Chapter 123

Data Matching – Optimal and Greedy

Introduction

This procedure is used to create treatment-control matches based on propensity scores and/or observed covariate variables. Both optimal and greedy matching algorithms are available (as two separate procedures), along with several options that allow the user to customize each algorithm for their specific needs. The user is able to choose the number of controls to match with each treatment (e.g., 1:1 matching, 1:*k* matching, and variable (full) matching), the distance calculation method (e.g., Mahalanobis distance, propensity score difference, sum of rank differences, etc.), and whether or not to use calipers for matching. The user is also able to specify variables whose values must match exactly for both treatment and controls in order to assign a match. **NCSS** outputs a list of matches by match number along with several informative reports and optionally saves the match numbers directly to the database for further analysis.

Matching Overview

Observational Studies

In observational studies, investigators do not control the assignment of treatments to subjects. Consequently, a difference in covariates may exist between treatment and control groups, possibly resulting in undesired biases. Matching is often used to balance the distributions of observed (and possibly confounding) covariates. Furthermore, in many observational studies, there exist a relatively small number of treatment group subjects as compared to control group subjects, and it is often the case that the costs associated with obtaining outcome or response data is high for both groups. Matching is used in this scenario to reduce the number of control subjects included in the study. Common matching methods include Mahalanobis metric matching, propensity score matching, and average rank sum matching. Each of these will be discussed later in this chapter. For a thorough treatment of data matching for observational studies, the reader is referred to chapter 1.2 of D'Agostino, Jr. (2004).

The Propensity Score

Ideally, one would match each treatment subject with a control subject (or subjects) that was an exact match on each of the observed covariates. As the number of covariates increases or the ratio of the number of control subjects to treatment subjects decreases, it becomes less and less likely that an exact match will be found for each treatment subject. *Propensity scores* can be used in this situation to simultaneously control for the presence of several covariate factors. The propensity score was introduced by Rosenbaum and Rubin (1983). The propensity score for subject i (i = 1, …, N) is defined as the conditional probability of assignment to a treatment (Z_i = 1)

versus the control ($Z_i = 0$), given a set (or vector) of observed covariates, $\mathbf{x}_i$. Mathematically, the propensity score for subject *i* can be expressed as

$$e(\mathbf{x}_i) = \text{pr}(Z_i = 1 \,|\, \mathbf{X}_i = \mathbf{x}_i) .$$

It is assumed that the Z_i's are independent, given the X's. The observed covariates, $\mathbf{x}_i$, are not necessarily the same covariates used in the matching algorithm, $\mathbf{y}_i$, although they could be. Rosenbaum and Rubin (1985a) suggest using the logit of the estimated propensity score for matching because the distribution of transformed scores is often approximately normal. The logit of the propensity score is defined as

$$q(\mathbf{x}) = \log\left(\frac{1-e(\mathbf{x})}{e(\mathbf{x})}\right),$$

Matching on the observed propensity score (or logit propensity score) can balance the overall distribution of observed covariates between the treatment and control groups. The propensity score is often calculated using logistic regression or discriminant analysis with the treatment variable as the dependent (group) variable and the background covariates as the independent variables. Research suggests that care must be taken when creating the propensity score model (see Austin et al. (2007)). For more information about logistic regression or discriminant analysis, see the corresponding chapters in the **NCSS** manuals.

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.

The optimal and greedy algorithms differ in three fundamental ways:

1. Treatment of Previously Matched Subjects
2. Complete vs. Incomplete Matched-Pair Samples
3. Variable (Full) Matching

Treatment of Previously Matched Subjects

Optimal matching refers to the use of an optimization method based on the Relax-IV algorithm written by Dimitri P. Bertsekas (see Bertsekas (1991)), which minimizes the overall sum of pair-wise distances between treatment subjects and matched control subjects. The Relax-IV algorithm is based on network flow theory, and matching is just one of its many uses. Optimal matching is not a linear matching algorithm in the sense that as the algorithm proceeds, matches are created, broken, and rearranged in order to minimize the overall sum of match distances.

Greedy matching, on the other hand, is a linear matching algorithm: when a match between a treatment and control is created, the control subject is removed from any further consideration for matching. When the number of matches per treatment is greater than one (i.e., 1:*k* matching), the greedy algorithm finds the best match (if possible) for each treatment before returning and creating the second match, third match, etc. Once a treatment subject has been matched with the user-specified number of control subjects, the treatment subject is also removed from further consideration. A familiar example of a greedy algorithm is forward selection used in multiple regression model creation.

Complete vs. Incomplete Matched-Pair Samples

Optimal matching only allows for *complete matched-pair samples*, while greedy matching also allows for *incomplete matched-pair samples*. A complete matched-pair sample is a sample for which every treatment is matched with at least one control. An incomplete matched-pair sample is a sample for which the number of treatment subjects matched is less than the total number of treatment subjects in the reservoir. Rosenbaum and Rubin (1985b) present strong reasons for avoiding incomplete matched-pair samples.

