calculation of ATE following the formula of Theorem 3**

For $E_1$ where e(x) = 0.5 the inner sum is: 1*0.333 + 2*0.167 … where y(1) = 1, 2

For $E_2$ where e(x) = 0.333 the inner sum is 2*0.25 + 3*0.25 … where y(1) = 2, 3

Reverse the order of summation. The outside sum is over values of y(1) and, for a value of y(1), the inside sum is over the disjoint sets $E_j$ where this y(1) appears.

= $\sum_{\mathrm{y(1)}} \sum_{\mathrm{y(1):Ej}}$ y(1) * P( Y(1)=y(1), $E_j$ )

For this example: 1*0.333 + (2*0.167 + 2*0.25) + 3*0.25 = 1.917

Summing probabilities across the $E_j$ for common values of y(1) gives marginal probabilities.

= $\sum_{\mathrm{y(1)}}$ y(1) * P( Y(1)=y(1) ) … potential treatment outcomes y(1), where t=0, contribute zero to the sum.

1*0.333 + 2*(0.167+0.25) + 3*0.25 = 1.917

So, finally: $\sum_{\mathrm{y(1)}}$ y(1) * P( Y(1)=y(1) ) = E[Y(1)]

This sum is over all values of y(1) as required by the definition of E[Y(1)]. The factors P(Y(1)=y(1)) are not dependent on the treatment T. Potential treatment outcomes y(1), where t=0, contribute zero to the sum.

Still seems magical.

The proof that $\sum_{i=1}^{n}$ $y_i$*(1-$t_i$) / (1 - e($x_i$)) } = E[Y(0)] is similar.

## APPENDIX B

In this example of a causal analysis the covariate set will omit a covariate which is strongly related to both the treatment and the outcome.

The treatment has two values (with non-random selection into the treatment):

T=0: Attending State University
T=1: Attending Ivy University

The outcome Y will be the graduate's annual income when the graduate is age 35.

Ivy University has high tuition.

Does Ivy University improve the graduates outlook for having higher income at age 35?

The two covariates to support this analysis are:

X1: Gender
X2: Religion

It is certainly true that comparisons of the outcome across the treatments could be made for subgroups of subjects with exactly the same values of X1 and X2.

However: X3 = Parent's Income is not included. X3 is strongly related both to treatment (high tuition requires high parent income) and Outcome (high parent income leads to high paying jobs / businesses for the graduate).

So the "treatment effect" may be nothing other than measuring the effect of parent income on the outcome, and not the effect of the university. Expressed in the framework of SIA the condition (a) would fail:

$$P(Y(1)=\text{high} \mid T=1, X=(x1, x2)) > P(Y(1)=\text{high} \mid X=(x1, x2))$$

Having T=1 (Ivy U.) implies high parent income as well as high graduate income at age 35 … so a high value of P(Y(1)=high | T=1, X=(x1, x2)).

But for the right side of the inequality, there are some subjects with T=1 and some with T=0. Parent Income would be lower for T=0. So the potential treated outcome would tend to be lower … so a lower value for P(Y(1)=high | X=(x1, x2)).