Chapter 124

Data Stratification

Introduction

This procedure is used to create stratum assignments based on quantiles from a numeric stratification variable. The user is able to choose the number of strata to create and the amount of data used in the quantile calculations. Stratification is commonly used in the analysis of data from observational studies where covariates are not controlled. This procedure is based on the results given in D'Agostino, R.B., Jr. (2004), chapter 1.2.

Observational Studies

In observational studies, investigators do not control the assignment of treatments to subjects. Consequently, a difference in covariates may exist among treatment groups. Stratification (or subclassification) is often used to control for these differences in background characteristics. Strata are created by dividing subjects into groups based on observed covariates. However, as the number of covariates increases, the number of required strata grows exponentially. Propensity scores, defined as the conditional probability of treatment given a set of covariates, can be used in this situation to account for the presence of uncontrollable covariate factors. Stratification on the propensity score alone can balance the distributions of covariates among groups without the exponential increase in the number of strata. Rosenbaum and Rubin (1984) suggest that the use of five strata often removes 90% or more of the bias in each of the covariates used in the calculation of the propensity score. The propensity score is usually calculated using logistic regression or discriminant analysis with the treatment variable as the dependent (group) variable and the background covariates as the independent variables. For further information about propensity scores, their calculation, and uses, we refer you to the chapter entitled “Data Matching for Observational Studies” in this manual, or chapter 1.2 (pages 67 - 83) of D'Agostino, R.B., Jr. (2004). For more information about logistic regression or discriminant analysis, see the corresponding chapters in the ***NCSS*** manuals.

Data Structure

The data values for stratification must be entered in a single variable (column). Only numeric values are allowed. Missing values are represented by blanks. Text values are treated as missing values. Optional data label and grouping variables may also be used, with each variable representing a single column in the data file. The following is a subset of the Propensity dataset, which will be used in the tutorials that follow.

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

Procedure Options

This section describes the options available in this procedure.

Variables Tab

Specify the variables to be analyzed.

Data Variables

Data Stratification Variable

Specify the variable that contains the numeric data to be used for stratification. In observational studies, propensity scores are commonly used for stratification. Propensity scores are often obtained using logistic regression or discriminant analysis. This variable is required. Only numeric values are analyzed. Text values are treated as missing values in the reports.

Data Label Variable

The values in this variable contain text (or numbers) and are used to identify each row. This variable is optional.

Storage Variable

Store Stratum Numbers In

Specify a variable to store the stratum number assignments for each row. This variable is optional.

Options

Number of Strata

Specify the number of strata to create. The number of strata must be less than the number of rows with non-missing data in the database (or quantile calculation group). Rosenbaum and Rubin (1984) suggest that the use of five strata often removes 90% or more of the bias in each of the covariates used in the calculation of the propensity score.

Calculate Quantiles Using

Select the data that will be used in quantile calculations for stratification. All rows with non-missing data will be assigned to a stratum based on the calculated quantiles. The options are:

- **All Data**
  Use all data for quantile calculations.

- **Data from Quantile Calculation Group**
  Use only the data from the Quantile Calculation Group in quantile calculations. A Grouping Variable and a Quantile Calculation Group must also be specified.

Options – Group Options

Grouping Variable

Specify the variable that contains the quantile calculation group information. The response variable that was used in logistic regression or discriminant analysis to produce the propensity scores is often used as the grouping variable. This variable is only used if Calculate Quantiles Using is set to 'Data from Quantile Calculation Group'. The Quantile Calculation Group must also be specified.

Quantile Calculation Group

Specify the group that is to be used in quantile calculations. The propensity scores in this group only will be used to calculate the quantiles for stratification of the entire database. This option is only used if Determine Quantiles Using is set to 'Data from Quantile Calculation Group'. The Grouping Variable must also be specified.

Reports Tab

The following options control the format of the reports that are displayed.

Select Reports

Run Summary Report ... Strata Detail Report - Sorted by Stratum

Indicate whether to display the indicated reports.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimals

Quantiles and Data Values

Specify the number of digits after the decimal point to be displayed on output values of the type indicated.

Example 1 – Creating Strata Assignments

This section presents an example of how to create a column of stratum assignment numbers from a set of propensity scores. The data used in this example are contained in the Propensity dataset. 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 represent the probability of being exposed given the observed covariate values.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu.

