TWEDA USERS MANUAL

by

R. D. Cook, J. Huang, L. Thibodeau,
S. Weisberg

Technical Report No. 437
July 1984

University of Minnesota
School of Statistics
St. Paul, Minnesota

TWEDA USER'S MANUAL, VERSION 2.0

by

R. D. Cook, J. Huang, L. Thibedeau, S. Weisberg

Revised, July 1984

Introduction

TWEDA is an ancronym for two-way exploratory data analysis. This is a program for the analysis of interactions in a two-way unreplicated layout. The program includes several models for nonadditivity, and score tests for the need to transform and for constant variance. Most of the techniques included in TWEDA are outlined in Cook and Weisberg (1983) and their book (1982, section 2.5), which is based on an earlier report written in 1975 by R. D. Cook. The program was written by R. D. Cook and by L. A. Thibodeau, with some assistance from Christopher Bingham. The current version of the program was modified by Joanna Huang and S. Weisberg in 1984.

The program is written in MNF FORTRAN. On the University of Minnesota MERITSS computer (ME), it is accessed by typing:

```
X,GET,TWEDA/UN=2051999
X,TWEDA
```

For use on MIRJE (CA), please contact S. Weisberg, 373-1068.

Disclaimer. TWEDA has been tested for accuracy, and to the best of our knowledge, is correct. However, neither the University of Minnesota nor any of the authors claim any reponsibility for any errors, or for their consequences.

1. General Information

1.1 Nonadditivity Models and Analysis

Consider an r×c two-way table of response variables y_{ij}, i = 1, 2, ..., r; j = 1, 2, ..., c. Then the additive model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (1)$$

where μ is a grand mean, α_i is a row effect, β_j is a column effect and $\Sigma\alpha_i = \Sigma\beta_j = 0$. Further, the ε_{ij} are uncorrelated errors with mean zero and unknown finite variance σ^2.

Non-additivity analysis is based on the addition of interaction terms to the additive model. For example, the most common and simplest method of analysis was proposed by Tukey (1949) and results from the addition of an interaction term of the form $\tau\alpha_i\beta_j$. The model with this term added is referred to as Tukey's model:

$$y_{ij} = \mu + \alpha_i + \beta_j + \tau\alpha_i\beta_j + \varepsilon_{ij} \ .$$

The term $\tau\alpha_i\beta_j$ has a single degree of freedom associated with it and thus this analysis is referred to as Tukey's one-degree of freedom for non-additivity. Johnson and Graybill . (1972) also provided valuable contributions to the analysis of a generalization of this model.

Mandel (1969, 1971) discussed several models for non-additive data, each allowing several degrees of freedom for interaction. Two important special cases are the row-model (row-regression-model):

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_j\alpha_i + \varepsilon_{ij}$$

and the column-model (column-regression-model):

$$y_{ij} = \mu + \alpha_i + \beta_j + \delta_i\beta_j + \varepsilon_{ij} \quad .$$

The parameters γ_j and δ_i are referred to as the column and row slopes, respectively. It can be shown that the terms $\gamma_j\alpha_i$ and $\delta_i\beta_j$ are orthogonal on thus the combined or slopes-model

$$y_{ij} = \mu + \alpha_i + \beta_j + \gamma_j\alpha_i + \delta_i\beta_j + \varepsilon_{ij}$$

is also of interest.

Initially, concern usually centers on testing one or more of the hypotheses:

Tukey's: H_0: $\tau = 0$,

Row: H_0: $\gamma_j = 0$ for all j,

Column: H_0: $\delta_i = 0$ for all i.

All of these hypotheses lead to F-tests under H_0, but the alternative distributions are very complex; see Millikan and Graybill (1970).

If a hypothesis is rejected then the analyst is usually interested in the reason for the rejection. Discussions of these reasons may be found in Cook and Weisberg (1982), Johnson and Graybill (1972), Mandel (1971) and Tukey (1949, 1965).

