SESUG Paper HOW213-2019

Introduction to Propensity Score Modeling and Treatment Effect Estimation

Jason Brinkley, Abt Associates

ABSTRACT

Measuring the impact of treatments and interventions is an important aspect to all areas of evaluation and research. While a randomized experiment or clinical trial may be the gold standard for providing causal information regarding effectiveness, it is often the case that we must rely on observational data and secondary sources for measuring effectiveness. Confounding plays a huge role in measuring and evaluating effectiveness and statistical adjustments for confounding are a focal point of modern analytics. In many cases it is not enough to perform simple adjustments via outcome regression models to control for confounding, especially in cases where such confounding plays a role in the very selection of who gets which intervention of interest. Propensity score models have become a popular route for creating balance in observational data by implementing a pseudo or quasi-experimental set of conditions on the data of interest. This workshop is designed to give a soft introduction and background to propensity score modeling and associated treatment effect estimation. We will discuss the motivation for counterfactual data analysis and apply real world data (from the medical literature) to explore how to use the PSMATCH and CAUSALTRT procedures in SAS. Attendees need only have a working knowledge of multiple regression analyses.

INTRODUCTION

The gold standard for assessing causal effects is the randomized experiment. However, there arise situations in which assessments of causality are needed from data that are observational in nature. The area of causal inference has taken shape from this need, and while there are many different approaches to causal inference, most methods form from the same basic premises.

Cofounding is the fundamental problem in observational studies. Without the benefit of experimental control it is unclear at times whether the associations that are seen in observational studies illustrate a causal relationship. In fact we often say that association does not imply causation. However, in what cases can we reasonable make that leap?

THOUGHT EXPERIMENT

Suppose I have access to a large observational database that houses a registry of treatment information for patients with a particular disorder. Think of the big ones: Cancer, Heart Disease, etc. In these large databases we have lots of data on patients, which treatments they received (therapies, surgeries, etc.) as well as patient demographics, family history, biomarkers, etc. Is it possible that with thorough observational data we can talk about causal effect of specific interventions or therapies?

In a study like this, suppose we were able to see the outcome of interest for the same patient who was both given the intervention and denied the intervention at the same time. If everything else was held constant then the only difference in the two outcomes would depend only on the choice to give or withhold that intervention. So in essence the conclusion would be causal in nature.

COUNTERFACTUALS

For this paper we will work with a naïve example. Suppose there is binary response variable Y (success or fail, death or no death, etc.); further suppose there is a binary treatment/intervention of interest T (T=0 is standard treatment, T=1 is novel treatment)

Suppose there is a constellation of covariates X, which could be a mix of categorical or quantitative variables.

Now imagine two hypothetical worlds:

- In world 0 (subscripted as $_0$ onward), everyone who needs treatment receives treatment 0.
- In world 1 (subscripted as $_1$ onward), everyone who needs treatment receives treatment 1.

In fact everything up to the moment when treatment was decided was exactly the same and the only difference is that all patients get one treatment or the other.

In reality what we have is a mixture of these two worlds. Patients either got one treatment or the other. Now let's make a basic assumption: the response that is observed on a person is the same as the response in the hypothetical world where everyone got the treatment the patient received in the observed world.

That is to say that if T=0 then $Y = Y_0$ and T=1 then $Y = Y_1$.

We call this the Stable Unit Treatment Value Assumption and write it mathematically as:

$$\boldsymbol{Y = TY_1 + (1 - T)Y_0}$$

We call the quantities (Y_0, Y_1) potential outcomes or counterfactuals. They are potential outcomes because only one can be realized and the other value is counter to what we actually find in the data. However, from a philosophical standpoint we can boil the problem of confounding and observational studies down to a missing data problem.

The problem here is we only get to observe one potential outcome on each individual. If we could observe both values on everyone then we could make causal conclusions about the effect of treatment. It turns out that if we make one more assumption then we can make everything shake out in terms of expectations.

But in order to do that we need to make a very strong assumption about the structure of our observed data. No unmeasured confounding (or strong ignorability) assumption states that treatment assignment to an individual is independent of potential outcomes (Y_0 , Y_1) given X.

In an observational study the treatment that a physician chooses to give a patient can be reasonably assumed to only be based on characteristics of the patient at the time of treatment and not on the patient's potential outcome (which of course is not known at the time of treatment).

Consequently, this second assumption will be tenable if the key factors influencing the decision making process for a physician was captured in the data X that were collected. If, however, there are additional factors beyond the data X that influence treatment decisions, then this assumption may not hold.

So what does this get us?

$$\begin{aligned}\boldsymbol{P(Y_1 = 1)} &= \boldsymbol{E_X\{E(Y_1)|\mathrm{X}\}} \\ &= \boldsymbol{E_X\{E(Y_1)|\mathrm{X},\ \mathrm{T} = 1\}} \quad \textbf{(STRONG IGNORABILITY)} \\ &= \boldsymbol{E_X\{Y|\mathrm{X},\ \mathrm{T} = 1\}} \quad \textbf{(SUTVA)}\end{aligned}$$

Thus we can denote a causal quantity in terms of observed data:

Average Treatment Effect = ATE = $\boldsymbol{P(Y_1 = 1) - P(Y_0 = 1) = E_X\{Y|\mathbf{X},\ \mathbf{T} = 1\} - E_X\{Y|\mathbf{X},\ \mathbf{T} = 0\}}$

The last quantity is something that we can estimate based on data, and is a focal point of the estimation routines in the CASUALTRT procedure. The interested reader should see Rubin (1974) and Jewell (2004) for a more thorough introduction to this topic.