Variable (Full) Matching

Variable (or "Full") matching is only available using the optimal matching algorithm. In variable matching, a different number of controls may be matched with each treatment. Each control is used only once, and each treatment receives at least one control. All eligible controls (e.g. all controls for which at least one treatment-control distance is non-infinite) are matched. Results from Gu and Rosenbaum (1993) suggest that in terms of bias reduction, full matching performs much better than 1:*k* matching. If we require that every treatment have the same number of controls, and the distributions between the two groups of covariates are not the same, then some treatments will be paired with controls that are not good matches. Variable matching, on the other hand, is more flexible in allowing control subjects to pair with the closest treatment subject in every case.

The gains in bias reduction for variable matching over 1:*k* matching, however, must be weighed against other considerations such as simplicity and aesthetics. The analysis after 1:*k* matching would arguably be simpler; a more complex analysis method (e.g. stratified analysis) would be employed after variable matching than would be after 1:*k* matching.

The Distance Calculation Method

Several different distance calculation methods are available in the matching procedures in **NCSS**. The different methods are really variations of three common distance measures:

1. Mahalanobis Distance
2. Propensity Score Difference
3. Sum of Rank Differences

The variations arise when using calipers for matching or when using forced match variables. A *caliper* is defined in this context a restricted subset of controls whose propensity score is within a specified amount (*c*) of the treatment subject's propensity score. A *forced match variable* contains values which must match exactly in the treatment and control for the subjects to be considered for matching. If the values for the forced match variables do not agree, then the distance between the two subjects is set equal to ∞ (infinity), and a match between the two is not allowed.

Distance Measures

The complete list of possible distance measures available in **NCSS** is as follows:

1. Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)
2. Mahalanobis Distance within Propensity Score Calipers (matches allowed outside calipers)
3. Mahalanobis Distance including the Propensity Score (if specified)
4. Propensity Score Difference within Propensity Score Calipers (no matches outside calipers)
5. Propensity Score Difference
6. Sum of Rank Differences within Propensity Score Calipers (no matches outside calipers)
7. Sum of Rank Differences within Propensity Score Calipers (matches allowed outside calipers)
8. Sum of Rank Differences including the Propensity Score (if specified)

Distance measures #2 and #7, where matches are allowed outside calipers in caliper matching, are only available with greedy matching. All others can be used with both the greedy and optimal matching algorithms.

For distance measures that involve propensity score calipers, the caliper size is determined by the user-specified radius, *c*. For any treatment subject, *i*, the j^{th}, control subject is included in the i^{th} treatment caliper if

$$|q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c$$

where $q(\mathbf{x}_i) = e(\mathbf{x}_i)$ is the propensity score based on the covariates $\mathbf{x}_i$. If the logit transformation is used in the analysis, then $q(\mathbf{x}) = \log((1-e(\mathbf{x}))/e(\mathbf{x}))$. The width of each caliper is equal to 2*c*.

Distance Calculation

Eight different distance measures are available in **NCSS**. Specify the method to be used in calculating distances between treatment and control subjects. If the distance method involves propensity score calipers, then a Propensity Score Variable must also be specified. For the formulas that follow, we will adopt the following notation:

1. The subscript *i* refers to the i^{th} treatment subject.
2. The subscript *j* refers to the j^{th} control subject.
3. $d(i, j)$ is the estimated distance between subjects *i* and *j*.
4. **x** is the vector of observed covariates used to estimate the propensity score.
5. $q(\mathbf{x}) = e(\mathbf{x})$ is the propensity score based on the covariates **x**. If the logit transformation is used in the analysis, then $q(\mathbf{x}) = \log((1-e(\mathbf{x}))/e(\mathbf{x}))$.
6. **y** is the vector of observed covariates used in the distance calculation. **y** is not necessary equivalent to **x**, although it could be.
7. $\mathbf{u} = (\mathbf{y}, q(\mathbf{x}))$ is the vector of observed covariates and the propensity score (or logit propensity score).
8. *C* is the sample covariance matrix of the matching variables (including the propensity score) from the full set of control subjects.
9. *c* is the caliper radius. The width of each caliper is 2*c*.
10. $FM_{i,l}$ and $FM_{j,l}$ are the values of the l^{th} forced match variable for subjects *i* and *j*, respectively. If no forced match variables are specified, then $FM_{i,l} = FM_{j,l}$ for all *l*.
11. $R_{i,p}$ and $R_{j,p}$ are the ranks of the p^{th} covariate values or propensity score for subjects *i* and *j*, respectively. Average ranks are used in the case of ties.

The options are:

- **Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} (\mathbf{u}_i - \mathbf{u}_j)^T C^{-1} (\mathbf{u}_i - \mathbf{u}_j) & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Mahalanobis Distance within Propensity Score Calipers (matches allowed outside calipers)**

$$d(i,j) = \begin{cases} (\mathbf{u}_i - \mathbf{u}_j)^T C^{-1} (\mathbf{u}_i - \mathbf{u}_j) & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| > c \text{ for all unmatched } j \\ & \text{and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

The absolute difference, $|q(\mathbf{x}_i) - q(\mathbf{x}_j)|$, is only used in assigning matches if there are no available controls for which $|q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c$.

