Journal of Statistical Software

January 2005, Volume 12, Issue 2. http://www.jstatsoft.org/

SimReg: A Software Includinge Some New Developments in Multiple Comparison and Simultaneous Confidence Bands for Linear Regression Models

Mortaza Jamshidian
California State University

Wei Liu
University of Southampton

Ying Zhang
University of Iowa

Farid Jamshidian
University of California at Los Angeles

Abstract

The problem of simultaneous inference and multiple comparison for comparing means of $k(\geq 3)$ populations has been long studied in the statistics literature and is widely available in statistics literature. However to-date, the problem of multiple comparison of regression models has not found its way to the software. It is only recently that the computational aspects of this problem have been resolved in a general setting. **SimReg** employs this new methodology and provides users with software for multiple regression of several regression models. The comparisons can be among any set of pairs, and moreover any number of predictors can be included in the model. More importantly predictors can be constrained to their natural boundaries, if known.

Computational methods for the problem of simultaneous confidence bands when predictors are constrained to intervals has also recently been addressed. **SimReg** utilizes this recent development to offer simultaneous confidence bands for regression models with any number of predictor variables. Again, the predictors can be constrained to their natural boundaries which results in narrower bands, as compared to the case where no restriction is imposed. A by-product of these confidence bands is a new method for comparing two regression surfaces, that is more informative than the usual partial F test.

Keywords: linear regression, multiple comparison, simultaneous confidence bands, partial F test, statistic software.

1. Introduction

The problem of simultaneous inference and multiple comparison for comparing means of $k(\geq 3)$ populations has been long studied in the statistics literature. Miller (1981), Hochberg and Tamhane (1987), and Hsu (1996) have provided excellent summaries of the work in this area. Spurrier (1999) seems to be the pioneering work to extend this problem to simultaneous comparison of several regression lines. In his work, a set of simultaneous confidence bands for all the contrasts of several simple linear regression lines over the entire range of the single predictor $(-\infty, \infty)$ is constructed when the design matrices of the regression lines are the same. The recent work of Liu, Jamshidian, and Zhang (2004) [hereafter referred to as LJZ (2004)] extends Spurrier's work in many directions. Specifically, their work allows multiple comparison of several regression models with (1) any number of predictor variables, (2) design matrices that are not necessarily equal, (3) predictors that can be restricted in a finite or infinite range, and (4) comparison of any desired set of pairs of groups. The software **SimReg**, presented in this paper, implements this new methodology and provides a user-friendly interface for data input and analysis.

SimReg also provides simultaneous confidence bands for the linear regression model when predictors are constrained to intervals. This problems has a history going back to Working and Hotelling (1929). Scheffé (1953) provides a simultaneous confidence band when predictors are not constrained, i.e. they range in the interval $(-\infty, \infty)$. Since then there is a fair amount of literature that provide exact simultaneous confidence bands for simple linear regression model with the single predictor constrained to intervals. There are also works that offer conservative simultaneous confidence bands for the cases where predictors are constrained in intervals or other types of regions. Liu, Jamshidian, and Zhang (2005a) [hereafter referred to as LJZ (2005)] seem to be the first paper to offer "exact" simultaneous confidence bands for multiple regression when predictors are constrained in finite intervals. As mentioned, **SimReg** provides these intervals.

The remaining sections are organized as follows: Section 2 describes the mathematical model for the multiple comparison problem, and Section 3 does the same for the simultaneous confidence band problem. Section 4 gives a brief discussion of the algorithm and parameter settings. Section 5 describes how the software can be installed and used. Finally Section 5 gives a few items that are planned for future versions of the software.

2. The multiple comparison problem

Suppose that data are observed for k groups of subjects and let the model

$$\mathbf{Y}_i = X_i \mathbf{b}_i + \mathbf{e}_i, \quad i = 1, \cdots, k$$

be the linear regression model for the ith group. Here, for the ith group, $\mathbf{Y}_i^T = (y_{i1}, \cdots, y_{in_i})$ denotes the vector of responses, X_i is an $n_i \times (p+1)$ full column rank design matrix with the first column given by $(1, \cdots, 1)^T$ and the $l(\geq 2)$th column given by $(x_{1,l-1}^i, \cdots, x_{n_i,l-1}^i)^T$, $\mathbf{b}_i^T = (b_0^i, \cdots, b_p^i)$ is the vector of regression coefficients, and $\mathbf{e}_i^T = (e_{i1}, \cdots, e_{in_i})$ denotes the vector of regression errors with all the $\{e_{ij}, j = 1, \cdots, n_i, i = 1, \cdots, k\}$ being iid $N(0, \sigma^2)$. Since $X_i^T X_i$ is non-singular, the least squares estimator of $\mathbf{b}_i$ is given by $\hat{\mathbf{b}}_i = (X_i^T X_i)^{-1} X_i^T Y_i, i = 1, \cdots, k$. Let $\hat{\sigma}^2$ denote the pooled error mean square with degrees of freedom $\nu = \sum_{i=1}^k (n_i - p - 1)$; $\hat{\sigma}^2$ is independent of the $\hat{\mathbf{b}}_i$.

LJZ(2004) proposed a method to construct a set of simultaneous confidence bands for

$$\mathbf{x}^T\mathbf{b}_i - \mathbf{x}^T\mathbf{b}_j = (1, x_1, \cdots, x_p)\mathbf{b}_i - (1, x_1, \cdots, x_p)\mathbf{b}_j, \quad (i,j) \in \Lambda$$

over a given range $x_l \in [a_l, b_l], l = 1, \cdots, p$, where Λ is an index set that determines the comparison of interest. For example, if the pairwise comparison is of interest then $\Lambda = \{(i,j) : 1 \leq i \neq j \leq k\}$; if the comparisons of the second to kth regression models with the first regression model are of interest then $\Lambda = \{(i,j) : 2 \leq i \leq k, j = 1\}$; if the successive comparison of the k regression models is of interest then $\Lambda = \{(i, i+1) : 1 \leq i \leq k-1\}$. Specifically, they construct the following set of simultaneous confidence bands

$$\mathbf{x}^T\mathbf{b}_i - \mathbf{x}^T\mathbf{b}_j \in \mathbf{x}^T\hat{\mathbf{b}}_i - \mathbf{x}^T\hat{\mathbf{b}}_j \pm c\hat{\sigma}\sqrt{\mathbf{x}^T\Delta_{ij}\mathbf{x}}\ , \forall x_l \in [a_l, b_l] \ \text{ for } l = 1, \cdots, p, \ \text{ and } \forall (i,j) \in \Lambda \tag{2.1}$$

where $\Delta_{ij} = (X_i^T X_i)^{-1} + (X_j^T X_j)^{-1}$, and c is the critical constant required so that the confidence level of this set of simultaneous confidence bands is equal to $1-\alpha$. The confidence level of the bands in (1.1) is given by $P\{T < c\}$ where

$$T = \sup_{(i,j)\in\Lambda} \sup_{x_l \in [a_l, b_l], l=1,\cdots,p} \frac{|\mathbf{x}^T[(\hat{\mathbf{b}}_i - \mathbf{b}_i) - (\hat{\mathbf{b}}_j - \mathbf{b}_j)]|}{\hat{\sigma}\sqrt{\mathbf{x}^T\Delta_{ij}\mathbf{x}}}. \tag{2.2}$$

LJZ (2004) provide a simulation procedure to simulate T and obtain the critical constant. **SimReg** adopts this methodology to obtain c, however, as we will explain in Section 4, we adopt a different algorithm than that proposed by LJZ (2004).

