Chapter 129

Group By (Break Variables)

Introduction

Group By columns (also known as Break Variables) are used to quickly divide your data into groups without using a Filter. When running a procedure, a separate analysis is conducted for each group, and separate statistical reports and plots are generated for each group. Group By columns should contain categorical data.

If a single Group By column is selected, groups are generated for each unique value within the column. If multiple Group By columns are selected, groups are generated as combinations of the unique values within each column.

Data filters may be used in conjunction with Group By columns to further refine groups. For more information about specifying Filters, see the Filters chapter in the Help Documentation.

Selecting the Group By Columns

It is easy to specify Group By columns using the Group By Editor window. The Group By Editor is a simple point-and-click tool for selecting Group By columns.

There are several ways to launch the Group By Editor.

1.   From the Data Window (or Procedure Window) main toolbar, click on the *Group By* dropdown button and select *Load the Group By Editor*.

2.   From the Data Window menu, select *Data > Group By > Edit Group By Columns*.

3.   From the Data Window menu, select *Data > Column X Info > Group By* to edit the Group By status for the column containing the active cell.

4.   Click on the button at the right side of a Group By cell when it is selected in the Column Info Table to edit the Group By status for the column containing the active cell.

5.   Select a cell in the Data Table, right-click, and select *Column X Info > Edit Group By* from the popup menu.

To quickly edit the Group By columns without using the Group By Editor, click on the "Column Info" tab on the Data Window. In the row (or column) labeled "Group By", select a cell and type "Yes", "Y", "True", "T", or "X". You can also click the associated edit button to assign the column as a Group By column.

Specifying the Order in which Group By Columns are Applied

The Group By Editor automatically returns appropriate ordering values when multiple Group By columns are selected. To specify the order in which to apply the Group By columns in the Column Info table, enter a numeric value in parentheses after the Group By indicator (e.g. "Yes (1)").

For example, suppose columns C5 (Values: 1, 2) and C6 (Values: A, B) to be designated as Group By columns. To assign C5 as the first Group By column and C6 as the second Group By column you would enter "Yes (1)" in the Group By cell for C5 and "Yes (2)" in the Group By cell for C6. The 4 resulting groups would be

> C5 = 1, C6 = A
> C5 = 1, C6 = B
> C5 = 2, C6 = A
> C5 = 2, C6 = B

and reports would be generated in this order. However, if you wanted to change the order in which reports are generated such that C6 is displayed first, you would enter "Yes (2)" in the Group By cell for C5 and "Yes (1)" in the Group By cell for C6.

The 4 resulting groups would be

> C6 = A, C5 = 1
> C6 = A, C5 = 2
> C6 = B, C5 = 1
> C6 = B, C5 = 2

Note that the total number of groups is the same, but they are assigned in a different order.

Activating and Deactivating the Group By System

The *Group By Active* option in the Group By button dropdown on the Data Window (or Procedure Window) main toolbar controls whether the Group By system is active. When running any procedure with Group By Active checked, a separate analysis will be conducted for each group.

Activating and Deactivating the Group By system will not delete the information about columns designated as Group By columns. This option just provides a simple way to indicate whether to apply data grouping when running an analysis. You can activate and deactivate the Group By system in between analyses as often as you would like.

Example 1 – Data Analysis using Group By Columns

In this section, we will use the Fisher example dataset to demonstrate the use of a Group By column to quickly analyze smaller groups of data. First, we will run the analysis without specifying a Group By column.

Analysis with No Group By Column

To run this example, complete the following steps:

1   Open the Fisher example dataset
- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Fisher** and click **OK**.

2   Specify the Linear Regression and Correlation procedure options
- Find and open the **Linear Regression and Correlation** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below.