Johnson and Graybill (1972) suggested an alternative form for modelling

interaction via the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \delta w_i u_i + \varepsilon_{ij} \qquad (2)$$

where $w^T = (w_1, \ldots, w_r)$ and $u^T = (u_1, \ldots, u_c)$ are latent row and column variables, $u^T u = w^T w = 1$. As a matrix expression, if Y is an r × c matrix if $\alpha^T = (\alpha_1, \ldots, \alpha_r)$, $\beta^T = (\beta_1, \ldots, \beta_c)$, we can write

$$Y = \mu 1 1^T + \alpha 1^T + 1\beta^T + \delta w u^T + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{\varepsilon}$ is also r × c. We see that this model requires interaction to be of the form of a rank one matrix. The estimates of δ, w, u simply correspond to the best rank one approximation to the matrix $R = (r_{ij})$ of residuals from the additive model: $\hat{\delta}$ is the largest singular value of R, and $\hat{w}$, $\hat{u}$ are the corresponding left and right singular vectors. A test of $\delta = 0$ is given by $\Delta = \hat{\delta}^2/tr(R^T R)$; critical values are given by Cook and Weisberg (1982 Table 2.5.2). The hypothesis of $\delta = 0$ is rejceted for large Δ. These statistics are computed by the DECOMP command.

TWEDA provides test statistics for all of the above hypotheses as well as graphical, semi-graphical, and numerical devices to illuminate the source of non-additivity.

1.2 Heteroscedasticity (command HSCORE)

The non-additivity models modify the additive model (1) by changing mean structure. Alternatively, we could contemplate modifying (1) by assuming a heteroscedastic model,

$$\mathrm{Var}(\varepsilon_{ij}) = \sigma^2 e^{\lambda^T h_{ij}}$$

for a suitably defined vector h_{ij}, and for an unknown parameter vector λ. Clearly, if $\lambda = 0$, we get constant variance. TWEDA includes 4 score tests for constant variance, essentially as follows:

(1) ROWS: $h_{ij} = h_{i.}$ = vector of all zeros except a one in the i-th place---tests for different variance in each row.

(2) COLS: $h_{ij} = h_{.j}$ = vector of all zeros except a one in the j-th place---tests for different variance in each column.

(3) ROWS + COLS: h_{ij} = vector of length r + c, concatenation of $h_{i.}$ to $h_{.j}$ tests for separate variances in each cell.

(4) FITTED VALUES tests

$$\mathrm{Var}(\varepsilon_{ij}) = e^{\lambda E(y_{ij})}$$

The methodology for the score test is given by Cook and Weisberg (1983). A graphical equivalent is available in the PLOT command.

1.3 Transformations (command LSCORE)

TWEDA computes Atkinson's (1971) score test, t_D for the Box-Cox model

$$y_{ij}^{(\lambda)} = \mu + \alpha_i + \beta_i + \varepsilon_{ij}$$

where $y^{(\lambda)}$ is a power family. Details are given by Cook and Weisberg (1982, Sec 2.4). An added variable plot (ibid., sec. 2.4.3) is also

available in PLOT.

1.4 Data Files

TWEDA works on an array of numbers with r rows and c columns. The data file must have exactly r rows and c columns. In addition, the file must be a formatted file as follows:

1. The first line of the file consists of an arbitrary identifier (in columns 1 to 10), then the number of rows (ending in column 19), and then the number of columns (ending in column 29). The word FORMAT may appear after the number of columns.

2. The second line of the file gives the format statement, beginning in column 1. Only E, F, G, O, and X formats are permitted.

3. The data begin on the third line of the file.

Files written by the program MULTREG can be read directly by TWEDA. Files written by the program MATTER cannot be read directly, since the first line of the file is in a slightly different format. Thus, the first line of a written by MATTER must be changed either by (1) reading the file into MULTREG and writing a new file, or (2) by modifying the first line using an editing program.

TWEDA is limited to tables of size 40 by 40 maximum, and 3 by 3 minimum. A printer with 132 columns is recommented for tables with more than 25 columns.

2. Computations

2.1 Models and Analysis

Several different forms of calculations are of available: the sums-of-squares for the additive, row, column, and Tukey's models and the singular

value decompositim of R for the Johnson-Graybill approach. Cook and Weisberg (1982) gives a detailed explanation of the computational method used for each term.

The singular value decomposition of R uses a Jacobi threshold algorithm from the IMSL(1975) package. The algorithm has been shown to be extremely accurate and efficient but could be replaced by any algorithm for the singlar values and vectors of a non-negative definite matrix, such as SSVDC in LINPACK.

2.2 Tests for Heteroscedasticity and Estimated Transformation

Let $U_{ij}^2 = r_{ij}^2/\Sigma\Sigma r_{ij}$, where the r_{ij} are residuals from the additive model. Four tests for heteroscedasticity are available. The basic technique is to regress U_{ij} on h_{ij}. Then, 1/2 (sum of squares for regression of U_{ij} on h_{ij}) is asympotically distributed as χ^2 with q degrees of freedom where q is the dimension of h_{ij}.

