Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

A99.9
57644
Cp. 2

EVENT PROBABILITIES ESTIMATED BY REGRESSION

David A. Hamilton, Jr.

YES

NO

BAD

CONDITION

GOOD

USDA Forest Service Research Paper INT-152, 1974
INTERMOUNTAIN FOREST & RANGE EXPERIMENT STATION
Ogden, Utah 84401

THE AUTHOR

DAVID A. HAMILTON, JR., is a Research Forester working in timber measurements and management planning research at the Forestry Sciences Laboratory, Moscow, Idaho. Since receiving his Ph.D. in forestry from Iowa State University in 1970, he has been working on the sampling and modeling of forest mortality.

USDA Forest Service
Research Paper INT-152
May 1974

EVENT PROBABILITIES ESTIMATED BY REGRESSION

David A. Hamilton, Jr.

INTERMOUNTAIN FOREST AND RANGE EXPERIMENT STATION
Forest Service
U.S. Department of Agriculture
Ogden, Utah 84401
Roger R. Bay, Director

CONTENTS

	Page
INTRODUCTION	1
APPLICATIONS OF MODIFIED PROGRAM.	3
Mortality	3
Cull Estimation	4
Canker Inactivation	4
SAMPLING ASSUMPTIONS OF ALGORITHM	5
WALKER-DUNCAN ALGORITHM	6
MODIFICATION OF ALGORITHM.	8
STATISTICAL CRITERIA FOR GOODNESS-OF-FIT	10
GUIDES FOR USE OF RISK	12
Control Cards	12
Data Preparation	15
Program Output	16
Error Diagnostics	16
DISCUSSION AND CONCLUSIONS	18
LITERATURE CITED	18

ABSTRACT

Describes an algorithm for fitting the relationship between a dichotomous dependent variable and a number of independent variables. The specific functional form utilized is the logistic function. The algorithm, proposed and programmed by Walker and Duncan (1967), has been generalized by adding a subroutine to handle data transformations and by introducing procedures to handle data sampled with unequal probability. A user's guide discusses makeup of program control cards, data transformation capabilities, program output, and program-generated errors.

INTRODUCTION

The procedure discussed in this paper evolved from efforts to develop a model for estimating the probability of mortality for an individual tree, as a function of individual tree and crown characteristics. However, the algorithm discussed may be applied to describe the relationship between dependent and independent variables when the dependent variable is dichotomous.

In mathematics, a dichotomous variable is one that may assume only one of two distinct values. For estimating the probability of individual tree mortality, the dichotomous dependent variable assumes the value 1 when a tree dies during the measurement interval and otherwise assumes the value 0. The independent variables are diameter at breast high (d.b.h.), crown ratio, crown width, tip character, and insect or disease damage.

A dichotomous dependent variable may pose difficulties in the analysis of data.^{1/} The model frequently applied in this situation is a linear regression model of the form

$$E(y|x_1, \dots, x_n) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

where

x_i = i th independent variable ($i = 1, \dots, n$)

β_j = j th regression coefficient ($j = 0, 1, \dots, n$)

$$E(y|x_1, \dots, x_n) = P(y=1|x_1, \dots, x_n)$$

A major difficulty in using a linear regression model is that probabilities greater than 1 or less than 0 may occur. This is most likely to occur when an independent variable assumes a value near or outside the limits of the initial data set.

^{1/}This analysis is different from discriminant analysis, which is used to assign individual observations to one of a finite number of classes. The analysis discussed in this paper, however, is designed to estimate the probability that any given observation will assume one of two distinct classes.

Neter and Maynes (1970) have cited several reasons for expecting a curvilinear rather than a linear relationship when the dependent variable is dichotomous. Thus, a nonlinear model appears to be preferable.

Neter and Maynes have also questioned the meaning or usefulness of the correlation coefficient when the dependent variable is dichotomous. When the true relationship is nonlinear, the correlation coefficient does not apply. Even when the true relationship is linear, the meaning of the correlation coefficient is not clear. Neter and Maynes (1970) have demonstrated that when the dependent variable is dichotomous the correlation coefficient can attain a value of 1 only when the independent variable is also dichotomous. Thus, we are usually unable to judge an estimated correlation coefficient against the maximum value 1 that indicates a perfect relationship.

