

**10<sup>th</sup> National Convention on Statistics (NCS)**  
EDSA Shangri-La Hotel  
October 1-2, 2007

**The Propensity Score Matching for Correcting Sample Selection Bias**

by

Emeterio S. Solivas, Girly M. Ramirez and Allan N. Manalo

For additional information, please contact:

Author's name	:	Emeterio S. Solivas
Designation	:	Associate Professor
Affiliation	:	University of the Philippines at Los Baños
Address	:	College, Laguna
Tel. no.	:	(06349) 536-2381
E-mail	:	<a href="mailto:essolivas@uplb.edu.ph">essolivas@uplb.edu.ph</a>
Co-author's name	:	Girly M. Ramirez/ Allan N. Manalo
Affiliation	:	University of the Philippines at Los Baños/ Accenture (Phil.)
Address	:	College, Laguna
Tel. no.	:	(06349) 536-2381
E-mail	:	

# **The Propensity Score Matching for Correcting Sample Selection Bias**

by

Emeterio S. Solivas, Girly M. Ramirez and Allan N. Manalo

## **ABSTRACT**

Development program evaluation studies often deal with the effect of the program intervention into the subjects of the program. Subjects are assigned into two groups, the program group and the control group that serves as the comparison group. Sometimes assigning of the subjects are not randomized, leading to bias results due to self-selection. Using matching on the propensity score may reduce the systematic bias present in the data. Moreover, the use of propensity score matching (PSM) provides more reliable and accurate results.

As an applied example, this paper used the data from baseline and endline evaluation surveys for the Early Childhood Development Program of the DSWD. The results showed that the Propensity Score Matching was successful in reducing the bias on the covariates.

**Keywords:** Selection bias, logistic regression, propensity score, standardized bias

## **I. Introduction**

In the evaluation of a development program, the “before and after type” method of analysis is usually employed. This paired comparison method can be easily used in the usual scenario where the subjects that are observed before the program are the same subjects that are observed after the program. Situations may arise where the subjects observed before the program cannot be the same subjects observed after the program. For instance, in the early childhood development program of the DSWD, the baseline evaluation measured sample of children below 72 months. The program was carried out for five years. At the endline evaluation, the same children in the baseline were no longer on early childhood. Thus, another group of children below 72 months had to be measured. A bias in the use of the usual comparisons between the baseline and endline groups may arise when there are some systematic differences between the two groups. These systematic differences are due to some covariates that are different between the two groups. Failure to account for these differences in baseline characteristics can lead to biased results (Posner et al, 2002).

In a situation like this, where non-randomization of sampling units are not performed, selection bias generally occur due to individual self-selection. The bias may be partially avoided by incorporating the covariates into the study such as matching, or into the estimation of the program effect such as covariance adjustment. However, covariance adjustments are difficult to perform when there are large number of covariates to adjust. One way of handling this situation

is by statistical matching. Statistical matching matches the observations that are similar between the two groups. With the use of statistical matching, subjects in the baseline group are paired with subjects from the endline group on some covariates that cause the systematic difference between the groups thereby reducing the systematic bias.

There are several techniques developed under the statistical matching depending on the measurement of similarity used in matching the units, such as the Mahalanobis distances. Another is the use of what is called the propensity score which is a scalar summary of the covariates. The propensity score for an individual is the probability of being assigned to the program group conditional on the individual's covariate values. Intuitively, the propensity score is a measure of the likelihood that a person would have been in the program using only their covariate scores (D'Agustino, 1998).

Propensity Score Matching (PSM) is a way of improving the ability of the regression to generate accurate causal estimate by virtue of its nonparametric approach to the balancing of the covariates between the program and control group, which removes bias due to observable variables. It provides an estimate of the effect of a program variable on an outcome variable that is largely free of bias arising from association between status and observable variables.

For applied example, this study will use the data from the Early Childhood Development Program Evaluation Surveys. In this evaluation study, the children before the program were compared to the children after the program. The units of study are the households.

## **II. Theoretical Framework**

Suppose a researcher wants an answer to the counterfactual question like what would happen to a child's development if he or she were subject to an alternative development program  $P$ ? In order to answer this question, there is a need to measure the difference between the child's development with and without  $P$ .

Let  $Y_i^P \equiv$  development index of child  $i$  in the program group, and

$Y_i^C \equiv$  development index of child  $i$  in the control (with program) group.

Then  $Y_i^P - Y_i^C \equiv$  effect of  $P$  for a given child. The problem is that no child can be with  $P$  and  $C$  at the same time.