PROPENSITY SCORES

Long before the ideas of counterfactual data analysis were formed, others were still wrestling with ways to deal with the biases in observational data of the sort described in the earlier treatment example. The major problem (which still exists) is without experimental control, there is a real possibility of having inherent biases in observational databases. For example if sicker patients are all given a particular treatment and they have a poor outcome, can we attribute that effect to the treatment or to the fact the patients were very sick?

Propensity scores have emerged as a popular way to help ‘adjust’ away some of the potential biases that arise from observational data.

For the binary T, Y and vector X scenario discussed earlier define the propensity score as

$$\boldsymbol{\pi_T}=\mathrm{P(T=1|X)}$$

Why model treatment selection?

The general theory goes that two patients with similar propensity scores are somehow more ‘alike’ than those with very different propensities. We calculate these propensities and then try to use them to bring a sort of ‘balance’ to the sample. The examples below will demonstrate the utility of propensity scores, but conceptually this analysis makes a better adjustment then just regression modeling for adjusted estimates along can make. Think of evaluating a medical procedure (for example - surgery) that might be lifesaving, it is possible that older patients are more likely to die of the condition for which the procedure is looking to provide protection. However, it is also possible that the procedure has risks and those risks increase with age. Therefore, evaluating the effectiveness of the procedure adjusting for age in an outcome only framework will only adjust for the added risk of a poor outcome and not the bias in treatment selection. If a patient is more likely to die because they are older but also less likely to receive surgery because they are older then looking at death rates between surgery and non-surgery patients adjusted for age alone will not also adjust for the fact that older patients were less likely to receive the surgery.

Ways we use propensity scores:

- Stratification. Take your data and estimate the propensity scores for treatment. Then divide the sample into propensity based strata (usually quintiles) and do the analysis of treatment versus outcome (adjusted for covariates) for each quintile. Differences in the results among quintile groups may be attributed to observational study biases and help elucidate the effect of treatment on response.

- Matching – split the data into treatment 0 and 1 groups. Match them by propensity score. What is the difference in outcomes between ‘similar’ individuals whose difference is in treatment? This can be especially hard if matches are hard to come by.
- Weighting – We sometimes use the inverse of the propensity score as weights for analyses. This can be a useful mechanism in cases where the data don’t have an adequate sample for matching purposes. We tend to look for ‘overlap’ between treatment 0 and 1 groups as a basis for matching and when that overlap is low, weighting can be a good option.

- Control – Use the propensity score as another covariate in your analyses. It can be done but is usually not the preferred route, as it does little to mitigate observational biases.

Guo and Fraser (2015) provide an excellent text on working with Propensity Scores; some of the examples will be used for this paper.

OTHER CONSIDERATIONS

There is a standard problem of using propensities that are very small because it is difficult to find matches or inverse weights get very large. So looking at the statistical distribution of propensity scores can be very important in their appropriate use. In addition, there are many ways by which propensity scores can be used for matching and calibration and there is no one set of rules that say when to use which technique. The examples used here follow an exploratory approach, try a few different matching or estimation techniques and see which ones fit the data the best.

IMPLEMENTATION IN SAS®

Beginning in SAS/STAT version 14.2, the PSMATCH and CAUSALTRT procedures were made available for SAS users to implement causal inference. While the procedures have many different features, the two most important for this paper are the PSMATCH procedure's ability to perform propensity score matching and weighting and the CAUSALTRT procedure's ability to perform ATE calculations. For ease of discussion we will limit our examples to binary interventions and binary outcome measures. This means that the general implementation of these methods will be based on logistic regression models; however, this need not be the case.

PSMATCH PROCEDURE

The PSMATCH Procedure allows for the creation of propensity score matching. There are a variety of ways that this can be done. In general, PSMATCH code implementation is written as follows:

```
proc psmatch data=*dataset*;
class *class variables including treatment*;
psmodel *treatment = covariates*;
match distance = *measure* method=*method* *other options*;
output out=*output* *other options*;

run;
quit;
```

For those that are familiar with other modeling procedures such as the LOGISTIC Procedure, the format is very similar. We specify a class statement with all of our categorical variables, with a key difference being that the treatment (i.e. the outcome variable in our propensity model) is also included. The psmodel statement is structured as a traditional model statement with response as a combination of covariates of interest. The match statement is where the procedures real work is done and here there are two main options; the distance command tells SAS which metric to use for matching. One could match by propensity score, log propensity score, or Mahalanobis distance (not discussed here). The method command breaks down into three groupings (summarized below but more details can be found here):

1. Greedy Matching – Distance measures are sorted and each treated individual is matched with the nearest untreated individual with similar propensity score. Designed to allow for the most matches between groups.
2. Replacement Matching – Each treated individual is matched with an untreated individual whose distance score is *close* to the treated individual. Close is defined by other options; but an untreated individual may be matched to multiple treated individuals since the matches are done with replacement. Designed to allow for the most matches to treated individuals.
3. Optimal Matching – Designed to match treatment and control groups simultaneously to minimize distance measure for some criteria. Criteria specified in other options.

Traditionally, greedy matching has been a very popular mechanism for matching as it tends to produce the largest set of matches without replacement. Replacement matching is popular when it is desired to keep as many treated individuals as possible in the analysis. Optimal matching may serve best when both the treated and untreated groups are important for analysis and it is known that overlap may be poor and many low and/or high propensity scores will not have matches. See the example below.