Walker and Duncan (1967), and others have suggested the logistic function in place of the linear model. This model has the form

$$E(y|x_1, \dots, x_n) = \{1 + \exp[-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)]\}^{-1} \quad (1)$$

and restricts probability estimates to the closed interval [0,1].

Another difficulty in analyzing a dichotomous dependent variable is that we must assume that the variance is nonhomogeneous. Thus, weighted regression must be used to estimate the regression coefficients.

Walker and Duncan (1967) have suggested an algorithm and computer program for fitting the logistic model. Modifications of the computer program and of the algorithm to handle data sampled with unequal probability and four sample analyses are discussed in this paper.

The discussion of the algorithm will utilize matrix notation. Equation (1) is rewritten in matrix notation as

$$\hat{P} = [1 + \exp(-x'\beta)]^{-1} + \epsilon \quad (2)$$

where

P = probability of the occurrence of an event

$\hat{P}$ = estimate of P

x = vector of independent variables

β = vector of regression coefficients

$E(\epsilon) = 0$

$\text{Var}(\epsilon) = P(1-P)$

APPLICATIONS OF MODIFIED PROGRAM

Mortality

The program has been used in two independent situations to estimate the probability of individual tree mortality. Permanent sample plots established in western white pine (*Pinus monticola* Dougl.) on the Deception Creek Experimental Forest and on the Kaniksu National Forest were measured at 5-year intervals for 20 years. A preliminary analysis of these data produced the following model for predicting mortality of individual trees:

$$\hat{P} = \{1 + \exp[-0.2735 + 0.2746(CR) + 2.3054(PCT)]\}^{-1}$$

where

CR = crown ratio

PCT = position in the tree basal area distribution

Lee (1971) tabulated mortality rates for lodgepole pine by 2-inch-diameter classes. A functional expression describing this relationship must be developed if it is to be used in a computer program. Such an expression demonstrates another use for this analysis and illustrates the versatility of the logistic function.

The logistic function is expressed graphically as a sigmoid shaped curve. Because Lee's data indicate that a U-shaped curve describes the relationship between mean stand d.b.h. and mortality, the independent variable must be transformed to fit the relationship. The sets of transformed variables x and x^2 , x and $1/x$, and x and $\ln x$ have the potential of fitting either this U-shaped distribution or a bell-shaped distribution. For Lee's data the model is

$$\hat{P} = \{1 + \exp[3.002 + 0.30775(\overline{DBH}) - 0.0174(\overline{DBH})^2]\}^{-1}$$

where

$\overline{DBH}$ = mean stand d.b.h.

Cull Estimation

The logistic function was used twice to estimate cull volume on the Kaniksu National Forest. First it was used to predict the probability that a tree of given characteristics was a cull tree. The dependent variable assumed the value 1 when the tree was a cull tree and otherwise assumed the value 0. The independent variables were DBH and a measure of stand density, $\frac{V/a}{DBH}$. (V/a is gross cubic foot volume per acre.) For hemlock the resulting model is

$$\hat{P} = \{1 + \exp[9.1010 - 0.3696(DBH) - 0.00348(\frac{V/a}{DBH})]\}^{-1}$$

The second use of the function was to estimate the probability that a noncull tree had some cull volume. The dependent variable assumes the value 1 if a tree classified as noncull has some cull volume and otherwise assumes the value 0. The independent variables were DBH and $(\frac{V/a}{DBH})$. For the cedar-hemlock species mix, the model for predicting the probability of cubic-foot defect in noncull trees is

$$\hat{P} = \{1 + \exp[4.494 - 0.3776(DBH)]\}^{-1}$$

Cubic-foot defect in noncull trees was estimated by the product of $\hat{P}$ and a linear function that estimates cubic-foot defect in trees having cull as a function of DBH and $(\frac{V/a}{DBH})$.

