

DISTLM *forward*

Distance-based multivariate analysis for a linear model
using forward selection

A computer program
by Marti J. Anderson



Department of Statistics
University of Auckland
(2003)

DISCLAIMER

This FORTRAN 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 can use this program for scientific and research purposes, but intellectual property resides with Marti Jane Anderson and Brian McArdle and copyright for the source code and program remains the property of Marti Jane Anderson. This program relies on several canned FORTRAN routines from Numerical Recipes (Press et al. 1992). Namely, for finding the eigenvalues of a real symmetric distance matrix, using householder reduction, DISTLM uses routines “tred2” and “tqli.” Also, to find the solution to linear equations (i.e. the inverse of a matrix) by Gauss-Jordan elimination, DISTLM uses the “gaussj” subroutine. Also, to obtain Monte Carlo samples from the asymptotic permutation distribution (if requested by the user), the “genchi” function from the RanLib library was used (courtesy of B. W. Brown and J. Lovato, Department of Biomathematics, University of Texas, accessed through webpage <http://lib.stat.cmu.edu/general/Utexas/>).

Research publications that use this method should cite the paper by McArdle and Anderson (2001). Users of the computer program may also refer to the present user’s guide as follows:

Anderson, M.J. 2003. DISTLM *forward*: a FORTRAN computer program to calculate a distance-based multivariate analysis for a linear model using forward selection. Department of Statistics, University of Auckland, New Zealand.

Author’s contact details:

Dr Marti J. Anderson
Department of Statistics
Tamaki Campus
University of Auckland
Private Bag 92019
Auckland, New Zealand
Tel: 64-9-373-7599 ext 85052
Fax: 64-9-373-7000
Email: mja@stat.auckland.ac.nz
Webpage: <http://www.stat.auckland.ac.nz/~mja>

I. Description

DISTLM *forward* is a computer program that does a multivariate multiple regression on the basis of any distance measure and does a forward selection of the predictor variables, either individually or in sets, with tests by permutation. It is a straightforward modification of the DISTLM program to perform a forward selection of predictor variables. For details of the DISTLM procedure, see the user notes for DISTLM.

II. Fitting individual variables or sets of variables

The program offers essentially two options: one can either ask for a forward selection of individual variables, or for a forward selection of sets of variables. The first is useful in the general case, e.g. for fitting individual environmental variables sequentially in the linear model. The second is useful for the situation where one wishes to fit a sequential model of whole sets of variables. For example, in the paper by Anderson et al. (2004), there were seven sets of environmental variables of interest:

- (1) Ambient sediment grain size variables (GS1 – GS4),
- (2) Depositional environment classification (contrasts between High, Medium and Low depositional environments, labeled HvML and MvL).
- (3) Trapped sediment characteristics (Sdep, gt125, Perfin)
- (4) Erosion variables (bed height movement, labeled BH and sdBH)
- (5) Distance from the mouth of the estuary (D and D²).
- (6) Chlorophyll a (Chla) and
- (7) Organics (Org)

Interest lies in characterizing and modeling the soft-sediment faunal assemblages in the Okura estuary in terms of these environmental variables, individually and in sets.

The program uses the proportion of the total sum of squares that is explained by the individual variable (or the set) as the criterion for the forward selection. Output from the program then includes: (a) the results of the marginal tests (i.e. fitting each variable (or set) individually, ignoring other variables (or sets)) followed by (b) the results of the forward selection procedure with the conditional tests, (i.e. fitting each variable one at a time, conditional on the variables that are already included in the model). Also included in the output file is information on the correlation among all pairs of explanatory variables. This provides a further check on issues of multi-collinearity. It is well-known that the forward selection procedure does not result in a “best” fitted model from the variables available. However, it does give a reasonable starting point, at least, for further investigation. See Neter et al. (1996) for further details concerning multi-collinearity and different variable selection procedures for linear models.

III. Input files

The program allows the user to input a distance matrix **D** directly or a raw multivariate data matrix of response variables **Y**. In either case, the file should be saved as ASCII *.txt, with no column or row headings. This is the first input file for the program. If a file with a raw data matrix is input, either the rows or the columns may correspond to the variables for the analysis (the user will be given the option to choose one of these). Then, the user has several options for transformation, standardisation and choice of distance measure.

The other input file is an **X** matrix containing one or more explanatory (predictor) variables of interest. This matrix needs to have the *columns* as variables. There should then be as many rows of data in the **X** matrix as there are observations in the original **Y** (or **D**) matrix. In addition, this **X** matrix should have a name at the top of each column for each variable in the analysis. ***These names must not contain any spaces and must have less than 8 characters.*** The file must be saved as an ASCII *.txt file for input into the program.

