Paper 11901-2016

Utilizing Propensity Score Analyses to Adjust for Selection Bias: A Study of Adolescent Mental Illness and Substance Use

Deanna Schreiber-Gregory, National University

Abstract

An important strength of observational studies is the ability to estimate a key behavior or treatment's effect on a specific health outcome. This is a crucial strength as most health outcomes research studies are unable to use experimental designs due to ethical and other constraints. Keeping this in mind, one drawback of observational studies (that experimental studies naturally control for) is that they lack the ability to randomize their participants into treatment groups. This can result in the unwanted inclusion of a selection bias. One way to adjust for a selection bias is through the utilization of a propensity score analysis. In this study we provide an example of how to utilize these types of analyses. Our concern is whether recent substance abuse has an effect on an adolescent's identification of suicidal thoughts. In order to conduct this analysis, a selection bias was identified and adjustment was sought through three common forms of propensity scoring: stratification, matching, and regression adjustment. Each form is separately conducted, reviewed, and assessed as to its effectiveness in improving the model. Data for this study was gathered through the Youth Risk Behavior Surveillance System, an ongoing nationwide project of the Centers for Disease Control and Prevention. This presentation is designed for any level of statistician, SAS® programmer, or data analyst with an interest in controlling for selection bias, as well as for anyone who has an interest in the effects of substance abuse on mental illness.

Introduction to the Data Set

The Youth Risk Behavior Surveillance System (YRBSS) was developed as a tool to help monitor priority risk behaviors that contribute substantially to death, disability, and social issues among American youth and young adults today. The YRBSS has been conducted biennially since 1991 and contains survey data from national, state, and local levels. The national Youth Risk Behavior Survey (YRBS) provides the public with data representative of the United States high school students. On the other hand, the state and local surveys provide data representative of high school students in states and school districts who also receive funding from the CDC through specified cooperative agreements. The YRBSS serves a number of different purposes. The system was originally designed to measure the prevalence of health-risk behaviors among high school students. It was also designed to assess whether these behaviors would increase, decrease, or stay the same over time. An additional purpose for the YRBSS is to have it examine the co-occurrence of different health-risk behaviors. This particular study exams the co-occurrence of suicidal ideation as an indicator of psychological unrest with other health-risk behaviors. The purpose of this study is to serve as an exercise in correlating two different variables across multiple years with large data sets.

Methods

YRBSS provided data sets free to the public online and instructions on how to download the data sets, as well as how to apply the formatting. In order to apply the formatting, the researcher needed only to specify libraries for the data sets and formats:

```
libname mydata 'I:\RRSC_MWSUG_Analytics_2012\MWSUG_2012';   /* Tells SAS® where
the data is */
libname library 'I:\RRSC_MWSUG_Analytics_2012\MWSUG_2012';  /* Tells SAS® where
the formats are */
```

This enabled SAS® to read all the formatting as well as output the variable names, questions, and answers in a very clean manner

Concatenating Data Sets

In order for data from all of the years to be used in this analysis, concatenating the 11 data sets was necessary. The researcher chose which questions would be used in the analysis based on the whether or not the questions in all of the national surveys. All questions asked between the years of 1991 and 2011 were included in the model and separated into categories based on risk behavior type. The questions used were then given new names in order for the appropriate questions to be concatenated together. This was necessary because even though the questions used were present in all the surveys, the order in which the questions appeared in the survey differed between each year.

```
data YRBS1991;
       set mydata.YRBS1991;

       alcohol1=q33; alcohol2=q32; alcohol3=q34; alcohol4=q35;
       drugs1=q37;         drugs2=q36;         drugs3=q38;         drugs5=q39;
       drugs6=q40;  drugs7=q41;         drugs8=q43;         drugs15=q44;
       drugs16=q45; mood2=q19;          mood3=q20;          mood4=q21;
       mood5=q22;          sexuality2=q48;     sexuality3=q49;
       sexuality4=q50;     sexuality5=q51;     sexuality6=q52;     tobacco1=q23;
       tobacco2=q25; tobacco3=q24; tobacco4=q27; tobacco5=q28; tobacco6=q29;
       tobacco7=q30; tobacco11=q26;     tobacco12=q31;      vehicle1=q9;
       vehicle2=q10; vehicle3=q6;  vehicle5=q11; vehicle6=q12; vehicle8=q7;
       vehicle9=q8; violence1=q14;      violence2=q15;      violence10=q16;
       violence11=q17;     violence12=q18;     drop q1-q97;
run;
```