Define  $E[Y_i^P - Y_i^C] \equiv$  mean effect of  $P$  on the given child. Looking at the average effect of  $P$  on the children, then the mean difference  $D$  between the  $P$  children and  $C$  children is

$$\begin{aligned} D &= E[Y_i^P | P] - E[Y_i^P | C] = E[(Y_i^P | P - Y_i^C | C)] \\ &= \{E[Y_i^P | P] - E[Y_i^C | P]\} + \{E[(Y_i^C | P) - E[Y_i^C | C]]\} \end{aligned}$$

The term  $t_i = \{E[Y_i^P | P] - E[Y_i^C | P]\} \equiv$  average difference the  $P$  makes among  $P$  children;

$$d_i = \{E[(Y_i^C | P) - E[Y_i^C | C]]\} \equiv \text{selection bias due to systematic differences between P children and C children.}$$

Define  $T \equiv$  program of the study ( $1 = P, 0 = C$ ). Let  $p(X) \equiv \text{Prob}[T = 1 | X]$ . Then the term

$$\{E[(Y_i^C | P) - E[Y_i^C | C]]\} \text{ implies that } E[(Y_i^C | p(x), T) - E[Y_i^C | p(x), C]] = 0.$$

In order to investigate the effect of  $P$ , the strategy is to find ways of reducing the term  $\{E[(Y_i^C | P) - E[Y_i^C | C]]\}$ . One way is by the Propensity Score Matching technique.

By using the estimate of the propensity score and then comparing observations on children that have similar propensity score, then the observable selection bias is reduced and the program effect  $\tau$  is isolated from  $D$ . (Rubin, 1997).

*Key Assumption: Conditional Independence*

Foster (2003) said that propensity score matching rests on a critical assumption known as the conditional independence assumption. The conditional independence assumption states that for a given set of covariates, participation is independent of potential outcomes. Among individuals with the same characteristics used for matching, the model assumes that these individuals are sorted into different programs as if randomly assigned.

### III. Methodology

#### 3.1. The Logistic Regression and Propensity Score

A propensity score is obtained as the probability score of an individual from the fitted simple logistic regression model of the two groups as the dependent variable to the set of covariates. Logistic regression is applied when the dependent variable is dichotomous.

In this study, the dependent variable is  $T \equiv$  program of the study ( $1 = P, 0 = C$ ). Let  $p(X) \equiv \text{Prob}[T = 1|X]$ . The logistic regression is defined as

$$p(\mathbf{x}) = \frac{e^{a + \sum b_i X_i}}{1 + e^{a + \sum b_i X_i}}$$

where:  $a$  = constant of the equation; and  
 $\beta_i$  = coefficients of the covariates  $X_i$  ;

Alternatively, the equation above can take on a linear form through logit transformation, also known as logit regression equation,

$$L(\mathbf{x}) = \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = a + \sum b_i X_i$$

The propensity score is the probability of an individual to participate in the program given his observed covariates  $X$ . This can be computed by substituting the values of the covariates in the logistic regression equation.

### **3.2. Nearest-Neighbor (NN) Matching**

This is the most straightforward matching estimator. The individual from the comparison group is chosen as a matching partner for a program individual that is closest in terms of propensity score. The following steps were done in the nearest-neighborhood matching:

1. The computed propensity scores were grouped according to program assignment.
2. The propensity scores were then arranged randomly.
3. The first household in the program group was matched to the household in the control group with the nearest propensity score. The matched households were then discarded from the list.
4. Step 3 continued until the last household in the program group finds its matched household in the control group.

### 3.3. Assessing the Matched Data

#### 3.3.1. The standardized bias

One suitable indicator to assess the distance in marginal distribution of the covariate is the standardized bias suggested by Rosenbaum and Rubin (1985). The standardized bias is computed as:

$$SB = \frac{100(\bar{x}_C - \bar{x}_P)}{\sqrt{(s_C^2 + s_P^2)/2}}$$

where:  $\bar{x}_C$  = mean of the control group;  $\bar{x}_P$  = mean of the program group,

$s_C^2$  = variance of the control group;  $s_P^2$  = variance of the program group.

For binary data, the standardized bias is computed as:

$$SB = \frac{100(p_C - p_P)}{\sqrt{[p_P(1 - p_P) + p_C(1 - p_C)]/2}}$$

where:  $p_C$  = proportion of the covariate in the control group, and

$p_P$  = proportion of the covariate in the program group.