3. Simultaneous confidence bands

Consider the multiple linear regression model

$$\mathbf{Y} = X\mathbf{b} + \mathbf{e} \tag{3.1}$$

where $\mathbf{Y}_{n\times 1}$ is the vector of observed responses, $X_{n\times(p+1)}$ is the design matrix with the first column given by $(1, \cdots, 1)^T$ and the jth $(2 \leq j \leq p+1)$ column given by $(x_{1,j-1}, \cdots, x_{n,j-1})^T$, $\mathbf{b}_{(p+1)\times 1} = (b_0, \cdots, b_p)^T$ is the vector of regression coefficients, and $\mathbf{e}_{n\times 1}$ be the error vector with $\mathbf{e} \sim N(0, \sigma^2 I)$ and σ^2 unknown. Assume $X^T X$ is non-singular, so the least squares estimator of $\mathbf{b}$ is given by $\hat{\mathbf{b}} = (X^T X)^{-1} X^T \mathbf{Y}$. Let $\hat{\sigma}^2$ denote the mean square error with degrees of freedom $\nu = n - p - 1$; $\hat{\sigma}^2 \sim \sigma^2 \chi^2_\nu / \nu$ and is independent of the $\hat{\mathbf{b}}$.

It has been argued by several authors that a rectangular $\mathcal{X}$ is often one of the most useful shapes for practical purposes (see e.g., Casella and Strawderman (1980), and Naiman (1987)). In many applications of the linear regression model the experimenter can specify reasonable constraints on the predictor variables in terms of a lower and an upper bound for each of the predictor variables. LJZ (2005) construct a confidence band of the form

$$\mathbf{x}^T\mathbf{b} \in \mathbf{x}^T\hat{\mathbf{b}} \pm c\hat{\sigma}\sqrt{\mathbf{x}^T(X^T X)^{-1}\mathbf{x}} \quad \text{for all } (x_1, \cdots, x_p)^T \in \mathcal{X}, \tag{3.2}$$

where $\mathcal{X}$ is a rectangular region given by

$$\mathcal{X} = \{(x_1, \cdots, x_p)^T : \ a_i \leq x_i \leq b_i, i = 1, \cdots, p\}.$$

Here $-\infty \le a_i < b_i \le \infty, i = 1, \cdots, p$ are given constants. The main task is to determine the critical constant c in (3.2) so that the confidence band has a confidence level equal to $1 - \alpha$. The confidence level of the band in (3.2) is given by $P\{T < c\}$, where

$$T = \sup_{x_i \in [a_i, b_i], i=1,\cdots,p} \frac{|\mathbf{x}^T(\hat{\mathbf{b}} - \mathbf{b})|}{\hat{\sigma}\sqrt{\mathbf{x}^T(X^TX)^{-1}\mathbf{x}}}. \tag{3.3}$$

The distribution of T does not depend on $\mathbf{b}$ or σ, however, it does depend on the bounds (a_i, b_i) and the design matrix X in a complicated manner. The latter makes it difficult to derive a formula for the distribution of T in this general setting. LJZ (2005) provide a method to determine c via simulating T. **SimReg** uses the active set method described by LJZ (2005) to compute T. A random pivot T can be simulated efficiently, and thus c can be determined as accurately as one wishes by simulating a sufficiently large number of T's.

A simultaneous confidence band provides information on the whereabouts of the true regression model $\mathbf{x}^T\mathbf{b}$. Any regression model that is contained in the simultaneous confidence band for all $\mathbf{x} \in \mathcal{X}$ is deemed by the confidence band as a plausible candidate for the true model. Of course the true model $\mathbf{x}^T\mathbf{b}$ is included in the confidence band with $1 - \alpha$ probability.

4. Algorithms

The main computing task in the multiple comparison and the simultaneous confidence band problems described in Sections 2 and 3 is that of solving the optimization problems (2.2) and (2.3). In (2.2), the optimization over Λ will be performed by looking at the corresponding objective value for each pair in Λ and obviously choosing the one with the highest value. The sup of the quantity shown in (2.2) over $x_l \in [a_l, b_l]$ for $l = 1, \cdots, p$ is less trivial however. LJZ (2004) show that the T in (2.2) is equivalent to

$$\sup_{(i,j)\in\Lambda} \sup_{x_l \in [a_l, b_l], l=1,\cdots,p} \frac{|(P_{ij}\mathbf{x})^T\mathbf{Z}_{ij}|}{(\hat{\sigma}/\sigma)\sqrt{(P_{ij}\mathbf{x})^T(P_{ij}\mathbf{x})}} = \sup_{(i,j)\in\Lambda} Q_{ij} \frac{\parallel \mathbf{Z}_{ij} \parallel}{(\hat{\sigma}/\sigma)}, \tag{4.1}$$

where

$$Q_{ij} = \sup_{x_l \in [a_l, b_l], l=1,\cdots,p} \frac{|(P_{ij}\mathbf{x})^T\mathbf{Z}_{ij}|}{\parallel P_{ij}\mathbf{x} \parallel \parallel \mathbf{Z}_{ij} \parallel},$$

P_{ij} is the matrix square root of Δ_{ij}, and $\mathbf{Z}_{ij} = P_{ij}^{-1}(\mathbf{Z}_i - \mathbf{Z}_j), 1 \le i \ne j \le k$ with $\mathbf{Z}_i \sim N(\mathbf{0}, (X_i^TX_i)^{-1})$. Similarly, LJZ (2005) show that the problem (3.3) is equivalent to

$$T = \sup_{x_i \in [a_i, b_i], i=1,\cdots,p} \frac{|(P\mathbf{x})^T\mathbf{Z}|}{(\hat{\sigma}/\sigma)\sqrt{(P\mathbf{x})^T(P\mathbf{x})}} = Q \frac{\parallel \mathbf{Z} \parallel}{(\hat{\sigma}/\sigma)}, \tag{4.2}$$

where

$$Q = \sup_{x_i \in [a_i, b_i], i=1,\cdots,p} \frac{|(P\mathbf{x})^T\mathbf{Z}|}{\parallel P\mathbf{x} \parallel \parallel \mathbf{Z} \parallel},$$

P is the square root of $(X'X)^{-1}$, and $\mathbf{Z} \sim N(0, I_{p+1})$. Thus to simulate a realization of T for the multiple comparison problem a $\mathbf{Z}_{ij}$ is generated, Q_{ij} is computed, and then (4.1) is used to obtain T. This process is similar for the simultaneous confidence band problem where a $\mathbf{Z}$

is generated, Q is computed and then (4.2) is utilized. The overall process is then to generate a large number of copies of T and use the $100(1-\alpha)\%$ quantile of the generated values to approximate c.

The problem of obtaining Q_{ij} and Q in (4.1) and (4.2) is mathematically equivalent. So hereafter we will refer to both problems as that of obtaining Q. It turns out that the most time consuming and difficult part of the simulation process described above is that of computing Q. LJZ (2004) took a general approach of applying a smooth optimization technique to obtain Q. As they point out, a shortcoming of their method is that it can lead to local maxima that are not the global solution to the maximization problem. LJZ (2005) provide two other approaches, one which is a branching method, and another which is an active set method. The former is appropriate for cases where the number of predictors is small, say two or three, and the latter works best overall. It can be shown that Q can be obtained by solving a problem of the form

$$Q = \sup_{\mathbf{s}\in\Omega} \frac{|\mathbf{s}^T\mathbf{Z}|}{\| \mathbf{s} \| \| \mathbf{N} \|}, \tag{4.3}$$

where $\Omega = \{\mathbf{s} : \mathbf{s} = \gamma\mathbf{v},\ \mathbf{v} \in L,\ \gamma > 0\}$ and $L = \{P\mathbf{x} :\ x_i \in [a_i, b_i], i = 1, \cdots, p\}$. Furthermore, it can be shown that if $\hat{\mathbf{s}} \in \Omega$ is the solution to

$$\inf_{\mathbf{s}\in\Omega} \|\mathbf{s} - N\|^2, \tag{4.4}$$

then $\hat{\mathbf{s}}$ is also the solution to (4.3). This is a quadratic programming problem for which LJZ (2005) suggest using an active set method. **SimReg** has adopted the algorithm described in LJZ (2005) for both the multiple comparison and the simultaneous confidence band problems.

