Paper 2382-2015

Reducing the Bias: Practical Application of Propensity Score Matching in Healthcare Program Evaluation

Amber Schmitz MS, Optum; Jessica Navratil-Strawn MS MBA, Optum; Stephen Hartley BS, Optum; Ronald Ozminkowski PhD, Optum

ABSTRACT

To stay competitive in the marketplace, healthcare programs must be capable of reporting the true savings to clients. This is a tall order considering most healthcare programs are setup to be available to the client's entire population; thus, the program cannot be conducted as a randomized control trial. In order to evaluate the performance of the program for the client, we use an observational study design which has inherent selection bias due to its inability to randomly assign participants. To reduce the impact of bias, we apply propensity score matching to the analysis. This technique is beneficial to healthcare program evaluations because it helps reduce selection bias in the observational analysis and in turn provides a clearer view of the client's savings. This paper will explore how to develop a propensity score, evaluate the use of inverse propensity weighting versus propensity matching, and determine the overall impact of the propensity score matching method on the observational study population. All results shown are drawn from a savings analysis using a participant (cases) versus non-participant (controls) observational study design for a healthcare decision support program aiming to reduce emergency room visits.

INTRODUCTION

Randomized control studies are the gold-standard in experimental design because they allow subjects to be randomly assigned to a participant group, effectively removing selection bias. Nevertheless, most real-world studies do not allow participant randomization and are carried out as observational studies where participant assignment is by choice and certainly not random. This non-random participant assignment may introduce selection bias into the study; this in turn may lead to misinterpretation of the results unless the selection bias is accounted for during the statistical analyses.

Observational studies frequently utilize propensity score methods to help reduce the inherent selection bias which in turn gives a clearer view of the participant effect. A propensity score is the probability from 0 to 1 that a subject would be assigned to the participant versus non-participant group in the observational study (Fraeman, 2010). The quality of the propensity score itself is dependent on the balancing effect of the observed and measurable covariates, and on the importance of the measured variables in determining the decision to participate or not in the intervention of interest (Heckman and Navarro-Lozano, 2004). Depending on the study design and sample sizes of the respective participant and non-participant groups, the propensity scores can be used in a variety of ways but are primarily used to balance study covariates through inverse propensity weighting or matching (Leslie, 2007). Once the observed characteristics of the participant and non-participant groups are balanced, the participant effect is estimated as the difference between the average outcomes of the two groups. Some analysts will also conduct another regression analysis with the matched or weighted data, using as independent variables the ones theoretically most important to determining the savings metric that is used to estimate savings, along with any variables not sufficiently balanced by the propensity methods (Faries et al, 2010). Others will use residual inclusion methods or other approaches at this stage to help account for selection bias introduced by variables that cannot be measured but which also influence program participation and outcomes (Terza et al, 2008).

This paper discusses how to develop a propensity score, evaluates the use of inverse propensity weighting versus propensity matching, and determines the overall impact of the propensity score matching method on the observational study population. All results shown are drawn from a savings analysis using a participant (case) versus non-participant (control) observational study design for a healthcare decision support program aiming to reduce emergency room visits.

DESCRIPTION OF STUDY

Emergency Room Decision Support (ERDS) is a healthcare decision support program that identifies subjects who have visited the emergency room one or more times within the past 12 months. The service provides subjects with education and information with the overall product goal of reducing unnecessary emergency room visits for non-emergency healthcare conditions. These subjects are telephonically outreached to by an intervention specialist who: 1) discusses the emergency room visit(s) with the subject, 2) reviews other healthcare resources available to the subject for non-emergency healthcare conditions, and 3) verifies engagement and help the subject find a primary care provider (if the subject does not already have one).

A retrospective cohort study applying propensity score matching was used to evaluate the effectiveness of the ERDS program on healthcare expenditures and emergency room utilization rates between the participant and non-participant study cohorts. Subjects that completed the telephonic outreach were considered for the participant (case) group while those subjects that declined or did not complete the telephonic outreach were considered to be in the non-participant (control) group.