The output statement outputs a new dataset that (depending on options) will give the original observations, which observations were matched together, propensity scores, and inverse propensity scores that can be used as weights in follow-up analysis.

CAUSALTRT PROCEDURE

The CAUSALTRT Procedure allows for the estimation of measures such as average treatment effect. There are a variety of ways that this can be done. In general, CAUSALTRT code implementation is written as follows (more details can be found here):

```
proc causaltrt data=*dataset* *other options*;
class *class variables*;
model *outcome = covariates without treatment*;
psmodel *treatment = covariates*;
run;
quit;
```

The procedure fits two sets of models, an outcome regression model and a propensity score model and will use the two to produce measures such as ATE. There are a variety of estimators for ATE, each with significant theoretical implications. Lamm and Yung (2017) have a full discussion of these estimators but they tend to break down into three types:

1. Outcome regression model based adjustment with no propensity score balancing
2. Propensity score adjustment using inverse propensity weighting
3. *Doubly-Robust* adjustment (also known as Augmented Inverse Propensity Weight adjustment) that uses a combination of outcome regression and propensity score modeling to estimate causal effects.

Outcome regression only does not benefit from the gains in using a propensity score while inverse propensity weighting can be sensitive to extreme propensity scores. Augmented Inverse Propensity Weight (AIPW) adjustment is a popular route because using both the outcome and propensity models offer two opportunities at robust and unbiased estimation of intervention impact. However, AIPW estimation is inefficient and will usually have larger standard errors because of the trade-off in variance for additional protection from bias.

EXAMPLE 1 -

For the first example we will generate some synthetic data to illustrate the need for adjusted treatment effect considerations. The code below can be manipulated in many ways but the point is that we will generate an outcome variable, Y, from a logistic model that has four inputs: X1 and X2 that are continuous normal variables with mean 50, standard deviation 10, and a common correlation of 0.25. In addition an uncorrelated covariate that is binary with chance of success 0.7 is generated and an intervention variable, E, that is a logistic model of X1, X2, and X3. The betas for our logit model are shown in the code below. The user may feel free to tweak any of the numeric inputs for a scenario where X1, X2, and X3 have more or less impact on intervention E or outcome Y. These inputs were chosen to be illustrative here and it is expected that in a full demonstration that the parameters would be moved around some to illustrate how the relationship to covariates and distribution of propensities would impact treatment estimation or propensity score matching.

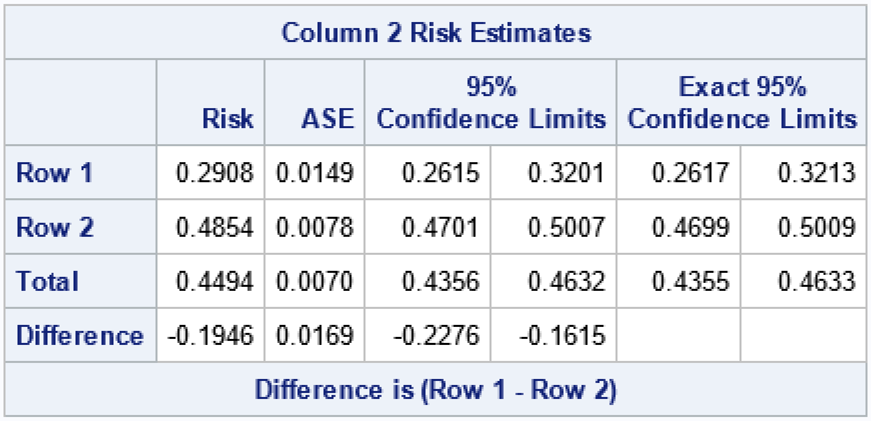
```
Data Example1;
call streaminit(34567);
keep ID x1 x2 x3 E Y;
mu1=50; mu2=50; var1=100; var2=100; rho=.25;
do ID = 1 to 5000;
x1 = rand("Normal", 0,1);
x2 = rho*x1+sqrt(1-rho**2)*rand("Normal",0,1);

x1 = mu1 + sqrt(var1)*x1;
x2 = mu2 + sqrt(var2)*x2;
x3 = rand("Bernoulli", .7);

*Exposure Modeling;
p1 = .5;
m1 = exp(-2 + .01*x1 + .04*x2 + 1.8*x3);
p1 = m1/(1+m1);
E = rand("Bernoulli", p1);

*Outcome Modeling;
m2=exp(-6 + (.5*E) + (.05*x1) + (.05*x2) + .5*x3);
p2 = m2/(1+m2);
Y = rand("Bernoulli", p2);
output;
end;
Run;
```

From here a simple Proc Freq command with a risk difference table statement yields the following intervention by outcome rate comparison.



Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.2908	0.0149	0.2615	0.3201	0.2617	0.3213
Row 2	0.4854	0.0078	0.4701	0.5007	0.4699	0.5009
Total	0.4494	0.0070	0.4356	0.4632	0.4355	0.4633
Difference	-0.1946	0.0169	-0.2276	-0.1615		
Difference is (Row 1 - Row 2)						

Figure 1 – Risk Difference Estimates from Proc Freq for Unadjusted Comparisons.

This is to say that rate of poor outcomes among the exposure group (E=1) is 0.4854 and the rate among unexposed group (E=0) is 0.2908 and the difference is -0.1946. We would say that there is about a 19.5% reduction in poor outcomes among the unexposed group, not adjusting for any other variables.