5. The SimReg software

SimReg was developed using MATLAB Version 7.0.1, Release 14 on Windows XP. It can, however, be easily adopted in other platforms. **SimReg** performs two main tasks of performing multiple comparison of several regression lines and obtaining simultaneous confidence bands for a multiple regression model.

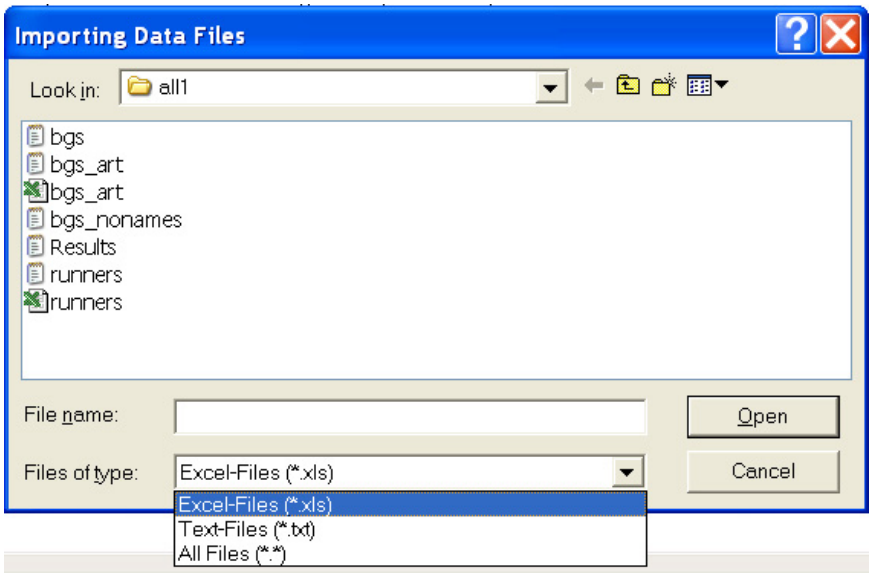
What sets **SimReg** apart from the currently existing software is that it performs these tasks with the option of allowing the user to set natural boundaries (constraints) for the predictors. It is clear that a linear regression model often does not hold for the whole range of $(-\infty, \infty)$ range of the predictors, and it is often the case that the predictors have reasonable finite range. Imposing the constraints will result in simultaneous confidence bands that are narrower (see, LJZ 2005), and results in more reliable multiple comparisons. Additionally, for the multiple comparison problem, it allows any type of pair comparison desired by the user.

The software includes a user interface for importing data and setting the desired parameters. It produces output that can be used for inference, and includes graphs for for models with one or two predictors. Below we explain how **SimReg** can be used and what its capabilities are.

5.1. Getting started and data input

SimReg is archived in the file "Simreg.rar" using WinRAR 3.40. Use the following three steps to get started with the **SimReg** program:

Figure 1: **SimReg** window to import data files



1. Unpack **SimReg** in any desired directory. To unpack, you can download WinRAR from http://www.rarsoft.com/download.htm. As an example, suppose that you install the software in the directory "`C:\Program Files\Simreg`".

2. Start MATLAB and add the path of the directory where you installed the software to your current MATLAB path. This can be done using the "`addpath`" command in MATLAB. For example, at the MATLAB command prompt you type "`addpath 'C:\Program Files\Simreg' `".

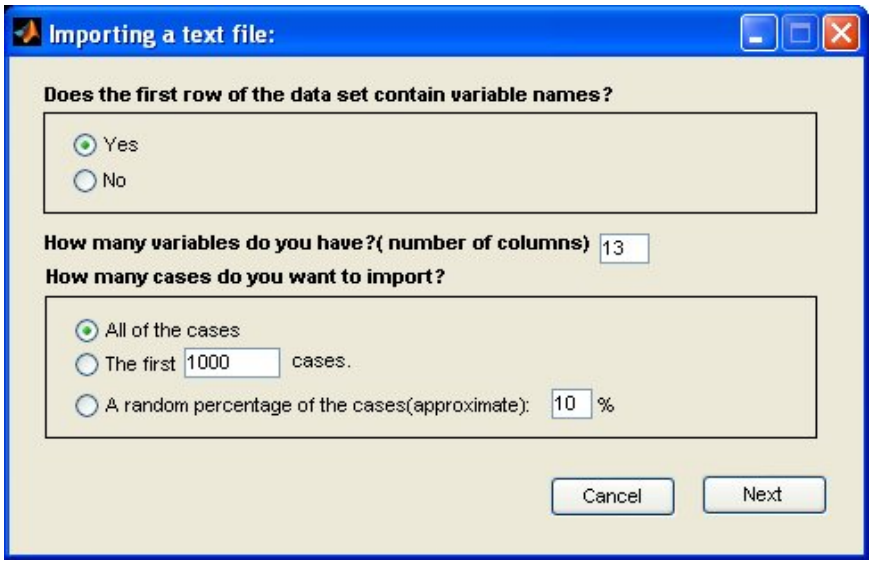
3. At the MATLAB command prompt, type `simreg`.

When the command `simreg` is issued at the MATLAB command prompt, a window (shown in Figure 1) pops up where the name of the data file to be analyzed should be entered. **SimReg** can import Excel files[1] with the extension "`xls`"or data in ASCII format saved in files with the extension "`txt`". In either case rows should represent cases and columns represent variables. The first row of the data file can include variable names, or it could be the first case observed. Note that in this version of **SimReg** no missing data handling capability is available, therefore the user should remove all cases with missing data before uploading the data file.

To illustrate the ideas we use, and have included, a modified version of the data used in the Berkeley Guidance Study published by Tuddenham and Snyder (1954). The study was a longitudinal monitoring of 136 boys and girls born in Berkeley, California between January 1928 and June 1929, and followed for at least 18 years. Our modification to this data is addition of an artificial categorical variable "Race" with the aim of including a categorical variable which has more than two levels. One of the races "White", "Black", "Asian", or "Other" was arbitrarily assigned to each case. The modified data set includes a total of 13

[1]Excel files can be imported if version 7.x of Matlab is used. This option is not available if version 6.x of Matlab is used.

Figure 2: Options for importing a txt file



variables. The file "`bgs_art.txt`" is the ASCII version of this data set, and "`bgs_art.xls`" is the Excel version, both included in the **SimReg** package.