DEVELOPING A PROPENSITY SCORE

A propensity score is the probability of each subject participating in the particular program based on measurable variables from the time period prior to participant status assignment. The basic syntax used is as follows:

```
proc logistic data=analysis_ds;
class <Class Baseline Variables>;
model participant_grp (event='1') = <Baseline Variables>;
output out=propensity_scores predicted=score;
run;
```

The PROC LOGISTIC statement is interpreted as follows:

- Analysis_ds is the subject-level dataset that includes a binary indicator variable for participant group assignment along with the baseline variables that the propensity score model requires.
- The CLASS statement contains all of the class baseline variables to be used in propensity score assignment.
- Participant_grp is the binary variable indicating which participant group the subject is a member of and the event='1' statement models the probability that the outcome will be 1 (assigned to participant group).
- <Baseline Variables> are all the variables being used in the propensity score model to generate the propensity score. These are measurable variables that occurred prior to participant assignment and should be theoretically related to the decision to participate in the intervention of interest or at least shown empirically to influence that decision in previous analyses.
- Propensity_scores is the output dataset containing the original Analysis_ds variables along with the new propensity score value (score variable).
- Score is the generated propensity score from the model with a value between 0 and 1.

PROPENSITY SCORE VARIABLE SELECTION

Selecting the baseline variables needed to generate the propensity score can be complex. As Parsons (2001) noted, the propensity score is only as good as its model. A bad model will produce unreliable propensity scores, so all efforts should be made to develop the best propensity score model for your study. A recommended approach is to collect as much information as possible on the subject and his/her health status and refine the model from there focusing on the variables most likely to influence program participation and associated medical care utilization. Most of the primary predictor variables investigated for propensity score fall into the following categories (examples are provided):

- Demographics: Age, gender, urban residence, region (geo-coded by zip code), income (geo-coded by zip code), work industry [Geo-coded variables from U.S. Census Bureau (2014)].
- Health Status: Mental health claims, pharmacy claims, IPRO future risk score, preventative health measures, number of Emergency Room visits in the last 12 months.
- Healthcare Supply Side Measures: Number of Primary Care Providers per 100,000 residents of health service area, Number of Specialty Care Providers per 100,000 residents of health service area, Number of Hospital Beds per 1,000 residents of health service area [Supply variables from The Dartmouth Atlas of Health Care (2014)].
- Other: Medical health plan benefit characteristics such as paid to allow ratio, capitated versus Fee-for-Service plan type.

In order to determine the model diagnostics, it is important to look at a correlation matrix to test the appropriateness of the variables in the propensity model along with multicollinearity testing between the predictor variables. The first step of verifying variables for inclusion in the propensity score model is correlation testing. Correlation testing is completed by using PROC CORR and the results are printed for documentation using PROC PRINT. The correlation matrix is verified to ensure that none of the predictor variables are correlated with each other. If correlation is present, suggested by value of 0.55 or higher in this example, one of the predictor variables is removed from the baseline variables and the testing for variable correlation is repeated. The sample code is as follows:

```
proc corr data=propensity_scores outp=correlation_out;
var <Baseline Variables>;
run;
title 'Output Dataset from PROC CORR';
proc print data=correlation_out noobs;
run;
```

After correlation testing, the second critical diagnostic check is to complete multicollinearity testing which assures that two or more of the predictor variables are not highly collinear. There are multiple ways to do this. An easy approach is to estimate the value of the Variance Inflation Factor (VIF); high values of the VIF for two or more variables suggest high collinearity. Researchers point to VIF values of 10 or higher as evidence of potentially harmful collinearity (Hair, 1995; Kennedy, 1992), but we have seen values much lower than that (e.g., as low as 3.0) provide evidence that collinearity may be problematic.

A second test for collinearity is based upon the Condition Index of the X;X matrix constructed from the independent variables used in the analysis. Condition Index values greater than 30 may indicate harmful colinearity. An advantage of using the condition index approach is that it provides more direct evidence of collinearity between three or more variables than does the VIF statistic or other collinearity detectors. This will help the user decide how many variables to keep in the final regression model -- those that are not highly correlated with each other or not highly linearly related to each other (Belsley, Kuh, and Welsch, 2004).

