RegGen: A Program for Generating Regression Data Sets

Version 1.2

Jeff Miller
Department of Psychology
University of Otago
Dunedin, New Zealand

November, 2004

Here are my contact details:
Prof Jeff Miller
Department of Psychology
Univ of Otago
Dunedin, New Zealand
email address: miller@psy.otago.ac.nz

Contents

1 Introduction

This program was designed for use in teaching the statistical procedure known as *Regression Analysis*. It generates data sets for use as examples or practice problems.

The user specifies the number of cases (i.e., sample size), the number of variables per case, the mean and standard deviation of each variable, and the matrix of correlations between variables. The program then generates set of data satisfying these conditions exactly (up to some rounding error). The data are then written to a file for subsequent analysis by a statistical package.

A critical feature of RegGen is that the generated data satisfy the specified conditions almost exactly. For example, if you specify that a certain variable should have a mean of 100 and an SD of 10, the sample will have exactly that mean and SD, up to some small rounding error associated with the number of decimal places in the output. Thus, you specify the sample characteristics directly rather than specifying the underlying population values from which random samples are taken.

2 Installation

Copy the program RegGen.exe to any directory in your path.

3 Running RegGen Interactively

To start RegGen for an interactive run, simply type "reggen" at a command prompt.

You will see a brief warning message suggesting that it is better to use a parameter file. Ignore that message for now. The reason for it is explained in the next section.

The program will now ask you to specify the parameters of the desired data set, as in the following example.

```
Enter the desired number of cases : 20
Enter the desired number of variables : 2
Enter the desired correlation of var 1 with var 2 : 0.3
Variable 1: Desired mean : 100
Variable 1: Desired std dev : 10
Variable 1: Desired number of decimal places : 0
Variable 2: Desired mean : 100
Variable 2: Desired std dev : 10
Variable 2: Desired number of decimal places : 0
```

The user types in only the numbers at the far right of each line. I hope that most of this is self-explanatory. The only point that may deserve comment is the "number of decimal places." This option allows the user to specify whether the generated data values are to be whole numbers (0 decimal places), or to contain fractional parts measured to 1, 2, 3, etc decimal places.

After you have specified the parameters, RegGen will try to generate the data set. It will fail if the requested correlation matrix is impossible (as discussed further in the next section). In that case it halts with this message:

```
Requested correlation matrix is impossible.
```

If RegGen succeeds in generating the data set, it next allows you to display and write out the data set is has generated, as described in section 5.

4 Running RegGen With A Parameter File

The parameters of the data set can be specified in an ASCII file instead of interactively. This is useful when you want to create several data sets that differ in only a few parameters: For example, you might want to generate several 5-variable data sets varying only the correlation of variables 3 and 4. It can also be useful even if you only want one data set: Sometimes, it is difficult to be sure whether the correlation matrix you want is possible or impossible. If you specify an impossible data set interactively, you have to start over from the beginning. If your parameters are specified in a file, you can just change one or two correlations in the file and rerun reggen.

The file RegGen.Smp is a sample parameters file. The comments to the right of each line are optional, but are included to enable you to figure out the format. The file can be created with any ASCII editor. Be sure to use spaces rather than tabs between numbers.

To use a parameters file, invoke RegGen with a command line parameter giving the name of the parameters file, like this:

```
C:> reggen reggen.smp
```

There is an additional option which allows you to specify the names of the variables for an output file in the MTab format, mentioned below. Within the input parameter file, you are allowed to put the name of each variable at the beginning of the line on which you specify its mean, sd, and number of decimal places, like this:

```
Height 100 10 0  { mean, sd, & number of decimal places for var 1 }
Weight 500 10 0  { mean, sd, & number of decimal places for var 2 }
Time   100 20 0  { mean, sd, & number of decimal places for var 3 }
Age      0  1 3  { mean, sd, & number of decimal places for var 4 }
```

Variable names can contain any nonblank characters, and they are limited to 20 characters.

5 Getting Output from RegGen

After RegGen has generated the desired data set, it will produce a display more or less like this:

```
RegGen: Program to generate regression data sets  Version 1.1:  March 1998
@Copyright 1998..2000    Jeff Miller              Dept of Psychology
                         Univ of Otago            Dunedin, New Zealand
N of cases = 10.  N of variables = 4

ACTION MENU:
 S = list data to Screen
 F = list data to File
 M = write Mtab data file
 N = make New data

 action or Quit :
```

To select one of these options, simply type the single capitalized letter. Here are descriptions of the individual options:

S List the data to the screen so that you can look at it.

F List the data to an output file, in just the same format as to the screen.

M Create an output file in the proper format to be read by the MTab program, a free DOS program for doing regression analysis and descriptive statistics.

N Generate a new data set. The parameters of the data set will be the same as the current one, you just get another example of a data set with these parameters.

Q Quit to DOS.

6 Fully Batch Operation

It is also possible to run RegGen in fully batch mode, without operator intervention. This may be desirable, for example, if you want to interface RegGen to a simulation program of some sort. To do this, you must of course first prepare the input parameters file as discussed above. Then, you can invoke RegGen with an additional parameter that is the name of the output file. For example,

```
C:> reggen reggen.smp reggen.out
```

would automatically write an output file called reggen.out without any user input at all. Several optional switches are available when you run RegGen. These are listed after the output file name in any order, like this:

```
C:> reggen reggen.smp reggen.out -m
```

The switches can be any of the following:

-m Indicates that the output file should have the MTab format.

-g Generate a random correlation matrix rather than using the one specified in the input file. In this case it does not matter what is on the correlation lines in the input file, but they must still be present. For example, the correct number of blank lines will suffice.

A possibly useful trick: By default, RegGen always asks for user confirmation before writing over an existing file. To avoid that confirmation check, use an exclamation point as the first character of the file name. If you do that, RegGen will (a) delete the exclamation point from the output file name, and (b) write the output to a file with the indicated name, *automatically overwriting the file if it already exists.*