5.2. Reading the data

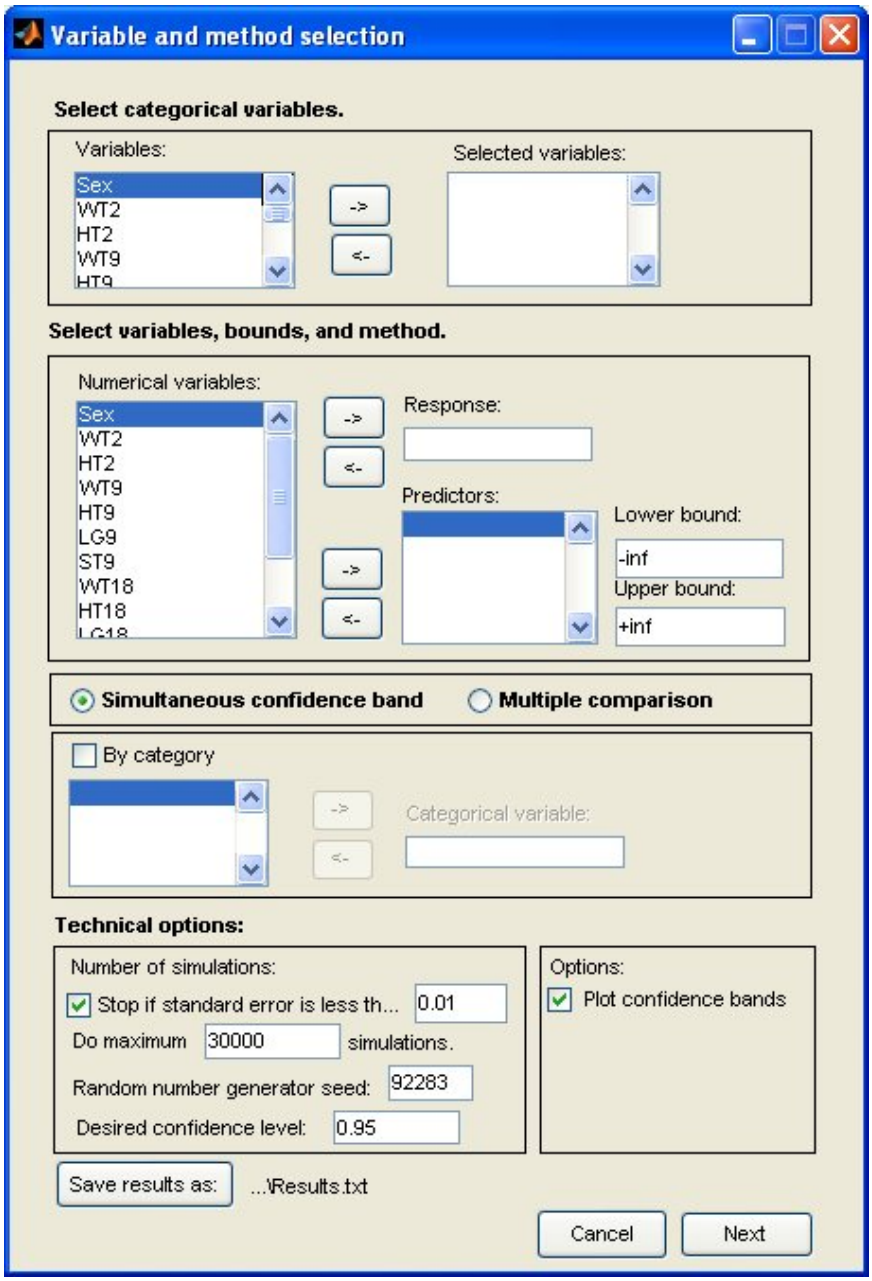
Reading a txt file

If we select a "`txt`" file (e.g., "`bgs_art.txt`") in the *Importing Data Files* menu and click on open, then a second menu titled "*Importing a Text File*" will pop up (see Figure 2). On this menu you must input information such as whether the first row of the data includes variable names, and the number of variables included in your data set (this must be equal to the number of columns in your data set). Additionally you have the option of selecting a subset of data for analysis. In particular, you can use all, a subset of first few cases, or a randomly selected subset of the data. If the first row of your data file does not include variable names, then the program assigns variables names "`Var#`" where "#" is the column number. Once "`Next`"is pressed we get a menu titled "*Variable and method selection*", shown in Figure 3.

Reading an Excel file

As noted earlier, Excel files can be read only in Version 7.x of Matlab and this option is not available in lower versions. To read an Excel file, in the *Importing Data Files* menu choose the option of '*.xls" and select your input file and press open. This will lead directly to the menu titled '*Variable and method selection*", shown in Figure 3. Note that when importing an Excel file, the first row of the data must contain variable names. Additionally, the menu "*Importing a Text File*" and the options in that menu for subseting the data are not available, and all the data in the file are used for analysis.

Figure 3: Menu to select regression variables and the method of analysis



5.3. Analysis of the data

The menu "*Variable and method selection*", shown in Figure 3, is used to specify the model and select one of the options of "simultaneous confidence band" or "multiple comparison". The first step, however, consists of specifying the categorical variables in the study, if any. Initially, the menus on the right hand side of the first and the second panels of the menu in Figure 3 consist of all the names of the variables read. On the top panel, the user selects the name of the categorical variables. If multiple comparison is to be performed, obviously, at least one categorical variable must exist. For our example, two categorical variables of "Sex" and "Race" were selected, as shown on Figure 4. Once the categorical variables are selected, then these variables are added to the fourth panel where the option "*By category*" appears and they are removed from the second panel titled "*Numerical variables*". As shown on Figure 4, for our example the variables "Sex" and "Race" were removed from the middle panel, and added to the fourth panel.

The next step is to specify the regression model. The response and the predictor variables for the model are selected in the second panel. In the example shown "`HT18`" is selected as response and "`WT9`" and "`HT9`" are selected as predictors. At this stage lower and upper bounds for the predictors can be set. The default is $(-\infty, \infty)$, which is written as "`-inf`" and "`+inf`". For our example we have set the bound on `WT9` to $(10, 60)$ and that for `Ht9` to $(120, 160)$.

On the third panel of the menu shown on the Figure 4, the user will choose the type of analysis to be performed. As mentioned earlier, **SimReg** provides two methods of simultaneous confidence bands and multiple comparison.

The confidence band option

When the "Simultaneous confidence band" option in the third panel of the "*Variable and method selection*" menu is selected, then the critical value c for the simultaneous confidence band (3.2) is computed. The user has the option of performing this analysis for each level of a selected categorical variable, by choosing the "By category" option and selecting the desired categorical variable. If no category is selected, then the analysis will be performed for the whole data set, ignoring categories. If a category is selected, then a separate simultaneous confidence band will be given for each of the levels of the selected category.

As shown in Figure 4, for our example, we request a simultaneous confidence band for the regression model for each level of the category "Sex". When the button "Next" is pressed, then the software runs a simulation and when the computation is done a menu pops up which states "please see Result.txt for the result." If the plot option is selected (available only if $p = 1, 2$), then plots of the confidence bands and the regression plane (line) will also be given.

Table 1 shows the content of the file "Results.txt" for our example. In general, the output consists of summary statistics, regression coefficients and standard T-test and p-values, and the residual mean square for all observed values or for each level of the selected categorical variable, if one was selected. Most importantly, it gives the critical value c in (3.2) for the confidence band. Finally, if the plot option is selected, a plot of confidence bands, the regression line (or plane) and the observed data is produced for all the data or each category level, if a categorical variable is selected. For our example, these are given for the two levels of the variable "Sex", namely "Female" and 'Male". Additionally, the number of simulations performed and the lower and upper bounds for the predictors are echoed. For our example,

Figure 4: Categorical, response, predictors, and their bounds selected.