We test for multicollinearity using PROC REG. Please note that PROC REG requires binary variables instead of categorical variables, so all of our baseline categorical variables were transitioned to variables with a binary value. The code below demonstrates the basics of testing for multicollinearity:

```
proc reg data=analysis_ds_binary;
model outcome_var=<Baseline Variables Binary> / vif collin;
run;
```

In the 'model' statement above, 'vif' and 'collin' will produce the VIF and condition index output for review. Once correlation and multicollinearity testing have been completed, the remaining baseline variables are deemed to be the final <Baseline Variables> components and can be implemented into the propensity score creation code as previously discussed.

EVALUATION OF PROPENSITY APPLICATION METHODS

Once the propensity scores are generated, the most appropriate application technique for your study must be determined. Two main propensity score applications include inverse propensity weighting and propensity matching.

INVERSE PROPENSITY SCORE WEIGHTING

Inverse propensity weighting uses the propensity score to weight the subject in the outcome variable modeling. Each subject in the participant group is assigned a value of the direct inverse of his/her propensity score while each subject in the control group is assigned the inverse of the propensity score subtracted from 1. The inverse propensity weight is assigned in SAS® as follows:

```
data ps_weight;
set propensity_scores;
if participant_grp=1 then ps_weight=(1/score);
    else ps_weight=1/(1-score);
run;
```

Basically, what the weighting does is make the average values of variables measured for the participant group members look more like the average values of those variables for control group members, and vice versa, once the data are weighted. Since all of these variables were constructed prior to engagement in the ERDS program, propensity score weighting helps avoid selection bias due to pre-existing differences between these two groups of people.

The inverse propensity weight is used as a case weight when estimating the impact of the intervention on the outcome(s) of interest. Prior to completing the outcome variable modeling, the inverse propensity weight must be normalized in order to stabilize extreme weights (Lanehart, 2012). The need for stabilization techniques to account for extreme inverse propensity weights is one drawback of this technique. Another issue to consider when selecting between propensity weighting and matching is the distribution of the overall sample between the participant and control groups. If the participant group is disproportionately small compared to the control group, propensity score matching is most likely more appropriate for analysis due to the fact that the matching balances the overall analyzed population while propensity weighting could flood the analysis with data from a sample that was not as similar to the participant group as could be found if matching was used.

PROPENSITY MATCHING

The second primary application option for propensity scores is that of propensity score matching. The purpose of propensity score matching is to match participant group subjects to control group subjects based on their respective propensity scores in order to achieve similar groups for the final outcomes comparison. There are several ways to do this, but the most-often used methods are called 'optimal match' and 'greedy match'. The differences between these methods are as follows:

- The greedy match uses the nearest-neighbor approach, which means that the first match that is within an established caliper distance or minimum acceptable distance between propensity scores is chosen and maintained.
- The optimal match looks for the closest distance between any matched combinations and thus reconsiders matches until the closest or optimal match is established.

For the application in our ERDS analysis, we have chosen to proceed with propensity score matching via the greedy method over both inverse propensity score weighting and optimal propensity matching due to

a desire to avoid an overabundance of control group members, given the small sample size of the participant group in the study. The algorithm applied is based on that presented in Parsons (2001) which samples for best matches and then iteratively loops to create next-best matches until the matching loop completes the matching through the chosen caliper levels via a 6 to 2 greedy matching algorithm. The caliper level is the digit of the propensity score value that is indicated for matching.

A 6 to 2 greedy match indicates that we start our greedy looping algorithm with participant subjects searching for the control subjects that have the closest propensity score match at the sixth digit caliper (i.e. 0.000001). If the participant subject finds a match at the sixth digit caliper, his/her match is established and not reconsidered. If the participant subject does not find a match at the sixth digit caliper, the participant subject repeats the propensity score match assessment to the fifth digit caliper. This scenario continues to loop until the matching algorithm completes establishing matches to the second digit caliper (i.e. 0.01). Once the second digit caliper loop completes, the matched pairs from all of the iterative loops are compiled into the final matched dataset for the outcomes analysis.