Canker Inactivation

The algorithm was also used to investigate levels and trends of natural inactivation of western white pine blister rust cankers and thereby to determine causal relationships for natural inactivation and to explain the variation that occurs in natural inactivation.

The following model demonstrates the relationship between the probability of a canker becoming inactive in the coming year ($\hat{P}$), stand elevation (EL), canker age (CA), and branch diameter (BD):

$$\hat{P} = \{1 + \exp[0.883 + 0.0323(EL) - 0.304(CA) + 1.617(BD)]\}^{-1}$$

Readers unconcerned with the theoretical development of the algorithm and its modifications may omit reading the next four sections. Enough detail has been provided in the section "Guides for the Use of Risk", page 12, to run the program.

SAMPLING ASSUMPTIONS OF ALGORITHM

When data have been collected according to a sampling design, each observation represents a number of similar observations. When this number (weight or blowup factor) varies from observation to observation (i.e., when sampling is other than equal probability sampling), the effect on the relationship of interest must be recognized, particularly in estimating individual tree mortality.

Because of the distribution of mortality in endemic situations, wide areas must be sampled to assure an adequate representation of dead trees. It would not be feasible to measure each green tree in the area; therefore, only a subsample of the green trees is measured. In this situation not all trees are sampled with equal probability. For example, assume 5-acre plots have been selected for obtaining mortality counts. A 1/5-acre plot is established within each 5-acre mortality plot to obtain counts of green trees. Thus, dead trees have a relative weight of 1 and green trees have a relative weight of $5/(1/5) = 25$.

WALKER-DUNCAN ALGORITHM

RISK, the computer version of this algorithm, uses weighted least squares parameter estimation. Weights are estimated as the inverse of the standard deviation. A first-degree Taylor series expansion around a guessed value of β provides the linear function of the parameters required to make least squares a manageable estimation procedure.

An alternative form of the logistic function is given by

$$\ln [P/(1-P)] = x'\beta$$

Although this form of the logistic function can be fit by any linear regression program, it has disadvantages. The independent variables must either be discrete or must be partitioned into discrete classes. The set of all possible combinations of these classes defines a set of categories to which each observation is assigned. P is the proportion of the observations in each category for which the dependent variable is 1. Another disadvantage is that each category must contain enough observations to insure a meaningful value of P for that category. Thus, the form of independent variables is not only restricted by the alternative model, but also the data requirements are frequently greater than if equation (2) were used.

A unique feature of RISK is that parameter estimates are updated when each observation is added rather than only when each iteration is completed. Updating utilizes the following relationship reported by Bartlett (1951):

$$(X'X + xx')^{-1} = (X'X)^{-1} - \frac{(X'X)^{-1}xx' (X'X)^{-1}}{1 + x' (X'X)^{-1}x} \quad (3)$$

$(X'X)$ is the sum of squares matrix for those observations already in the solution, and xx' is the matrix of components added to each element of $(X'X)$ by the addition of the next observation. This relationship is verified by demonstrating that multiplication of the right-hand side of (3) by $(X'X + xx')$ yields the identity matrix.

This algorithm improves on alternative nonlinear regression algorithms because starting values for the parameters are not required. Instead β_1 through β_n of equation (1) are set equal to 0. β_0 is calculated in RISK by solving the equation

$$\bar{p} = [1 + \exp(-\beta_0)]^{-1}$$

for β_0 . $\bar{p}$ is the mean value of the dependent variable in the population (i.e., proportion of the observations for which the dependent variable is 1). If it is desirable to start the iteration at some other value of β_0 , a preferred value of β_0 may be specified.

Following each pass through the data, the resulting $(X'X)^{-1}$ matrix is corrected to remove the effects of the $(X'X)$ matrix used to start that pass. The parameter estimates are similarly corrected. These corrections are made by using the following formulas:

$$(X'X)_c^{-1} = [(X'X)^* - (X'X)_0]^{-1}$$

$$b_c = (X'X)_c^{-1} [(X'X)^* b^* - (X'X)_0 b_0]$$

where

$(X'X)_c^{-1}$ = corrected matrix

$(X'X)^*$ = inverse of the uncorrected $(X'X)^{-1}$ matrix

$(X'X)_0$ = prior $(X'X)$ matrix

b_c = corrected vector of parameter estimates

b^* = uncorrected vector of parameter estimates

b_0 = vector of starting values for parameter estimates

Matrix correction, which reduces the number of passes required for convergence, can be requested once within each pass through the data. Experience indicates, however, that savings in computations resulting from fewer passes are frequently offset by increased computations arising from matrix inversion.

