Linear and Logistic Regression with Interactions

Author: Alison Figueira, Golden Helix, Inc.

Overview

This script will output the results from either a Linear or Logistic Regression Analysis run with one dependent variable, multiple interacting, and non-interacting covariates on all numeric columns. This script uses the *numpy*, *scipy*, and *statsmodels* packages.

Recommended Directory Location

Save the script to the following directory:
*..\AppData\Local\Golden Helix SVS\UserScripts\Spreadsheet\Numeric*

Note: The *AppData (or Application Data)* folder is a hidden folder on Windows operating systems and its location varies between various versions. The easiest way to locate this directory on your computer is to open SVS and select the **Tools >Open Folder > UserScripts Folder** menu option and save the script in the \Spreadsheet\Numeric\ folder. If saved to the proper folder, this script will be accessible from the spreadsheet menu.

Using the Script

1. Open the spreadsheet containing the data to be analyzed. The data should be by column, such as the example below.

Figure 1: Example Spreadsheet with the data column wise.
Make sure to inactive (gray) any columns that you do not wish to include in your analysis. Also, the dependent column can be chosen here by setting the column to dependent (magenta).

2. While in the spreadsheet window, select **Numeric > Linear and Logistic Regression with Interactions**

3. In the first box of the prompt window, add the column from your spreadsheet that contains the dependent, (if it wasn’t already selected in the spreadsheet window). In the second box, select covariates that do not interact with the other numeric columns. In the third box, select covariates that do interact with the other numeric columns.

![Figure 2: Prompt Dialog with dependent and covariates chosen.](image)
4. Click **OK** to run the analysis.

5. During the analysis, if there are any columns with non-unique values, they will not be included in the analysis and their row in the results spreadsheet will contain missing values.

6. When done, a results spreadsheet, called “Regression Results,” will be created.

7. The Marker Map from the original spreadsheet will try to be preserved and it will appear row oriented.

8. The final spreadsheet will have the predictors by row with results by column. Output will depend on the test (Linear or Logistic) and whether interacting and/or non-interacting covariates were chosen. If covariates were chosen, Full vs. Reduced results will be included. If interacting covariates were chosen, then Interaction terms and standard errors for those terms will be included.

![Figure 3: Example Results](image)

**Regression Model:**

The basic regression model used in this script is as follows:

\[ Y = B_0 + B_1x + B_2x + SNPx + B_3SNPx + e \]

- **Y** is the dependent variable
- **B_0** is the intercept term
$B_1$ are the covariates that do not interact with the column or SNP that is being examined.
$B_2$ are the covariates that do interact with the column or SNP that is being examined.
SNP is the column or SNP that is being examined.
$B_2$SNP are the interaction term(s).
e is the error term.

Covariates are added into the model in the same order they appear in the first dialog window (or similarly, the same order they appear in the spreadsheet). The interaction terms are in the same order as the $B_2$ (covariates that interact) terms.
For categorical covariates, the betas apply the categories in alphabetical order with the reference being the first one alphabetically.