1 Open the Propensity dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Propensity.NCSS**.
- Click **Open**.

2 Open the Data Stratification window.

- Using the Data or Tools menu or the Procedure Navigator, find and select the **Data Stratification** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Data Stratification window, select the **Variables tab**.
- Enter **Propensity** in the **Data Stratification Variable** box.
- Enter **ID** in the **Data Label Variable** box.
- Enter **C11** in the **Store Stratum Numbers In** box.
- Leave all other options at their default values.

4 Specify the reports.

- On the Data Stratification window, select the **Reports tab**.
- Put a check mark next to **Strata Detail Report - Sorted by Row** and **Strata Detail Report - Sorted by Stratum**. Leave all other options at their default values.

5 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Report

Run Summary Report

Data Stratification Variable	Propensity
Data Label Variable	ID
Stratum Number Storage Variable	C11
Quantiles Calculated Using	All Data
Total Number of Rows Read	30
Rows with Non-Missing Data	30
Rows with Missing Data	0
Rows Used in Quantile Calculations	30
Number of Strata Created	5

This report gives a summary of the variables and parameters used in the creation of the strata.

Quantile Report

Quantile	Value
0.20	0.02939
0.40	0.08015
0.60	0.25289
0.80	0.57273

This report shows the values of the four quantiles necessary to create five strata. The length of this report depends on the number of strata desired.

Quantile

This is the quantile calculated. The number of quantiles required is equal to the number of strata minus one.

Value

This is the value of the q^{th} quantile. The $100q^{th}$ quantile is computed as

$$Z_q = (1-g)X[k_1] + gX[k_2]$$

where

Z_q is the value of the quantile,

q is the fractional value of the quantile (for example, for the 75th quantile, q = .75),

$X[k]$ is the k^{th} observation when the data are sorted from lowest to highest,

k_1 is the integer part of $q(n+1)$,

$k_2 = k_1 + 1$,

g is the fractional part of $q(n+1)$ (for example, if $q(n+1) = 23.42$, then g = .42),

n is the total sample size.

Strata Summary Report

Stratum Number	Size	Range
1	6	Propensity <= 0.02939
2	6	0.02939 < Propensity <= 0.08015
3	6	0.08015 < Propensity <= 0.25289
4	6	0.25289 < Propensity <= 0.57273
5	6	Propensity > 0.57273

This report provides a summary of the strata created.

Stratum Number

This is the number assigned to the stratum. These represent the values stored on the database in the stratum storage variable (if specified).

Size

This is the number of rows (or subjects) in each stratum.

Range

This is the propensity score interval associated with each stratum.

Strata Detail Report – Sorted by Row

Strata Detail Report - Sorted by Row

Row	Stratum Number	Propensity	ID
1	5	0.74181	A
2	1	0.01079	B
3	1	0.00087	C
4	5	0.58614	D
5	3	0.11743	E
.	.	.	.
.	.	.	.
.	.	.	.

This report provides a row-by-row list of the assigned stratum numbers, sorted by row.

Row

This is the row on the database.

Stratum Number

This is the number of the stratum to which the observation was assigned. This represents the value stored on the database in the stratum storage variable (if specified).

Data Value (e.g. Propensity)

This is the data value for this row. The title of this column depends on the name (or label) of the Data Stratification Variable.

Data Label (e.g. ID)

This is the data label value for this row. The title of this column depends on the name (or label) of the Data Label Variable.

Strata Detail Report – Sorted by Stratum

Strata Detail Report - Sorted by Stratum

Row	Stratum Number	Propensity	ID
3	1	0.00087	C
25	1	0.00267	Y
13	1	0.00379	M
7	1	0.00829	G
2	1	0.01079	B
17	1	0.02861	Q
20	2	0.03253	T
16	2	0.03604	P
18	2	0.04800	R
21	2	0.05300	U
12	2	0.06839	L
6	2	0.07539	F
15	3	0.08730	O
5	3	0.11743	E
.	.	.	.
.	.	.	.
.	.	.	.

This report provides a row-by-row list of the assigned stratum numbers, sorted by stratum number and data value (e.g. “Propensity”).