

DISTLM v.5

Distance-based multivariate analysis for a linear model

A computer program
by Marti J. Anderson



Department of Statistics
University of Auckland
(2004)

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. To obtain Monte Carlo samples from the asymptotic permutation distribution (if requested by the user), the function “ppchi2” is used, which is Algorithm AS 91 (Appl. Statist. (1975) Vol.24, p. 35). This also relies on the Numerical Recipes random number generator function “ran2”.

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

Anderson, M.J. 2004. DISTLM v.5: a FORTRAN computer program to calculate a distance-based multivariate analysis for a linear model. 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 is a computer program that calculates a multivariate multiple regression analysis of any symmetric distance matrix, including a test by permutation, as described by McArdle and Anderson (2001). Consider an $(N \times p)$ matrix of response variables \mathbf{Y} (N = the number of observational units and p = the number of variables). Consider also an $(N \times q)$ matrix, \mathbf{X} , which is of interest for a multivariate hypothesis. The purpose of DISTLM is to perform a permutational test for the multivariate null hypothesis of no relationship between matrices \mathbf{Y} and \mathbf{X} on the basis of any distance measure of choice, using permutations of the observations. Note that \mathbf{X} may contain the codes of an ANOVA model (a design matrix), or it may contain one or more explanatory (predictor) variables of interest (e.g. environmental variables).

To do the analysis, the first step is to let $\mathbf{D} = (d_{ij})$ be an $(N \times N)$ distance matrix calculated from observation units of \mathbf{Y} , using some chosen appropriate distance measure. Let $\mathbf{A} = (a_{ij}) = (-\frac{1}{2}d_{ij}^2)$, then calculate Gower's (1966) centered matrix (\mathbf{G}) by centering the elements of \mathbf{A} , i.e.

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)$$

where $\mathbf{1}$ is a column of 1's of length N and \mathbf{I} is an $(N \times N)$ identity matrix. Next, we calculate the "hat" or projection matrix $\mathbf{H} = \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$ (e.g. Johnson and Wichern 1992). Then, an appropriate test statistic associated with the null hypothesis is the pseudo F -statistic:

$$F = \frac{tr(\mathbf{HGH})}{tr[(\mathbf{I} - \mathbf{H})\mathbf{G}(\mathbf{I} - \mathbf{H})]}.$$

A P -value is then obtained by recalculating this statistic for a large number of random re-orderings of the observations (i.e. the rows and columns of \mathbf{G}), while keeping \mathbf{X} (and \mathbf{H}) constant. If the analysis is done on only one response variable and the Euclidean distance is used, then the above statistic is the traditional univariate F -statistic, although the program will still perform the test using permutations, rather than relying on tabled P -values. For more details and background to this theory, see McArdle and Anderson (2001), Anderson (2001a) and Legendre and Anderson (1999).

II. More Complicated Models and Designs

The above description is for the simplest scenario of only one predictor matrix, \mathbf{X} . There are several situations where more complicated analyses are necessary. Indeed, one of the nice features of this program is that it can be used for tests in the presence of covariables, tests of unbalanced ANOVA designs, and tests of individual terms in multi-factorial ANOVA.

In the context of multiple regression, one may wish to control for covariables in the model. A third matrix, \mathbf{X}_C , containing covariables, can be specified for this situation. In this case, one then has the further option of performing the permutation test using (i) unrestricted permutation of raw data, (ii) permutation of residuals under a reduced model or (iii) permutation of residuals under a full model. These three options are asymptotically equivalent and should give similar results. For more information and a comparison of these methods of permutation, see Anderson and Legendre (1999) and Anderson and Robinson (2001).

In the context of analysis of variance, a test for an unbalanced design can be done by simply coding the relevant \mathbf{X} matrix accordingly. For information on constructing \mathbf{X} matrices with appropriate codes for ANOVA designs (including interactions, etc.) see Appendix C in Legendre and Anderson (1999) or Neter et al. (1996).

Some terms in a multi-factorial ANOVA do not use the residual mean square as the denominator of the F -ratio for their test, but use the mean square of some other term in the model. The multivariate analogue also

has this requirement in a directly analogous fashion (McArdle and Anderson 2001). Although the program PERMANOVA_2factor (Anderson 2004) will perform the correct tests for all terms in any two-way design, many experiments have more than two factors. There is currently no existing software to deal with this issue other than DISTLM. In the ANOVA context, DISTLM will allow you to enter a third matrix (\mathbf{X}_{denom}) which contains the codes for the denominator term for the individual test you are performing. As an example, consider that you have a design with two factors: A and B, nested in A. When you wish to do the test for factor A, you need to construct the pseudo F -ratio as follows:

$$F = \frac{tr(\mathbf{H}_A \mathbf{G} \mathbf{H}_A)}{tr(\mathbf{H}_B \mathbf{G} \mathbf{H}_B)}$$

where \mathbf{H}_A is the hat matrix corresponding to the \mathbf{X} matrix that codes for factor A and \mathbf{H}_B is the hat matrix corresponding to the \mathbf{X} matrix that codes for factor B.