- **Mahalanobis Distance including the Propensity Score (if specified)**

$$d(i,j) = \begin{cases} \left(\mathbf{u}_i - \mathbf{u}_j\right)^T C^{-1} \left(\mathbf{u}_i - \mathbf{u}_j\right) & \text{if } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Propensity Score Difference within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Propensity Score Difference**

$$d(i,j) = \begin{cases} |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Sum of Rank Differences within Propensity Score Calipers (no matches outside calipers)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

- **Sum of Rank Differences within Propensity Score Calipers (matches allowed outside calipers)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c \text{ and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ |q(\mathbf{x}_i) - q(\mathbf{x}_j)| & \text{if } |q(\mathbf{x}_i) - q(\mathbf{x}_j)| > c \text{ for all unmatched } j \\ & \quad \text{and } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

The absolute difference, $|q(\mathbf{x}_i) - q(\mathbf{x}_j)|$, is only used in assigning matches if there are no available controls for which $|q(\mathbf{x}_i) - q(\mathbf{x}_j)| \leq c$.

- **Sum of Rank Differences including the Propensity Score (if specified)**

$$d(i,j) = \begin{cases} \sum_p |R_{i,p} - R_{j,p}| & \text{if } FM_{i,l} = FM_{j,l} \text{ for all } l \\ \infty & \text{otherwise} \end{cases}$$

In the Greedy Data Matching procedure, two distance calculation methods are available that are not in the Optimal Data Matching procedure (option #2 and option #7). Both involve caliper matching with matches allowed outside calipers. When matches are allowed outside calipers, the algorithm always tries to find matches inside the calipers first, and only assigns matches outside calipers if a match was not found inside. Matches outside calipers are created based solely on the propensity score, i.e., if matches outside calipers are allowed and no available control subject exists that is within c propensity score units of a treatment subject, then the control subject with the nearest propensity score is matched with the treatment. This type of matching algorithm is described in Rosenbaum and Rubin (1985a).

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 2 and 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.

What Caliper Radius to Use?

The performance of distance metrics involving calipers depends to some extent on the caliper radius used. For instances in the literature where we found reports, comparisons, or studies based on caliper matching, Cochran and Rubin (1973) was nearly always mentioned as the literature used in determining the caliper radius (or "caliper width" as they call it) for the study. The following table (Table 2.3.1 from Cochran and Rubin (1973)) can be used to determine the appropriate coefficient and/or caliper radius to use:

Table 2.3.1 from Cochran and Rubin (1973). Percent Reduction in bias of *x* for caliper matching to within $\pm a\sqrt{(\sigma_1^2+\sigma_2^2)/2}$

a	$\sigma_1^2/\sigma_2^2 = 1/2$	$\sigma_1^2/\sigma_2^2 = 1$	$\sigma_1^2/\sigma_2^2 = 2$
0.2	0.99	0.99	0.98
0.4	0.96	0.95	0.93
0.6	0.91	0.89	0.86
0.8	0.86	0.82	0.77
1.0	0.79	0.74	0.69

The caliper radius to use depends on the desired bias reduction (table body), the coefficient *a*, and the ratio of the treatment group sample variance of $q(\mathbf{x})$, σ_1^2, to the control group sample variance of $q(\mathbf{x})$, σ_2^2. "Loose Matching" corresponds to $a \geq 1.0$, while "Tight Matching" corresponds to $a \leq 0.2$. The caliper radius is calculated as

$$c = a\sqrt{(\sigma_1^2+\sigma_2^2)/2} = a \times SIGMA$$

NCSS allows you to choose the caliper radius using the syntax "*a**SIGMA", where you specify the value for *a* (e.g., "0.2*SIGMA") or by entering the actual value directly for *c* (e.g., "0.5"). In the case of the former, the program calculates the variances of the treatment and control group propensity scores for you and determines the pooled standard deviation, sigma. You may want to run descriptive statistics on the treatment and control group propensity scores to determine the variance ratio of your data in order to find the appropriate value of *a* (from the table above) for your research objectives.

Data Structure

The propensity scores and covariate variables must each be entered in individual columns in the database. Only numeric values are allowed in propensity score and covariate variables. Blank cells or non-numeric (text) entries are treated as missing values. If the logit transformation is used, values in the propensity score variable that are not between zero and one are also treated as missing. A grouping variable containing two (and only two) unique groups must be present. A data label variable is optional. The following is a subset of the Propensity dataset, which illustrates the data format required for the greedy and optimal data matching procedures.