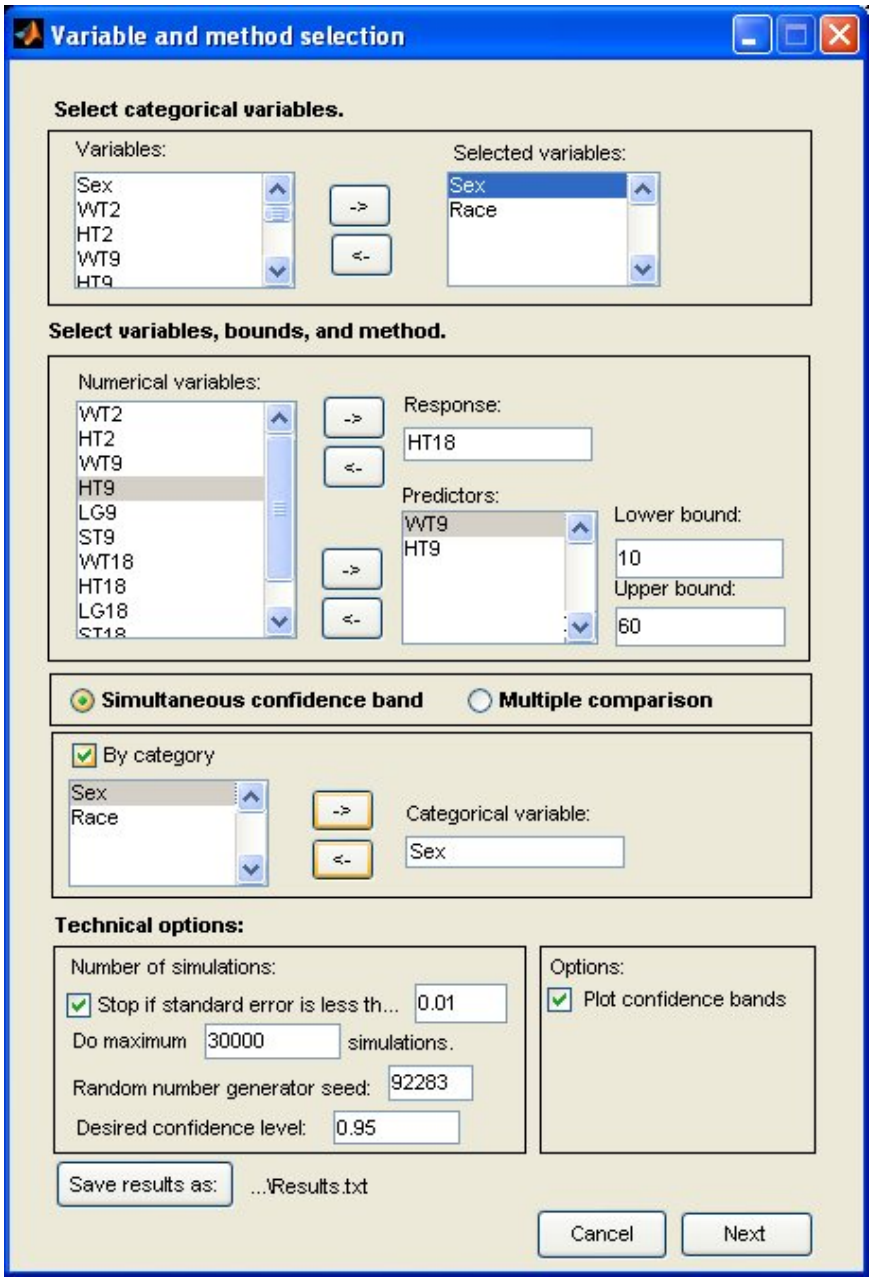


Table 1: The output for the simultaneous confidence band example

```
Summary statistics and results of the analysis
==============================================

Summary statistics:
All observed values:

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18    136    153.600     172.579       172.500       8.844         195.100
      WT9    136     19.900      31.626        30.900       5.969          66.800
      HT9    136    121.400     135.493       135.700       5.496         152.500

======================================================================

Summary statistics:
 Group:      Female

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     70    153.600     166.544       166.750       6.075         183.200
      WT9     70     22.000      31.621        30.650       5.824          47.400
      HT9     70    121.400     135.120       135.700       5.613         152.500

======================================================================

 Regression Analysis
 Group:      Female

 Response:       HT18
Residual Mean Square =  11.0927

Variable         Coeff          s.e.          T-stat         p-value
 Constant      23.1721        11.9682          1.936         0.0285
      WT9      -0.3585         0.1004         -3.572         0.0003
      HT9       1.1450         0.1041         10.995         0.0000

 ================================================================
Group:    Female

 Number of simulations  =    30000

The lower and upper bounds imposed on the predictors  are:
      WT9    Lower Bound = 10.000     Upper Bound = 60.000
      HT9    Lower Bound = 120.000     Upper Bound = 160.000

 The simulated critical value for 0.950 confidence level =  2.8768
 The standard error  =  0.0096

 ================================================================

Summary statistics:
 Group:        Male

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     66    160.900     178.979       178.900       6.517         195.100
      WT9     66     19.900      31.632        31.000       6.164          66.800
      HT9     66    122.000     135.889       135.600       5.385         147.500

======================================================================

 Regression Analysis
 Group:        Male

 Response:       HT18
Residual Mean Square =   9.3265

Variable         Coeff          s.e.          T-stat         p-value
 Constant      20.7883        11.0985          1.873         0.0328
      WT9      -0.2055         0.0809         -2.540         0.0068
      HT9       1.2120         0.0926         13.087         0.0000

 ================================================================
Group:      Male

 Number of simulations  =    30000

The lower and upper bounds imposed on the predictors  are:
      WT9    Lower Bound = 10.000     Upper Bound = 60.000
      HT9    Lower Bound = 120.000     Upper Bound = 160.000

 The simulated critical value for 0.950 confidence level =  2.8640
 The standard error  =  0.0116

 ================================================================

 The pooled variance is  10.2368

 The degrees of freedom is    130
======================================================================
```

the critical value c for females is 2.8768 with standard error .0096, and that for the "Male" is 2.8640 with standard error .0116. Finally, Figures 5 shows the 95% confidence bands, and regression plane superimposed by a scatter of the observed data for each of the levels of the variable Sex for our example. These plots can be navigated by the tools available in MATLAB.

The multiple comparison option

To perform multiple comparison, this option must be selected in the "*Variable and method selection*" menu and the user must choose a categorical variable, using the "*By category*" option, whose levels will be compared. To give an example, we will perform a multiple comparison of the same regression model that we specified in the previous subsection, but using the levels of the categorical variable "`Race`". Figure 6 shows the relevant specifications. In this case when the button "next" is pressed, then the "*Pair comparison selection*" menu, shown in Figure 7, pops up.

The "*Pair comparison selection*" menu allows the user to provide the set Λ, defined in Section 2. The user has the option of selecting (comparing) either all pair comparison in the top panel, comparison to a specific control level in the middle panel, or any arbitrary pairs on the bottom panel. In the example shown in Figure 7, comparison to a control is selected, and specifically we are making comparison of the level `White` to all other levels of the variable `Race`. Once again, when the computation is complete, a window pops-up which states "Please see Results.txt for the result". Additionally, if the plot option is on, plots will be generated as well.

The content of the file `Results.txt` for our example is given in Tables 2 and 3. This output in general consists of summary statistics for the observed variables, and result of the fitted regression model for each levels of selected categorical variable. More importantly, it provides the critical value c in Section 2.2, the statistics T in (2.2), and its corresponding p-value for the multiple comparison specified. For our example, the critical value is $c = 3.1947$ with the standard error of 0.0107, and the test statistics is $T = 11.917$ with the p-value nearly zero, indicating a significant difference between the pairs.

In addition to multiple comparison, the test statistics and p-values for single pair comparison of the pairs specified in the "*Pair comparison selection*" menu (Λ)is given. For our example, single pair comparisons indicate a significant difference between the pairs (Asian, White) and (Other, White), but not a significant difference between the pair (Black, White). Corresponding to each of these tests of significance of difference of pairs, the program generates plots that can be used for this purpose. These plots are more informative than simply the p-values (see Liu, Jamshidian, Zhang, Bertz, and Han (2005b)). Specifically, the plots generated are simultaneous confidence bands for pair differences; that is, they test the equality of the regression model between single pairs of categorical levels selected. If the bands cross the zero line (or plane), then the hypothesis of equality of regression model for the pair indicated on the plot is rejected. Figures 8 and 9 shows a snapshot of the plots just mentioned for our example. As expected (from the p-values observed earlier), the bands for the pairs (Asian, White) and (Other, White) intersect the zero plane, but that for the pair (Black, White) does not intersect the zero plane.