III. Input files

The program allows the user to input a distance matrix \mathbf{D} directly or a raw multivariate data matrix of response variables \mathbf{Y} . In either case, the file should be saved as ASCII *.txt, with no column or row headings. If you are using a Macintosh, then save the file (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.

Either the \mathbf{Y} or \mathbf{D} matrix corresponding to the response variables will be 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.

Note 1:

For Regression, other input files include:

- a) an \mathbf{X} matrix containing one or more predictor variables of interest
- b) as an option, an \mathbf{X}_C matrix containing one or more covariables, if required.

Note 2:

For ANOVA, other input files include:

- a) an \mathbf{X} design matrix corresponding to the ANOVA term of interest
- b) an \mathbf{X}_{denom} design matrix corresponding to the ANOVA term that should form the denominator mean square for the test, if necessary (i.e. if the residual mean square is NOT the correct denominator for the test).
- c) as an option, an \mathbf{X}_C matrix containing one or more covariables, if required.

If the residual mean square is to be the denominator, then one must either provide a further file (a fourth matrix) with the design matrix (\mathbf{X}_{full}) that corresponds to the full model, or simply specify that there are no other terms in the model than the one being tested.

Note 3:

The program requires you to choose a desired method of permutation for the test of interest, which will include one or more of the following options:

- a) Unrestricted permutation of raw data
- b) Permutation of residuals under a reduced model
- c) Permutation of residuals under a full model

If one chooses either b or c , there are situations where one must provide an input file with a matrix that codes for either the full model (\mathbf{X}_{full}) or the reduced model ($\mathbf{X}_{reduced}$). Whether or not this extra matrix will be needed for the analysis depends on whether the information already given about the model and the term being tested is sufficient to determine these or not.

Note 4:

For any of the above \mathbf{X} input matrices, variables must be columns and observation units must be rows. The file must be saved in ASCII *.txt format with no headers or labels of any kind. The user will specify the number of columns, while the number of rows must be equal to the number of observational units already specified for \mathbf{Y} (or \mathbf{D}). In the case of an ANOVA design, the number of columns of each \mathbf{X} matrix will correspond to the degrees of freedom associated with that term in the analysis. \mathbf{X} matrices that are not of full rank will simply choke the program (yielding the error message “singular matrix” when the program tries to invert $\mathbf{X}'\mathbf{X}$), so make sure the coding is efficient! Also, do not include an intercept in any of the \mathbf{X} matrices (i.e. a column containing all 1's), as the program already provides for this.

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.exe (or DISTLM.mac) file, for use with the program. Double-click on the “DISTLM.exe” (or “DISTLM.mac”) file to run the program.

IV. Changes from the Previous Version

- DISTLM version 5 includes the option to include covariables in any ANOVA model or Regression model.
- Additional error checking and appropriate error messages have been added, making the program a wee bit more friendly.
- DISTLM version 5 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.
- DISTLM version 5 uses routines from Applied Statistics to generate random chi-square variables.

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.

- A more efficient algorithm is used to calculate the test-statistic, which has increased the speed of the program considerably.
- More information is given in the output file concerning choices made for the procedure and the nature of the permutation test done.
- In the case of the test of an ANOVA term, there is a new additional option to obtain a P -value for the test from a random Monte Carlo sample from the asymptotic distribution of the pseudo F -statistic under permutation. The nature of this distribution depends on the particular data set at hand, the distance measure chosen and the nature of the hypothesis. This can be particularly useful in the case of a test where the number of unique permutations is limited because of the number of permutable units. The correct permutable units for any particular term in an ANOVA design are dictated by the denominator used to construct the F -statistic; see Anderson (2001b) and Anderson

and ter Braak (2003). In these situations, a valid P -value can be obtained using the Monte Carlo approach. The theoretical details of this distribution under permutation are given in Anderson and Robinson (2003).

- There is an additional option to output the values of the numerator and denominator under permutation. This provides a method of constructing, for example, linear combinations of mean squares under permutation, if required (i.e. by using the same seed and running the program several times). This is useful in the case of an ANOVA where there is not a single term that can be used as the denominator for the test of the term of interest (e.g. see Searle et al. 1992).

V. 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 survey of prawns across the northern Great Barrier Reef (Anderson and Gribble 1998). The variables consisted of biomass recorded for 14 species of prawns from trawls at 126 stations. Environmental variables thought to be important in structuring assemblages of prawns were also recorded, including habitat type, percentage of mud, mean water depth, rugosity of the sea floor and fishing effort. In addition, temporal variables thought to be important in structuring prawn distributions were recorded, including daily variation (Julian date), lunar periodicity and diel periodicity. See Anderson and Gribble (1998) for details. The hypothesis of interest (examined below) was to test for a significant relationship between the set of environmental variables and the multivariate prawn biomass, given the temporal variability inherent in conducting such a large survey.

