Overview of SAS Matching Programs Developed by the Ohio Valley Node of the National Institute on Drug Abuse (NIDA) National Drug Abuse Treatment Clinical Trials Network (CTN)

Daniel Lewis

lewisdw@ucmail.uc.edu

March 4, 2019

Table of Contents

Statement of the Problem

Often when analyzing large observational data sets such as electronic health records, one wants to study the impact on one or more outcomes of a binary variable which divides a large data sample into a relatively small base cohort and a much larger control cohort. One such binary variable might be whether or not a person received a given treatment; another might be whether or not a person was diagnosed with a given condition.

Unlike clinical trials where participants are randomly assigned to cohorts, in observational data such as medical records, people rarely end up in their respective cohorts arbitrarily. There are usually factors (observable and unobservable) which determine, at least in part, into which cohort a given person ends up. If a statistically significant effect is found for the binary variable which defines the cohorts, it is impossible to determine whether that effect is due to the binary variable, itself, or due to one or more confounding factors.

One way to at least partially address this problem is to find the subset of people in the control cohort that mirror the people in the base cohort relative to an observed set of background characteristics expressed as numeric variables. When this process is successful, the analyst has cohorts that are reasonably mutually consistent, at least on known factors. This increases the likelihood that a significant by-cohort effect is directly due to the cohort partitioning, itself, and not due to some other confounding factor.

Computer programs that perform this mirroring process are referred to here as “matching” programs. A matching program, as defined in this document, matches each member in the base cohort to a distinct member in the control cohort that is “similar” to it where the “similarity” of two subjects is indicated by a single numeric distance measure derived by comparing the background characteristics for the two subjects. The smaller the value of the distance measure for a given pair of subjects, the more similar those subjects are to each other relative to the background characteristics on which the distance measure is based.

Brief Description of the Current Ohio Valley Node-Developed Matching Programs

Introduction

Recently, the Ohio Valley Node (OVN) of the NIDA Clinical Trials Network (based at the University of Cincinnati) developed matching programs in SAS for use in its analyses of electronic health data. These programs were developed with the following objectives:

- The matching programs would use documented techniques that were as effective as possible over as wide a range of applications as possible with as little controversy as possible.
- Researchers with no more than basic knowledge of SAS, the matching process, propensity scores, and calipers would be able to understand the programs well enough to use them.
- The matching programs would be efficient with respect to time and computer resources.

The OVN-developed programs being discussed also have the advantage of being open-source and free of charge.

Programs Available

The current OVN-developed data-matching package consists of four alternative SAS macro functions:

%Mahalanobis_optimal which performs optimal matching using the Mahalanobis distance as the distance measure

%Propensity_score_optimal which performs optimal matching using the difference between logistic propensity scores as the distance measure:

%Mahalanobis_greedy which performs greedy matching using the Mahalanobis distance as the distance measure

%Propensity_score_greedy which performs greedy matching using the distance between logistic propensity scores as the distance measure

Brief Description

Matching Processes (Optimal vs. Greedy)

Each of these macro functions performs a 1:1 match matching each member in the base cohort with a distinct member in the control cohort. For each matched pair, there is a number indicating how similar the base-cohort member is to its selected match based on a given set of numeric background characteristics. This number is called the "distance" for that pair. A small distance indicates that the pair are very similar while a

large distance indicated the pair are very dissimilar. The next section discusses distance measures in more detail.

The macro functions discussed in this document allow for two different methods for matching base-cohort subjects to control-cohort subjects; one method is called "optimal" matching, and one is called "greedy" matching. Optimal matching matches the base people to respective control people so as to minimize the sum of the distances over all pairs. "Greedy" matching is a faster and less computer-intensive approximation to optimal matching. Greedy matching can be used when optimal matching is not practical or not considered reasonable [Ref. 5 page 123-3] [Ref. 6].

