Program for multiple linear regression (ordinary or through the origin) with permutation test – User's notes

Pierre Legendre
Département de sciences biologiques
Université de Montréal

March 2002

This program computes a multiple linear regression and performs tests of significance of the equation parameters using permutations. In this new version, the regression line can be forced through the origin. Permutation testing is recommended when the residuals of the regression equation are not normally distributed; it does not solve the problems caused by heteroscedasticity. Two permutation methods are available in the program:

(1) Permutation of the values of **y** (recommended in multiple regression through the origin).

(2) Permutation of residuals of the full regression model (ter Braak 1990, 1992) recommended in ordinary multiple regression.

Details about these permutation methods are given in Legendre & Legendre (1998, pp. 606-612) and in Anderson & Legendre (1999). For regression through the origin, the methods for permutation testing are described in Legendre & Desdevises (manuscript).

For ordinary multiple regression, Monte Carlo simulations conducted by Anderson & Legendre (1999) concluded that:

• For data with non-normal error structure, permutation tests had type I error closer to the nominal significance level α. They also have greater power than the normal-theory *t*-test.

• Permutation of raw data and permutation of residuals gave asymptotically equivalent results in most situations and provided good approximative tests for partial regression coefficients.

• Permutation of raw data resulted in unstable (and often inflated) type I error when the covariable contained an extreme outlier, whether or not there was collinearity between predictor variables, or the data were normal or non-normal. This problem was not amended with increasing sample sizes.

• The presence of outliers in the covariable did not adversely affect the tests based on permutation of residuals.

Thus, permutation of raw data should not be used when the covariables contain (or may contain) outliers; permutation of residuals should be used in that case.

Input file(s)

Either a single data table containing the predictor (X) and response (y) variables, or two data files: one for X and one for y. Each data table is written to an ASCII (text) file without identifiers for the rows or columns.

1. A single rectangular data table with rows as objects and columns as variables. The response variable (y) may come first or last among the columns. The program asks about the position of the y variable in the file (first or last), and also the number of objects and predictor variables.

2. Two data files: one for the predictor variables (X) and another for the response variable (y). The program asks for the names of these two files in turn, as well as the number of objects and variables in the file containing the X variables. The number of objects of the two files is assumed to be the same.

Output files

1. Multiple regression equation and tests of significance. The tests of individual regression coefficients can be one-tailed or two-tailed.

2. Optional file with fitted values and regression residuals.

See example below.

Example 1

Input file

The following data are from Table 16.1 of Sokal & Rohlf (1995). They concern 41 cities in the US. The first two columns are two environmental variables (X). The third one (y) is SO_2 content of the air.

x_1	x_2	y
70.3	213	10
61.0	91	13
56.7	453	12
51.9	454	17
49.1	412	56
54.0	80	36
57.3	434	29
68.4	136	14
75.5	207	10
61.5	368	24
50.6	3344	110
52.3	361	28
49.0	104	17
56.6	125	8
55.6	291	30
68.3	204	9

55.0	625	47
49.9	1064	35
43.5	699	29
54.5	381	14
55.9	775	56
51.5	181	14
56.8	46	11
47.6	44	46
47.1	391	11
54.0	462	23
49.7	1007	65
51.5	266	26
54.6	1692	69
50.4	347	61
50.0	343	94
61.6	337	10
59.4	275	18
66.2	641	9
68.9	721	10
51.0	137	28
59.3	96	31
57.8	197	26
51.1	379	29
55.2	35	31
45.7	569	16