Propensity dataset (subset)

ID	Exposure	X1	…	Age	Race	Gender	Propensity
A	Exposed	50	…	45	Hispanic	Male	0.7418116515
B	Not Exposed	4	…	71	Hispanic	Male	0.01078557025
C	Not Exposed	81	…	70	Caucasian	Male	0.0008716385678
D	Exposed	31	…	33	Hispanic	Female	0.5861360724
E	Not Exposed	65	…	38	Black	Male	0.1174339761
F	Exposed	22	…	29	Black	Female	0.07538899371
G	Not Exposed	36	…	57	Black	Female	0.008287371892
H	Not Exposed	31	…	52	Caucasian	Male	0.4250166047
I	Not Exposed	46	…	39	Hispanic	Female	0.2630767334
J	Exposed	3	…	58	Hispanic	Male	0.4858799526
K	Not Exposed	84	…	24	Black	Female	0.1251753736

Example 1 – Optimal (1:1) Matching using the Mahalanobis Distance within Propensity Score Calipers

This tutorial describes how to create 1:1 treatment-control matches using the Mahalanobis Distance within Propensity Score Calipers distance metric. The data used in this example are contained in the PROPENSITY database. The propensity scores were created using logistic regression with Exposure as the dependent variable, X1 – Age as numeric independent variables, and Race and Gender as categorical independent variables. The propensity score represents the probability of being exposed given the observed covariate values. The optimal matching algorithm will always produce a complete matched-pair sample.

Setup

To run this example, complete the following steps:

1 Open the Propensity example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Propensity** and click **OK**.

2 Specify the Data Matching – Optimal and Greedy procedure options

- Find and open the **Data Matching – Optimal and Greedy** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Grouping Variable	**Exposure**
Treatment Group	**Exposed**
Propensity Score Variable	**Propensity**
Use Logit	**Checked**
Covariate Variable(s)	**X1-Age**
Data Label Variable	**ID**
Store Match Numbers In	**C11**
Caliper Radius	**1.5*Sigma**
Reports Tab	
Matching Detail Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

The following reports will be generated for both optimal and greedy matching with slight variations depending on the algorithm selected.

Data Summary Report

Data Summary Report

Rows Read	30
Rows with Missing Data	0
Treatment Rows	8
Control Rows	22

Data Variables

Grouping Variable	Exposure
- Treatment Group	"Exposed"
- Control Group	"Not Exposed"
Data Label Variable	ID

Variables Used in Distance Calculations

Propensity Score Variable	Logit(Propensity)
Covariate Variable 1	X1
Covariate Variable 2	X2
Covariate Variable 3	X3
Covariate Variable 4	X4
Covariate Variable 5	X5
Covariate Variable 6	X6
Covariate Variable 7	Age

Storage Variable

Match Number Storage Variable	C11

This report gives a summary of the data and variables used for matching.

Matching Summary Report

Matching Summary Report

Distance Calculation Method	Mahalanobis Distance within Propensity Score Calipers (no matches outside calipers)
Caliper Half-Width	2.63288
Order For Matching	Random (Computer-Generated Random Seed: 3275323)
Controls Matched per Treatment	1
Sum of Match Mahalanobis Distances	53.94887
Average Match Mahalanobis Distance	6.74361

Exposure	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Exposed	8	8	100.00%	0	0.00%
Not Exposed	22	8	36.36%	14	63.64%

This report gives a summary of the matches created, as well as a summary of the matching parameters used by the matching algorithm.

Distance Calculation Method

This is the method used to calculate distances between treatment and control subjects.

Caliper Radius

This is the caliper radius entered or calculated by the program. This line is only displayed if caliper matching based on propensity scores was used.

Order for Matching

This is the order used in matching as selected on the procedure window.

Controls Matched per Treatment

This is the target number of controls to match with each treatment. This value is specified on the procedure window.

Sum of Match Mahalanobis Distances (Sum of Match Propensity Score Differences or Sum of Match Rank Differences)

This is the sum of Mahalanobis distances, propensity score differences, or rank differences (depending on the distance calculation method selected) for all matched pairs.

Average Match Mahalanobis Distance (Average Match Propensity Score Difference or Average Match Rank Differences)

This is the average Mahalanobis distances, propensity score difference, or rank difference (depending on the distance calculation method selected) for all matched pairs. This is calculated as the [Sum of Match Distances (or Differences)]/[Number of Matches Formed].

Group (e.g. Exposure)

This specifies either the treatment or the control group. The title of this column is the Grouping Variable name (or label).

N

This is the number of candidates for matching in each group, i.e. the number of subjects with non-missing values for all matching variables in each group.

Matched (Unmatched)

This is the number of subjects that were matched (unmatched) from each group.

Percent Matched (Percent Unmatched)

This is the percent of subjects that were matched (unmatched) from each group.

Group Comparison Reports

Group Comparison Report for Variable = Logit(Propensity)

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	22	2.63066	2.06	-2.81410	-160.32%
After Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	8	0.82159	1.33	-1.00503	-73.88%

Group Comparison Report for Variable = X1

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	39.50000	20.96		
	Not Exposed	22	45.90909	26.11	-6.40909	-27.07%
After Matching	Exposed	8	39.50000	20.96		
	Not Exposed	8	26.50000	13.60	13.00000	73.58%

.
.
.

(output reports continue for each covariate variable specified)

This report provides summary statistics by group for the data in the propensity score variable and each covariate variable both before and after matching. Notice that the matching seemed to improve the balance of the propensity scores (Standardized Difference dropped from –160% to –73%) between the treatment and control groups, but worsened the balance for the covariate X1 (Standardized Difference increased from –27% to 73.58%).

Group Type

This specifies whether the summary statistics refer to groups before or after matching.