When one of the "greedy" macro functions is selected, the program repeats the matching process for a user-selected number of times where each iteration uses a distinct random ordering of the base cohort and, therefore, produce a different set of matched pairs. The program selects a single, final set of matched pairs from these multiple sets; this selection is based first on which set of matched pairs comes closest to matching everybody in the base cohort (leaving no base member unmatched), and second on which set of matched pairs has the smallest sum-of-the-distances.

The Matching Process Distance Measures (Mahalanobis vs. Logistic Propensity Scores)

The optimal and greedy matching processes both select a set of base-cohort-to-control-cohort matched pairs based on a numeric "distance measure". As was stated in the previous section, a distance measure is defined as a number that indicates how similar a pair of subjects are with respect to a numeric set of background characteristics. The SAS macro functions being presented here allow the user to choose between two popular distance measures: the Mahalanobis distance and the within-pair logistic propensity score differences. Though a definition of these distance measures would require some advanced mathematics and is beyond the scope and intent of this document, the Mahalanobis distance can generally be considered a more refined distance measure which usually works well with fewer than 20 background characteristic variables. For more specific recommendations, refer to the references for this document [Ref. 1], [Ref. 5 pages 123-4], [Ref. 3 page 171], [Ref. 4]

The Preselection Process

Regardless of whether the matching process is "optimal" or "greedy", and regardless of the distance measure used within the matching process, the matching process, itself is always preceded by a preselection of potential matches based on the respective differences in the logistic propensity scores within each potential match. This preselection can occur in one of two ways: the user can choose to use an elimination caliper which eliminates potential matches with propensity score differences beyond caliper, or the user can choose both a punishing caliper (which forces potential matches with larger propensity score differences to be used only as necessary [Ref.3 page 174]) and a wider elimination caliper (which assures that all final matches will have propensity score differences within reasonable boundaries).

Program Options (SAS Macro Function Passing Parameters)

Each of the presented SAS macro functions has five critical input parameters:

- The name of the input data set
- The name of the output data set containing the base cohort
- The name of the output data set containing the selected subset of the control cohort
- A list of numeric background characteristic variables to be used in calculating the distance measure
- The name of the binary (0 / 1) variable which indicates the cohorts (1=base, 0= control)

The macro functions also have the following optional input parameters:

- All four programs allow the user to select the size of the primary caliper expressed as a fraction of the logistic propensity score standard deviation over both cohorts (If not included, this value is set to 0.2 giving a caliper that is 20% of the propensity score standard deviation over both cohorts.)

- All four programs allow the user to select the size of the wider secondary caliper also expressed as a fraction of the logistic propensity score standard deviation (If not included, there will be no secondary caliper. If the secondary caliper is not greater than the primary caliper, it is ignored.)

  Note: If only one caliper is used it is an elimination caliper. If two calipers are used the primary caliper is a punishing caliper, and the larger secondary caliper is an elimination caliper.

- The programs using the Mahalanobis measure in their matching process have an optional parameter that lists background characteristics for calculating the logistic propensity scores used in preselection. (If not included, the list of background characteristics for the distance measure will be used.)

- The programs using the Mahalanobis measure in their matching process have an optional parameter which allows the user to substitute ranks for the actual values of the background characteristic variables. (If not included, no substitution occurs)

  This substitution can be useful in reducing the impact of outliers [Ref. 3 page 171] A more effective and appropriate option might be to identify and deal with potential outliers before matching.

- The programs using a greedy matching process have an optional parameter indicating the number of sets of matches from which to select the best set of matches. (If not included, this parameter is set to a default = 1000.)

  Obviously, the more replicate greedy matches performed, the closer the final selection will approximate an optimal match.

Discussion

Limits of the Matching Process

Any matching process assumes that there is a subset of subjects in the control cohort that are similar to the subjects in the base cohort. Obviously, if this not true, the matching process will be of limited value in balancing the cohorts.