Dialogue with the program

```
Français: tapez (1)
English:  type  (2)
2

Program for multiple regression
with permutation tests.
Option: intercept forced to 0

Pierre Legendre
Département de sciences biologiques
Université de Montréal
      © Pierre Legendre, 1999, 2002

Data file(s):
(1) A single file containing variables X and y?
(2) Separate files for X (predictor variables) and y (response variable)?
1

(0) Intercept forced to 0
(1) Ordinary regression model with an intercept
1

Data file:
(1) Predictors X first, followed by response variable y?
```

```
(2) Response variable y first, followed by predictors X?
1

Name of the data file?
Input data file: S&R Table 16.1/41x3/X1,X2,y

How many objects and predictor variables?
(Do not count the response variable y)
41 2

Input data file:      S&R Table 16.1/41x3/X1,X2,y
   41 objects
    1 response variable
    2 predictor variables

Regression coefficients: (1) one-tailed or (2) two-tailed test?
1

To test the regression coefficients:
  (0) no permutation test
  (1) permute the raw data
  (2) permute the residuals of the full regression model (recommended)
2

How many permutations? (e.g. 999, 9999, ...)
9999

Do you want the fitted values and residuals?
(0) no,  (1) yes
1
They will be written out to the file 'Adjusted values and residuals'

R^2 =  0.51611    F =    20.26553    prob (param.) =  0.00000 *
                                     prob (perm.)  =  0.00010 *
Number of permutations of raw data:  9999

Regression coefficient(s): One-tailed test in direction of sign
Number of permutations of residuals:  9999

-----------------------------------------------------------
Variable            b         t       P-perm     P-param
-----------------------------------------------------------
Intercept       77.23671   3.59122               0.00046 *
    1           -1.04805  -2.80777   0.00200 *   0.00392 *
    2            0.02430   5.07607   0.00100 *   0.00001 *
-----------------------------------------------------------
* the parameter estimate is significant at the 0.05 level

Computation time:      1.97 sec.

Results are also found in file 'Regression.out'
End of the program.
```

Files of results

File 'Regression.out': This file contains the coefficient of determination, the multiple regression equation, as well as the tests of significance. The correlation coefficient is also given in cases of simple linear regressions (a single predictor variable).

```
Program for multiple regression
with permutation tests.
Option: intercept forced to 0

Pierre Legendre
Département de sciences biologiques
Université de Montréal
      © Pierre Legendre, 1999, 2002


Input data file:      S&R Table 16.1/41x3/X1,X2,y                    [1]
   41 objects
    1 response variable
    2 predictor variables

R^2 =  0.51611    F =    20.26553    prob (param.) =  0.00000 *      [2]
                                     prob (perm.)  =  0.00010 *
Number of permutations of raw data:  9999                           [3]

Regression coefficient(s): One-tailed test in direction of sign
Number of permutations of residuals:  9999                          [4]

-----------------------------------------------------------
Variable            b         t       P-perm     P-param
-----------------------------------------------------------
Intercept       77.23671   3.59122               0.00046 *
    1           -1.04805  -2.80777   0.00200 *   0.00392 *
    2            0.02430   5.07607   0.00100 *   0.00001 *
-----------------------------------------------------------
* the parameter estimate is significant at the 0.05 level

Computation time:      1.97 sec.
```

The explanations that follow refer to the numbers [in brackets] at the right of the output listed above.

[1] Identification of the file or files used for input.

[2] Coefficient of multiple determination (R^2, noted R^2), *F*-statistic, and associated parametric and permutational probabilities.

[3] Permutation method and number of permutations used to test the significance of the coefficient of determination (R^2).

[4] Permutation method and number of permutations used to test the significance of the regression coefficients.

[5] The parameters of the model (intercept and partial regression coefficients) are listed in column 'b', which is followed by the associated *t*-statistics. The next column gives the permutational probabilities for the slope parameters; the last column contains the parametric probabilities. The tests of individual regression coefficients are one-tailed or two-tailed depending on user's choice. Probabilities equal to or smaller than 0.05 are identified by asterisks.

File 'Adjusted values and residuals': This file contains the fitted values and regression residuals.