For example, in the case of the example dataset from Anderson et al. (2004), the text file containing the X variables of interest looks like this:

```

GS1    GS2    GS3    GS4    HvML    MvL    Sdep    gt125    Perfin    BH    sdBH    D    D2    Chla    Org
23.59734  8.702262  26.05058  18.55831  -0.5    -1    0.020654  8.010118  73.40201  -0.66908  3.164503  1    1    6.240836  2.647367
 15.3906  18.14612  62.03663  4.42666  -0.5    1    0.043621  22.92207  54.73604  0.052879  0.642684  4    16    6.195168  1.714974
 57.3087  19.60367  17.71087  4.309605  1    0    0.023211  17.53434  59.05184  -0.263372  1.223297  5    25    4.874898  3.710351
19.65692  20.48721  55.57069  4.28518  -0.5    -1    0.046061  26.71051  40.36132  -0.767586  3.889855  6    36    4.739551  1.39717
19.76767  16.713  58.90293  4.616395  -0.5    1    0.051658  11.28514  68.05619  -0.186799  1.412287  9    81    6.94297  0.595493
20.24753  18.84764  53.29374  6.419465  -0.5    -1    0.036735  7.15513  72.74136  0.47967  4.315259  10   100   5.942024  1.585761
38.15989  10.94259  34.31177  14.3252  -0.5    1    0.031428  19.86356  56.01299  -0.041274  2.921794  11   121   4.462916  2.863418
41.97064  30.69571  25.36141  1.819522  1    0    0.038238  13.20201  58.40271  -0.33238  1.684918  14   196   5.053443  0.943299
47.18145  31.18641  21.48586  0.146285  1    0    0.022882  11.48004  74.52253  -0.20468  1.453698  15   225   6.800356  3.569125
50.55069  21.17107  21.14967  5.798005  1    0    0.027657  7.304065  76.95846  -0.014446  1.142638  7    49    6.282705  2.084369
17.31279  10.62196  52.62707  16.21854  -0.5    1    0.072395  58.44392  19.14886  0.005563  2.671184  8    64    8.299199  2.862428
13.82543  13.50158  58.90341  13.30441  -0.5    -1    0.064342  30.63321  41.92901  0.30757  2.128036  2    4    4.064562  1.557842
5.875537  8.350132  72.14357  10.87457  -0.5    -1    0.031218  42.76244  36.61735  -0.267318  1.045565  3    9    3.152648  1.091007
etc...

```

Note:

If you are fitting *sets* of variables, then the program will require *another* text file that contains the information shown in the Figure below:

name	No. of sets	Columns 1 through 4 in the X data file correspond to the 4 variables in the set called "Ambient"
Ambient	7	1 4
Dep	5	6
Trapped	7	9
Erosion	10	11
Dist	12	13
Chla	14	14
Org	15	15

1 line for each set, giving the name and then the positions of the first and last variable in the set as it occurs in the X input file.

The first entry in the file will be the number of sets, then there will be a line for each. In each line will be given first the name of the set and then the numbers that correspond to the first and last columns in the X data file that contain those sets. Note that variables intended to be in the same set must be listed together within the X data file. There should be tab or space delimiters between the names and the numbers in this file and the file should be saved as an ASCII *.txt file.

To avoid dealing with long file names and paths to locate files, place all relevant input files in the same location on your computer (i.e. the same directory) as the DISTLM_forward.exe (or DISTLM_forward.mac) file, for use with the program. Double-click on the "DISTLM_forward.exe" (or "DISTLM_forward.mac") file to run the program.

If you are using a Macintosh, then save all input files (for example, from Excel) as "Text (Windows)". If you save it as a simple text file from Excel, it won't run unless you open up the file again (for example, using a text editor such as BBEdit) and manually add a carriage return ("Enter") to the end of the last line. This is a weird quirk of the way the Mac version of Excel saves text files that, for some reason unknown to me, causes my program to crash. You should not experience any such problem, however, if the file is saved as "Text (Windows)" or if the file was created using a different text editor.

DISTLM_forward is written using dynamic allocation of memory and thus (theoretically) there are no limits to the sizes of the matrices (rows or columns) that are entered for analysis. Please contact the author if problems (other than the time required) are encountered in the analysis of large matrices.