MODIFICATION OF ALGORITHM

The original algorithm was modified to deal with unequal weighting of observations. An extension of the work of Bartlett (1951) demonstrates the inverse of the expression $(X'X + mXX')$ may be written as

$$(X'X + mXX')^{-1} = (X'X)^{-1} - \frac{(X'X)^{-1} mXX' (X'X)^{-1}}{1 + mX' (X'X)^{-1} X} \quad (4)$$

where m is the weight or blowup factor associated with each observation, x . This relationship may be verified by showing that the product of $(X'X + mXX')$ and the right-hand side of (4) yields the identity matrix.

The first degree Taylor Series expansion of the logistic function around a guessed value of the parameter vector, $\bar{\beta}$, may be expressed as

$$y_n^* = x_n' \beta / w_n + \epsilon_n$$

where for the n th observation in the population

$$y_n^* = p_n - (1 + \exp(x_n' \bar{\beta}))^{-1} + x_n' \bar{\beta} / w_n \quad (5)$$

p_n = observed value of dependent variable (either 0 or 1)

$$w_n = 1 / \hat{P}_n \hat{Q}_n$$

$$E(\epsilon_n) = 0$$

$$\text{Var}(\epsilon_n) = \sigma^2 P_n Q_n$$

Let $Y = WY^*$ so that the weighted Normal Equations may be written as

$$X'W^{-1}X\beta = X'W^{-1}Y$$

which implies that the least squares estimator of β is

$$b = (X'W^{-1}X)^{-1}X'W^{-1}Y$$

Since W^{-1} is diagonal we can define W^* such that

$$W^*W^* = W \text{ (i.e., } w_n = 1/\hat{P}_n\hat{Q}_n \text{ and } w_n^* = \sqrt{w_n} \text{)}$$

Then let $X^* = W^*X$ and

$$b = (X^{*'}X^*)^{-1}X^{*'}W^{-1}Y \quad (6)$$

A recursive relationship for estimating b_n results when equations (4), (5), and (6) are combined in the following manner:

$$\begin{aligned}
 b_n &= [(X^{*'}X^*)_{n-1} + mX_n^{*'}X_n^{*}]^{-1}[(X'W^{-1}Y)_{n-1} + mX_nY_n/w_n] \\
 &= \left\{ (X^{*'}X^*)_{n-1}^{-1} - \frac{(X^{*'}X^*)_{n-1}^{-1} mX_n^{*'}X_n^{*} (X^{*'}X^*)_{n-1}^{-1}}{1 + mX_n^{*'}(X^{*'}X^*)_{n-1}^{-1}X_n^{*}} \right\} [(X'W^{-1}Y)_{n-1} + mX_nY_n/w_n] \\
 &= b_{n-1} + \frac{(X'W^{-1}X)_{n-1}^{-1} mX_n (p_n - \hat{p}_n)}{1 + mX_n^{*'}(X^{*'}X^*)_{n-1}^{-1}X_n^{*}/w_n} \quad (7)
 \end{aligned}$$

Iteration continues until the change in regression coefficients following two successive passes through the data satisfies the convergence criteria

$$\delta_i/[\tau + |b_i|] < \eta$$

where

δ_i = change in i th regression coefficient following successive passes through the data.

b_i = i th regression coefficient

η = convergence criteria

τ = small number to handle case where b_i approaches 0.

Applying equation (7) is similar to adding m identical observations to the solution between successive estimations of the set of parameters. Adding an observation with a large relative weight frequently drives the solution so far that the computer cannot distinguish the resulting estimate of P_n from either 0 or 1, which will terminate the iteration abnormally.