While we know from the data generation code that there is a multivariate relationship among the data, simple Proc Logistic code can be used to illustrate those models. Instead of focusing on logistic regression models on outcome, instead consider the logistic regression model on treatment. Specifically consider the following code:

```
*Exposure Modeling;
proc logistic descending;
class x3;
model e = x1 x2 x3;
*output propensities;
output out=temp p=phat;
run;
```

This creates a logistic regression model with all covariate inputs, one of which is categorical. We also output a new dataset call *temp* that has the probability of exposure listed as the variable *phat*. Then we run a Proc Univariate run of phat with histograms stratified by exposure, E, and we get the following:

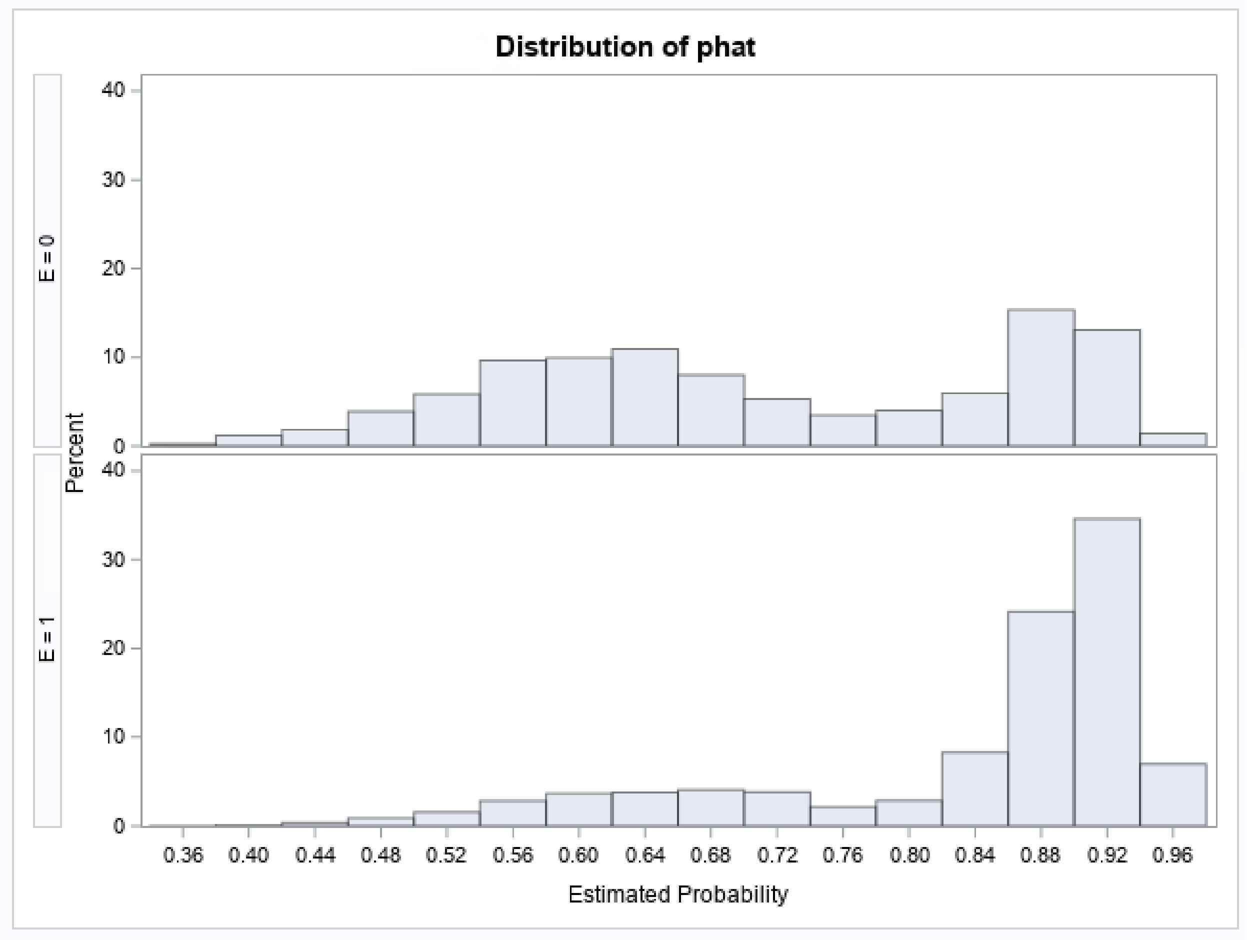


Figure 2 – Risk Difference Estimates from Proc Freq for Unadjusted Comparisons.