Variables Tab

Y Dependent Variable(s) ....**SepalLength**
X Independent Variable ....**PetalLength**

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

Output

Linear Regression Plot Section
_____

Regression Estimation Section
_____

Parameter	Intercept B(0)	Slope B(1)
Regression Coefficients	43.0660	0.4089
Lower 95% Confidence Limit	41.5170	0.3716
Upper 95% Confidence Limit	44.6151	0.4463
Standard Error	0.7839	0.0189
Standardized Coefficient	0.0000	0.8718
T Value	54.9389	21.6460
Prob Level (T Test)	0.0000	0.0000
Reject H0 (Alpha = 0.0500)	Yes	Yes
Power (Alpha = 0.0500)	1.0000	1.0000
Regression of Y on X	43.0660	0.4089
Inverse Regression from X on Y	38.2120	0.5381
Orthogonal Regression of Y and X	42.3428	0.4282

Estimated Model
_____

(43.0660341504758) + (0.408922277351185) * (Petal Length)
_____

Notes:
The above report shows the least-squares estimates of the intercept and slope followed by the corresponding standard errors, confidence intervals, and hypothesis tests. Note that these results are based on several assumptions that should be validated before they are used.

A single report is generated for all the data.

Analysis with Iris (Column 5) as the Group By Column

Now, we will assign Iris, which has three unique values (1, 2, 3), as a Group By column and rerun the analysis to get the results for each individual group. (If you have already followed the first part of this example above, steps 2 and 5 are all you need to complete now.)

To run this example, complete the following steps:

1   Open the Fisher example dataset

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

2   Assign Iris as a Group By column

- On the Data Window main toolbar, select the **Group By** dropdown button and click **Load the Group By Editor**. On the Group By Editor window, select **Iris** as the Group By column.

or

- On the Data Window in the **Column Info** table, enter **Yes** in the **Group By** cell for **Iris** (column 5).

3   Specify the Linear Regression and Correlation procedure options

- Find and open the **Linear Regression and Correlation** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below.

Variables Tab

Y Dependent Variable(s) ....**SepalLength**
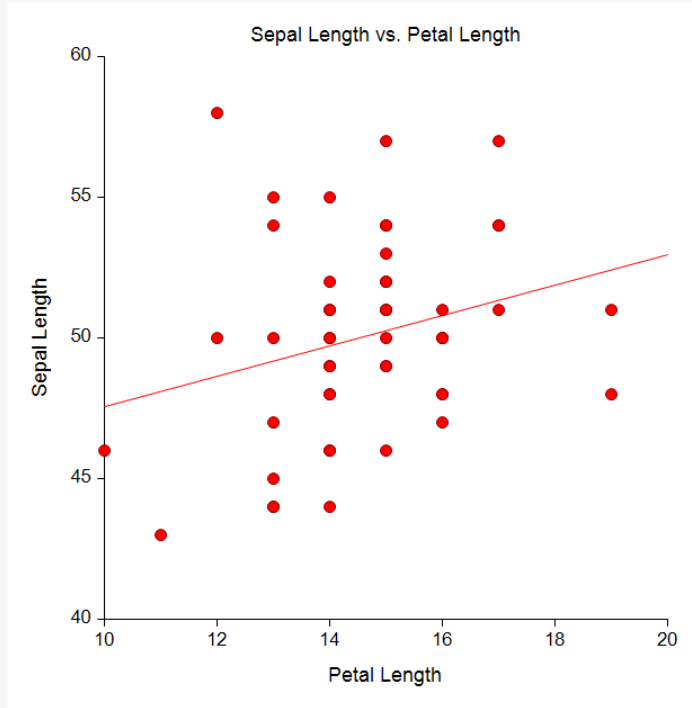X Independent Variable ....**PetalLength**

4   Run the procedure

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

Output

Group   Iris = Setosa

Linear Regression Plot Section
_____

Regression Estimation Section
_____

Parameter	Intercept B(0)	Slope B(1)
Regression Coefficients	42.1317	0.5423
Lower 95% Confidence Limit	33.7757	-0.0253
Upper 95% Confidence Limit	50.4877	1.1099
Standard Error	4.1559	0.2823
Standardized Coefficient	0.0000	0.2672
T Value	10.1378	1.9209
Prob Level (T Test)	0.0000	0.0607
Reject H0 (Alpha = 0.0500)	Yes	No
Power (Alpha = 0.0500)	1.0000	0.4691
Regression of Y on X	42.1317	0.5423
Inverse Regression from X on Y	-61.0075	7.5970
Orthogonal Regression of Y and X	-36.5167	5.9218

Estimated Model
_____
(42.1316822303424) + (0.542292597103796) * (Petal Length)
_____

Notes:
The above report shows the least-squares estimates of the intercept and slope followed by the corresponding standard errors, confidence intervals, and hypothesis tests. Note that these results are based on several assumptions that should be validated before they are used.

