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.

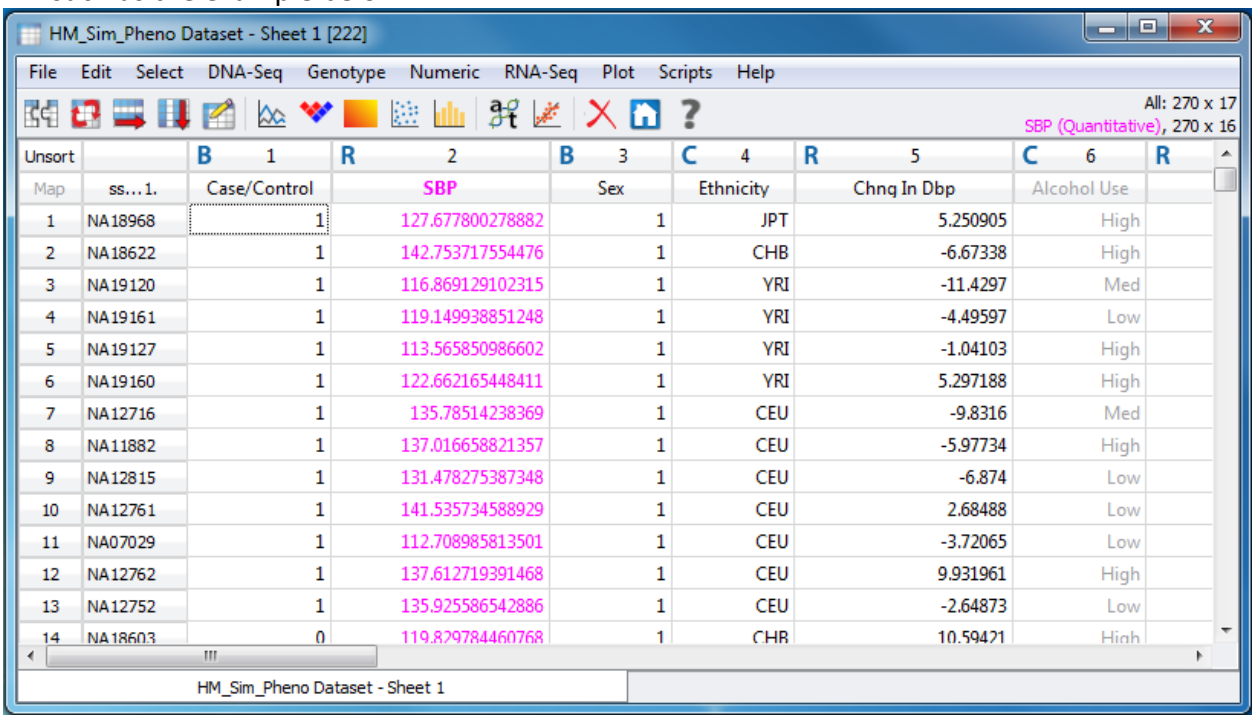
Recommended Directory Location

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.



Unsort		B 1	R 2	B 3	C 4	R 5	C 6	R
Map	ss...1.	Case/Control	SBP	Sex	Ethnicity	Chng In Dbp	Alcohol Use	
1	NA18968	1	127.677800278882	1	JPT	5.250905	High	
2	NA18622	1	142.753717554476	1	CHB	-6.67338	High	
3	NA19120	1	116.869129102315	1	YRI	-11.4297	Med	
4	NA19161	1	119.149938851248	1	YRI	-4.49597	Low	
5	NA19127	1	113.565850986602	1	YRI	-1.04103	High	
6	NA19160	1	122.662165448411	1	YRI	5.297188	High	
7	NA12716	1	135.78514238369	1	CEU	-9.8316	Med	
8	NA11882	1	137.016658821357	1	CEU	-5.97734	High	
9	NA12815	1	131.478275387348	1	CEU	-6.874	Low	
10	NA12761	1	141.535734588929	1	CEU	2.68488	Low	
11	NA07029	1	112.708985813501	1	CEU	-3.72065	Low	
12	NA12762	1	137.612719391468	1	CEU	9.931961	High	
13	NA12752	1	135.925586542886	1	CEU	-2.64873	Low	
14	NA18603	0	119.829784460768	1	CHB	10.59421	High	

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.