The computational difficulty just discussed has been minimized by setting a limit for the sum of successive weights. The effect of adding each weighted observation to the solution is accumulated. However, the parameter estimates used in the iterative estimation procedure are updated only when the sum of successive weights exceeds the set limit. The same procedures may be used to handle similar problems that arise when data sorting results in runs of 0's or 1's.

STATISTICAL CRITERIA FOR GOODNESS-OF-FIT

When the dependent variable is dichotomous, mean square error (the customary measure of goodness-of-fit for regression relationships) is not appropriate. The definition of mean square error demonstrates its inappropriateness as a measure of goodness-of-fit.

$$MSE = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{N-r}$$

where N is the population size and r is the number of parameters to be estimated in the regression. y_i is always either 0 or 1 while $\hat{y}_i$ is a continuous variable in the range $(0 \leq \hat{y}_i \leq 1)$. The following proof shows that as N increases, the limiting value for the mean square error is 1 rather than 0, as is expected in most regression relationships.

When Y is a (0,1) variable, $\bar{Y} = \sum_{i=1}^N y_i / N = P$ and $\sum_{i=1}^N (y_i - \bar{Y})^2 = NPQ$ (Cochran 1963).

$$SSE = \sum_{j=1}^L \sum_{i=1}^{N_j} \frac{(Y_{ij} - \hat{Y}_{ij})^2}{P_{ij} Q_{ij}}$$

where L represents the number of arbitrary classes defined such that in each class P_{ij} is approximately constant (i.e., $\equiv P_{.j}$).

Thus

$$SSE \doteq \sum_{j=1}^L \sum_{i=1}^{N_j} \frac{(Y_{ij} - \hat{Y}_{ij})^2}{P_{.j} Q_{.j}}$$

$$\text{Within each class } \hat{Y}_{ij} \doteq \sum_{i=1}^{N_j} Y_{ij} / N_j = \bar{Y}_{.j} = P_{.j}$$

so

$$SSE \doteq \sum_{j=1}^L \frac{1}{P_{.j} Q_{.j}} \sum_{i=1}^{N_j} (Y_{ij} - \bar{Y}_{.j})^2 = \sum_{j=1}^L \frac{1}{P_{.j} Q_{.j}} N_j P_{.j} Q_{.j} = \sum_{j=1}^L N_j = N$$

This approximation improves as the number of classes increases and as the width of each class decreases.

Thus, $MSE \doteq \frac{N}{N-r}$ where r = number of parameters to be estimated and

$$\lim_{N \rightarrow \infty} MSE \doteq \lim_{N \rightarrow \infty} \frac{N}{N-r} = 1$$

RISK output includes three types of statistics that might be used to measure goodness-of-fit. These are a set of t tests that test whether each of the estimated parameters is significantly different from 0, an analysis of variance F test that tests the significance of the amount of variation explained by regression, and a chi-square table that evaluates goodness-of-fit over the range of predictions (e.g., $0 \leq \hat{y}_{ij} \leq 1$).

Each of these statistics tells us something different about the goodness-of-fit of a model. Walker and Duncan (1967) imply that the F statistic is the appropriate measure of goodness-of-fit for screening alternative models. Although I feel that the chi-square statistic is preferable for screening alternative models, the method used to measure goodness-of-fit is left to the discretion of the user.

RISK is not an efficient procedure for screening alternative sets of independent variables, especially when the data set is large. RISK seems to be most efficient for estimating parameters after the optimal set of predictor variables has been selected. The method proposed by Sterling and others (1969) based on information theory has been successfully used to select an optimal set of independent variables.^{2/}

^{2/}A computer program of the screening algorithm is documented by David A. Hamilton, Jr., and Donna Wendt, to be published as USDA Forest Service General Technical Bulletin.