Also, the effectiveness of the matching process is limited by the size of the control cohort with respect to the base cohort. In the extreme case, if the control cohort is the same size as the base cohort, then all of the original members of the control cohort will be selected by the matching process, which renders the process of dubious value. In the experience of the author, the matching programs presented here work best when the control cohort is at least three times as big as the base cohort.

Avoiding Incomplete Matches

A complete match is where every subject in the base cohort is matched to a distinct subject in the control cohort. An incomplete match is where one or more subjects in the base cohort are not matched. The programs presented here will produce incomplete matches if the elimination caliper in the preselection is not wide enough. (See the discussion below on calipers.) Incomplete matches should be avoided for statistical reasons [Ref. 3 page 85], [Ref. 7].

Choosing Optimal Matching vs. Greedy Matching

An "Optimal matching" algorithm is, by definition, a matching algorithm that minimizes the sum of the within-pair distances over all the selected matched pairs. Optimal matching is the gold standard for matching algorithms, and as such, is recommended by at least some sources whenever possible [Ref. 5 page 123-2] [Ref. 6].

However, because it is computationally intensive, optimal matching can overwhelm computer resources when dealing with extremely large data sets, in dealing with data sets where the ratio of control people to base people is not sufficiently large, and when matching distances are based on a large number of background characteristics. In these situations, one might want to consider the following suggestions before abandoning optimal matching:

- Using small calipers in the preselection process can (along with producing better matches) substantially reduce pressure on computer resources by vastly reducing the number of potential matched pairs being considered by the matching process, itself.

- The default SAS system settings for memory usage can be adjusted to better use available resources. (Contact SAS support.)

- If the analyst is planning to routinely analyze large data sets, an upgrade in computing resources may be reasonable.

If after considering these suggestions, optimal matching is not reasonable, greedy matching is statistically valid, and at least one simulation study found greedy matching to be as effective as optimal matching at producing statistically balanced cohorts suitable for exploring by-cohort effects [Ref 2 page 10] [Ref. 5 page 123-2]. (Note that simulations are artificial and may not reflect real-world situations.) As was stated earlier, the programs presented here improve on the greedy matching process by allowing the user to select the best set of matches from replicate sets.

Choosing Mahalanobis Distances vs. Logistic Propensity Score Differences

The Mahalanobis distance and logistic propensity score difference are both commonly used distance measures in matching processes though neither one is totally free of controversy [(Ref. 3 page 171)] [Ref. 4]. The Mahalanobis distance is generally be considered a more refined distance measure and usually works well with fewer than 20 background characteristic variables. For specific recommendations, refer to the references for this document [Ref. 5 pages 123-4], [Ref. 3 page 171] [Ref. 4].

Below are some issues related to the Mahalanobis measure that were addressed in the programs being presented:

- The Mahalanobis distance does not always treat variables with dramatically different variances in a balanced manner. Paul Rosenbaum pointed out that this could be a particular problem for Mahalanobis measure involving binary variables. (Ref. 3 page 171)

  The programs being presented address this problem by standardizing each variable before calculations. (In this context, "standardizing" means subtracting the variable's mean from each value and then dividing by standard deviation.) (Ref. 1) One of the results of standardizing is that all standardized variables have the same variance.

- The Mahalanobis measure involves matrix calculations which can fail due to computer round-off error.

  The programs being presented address this problem in two ways. First, the variables involved are transformed into their principal components before performing the Mahalanobis calculations. (Roughly speaking, this means transforming variables into their inherent, uncorrelated sources of variability.) (Ref. 1) This transformation greatly simplifies the Mahalanobis calculations and reduces the potential for matrix algebra failure.

  Second, the program continually checks for and sets to zero values that are close enough to zero as to be almost certainly due to round-off error. These adjustments have an insignificant effect on final calculations while they almost totally eliminate the potential for matrix algebra failure due to computer round-off error.