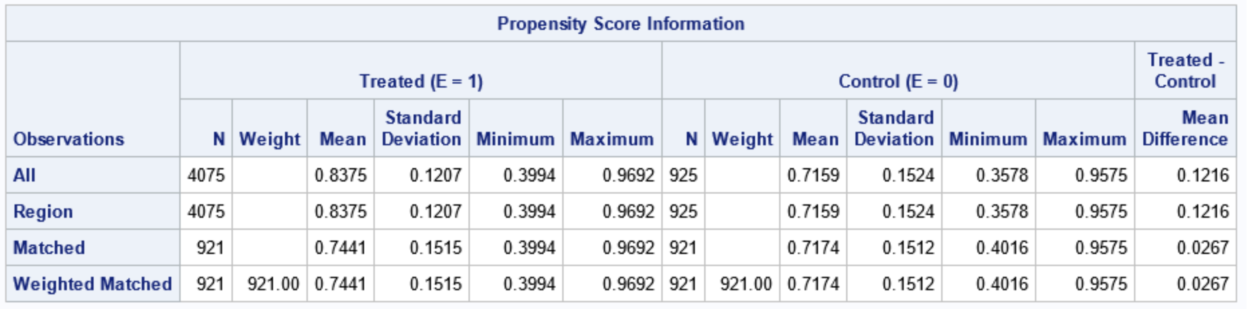
This basic output shows us the potential for overlap from this specific regression model. The PSMATCH procedure would use such a model to explore matching options. Note propensity models do not always give skewed propensity output, with this specific example we see that all individuals have some chance of exposure but that there is a large subgroup of individuals in the E=1 subgroup that have a very high likelihood of exposure. In a greedy matching scenario, it would be potentially hard to find a match for everyone here but matching with replacement would also tend to use (and reuse) those in the E=0 subgroup that have high propensity. The reader is encourage to tweak the variable inputs for x1, x2, and x3 to explore how they influence the distribution of propensities and the potential for overlap.

Moving on to implementing propensity score matching to this data, consider the following PSMATCH code:

```
*Propensity Score Matching - Greedy with Caliper;
ods graphics on;
proc psmatch data=Example1;
class E x3;
psmodel E (Treated='1')= x1 x2 x3;
match method=greedy distance=ps caliper=0.20;
assess ps allcov / plots=(barchart boxplot);
output out(obs=match)=PSmatched  matchid=_MatchID;
run;
ods graphics off;
```

There are many interesting aspects of the output, we will focus on two part, first this code has run a greedy propensity score matched with a caliper of 0.20 (meaning matches can be within plus/minus 0.20 of propensity scores in order to be assessed as a match) and we have turned the graphics option on so that we can see a variety of assessment output.

First the main output:



Propensity Score Information													
	Treated (E = 1)						Control (E = 0)						Treated - Control
Observations	N	Weight	Mean	Standard Deviation	Minimum	Maximum	N	Weight	Mean	Standard Deviation	Minimum	Maximum	Mean Difference
All	4075		0.8375	0.1207	0.3994	0.9692	925		0.7159	0.1524	0.3578	0.9575	0.1216
Region	4075		0.8375	0.1207	0.3994	0.9692	925		0.7159	0.1524	0.3578	0.9575	0.1216
Matched	921		0.7441	0.1515	0.3994	0.9692	921		0.7174	0.1512	0.4016	0.9575	0.0267
Weighted Matched	921	921.00	0.7441	0.1515	0.3994	0.9692	921	921.00	0.7174	0.1512	0.4016	0.9575	0.0267

Figure 3 – PSMATCH Output for Example Code

Here we see that greedy matching took 921 individuals from the E=0 group and matched them to 921 cases from the E=1 group. The mean difference in propensity scores between the matched groups is reduced from 0.1216 in the full sample to 0.0267 in the reduced sample. Note that adjustments in the caliper will both reduced the matched sample size but also the difference in the matched and unmatched sample. Interestingly, a key graphic in the output shows that the matched data align well on propensities:

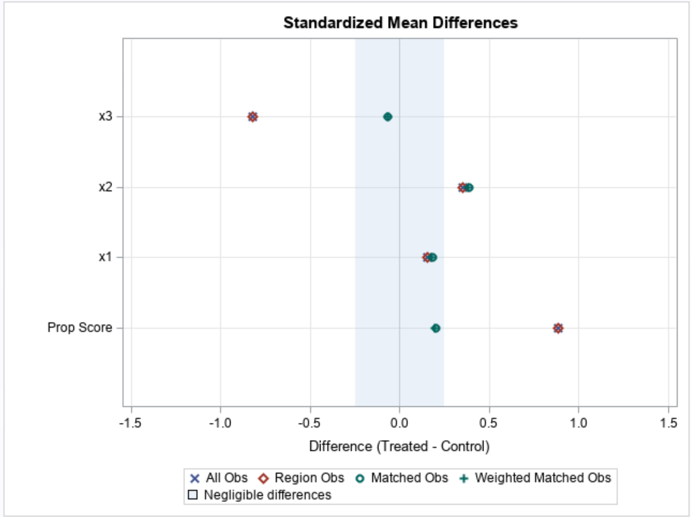
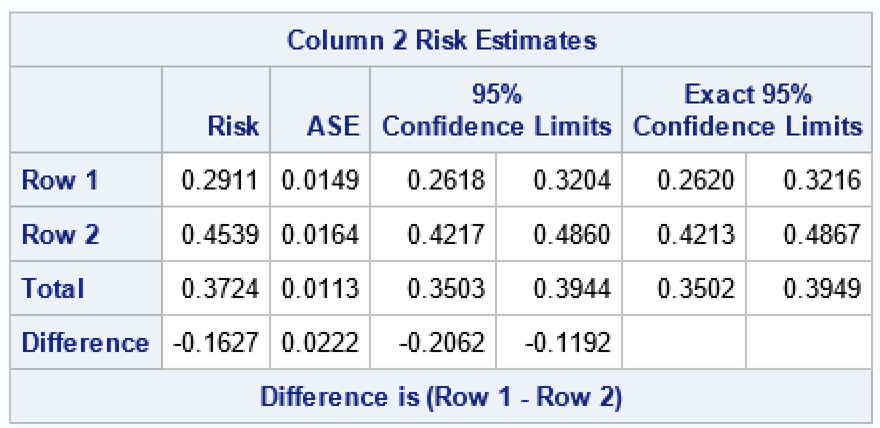