Technical options

The "*Variable and method selection*" menu, shown in Figure 3, consists of a "*Technical options*"

Figure 5: The observed data, simultaneous confidence bands, and the regression plane for Females and Males

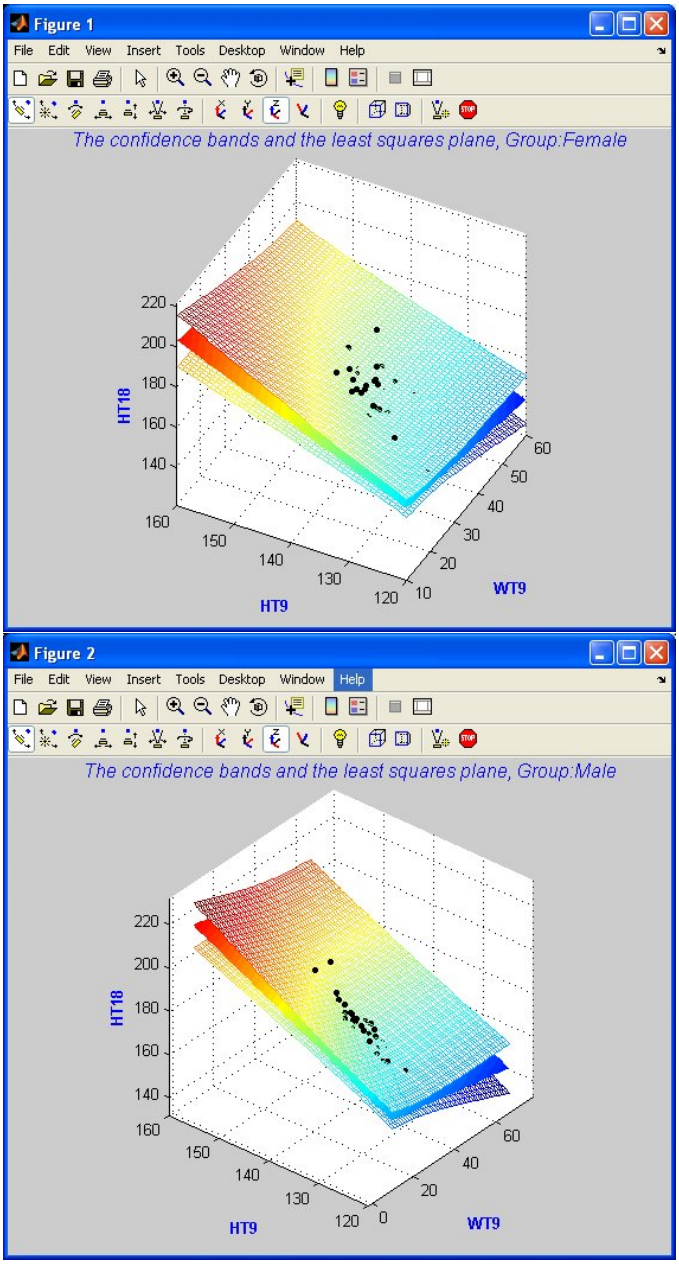


Figure 6: A snapshot of the "variable and method selection menu", specifying multiple comparison