References and Quotes

1. Feng, W. W., Yu Jun, M. S. "A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study." (WW Feng, Y Jun, R Xu - SAS PharmaSUG 2006 Conference, 2006) https://www.lexjansen.com/pharmasug/2006/PublicHealthResearch/PR05.pdf

   This SUGI paper explains and presents code for a matching program that uses Mahalanobis distances and propensity score calipers.

2. Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." *Statistical science: a review journal of the Institute of Mathematical Statistics* 25.1 (2010): 1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2943670/

   (page 2):
   " Alternatives to matching methods include adjusting for background variables in a regression model, instrumental variables, structural equation modeling, or selection models. Matching methods have a few key advantages over those other approaches. First, matching methods should not be seen in conflict with regression adjustment and in fact the two methods are complementary and best used in combination. Second, matching methods highlight areas of the covariate distribution where there is not sufficient overlap between the treatment and control groups, such that the resulting treatment effect estimates would rely heavily on extrapolation. Selection models and regression models have been shown to perform poorly in situations where there is insufficient overlap, but their standard diagnostics do not involve checking this overlap (Dehejia and Wahba, 1999, 2002; Glazerman *et al.*, 2003). Matching methods in part serve to make researchers aware of the quality of resulting inferences. Third, matching methods have straightforward diagnostics by which their performance can be assessed."

   (page 3):
   For efficient causal inference and good estimation of the unobserved potential outcomes, we would like to compare treated and control groups that are as similar as possible.

   (page 10):
   "**3.1.1 Optimal Matching—**One complication of simple ("greedy") nearest neighbor matching is that the order in which the treated subjects are matched may change the quality of the matches. Optimal matching avoids this issue by taking into account the overall set of matches when choosing individual matches, minimizing a global distance measure (Rosenbaum, 2002). Generally, greedy matching performs poorly when there is intense competition for controls, and performs well when there is little competition (Gu and Rosenbaum, 1993). Gu and Rosenbaum (1993) find that optimal matching does not in general perform any better than greedy matching in terms of creating groups with good balance, but does do better at reducing the distance within pairs (p. 413): "…optimal matching picks about the same controls [as greedy matching] but does a better job of assigning them to treated units." Thus, if the goal is simply to find well-matched groups, greedy matching may be sufficient. However, if the goal is well-matched pairs, then optimal matching may be preferable."

3. Rosenbaum, Paul R. "Basic Tools of Multivariate Matching." *Design of observational studies*. Springer New York, 2010. 163-185.

   (page 85):
   "**3.6 Bias Due to Incomplete Matching**
   Chapter 3 has followed Chapter 2 in focusing on internal validity, as discussed in 2.6. That is, the focus has been on inference about treatment effects for the 2I matched individuals, not on whether the same treatment effects would be found in the population of L individuals. When treatment effects vary from person to person, as seemed to be the case in § 2.5, changing the individuals under study may change the magnitude of the effect. Although not required for internal validity, it is common in practice to match all of the treated subjects in the population, so that the number of matched pairs, I, equals the number of treated subjects $\Sigma Z_$ in the available population of L individuals. The goal here is an aspect of external validity, specifically the ability to speak about treatment effects in the original population of L individuals. If the population of L individuals were itself a random sample from an infinite population, and if all treated subjects were matched, then under the na¨ıve model (3.5)−(3.8), the average treated-minus-control difference in observed responses, $(1/I)\ \Sigma$ Yi, would be unbiased for the

expected treatment effect on people who typically receive the treatment, namely E (rT −rC | Z = 1); however, this is typically untrue if some treated subjects are deleted, creating what is known as the ‘bias due to incomplete matching’ [72]. “