Group   Iris = Versicolor

Linear Regression Plot Section
_____
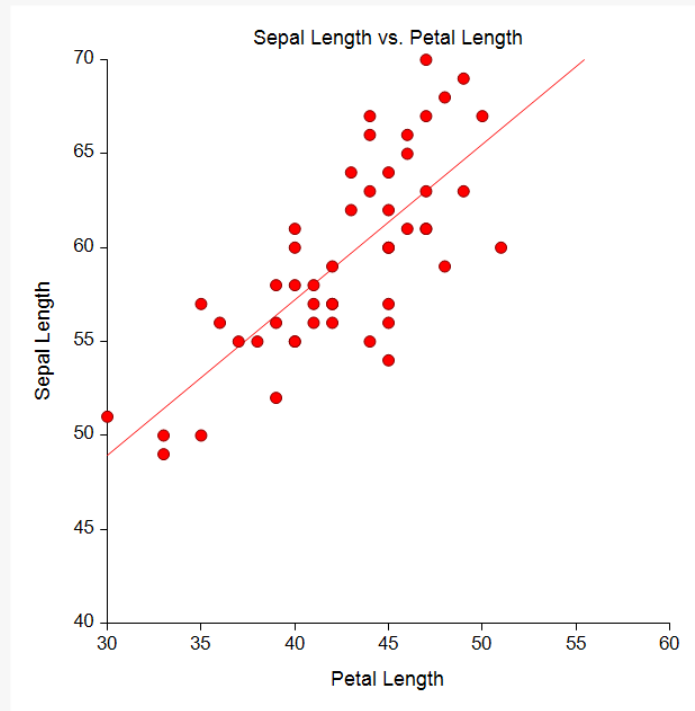
Sepal Length vs. Petal Length

Regression Estimation Section
_____

Parameter	Intercept B(0)	Slope B(1)
Regression Coefficients	24.0752	0.8283
Lower 95% Confidence Limit	15.1026	0.6189
Upper 95% Confidence Limit	33.0479	1.0377
Standard Error	4.4626	0.1041
Standardized Coefficient	0.0000	0.7540
T Value	5.3949	7.9538
Prob Level (T Test)	0.0000	0.0000
Reject H0 (Alpha = 0.0500)	Yes	Yes
Power (Alpha = 0.0500)	0.9996	1.0000
Regression of Y on X	24.0752	0.8283
Inverse Regression from X on Y	-2.6966	1.4567
Orthogonal Regression of Y and X	11.1176	1.1324

Estimated Model
_____

(24.0752310536045) + (0.828280961182998) * (Petal Length)

_____

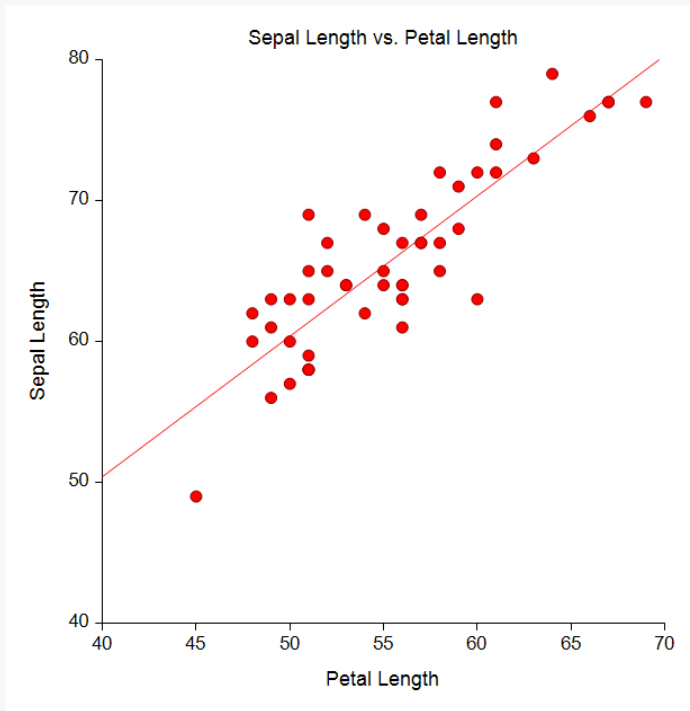Notes:
The above report shows the least-squares estimates of the intercept and slope followed by the corresponding standard errors, confidence intervals, and hypothesis tests. Note that these results are based on several assumptions that should be validated before they are used.