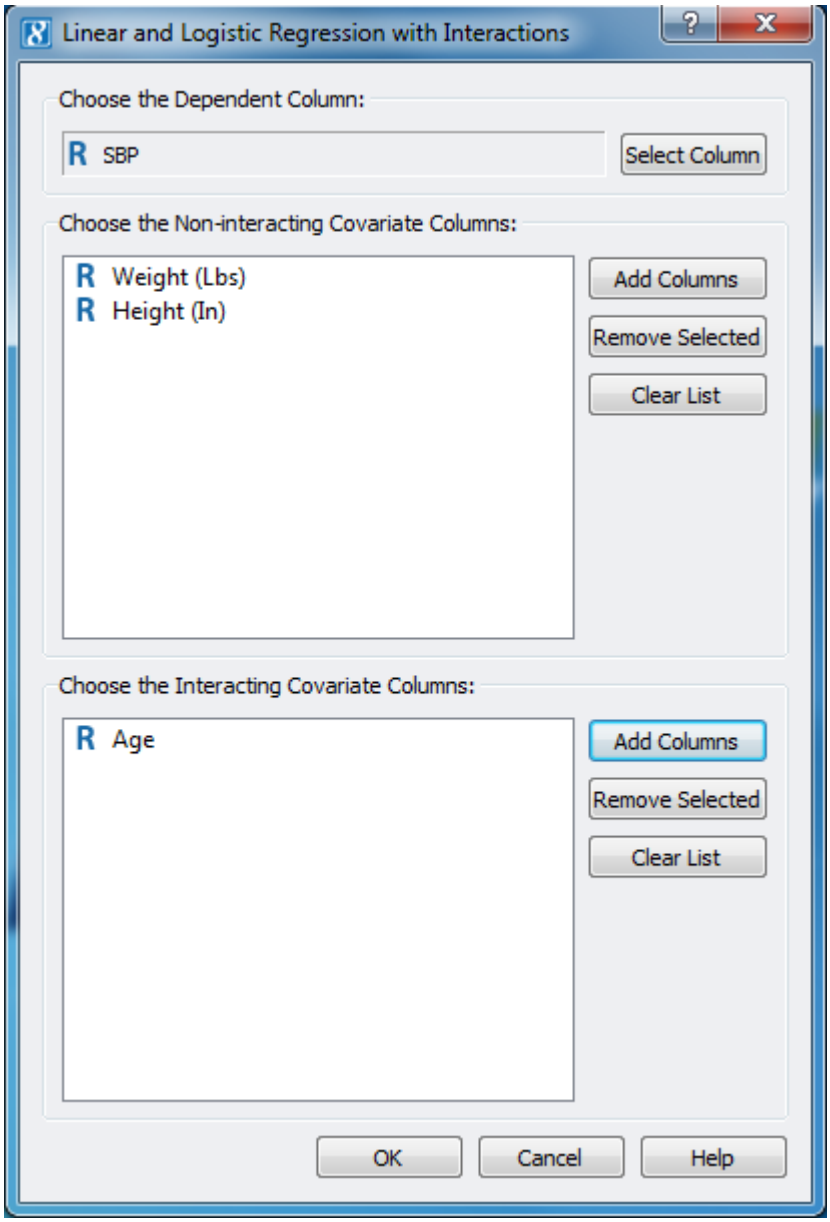


Figure 2: Prompt Dialog with dependent and covariates chosen.