For Atkinson's test, set

$$g_{ij} = -y_{ij}\,[1-\log(y_{ij}/\dot{y})] + \log(\dot{y}) + 1$$

where $\dot{y}$ is the geometric mean of y_{ij}'s. Then fit an additive model for g_{ij}. Let $r_{ij}(g)$ be the residuals from this model. Next, fit the regression of r_{ij} on $r_{ij}(g)$,

$$r_{ij} = \alpha + \beta r_{ij}(g) + \xi_{ij} \quad .$$

The usual t-statistic for testing $\beta = 0$, say t_D, is the test for transformation t_D is asympotically N(0,1). A quick estimate of the

transformation; is to raise y_{ij} to power $1-\hat{\beta}$ or to take logarithms if $\hat{\beta}$ is near 1.

3. Commands

TWEDA is an interactive program and provides analysis as requested, so only the options entered will be computed and printed. All options are order independent. The example given here is also given by Cook and Weisberg (1982, p.96), who give a fairly complete analysis; it is used here only to illustrate most of the commands available. The example has 7 rows (treatments) and 3 columns (soils). The data set is given by:

```
SOURCEDATA        7        3  FORMAT
(F4.1,2F5.1)
11.1 32.6 63.3
15.3 40.8 65.0
22.7 52.1 58.8
23.8 52.8 61.4
25.6 63.1 41.1
31.2 59.5 78.1
25.8 55.3 60.2
```

To initate the analysis respond to NEXT? with any of the following commands.

3.1 HELP

Lists commands with a short explanation of options.

3.2 LIST

Lists the current data matrix. It is recommended that the data be listed upon entry and after each modification using CHANGE.

```
x,tweda
T W E D A  VER. 1.1 84/05/08. 14.00.28.
NAME YOUR DATA FILE
? agdata
TABLE HAS  7 ROW AND  3 COLUMNS.
FOR HELP, TYPE HELP

       NEXT? list
```

11.100	32.600	63.300
15.300	40.800	65.000
22.700	52.100	58.800
23.800	52.800	61.400
25.600	63.100	41.100
31.200	59.500	78.100
25.800	55.300	60.200

3.3 ANOVA

Results in the computation and output of an analysis of variance table for the slope-model and for the additive-model. The F-statistics given are for the apporpriate mean square divided by the residual mean square from the slope-model. The slope-model F-statistics are for the hypotheses of no significant regression; H_0: $\gamma_j = 0$ for all j in the row-model, H_0: $\delta_i = 0$ for all i in the column-model. The F-statistic corresponding to "TUKEY" is for the one-degree of freedom for non-additivity test; H_0: $\tau = 0$. The interpretation of "ROW" and "COLUMN" effects in the presence of significant non-additivity may not be straightforward.

```
    NEXT? anova
            TWO-WAY ANALYSIS OF VARIANCE
SOURCE        D.F.     SS              MS              F     P-VALUE
-------------------------------------------------------------------
ADDITIVE MODEL
ROWS             6    731.05810     121.84302          .882   .5664
COLUMNS          2    5696.3400     2848.1700        20.610   .0038
RES(ADD)        12    947.43333     78.952778
-------------------------------------------------------------------
NON-ADDITIVITY MODELS
COL-MODEL        5    228.40484     45.680968          :331   .8751
ROW-MODEL        1    26.922495     26.922495          .195   .6774
TUKEY MODEL      1    1.1350509     1.1350509          .008   .9313
RESIDUAL         5    690.97095     138.19419
-------------------------------------------------------------------
TOTAL           20    7374.8314
GRAND AVERAGE         44.742857
```

From this table, we have no reason to suspect interaction of the types given by the models used in this program, as all the F-tests are small.