Group (e.g. Exposure)

This specifies either the treatment or the control group. The title of this column is the grouping variable name (or label).

N

This is the number of non-missing values in each variable by group. If there are missing values in covariates that were not used for matching, then these numbers may be different from the total number of subjects in each group.

Mean

This is the average value for each variable by group.

SD

This is the standard deviation for each variable by group.

Mean Difference

This is the difference between the mean of the treatment group and the mean of the control group.

Standardized Difference (%)

The standardized difference can be used to measure the balance between the treatment and control groups before and after matching. If a variable is balanced, then the standardized difference should be close to zero. The standardized difference is the mean difference as a percentage of the average standard deviation

$$\text{Standardized Difference (\%)} = \frac{100\left(\bar{x}_{t,p} - \bar{x}_{c,p}\right)}{\sqrt{\left(s_{t,p}^2 + s_{c,p}^2\right)/2}}$$

where $\bar{x}_{t,p}$ and $\bar{x}_{c,p}$ are the treatment and control group means for the p^{th} covariate variable, respectively, and $s_{t,p}^2$ and $s_{c,p}^2$ are the treatment and control group sample variances for the p^{th} covariate variable, respectively.

Matching Detail Report

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

		Treatment			Matched Control		
Match Number	Mahalanobis Distance	Row	Logit Propensity	ID	Row	Logit Propensity	ID
1	4.32807	1	-1.05541	A	8	0.30221	H
2	5.05385	4	-0.34801	D	22	-1.28232	V
3	9.07686	6	2.50671	F	16	3.28652	P
4	3.99318	10	0.05650	J	24	1.73357	X
5	13.85904	14	-1.11718	N	28	-0.07642	BB
6	9.25961	19	-1.31100	S	27	0.85319	AA
7	5.06011	26	1.16584	Z	29	0.72590	CC
8	3.31815	30	-1.36499	DD	9	1.03004	I

This report provides a list of all matches created and important information about each match.

Match

This is the match number assigned by the program to each match and stored to the database (if a storage variable was specified).

Mahalanobis Distance (Propensity Score |Difference| or Sum of Rank |Differences|)

This is the estimated distance between the treatment and matched control. The column title depends on the distance calculation method selected.

Row

This is the row of the treatment or control subject in the database.

Propensity Score (or first covariate variable)

This is the value of the propensity score (or logit propensity score if 'Use Logit' was selected). If no propensity score variable was used in distance calculations, then this is the value of first covariate variable specified. The title of this column is based on the propensity score variable name (or label) or the first covariate variable name (or label).

Data Label (e.g. ID)

This is the identification label of the row in the database. The title of this column is the data label variable name (or label).

Example 2 – Greedy (1:2) Matching using the Propensity Score Difference with Forced Match Variables

Continuing with Example 1, we will now use the greedy matching algorithm to create matches while using race and gender as forced match variables. This will force the algorithm to find control matches for treatments where the gender and race match exactly, i.e., a male can only be matched with a male, and a female can only be matched with a female, etc. Please note that the optimal matching algorithm can also be used with forced match variables, but we use the greedy matching algorithm here to display the incomplete matched-pair sample that results.

Setup

To run this example, complete the following steps:

1 Open the Propensity example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Propensity** and click **OK**.

2 Specify the Data Matching – Optimal and Greedy procedure options

- Find and open the **Data Matching – Optimal and Greedy** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Grouping Variable	**Exposure**
Treatment Group	**Exposed**
Propensity Score Variable	**Propensity**
Use Logit	**Checked**
Forced Match Variable(s)	**Race-Gender**
Covariate Variable(s)	**X1-Age**
Data Label Variable	**ID**
Store Match Numbers In	**C11**
Distance Calculation Method	**Propensity Score Difference**
Matches per Treatment	**2**
Reports Tab	
Matching Detail Report	**Checked**
Incomplete Matching Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Greedy Data Matching Output

Data Summary Report

Rows Read	30
Rows with Missing Data	0
Treatment Rows	8
Control Rows	22

Data Variables

Grouping Variable	Exposure
- Treatment Group	"Exposed"
- Control Group	"Not Exposed"
Data Label Variable	ID

Variables Used in Distance Calculations

Propensity Score Variable	Logit(Propensity)
Forced Match Variable 1	Race
Forced Match Variable 2	Gender

Storage Variable

Match Number Storage Variable	C11

Matching Summary Report

Distance Calculation Method	Propensity Score Difference
Order For Matching	Sorted by Distance
Controls Matched per Treatment	2
Sum of Match Propensity Score Differences	14.63954
Average Match Propensity Score Difference	1.46395

Exposure	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Exposed	8	6	75.00%	2	25.00%
Not Exposed	22	10	45.45%	12	54.55%

Group Comparison Report for Variable = Logit(Propensity)

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	-0.18344	1.39		
	Not Exposed	22	2.63066	2.06	-2.81410	-160.32%
After Matching	Exposed	6	0.11751	1.50		
	Not Exposed	10	1.48046	1.55	-1.36296	-89.21%

Group Comparison Report for Variable = X1

Group Type	Exposure	N	Mean	SD	Mean Difference	Standardized Difference (%)
Before Matching	Exposed	8	39.50000	20.96		
	Not Exposed	22	45.90909	26.11	-6.40909	-27.07%
After Matching	Exposed	6	33.66667	20.79		
	Not Exposed	10	44.90000	30.57	-11.23333	-42.97%

.
.
.