(page 171):
“The Mahalanobis distance was originally developed for use with multivariate Normal data, and for data of that type it works fine. With data that are not Normal, the Mahalanobis distance can exhibit some rather odd behavior. If one covariate contains extreme outliers or has a long-tailed distribution, its standard deviation will be inflated, and the Mahalanobis distance will tend to ignore that covariate in matching. With binary indicators, the variance is largest for events that occur about half the time, and it is smallest for events with probabilities near zero and one. In consequence, the Mahalanobis distance gives greater weight to binary variables with probabilities near zero or one than to binary variables with probabilities closer to one half. In the welder data, two individuals with the same age and race but different smoking behavior have a Mahalanobis distance of 4.407, whereas two individuals with the same age and smoking behavior but different race have a Mahalanobis distance of 8.127, so a mismatch on race is counted as almost twice as bad as a mismatch on smoking. Two people who differed by 20 years in age with the same race and smoking would have a Mahalanobis distance of 0.021 × 202 =8.4, so a difference in race counts about as much as a 20-year difference in age. In a different context, if there were binary indicators for the states of the United States, then the Mahalanobis distance would regard matching for Wyoming as vastly more important than matching for California, simply because fewer people live in Wyoming. In many contexts, rare binary covariates are not of overriding importance, and outliers do not make a covariate unimportant, so the Mahalanobis distance may not be appropriate with covariates of this kind.

A simple alternative to the Mahalanobis distance (i) replaces each of the covariates, one at a time, by its ranks, with average ranks for ties, (ii) premultiplies and postmultiplies the covariance matrix of the ranks by a diagonal matrix whose diagonal elements are the ratios of the standard deviation of untied ranks, 1, . . . ,L, to the standard deviations of the tied ranks of the covariates, and (iii) computes the Mahalanobis distance using the ranks and this adjusted covariance matrix. Call this the ‘ rank-based Mahalanobis distance.’ Step (i) limits the influence of outliers. After step (ii) is complete, the adjusted covariance matrix has a constant diagonal. Step (ii) prevents heavily tied covariates, such as rare binary variables, from having increased influence due to reduced variance. “

(page 174):
Although it is not obvious from Tables 8.3, 8.5, and 8.6, because of competition for controls, there is no pair matching in which the caliper on the propensity score, |_e (xk)−_e (x_)| ≤ w with w = 0.086, is respected for all 21 matched pairs. This is true because of competition among welders for the same controls, despite the fact that each welder has a propensity score within w = 0.086 of at least one potential control. For this reason, the infinities in Tables 8.5 and 8.6 are not used; they are replaced by the addition of a ‘penalty function,’ which exacts a large but finite penalty for violations of the constraint; see, for instance, [2, pages 372−373] or [21, Chapter 6]. The penalty used here is 1000 × max(0, |_e (xk)−_e (x_)|−w), so if |_e (xk)−_e (x_)| ≤ w then the penalty is zero, but if |_e (xk)−_e (x_)| > w then the penalty is 1000 × (|_e (xk)−_e (x_)|−w). For instance, the penalty for matching welder #1 to potential control #1 is 1000 × (|.4587−0.1437|−0.0860) = 229. This penalty is added to the Mahalanobis distance or rank-based Mahalanobis distance for the corresponding pair. Optimal matching will try to avoid the penalties by respecting the caliper, but when that is not possible, it will prefer to match so the caliper is only slightly violated for a few matched pairs.

4. King, Gary, and Richard Nielsen. "Why propensity scores should not be used for matching." *Copy at http://j.mp/1sexgVw* 378 (2016).

(Abstract):
“We show that propensity score matching (PSM), an enormously popular method of preprocessing data for causal inference, often accomplishes the opposite of its intended goal—increasing imbalance, inefficiency, model dependence, and bias. . . . Although these results suggest that researchers replace PSM with one of the other available methods when performing matching, propensity scores have many other productive uses.”

