

Propensity Score Analysis of Complex Survey Data:
Randomization Inference in Non-Randomized Studies

Frank Yoon

Submitted to Prof Elaine Zanutto

May 5, 2006

Introduction

Propensity scoring methods have been proposed as a way to make adjustments for covariates in observational studies, when we wish to perform inference on treatment effects. In randomized experiments the propensity score is known and equal for all subjects in treatment and control groups, and thus we are able to perform reasonable inference in such experiments, because the only stochastic quantities are those produced by randomization (Rosenbaum 2002). In observational studies, however, the propensity score is almost always unknown, and the comparison groups under study may not be comparable prior to treatment with respect to their covariates. When these pre-treatment characteristics are known to be different, such *overt bias* between treatment groups is removed by a variety of adjustment methods, such as matching or subclassification on the estimated propensity score.

Some studies may encounter more than two treatment groups, such as when subjects may be exposed to varying levels of treatment. In the case of *multiple treatment doses*, subclassification on the propensity score allows for the formation of comparison groups which are similar in their distribution of covariates, thus allowing for reasonable inference of treatment effects. In this paper I will explore methodology proposed by Zanutto, Lu, and Hornik (2005) and Zanutto (2006) for propensity score methodology on complex survey data.

Objective and Data

In September 1999 the U.S. Office of National Drug Control Policy (ONDCP) launched Phase III of the National Youth Anti-Drug Media Campaign, which ran through

the spring of 2004. An extensive range of media was used to disseminate the campaign messages to a national audience; advertising space had been purchased on television, radio, newspapers, magazines, billboards, transit ads, movie theaters, video rentals, and other venues. When Congress appropriated the funding for the campaign in 1998, it required that ONDCP conduct an evaluation of the campaign to assess its effectiveness of preventing drug use before it started as well as convincing youth who were occasional users to stop drug use. The National Survey of Parents and Youth (NSPY) was conducted for this matter, which measured covariates on survey participants, their past behavior with regard to drugs and other substances, and their exposure to the media campaign. Exposure to the media campaign was measured on a three-level index, which are taken as the multiple treatment doses in this paper. I will implement propensity score methods that accommodate this feature of the data from the evaluation. The purpose is to illustrate methodology, not to establish results regarding the effect of the media campaign.

The three-level general exposure index measures exposure to anti-drug ads, not limited to those of the media campaign, through many different media. Survey respondents were initially asked to answer four questions regarding exposure to anti-drug advertising on television, radio, billboards, public displays, movie theaters, and rental videos. Scores were defined for these responses, and were then combined into a continuous measure, which finally was categorized into three classes for the general index. The index is defined as the following responses to the hypothetical question "In recent months, how many times a month have you been exposed to anti-drug

advertisements?": (1) Less than 4 times per month; (2) 4 – 11 times per month; or (3) 12 or more times per month

The covariates in the survey data were recorded as categorical variables. A list of the covariates that were controlled for in this analysis is given in Table 1. For these covariates, I aimed to imitate the list as in Zanutto et al. (2005). One of the main differences in the complex survey data is that continuous covariates which were measured in the pilot study, such as those which measured the number of hours of television watched, are coded as categorical variables (where each level corresponds to a range of hours, for example). A second difference is that the complex survey data did not record after-school activities, so instead, I used a proxy variable which indicated whether or not a survey respondent had plans for college. Also, whereas the pilot study recorded the mother's education level, the complex survey data provided information about whether or not the survey respondent lived in a dual-parent household and about the education level of the parents.

The outcomes of interest are answers to questions related to intended marijuana and inhalant use:

- (a) How likely is it that you will use marijuana, even once or twice, in the next 12 months?
- (b) How likely is it that you will use marijuana nearly every month for the next 12 months?
- (c) How likely is it that you will use inhalants, even once or twice, in the next 12 months?

(d) How likely is it that you will use inhalants nearly every month for the next 12 months?

To each of these questions, the response options were: (1) I definitely will not; (2) I probably will not; (3) I probably will; and (4) I definitely will. It was expected that a higher level of exposure to anti-drug advertisements would lead to less intention to use marijuana or inhalants.

Methodology

The initial imbalance in covariates is summarized in Table 1. The p-values are based on the test of $H: \beta = 0$ of the logistic regression model of the covariate (dependent variable) on the exposure level (independent variable); significance of the parameter implies a dependency between exposure level and that covariate, indicating that the covariate is not balanced across treatment groups. Note that many of the covariates to be controlled for in this analysis are imbalanced.