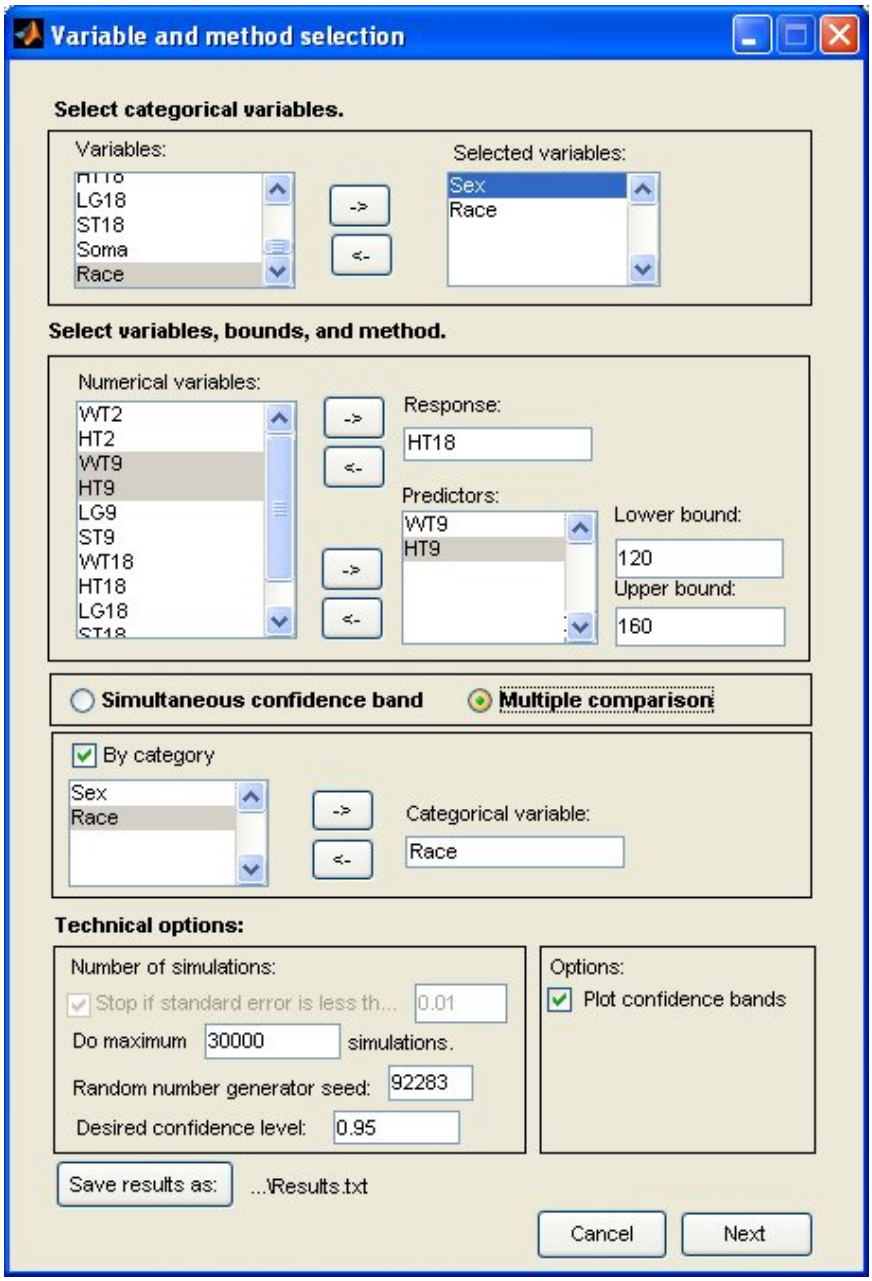


Figure 7: A snapshot of the "Pair Comparison Selection" menu

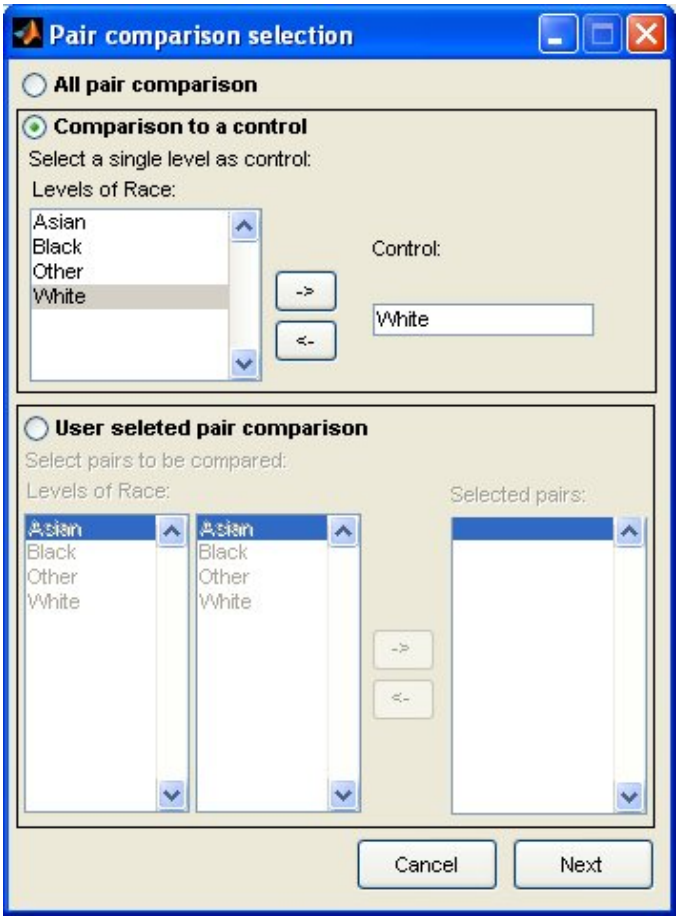


Table 2: Output for the multiple comparison of `White` with other levels of the variable `Race`.

```
Summary statistics and results of the analysis
==============================================

Summary statistics:
 Group:       Asian

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     45    154.600     168.804       168.100       7.875         188.000
      WT9     45     22.200      32.364        30.800       7.906          66.800
      HT9     45    123.200     135.556       135.800       6.010         152.500

=====================================================================

 Regression Analysis
 Group:       Asian

 Response:       HT18
Residual Mean Square =  29.0842

Variable         Coeff          s.e.          T-stat         p-value
 Constant      33.0710        23.9793          1.379         0.0876
      WT9      -0.0381         0.1539         -0.248         0.4027
      HT9       1.0104         0.2024          4.992         0.0000

 ===============================================================

Summary statistics:
 Group:       Black

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     24    160.900     178.096       177.850       6.945         194.300
      WT9     24     19.900      30.642        30.900       4.102          43.200
      HT9     24    122.000     135.129       134.300       5.218         147.400

=====================================================================

 Regression Analysis
 Group:       Black

 Response:       HT18
Residual Mean Square =   8.0487

Variable         Coeff          s.e.          T-stat         p-value
 Constant       8.1853        16.1141          0.508         0.3084
      WT9      -0.1775         0.1737         -1.022         0.1592
      HT9       1.2976         0.1366          9.502         0.0000

 ===============================================================

Summary statistics:
 Group:       Other

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     34    153.600     166.706       166.750       5.790         177.500
      WT9     34     22.000      31.538        30.900       5.485          42.400
      HT9     34    121.400     135.100       135.650       5.387         144.800

=====================================================================

 Regression Analysis
 Group:       Other

 Response:       HT18
Residual Mean Square =  10.5821

Variable         Coeff          s.e.          T-stat         p-value
 Constant      24.7002        16.8182          1.469         0.0760
      WT9      -0.4657         0.1424         -3.271         0.0013
      HT9       1.1598         0.1449          8.002         0.0000

 ===============================================================

Summary statistics:
 Group:       White

 Variable      N      Min         Mean         Median      Std. Dev.        Max
     HT18     33    167.000     179.764       180.200       6.538         195.100
      WT9     33     24.400      31.427        31.000       4.439          43.100
      HT9     33    125.400     136.079       136.000       5.259         146.000

=====================================================================
```

Table 3: Table 2 (Continued)

```
 Regression Analysis
 Group:       White

 Response:       HT18
Residual Mean Square =   9.6432

Variable         Coeff          s.e.          T-stat         p-value
 Constant      12.4400        16.8338          0.739         0.2328
      WT9      -0.4061         0.1772         -2.291         0.0146
      HT9       1.3234         0.1496          8.845         0.0000

 ================================================================

 The pooled variance is  16.1927

 The degrees of freedom is    124
=====================================================================
 Number of simulations  =    30000

The lower and upper bounds imposed on the predictors  are:
      WT9    Lower Bound = 10.000     Upper Bound = 60.000
      HT9    Lower Bound = 120.000     Upper Bound = 160.000

 The simulated critical value for 0.950 confidence level =  3.1947
 The standard error  =  0.0107

 ================================================================
 Observed statistics and p-values for single pair comparison:
----------------------------------------------------------------
Pair {    Asian,    White} Observed Statistics = 11.0953    p-value = 0.0000
Pair {    Black,    White} Observed Statistics =  1.0176    p-value = 1.0000
Pair {    Other,    White} Observed Statistics = 11.9170    p-value = 0.0000


 Observed statistics and p-values for all selected pairs comparison:
---------------------------------------------------------------------
Observed Statistics = 11.9170    p-value = 0.0000

 ================================================================
```

Figure 8: Single pair comparison plots

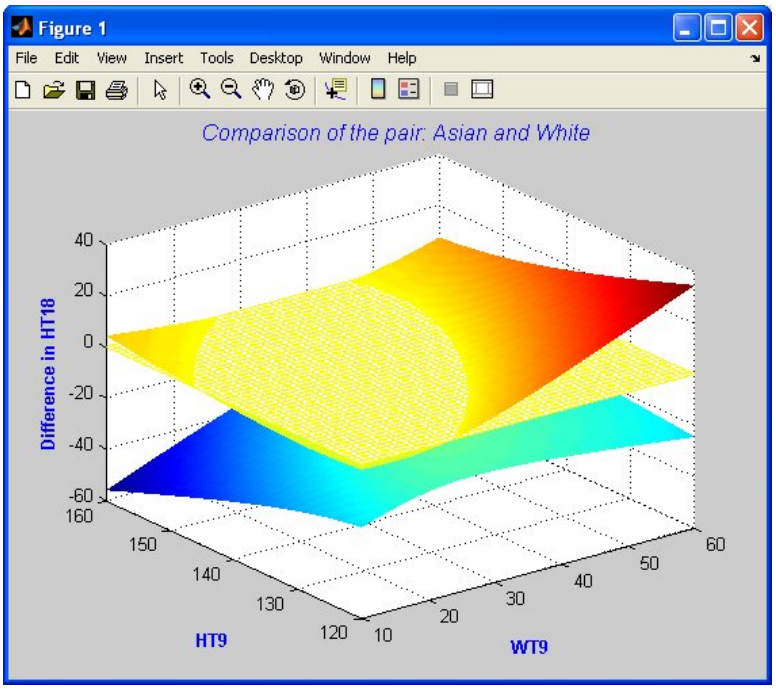
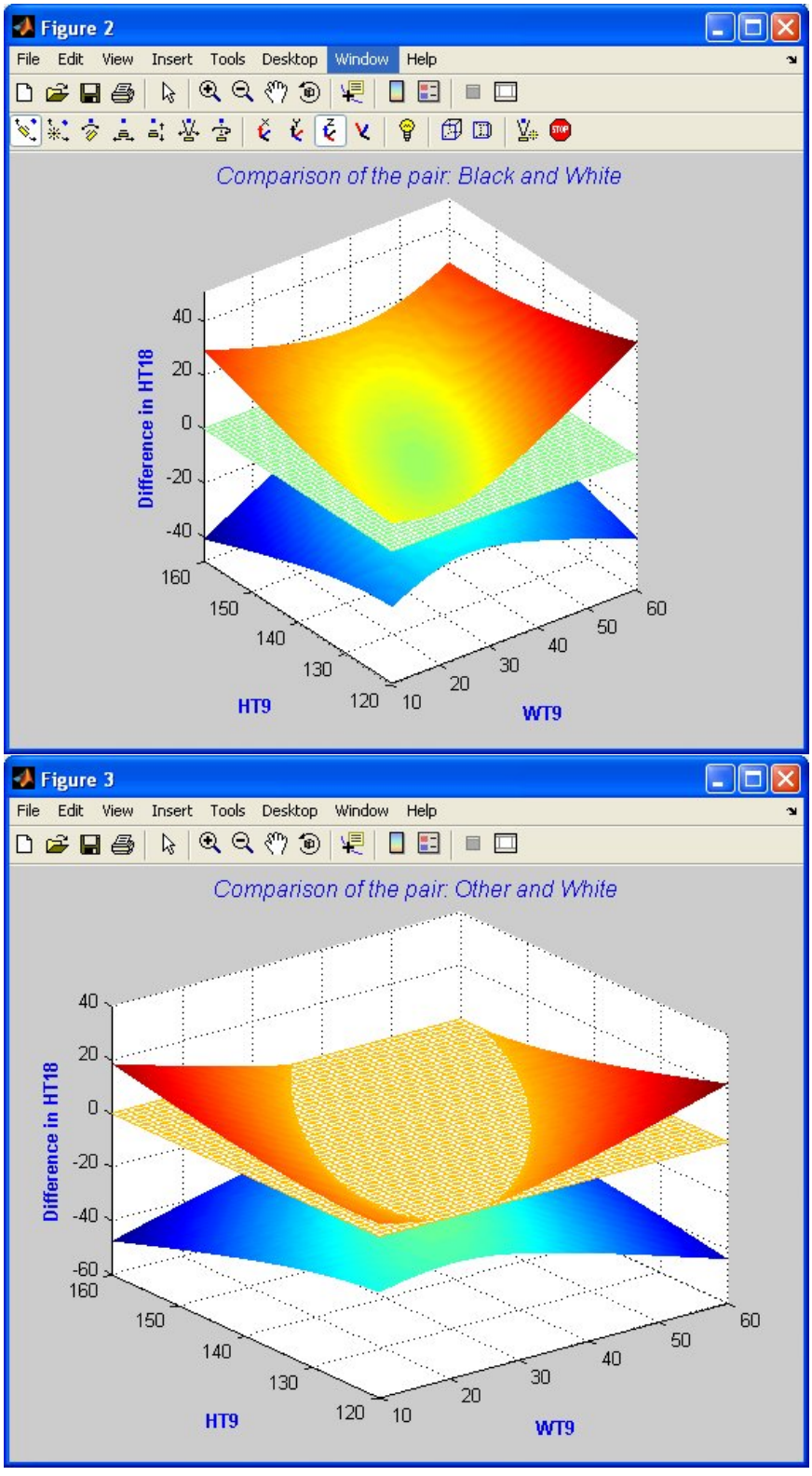