(output reports continue for each covariate variable specified)

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

Match Number	Logit Propensity |Difference|	Treatment Row	Treatment Logit Propensity	Treatment ID	Matched Control Row	Matched Control Logit Propensity	Matched Control ID
1	1.20120	4	-0.34801	D	27	0.85319	AA
1	1.37805	4	-0.34801	D	9	1.03004	I
2	0.15966	6	2.50671	F	15	2.34705	O
2	0.56240	6	2.50671	F	11	1.94431	K
3	0.66941	10	0.05650	J	29	0.72590	CC
3	4.46221	10	0.05650	J	2	4.51870	B
4	1.61321	19	-1.31100	S	8	0.30221	H
4	3.92267	19	-1.31100	S	12	2.61167	L
5	0.58806	26	1.16584	Z	23	1.75390	W
6	0.08266	30	-1.36499	DD	22	-1.28232	V

Notice that only the propensity score variable was used in distance calculations, but group comparison reports were generated for each covariate variable specified. In the Matching Detail Report, you can see that not all treatments were matched (incomplete matching). Finally, notice that race and gender were both used as Forced Match variables.

If you go back to the spreadsheet and sort the data on C11 (click on **Data** > **Sort** from the **NCSS Home** window), you will notice that matches were only created where the race and gender were identical for both the treatment and control.

	Exposure	ID	X1	X2	X3	X4	X5	X6	Age	Race	Gender	Propensity	C11
11	Not Exposed	U	64	48	48	59	56	50	60	Hispanic	Female	0.05299638	
12	Not Exposed	X	13	25	35	8	6	9	52	Hispanic	Female	0.15013131	
13	Not Exposed	Y	30	100	105	53	60	94	72	Black	Female	0.00267398	
14	Not Exposed	BB	20	36	44	17	16	22	33	Black	Female	0.51909545	
15	Exposed	D	31	81	86	46	50	74	33	Hispanic	Female	0.58613607	1
16	Not Exposed	I	46	36	40	39	36	32	39	Hispanic	Female	0.26307673	1
17	Not Exposed	AA	31	22	28	21	17	11	60	Hispanic	Female	0.2987649	1
18	Exposed	F	22	30	39	17	15	17	29	Black	Female	0.07538899	2
19	Not Exposed	K	84	86	82	90	91	97	24	Black	Female	0.12517537	2
20	Not Exposed	O	82	79	75	86	86	89	36	Black	Female	0.08730072	2
21	Not Exposed	B	4	2	5	11	11	5	71	Hispanic	Male	0.01078557	3
22	Exposed	J	3	38	49	5	6	19	58	Hispanic	Male	0.48587995	3
23	Not Exposed	CC	15	35	44	14	13	20	65	Hispanic	Male	0.32609454	3
24	Not Exposed	H	31	91	96	50	56	85	52	Caucasian	Male	0.4250166	4
25	Not Exposed	L	73	88	86	83	85	96	63	Caucasian	Male	0.06839106	4
26	Exposed	S	64	36	36	53	49	37	47	Caucasian	Male	0.78768042	4
27	Not Exposed	W	71	5	5	45	37	7	53	Caucasian	Female	0.14755618	5
28	Exposed	Z	36	67	71	44	46	61	62	Caucasian	Female	0.23760766	5
29	Not Exposed	V	12	63	72	23	26	48	29	Caucasian	Female	0.78284523	6
30	Exposed	DD	46	60	63	49	49	57	28	Caucasian	Female	0.79656871	6
31													

Incomplete Matching Report

Incomplete Matching Report
Exposure = "Exposed"

Row	Matches (Target = 2)	Logit Propensity	ID
1	0	-1.05541	A
14	0	-1.11718	N
26	1	1.16584	Z
30	1	-1.36499	DD

This report lists the treatments that were not paired with the target number of controls (2 in this case). Rows 1 and 14 were not paired with any controls. Rows 26 and 30 were only paired with 1 control. All other treatment rows

were paired with 2 treatments. Incomplete matching is usually due to the use of forced match variables, using caliper matching, or setting Matches per Treatment to 'Maximum Possible'.

Treatment Row

This is the row in the database containing the treatment subject that was not fully matched.

Matches (Target = k)

This is the number of matches that were found for each treatment. The target represents the number of Matches per Treatment specified on the input window.

Propensity Score (or first covariate variable)

This is the value of the propensity score (or logit propensity score if 'Use Logit' was selected) for the incompletely matched treatment. If no propensity score variable was used in distance calculations, then this is the value of first covariate variable specified. The title of this column is based on the propensity score variable name (or label) or the first covariate variable name (or label).

Data Label (e.g. ID)

This is the identification label of the incompletely matched row in the database. The title of this column is the data label variable name (or label).

