Programme de régression linéaire multiple (ordinaire ou par l'origine) avec tests par permutation – Guide

Pierre Legendre
Département de sciences biologiques
Université de Montréal

Mars 2002

Ce programme calcule des régressions linéaires multiples et teste la signification des paramètres par permutation. Dans cette nouvelle version, on peut forcer la droite de régression à passer par l'origine. Les tests par permutation sont recommandés lorsque les résidus de la régression ne sont pas distribués normalement; ils ne solutionnent pas les problèmes posés par l'hétéroscédasticité. Deux méthodes de permutation sont disponibles dans le programme:

(1) Permutation des données brutes.

(2) Permutation des résidus du modèle de régression (ter Braak 1990, 1992).

On trouvera une explication détaillée de ces méthodes de permutation dans Legendre & Legendre (1998, pp. 606-612) ainsi que dans Anderson & Legendre (1999). Pour la régression forcée à l'origine, les méthodes de test par permutation sont décrites dans l'article de Legendre & Desdevises (manuscrit).

Des simulations de Monte Carlo réalisées par by Anderson & Legendre (1999) ont conduit aux observations suivantes:

• Pour des données générées avec un terme d'erreur non normal, les tests par permutation ont une erreur de type I plus près du seuil de signification α. Ils ont aussi davantage de puissance que les tests *t* paramétriques.

• La permutation des données brutes et la permutation des résidus produisent des résultats asymptotiquement équivalents. Les deux méthodes fournissent un bon test approximatif des coefficients de régression partielle.

• Lorsque la covariable contient des observations aberrantes, la méthode de permutation des données brutes produit des résultats instables en termes d'erreur de type I; celle-ci est surestimée la plupart du temps. Cela se produit pour des données normales ou non, avec ou sans colinéarité entre les variables prédictives. D'augmenter le nombre d'observations ne règle pas ce problème.

• La présence d'observations aberrantes dans la covariable n'a aucun effet néfaste sur les résultats obtenus par permutation des résidus.

On en conclut que la méthode de permutation des données brutes ne doit pas être employée lorsque les covariables contiennent (ou peuvent contenir) des observations aberrantes. On doit plutôt employer la méthode de permutation des résidus dans ce cas.

Fichier(s) de données

On peut fournir au programme un seul fichier contenant les variables prédictives (X) et la variable réponse (y), ou encore deux fichiers de données: l'un contenant les variables X et l'autre la variable y. Chaque tableau de données constitue un fichier ASCII (texte) sans identificateurs, ni pour les lignes, ni pour les colonnes.

1. Toutes les données peuvent se trouver dans un seul tableau rectangulaire de données dont les lignes correspondent aux objets et les colonnes aux variables. La variable réponse (y) peut être en première ou en dernière colonne. Le programme demande à l'utilisateur la position de la variable y dans le tableau. Il demande également combien il y a d'objets et de variables prédictives.

2. On peut fournir deux fichiers de données: l'un pour les variables prédictives (X) et l'autre pour la variable réponse (y). Le programme demande le nom de chacun des deux fichiers, de même que le nombre d'objets et le nombre de variables dans le tableau contenant les variables X. Le programme suppose que le nombre d'objets des deux fichiers est le même.

Fichiers de résultats

1. L'équation de régression multiple et les tests de signification. On peut choisir de faire des tests unilatéraux ou bilatéraux des coefficients de régression partielle.

2. Un second fichier (optionnel) contient les valeurs ajustées ainsi que les résidus de la régression.

Voir l'exemple ci-dessous.

Exemple 1

Fichier de données

Les données ci-dessous sont extraites du tableau 16.1 de Sokal & Rohlf (1995). Les lignes correspondent à 41 villes des USA. Les deux premières colonnes contiennent des variables environnementales (X). La troisième colonne (y) représente la teneur en SO_2 de l'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 avec le programme

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

Programme de régression multiple
avec tests par permutations.
Option: ordonnée à l'origine forcée à 0

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

Fichier(s) de données:
(1) Un seul fichier contenant les variables X et y?
(2) Fichiers séparés pour les variables prédictives X et la variable
réponse y?
1

(0) Ordonnée à l'origine forcée à 0
(1) Modèle de régression ordinaire avec une ordonnée à l'origine
1
```

```
Fichier de données:
(1) Les variables X d'abord, la variable réponse y ensuite?
(2) La variable réponse y d'abord, les variables X ensuite?
1