Figure 4 – Standardized Mean Differences

The output shows that the matched data align well on all covariates and propensity score because mean standardized mean differences are close to zero. However, the output dataset of this matched sample only has 1842 observations and we lose a lot of data with this strict a matching setup. There are ways to work to have a larger matched sample and those would be explored in a workshop format. The reader should play with the PSMATCH procedure options for number of matches (indicator is k) and reverse the intervention and control group as to have more of the E=1 group in a matched dataset. However, moving forward from here we can take the temporary output dataset which only has matched individuals (from the (obs=matched) option) and run another basic Proc Freq to look at the risk differences in this matched dataset. The output would include the following table:



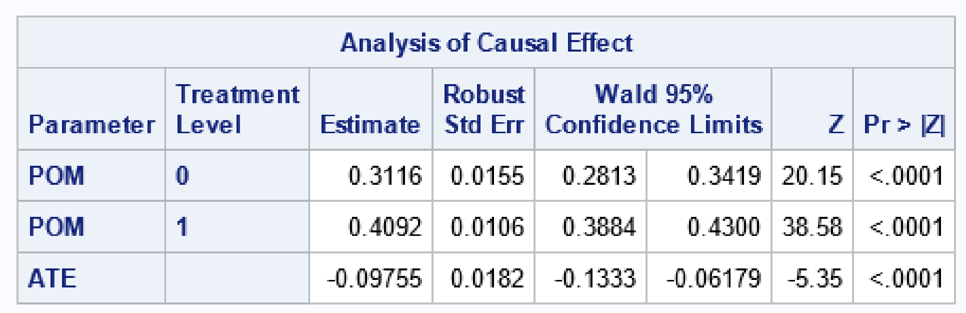
Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.2911	0.0149	0.2618	0.3204	0.2620	0.3216
Row 2	0.4539	0.0164	0.4217	0.4860	0.4213	0.4867
Total	0.3724	0.0113	0.3503	0.3944	0.3502	0.3949
Difference	-0.1627	0.0222	-0.2062	-0.1192		
Difference is (Row 1 - Row 2)						

Figure 5 – Risk Difference Estimates from Proc Freq for Unadjusted Comparisons.

Comparing this to the earlier output and we already see that the unadjusted risk difference is already lower in a matched cohort versus taking the entire dataset altogether. But since much of the data has been lost at this point, it would be better to have multiple matches in this imbalanced data or to take a weighted approach. The following CAUSALTRT code will perform these adjustments with both a propensity score model and an outcome regression model and give us an adjusted difference in its Average Causal Treatment Effect output:

```
Proc causaltrt method=aipw;
class e x3;
model Y= x1 x2 x3;
psmodel e = x1 x2 x3;
run;
```

Here we estimate treatment effects using the augmented inverse propensity weight estimator, which is encouraged in cases in which the sample size is moderate (several hundred cases) and there has not been an in-depth exploration of different regression model fits. The reader is encouraged to explore both the regression adjusted (regadj) and inverse propensity (ipw) estimators of treatment effect. The most important output of the above code is the following:



Analysis of Causal Effect							
Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	0	0.3116	0.0155	0.2813	0.3419	20.15	<.0001
POM	1	0.4092	0.0106	0.3884	0.4300	38.58	<.0001
ATE		-0.09755	0.0182	-0.1333	-0.06179	-5.35	<.0001

Figure 6 – CASUALTRT Output

The output shows reductions again in the difference in poor outcomes between the E=1 and E=0 groups. While the original unadjusted estimates of difference between groups was around -0.195 which dropped to -0.163 in a propensity matched group analysis all the way to only a -0.098 drop in a full model based analysis that considers all covariates and both outcome and propensity models. We see no overlap in the confidence intervals between the original unadjusted differences and the confidence intervals in the ATE of the output above. Overall, we see that there are still significant differences between the E=1 and E=0 groups but we have adjusted for a significant amount of covariate confounding and have obtain a better estimate of the true potential causal impact of E on Y.

EXAMPLE 2 -

The second example comes directly from Guo and Fraser (2015, Chapter 5); a sample of children from the Social and Character Development (SACD) program which was a joint program between the US Department of Education and the US Centers for Disease Control and Prevention. The SACD program focused on the school-wide implementation and the intervention curriculum was randomized to a number of schools in several states. Students were followed from third grade to fifth grade and had slightly different implementation across the different states. The data used here focus on 14 schools that were part of the North Carolina implementation with 7 intervention sites and 7 control sites. The intervention curriculum was called *Making Choices* and designed to increase social competence and reduce aggressive behaviors.

While the study had a group randomized design, certain features such as achievement on statewide tests and the distribution of race/ethnicity were different between the intervention and control sites. These school-wide systematic differences pushed down to individual level differences for students between intervention and control schools and as such straight intervention versus control comparisons would be biased and school-wide difference mean that there are differences in which students received the intervention and which received the control curriculum. The data and sample scripts can be found on the Guo and Fraser companion website; the textbook data includes all of the original SACD program data and some results of additional models that are fit based on discussions in the textbook.