Example 3 – Matching on Forced Match Variables Only

Continuing with Example 2, suppose we wanted to form matches based solely on forced match variables, i.e., we want the matches to have exactly the same values for each covariate. We could enter all of the covariates in as forced match variables, but with a database as small as we are using, we are unlikely to find any matches. We will use the greedy data matching procedure to illustrate how you can assign matches based on the gender and race forced match variables only. Random ordering is used to ensure that the treatments are randomly paired with controls (where the forced match variable values match).

In order to complete this task, you must first create a new column in the database filled with 1's. You can do this by clicking on the first cell in an empty column and selecting **Edit** > **Fill** from the **NCSS Home** window (for **Fill Value(s)** enter **1**, for **Increment** enter **0**, and click **OK**). A column of ones has already been created for you in the Propensity dataset. This column of ones is necessary because the matching procedure requires either a propensity score variable or a covariate variable to run.

Setup

To run this example, complete the following steps:

1 Open the Propensity example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Propensity** and click **OK**.

2 Specify the Data Matching – Optimal and Greedy procedure options

- Find and open the **Data Matching – Optimal and Greedy** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Grouping Variable	**Exposure**
Treatment Group	**Exposed**
Propensity Score Variable	**Ones**
Use Logit	**Unchecked**
Forced Match Variable(s)	**Race-Gender**
Data Label Variable	**ID**
Store Match Numbers In	**C11**
Distance Calculation Method	**Propensity Score Difference**
Matches per Treatment	**2**
Order for Matching	**Random**
Reports Tab	
Matching Detail Report	**Checked**
Incomplete Matching Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Matching Reports

Matching Detail Report
Treatment = "Exposed", Control = "Not Exposed"

		Treatment			Matched Control		
Match Number	Ones |Difference|	Row	Ones	ID	Row	Ones	ID
1	0.00000	1	1.00000	A	2	1.00000	B
2	0.00000	4	1.00000	D	13	1.00000	M
2	0.00000	4	1.00000	D	27	1.00000	AA
3	0.00000	6	1.00000	F	15	1.00000	O
3	0.00000	6	1.00000	F	17	1.00000	Q
4	0.00000	10	1.00000	J	29	1.00000	CC
5	0.00000	14	1.00000	N	23	1.00000	W
6	0.00000	19	1.00000	S	8	1.00000	H
6	0.00000	19	1.00000	S	12	1.00000	L
7	0.00000	30	1.00000	DD	22	1.00000	V

Incomplete Matching Report
Exposure = "Exposed"

Row	Matches (Target = 2)	Ones	ID
1	1	1.00000	A
10	1	1.00000	J
14	1	1.00000	N
26	0	1.00000	Z
30	1	1.00000	DD

The matching detail report is not very informative because all of the propensity scores are equal to 1. If you run the procedure several times, you will notice that the controls are randomly pairing with the treatments when the race and gender are the same. Your report may be slightly different from this report because random ordering was used. If you sort on C11, you will see that all matched pairs have the same value for race and gender.

13	Exposed	Z	36	67	71	44	46	61	62	Caucasian	Female	0.23760766	
14	Exposed	A	50	102	103	70	75	102	45	Hispanic	Male	0.74181165	1
15	Not Exposed	CC	15	35	44	14	13	20	65	Hispanic	Male	0.32609454	1
16	Exposed	N	64	1	2	38	29	0	39	Caucasian	Female	0.75346448	2
17	Not Exposed	W	71	5	5	45	37	7	53	Caucasian	Female	0.14755618	2
18	Exposed	D	31	81	86	46	50	74	33	Hispanic	Female	0.58613607	3
19	Not Exposed	I	46	36	40	39	36	32	39	Hispanic	Female	0.26307673	3
20	Not Exposed	AA	31	22	28	21	17	11	60	Hispanic	Female	0.2987649	3

Example 4 – Validation of the Optimal Data Matching Algorithm using Rosenbaum (1989)

Rosenbaum (1989) provides an example of both optimal and greedy matching using a well-known dataset from Cox and Snell (1981), which involves 26 U.S. light water nuclear power plants (six "partial turnkey" plants are excluded in the analysis). Seven of the plants were constructed on sites where a light water reactor had existed previously; these are the treatments. The 19 remaining plants serve as the controls. The sum of rank differences was used to calculate distances between treatment and control plants. Two covariate variables were used in the analysis: the date the construction permit was issued (Date), and the capacity of the plant (Capacity). Site was used as the grouping variable with "Existing" as the treatment group. Rosenbaum (1989) reports the following optimal pairings by plant number (treatment, control):

(3,2), (3,21), (5,4), (5,7), (9,7), (9,10), (18,8), (18,13), (20,14), (20,15), (22,17), (22,26), (24,23), (24,25)

The data used in this example are contained in the CoxSnell dataset.