Nom du fichier de données?
Fichier de données: S&R Table 16.1/41x3/X1,X2,y

Combien d'objets et de variables prédictives?
(Ne pas compter la variable réponse y)
41 2

Fichier de données:      S&R Table 16.1/41x3/X1,X2,y
   41 objets
    1 variable réponse
    2 variables prédictives

Coefficients de régression: Test (1) unilatéral ou (2) bilatéral?
1

Pour le test des coefficients de régression:
  (0) pas de test par permutation
  (1) permuter les données brutes
  (2) permuter les résidus du modèle de régression complet (recommandé)
2

Combien de permutations? (e.g. 999, 9999, ...)
9999

Voulez-vous les valeurs ajustées ainsi que les résidus?
(0) non,  (1) oui
1
Ces valeurs se trouveront dans le fichier 'Valeurs ajustées et résidus'

R^2 =  0.51611    F =    20.26553    prob (param.) =  0.00000 *
                                     prob (perm.)  =  0.00010 *
Nombre de permutations des données brutes:  9999

Coefficients de régression: Test unilatéral dans la direction du signe
Nombre de permutations des résidus:  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   0.00050 *    0.00001 *
-------------------------------------------------------------
* l'estimation du paramètre est significative au seuil 0.05

Durée du calcul:      1.97 sec.

Les résultats se trouvent également dans le fichier 'Regression.out'
Fin du programme.
```

Fichiers de résultats

Fichier 'Regression.out': Ce fichier contient le coefficient de détermination, l'équation de régression, ainsi que les résultats des tests de signification. Le coefficient de corrélation est également fourni pour les régressions linéaires simples (une seule variable prédictive).

```
Programme de régression multiple
avec tests par permutations.
Option: ordonnée à l'origine forcée à 0

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


Fichier de données:      S&R Table 16.1/41x3/X1,X2,y                [1]
   41 objets
    1 variable réponse
    2 variables prédictives

R^2 =  0.51611    F =    20.26553    prob (param.) =  0.00000 *     [2]
                                     prob (perm.)  =  0.00010 *
Nombre de permutations des données brutes:  9999                    [3]

Coefficients de régression: Test unilatéral dans la direction du signe
Nombre de permutations des résidus:  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   0.00050 *   0.00001 *
------------------------------------------------------------
* l'estimation du paramètre est significative au seuil 0.05

Durée du calcul:      1.97 sec.
```

Les explications qui suivent se rapportent aux numéros [entre crochets] à droite du tableau de résultats ci-dessus.

[1] Identification du fichier ou des fichiers de données.

[2] Coefficient de détermination multiple (R^2, noté R^2), statistique *F*, probabilité paramétrique et probabilité permutationnelle associées.

[3] Méthode utilisée pour le test du coefficients de détermination (R^2) par permutation, ainsi que le nombre de permutations.

[4] Méthode utilisée pour le test des coefficients de régression par permutation, ainsi que le nombre de permutations.

[5] Les paramètres du modèle (l'ordonnée à l'origine et les coefficients de régression partielle) se trouvent dans la colonne 'b' qui est suivie de la colonne des statistiques *t* associées. La colonne suivante fournit la probabilité permutationnelle pour les paramètres de pente. La dernière colonne contient les probabilités paramétriques. Les tests des coefficients de régression individuels sont unilatéraux ou bilatéraux, selon le choix de l'utilisateur. Les probabilités inférieures ou égales à 0.05 sont identifiées par un astérisque.

Fichier 'Valeurs ajustées et résidus': Ce fichier contient les valeurs ajustées et les résidus de la régression.

Val.ajustées	Résidus
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

Exemple 2

Fichier de données

Les données qui suivent proviennent de l'article de Legendre & Desdevises (manuscrit). Des contrastes indépendants ont été calculés pour deux variables sur l'arbre phylogénétique de parasites du genre *Lamellodiscus*: l'indice de non-spécificité (NSI, variable réponse) et la taille maximum de l'hôte (variable explicative). La régression par l'origine a été utilisée pour la modélisation (Garland et al. 1992). Nombre de paires de contrastes: $n = 17$.

NSI	Taille maximum de l'hôte
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 avec le programme

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

Programme de régression multiple
avec tests par permutations.
Option: ordonnée à l'origine forcée à 0

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

Fichier(s) de données:
(1) Un seul fichier contenant les variables X et y?
(2) Fichiers séparés pour les variables prédictives X et la variable
réponse y?
1
```

```
(0) Ordonnée à l'origine forcée à 0
(1) Modèle de régression ordinaire avec une ordonnée à l'origine
0

Fichier de données:
(1) Les variables X d'abord, la variable réponse y ensuite?
(2) La variable réponse y d'abord, les variables X ensuite?
2

Nom du fichier de données?
Fichier de données: NSI=f(HostSize).txt

Combien d'objets et de variables prédictives?
(Ne pas compter la variable réponse y)
41 2

Fichier de données:      NSI=f(HostSize).txt
   17 objets
    1 variable réponse
    1 variables prédictives

Coefficients de régression: Test (1) unilatéral ou (2) bilatéral?
1

Pour le test des coefficients de régression:
  (0) pas de test par permutation
  (1) permuter les données brutes
1
Combien de permutations? (e.g. 999, 9999, ...)
99999

Voulez-vous les valeurs ajustées ainsi que les résidus?
(0) non,  (1) oui
0

r   = -0.63078                        Test bilatéral:
R^2 =  0.39788    F =    10.57295    prob (param.) =  0.00500 *
                                     prob (perm.)  =  0.00763 *
Nombre de permutations des données brutes:  99999

Coefficients de régression: Test unilatéral dans la direction du signe
Nombre de permutations des résidus:  99999

-------------------------------------------------------------
Variable            b         t       P-perm     P-param
-------------------------------------------------------------
    1           -1.30324  -3.25161   0.00380 *   0.00250 *
-------------------------------------------------------------
* l'estimation du paramètre est significative au seuil 0.05

Durée du calcul:      1.97 sec.
Les résultats se trouvent également dans le fichier 'Regression.out'
Fin du programme.
```