GUIDES FOR USE OF RISK

RISK is a computer program written by Walker and Duncan (1967) that calculates the nonlinear regression coefficients of the logistic function utilizing the algorithm just discussed. The program has been modified to handle data with unequal sampling weights attached to each observation. The program is written in FORTRAN IV and has been run on an IBM 360/67. Requests for the program should be directed to:

Intermountain Forest and Range Experiment Station
Forestry Sciences Laboratory
Attention: David A. Hamilton, Jr.
1221 S. Main
Moscow, Idaho 83843

Control Cards

<i>Field Function</i>	<i>Value</i>	<i>Column</i>	<i>Name</i>	<i>Format</i>
<u>Card 1</u>				
Problem identification		1-80	IDENT	20A4
<u>Card 2</u>				
Regressor set number	xx...x	1-10	ANRS	F10.0
Run number	xx...x	11-20	AN	F10.0
Number of parameters to be estimated (≤ 30)	xx...x	21-30	ANR	F10.0
Number of observations in population	xx...x	31-40	ANØ	F10.0
Is input file on cards?	Yes = 0	41-50	ATAPE	F10.0
No = input unit number containing input file				

Field Function	Value	Column	Name	Format
Is default value for η , (.0002), in the convergence criteria to be overridden? ^{3/}	Yes = desired value for η No = blank	51-60	ERR	F10.0
Is default value for τ , (.01), in the convergence criteria to be overridden?	Yes = desired value for τ No = blank	61-70	TAU	F10.0

Card 3

Are the starting values for parameter estimates to be provided? ^{4/}	No = 0 Yes = desired value of β_0	1-10	PO	F10.0
Code designating a missing value in an observation ^{5/}	xx...x	11-20	U	F10.0
Are the variables to be labeled on the computer output?	No = 0 Yes = number of labels provided	21-30	VARLAB	F10.0
Limit for sum of weights ^{6/}	xx...x	31-40	ELIM	F10.0

Card 4

Number of variables to be read for each observation (≤ 32)	xxxxx	1-5	LIM	I5
---	-------	-----	-----	----

Card 5

NOTE: On card 5, the absolute value of A is the observation number at which matrix correction would occur.

^{3/}The default values for the parameters of the convergence criteria are satisfactory for most uses of RISK. If increased precision is desired, however, the value of η should be decreased. Conversely, if less precision is required, η can be increased.

^{4/}Starting values for the parameter estimates are generated by RISK. These are:

$$\beta_j = 0 \quad (j=1, \dots, n)$$

$$\frac{\beta_0}{\bar{p}} = \text{solution of the equation } \bar{p} = [1 + \exp(-\beta_0)]^{-1}$$

$$\bar{p} = \text{mean value of dependent variable in the population.}$$

~An alternative value for β_0 may be specified on control card 3.

^{5/}Each missing value in the data should be coded. RISK deletes any observation that has one of the independent variables set equal to the missing value code.

^{6/}The limit for the sum of weights should normally be set equal to 0.5. However, if the data are sorted or if the observations have unequal weights, the resulting estimates of P may be driven to 0 or 1. When this occurs, the limit for the sum of weights should be increased until the difficulty is overcome.

<i>Field Function</i>	<i>Value</i>	<i>Column</i>	<i>Name</i>	<i>Format</i>
For J=1,...,16, should matrix correction occur within Jth pass through data? If so, is intermediate output desired? ^{2/}	No = 0	1-5 6-10	NIT(J)	I5
	Correction with intermediate output = -A	.		
	Correction without intermediate output = A	.		

NOTE: For passes J = 2,...,16, if no additional passes through the data are desired, NIT(J) is set equal to 9999.

Card 6

Format for reading card or tape input	1-80	F1000	20A4
---------------------------------------	------	-------	------

Card 7

Leader label for input file	1-80	ABEL	20A4
-----------------------------	------	------	------

NOTE: Omit card 7 if card input file is used.

Card 8

For J=1,...,VARLAB what is the Jth variable name or label?	xxxx	2-5 7-10	SX(J)	(1X,A4)
		.		
		.		
		.		

NOTE: Omit this card if VARLAB = 0

Card 9

Which field on card 8 contains the label for the dependent variable?	xxxxx	1-5	NVAR(NR1)	I5
--	-------	-----	-----------	----

For K=2,...,NR what field on card 8 contains the label for the Kth independent variable?	xxxxx	6-10 11-15	NVAR(K)	I5
		.		
		.		
		.		

NOTE: Omit this card if VARLAB = 0.