Group   Iris = Virginica

Linear Regression Plot Section

Sepal Length vs. Petal Length

Regression Estimation Section

Parameter	Intercept B(0)	Slope B(1)
Regression Coefficients	10.5966	0.9957
Lower 95% Confidence Limit	1.2116	0.8275
Upper 95% Confidence Limit	19.9816	1.1640
Standard Error	4.6677	0.0837
Standardized Coefficient	0.0000	0.8642
T Value	2.2702	11.9011
Prob Level (T Test)	0.0277	0.0000
Reject H0 (Alpha = 0.0500)	Yes	Yes
Power (Alpha = 0.0500)	0.6044	1.0000
Regression of Y on X	10.5966	0.9957
Inverse Regression from X on Y	-8.1387	1.3332
Orthogonal Regression of Y and X	0.4837	1.1779

Estimated Model

(10.5965909090901) + (0.995738636363612) * (Petal Length)

Notes:
The above report shows the least-squares estimates of the intercept and slope followed by the corresponding standard errors, confidence intervals, and hypothesis tests. Note that these results are based on several assumptions that should be validated before they are used.

A separate report is generated for each of the three groups (Iris = 1, Iris = 2, Iris = 3).

Example 2 – Reordering Groups in a Group By Column and Applying a Filter to Refine the Analysis and Restrict Output

In this section, we will use the Exercise example dataset to demonstrate how to reorder groups within a Group By column and also demonstrate how to apply a filter in conjunction with Group By columns to restrict the analysis to a smaller number of the available groups.

Suppose you want to group analyses by Exercise (with three levels: 0 - None, 1 - Weekly, 2 - Daily) together with Time (with three levels: 0, 10, 20). Combining the levels of these two columns creates 9 groups:

1: Exercise = 0 - None, Time = 0
2: Exercise = 0 - None, Time = 10
3: Exercise = 0 - None, Time = 20
4: Exercise = 1 - Weekly, Time = 0
5: Exercise = 1 - Weekly, Time = 10
6: Exercise = 1 - Weekly, Time = 20
7: Exercise = 2 - Daily, Time = 0
8: Exercise = 2 - Daily, Time = 10
9: Exercise = 2 - Daily, Time = 20

These 9 groups could potentially produce a lot of output to sift through. Suppose you only want to see output for groups with daily or weekly exercise and a time of 10 or 20. This would restrict the analysis to 4 groups:

1: Exercise = 1 - Weekly, Time = 10
2: Exercise = 1 - Weekly, Time = 20
3: Exercise = 2 - Daily, Time = 10
4: Exercise = 2 - Daily, Time = 20

Finally, suppose you want to reorganize the groups so that time is displayed first in descending order (i.e. 20 before 10) and exercise is displayed second in default order. This would result in 4 reordered groups:

1: Time = 20, Exercise = 1 - Weekly
2: Time = 20, Exercise = 2 - Daily
3: Time = 10, Exercise = 1 - Weekly
4: Time = 10, Exercise = 2 - Daily

Note that the groups are the same, they are just displayed in different order.

The steps below will show you how to set up the data and analysis to get this ordering. **To skip the steps that setup the dataset (steps 2 through 4), simply load the "Exercise (Group By)" example dataset in step 1.**

1   Open the Exercise example dataset
- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Exercise** and click **OK**. (Select **Exercise (Group By)** to skip to step 5.)