Fitted val.	Residuals
8.73533	1.26467
15.51714	-2.51714
28.82178	-16.82178
33.87674	-16.87674
35.79052	20.20948
22.58617	13.41383
27.73117	1.26883
8.85523	5.14477
3.13963	6.86037
21.72529	2.27471
105.47747	4.52253
31.19725	-3.19725
28.40973	-11.40973
20.95491	-12.95491
26.03741	3.96259
10.61270	-1.61270
34.78374	12.21626
50.79822	-15.79822
48.63484	-19.63484
29.37762	-15.37762
37.48608	18.51392
27.66099	-13.66099
18.82529	-7.82529
28.41877	17.58123
37.37625	-26.37625
31.87026	-8.87026
49.62251	15.37749
29.72683	-3.72683
61.13522	7.86478
32.84830	28.15170
33.17030	60.82970
20.86707	-10.86707
21.66594	-3.66594
23.43441	-14.43441
22.54898	-12.54898
27.11565	0.88435
17.42035	13.57965
21.44712	4.55288
32.89239	-3.89239
20.23483	10.76517
43.16961	-27.16961

Example 2

Input file

The following data are from Legendre & Desdevises (manuscript). Independent contrasts for two variables were computed on the phylogenetic tree of *Lamellodiscus* parasites: non-specificity index (NSI, response variable) and maximum host size (explanatory variable). They were modelled using regression through the origin (Garland et al. 1992). Number of pairs of contrasts: $n = 17$.

NSI	Maximum host size
0.04152	0.00405
0.00000	0.00000
0.01716	0.03023
0.00000	0.00000
0.16553	0.09463
0.45859	-0.20256
0.18470	-0.11613
0.00000	0.09224
0.00000	-0.03719
-0.08754	-0.08257
0.11719	-0.02669
0.16120	-0.00904
0.25614	-0.07076
0.12913	-0.08901
0.09254	-0.04756
0.11082	-0.00829
0.01401	0.04786

Dialogue with the program

```
Français: tapez (1)
English:  type  (2)
2

Program for multiple regression
with permutation tests.
Option: intercept forced to 0

Pierre Legendre
Département de sciences biologiques
Université de Montréal
      © Pierre Legendre, 1999, 2002

Data file(s):
(1) A single file containing variables X and y?
(2) Separate files for X (predictor variables) and y (response variable)?
1
```

```
(0) Intercept forced to 0
(1) Ordinary regression model with an intercept
0

Data file:
(1) Predictors X first, followed by response variable y?
(2) Response variable y first, followed by predictors X?
2

Name of the data file?
Input data file: NSI=f(HostSize).txt

How many objects and predictor variables?
(Do not count the response variable y)
17 1

Input data file: NSI=f(HostSize).txt
   17 objects
    1 response variable
    1 predictor variables

Regression coefficients: (1) one-tailed or (2) two-tailed test?
1

To test the regression coefficients:
  (0) no permutation test
  (1) permute the raw data
1

How many permutations? (e.g. 999, 9999, ...)
99999

Do you want the fitted values and residuals?
(0) no,  (1) yes
0

r   = -0.63078
R^2 =  0.39788    F =    10.57295    prob (param.) =  0.00500 *
                                     prob (perm.)  =  0.00763 *
Number of permutations of raw data:  99999

Regression coefficient(s): One-tailed test in direction of sign
Number of permutations of residuals:  99999

------------------------------------------------------------
Variable            b         t       P-perm     P-param
------------------------------------------------------------
    1           -1.30324  -3.25161   0.00380 *   0.00250 *
------------------------------------------------------------
* the parameter estimate is significant at the 0.05 level

Computation time:      4.18 sec.
Results are also found in file 'Regression.out'
End of the program.
```

Files of results

File 'Regression.out': This file contains the coefficient of determination, the multiple regression equation, as well as the tests of significance. The correlation coefficient is also given in cases of simple linear regressions (a single predictor variable), as it is the case in this example.

```
Program for multiple regression
with permutation tests.
Option: intercept forced to 0

Pierre Legendre
Département de sciences biologiques
Université de Montréal
      © Pierre Legendre, 1999, 2002


Input data file:         NSI=f(HostSize)/17x2.txt
   17 objects
    1 response variable
    1 predictor variables

r   = -0.63078                          Two-tailed test:
R^2 =  0.39788    F =    10.57295       prob (param.) =  0.00500 *
                                        prob (perm.)  =  0.00763 *
Number of permutations of raw data:  99999

Regression coefficient(s): One-tailed test in direction of sign
Number of permutations of residuals:  99999

-----------------------------------------------------------
Variable            b         t       P-perm     P-param
-----------------------------------------------------------
    1           -1.30324  -3.25161   0.00380 *   0.00250 *
-----------------------------------------------------------
* the parameter estimate is significant at the 0.05 level

Computation time:      4.18 sec.
```