If you are using a Macintosh and you get an error that reads:

```

BUFFER allocation failed
REWIND(UNIT=*,...

```

then you need to increase the memory allocated to the program. To do this, click on the program's icon and type "i" while holding the apple key, then choose "Memory" (or choose "Get Info > Memory" from the File menu). Depending on the size of your matrices, you might even have to increase this value to something obnoxious, like 50000 or so. However, if you have a large input file and cannot seem to get the program to work even after doing this, then please contact the author.

IV. Questions Asked by the Program

The questions asked by the program are best demonstrated by the use of an example. The data analysed here are from a monitoring program of soft sediment assemblages in the Okura estuary, north of Auckland, New Zealand (Anderson et al. 2004). The **Y** matrix consisted of a total of 88 observations of the abundances of 73 soft-sediment taxa. Environmental variables thought to be important in structuring assemblages were also recorded, as listed in the above section. The purpose of the analysis was to test hypotheses about the relationship between the fauna and the environmental variables and to build a model, using forward selection, using (a) individual variables and (b) sets of variables. One particular hypothesis of interest was to determine whether information from trapped sediments added anything significant toward our ability to model these assemblages, given the strong known relationship of faunal assemblages with grain sizes of ambient sediments.

The response variable matrix **Y** (88 rows x 73 columns) was contained in an ASCII text file called "OkuraY.txt". The file containing the matrix of environmental variables of interest for the test (88 rows + a header row x 15 columns) was called "OkuraX.txt." The file containing the variable set information is called "SetsOkura.txt".

The questions asked by the program are given in Courier font, while responses given for this data set are in *Times bold italics*. Note that you can stop the program at any time by typing Ctrl-C (as for any DOS program) or, if you are a Mac user and this does not work, then try typing a period (a full stop) while holding down the apple key.

DISTLM forward

 A program for analysing multivariate data
 on the basis of any distance measure,
 with forward selection of explanatory variables
 in a linear regression model,
 using permutations.

by M.J. Anderson
 Department of Statistics
 University of Auckland (2003)

Type the name of the input file containing your data
 (i.e. response variables).

OkuraY.txt

Type a name for the output file of results (*.txt)

Results.txt

Note: If the program crashes it may be because you already have a file in the directory with the same name as what you are typing here for the output file. If so, then you need to delete the previous file or type a new name. Unfortunately, the program is not smart enough to write over the top of an existing file.

Nature of the data in the input file:

- 1) raw data (n x p)
- 2) distance matrix (n x n)

I

Structure of the input file:

- 1) rows are samples and columns are variables
- 2) columns are samples and rows are variables

1

How many variables (columns) are there?

73

How many observations (rows) are there?

88

Choice of transformation:

- 1) none
- 2) square-root
- 3) fourth-root
- 4) $\ln(x)$
- 5) $\ln(x+1)$
- 6) $\log_{10}(x)$
- 7) $\log_{10}(x+1)$
- 8) presence/absence

5

Choice of standardisation:

- 1) none
- 2) standardise by row (sample) sums
- 3) standardise by column (variable) sums
- 4) double standardise by row and column sums
- 5) standardise each variable to z-scores (normalize)
- 6) standardise each variable by dividing by its s.d.
- 7) standardise each variable by dividing by its range

1

Choice of distance measure:

- 1) Bray-Curtis dissimilarity
- 2) square root of Bray-Curtis
- 3) Euclidean distance
- 4) Orloci's Chord distance
- 5) Chi-square metric
- 6) Chi-square distance
- 7) Hellinger distance
- 8) Gower dissimilarity
- 9) Gower, excluding double zeros
- 10) Canberra distance
- 11) square root of Canberra distance
- 12) CY dissimilarity
- 13) Deviance based on the binomial
- 14) Deviance per observation (scale invariant)
- 15) Kulczynski dissimilarity
- 16) McArdle-Gower dissimilarity
- 17) Manhattan (city block) distance
- 18) Jaccard dissimilarity

1

Please wait while I calculate distances...

Type the name of the file containing the X matrix of regression (predictor) variables

(Note1: rows correspond to the n observations)

(Note2: cols have names < 8 char. in first row)

OkuraX.txt

What is the number of columns for this X matrix?'

15

Do you wish to test:

- 1) X variables individually, one at a time
- 2) X variables in specified sets

2

Note: We will start by considering the sets of variables and then go on to examine the variables individually in a separate second analysis.

Type the name of the file containing the variable set info

SetsOkura.txt

How many permutations do you want for the tests?
(i.e. 99, 499, 999, 4999, etc.)