The response variable matrix Y (126 rows x 14 columns) was contained in an ASCII text file called "Species.txt". The file containing the matrix of environmental variables of interest for the test (126 rows x 5 columns) was called "Env.txt." The file containing the matrix of temporal covariables (126 rows x 3 columns) was called "Time.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 v.5

 A program for analysing multivariate data
 on the basis of any distance measure,
 according to any linear ANOVA or regression model,
 using permutations.

by M.J. Anderson
 Department of Statistics
 University of Auckland (2004)

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

Species.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?

14

How many observations (rows) are there?

126

Choice of transformation:

- 1) none
- 2) square-root
- 3) fourth-root
- 4) ln(x)
- 5) ln(x+1)
- 6) log10(x)
- 7) log10(x+1)
- 8) presence/absence

1

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 (normalise)
- 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

6

Note: The chi-square distance is a reasonable choice for these data, which spanned a large geographic area and were expected to show unimodal response curves across the Great Barrier Reef.

Do you want to output the distance matrix?

- 1) Yes
- 2) No

2

Note: This is usually not necessary unless you want to see particular distance values.

Do you wish to output the mean squares for the numerator and denominator under permutation?

- 1) No
- 2) Yes

1

Is this test for an ANOVA design or a regression model?

- 1) Regression model
- 2) ANOVA design

1

Please wait while I do a few preliminary calculations...

Type the name of the file containing the X matrix of regression (predictor) variables (and for which rows correspond to the n observations)

Env.txt

What is the number of columns for this X matrix?

5

Are there covariables in the model?

- 1) No
- 2) Yes

2

Type the name of the file containing the covariables

Time.txt

How many columns does it have?

3

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

9999

Which general method of permutation do you want?

- 1) Permutation of residuals (reduced model)
- 2) Permutation of residuals (full model)

1

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

5

Note: Any random integer seed will do here. If you wish to re-calculate the exact same permutation test at a later date (i.e. with the exact same set of re-orderings), make note of this seed and use it again.

Do you wish to permute units other than the individual observation units (i.e. groups of units)?

- 1) No
- 2) Yes

1

Calculating...

Please wait while I do the permutations...

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

End of permutations.
 The results have been sent to the output file.
 End of the program.
 Press q to quit.

q

When the program has completed the analysis, it will print the results on the screen. It will simultaneously send the results to the named output file, with more details concerning the choices you made for the analysis, for you to look at more closely. In this case, the output file looks like this:

```

DISTLM v.5
-----
A program for analysing multivariate data
on the basis of any distance measure,
according to any linear ANOVA or regression model,
using permutations.

by M.J. Anderson
Department of Statistics
University of Auckland (2004)

Input file of data: Prawns.txt
Input file of X matrix containing predictor variables: Env.txt
Input file of X matrix containing covariables: Time.txt

The no. of observations = 126
The no. of variables = 14

--- Results ---
*Regression SS, Residual SS and Residual df were calculated
  after removing SS and df due to covariables*
-----
Covariables      df      Sum of squares      Mean square
Regression       5      150.87002           30.17400
Residual        117     677.97962           5.79470
Total           125     887.95024
-----
                                pseudo-F = 5.20717
                                permutation P = 0.00010
-----
No transformation
No standardisation
Analysis based on Chi-square distances
Permutation of residuals(reduced model)
No. of permutations used = 9999
The integer chosen as the seed for the permutations = 5
Proportion of variation explained = 0.1699

```

These results indicate that the chosen set of environmental variables together explain a significant proportion of the multivariate variation in the species data, even after the temporal variability at various scales is taken into account in the form of covariables in the analysis.

VI. References

- Anderson, M.J. (2004). PERMANOVA_2factor: a FORTRAN computer program for permutational multivariate analysis of variance (for any two-factor ANOVA design) using permutation tests. Department of Statistics, University of Auckland, New Zealand.
- Anderson, M.J. (2001a). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26: 32-46.
- Anderson, M.J. (2001b). Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences* 58: 626-639.
- Anderson, M.J. and Gribble, N.A. (1998). Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. *Australian Journal of Ecology* 23: 158-167.
- 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. and Robinson, J. (2003). Generalised discriminant analysis based on distances. *Australian & New Zealand Journal of Statistics* 45(3): 301-318.
- Anderson, M.J. and ter Braak, C.J.F. (2003). Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* 73: 85-113.
- Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53: 325-338.
- Johnson, R.A. & Wichern, D.W. (1992). *Applied multivariate statistical analysis*, 3rd edition. Prentice-Hall, Englewood Cliffs, New Jersey.
- Legendre, P. and Anderson, M.J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69: 1-24.
- 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.
- Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance components*. John Wiley and Sons, Toronto.