The propensity score is modeled using McCullagh's ordinal logit model ("proportional odds model," Agresti 1996). For a survey respondent i , denote by the variable Z_i his/her exposure level to the campaign, and by the vector \mathbf{x}_i that respondent's observed covariates; denote by $j = 2,3$ the exposure level. The propensity score is modeled as:

$$\log\left(\frac{\Pr[Z_i \geq d]}{\Pr[Z_i < d]}\right) = \alpha_d + \beta^T \mathbf{x}_i \quad \text{for } j=2,3$$

For each respondent i , the propensity score is determined by the linear combination of the covariates $\beta^T \mathbf{x}_i = b(\mathbf{x}_i)$, which itself is a balancing score. It is on $b(\mathbf{x}_i)$ that subclasses of survey subjects are created. The idea is that we can consider each subject within a given

subclass as having been randomly assigned to an exposure level; across exposure levels within a given subclass, the distribution of covariates is balanced – conditionally on $b(\mathbf{x})$ the observed covariates \mathbf{x} and treatment doses Z are independent.

Subclasses are formed on the quintiles of $b(\mathbf{x})$, and the success of subclassification is assessed by checking for covariate balance within exposure levels. Specifically, for each categorical covariate, a logistic regression model is formed after subclassification: as before, the dependent variable is the covariate, while the independent variables are the exposure level and the quintile of the balancing score, with an interaction term. Significance of the main effect of exposure level or of the interaction term with the quintile suggests that the covariate is not appropriately balanced. In this case the propensity score model should be refitted, adding higher order terms and interactions. We are not worried about overfitting the propensity score model, as our aim is to achieve covariate balance rather than a parsimonious model. Tests for the main effect of exposure level or interaction term can be done by taking differences of deviances of a saturated versus a reduced model and comparing it to the χ^2 -distribution with degrees of freedom being the difference of the degrees of freedom between the two models.

Once subclasses are formed, we can estimate the response levels by a weighted mean involving the survey weights. In essence each survey respondent represents a number of people in the population who were exposed to the media campaign; the survey weight accounts for this. Let y_i be the response and w_i be the survey weight of subject i , let S_k be the set of observations in balancing score stratum k . For each exposure level $j = 1, 2, 3$, denote the weighted mean of the response vector \mathbf{y} by:

$$\bar{y}_{*j} = \sum_{k=1}^5 \left(\frac{\sum_{i \in S_k} w_i}{\sum_{k=1}^5 \sum_{i \in S_k} w_i} \right) \left(\frac{\sum_{i \in S_k} w_i y_i}{\sum_{i \in S_k} w_i} \right)$$

The preceding formula is a linear combination of ratio estimators, assuming unequal probability sampling without replacement with overall inclusion probabilities $1/w_i$. Thus, we can approximate the corresponding standard error as:

$$se(\bar{y}_{*j}) = \sqrt{\sum_{k=1}^5 \left(\frac{\sum_{i \in S_k} w_i}{\sum_{k=1}^5 \sum_{i \in S_k} w_i} \right)^2 s_k^2}$$

where

$$s_k^2 = \frac{n}{n-1} \sum_{i=1}^n \left(z_{ik} - \frac{1}{n} \sum_{j=1}^n z_{jk} \right)^2$$

and

$$\begin{aligned} z_{ik} &= \frac{w_i}{\sum_{i \in S_k} w_i} \left(y_i - \frac{\sum_{i \in S_k} w_i y_i}{\sum_{i \in S_k} w_i} \right) & i \in S_k \\ &= 0 & i \notin S_k \end{aligned}$$

With the point estimate and standard error, confidence intervals for the mean response within the exposure levels can be generated.

Note that this estimate of the standard error does not account for non-response or poststratification adjustments to the survey weights, which replication methods can account for (Zanutto, 2006).

Table 1: Covariates – Initial imbalance; distribution of counts following subclassification