5. Hintze J. Data matching – Optimal and Greedy. In *NCSS User's Guide*. Kaysville, UT, NCSS, 2007, Chapter 123 https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Data_Matching-Optimal_and_Greedy.pdf

(page 123-2):

"**Optimal vs. Greedy Matching** Two separate procedures are documented in this chapter, *Optimal Data Matching* and *Greedy Data Matching*. The goal of both algorithms is to produce a matched sample that balances the distribution of observed covariates between the treatment and matched-control groups. Both algorithms allow for the creation of 1:1 or 1:*k* matched pairings. Gu and Rosenbaum (1993) compared the greedy and optimal algorithms and found that "optimal matching is sometimes noticeably better than greedy matching in the sense of producing closely matched pairs, sometimes only marginally better, but it is no better than greedy matching in the sense of producing balanced matched samples." The choice of the algorithm depends on the research objectives, the desired analysis, and cost considerations. We recommend using the optimal matching algorithm where possible."

(page 123-4):

"**Which Distance Measure to Use?** The best distance measure depends on the number of covariate variables, the variability within the covariate variables, and possibly other factors. Gu and Rosenbaum (1993) compared the imbalance of Mahalanobis distance metrics versus the propensity score difference in optimal 1:1 matching for numbers of covariates (*P*) between 2and 20 and control/treatment subject ratios between 2 and 6. Mahalanobis distance within propensity score calipers was always best or second best. When there are many covariates (*P* = 20), the article suggests that matching on the propensity score difference is best. The use of Mahalanobis distance (with or without calipers) is best when there are few covariates on which to match (*P* = 2). In all cases considered by Gu and Rosenbaum (1993), the Mahalanobis distance within propensity score calipers was never the worst method of the three. Rosenbaum and Rubin (1985a) conducted a study of the performance of three different matching methods (Mahalanobis distance, Mahalanobis distance within propensity score calipers, and propensity score difference) in a greedy algorithm with matches allowed outside calipers and concluded that the Mahalanobis distance within propensity score calipers is the best technique among the three. Finally, Rosenbaum (1989) reports parenthetically that he has had "unpleasant experiences using standard deviations to scale covariates in multivariate matching, and [he] is inclined to think that either ranks or some more resistant measure of spread should routinely be used instead.

Based on these results and suggestions, we recommend using the Mahalanobis Distance within Propensity Score Calipers as the distance calculation method where possible. The caliper radius to use is based on the amount of bias that you want removed"

6. Rosenbaum, Paul R. "Optimal matching for observational studies." Journal of the American Statistical Association 84.408 (1989): 1024-1032.

"**1.4 Comparing Greedy and Optimal Matching**

Why prefer the optimal match? There are several rea- sons. First, there is the obvious point that the optimal match is always as good as and often better than the greedy match. In the example, the loss due to greedy was 11%; not a disaster, but worth avoiding. The second point, though distinct, is closely related to the first. Although a greedy algorithm (like forward step-wise regression) may provide a tolerable answer, it rarely comes with a guarantee that the answer is in fact tolerable. In particular, greedy matching can be arbitrarily poor compared to optimal matching"

7. Rosenbaum, Paul R., and Donald B. Rubin. "The Bias Due to Incomplete Matching." *Biometrics*, vol. 41, no. 1, 1985, pp. 103–116., www.jstor.org/stable/2530647.

8. Rubin, Donald B. "The use of matched sampling and regression adjustment to remove bias in observational studies." *Biometrics* (1973): 185-203.

"**SUMMARY**

The ability of matched sampling and linear regression adjustment to reduce the bias of an estimate of the treatment effect in two sample observational studies is investigated for a simple matching method and five simple estimates. Monte Carlo results are given for moderately linear exponential response surfaces and analytic results are presented for quadratic response surfaces. The conclusions are (1) in general both matched sampling and regression adjustment can be expected to reduce bias, (2) in some cases when the variance of the matching variable differs in the two populations both matching and regression

adjustment can increase bias, (3) when the variance of the matching variable is the same in the two populations and the distributions of the matching variable are symmetric the usual co- variance adjusted estimate based on random samples is almost unbiased, and (4) the combination of regression adjustment in matched samples generally produces the least biased estimate."