One thing to consider is that matching methods will inevitably lead to incomplete matches. An incomplete match may stem from 1) subjects having missing data that is required for the propensity score in which case no propensity score will be calculated or 2) disjointed propensity score ranges between the participant and control groups. We typically strive for a 90% match of the participant group subjects. If a 90% match cannot be met, the propensity score algorithm will need to be reassessed. Further information on how to evaluate the performance of the chosen caliper level greedy match can be found in Parsons (2001).

OUTCOME OF APPLICATION

The results of the application of the propensity score matching algorithm for a 6 to 2 greedy match are listed below. Group differences were evaluated using student t-tests and chi-squared tests while effect size/significance was evaluated using standardized differences. Standardized differences were used instead of p-values since they are not influenced by sample size (Faries, 2010). Standardized differences with an absolute value of 0.1 or greater were considered significant. For brevity, the results shown are for the covariates that were unbalanced in the original population. All covariates in the propensity score model and final outcomes assessment model should be evaluated for balance upon the application of propensity score matching.

Table 1 Displays significantly different covariates between the participant and control groups before matching.

Variable - Level		Table 1: Original Population		
		Participant N=1,007	Non-Participant N=19,165	Standardized Difference
Age	Under 40 (%)	38.1	48.3	**0.207**
	40 & Older (%)	61.9	51.7	**0.207**
ER to Index Date Lag	Less than 165 Days (%)	79.3	86.9	**0.202**
	165 Days or More (%)	20.7	13.1	**0.202**
Metropolitan Statistical Area (MSA)	No (%)	11.4	8.1	**0.111**
	Yes (%)	88.6	91.9	**0.111**
Mammogram Preventative Screening Eligible	No (%)	74.0	80.4	**0.152**
	Yes (%)	26.0	19.6	**0.152**
Cervical Preventative Screening Eligible	No (%)	72.4	77.5	**0.118**
	Yes (%)	27.6	22.5	**0.118**
Colorectal Preventative Screening Eligible	No (%)	87.5	91.8	**0.140**
	Yes (%)	12.5	8.2	**0.140**

Table 1. Unbalanced covariates in original population.

Table 2 Displays the results based on the final 6 to 2 greedy matched pairs dataset. A 90% match of the participant group was achieved. For each of the covariates that had a significant difference at baseline, there is no longer a significant difference in the matched pairs dataset. The matched dataset is now ready to perform the outcomes modeling to complete the ERDS savings evaluation.

Variable - Level		Table 2: 6 to 2 Greedy Matched Population		
		Participant N=893	Non-Participant N=893	Standardized Difference
Age	Under 40 (%)	40.6	41.1	0.010
	40 & Older (%)	59.4	58.9	0.010
ER to Index Date Lag	Less than 165 Days (%)	83.9	83.8	0.004
	165 Days or More (%)	16.1	16.2	0.004
Metropolitan Statistical Area (MSA)	No (%)	8.8	7.6	0.044
	Yes (%)	91.2	92.4	0.044
Mammogram Preventative Screening Eligible	No (%)	74.4	77.4	0.071
	Yes (%)	25.6	22.6	0.071
Cervical Preventative Screening Eligible	No (%)	70.9	71.9	0.023
	Yes (%)	29.1	28.1	0.023
Colorectal Preventative Screening Eligible	No (%)	88.6	90.0	0.046
	Yes (%)	11.4	10.0	0.046

Table 2. Balanced covariates in matched population.

Table 3 Displays the impact of the propensity matching on the ERDS program savings. As expected, the savings value was different for the No Propensity model due to the difference in baseline characteristics of the participant (case) and non-participant (control) groups. Once propensity score matching was applied, savings increased for the client. Without controlling for the differences in characteristics between the case and control groups, the savings would have been misrepresenting the actual performance of the program.