We focus on the change in the Prosocial Behavior from the Carolina Child Checklist (CCCPROS) and we have data from the fall (beginning) and spring (end) of fifth grade. While we have the individual scores we are concerned with the change from fall to spring and in particular the outcome of interest will be whether a student had a positive change score (spring score minus fall score greater than 0) indicating an increase in prosocial behaviors and social awareness. Variables in the data include:

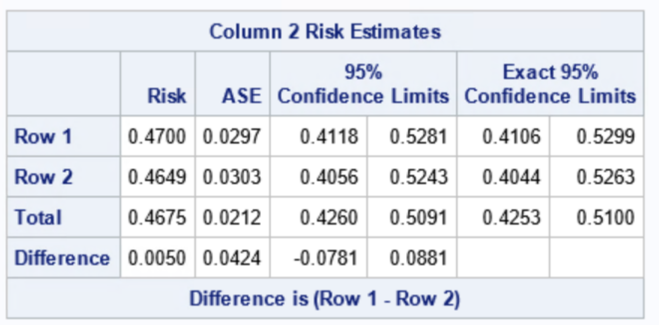
- intschb = Intervention Indicator (0 – Control, 1 - Intervention)
- agey = Child Age (Numeric)
- pcedu = Highest Grade of Primary Caregiver (Categorical)
- incpovlr = Income variable
- female = Gender Indicator
- black = African American Indicator
- hisp = Hispanic Indicator
- father = Father or Step-Father Lives in Household Indicator
- AYP05Cs = Academic Progress Score (Numeric)
- pmin05s = Percent Minority at School (Numeric)
- pfrd05s = Percent Free/Reduced Lunch at School (Numeric)
- schbl = School ID
- outcome = Indicator for Positive Change in Prosocial Behavior Score (Fall to Spring)

The Proc Freq code below shows the unadjusted relationship between intervention and outcome:

```
Proc freq;
tables intschb*Outcome/nocol nopercent riskdiff;
```

```
        run;
```

The associated output is in the figure (note that *Row 2* is the intervention group):



Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.4700	0.0297	0.4118	0.5281	0.4106	0.5299
Row 2	0.4649	0.0303	0.4056	0.5243	0.4044	0.5263
Total	0.4675	0.0212	0.4260	0.5091	0.4253	0.5100
Difference	0.0050	0.0424	-0.0781	0.0881		
Difference is (Row 1 - Row 2)						

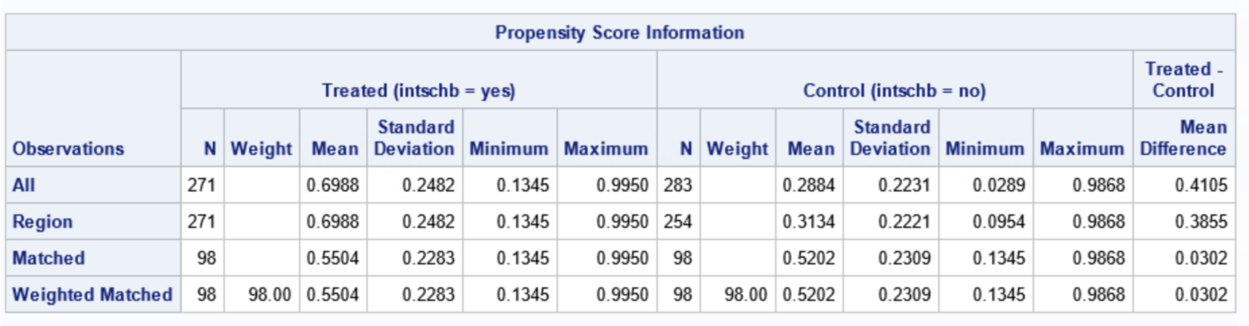
Figure 7 – FREQ Output

We see that with no covariate adjustment that there seems to be a negligible difference between the intervention and control groups. But a logistic regression model predicting intervention assignment shows a significant bias on Academic Progress Score (using all individual level covariates listed above). The model fit had a strong prediction (with area under an ROC curve of 0.89).

Performing a propensity score matching on this data with the following PSMatch code:

```
ods graphics on;
Proc PSMatch data=cccpros57;
class intschb female black hisp pcedu fatherr;
psmodel intschb (treated='yes') = agey female black hisp pcedu incpovlr
pcempf fatherr AYP05Cs;
match method=greedy distance=ps caliper=0.25;
assess ps allcov / plots=(barchart boxplot);
output out(obs=match)=PSmatched  matchid=_MatchID;
run;
ods graphics off;
```

Yields the following output:

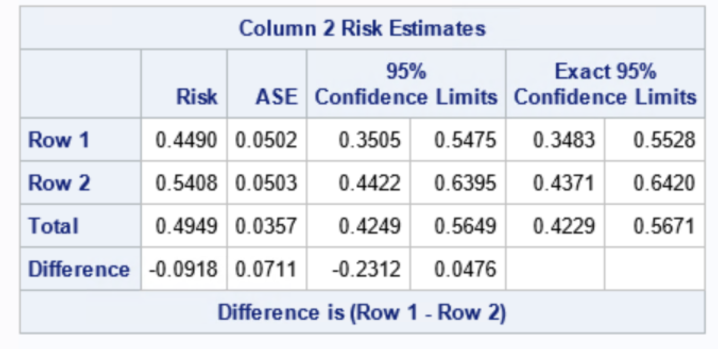


Propensity Score Information													
	Treated (intschb = yes)						Control (intschb = no)						Treated - Control
Observations	N	Weight	Mean	Standard Deviation	Minimum	Maximum	N	Weight	Mean	Standard Deviation	Minimum	Maximum	Mean Difference
All	271		0.6988	0.2482	0.1345	0.9950	283		0.2884	0.2231	0.0289	0.9868	0.4105
Region	271		0.6988	0.2482	0.1345	0.9950	254		0.3134	0.2221	0.0954	0.9868	0.3855
Matched	98		0.5504	0.2283	0.1345	0.9950	98		0.5202	0.2309	0.1345	0.9868	0.0302
Weighted Matched	98	98.00	0.5504	0.2283	0.1345	0.9950	98	98.00	0.5202	0.2309	0.1345	0.9868	0.0302