NOTE: Card input file inserted here if ATAPE = 0.

^{2/}Matrix correction removes the effects of the $(X'X)$ matrix used to start that pass through the data from both the resulting $(X'X)^{-1}$ matrix and the vector of regression coefficients. Requesting matrix correction within a pass through the data reduces the number of passes through the data required for convergence. However, experience has indicated that frequently any savings in computations resulting from fewer passes through the data are offset by increased computations arising from the additional matrix inversions.

<i>Field Function</i>	<i>Value</i>	<i>Column</i>	<i>Name</i>	<i>Format</i>
<u>Card 10,...,k</u>				

What is the Jth observation number for which intermediate output is desired?	xxxxx	1-5 6-10 . . .	NITI(J)	I5
--	-------	----------------------------	---------	----

NOTE: There must be at least one such card for each pass through the data for which $NIT(I) < 0$.

NOTE: An NITI(J) value set equal to the total number of observations or blank will terminate intermediate output for that pass through the data.

NOTE: Omit these cards if all $NIT(1) \geq 0$.

Data Preparation

A data transformation subroutine, titled TRNS, has been added to RISK to provide the user with maximum flexibility in input data format. A TRNS subroutine must be written for each use of RISK. Input to TRNS is the observation vector read according to the variable format specified by control card 6. The independent variable X(1) is set equal to 1 by the main program. The dependent variable, all additional independent variables, and the weight for the observation must be defined within TRNS. For the simplest case where no transformations are to be made, the following setup for TRNS would be adequate.

```

SUBROUTINE TRNS (D, NR1, NR, X, R)

DIMENSION D(32), X(32)

DO 1 I = 2,NR

1 X(I) = D(1)

X(NR1) = D(1)

R = D(NR + 1)

RETURN

END

```

The notation in this subroutine is defined as follows:

D = observation vector as it is read from cards or tape.

NR = number of parameters to be estimated.

NR1 = subscript of dependent variable in the X vector (set equal to NP+1 by main program).

X = vector of independent and dependent variables to be used in regression.

R = weight or blowup factor for each observation.

This sample subroutine assumes that the first variable in the observation vector is the dependent variable (0 or 1), the variables 2 through NR are independent variables, and variable (NR + 1) is the weight or blowup factor that describes the number of similar observations represented by the sample observation.

Program Output

RISK provides a summary of the input and a description of the regression results. The summary of input begins with a tabulation of control information the user has provided to run the program. Also included are tables of means and standard deviations of all variables in the population. The final portion of this summary of input consists of three matrices: a matrix of correlation coefficients, an associated matrix of t statistics indicating the significance of the correlation coefficients, and a product moment matrix.

The output dealing with regression results may be broken into two sections: (a) A list of the parameter estimates and error mean square resulting from each of the passes through the data followed by the final parameter estimates and (b) the measures of goodness-of-fit. These measures are a table of t statistics, an analysis of variance table, and a table of chi-square statistics.

The first table of t statistics were calculated on the assumption that error mean square is 1 (i.e., the model provides a perfect fit of the data). The second table is calculated using the actual error mean square provided by the analysis of variance. Neither of these statistics is truly distributed according to the t distribution, because the $(X'X)^{-1}$ matrix is constructed utilizing estimated values of PQ as weights. The statistics in the first table are approximately normally distributed and the statistics in the second table are approximately distributed according to the t distribution. As the error in predicting P becomes negligible, these approximations improve.

The user may request intermediate output for any pass through the data that includes a matrix correction. This output consists of the $(X'X)^{-1}$ matrix and the vector of regression coefficients that exist at those points in the pass for which intermediate output is requested. This same output is provided following the final observation in any pass that included a matrix correction. The $(X'X)$ matrix is also printed at this time. The $(X'X)^{-1}$ and $(X'X)$ matrices both before and after correction, the regression coefficients before and after correction, and the regression coefficients that were used to start the pass are printed at the point of matrix correction.