2   Assign Exercise and Time as a Group By columns.

- On the Data Window main toolbar, select the **Group By** dropdown button and click **Load the Group By Editor**. On the Group By Editor window, select **Time** then **Exercise** as the Group By columns. (The resulting Group By Columns text should be **Time, Exercise**.).

  or

- On the Data Window in the **Column Info** table, enter **Yes (2)** in the **Group By** cell for **Exercise** (column 1) and enter **Yes (1)** in the **Group By** cell for **Time** (column 3).

3   Add Filters for Exercise and Time.

- On the Data Window main toolbar, select the **Filter** dropdown button and click **Load the Filter Editor**. Add a **List Condition** for **Exercise** that keeps only **1 – Weekly** and **2 – Daily**. Add a second **List Condition** for **Time** that keeps only **10** and **20**.

  or

- On the Data Window in the **Column Info** table, enter **"1 - Weekly", "2 - Daily"** in the **Filter** cell for **Exercise** (column 1) and enter **10, 20** in the **Filter** cell for **Time** (column 3).
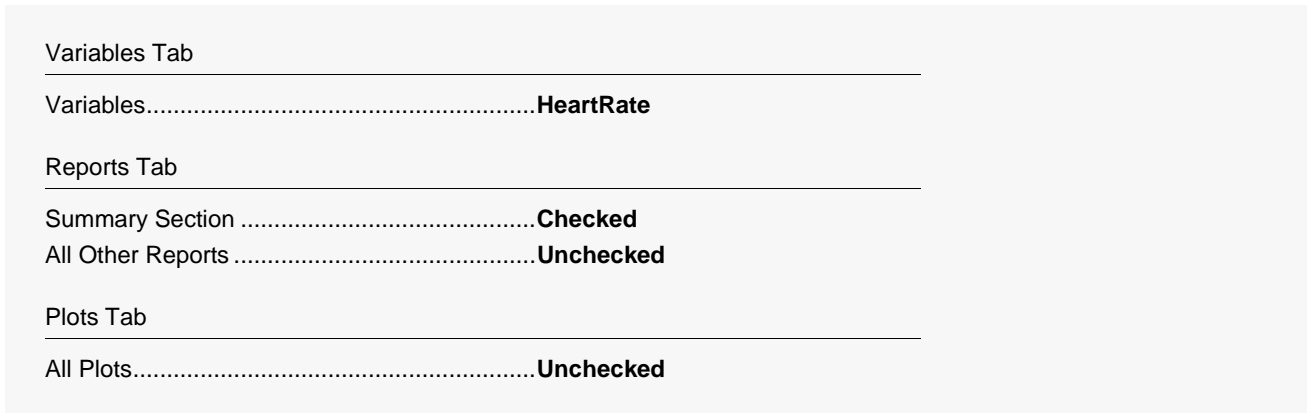
4   Change the Value Order for Time.

- On the Data Window in the **Column Info** table, click the **Edit Button** in the **Value Order** cell for **Time** (column 3) then select and change the **Custom Order** to **20 10 0**.

  or

- On the Data Window in the **Column Info** table, enter **20,10,0** in the **Value Order** cell for **Time** (column 3).

5   Specify the Descriptive Statistics procedure options

- Find and open the **Descriptive Statistics** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below.

Variables Tab
_____

Variables...**HeartRate**

Reports Tab
_____

Summary Section ....**Checked**
All Other Reports ....**Unchecked**

Plots Tab
_____

All Plots...**Unchecked**

---

6   Run the procedure

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

Output

Group   Time = 20, Exercise = 1 - Weekly

Summary Section of HeartRate

Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
6	64.16666	9.600348	3.919325	52	74	22

Group   Time = 20, Exercise = 2 - Daily

Summary Section of HeartRate

Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
6	60.83333	8.953584	3.655285	51	72	21

Group   Time = 10, Exercise = 1 - Weekly

Summary Section of HeartRate

Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
6	66.83334	11.30339	4.61459	53	78	25

Group   Time = 10, Exercise = 2 - Daily

Summary Section of HeartRate

Count	Mean	Standard Deviation	Standard Error	Minimum	Maximum	Range
6	61.5	10.74709	4.387482	49	76	27

A separate report is generated for each of the four groups. The reports are displayed in the desired order.