Table 3: Summary of Propensity Impact on Savings			
Metrics		No Propensity	6 to 2 Greedy Propensity Match
Sample Size	Control	19,165	893
	Case	1,007	893
Program Savings		$0.35 Million	$1.01 Million

Table 3. Summary of propensity impact on savings.

CONCLUSION

Observational data make healthcare program analysis much more timely and efficient; however, if not analyzed appropriately, the true value of the healthcare program may not be realized due to selection bias effecting participant group characteristics. Propensity scores can easily be generated in SAS® and applied to remove some of the selection bias that is inherent in observational studies. The example presented here accounts for selection bias by applying a propensity score matching technique due to the fact that the participant group is disproportionately smaller than the control group. The use of propensity score matching allows for a more balanced approach to assessing the true value of the savings of the ERDS program for our clients.

One of the limitations of any propensity score approach is that it removes selection bias due to only measurable differences between ERDS program participants and non-participants. There may still be some selection bias due to unmeasured differences, especially if these unmeasured differences are not highly correlated with variables we included in our analysis (Love, 2004). There are several ways to address unmeasured variables. James Heckman started the modern era of work in this area in the mid-1970s, eventually winning the Nobel Prize in economics for his work, in 2000. A description of his other many other models is beyond the scope of this paper but the reader is encouraged to investigate such modeling to determine if even better adjustments for selection bias are feasible. See papers by Heckman (1976 – the one that got it all started), Terza et al. (2008) and Vella (1998) for several examples.

REFERENCES

Belsley, D.A., Kuh, E. Welsch, R.E. *Regression Diagnostics: Identifiying Influention Data and Sources of Colinearity*. Hoboken, NJ: John Wiley & Sons, Inc., 2004.

Faries, D., Leon, A., Haro, J., Obenchain, R. Analysis of observational healthcare data using SAS. Cary, NC: SAS Institute, Inc., 2010.

Fraeman, K., "An Introduction to Implementing Propensity score Matching with SAS®", *Proceedings of the 2010 Northeast SAS User Group Conference,* Baltimore, MD. 2010.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C., Multivariate Data Analysis (3rd ed). New York: Macmillan, 1995.

Heckman, J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for Such Models, "*Annals of Economic and Social Measurement*, **5** (4), 475-492, 1976.

Heckman, J., and Navarro-Lozano, S. "Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics*, 86(1): 30-57, 2004.

Kennedy, P., A Guide to Econometrics. Oxford: Blackwell., 1992.

Lanehart, R., et al. "Propensity Score Analysis and Assessment of Propensity Score Approaches Using SAS® Procedures", *Proceedings of the SAS Global Forum 2012 Conference*, Orlando, FL. 2012.

Leslie, S. and Thiebaud, P., "Using Propensity Scores to Adjust for Treatment Selection Bias", *Proceedings of the SAS Global Forum 2007 Conference*, Orlando, FL. 2007.

Love TE. *Using Propensity Score Methods Effectively*. Copyright by Thomas E. Love, 2004.

Parsons, L., "Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques", *Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference,* Long, Beach, CA. 2001.

Terza, R., A. Basu, and P. Rathouz "Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling,"*Journal of Health Economics*, **27** (3), 531-543, 2008.

The Dartmouth Atlas of Health Care. The Darmouth Atlas of Health Care Tools. Available at: <http://www.dartmouthatlas.org/tools/>. Accessed July 28, 2014.

U.S. Census Bureau. United States Census Bureau Data. Available at: <https://www.census.gov/data.html>. Accessed July 28, 2014.

Vella F. "Estimating models with sample selection bias: A survey." *Journal of Human Resources*, **33** (1), 128-169, 1998.

ACKNOWLEDGMENTS

The methods and examples presented here are derived from Optum's Emergency Room Decision Support product support, conducted by the Consumer Solutions Group Healthcare Analytics team. The authors would like to thank Optum for its continued support of knowledge sharing in the healthcare industry.

CONTACT INFORMATION

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

Amber Schmitz
Consumer Solutions Group, Optum
6300 Olson Memorial Highway
Golden Valley, MN 55427
E-mail: amber.schmitz@optum.com
Web: www.optum.com