3.4 ESTIMATE

Prints the maximum likelihood (least squares) estimated coefficients from the additive-model. This command also prints the standard error of estimated effect and standard error of difference of estimated effects.

```
      NEXT? estimate

   ROW   ADDITIVE ROW EFFECT
    1     -9.076190
    2     -4.376190
    3     -.2095238
    4      1.257143
    5     -1.476190
    6      11.52381
    7      2.357143

  SE(EFFECT) =    5.130068
  SE(DIFF)   =    7.255011


  COLUMN   ADDITIVE COLUMN EFFECT
    1     -22.52857
    2      6.142857
    3      16.38571

  SE(EFFECT) =    3.358418
  SE(DIFF)   =    4.749520

  GRAND AVERAGE     =  44.742857
  SE(GRAND AVERAGE) = 1.93898326
```

3.5 RESID

Prints the residual matrix R from the additive-model. These are not Studentized residuals. Only residuals from the additive model are available. The statistic $\max(r_{ij})/[(r-1)(c-1)\hat{\sigma}^2]^{1/2}$ is used to test for a single outlier; critical values are given by Cook and Weisberg (1982, p. 88).

```
     NEXT? resid
-2.0381    -9.2095·    11.248
-2.5381    -5.7095     8.2476
```

```
  .69524     1.4238    -2.1190
  .32857      .65714    -.98571
  4.8619     13.690    -18.552
-2.5381    -2.9095     5.4476
  1.2286     2.0571    -3.2857
```

3.6 RTABL

Prints a semi-graphical, symbolic, standardized residual table. The residuals from the additive model in R are standardized by dividing by the estimate of σ^2 from the slope-model. The symbols are determined by the following code:

0 to 1.28	+	-1.28 to -0.0	-
1.28 to 2.33	++	-2.33 to -1.28	--
2.33 to 3.0	H	-3.00 to -2.33	L
3 or more	HI	-3 or less	LO

```
     NEXT? rtabl
-    -    +
-    -    +
+    +    -
+    +    -
+    +   --
-    -    +
+    +    -
```

3.7 DECOMPOSE

This command computes and prints the singular values and singular vectors for the matrix of residuals, R. These singular values and corresponding vectors are, as noted above, the MLE (LSE) of the parameters in the Johnson-Graybill model. This command also results in the printing of a statistic "LAMBDA" which is the largest singular value divided by the sum of squares of the singular values. This hypothesis of $\theta = 0$ in the Johnson-Graybill model is rejected for large values of LAMBDA; critical values are given by Cook and Weisberg (1982, p. 94).

```
      NEXT? decomp

   SINGLE VALUE DECOMPOSITION OF THE MATRIX RESIDUALS

   SINGULAR VALUE          RIGHT SINGULAR VECTOR(COLS)
   2.0957        -.79007    .57347     .21660
   30.709        -.20604   -.58120     .78724

   LAMDA .99536

SINGULAR VALUE            LEFT SINGULAR VECTOR(ROWS)
2.0957    .58928  -.24691  .09150      .04592      .00408  -.72372  .23984
30.709    -.47631 -.33652  .85935E-01  .39911E-01  .76733  -.21175  .13141
```

3.8 PLOT

Plot allows the user to display many of the statistics computed by TWEDA in scatter plots. Plots are specified by entering the command PLOT, and then responding to questions concerning choice for the vertical or y-axis and the horizontal or x-axis. The following plots are available:

Residuals versus fitted values: r_{ij} vs $\hat{y}_{ij}$

Residuals versus Tukey: r_{ij} vs $\hat{\alpha}_i\hat{\beta}_j$

Residuals versus Rankit: Normal probability plot of r_{ij}

Residuals versus row: r_{ij} vs $\hat{\alpha}_i$. The plotted quantities are the letters A (for cells in column 1), B (for cells in column 2), etc.

Residuals versus col: r_{ij} vs $\hat{\beta}_j$. The plotted quantities are the letters A (for cells in row 1), B (for cells in row 2), etc.

Residuals versus LSCORE: The added variable plot for the score vector for determining a transformation of the response.

SQRES versus ROW: Nonconstant variance plot as a function of row indicators.

SQRES versus COL: Nonconstant variance plot as a function of column indicators.

SQRES versus BOTH: Nonconstant variance plot as a function of row and

column indicators.

SQRES versus FITTED: Nonconstant variance plot as a function of the fitted response.

Singular vector versus ROW: Left singular vector vs the corresponding row effects from the additive model. Plotted quantities are the letters A, B, etc., correspoonding to column numbers 1, 2,.... Usually, the singular vector corresponding to the largest singular value is of interest.

Singular vector versus COL: Right singular vector vs the corresponding column effects from the additive model. Plotted quantities are the letters A, B, etc., corresponding to row numbers 1, 2, ... Usually, the singular vector corresponding to the largest singular value is of interest.