Disclaimer

This program is provided without any explicit or implicit warranty of correct functioning. It has been developed as part of a university-based research program. If, however, you should encounter problems with this program, the author will be happy to help solve them. Researchers may use this program for scientific purposes, but the source code remains the property of Pierre Legendre. Users of the program may refer to the present user's manual as follows:

Legendre, P. 2002. Program for multiple linear regression (ordinary or through the origin) with permutation test – User's notes. Département de sciences biologiques, Université de Montréal. 11 pages.
Available from the WWWeb site <http://www.fas.umontreal.ca/biol/legendre/>.

Program distribution

Computer programs written by P. Legendre are available from our base WWWeb site. They include FORTRAN77 source code, documentation, sample files and executable programs. Versions for MacOS (68k or PowerPC) and 32-bit DOS (suitable for DOS sessions under Windows 95/98/NT) are provided. WWWeb address: <http://www.fas.umontreal.ca/biol/legendre/>.

The programs are written in FORTRAN77 in order to facilitate diffusion. Indeed, there is a compiler, GNU FORTRAN (or g77; see http://www.gnu.org/software/fortran/), that is freely available for this level of FORTRAN, for the DOS/Windows, MacOS X, Unix, and Linux families of operating systems.

References

Anderson, M. J. & P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* 62: 271-303.

Garland, T. Jr., P. H. Harvey & A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic Biology* 41: 18-32.

Legendre, P. & L. Legendre. 1998. *Numerical ecology. 2nd English edition.* Elsevier Science BV, Amsterdam. xv + 853 pages.

Legendre, P. & Y. Desdevises. 2002. Independent contrasts and regression through the origin. Manuscript.

Sokal, R. R. & F. J. Rohlf. 1995. *Biometry – The principles and practice of statistics in biological research. 3rd edition.* W. H. Freeman, New York.

ter Braak, C. J. F. 1990. *Update notes: CANOCO version 3.10.* Agricultural Mathematics Group, Wageningen.

ter Braak, C. J. F. 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. 79-86 in: K.-H. Jöckel, G. Rothe & W. Sendler [eds.] *Bootstrapping and related techniques.* Springer-Verlag, Berlin.

Unix/DOS user's notes prepared by Philippe Casgrain

The Unix (including MacOS X) and DOS versions of this program were built with g77, the GNU FORTRAN compiler. They are command-line tools, which means that they must be started from the command line.

Furthermore, files created by the program, such as our "**.**out" files, cannot be deleted by the FORTRAN program. If, after launching, the program ends abruptly and a message is displayed, such as:

```
open: 'new' file exists
apparent state: unit 4 named NesAnova.out
lately writing direct unformatted external IO
Abort
```

this means that a file called "NesAnova.out" already exists in the current directory. Rename or remove that file before running the program again. This is a feature, not a bug.

Unix instructions

1. Open a new shell.
MacOS X users: open /Applications/Utilities/Terminal

2. At the prompt
(e.g. "[localhost:~] username%") type:
"cd /path/to/the/program/"
where "/path/to/the/program/" represents
the directory where the program is found.
Examples: /Applications, ~/Desktop, etc.

Don't forget that Unix systems are case-sensitive:
upper- and lowercase letters are different.

DOS instructions

1. Open a new shell: from the Start menu,
choose Programs → Accessories → MS-DOS

2. At the DOS prompt (e.g., C:\WINDOWS\>),
type "cd c:\path\to\the\program"
where "\path\to\the\program" represents
the directory where the program is found.

Examples: c:\tmp, c:\windows\desktop, etc.

3. Press the *Return* key.

4. Type the name of the program to start it.

Example: **.**/regressn
(the prefix "**./**" is essential if the program is not
part of your usual path for command-line utilities)

Example: regressn**.**exe

5. Press the *Return* key.

6. Follow the on-screen instructions.