#### 3.3.2. Bias Reduction

The reduction of bias due to matching is computed as

$$BR = 100\left(1 - \frac{B_M}{B_O}\right)$$

where:

$$B_M = \frac{100(\bar{x}_{MC} - \bar{x}_{MP})}{\sqrt{(s_{MC}^2 + s_{MP}^2)/2}} \equiv \text{standardized bias after matching, and}$$

$$B_O = \frac{100(\bar{x}_{OC} - \bar{x}_{OP})}{\sqrt{(s_{OC}^2 + s_{OP}^2)/2}} \equiv \text{standardized bias before matching.}$$

where: subscript  $M$  denotes after matching;  $O$  denotes before matching.

Note: Most empirical studies show that a bias reduction of 3 to 5% is seen as sufficient.

### **3.3.3. Test on Means**

According to Caliendo and Kepeinig (2005) the two sample t-test can also be used to check if there are significant differences in covariate means for both groups. Before matching, differences between the groups are expected; but after matching, the covariates should be balanced in both groups and hence no significant differences should be found. The t-test might be preferred if the evaluator is concerned with the statistical significance of the results.

## **4. Applied Example: ECD Program Evaluation**

This study used the propensity score matching in the ECD Project in Region 6 to eliminate the possible sample selection bias since the data was from a survey study. The data used was from the Early Childhood Development conducted in 2000 -2005.

There were 73 variables considered in the study, 64 of which were categorical variables. The variables were divided into 4 columns according to its category as shown in Annex Table 1. The first column shows the household characteristics whereas the second column contains the characteristics of the mother. The third column contains the father's characteristics, and the fourth column comprises the accessibility of the office to the household.

### ***Sampling Design***

Three regions namely Region 6 (Western Visayas), Region 7 (Central Visayas), and Region 12 (Central Mindanao) were the program areas and Region 8 (Eastern Visayas) has been designated as the control area. The program areas are regions in which the ECD intervention was implemented and the control area was the region that has no ECD program. A sample size of 8,000 households from the four regions was surveyed. And the following procedure has been implemented:

1. A responsible adult member of the household was asked if there were children or pregnant women in the household. If there were none, the interviewer went to the next household.
2. If the household has children, the age of the youngest child was first ascertained to ensure including children who are 0-6 years of age. If there are none, they go to the next household.
3. The resident status of eligible children and pregnant women were then verified. Residents were defined as those adults who had stayed for at least 6 months in the barangay and

children who were born to resident parents. If no one was a permanent resident, the interviewer proceeded to the next household.

4. If the household has permanent residents who were under 7 years old, they were included in the study.

#### *Data Used in this PSM Study*

Since the objective of this paper is to illustrate the use of PSM to reduce sample selection bias, only the data from Region 6 were used. A similar investigation was done using the data from all the regions in the program.

#### ***Variables Used***

There were three indicators considered in the study to assess the effectiveness of the ECD Project. These were used to identify if there was an improvement in the intervention. The indicators include:

##### *1. Hemoglobin Level and Anemia*

Blood samples were obtained on the eligible children (0 month and older), the cyanmethemoglobin technique was used in determining hemoglobin readings. Classification whether anemic or normal was based on the following cut-off levels by the World Health Organization (1972).

##### *2. Anthropometric indicators of nutrition and health status*

Nutritional status of children- whether the children were stunted/wasted/underweight or not were ascertained using the Food and Nutrition Research Institute – Philippine Pediatric Society (FNRI-PPS) reference standard for Filipino Children and the National Center for Health Statistics (NCHS) standard. Z scores of the children were computed. Stunting and wasting are present if the Z-scores are more than 2 standard deviations below the reference population mean for height-for-age, and weight-for-height.

##### *3. Over-all Developmental Index*

Refers to the development of children in seven domains, Gross Motor, Fine Motor, Self-Help, Receptive Language, Expressive Language, Cognitive, Socio-emotional. Scaled scores for each domain were derived and used to classify developmental indices of children. Table 1 shows the variables that were used to assess the performance of the ECD Project.

Table 1. Indicator variables in the ECD Project.

Indicator Variable	Description
%anemic	Proportion of anemic children in Region 6
%below average dev't	proportion of children whose dev't index was below average
Height for Age	proportion of normal children in height for age measurement
Weight for Age	proportion of normal children in weight for age measurement
Height for Weight	proportion of normal children in height for weight measurement

### ***Data analysis and PSM***

#### *Preliminary Statistical Analysis*

Prior to the logistic regression analysis, there is a need to measure the association between the response variable which is the ECD assignment and the explanatory variables. The measure of association used between the response variable and the categorical variable was Cramer's V while for the continuous variables; the point biserial coefficient was used. Testing of association was done to determine variables that have been related to the ECD Assignment.

After the testing for association, the PROC LOGISTIC procedure of STATA 8.2 was used in building the logistic regression model using the variables that are not related to the ECD assignment.

The propensity score was computed using the MS Excel, by substituting the values of the covariates of each individual household. Afterwards, using the C program on nearest-neighborhood matching, the matched households were selected.

The mean and variance of the covariates before and after matching were computed to calculate the standardized bias and the reduction of bias. The t-test was performed to test if there was a difference in the mean between the baseline and the endline study.

### ***Results and discussion***

#### *Correlation of ECD assignment with covariates*

Correlation analysis was performed to identify the variables that are not affected by the treatment assignment. There were seventy-three variables that were correlated to the ECD assignment. Forty-seven of these variables were found to be significantly correlated to ECD Assignment.

### *Logistic Regression on the ECD Project Participation*

Among the seventy-three explanatory variables, 26 variables were not correlated with the ECD assignment. Some of the variables were deleted because of multicollinearity and some because it had p-values equal to 1.0. Eight variables remained in the model and its intersections and quadratic terms were added. Table 2 shows the estimates, standard error, the p-values and the confidence interval of the logit model for ECD assignment at  $\alpha = 10\%$ .

Table 2. Logit regression model for ECD assignment

Covariate	Parameter Estimate	Standard Error	z	p-value	95% Confidence Interval	
CHILDREN6	5.3759	0.9096	5.91	0.000	3.593	7.158
					-	-
APPCOUNT	-0.2402	0.0786	-3.06	0.002	0.394	0.086
					-	-
CHILDREN6*CHILDREN6	-1.1599	0.3425	-3.39	0.001	1.831	0.488
Constant	1.0677	0.3006	3.55	0.000	0.478	1.657

Pseudo  $R^2 = 0.4608$

where: CHILDREN6 = no. of children below 6 y/o; APPCOUNT = appliance count.

### ***Propensity Score Matching***

The propensity score was computed by substituting the corresponding value of the predictors on the logistic regression equation.

The propensity score during the baseline and the endline was computed. There are 1649 and 765 cases of computed propensity score for the baseline and endline, respectively. After the propensity scores were randomized, the program created in Turbo C, performed the nearest-neighborhood matching. Table 3 shows some of the highlights of the matching procedure.

Table 3. Some Matched Household

baseline	p-score	endline	p-score
16	0.9980	200	0.9975
55	0.9936	9	0.9936
179	0.9936	835	0.9938
192	0.2987	53	0.3513
513	0.9869	855	0.9869

Household number 16 and 179 at the baseline whose propensity score were 0.9980 and 0.9936 were matched to household number 200 and 835 at the endline whose propensity score equal to 0.9975 and 0.9938, respectively. Household number 55 matched to 9 and 513 that was matched to 855 were exact matches. Exact matches are cases in which the difference of the two propensity score is 0. There are 721 exact matches on the data. Household number 192 in baseline and 53 in endline had the greatest difference on propensity score equal to 0.0526.

### ***Reduction Bias Before Matching***

Table 4 shows the descriptive statistics of the unmatched covariates. There was a slight difference between the covariates in the baseline and endline. *Children6* had the least standardized bias while *AppCount* had the highest standardised bias. The p-value cannot be rejected even at 5% level, indicating that there is no difference on the mean of the covariates.

Table 4. Descriptive Statistics of the Unmatched Covariates

Covariates	Baseline		Endline		Standardi sed Bias	t-test (p- value)
	Mean	SD	Mean	SD		
Children6	1.2965	0.8012	1.3556	0.7513	-7.5979	0.0861
AppCount	1.9788	3.0029	2.2013	2.7683	-7.7055	0.0827

### ***Reduction Bias After Matching***

The descriptive statistics of the matched covariates are shown on Table 5. Compared to the unmatched data, the means of the covariates in the baseline and endline were pretty similar as if the same households were used in the study. The standardized bias was reduced but still the *Children6*<sup>2</sup> had the least standardized bias, while *AppCount* had the highest standardized

bias. The p-values were larger in the matched covariates indicating higher level of acceptance that the means of the baseline and endline were equal.