Figure 9: Single pair comparison plots



section, where some technical parameters for the simulation process can be set. The user can specify the maximum number of simulations desired (see LJZ 2004 and 2005 for guidelines), or the simulation process can optionally be stopped by setting a threshold for the tolerance level of the standard error of the estimated critical value. The standard error is computed using the methods given in LJZ (2005). The standard error option is not available when multiple comparison is selected (see Figure 6). Also in this panel a seed for the random number generator can be set, and the desired confidence level, a value between 0.5 and 1, can be input. If the number of predictors is less than or equal to 2 , then there is a plot option, which is on by default, and results in the plots explained above. Finally, the button "Save results as" allows the user to input a filename where the results of the analysis will be stored in. The default filename for the results is "Results.txt" and it is stored in the current directory. This file is a txt file, and it can be best viewed either using the MATLAB editor or the Notepad. If Notepad is used, it would be best to set the font to "Courier New"with 12 point font size.

5.4. Access to a few useful variables

The user may be interested to perform further analyses after running the software. **SimReg** provides two files `useful_vars_scb.mat` and `useful_vars_mc.mat` that contain a few useful variables for this purpose. The file `useful_vars_scb.mat` is created when the simultaneous confidence band option is selected, and the `useful_vars_mc.mat` is created when the multiple comparison option is run. These files are stored in the current directory, where the user is running **SimReg**. The MATLAB's `load` *`filename`* command can be used to upload the content into the workspace area.

The following variables are stored in the file `useful_vars_scb.mat`, if no categorical variable is selected:

`B` This matrix contains $(X^T X)^{-1}$.

`RMS` The residual mean square from the regression analysis.

`se` Standard error of c.

`beta_ls` Regression coefficient from the regression analysis.

`c` The critical constant for the simultaneous confidence band.

`n` Number of cases.

`p` Number of predictors.

`pred` The predicted values from the regression model.

`Predictor_names` A vector of strings containing predictor names.

`resid` The residual values from the regression model.

`response_name` A string containing the name of the response variable.

`se` Standard error of c.

`y` A vector containing all the observations on the response variable.

The following variables are stored in the file `useful_vars_scb.mat`, if a categorical variable is selected:

`B_all` This matrix contains $(X^TX)^{-1}$ for all levels of the selected categorical variable. If there are k levels, then the size of the matrix has $k \times (p+1)$ rows and $(p+1)$ columns.

`RMS_all` The residual mean square from the regression analysis for each of the levels of the selected categorical variable.

`c_all` This is a vector containing the critical constants for the simultaneous confidence bands for each of the levels of the selected categorical variables.

`n_all` The number of cases in each of the levels of the selected categorical variable.

`Predictor_names` A vector of strings containing predictor names.

`se_all` This is a vector containing the standard error for the critical constants for the simultaneous confidence bands for each of the levels of the selected categorical variables.

The following variables are stored in the file `useful_vars_mc.mat`:

`B_all` This matrix contains $(X_i^TX_i)^{-1}$ for all levels of the selected categorical variable. If there are k levels, then the size of the matrix has $k \times (p+1)$ rows and $(p+1)$ columns.

`Bij_all` This matrix contains $(X_i^TX_i)^{-1} + (X_j^TX_j)^{-1}$ for each pair (i,j) in a row of Λ in order. If Λ has ℓ rows, then `Bij_all` consists of $\ell * (p+1)$ rows and $p+1$ columns.

`lambda` A two-column matrix with rows consisting of the pairs of levels of the categorical variable selected for multiple comparison.

`RMS_all` The residual mean square from the regression analysis for each of the levels of the selected categorical variable.

`beta_all` The regression coefficient from the regression analysis for each of the levels of the selected categorical variable.

`c` The critical value for multiple comparison.

`lb` A vector containing the lower-bound restrictions on predictors.

`n_all` The number of cases in each of the levels of the selected categorical variable.

`Predictor_names` A vector of strings containing predictor names.

`response_name` A string containing the name of the response variable.

`rnd_seed` The seed used for random number generation.

`se` Standard error of c.

`sighat` The pooled standard deviation $\hat{\sigma}$.

`tmax` A vector containing all simulated values of T.

`ub` A vector containing the upper-bound restrictions on predictors.

5.5. Future improvements of the software and copyright

The following includes a list of improvements that are planned to be made in **SimReg** in the future versions:

- Handle missing data.
- Add a context sensitive help to each of the menus.
- Add a "back" button to each menu.
- Provide html output.
- Provide one sided confidence bands.

Acknowledgements

Mortaza Jamshidian's research was supported in part by the National Science Foundation Grant DMS-0437258. The authors would like to thank the Associate Editor for suggestions that led to significant improvements of the user interface and this paper.

References

Casella G, Strawderman WE (1980). "Confidence Bands for Linear Regression with Restricted Predictor Variables." *Journal of the American Statistical Association*, **75**, 862–868.

Hochberg Y, Tamhane AC (1987). *Multiple Comparison Procedures*. John Wiley.

Hsu JC (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall.

Liu W, Jamshidian M, Zhang Y (2004). "Multiple Comparison of Several Linear Regression Lines." *Journal of the American Statistical Association*, **99**, 395–403.

Liu W, Jamshidian M, Zhang Y (2005a). "Exact Simultaneous Confidence Bands in Multiple Linear Regression with Predictor Variables Constrained in Intervals." *Journal of Computational and Graphical Statistics*. Forthcoming.

Liu W, Jamshidian M, Zhang Y, Bertz F, Han X (2005b). "Some New Methods for the Comparison of Two Linear Regression Models." Submitted.

Miller RG (1981). *Simultaneous Statistical Inference*. Springer-Verlag.

Naiman DQ (1987). "Simultaneous Confidence Bounds for Multiple Regression Functions Using Predictor Variable Constraints." *Journal of the American Statistical Association*, **82**, 214–19.

Scheffé H (1953). "A Method for Judging all Contrasts in the Analysis of Variance,." *Biometrika*, **40**, 87–104.

Spurrier JD (1999). "Exact Confidence Bounds for All Contrasts of Three or More Regression Lines." *Journal of the American Statistical Association*, **94**, 483–88.

Tuddenham RD, Snyder MM (1954). "Physical Growth of California Boys and Girls from Birth to Age 18." *California Publications on Child Development*, **1**, 183–364.

Working H, Hotelling H (1929). "Applications of the Theory of Error to the Interpretation of Trends." *Journal of the American Statistical Association*, **24**, 73–85.

Affiliation:

Mortaza Jamshidian
Department of Mathematics
California State University
800 N. State College Blvd.
Fullerton, CA 92834, United States of America
E-mail: mori@fullerton.edu
URL: http://math.fullerton.edu/mori

Journal of Statistical Software
January 2005, Volume 12, Issue 2.
http://www.jstatsoft.org/

Submitted: 2004-09-24
Accepted: 2005-01-07