999

Note that the method of permutation used by the program is permutation of raw data for the marginal tests (tests of individual sets or of individual variables), while for all conditional tests, the program uses permutation of residuals under a reduced model. See Anderson and Legendre (1999) and Anderson and Robinson (2001) for further details of these techniques in regression.

Type an integer to be used as the seed
for the random permutations

7

Calculating...

Please wait while I do the permutations...

This is for the marginals...

Working on perm. no.	1
Working on perm. no.	2
Working on perm. no.	3
etc...	

Once the program has finished doing the marginal tests, it will go on to do the conditional tests and the forward selection procedure.

Now doing the permutations for conditional term	2
Now doing the permutations for conditional term	3
Now doing the permutations for conditional term	4
etc...	

Results of the marginal and conditional tests are then both output to the screen and to the output file.

Program finished. Results are in the output file.

Type q to quit

q

In this case, the output file looks like this:

DISTLM forward

A program for analysing multivariate data
on the basis of any distance measure,
with forward selection of explanatory variables
in a linear regression model,
using permutations.

by M.J. Anderson
Department of Statistics
University of Auckland (2003)

Input file of data: OkuraY.txt

Input file of X matrix containing predictor variables: OkuraX.txt
 The no. of observations = 88
 The no. of variables = 73
 No. of permutations used = 999
 The integer chosen as the seed for the permutations = 7
 Data were transformed to $\ln(x+1)$
 No standardisation
 Analysis based on Bray-Curtis dissimilarities

RESULTS: MARGINAL TESTS

Variable	SS(Trace)	pseudo-F	P	prop
Ambient	3.9920	14.9149	0.0010	0.4182
Dep	3.4221	23.7499	0.0010	0.3585
Trapped	1.9331	7.1102	0.0010	0.2025
Erosion	1.1383	5.7539	0.0010	0.1192
Dist	1.8844	10.4536	0.0010	0.1974
Chla	0.1907	1.7529	0.1040	0.0200
Org	0.8708	8.6332	0.0010	0.0912

RESULTS: CONDITIONAL (SEQUENTIAL) TESTS

Variable	SS(Trace)	pseudo-F	P	prop	cumulative
Ambient	3.9920	14.9149	0.0010	0.4182	0.4182
Dep	1.0436	9.3711	0.0010	0.1093	0.5275
Trapped	0.6089	4.0581	0.0010	0.0638	0.5913
Erosion	0.6228	7.2192	0.0010	0.0652	0.6566
Dist	0.3220	4.0300	0.0010	0.0337	0.6903
Org	0.0955	2.4366	0.0040	0.0100	0.7003
Chla	0.0693	1.7862	0.0630	0.0073	0.7075

Other information in the output file includes the names of variables included in particular sets and the correlations between every pair of variables.

The first table in the output shows the relationship of each set of variables with the response data cloud, IGNORING all other variables. The second table shows the results of the forward selection procedure. In this case, it shows that the ambient grain size information explained the greatest proportion, at 41.8%. After fitting this, the Depositional contrasts were most useful. After fitting both Ambient and Depositional variables, the Trapped sediment information was most important, and so on. The above results indicated that each individual set of variables was important in explaining the variation in the species data, except for Chlorophyll a, which did not show a significant relationship with the species data even when considered alone. More particularly, the information from trapped sediments added a significant proportion of explained variation (although small) over and above the ability of ambient sediment to model the data. Overall, over 70% of the variability in the original distance matrix was explained by these sets of environmental variables together, omitting Chlorophyll a (i.e. see the second-to-last line in the "cumulative" column in the second table, above).

A second analysis was done of the above data, treating the variables individually, with the following results:

RESULTS: MARGINAL TESTS

Variable	SS(Trace)	pseudo-F	P	prop
GS1	2.4983	30.4873	0.0010	0.2617
GS2	2.3846	28.6364	0.0010	0.2498
GS3	2.1120	24.4337	0.0010	0.2213
GS4	2.1715	25.3242	0.0010	0.2275
HvML	3.0324	40.0381	0.0010	0.3177
MvL	0.3897	3.6602	0.0080	0.0408

Sdep	1.1838	12.1745	0.0010	0.1240
gt125	1.1144	11.3666	0.0010	0.1167
Perfin	1.3270	13.8861	0.0010	0.1390
BH	0.3748	3.5145	0.0040	0.0393
sdBH	0.6727	6.5205	0.0010	0.0705
D	1.5102	16.1626	0.0010	0.1582
D2	1.5830	17.0969	0.0010	0.1658
Ch1a	0.1907	1.7529	0.0950	0.0200
Org	0.8708	8.6332	0.0010	0.0912