Error Diagnostics

Six messages are generated by RISK to explain abnormal terminations of the program:

MORE THAN NC DATA CARDS HAVE BEEN INCLUDED IN THIS RUN

This diagnostic message is generated when there are more cards in the card input file than are specified, which could occur either from incorrectly counting the number of observations in the input data set or as the result of incorrectly preparing the set of control cards.

THE NUMBER OF REGRESSORS NR IS IN EXCESS OF THE PRESENT PROGRAM MAXIMUM OF 30

This diagnostic message is generated if the specified number of regressors, NR, is greater than 30. This check is made before the population of observations is read into the machine; if NR is greater than 30, program execution is terminated.

INCONSISTENCY BETWEEN ACTUAL AND SPECIFIC TAPE LABELS

When the population of observations are on tape or in a cataloged data set, if the leader label on input unit ATAPE is not identical to the label on control card 7, the above diagnostic message is generated and program execution is terminated. This provides some assurance that you are reading the appropriate input file.

MATRIX INVERSION FAILURE, NGO = ---, I = ----, SINGULARITY AT J = --, WITH A(J,J) = -----

This diagnostic message is generated when subroutine INVERT is unable to successfully invert a matrix. The variable NGO in this message indicates at what point in the program the inversion failed. When NGO equals 1, the inversion failure has occurred at statement RISK2610. When NGO equals 999, the inversion failure has occurred at statement RISK5360. In all other cases, NGO will be a two or three digit number. The final digit indicates where in the program inversion has failed; the initial one or two digits indicate the iteration on which inversion failed. The various values of NGO are:

<i>NGO</i>	<i>Location of Inversion Failure</i>
X2	RISK3460
X3	RISK3830
X0	RISK4520

where X is the iteration number on which matrix inversion fails. The variable I in this message is the number of the last observation added to the solution before inversion failed. J is the location in the matrix to be inverted that caused inversion failure, and A(J,J) is the actual value of the matrix element that caused inversion failure.

COEFFICIENTS HAVE DIVERGED TO POINT THAT P EQUALS ----

Occasionally for a sorted or weighted set of observations, the regression coefficients may be driven to the point where the estimates of P approach 0 or 1. This situation results in a division by 0, which would terminate program execution on the IBM 360/67; therefore, the value of P is checked and the above diagnostic message is generated any time P is equal to 0 or 1.

When this message occurs, the user has two courses of action that may allow completion of the analysis: (a) Rearranging the data to eliminate the effects of sorting (unfortunately, this is frequently not feasible), and (b) increasing the input parameter ELIM (the limit for the sum of weights), which reduces the effects both of large variability in the weights and of sorted data.

EXECUTION TERMINATED, SBX = ----

This diagnostic message is generated by circumstances similar to those that generate the previous message. However, in this case, the regression coefficients have been driven to the point that P cannot be evaluated by the computer. When this occurs, the user has the two alternatives discussed for the previous diagnostic message.

DISCUSSION AND CONCLUSIONS

The logistic function is preferable to a linear model for describing the relationship between a dichotomous dependent variable and a set of independent variables. The algorithm discussed in this paper provides a valid method of estimating the parameters of the logistic function.

Both the algorithm and the computer program have been modified to handle data collected with arbitrary probability. This modification is of particular value in cases such as estimating tree mortality when it would not be feasible to measure all observations with equal probability.

LITERATURE CITED

- Bartlett, M. S.
1951. An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Stat.* 22:107-111.
- Cochran, William G.
1963. *Sampling techniques*. 2nd ed. 413 p. New York: John Wiley and Sons, Inc.
- Lee, Y.
1971. Predicting mortality for even-aged stands of lodgepole pine. *For. Chron.* 47(2):29-32.
- Neter, John, and E. Scott Maynes
1970. On the appropriateness of the correlation coefficient with a 0,1 dependent variable. *J. Am. Stat. Assoc.* 65(330):501-509.
- Sterling, Theodor D., Randall G. Binks, Shelby Haberman, and Seymour V. Pollack
1969. Robot data screening--a ubiquitous automatic search technique. In: Milton, Roy C., and John A. Nelder (ed.), *Statistical Computation*, p. 319-333. New York: Academic Press. 462 p.
- Walker, