Genetic Matching: An Efficient Algorithm to Adjust Covariate Imbalance for Data Analysis and Modeling

Kao-Tai Tsai and Karl E. Peace
Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, Georgia

Abstract - *In causal-effect relationship research, similarity of groups being compared in terms of covariates or patient/disease characteristics is critical to ensure fairness of the comparison and unbiasedness of the findings. When dissimilarity is suspected, one can either adjust for imbalance or match the groups according to certain important covariates or characteristics. Regression analysis is commonly used to adjust the imbalance and matching techniques are usually used to match subjects between groups. Diamond and Sekhon [2] proposed a genetic matching algorithm to maximize the covariate balance. We describe the theory and conduct a simulation study to compare the relative performance of propensity score matching, Mahalanobis matching, and Genetic matching. Generally, Genetic matching achieves better covariate balance and produces more stable and unbiased treatment effect estimates. We also apply Genetic matching to a clinical study to investigate the treatment effects on rheumatoid arthritis.*

Keywords: propensity score, Mahalanobis matching, Genetic matching, Robbins-Munro stochastic approximation, randomized controlled clinical trials.

1 Introduction

In causal-effect relationship research, similarity of groups being compared in terms of covariates or patient/disease characteristics is critical to ensure fairness of the comparison and unbiasedness of the findings. When dissimilarity is suspected, one can either adjust for imbalance or match according to certain important covariates or characteristics. Regression analysis is commonly used to adjust for imbalance and matching techniques are usually used to match subjects between comparison groups. Therefore, matching has become an important method of causal-effect relationship inference in many fields including biomedicine, economics, social science, and statistics, to name a few.

Several matching procedures have been proposed in the literature by researchers since the early 1970s. Important differences between these proposed methods are the efficiency of the algorithms utilized and the effectiveness of the methods to reduce imbalance prior to subsequent inferences.

Propensity score matching based on logistic regression and multivariate matching based on Mahalanobis distance are among the more commonly used methods for this purpose. Several variations and combinations of these methods are also used frequently by practitioners.

When covariates have spherical or ellipsoidal distributions, these methods generally perform quite well; however, these methods can perform poorly when the distributions deviate substantially from this family of distributions. Therefore, it is highly desirable to have alternatives that can perform well even when the distributions of the covariates deviate substantially from this family of distributions.

Diamond and Sekhon [2] and Sekhon [15] proposed a genetic matching algorithm that imposes additional properties and generalizations to propensity score and Mahalanobis matching methods and maximizes the balance of observed covariates between the subject groups being compared. The method is nonparametric and does not depend on knowing or estimating the propensity score; however, when a propensity score is incorporated, the method can sometimes be improved by taking advantage of the information embedded in the propensity scores.

Genetic matching has been successfully utilized in social sciences to investigate causal-effect relationships (Diamond and Sekhon [3], Hopkins [5]); however, it has rarely been used in biomedical research to investigate between treatment group differences with covariate imbalance among subjects in the groups.

As stated by Peto, et al. [7], "There is simply no serious scientific alternative to the generation of large-scale randomized evidence.If trials can be vastly simplified, $\cdots$, and thereby made vastly larger, then they have a central role to play in the development of rational criteria for the planning of health care throughout the world." Recruitment of a large number of eligible patients from

a general population is both a major strength and weakness of large pragmatic trials.

Deliberately broadening the entry criteria means that the overall result can be difficult to apply to particular groups. However, in modern medical practice, physicians are often interested in individualized medicine and how best to use results of randomized clinical trials to maximize the wellbeing of each patient. Therefore, proper analyses of targeted subgroups of patients to investigate treatment efficacy has become increasingly necessary if heterogeneity of treatment effects is likely to occur.

Theoretically, the covariates of subjects should be well balanced in randomized controlled trials. However, in actual practice with small to moderate sample sizes, it is not uncommon to find subgroups of patients under study with covariate imbalance. This issue is a particular concern in many observational studies with long-term follow-up due to subject attrition. Therefore, it is critical to ensure similarity between subjects on important covariates in order to make the efficacy comparison of treatments meaningful and unbiased.

In section 2 of this article, we describe the theory of the propensity score, Mahalanobis distance matching and Genetic matching methods. In section 3, we describe a simulation study we conducted to compare the relative performance of these matching methods. In section 4, we apply Genetic matching to a dataset from a clinical study to investigate the relative effectiveness of two treatments for rheumatoid. Discussion and conclusions are presented in section 5.

2 Theory of Propensity Score, Mahalanobis Distance, and Genetic Matching

2.1 Propensity Score Matching

The concept of propensity scores is thoroughly discussed by Rosenbaum and Rubin [10] as well as by other authors. In the following, we describe a few key points for analytical purposes. Let Y_{i1} denote the response of the active treatment of subject $i, (1 \leq i \leq N)$, and Y_{i0} denote the response of the control treatment of subject i. Let X_i denote the vector of covariates associated with subject i and $T_i = 1(0)$ if subject i receives active (control) treatment. The observed outcome for subject i is then $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

If subjects were well randomized between treatment and control groups, then

$$E(Y_{ij}|T_i = 1) = E(Y_{ij}|T_i = 0), \quad j = 0, 1, \qquad (1)$$

even though $E(Y_{i0}|T_i = 1)$ of the treated group and $E(Y_{i1}|T_i = 0)$ in the control group cannot be estimated from the data since each subject can receive only either control or active treatment, but not both.

Under the well-randomized situation, the average treatment effect can be estimated using the observed data by

$$\tau = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) = a_1\tau_1 + b_1\tau_0, \quad (2)$$

where $a_1 > 0, b_1 > 0, a_1 + b_1 = 1$, and

$$\begin{aligned} \tau_1 &= [E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1)], \\ \tau_0 &= [E(Y_{i1}|T_i = 0) - E(Y_{i0}|T_i = 0)] \end{aligned} \qquad (3)$$

are the (unobserved) treatment effects from the treated and control groups, respectively.

When imbalance in covariates is suspected between the patient groups under study, proper matching of covariates is needed prior to subsequent inference in order to obtain a fair estimate of treatment effect or difference. Given covariate X_i, and following the results of Rubin [12, 14], one can show that

$$E(Y_{ij}|X_i, T_i = 1) = E(Y_{ij}|X_i, T_i = 0). \qquad (4)$$

Therefore, the treatment effect of the treated group can be estimated by

$$\tau_1 = E_{\{X_i|T_i=1\}}\{E(Y_i|X_i, T_i = 1) - E(Y_i|X_i, T_i = 0)\} \qquad (5)$$

where the expectation is taken over $\{X_i|T_i = 1\}$.

Define the propensity score as

$$e(X_i) = P(T_i = 1|X_i) = E(T_i|X_i), \qquad (6)$$

namely, the probability of patient i being assigned to active treatment given the covariate. Assume, given the subjects covariates, treatment assignments are not deterministic and are independent among study subjects, Rosenbaum and Rubin [9] had shown that

$$\begin{aligned} \tau_1 &= E_{\{e(X_i)|T_i=1\}}\{E(Y_i|e(X_i), T_i = 1) \\ &- E(Y_i|e(X_i), T_i = 0)|T_i = 1\}, \end{aligned} \qquad (7)$$

where the expectation is taken over $\{e(X_i)|T_i = 1\}$, and τ_0 can be estimated similarly. Therefore, the average treatment effect can be estimated by combining the results of τ_1 and τ_0. More details about the propensity score can be found in Rosenbaum [9] in addition to the papers mentioned herein.

Let $X_i = (x_{i1}, x_{i2}, \cdots, x_{ik})'$ and $m \leq k$ be the vector of covariates. A common method to estimate $e(X_i)$ is via the logit function, i.e.,

$$\text{logit}(e(X_i)) = \beta_0 + h_1(\eta_{1i}) + h_2(\eta_{2i}), \qquad (8)$$

where h_1 and h_2 are known functions and $\eta_{1i} = \sum_{r=1}^{m} f_r(x_{ir})$, $\eta_{2i} = \sum_{r,q=1}^{m} f_r(x_{ir}) f_q(x_{iq})$ represent the main effects and interactions, respectively. The parameters in Eq(1) can be estimated using MLE. Goodness-of-fit can be checked graphically via Landwehr, *et al* [6] or Tsai [16].

According to Rosenbaum & Rubin [10], it is advantageous to sub-classify or match not only on $e(x)$ but for other functions of x as well. In particular, such a refined procedure may be used to obtain estimates of the average treatment effect in a subpopulation defined by the components of X; for example, gender or different disease classifications.

2.2 Mahalanobis Matching and Genetic Matching

Given two covariates, X_i and X_j, the Mahalanobis and Genetic Matching are defined as following in terms of the distance between the covariates

$$md(X_i, X_j) = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{1/2}, \quad (9)$$

and

$$gmd(X_i, X_j) = \{(X_i - X_j)' S^{-1/2} W S^{-1/2} (X_i - X_j)\}^{1/2}, \quad (10)$$

respectively, where $S^{1/2}$ is the Cholesky decomposition of the covariance matrix of X, and W is a diagonal positive definite weight matrix. The elements of W are chosen to simultaneously minimize the distributional difference and location difference of covariates between the treatment and control groups based on the Kolmogorov-Smirnov test and t-test, respectively (Sekhon [15]).

The conventional test of covariate balance based on the t-test focuses only on location and can miss distributional differences between covariates. On the other hand, the Kolmogorov-Smirnov test compares distributional differences and can miss differences in locations. By combining these two tests, the covariates can be better matched in both location and other properties of the distributions.

3 Comparison of Matching Methods - a Simulation Study

3.1 Design of a Simulation Study

To investigate the performance of various matching methods, a simulation of 500 iterations was conducted under various scenarios. Specifically, the simulation plan was designed as follows:

1. Sample size: assume equal sample size ($N = 20, 30, 50, 100$) between treatment and control groups.

2. Assume 3 covariates (x_{i1}, x_{i2}, and x_{i3}) will be matched between treatment and control groups. The covariates were assumed to have somewhat different distributions between treatment and control. Four different distributions were assumed and are shown in the following table. They consist of standard normal distributions with possibly different means and variances, or contaminated normal distributions with either symmetric or asymmetric contaminations from either tail. The list of distributions is shown in the table below.

X_i	Group	*F:#1*	*F:#3*
x_{i1}	treated	$N(0,1)$	$0.9N(1,1) + 0.1N(1,3)$
	control	$N(0,1)$	$0.9N(0,1) + 0.1N(0,3)$
x_{i1}	treated	$N(0,1)$	$0.9N(0,2) + 0.1N(0,3)$
	control	$N(0,1)$	$0.9N(1,2) + 0.1N(1,3)$
x_{i1}	treated	$N(0,1)$	$0.9N(1,3) + 0.1N(1,4)$
	control	$N(0,1)$	$0.9N(0,3) + 0.1N(0,4)$
X_i	Group	*F:#2*	*F:#4*
x_{i1}	treated	$N(1,1)$	$.9N(1,1) + .1\|N(1,3)\|$
	control	$N(0,1)$	$.9N(0,1) + .1\|N(0,3)\|(-1)$
x_{i1}	treated	$N(0,2)$	$.9N(0,2) + .1N(0,3)$
	control	$N(1,2)$	$.9N(1,2) + .1N(1,3)$
x_{i1}	treated	$N(1,3)$	$.9N(1,3) + .1\|N(1,4)\|(-1)$
	control	$N(0,3)$	$.9N(0,3) + .1\|N(0,4)\|$

3. The response variable (Y) was assumed to follow two different models. The first model is

$$Y_i = \text{treatment effect} + \sum_{j=1}^{3} x_{ij} + error, \quad (11)$$

and the second model is

$$Y_i = \text{treatment effect} + \sum_{j=1}^{3} x_{ij} + \sum_{j \neq k=1}^{3} x_{ij} x_{ik} + error. \quad (12)$$

The treatment effect difference between treatment and control is assumed to be a constant, e.g., 1. The purpose of assuming two different models is to compare these methods when the model is incorrectly specified.

4. The statistical methods to be compared are:

   (a) Empirical mean difference,

   (b) Least squares (LS) fit (assuming the first model is correct),

   (c) LS fit (assuming the second model is correct),

   (d) Matching on the propensity score,

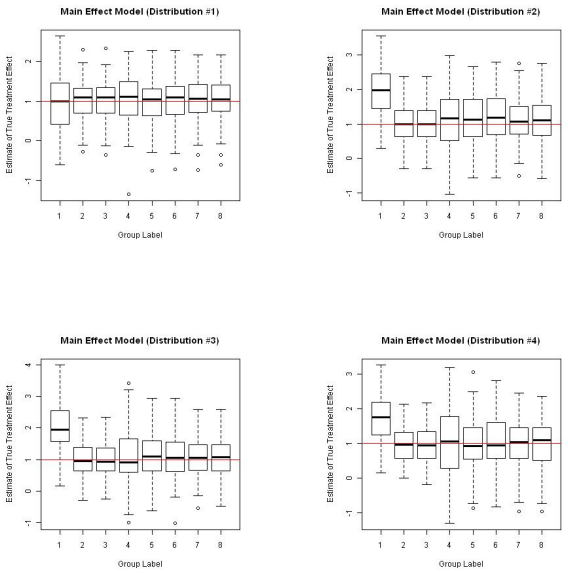

Figure 1: Estimation of treatment effect (=1): Main effect only. (Labels 1-8 = a-h of item 4 in Sec. 3.1)

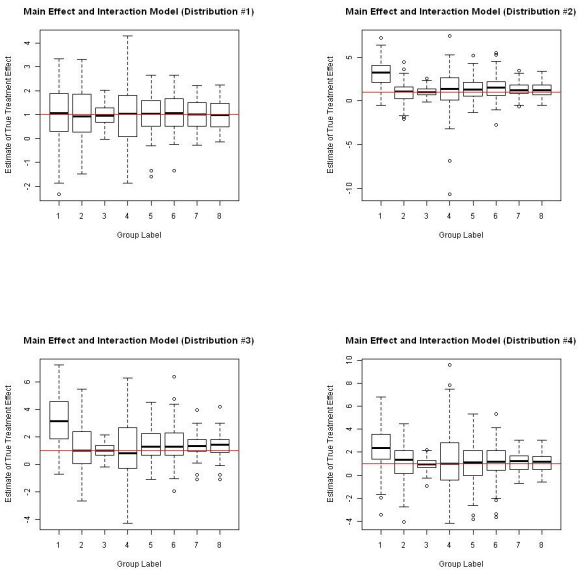

Figure 2: Estimation of treatment effect (=1): With interactions. (Labels 1-8 = a-h of item 4 in Sec. 3.1)

(e) Matching on x_{i1}, x_{i2}, and x_{i3} with all available data,

(f) Matching on x_{i1}, x_{i2}, x_{i3}, and the propensity score with all available data,

(g) Matching on x_{i1}, x_{i2}, and x_{i3} but excluding data in either tail outside of 2 times MAD (MAD is defined as $1.483\ \text{med}_i\{|x_{iu} - \text{med}_j(x_{ju})|\}$) from the median for each covariate (to mimic Tukey's robust trimmed estimate),

(h) Matching on x_{i1}, x_{i2}, x_{i3}, and the propensity score but excluding data in either tail outside of 2 times MAD from the median for each covariate.

5. Two criteria for comparisons are examined:

   (a) The estimates of the true treatment effect and the variation of the estimates,

   (b) Balancing the covariates between treatment and control groups. This will be assessed by examining the minimum p-value of the Kolmogorov-Smirnov test for equality of treatment and control groups distributions for each covariate, respectively, before and after matching. Large p-values are consistent with greater comparability of the treatment and control groups in terms of the covariates, and hence reflect better covariate balance among treatment and control groups.

3.2 Summary Results of the Simulation Study

By examining the median, the inter-quartile distance, and the overall range of the box plots of the estimated treatment effect, we make the following conclusions:

1. The simple observed treatment difference can be a very poor estimate when the covariate distributions are different and deviate from standard normal distributions as shown in panels 2 to 4 of Figures 1 and 2.

2. For the main effect model, the LS fit (when the model is correctly specified or even over-fitted with interaction terms) is generally better than other methods in estimating the treatment effect. But the LS fit with main effect only can perform poorly if the true model includes interactions; however, the LS fit with interactions (correct model) outperforms other methods.

3. Matching purely based on propensity scores usually performs worse than Genetic matching either with all available data or with the trimmed dataset in estimating the true treatment effect. The trimmed estimate using Genetic matching to match both covariates and propensity scores performs almost uniformly better than any other method regardless of model specification, except for the LS fit when the model is correctly specified as discussed in (b) above.

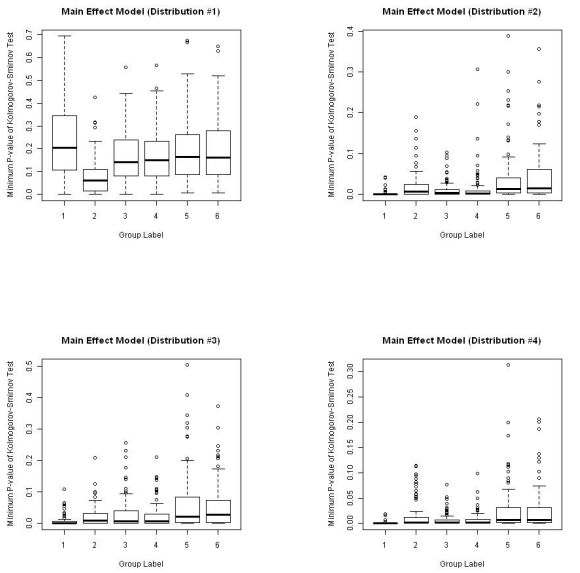

Figure 3: Minimum p-value of K-S test of equality: Main effect. (Labels 2-6 = d-h of item 4 in Sec. 3.1)

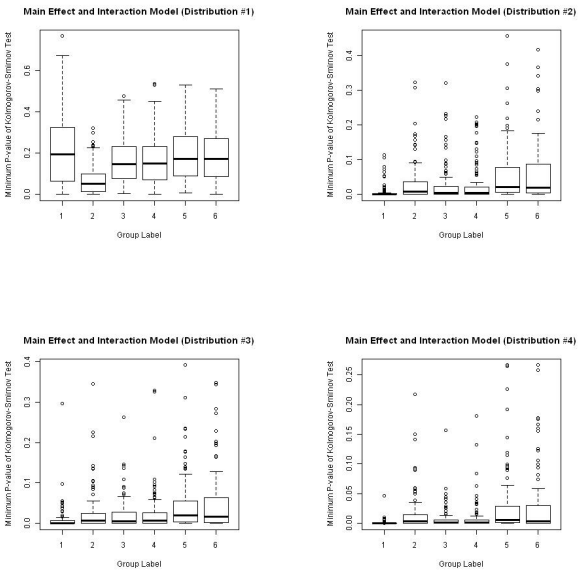

Figure 4: Minimum p-value of K-S test of equality: Interactions. (Labels 2-6 = d-h of item 4 in Sec. 3.1)

4. When the covariates of treatment and control groups have identical normal distributions, the LS method outperforms all other methods since there is no need for matching. Any effort to match is redundant. The propensity score matching seems to make the covariate matching worse more often than not. However, the Genetic matching seems to perform reasonably well, especially when the outliers were trimmed away (Panel 1 of Figures 3 and 4).

5. However, when the covariate distributions are different between treatment and control groups and deviate from the standard normal, the effect of matching from all methods becomes very visible. This can be seen in Panels 2-4 of Figures 3 and 4. Genetic matching with trimmed outliers tends to outperform all other methods either matched only on all covariates or with propensity score included. This is true for all distributions tested here.

As discussed above, when the model is correctly specified, the simple LS method outperforms other methods as expected. However, generally when analyzing data, one rarely knows the correct model or the distribution from which the data was generated. Therefore, the performance of LS method can be expected to diminish in the analysis of real data. On the other hand, the performance of Genetic Matching seems to be almost always comparable to the LS method when the model is correctly specified, and performs much better when the model is mis-specified as shown in Panels 1, 2, and 4 of Figure 2. Therefore, the Genetic Matching seems to serve as a "model mis-specification proof" tool for general data analysis.

It is interesting to note that Diamond, et al. [2] concluded that Genetic Matching is preferred over other matching methods because it is more efficient (smaller MSE) and is less biased.

4 Example

A phase III, multi-national randomized, double blind, placebo controlled clinical trial was conducted to compare the treatment effect of drug A and drug B to placebo in controlling disease activity in subjects with rheumatoid arthritis having an inadequate clinical response to methotrexate. The study was not originally designed to compare drug A and drug B directly. However, a post hoc analysis to compare these two drugs in a subgroup of countries of the original study is of clinical interest and also to meet the regulatory request. A total of 156 and 165 patients were randomized to drugs A and B in these countries, respectively. The primary endpoint of the study was the disease activity score based on 28 joints (DAS28).

Comparisons of several baseline covariates using the t-test did not show particular imbalance between the two treatment groups. However, a more in-depth investigation of the baseline distributions by quantile-quantile plots showed some deviations between the two popu-

lations. The objective in this analysis is to properly estimate the treatment difference under the situation of baseline imbalance.

The first step in this analysis is to match the patients from drugs A and B. Both the propensity score and the Genetic matching methods were used so that we can compare the relative performance of these two matching methods.

Several covariates were examined to compare the performance of propensity score and Genetic matching. The baseline pain scores between the treatment groups are compared and shown in Figure 5. The original Q-Q plot of pain scores between drug A and drug B is shown in Panel 1. The Q-Q plots of this covariate using propensity score matching and Genetic matching are shown in Panels 2 and 3, respectively. One can clearly see substantial improvement in covariate balance of Genetic matching over the propensity score matching.

Empirical permutation distributions of the treatment effect before and after Genetic matching were generated to determine the level of significance of the observed treatment effect among the randomly permutated samples. The observed treatment difference prior to matching is about -0.19. However, the magnitude of the treatment difference was reduced to -0.048 after matching. The treatment effect estimated after matching indicates the treatment difference is not as big as the original estimate. In other words, without this matching step, the treatment difference may have potentially been overestimated and the medical practice may be misguided. Even though the permutation test did not show a significant treatment difference in either pre or post matching; however, the treatment effect distributions from permutations seem to have some subtle difference and the test prior to matching showed a higher significance level than post matching. The 95% confidence interval of the treatment effect difference was also estimated using the stochastic approximation proposed by Robbins and Munro [8] and implemented by Garthwaite [4]. A total of 5000 randomized samples were generated and analyzed. The estimates fluctuate substantially in the beginning of the approximation process. The process began to stabilize after about 2500 randomizations. Figure 6 shows the stochastic approximation for the upper and lower limits of the confidence interval. The resulting 95% confidence interval is (-0.110, 0.4858).

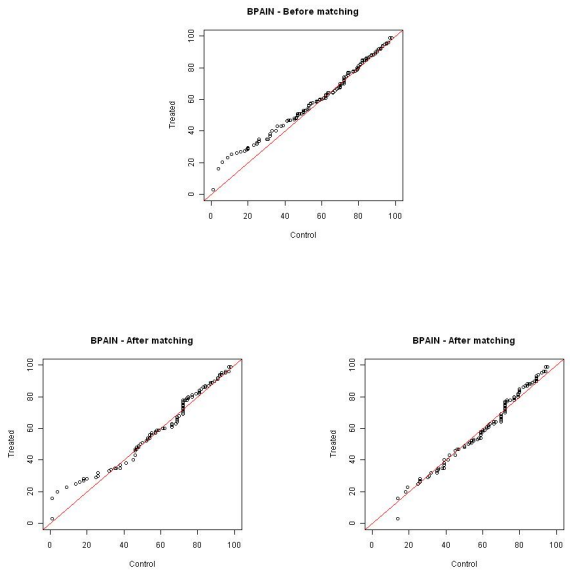

Figure 5: Comparison of covariate adjustment before and after propensity score and genetic matching, respectively.

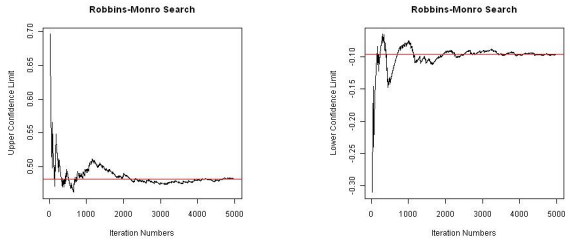

Figure 6: Stochastic approximation of the 95% confidence interval of treatment effect difference (based on 5000 simulated randomization)

5 Discussion

Statistical modeling and data analysis are important steps in advancing innovative scientific research in the fields of medicine, economics, social sciences, etc. To translate data into useful unbiased information is a critical endeavor for scientists and researchers as well. When data do not come from well-designed experiments, statistical modeling and data analysis to extract unbiased information can become much more challenging.