Table 5. Descriptive Statistics of the Matched Covariates

Covariates	Baseline		Endline		Standardised Bias	t-test (p-value)
	Mean	SD	Mean	SD		
Children6	1.3412	0.7361	1.3556	0.7513	-1.9333	0.7054
AppCount	2.1477	2.7686	2.2013	2.7683	-1.9359	0.7050

### ***Bias Reduction***

Table 6 shows the percentage reduction of the standardized bias before and after matching. Bias was reduced on every covariate. The *Children6*<sup>2</sup> had the least reduction of bias while *AppCount* had the highest reduction of the bias.

Table 6. The Reduction of Bias of the Covariates

Covariates	Standardized Bias		Bias Reduction
	Unmatched	Matched	
Children6	-7.5979	-1.9333	74.5546
AppCount	-7.7055	-1.9359	74.8764

## **V. Concluding Remarks**

The main objective of the study is to use Propensity Score Matching to reduce the self-selection bias involving the non-randomization in nature of a development program. Propensity score matching is one way of reducing the selection bias. This paper showed that the Propensity Score Matching was successful in reducing the bias on the covariates.

The logistic regression is just one method of obtaining propensity scores. Another method may be the use of probit model or discriminant model to predict the likelihood of being in the program group. The use of the different matching algorithm such as Caliper Matching, Mahalanobis Metric Matching, Kernel Matching, and Stratified Matching are also alternatives to the matching procedure.

## References

- Agostino Jr., Ralph B. Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-randomized Control Group. *Statistics in Medicine, Statist. Med.* 17, 2265-2281 (1998).
- Baser, Onnur. Using Propensity Score Matching Techniques for Average Treatment Effect: Application to Triptan Use.
- Barry, J. T. 1998. An Investigation of Statistical Matching. *Journal of Applied Statistics.* 15:3:275-283.111
- Behrman, Jose, Paulita Duazo, Sharon Ghuman, et al. Evaluating the Early Childhood Development Program in the Philippines. March 2005.
- Caliendo, Marco and Sabine Kopeinig. Some Practical Guidance for the Implementation of Propensity Score Matching. May 2005.
- Dehejia, Rajeev H. and Sadek Wahba. Propensity Score Matching Methods for Non-Experimental Causal Studies. NBER Working Paper 6829 (1998)
- Foster, EM. Propensity Score Matching: An Illustrative Analysis of Dose Response. *MEDICAL CARE.* Volume 41, Number 10, pp 1183-1192.
- Guo, Shenyang, Richard Barth and Claire Gibbons. Introduction to Propensity Score Matching: A New Device for Program Evaluation. Workshop presented at the Annual Conference of the Society for Social Work Research. New Orleans, January 2004.
- Rosenbaum, Paul R. and Donald B. Rubin, The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70 (April 1983), 41-55.
- Smith, Jeffrey and Petra E. Todd. Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods. *The American Economic Review* 91 (May 2001), 112-118.

**Annex Table 1. Covariates Considered in the Study**

<b>Household Characteristics</b>	<b>Mother's Characteristics</b>	<b>Father's Characteristics</b>	<b>Accessibility of the Office</b>
Province	Age	Age	Health center
Household type	Educational attainment	Educational attainment	Private clinic
Household size	Have has access to pre-natal care?	Have/ had diabetes?	Private dental office
Total income	Ever receive pre-natal care?	Have/had heart disease?	Government hospital
Number of rooms	Have/ had diabetes?	Have/ had cancer?	Private hospital
Number of Meals	Have/ had heart disease?	Have/ had tuberculosis?	Pharmacy
Number of children below 6yrs old	Have/ had cancer?	Have/ had asthma?	Public day care
Appliance count	Have/ had tuberculosis?	Have/ had hypertension?	Private day care
Appliance ratio	Have/ had asthma?	Have/ had goiter?	Public pre-school
Attended PES	Experience heavy menstrual bleeding?	Have/ had anemia?	Private childminding center
Quarrel?	Have/ had hypertension?	Have/ had hepatitis?	Public childminding center
Use iodized salt	Have/ had goiter?	Arthritis	Public elementary school
Treat water?	Have/ had anemia?	Have/ had urinary problem?	Private elementary school
Type of light	Have/ had hepatitis?	STD	
Type of fuel	Have/ had arthritis?	Smoke cigarette?	
Toilet location	Have/ had urinary problem?	Number of stick	
Toilet type	Have/ had pregnancy problem?	Alcoholic drink	
Garbage disposal	Smoke cigarette?	Drink alcoholic beverage?	
Type of roof	Number of stick		
Type of wall	Drink alcoholic beverage?		
Type of floor	Drinking frequency		