RESULTS: CONDITIONAL (SEQUENTIAL) TESTS

Variable	SS(Trace)	pseudo-F	P	prop	cumulative
HvML	3.0324	40.0381	0.0010	0.3177	0.3177
GS4	0.7104	10.4056	0.0010	0.0744	0.3921
GS3	0.5627	9.0193	0.0010	0.0589	0.4510
BH	0.4028	6.9111	0.0010	0.0422	0.4932
Sdep	0.3485	6.3652	0.0010	0.0365	0.5297
D2	0.2447	4.6700	0.0010	0.0256	0.5554
D	0.2064	4.0900	0.0010	0.0216	0.5770
GS2	0.1925	3.9556	0.0010	0.0202	0.5972
MvL	0.2125	4.5615	0.0010	0.0223	0.6194
gt125	0.2149	4.8402	0.0010	0.0225	0.6419
Perfin	0.2490	5.9725	0.0010	0.0261	0.6680
GS1	0.1417	3.5108	0.0020	0.0148	0.6829
Org	0.0925	2.3314	0.0100	0.0097	0.6925
sdBH	0.0739	1.8860	0.0470	0.0077	0.7003
Ch1a	0.0693	1.7862	0.0570	0.0073	0.7075

High correlations among many of the variables (e.g., HvML and GS1, gt125 and Perfin, D and D2) helps to explain the choice of certain variables over others in the forward selection procedure. The correlation matrix for these variables, also output by the program is shown below:

CORRELATIONS AMONG VARIABLES

	GS1	GS2	GS3	GS4	HvML	MvL	Sdep	gt125	Perfin	BH	sdBH	D	D2	Ch1a	Org
GS1	1.0000														
GS2	0.5509	1.0000													
GS3	-0.9204	-0.4884	1.0000												
GS4	-0.4411	-0.8816	0.2436	1.0000											
HvML	0.8486	0.7951	-0.7850	-0.6651	1.0000										
MvL	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000									
Sdep	-0.5558	-0.2856	0.6135	0.2746	-0.4932	0.2003	1.0000								
gt125	-0.5269	-0.4811	0.5672	0.4474	-0.4309	0.0930	0.6549	1.0000							
Perfin	0.5281	0.4214	-0.5793	-0.3874	0.4433	-0.0303	-0.7028	-0.9601	1.0000						
BH	-0.1687	-0.2789	0.3004	0.2165	-0.2702	0.2427	0.3555	0.1235	-0.0423	1.0000					
sdBH	-0.2119	-0.0500	0.0369	0.2536	-0.3202	-0.3066	0.0828	-0.0481	-0.0503	-0.1347	1.0000				
D	0.4262	0.6848	-0.3638	-0.5259	0.4750	0.4190	-0.1650	-0.2950	0.2977	0.0205	0.1403	1.0000			
D2	0.4250	0.7147	-0.4189	-0.5055	0.5128	0.3153	-0.2398	-0.3358	0.3435	-0.0391	0.1129	0.9724	1.0000		
Ch1a	0.0634	0.0803	-0.1047	-0.0572	-0.0074	0.1086	-0.0020	-0.1164	0.0981	-0.1191	0.1030	0.0259	0.0158	1.0000	
Org	0.5903	0.1135	-0.6184	-0.0420	0.4147	0.0037	-0.3676	-0.2391	0.2641	-0.1799	-0.1217	-0.0563	-0.0398	0.0932	1.0000

V. References

- Anderson, M. J. and Legendre, P. (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.
- Anderson, M. J. and Robinson, J. (2001). Permutation tests for linear models. *Australian and New Zealand Journal of Statistics* 43: 75-88.
- Anderson, M.J., Ford, R.B., Feary, D.A. and Honeywill, C. (2004). Quantitative measures of sedimentation in an estuarine system and its relationship with intertidal soft-sediment infauna. *Marine Ecology Progress Series* 272: 33-48.

- McArdle, B.H. and Anderson, M.J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82(1): 290-297.
- Neter, J., Kutner, M.H., Nachtsheim, C.J. & Wasserman, W. (1996). *Applied linear statistical models*, 4th edition. Irwin, Chicago, Illinois.
- Press, W.H., Teuklosky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). *Numerical Recipes in FORTRAN*, 2nd edition. Cambridge University Press, Cambridge.