In this paper, we described various matching techniques to make the subjects under consideration more comparable before statistical inference; we also conducted a simulation study to further investigate the performance of these methods under different scenarios in their relative ability to better balance the covariates between the subjects groups, and in obtaining the unbiased estimate of treatment effect. The methods we compared ranged from the usual linear regression, conventional matching techniques with all available data to more robust alter-

natives, which flexibly weights the outliers. Generally, Genetic matching is preferred to other methods under various covariate distributions in balancing the covariates and obtaining the true treatment effect.

Given its longer history, the propensity score matching has been the most well known and most commonly used method in casual-effect relationship research; however, the selection of variables to be incorporated into the logistic regression model to derive the propensity score is not a trivial matter.

Several authors have proposed various approaches to incorporate covariates to estimate the propensity score (e.g., Rubin & Thomas [11], Rubin [14], Brookhart et. al. [1]). The general findings are to incorporate covariates which are thought to be related to outcomes and are confounded with both treatment assignment and outcomes. The model which incorporates as many covariates as possible or the model which includes obvious covariates such as age, gender, and race do not seem to perform as well as one would expect. On the other hand, the Genetic matching method has the additional flexibility to allow the covariates to be assigned unequal weight and also takes into account the covariance of the variables incorporated into the distance calculation which can eliminate some modeling difficulties caused by co-linearity between covariates.

Estimation and comparison of treatment effects should only be conducted after careful examination of balance between the groups being compared. It is important to note that the research findings should be regarded as exploratory and be interpreted with care within the context of biological or scientific plausibility and relevance.

References

[1] Brookhart, M.A. et. al. Variable selection for propensity score models. American Journal of Epidemiology 2006; 163: 1149-1156.

[2] Diamond A, Sekhon, JS. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Working paper (2005).

[3] Diamond, A. and Sekhon, J. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. University of California, Berkeley, 2010.

[4] Gartwaite, P.H. Confidence intervals from randomization tests. Biometrics 1996; 52: 1387-1393.

[5] Hopkins, D. Politicized Places: Explaining Where and When Immigrants Provoke Local Opposition. American Political Science Review 2010, 104 (1): 4060.

[6] Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. Graphical Methods for Assessing Logistic Regression Models. Journal of the American Statistical Association 1984; 79: 61-71.

[7] Peto, R., Collins, R., and Gray, R. Large-scale randomized evidence: large, simple trials and overviews of trials. Journal of Clinical Epidemiology 1995; 48: 23-40.

[8] Robbins, H. and Munro, S. A stochastic approximation method. Annals of Mathematical Statistics 1951; 22 : 400-407.

[9] Rosenbaum, P.R. Observational Studies. New York: Springer-Verlag 1995.

[10] Rosenbaum, P.R. and Rubin, D.B. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 79 : 516-524.

[11] Rubin, D.B. and Thomas, N. Matching using estimated propensity score: relating theory to practice. Biometrics 1996; 52 : 249-264.

[12] Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 1974; 66 : 688-701.

[13] Rubin, D.B. Assignment to a treatment group on the basis of a covariate. Journal of Educational Statistics 1977; 2 : 1-26.

[14] Rubin, D.B. Estimating causal effects from large data sets using the propensity score. Annals of Internal Medicine 1997; 127 : 757-763.

[15] Sekhon, J.S. Alternative Balance Metrics for Bias Reduction in Matching Methods for Causal Inference. Working Paper. http://sekhon.berkeley.edu/papers /SekhonBalanceMetrics.pdf 2006.

[16] Tsai, K.T. Assessing Regression Modeling with Ordinal Responses. Presentation at the Joint Statistical Meetings of the American Statistical Association 2008.

Contact author: Kao-Tai Tsai
Email address: tsai0123@yahoo.com