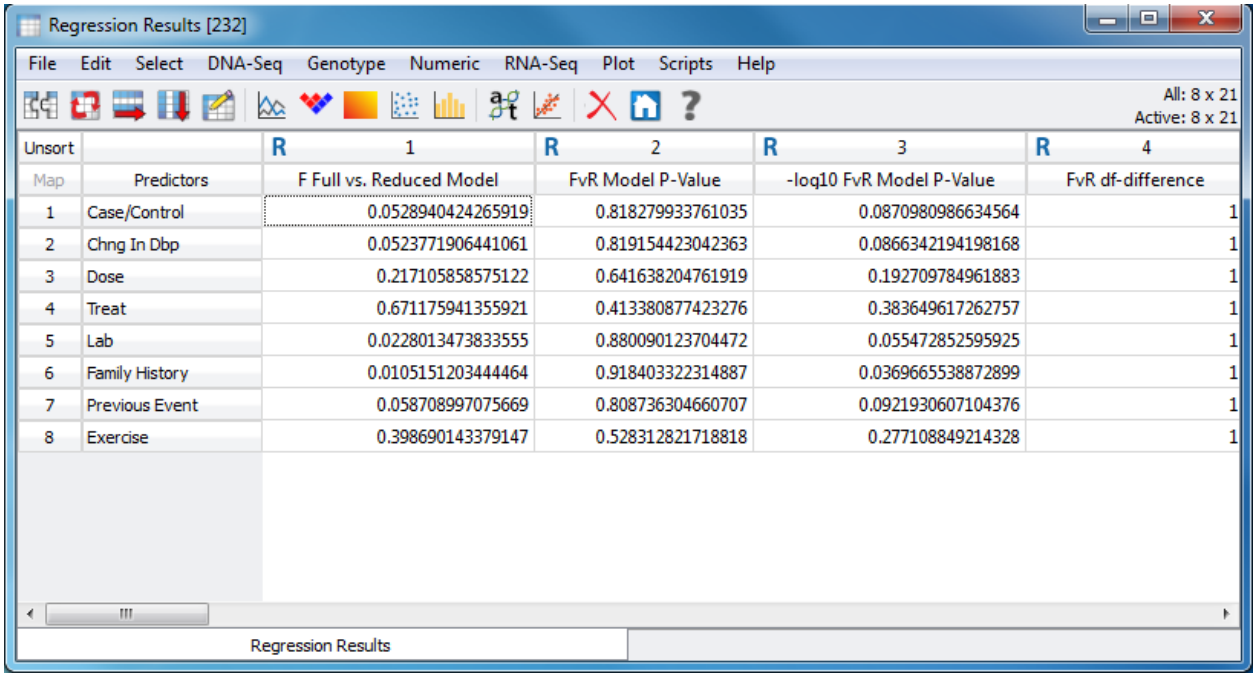
4. Click **OK** to run the analysis.

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.



Unsort		R 1	R 2	R 3	R 4
Map	Predictors	F Full vs. Reduced Model	FvR Model P-Value	-log10 FvR Model P-Value	FvR df-difference
1	Case/Control	0.0528940424265919	0.818279933761035	0.0870980986634564	1
2	Chng In Dbp	0.0523771906441061	0.819154423042363	0.0866342194198168	1
3	Dose	0.217105858575122	0.641638204761919	0.192709784961883	1
4	Treat	0.671175941355921	0.413380877423276	0.383649617262757	1
5	Lab	0.0228013473833555	0.880090123704472	0.055472852595925	1
6	Family History	0.0105151203444464	0.918403322314887	0.0369665538872899	1
7	Previous Event	0.058708997075669	0.808736304660707	0.0921930607104376	1
8	Exercise	0.398690143379147	0.528312821718818	0.277108849214328	1

Figure 3: Example Results

Regression Model:

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

$$Y = B_0 + B_1 x + B_2 x + SNPx + B_2 SNPx + e$$

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_2SNP 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.