Figure 8 – PSMATCH Output

We see that our total sample of over 500 observations is only matched to a common sample of 196 individuals. We encourage the reader to run this code and explore the visuals to understand why the overlap between these groups is so poor. In addition, to increase matches between groups, we would also advocate using the method=replacement option (with say, 5 replicates) in order to have a more robust set of comparisons.

The propensity score matching has immediate impact on estimating unadjusted effects in that the same Proc FREQ code that we ran above, rerun on this smaller matched set has a very different intervention story:



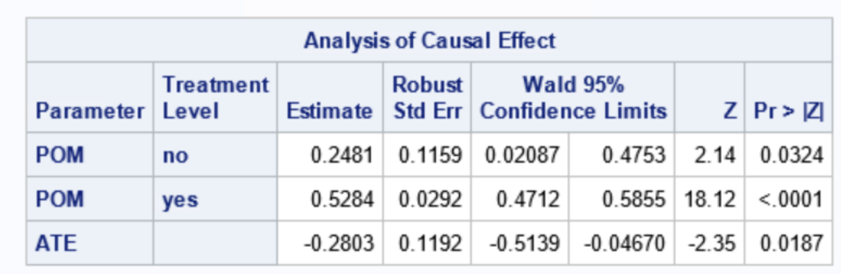
Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.4490	0.0502	0.3505	0.5475	0.3483	0.5528
Row 2	0.5408	0.0503	0.4422	0.6395	0.4371	0.6420
Total	0.4949	0.0357	0.4249	0.5649	0.4229	0.5671
Difference	-0.0918	0.0711	-0.2312	0.0476		
Difference is (Row 1 - Row 2)						

Figure 9 – FREQ Output

We see that in this smaller cohort that the impact of the intervention is that about 9% more students see growth in prosocial skills and awareness between fall and spring semesters. However, again we lose a lot of sample to the matching and are using less than half of the original data. If average causal treatment effect is what we want to look at then let's turn to the following CAUSALTRT code:

```
Proc causaltrt method=aipw data=cccpros57;
class female black hisp pcedu fatherr;
model outcome = agey female black hisp pcedu incpovlr pcempf fatherr
AYP05Cs;
psmodel intschb  = agey female black hisp pcedu incpovlr pcempf fatherr
AYP05Cs;
run;
```

Which yields the following output (using the augmented inverse propensity weighted estimator):



Analysis of Causal Effect							
Parameter	Treatment Level	Estimate	Robust Std Err	Wald 95% Confidence Limits		Z	Pr > |Z|
POM	no	0.2481	0.1159	0.02087	0.4753	2.14	0.0324
POM	yes	0.5284	0.0292	0.4712	0.5855	18.12	<.0001
ATE		-0.2803	0.1192	-0.5139	-0.04670	-2.35	0.0187

Figure 10 – CAUSALTRT Output

This is a very different estimate of intervention impact. Indeed, the reader should examine the code above using the regression adjusted, inverse propensity, and AIPW method options because the output is very different. This case shows that the various options and ways to handle these complex associations can have real impact. The AIPW estimator is likely the most robust option here but it does identify the largest potential intervention impact, in a case such as this reporting the confidence interval for percent whose change was positive (4.67%, 51.39%) is certainly advocated.

Lastly, Guo and Fraser again call attention to the problem that this is a nested design and we have done matching at the individual level. They provide a school matched propensity model in the data and a unique feature of the PSMATCH Procedure is the ability to take inputs from alternative models as propensities for matching. This means that hierarchical or multilevel models, models with variable selection features such as LASSO, and models using machine learning techniques could all be used as inputs. The reader is encouraged to pull the data from the website above and use the provided school based propensity scores that incorporate multilevel information.

CONCLUSION

The PSMATCH and CASUALTRT procedures provide a novel and powerful set of tools to help better understand the impact of interventions on outcomes in observational studies. These procedures have easy to use built in routines to provide powerful comparison groups for extended analysis and metrics for deeper covariate adjusted treatment effect. Beyond these key aspects, SAS has also provided powerful visual tools for assessing the impact of implementation and ways to incorporate more complex statistical modeling by bringing in propensity scores from other models.

REFERENCES

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Jewell, N.P. (2004) Statistics for Epidemiology. Chapman & Hall, London, U.K

Guo, S., & Fraser, M. W. (2015). Propensity score analysis. Sage.

Lamm, M & Yung YF (2017). Estimating Causal Effects from Observational Data with the CAUSALTRT

Procedure. Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc.

Available at https://support.sas.com/resources/papers/proceedings17/SAS0374-2017.pdf

CONTACT INFORMATION

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

Jason Brinkley
Senior Associate
919-294-7745
Jason_Brinkley@abtassoc.com