3.9 CHANGE

Allows for the editing of data and for the replacement of observations by generated values for specified models. No more than three new values may be generated at any time. There is no limit on the number of values edited. The generated values are obtained by numerically minimizing the residual sum of squares in the model specified. The model for the values may be specified by:

ADD - Additive-model,

ROW - Row-model,

COL - Column-model,

SLP - Slope-model.

As the generation of K, K = 1, 2, 3, replacement values requires the solution of K nonlinear equations in K unknowns the convergence to a solution cannot be assured. Care must be taken to insure that

1) The answers are reasonable.

2) The RSS for the model specified is, in fact, less than the original RSS.

If either of these conditions are not statisfied, the method of solution has not converged. This may be remedied by using better starting values for the replacement values. (A plot of residuals versus estimated row or column effects may be helpful.) To check the generated values type CHANGE using the generated values as starting values.

```
     NEXT? change
ENTER ROW AND COLUMN? 7 3
DO YOU HAVE A REPLACEMENT VALUE, YES OR NO.       ? no
IS THAT ALL, YES OR NO.       ? yes
WHICH MODEL:(ADD,ROW,COL,OR SLP),       ? add

       NEXT? list
  11.100      32.600      63.300
  15.300      40.800      65.000
  22.700      52.100      58.800
  23.800      52.800      61.400
  25.600      63.100      41.100
  31.200      59.500      78.100
  25.800      55.300      65.957
```

3.10 RENEW

Restores the data matrix to its original value.

3.11 PARTI

Allows the selection of a subtable of the original table for analysis. Each time PARTI is used the data are restored to their original values. The user will be asked to supply the rows and columns for the subtable of interest.

3.12 HSCORE

Results in the computation and output for a tests for non-constant variance table. Four tests are available in this table. The score test statistics given are asympotically distributed as χ^2, with df as shown.

```
     NEXT? hscore
                TESTS FOR NON-CONSTANT VARIANCE
```

MODEL	D.F.	SCORE TEST	P-VALUE
ROWS	6	22.993781	.0008
COLUMNS	2	5.1495147	.0762
ROWS+COLS	8	28.143296	.0004
FITTED VALUES	1	2.2746408	.1315

3.13 LSCORE

Prints Atkinson's score test, t_D, the asymptotic P-value and a quick estimate of a suggested transformation.

```
    NEXT? lscore
ATKINSON'S SCORE TD =  1.61     ASYMPTOTIC P-VALUE =   .1079
SUGGESTED TRANSFORMATION IS   .1136 POWER
```

3.15 END

Terminates the program.

5. References

[1] Andrews, D. F. and J. W. Tukey (1973). Teletypwriter plots for data analysis can be fast: 6-line plots including probability plots. Applied Statistics 22, 192-203.

[2] Anscombe, F. J. and John W. Tukey (1963). The examination and analysis of residuals, Technometrics 5, 141-160.

[3] Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression, New York and London: Chapman-Hall.

[4] Graybill, R. A. (1969). Introduction to Matrices with Applications in Statistics, Belmont, CA: Wadsworth Publishing Co.

[5] Hegemann, V. and D. E. Johnson (1976). On analyzing two-way AoV data with interation. Technometrics 18, 273-282.

[6] Hegemann, V. and D. E. Johnson (1976). The power of two tests for non-additivity. JASA, 71, 945-948.

[7] Johnson, D. E. and F. A. Graybill (1972). An analysis of a two-way model with interaction and no replication, JASA, 67, 862-868.

[8] Johnson, D. E. (1976). Some new multiple comparison procedures for the two-way AoV model with interactions. Biometrics, 32, 929-934.

[9] Mandel, John (1960). The partitioning of interaction in analysis of variance, Journal of Research of the Bureau of Standards - B. Mathematical Sciences, 73B, 309-328.

[10] Mandel, John (1970). The distribution of eigneevalues of covariance matrices of residuals in analysis of variance, Journal of Research of the Bureau of Standards - B. Mathematical Sciences, 74B, 149-154.

[11] Mandel, John (1971). A new analysis of variance model for non-additive data, Technometrics, 13, 1-18.

[12] Tukey, John W. (1949). One degree of freedom for non-additivity, Biometrics, 5, 232-242.

[13] Tukey, John W. (1965). Introduction to Exploratory Data Analysis, Limited edition, experimental version, Princeton University.

[14] Weisberg, S. and C. Bingham (1975). An Approximate Analysis of Variance Test for Non-normality Suitable for Machine Calculation. Technometrics, 17, 133-134.

[15] Yates, F. (1970). Experimental Design-Selected Papers, Griffin, London.