Setup

To run this example, complete the following steps:

1 Open the CoxSnell example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **CoxSnell** and click **OK**.

2 Specify the Data Matching – Optimal and Greedy procedure options

- Find and open the **Data Matching – Optimal and Greedy** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 4** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Grouping Variable	**Site**
Treatment Group	**Existing**
Propensity Score Variable	**<Empty>**
Covariate Variable(s)	**Date-Capacity**
Data Label Variable	**Plant**
Distance Calculation Method	**Sum of Rank Differences including the Propensity Score (if specified)**
Matches per Treatment	**2**
Reports Tab	
Matching Detail Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Matching Reports

Matching Summary Report

Distance Calculation Method	Sum of Rank Differences including the Propensity Score
Order For Matching	Random (Computer-Generated Random Seed: 3319004)
Controls Matched per Treatment	2
Sum of Match Rank Differences	74.00000
Average Match Rank Difference	5.28571

Site	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Existing	7	7	100.00%	0	0.00%
New	19	14	73.68%	5	26.32%

Matching Detail Report

Treatment = "Existing", Control = "New"

		Treatment			Matched Control		
Match Number	Sum of Rank |Differences|	Row	Date	Plant	Row	Date	Plant
1	18.50000	1	2.33000	3	23	3.75000	21
1	0.00000	1	2.33000	3	9	2.33000	2
2	10.50000	2	3.00000	5	12	3.17000	7
2	0.00000	2	3.00000	5	10	3.00000	4
3	5.50000	3	3.42000	9	20	3.42000	16
3	5.50000	3	3.42000	9	14	3.33000	10
4	0.00000	4	3.42000	18	17	3.42000	13
4	2.50000	4	3.42000	18	13	3.42000	8
5	0.00000	5	3.92000	20	18	3.92000	14
5	2.50000	5	3.92000	20	19	3.92000	15
6	5.00000	6	5.92000	22	21	4.50000	17
6	12.00000	6	5.92000	22	26	6.08000	26
7	8.00000	7	5.08000	24	25	5.42000	25
7	4.00000	7	5.08000	24	24	4.67000	23

The optimal match-pairings found by **NCSS** match those in Rosenbaum (1989) exactly. Notice, however, that the distances (Sum of Rank |Differences|) are slightly different in some instances from those given in Table 1 of the article. This is due to the fact that Rosenbaum (1989) rounds all non-integer distances in their reports. This rounding also affects the overall sum of match rank differences; **NCSS** calculates the overall sum as 74, while Rosenbaum (1989) calculates the overall sum as 71, with the difference due to rounding.

Example 5 – Validation of the Greedy Data Matching Algorithm using Rosenbaum (1989)

Continuing with Example 4, Rosenbaum (1989) also reports the results from the greedy matching algorithm, where the order for matching is sorted by distance. The article reports the following greedy pairings by plant number (treatment, control):

(3,2), (3,19), (5,4), (5,21), (9,10), (9,7), (18,8), (18,13), (20,14), (20,15), (22,17), (22,26), (24,23), (24,25)

Setup

To run this example, complete the following steps:

1 Open the CoxSnell example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **CoxSnell** and click **OK**.

2 Specify the Data Matching – Optimal and Greedy procedure options

- Find and open the **Data Matching – Optimal and Greedy** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 5** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Grouping Variable	**Site**
Treatment Group	**Existing**
Propensity Score Variable	**<Empty>**
Covariate Variable(s)	**Date-Capacity**
Data Label Variable	**Plant**
Distance Calculation Method	**Sum of Rank Differences including the Propensity Score (if specified)**
Matches per Treatment	**2**
Reports Tab	
Matching Detail Report	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Output

Matching Summary Report

Distance Calculation Method	Sum of Rank Differences including the Propensity Score
Order For Matching	Sorted by Distance
Controls Matched per Treatment	2
Sum of Match Rank Differences	80.00000
Average Match Rank Difference	5.71429

Site	N	Matched	Percent Matched	Unmatched	Percent Unmatched
Existing	7	7	100.00%	0	0.00%
New	19	14	73.68%	5	26.32%

Matching Detail Report

Treatment = "Existing", Control = "New"

		Treatment			Matched Control		
Match Number	Sum of Rank |Differences|	Row	Date	Plant	Row	Date	Plant
1	0.00000	1	2.33000	3	9	2.33000	2
1	21.00000	1	2.33000	3	22	4.17000	19
2	0.00000	2	3.00000	5	10	3.00000	4
2	15.50000	2	3.00000	5	23	3.75000	21
3	4.00000	3	3.42000	9	12	3.17000	7
3	5.50000	3	3.42000	9	14	3.33000	10
4	0.00000	4	3.42000	18	17	3.42000	13
4	2.50000	4	3.42000	18	13	3.42000	8
5	0.00000	5	3.92000	20	18	3.92000	14
5	2.50000	5	3.92000	20	19	3.92000	15
6	5.00000	6	5.92000	22	21	4.50000	17
6	12.00000	6	5.92000	22	26	6.08000	26
7	4.00000	7	5.08000	24	24	4.67000	23
7	8.00000	7	5.08000	24	25	5.42000	25

The greedy match-pairings found by **NCSS** match those in Rosenbaum (1989) exactly. Again, some of the distances are different from those in Table 1 of the article because of rounding. **NCSS** calculates the overall sum of rank differences as 80, while Rosenbaum (1989) calculates the overall sum as 79 with the difference due to rounding.