Covariate	p-value*	Cell Counts		
		small $b(x)$	Overlap region	large $b(x)$
Age	0.351			
12-13		2	716	5
14-16		5	1376	13
17-18		1	512	9
Race/Ethnicity	0.282			
White/Other		6	1823	17
Black		1	380	5
Hispanic		1	401	5
Gender	0.080			
Male		8	1339	5
Female		0	1265	22
Urbanity	0.699			
Urban		1	829	10
Suburban		2	703	6
Town or Rural		5	1072	11
Dual Parent Home	0.195			
No		1	755	9
Yes		8	1849	18
Parent Education	0.087			
Less than HS		3	359	4
HS graduate		3	829	5
Some college		0	747	11
College grad		2	669	7
Avg. TV on weekday	>0.001			
0-1 hr		5	842	1
2-4 hrs		3	1356	17
5 + hrs		0	406	9
Avg. TV on weekends	>0.001			
0-2 hrs		5	727	0
3-6 hrs		3	1150	12
7 + hrs		0	727	15
#Days of MTV in past 30 days	>0.001			
Never		7	564	0
1-4 days		1	752	1
5-14 days		0	574	7
15-30 days		0	714	19
#Days of ESPN in past 30 days	>0.001			
Never		5	1167	3
1-4 days		3	686	2
5-14 days		0	397	0
15-30 days		0	354	22
College Plans	>0.001			
No		5	587	2
Yes		3	2017	25
Importance of religion	0.005			
Low		6	780	7
High		2	1824	20
Ever smoked cigarettes	0.039			
No 0		5	1718	13
Yes 1		3	886	14
Ever used marijuana	0.023			
No 2		5	1984	12
Yes 1		3	620	15
Used marijuana in past year	0.099			
No 0		5	2146	18
Yes 1		3	458	9
Ever used inhalant	0.024			
No 0		7	2526	17
Yes 1		1	78	10
Used inhalant in past year	0.158			
No 0		7	2550	27
Yes 1		1	54	0
Friends used marijuana regularly in past year	0.093			
No 0				
Yes 1		4	758	14
		4	1846	13

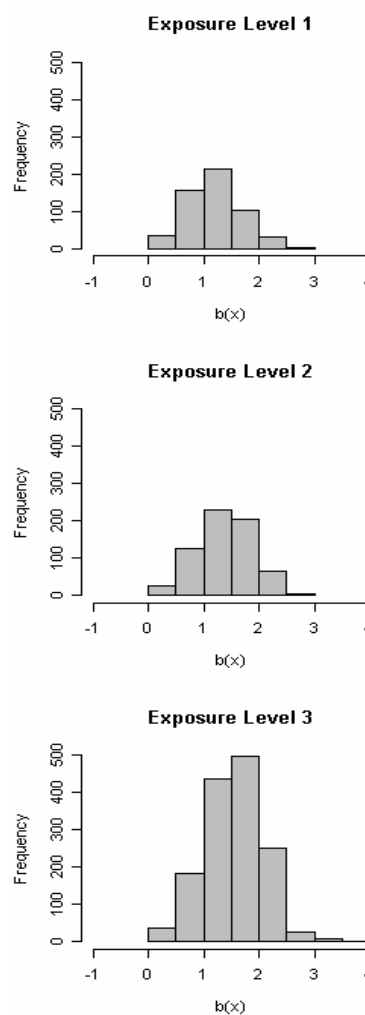
*p-value is based on the test of $H: \beta=0$, where β is the coefficient of the exposure effect of the logistic regression model testing for covariate imbalance.

Results and Discussion

My propensity score model included all the covariates as main effects with the following interaction terms: average number of television hours on weekdays with dual-parent household status; prior use of inhalants with gender; marijuana-using friends with importance of religion; expected college plans with parent education-level; marijuana-using friends with prior marijuana use; average number of television hours on weekends with expected college plans. From the model the estimated balancing score is obtained, from which subclasses are formed.

One limitation of the methodology is that we are restricted to regions of the balancing score that cover all exposure levels. This means that some extreme values of the balancing score will not be included in the estimation of response means. Visually we can inspect the overlap of the balancing scores among the exposure levels, to ensure that there is adequate coverage on which estimation can proceed. In my analysis, this region is restricted to the range $0.06199 \leq b(x) \leq 2.6113$. From Figure 1 it is seen that not many data were excluded from analysis, aside from the few observations that were in the upper range of $b(x)$. Cell counts for the levels of each covariate are shown in Table 1; the column denoted “small