The coding to concatenate the years together is given below:

```
data YRBS_Total;
set    YRBS1991   YRBS1993   YRBS1995   YRBS1997    YRBS1999    YRBS2001
       YRBS2003   YRBS2005   YRBS2007   YRBS2009    YRBS2011;
run;
```

Exploring the Data Set

To begin the analysis, the researcher used proc freq to find the frequency of occurrence for each variable response in the data set. Frequencies for demographics, risk behaviors, and mental health variables are all provided and reviewed. The appropriateness of weighting the variables involved in the model was explored using the results. An example of the code used is provided below:

```
proc freq data=YRBS_Total;
       tables year*mood1 year*mood2 year*mood3 year*mood4 year*mood5;
run;

proc sgpanel data=YRBS_Total;
        title "Yearly Mood1";
        panelby mood1/ novarname columns=1;
        vbar year;
run;
```

These frequencies showed very little change in each of the responses over the years. Also, when looking at the percentages of each response, the majority of students either denied participating in any unique risky behavior or reported participating in the behavior at a lower rate than other respondents. Given these results, the researcher sought to find out if participation in a particular set of risky behaviors, being that any unique risky behavior is avoided by the majority of the population, would contribute to suicidal ideation. This idea was formulated from the general idea that most risky behaviors are viewed as poor decisions or compensatory behaviors initiated by the environment or other stimuli.

Alternative R² and Model Fit Statistics

In this study, the max-rescaled r-square statistic (adjusted Cox-Snell) provided by SAS® as an option in the model statement of PROC SURVEYLOGISTIC is used as the main reference point in the explanation of how much each final model explains the occurrence of the dependent variable (and is therefore the means to which we evaluate the impact of the latent variables on the explanatory power of the model), however, there is some debate as to whether this is the most appropriately calculated statistic for the job. Paul D. Allison, in his talk on model fit statistics at SAS Global Forum 2014, touched on the possible preferential use of McFadden and Tjur tests. Allison argues that that the Cox-Snell R^2 has appeal as it is able to be naturally extended to regression models other than logistic, such as negative binomial regression and Weibull regression; however, the main limitation of the Cox-Snell test, and thus the reason we are exploring other options of estimating R^2, is its less-than-desirable short upper bound. The Cox Snell R^2 upper bound is less than 1.0. In fact, the upper bound for this test can oftentimes be a lot less than 1.0 depending on *p*, the marginal proportion of cases with events. Allison goes further to provide these examples of this gross deviance: if p=.5, the upper bound reaches a maximum of .75, but if p=.9 (or .1), the upper bound is a mere .48. This is why the max re-scaled R^2 is provided and used in this analysis, as it divides the original Cox-Snell R^2 by its upper bound, thus helping thus helping fix the problem of the “lower” upper bound. However, given that this deviation does exist, it would be beneficial to explore and record other R^2 alternatives when completing an in-depth analysis, however, for the sake of time and consistency, the Cox-Snell R^2 will continue to be reported in the analyses within this paper.

For model fit, the surveylogistic procedure provides three different model fit statistics: Akaike’s Information Criterion (AIC), Schwarz Criterion (SC), and the maximized value of the logarithm of the likelihood function multiplied by -2 (-2 Log L). When interpreting these model fit statistics, it is useful to note that lower values of each of these statistics indicates better fit; however, these statistics are open to interpretation and should be considered carefully as they are highly dependent on the structure of the model and sensitive to the number of variables and interactions that are included. These provided statistics are what this paper is primarily using for model fit; however, there are several ways to measure model fit in a logistic regression model. Paul D. Allison, in the same SAS Global Forum 2014 presentation on model fit statistics mentioned above, covers five alternative goodness-of-fit measures for logistic regression models: Hosmer-Lemeshow test, standardized Pearson sum of squared residuals, Stukel’s test, and the information matrix test. For the sake of consistency and time, we will continue to look to the model fit statistics provided by SURVEYLOGISTIC; but it is worth noting to explore these other alternatives when conducting a more thorough analysis. For anyone who would like to explore these alternative statistics, a GOFLOGIT macro is available at https://github.com/friendly/SAS-macros/blob/master/goflogit.sas was created as a comprehensive evaluation of the statistics available above (except for Hosmer-Lemeshow, which is simply indicating the lackfit option in model statement after a forward slash); however, Allison warns that a very fundamental problem in the application of a couple of these statistics is included in the model. In short, this macro is available to use if one so wishes but extreme caution should be taken in its interpretation. Please refer to Allison’s paper mentioned above for a more in depth explanation as to the limitation of the available macro.

Introduction to Propensity Scores

Randomized control trials (RCTs) measure the efficacy of treatment in controlled environments; however, this can often be restricted to subpopulations that limit generalizability of results. Observational studies, on the other hand, can evaluate treatment effectiveness in routine care settings or everyday use patterns. Considering this, a limitation of observational studies is the lack of treatment assignment. Non randomized groups usually differ in observed and unobserved characteristics causing selection bias when evaluating the effect of treatment.

Statistical techniques such as matching, stratification, and regression adjustment are commonly used to account for differences in treatment groups but may be limited if using too few covariates in the adjustment process. The use of propensity score techniques avoids this limitation because it can summarize more or all of the covariate information into a single score. The question now is, what is a propensity score? The propensity score is the conditional probability of being treated based on individual covariates. Rosenbaum and Rubin demonstrate that propensity scores can account for imbalances in treatment groups and reduce bias by resembling randomization of subjects into treatment groups.

By using propensity scores to balance groups, traditional adjustment methods can better estimate treatment effect on outcomes while adjusting for covariates. One method professed by Ralph B. D’Agostino, Jr. to adjust for the non-randomized treatment selection is to use a propensity score method in conjunction with traditional regression techniques. This process is performed using two steps, the first of which calculates propensity scores as the probability of patients being included in each treatment group based on pre-treatment observables. The aim of this step is to create balanced treatment groups that simulate random treatment allocation. The second step utilizes the created propensity scores with ANCOVA to more accurately estimate outcomes and study the possible covariate predictors.

Method Behind a Propensity Score Analysis

The logistic model describes the relationship of several independent variables to a dichotomous dependent variable. Furthermore, logistic regression is used to predict the probability of an event occurring as a function of independent variables (continuous and/or dichotomous). The logistic model can be represented as such:

$$P(X) = \frac{1}{1 + e^{-(\alpha + \Sigma \beta_i x_i)}}$$

Propensity scores are easily created through the LOGISTIC procedure in SAS®. In the case used and steps described in this paper, the dependent variable is treatment group (suicidal ideation/attempt) and the independent variables are substance abuse measures, demographics, and other environmental risk factors. In some cases, the dependent variable may be any dichotomous outcome (treated or untreated, uses drugs or does not, in an abusive relationship or not). The GENMOD procedure for generalized linear models may also create propensity scores by using the OUTPUT statement and keyword PREDICTED.

Propensity Score Creation Through PROC LOGISTIC

The following application illustrates the use of PROC LOGISTIC to create propensity scores. PROC LOGISTIC calculates propensity scores as the conditional probability of each adolescent participating in illicit drug use based on environmental variables and can output the propensity score to a data set. In this example, propensity scores were calculated based on a list of predefined covariates. The objective of this application was to balance the treatment groups so to reduce bias of treatment selection ad to obtain a better idea of treatment effect on the outcome of compliance. The logic function was specified in the LINK option to fit the binary logit model and the RSQUARE option assesses the amount of variation explained by the independent variables. The propensity score is outputted to the data set and named “psdataset”. The predicted probabilities are outputted to the variable named “ps”.

```
Proc logistic data=YRBS;
Class mhealth;
Model sabuse (event=’Yes’) = age gender race wviolence bviolence sviolence
sadness sactivity / link=logit rsquare;
Output out=psdataset pred=ps;
Run;
```

After creating the propensity scores, an evaluation of the distributions can check comparability of the treatment groups. Sizeable overlaps among the groups illustrate satisfactory overlap in covariate distributions and indicate that the groups are comparable.

The TABULATE procedure can then be used to create tables in order to demonstrate how propensity scoring can balance the groups. Unadjusted values (before propensity scores) for the two treatment groups can be displayed effectively in this way. Descriptors include demographic variables (age, gender, race), violence types (sexual, domestic, school, weapon), cigarette use, and alcohol use. PROC TTEST and PROC FREQ can then be used to test for differences between the groups.

Adjusted values (after propensity scores) for the two treatment groups can then be displayed in an additional table. PROC GLM can then be used to compare groups while adjusting for the propensity score. Differences between groups should be minimized when using the propensity score method.

Use of Propensity Scores

Once the propensity score is calculated what do you do with it then? As explained above, the 3 methods commonly used are matching on propensity score, stratification, and regression adjustment.

Regression Adjustment

Continuing with the examination of our study topic, the created propensity scores were used in regression adjustment where a propensity score weight, also referred to as the inverse probability of treatment weight (IPTW), was calculated as the inverse of the propensity score (Hogan and Lancaster). The treatment selection model above modeled the propensity to participate in illicit drug use. For those adolescents who did not participate in illegal drug use, the propensity score would be 1-ps and the propensity score weight would be the inverse of 1-ps.

```
Data psdataset
Set psdataset;
```

```
If sabuse=1 then ps_weight=1/ps;
Else ps_weight=1/(1-ps);
Run;
```

Next, a propensity score-weighted linear regression model, using the GLM procedure, was fitted to compare illicit drug use on the outcome of suicidal ideation while controlling for other covariates. The LSMEANS statement computes the least-squares means for the treatment variable allowing for multiple comparisons. The ADJUST=TUKEY option uses the Tukey-Kramer method to adjust the least-squares means and the PDIFF and CL options give the p values and corresponding confidence intervals for the differences in the least-squares means.

```
Proc glm data=psdataset;
Class sabuse mhealth gender race;
Model mhealth = sabuse gender age race violence / solution;
Lsmeans sabuse / OM ADJUST=TUKEY PDIFF CL;
Weight ps_weight;
Quit;
```

Stratification

Stratification, subclassification or binning using propensity scores involves grouping subjects into classes or strata based on the subject's observed characteristics. Once the propensity scores are calculated, subjects are placed into strata (Cochran states that 5 strata can remove 90% of the bias) with the idea that subjects in the same stratum are similar in the characteristics used in the propensity score development process. The tutorial by D'Agostino details how to perform this technique. Briefly, quintiles are used to group subjects into five strata after making sure that there is adequate propensity scores overlap between the treatment groups. To prove that the propensity scores removed any bias due to differences in covariates between treatment groups, t-tests or chi-square tests are conducted before and after propensity score creation. Finally, outcomes and treatment effects can be assessed using models while adjusting for the propensity scores. Continuing with the example and code above, subjects are divided into 5 classes based on the common propensity score overlap using the RANK procedure. Checking for difference between treatment group before and after stratifying subjects by propensity scores can be done using PROC FREQ, PROC TTEST and PROC GLM.

```
Proc rank data=psdataset groups=5 out=r;
Ranks rnks +1;
Var ps
Run;

Data quintile;
Set r;
Quintile = rnks +1;
Run;

/* check for differences in groups before propensity score*/
Proc freq data=quintile;
Tables sabuse*(gender mhealth race violence) / chisq;
Run;

Proc ttest data=quintile;
Var age sseverity;
Class sabuse;
Run;

/*check for differences in groups while adjusting for propensity scores*/
Class sabuse quintile;
Model age gender violence race mhealth = sabuse quintile;
Lsmeans sabuse;
Quit;
```

Matching

Matching groups by propensity scores is a common method to balance on covariates. Once the propensity score is calculated, subjects are matched by this single score as opposed to traditional direct matching by one or more covariates. A disadvantage of matching methods includes incomplete matching and inexact matching. That is, subjects may be excluded because of difficulty finding a match. Reducing this bias is well explained in Lori Parsons' papers (see reference section). Her papers offer code for performing case-control match using a greedy matching algorithm. As she explains in the paper, cases can be matched to controls based on propensity scores. TABULATE can be used again to create tables in order to display the results. Tables 1 and 2 should show how propensity scores were used to balance a treated (positive association with illicit drug use) and untreated (negative association with illicit drug use). Table 1 should contain the original population and should include results of rank-sum tests and chi-square tests showing differences between groups in many characteristics (only a few are shown). Table 2 should show how differences are eliminated after matching. This matched subset of patients can now be used to model outcomes and assess effect of treatment.

Summary and Review

Calculating the propensity score as the conditional probability of treatment summarizes observed values into a single score. The scores can then be used to control for selection bias by matching subjects, stratifying subjects, and/or as a regressor. All techniques have the purpose of balancing groups to remove bias when assessing treatment effect or outcomes.

Traditional techniques to control for bias may be limited if accounting for only a few covariates. Compared to multiple regression, the propensity score methodology summarizes many observables and is less sensitive to model misspecification (Perkins et al). Propensity scores can also diagnose if groups are comparable before moving onto the modeling stage of analysis. If distributions of the propensity scores fail to show much overlap in covariate values, the comparison groups are too different making it difficult to balance groups.

When creating propensity scores all covariates that affect both treatment and outcome must be included in the model and it is assumed that all patients have a non zero probability of receiving each treatment. The technique only looks at observed characteristics of the population thus does not account for unobserved factors, such as socioeconomic status, family relations, and friendships. This limitation is modified if unobserved covariates are correlated to observed factors. Analyses also need large sample sizes in order to establish adequate variance in covariate distributions.

Conclusion

This application introduces potential selection bias in observational studies and describes how propensity score methodology can control for overt bias when estimating treatment effectiveness. Propensity scoring methodology attempts to balance groups before comparing outcomes between treatment groups. Commonly used techniques via propensity scores include matching, stratification, and regression adjustment using the inverse of the propensity score. Each method can be used in conjunction with traditional risk adjustment techniques to reduce bias and better describe the effect of treatment on outcomes.

References

D'Agostino R.B. Sr, Kwan H. 1995. "Measuring Effectiveness: What to Expect Without a Randomized Control Group". Medical Care. 195:33 (4 suppl): AS95-AS105.

D'Agostino R.B., Jr, D'Agostino R.B., Sr. 2007. "Estimating Treatment Effects Using Observational Data". JAMA. 297 (3). 314-316

Rosenbaum P.R. and Rubin D.B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika, 70, 41-55.

D'Agostino, R.B. 1998. "Tutorial on Biostatistics: Propensity Score Methods for Bias Reduction in the comparison of a treatment to a non-randomized control group". Statistics in Medicine 17, 2265-2281.

Hogan, J.W., Lancaster, T. 2004. "Instrumental variable and propensity weighting for causal inference from longitudinal observational studies". Statistical Methods in Medical Research 13: 17-48.

Obenchain, R.L., Melfi, C.A., “Propensity Score and Heckman Adjustments for Treatment Selection Bias in Database Studies”.

Pasta, David J. 2000. “Using Propensity Scores to Adjust for Group Differences: Examples Comparing Alternative Surgical Methods”. Proceedings of the Twenty-Fifth Annual SAS® Users Group International Conference, Indianapolis, IN, 261-25.

Parsons, Lori. 2000. “Using SAS®® Software to Perform a Case Control Match on Propensity Score in an Observational Study”. Proceedings of the Twenty-Fifth Annual SAS® Users Group International Conference, Indianapolis, IN, 214-26.

SAS® Institute Inc. 2004. “SAS® Procedures: The LOGISTIC Procedure”. SAS® OnlineDoc® 9.1.3. Cary, NC: SAS® Institute Inc.

http://support.SAS®.com/documentation/onlinedoc/91pdf/SAS®doc_91/stat_ug_7313.pdf

Contact Information

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

Deanna Schreiber-Gregory, BS
Masters Student
Health and Life Science Analytics
National University
E-mail: d.n.schreibergregory@gmail.com