Fichiers de résultats

Fichier ‘Regression.out’: Ce fichier contient le coefficient de détermination, l’équation de régression, ainsi que les résultats des tests de signification. Le coefficient de corrélation est également fourni pour les régressions linéaires simples (une seule variable prédictive), comme c'est le cas dans cet exemple.

```
Programme de régression multiple
avec tests par permutations.
Option: ordonnée à l'origine forcée à 0

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


Fichier de données:      NSI=f(HostSize).txt
   17 objets
    1 variable réponse
    1 variables prédictives

r   = -0.63078                        Test bilatéral:
R^2 =  0.39788    F =    10.57295    prob (param.) =  0.00500 *
                                     prob (perm.)  =  0.00763 *
Nombre de permutations des données brutes: 99999

Coefficients de régression: Test unilatéral dans la direction du signe
Nombre de permutations des données brutes: 99999

-----------------------------------------------------------
Variable            b         t       P-perm     P-param
-----------------------------------------------------------
    1           -1.30324  -3.25161   0.00380 *   0.00250 *
-----------------------------------------------------------
* l'estimation du paramètre est significative au seuil 0.05

Durée du calcul:      4.18 sec.
```

Avertissement

Ce programmes est fourni sans aucune garantie implicite ou explicite de bon fonctionnement. Il a été mis au point dans le cadre de recherches universitaires. Cependant, si vous éprouvez des problèmes avec le programme, l'auteur se fera un plaisir de tenter de vous dépanner. Les chercheurs peuvent utiliser ce programme pour les fins de leurs recherches, mais le code-source demeure la propriété de Pierre Legendre. Les utilisateurs pourront référer au présent manuel comme suit:

Legendre, P. 2002. Programme de régression linéaire multiple (ordinaire ou par l'origine) avec tests par permutation – Guide. Département de sciences biologiques, Université de Montréal. 11 pages.
Distribué à partir du site WWWeb <http://www.fas.umontreal.ca/biol/legendre/>.

Distribution du programme

Les programmes écrits par P. Legendre sont disponibles sur notre site WWWeb. On y trouve les fichiers-source FORTRAN, la documentation, des fichiers de données pour des essais, de même que les programmes exécutables pour MacOS (68k ou PowerPC) et DOS 32-bits (approprié pour des sessions DOS sous Windows 95/98/NT). L'adresse WWWeb est la suivante: http://www.fas.umontreal.ca/biol/legendre/.

Les programmes sont écrits en FORTRAN77 afin d'en faciliter la diffusion. Il existe en effet un compilateur, GNU FORTRAN77, qui est disponible gratuitement pour cette évolution du langage FORTRAN, pour les familles de systèmes d'opération DOS/Windows, MacOS X, Unix et Linux.

Références

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.