Fig. 1: Balancing Score Regions



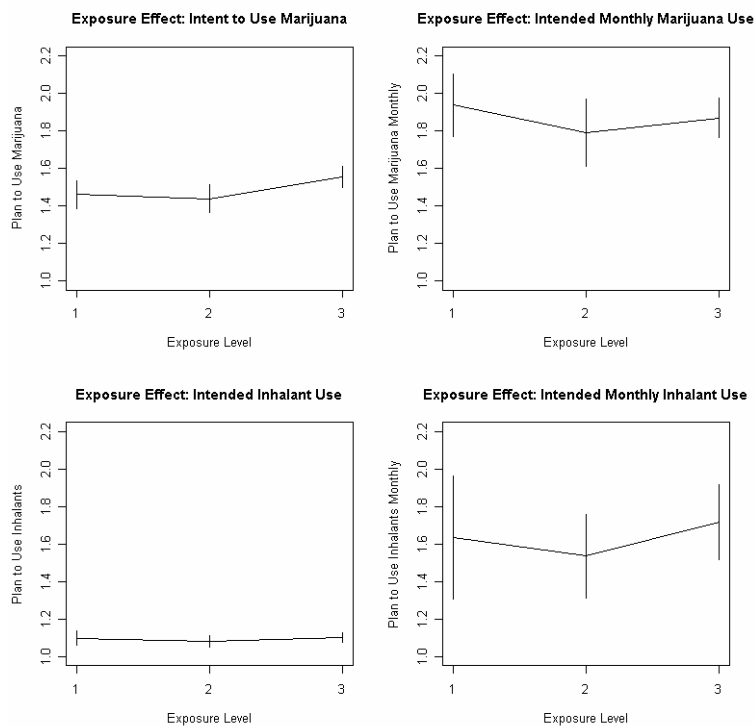
$b(x)$ ” indicates the range of the data that fall below $0.06199 \leq b(x)$, and the other denoted “large $b(x)$ ” indicates the range of data fall above $2.6113 \geq b(x)$.

		Exposure Level			Total
		1	2	3	
b(x) Quintile	1	182	133	205	521
	2	136	139	246	521
	3	105	134	282	521
	4	61	138	322	521
	5	51	104	367	521
Total		535	648	1422	2605

Table 2: Subclassification on b(x)

The subclassification is show in Table 2. Within each quintile there are representative subjects at each exposure level, indicating the good overlap that was achieved by subclassification on the balancing score.

Fig. 2: Effect* of Exposure to Media Campaign on Youth Intention to Use Marijuana/Inhalants



*bars indicate the 95% confidence interval based on the calculated standard error

In Figure 2 the estimated exposure effect is shown. It appears from this analysis that the media campaign had little effect on youths' intentions to use marijuana and inhalants. Note the relatively large margins of error in the confidence intervals for intentions to use these substances monthly; this is primarily due to the sparsity of responses for these questions – for example, if a survey respondent claimed that “I definitely will not” use marijuana occasionally (“even once or twice”) in the next 12 months, then his/her answer to the question of monthly use was deemed Not Applicable.

One issue we may wish to address in further analysis of these data is the distribution of exposure levels. Note from Table 2 the left-skew in the distribution – nearly 40% of the survey respondents were measured to have had a “high” (level 3) exposure to anti-drug advertisements. A reassessment of how the general exposure index was calculated may be important in future analyses. In particular it may be beneficial to consider how one would more accurately define the levels for a new exposure index, perhaps in order to account for those survey respondents who may have been exposed to much more advertising than others; as it stands, level 3 of the exposure index does not distinguish extreme observations.

As with many complex survey data, methods for dealing with non-response could be incorporated into an analysis such as this one. In this analysis the initial dataset contained 3561 observations, of which 934 were removed due to incomplete information in the 18 covariates that I wished to control for. As mentioned before, methods using replicate weights can account for extra variability due to non-response; in addition, multiple-imputation methods could be used to deal with missing values.

Conclusion

A key feature of propensity score methodology for covariate adjustment in observational studies is that it does not require a specified relationship between outcomes of interest (in this case, responses to questions on intended drug use) and the covariates. With propensity scores, one can simulate randomization in observational studies, the heavenly result being, as Fisher noted, that randomization is “the reasoned basis for inference in experiments.” Subclassification on the propensity score provides a natural framework to incorporate survey weights such that population-level inference is possible, based on survey samples. The statistical implications of complex survey design on propensity scoring methods in inference of treatment effects have not been fully explored in the literature; however, as it stands, there exists sufficient motivation to pursue work in this line of methodologies.

References

- Agresti, Alan (1996). An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc.: New York.
- Rosenbaum, Paul R. (2002). "Covariance Adjustment in Randomized Experiments and Observational Studies." Statistical Science (17:3).
- WESTAT (2004). "User's Guide for the Evaluation of the National Youth Anti-Drug Media Campaign." WESTAT: Rockville, Maryland.
<http://www.drugabuse.gov/dspr/westat/>
- Yanovitzky, I., Zanutto, E., Hornik, R. (2005). "Estimating causal effects of public health education campaigns using propensity score methodology." Evaluation and Program Planning (28:209-220).
- Zanutto, E. (2006). "A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data." *Journal of Data Science*, 67-91.
- Zanutto, E., Lu, B., and Hornik, R. (2005). "Using Propensity Score Subclassification for Multiple Treatment Doses to Evaluate a National Anti-Drug Media Campaign." *Journal of Educational and Behavioral Statistics*, 30: 59-73.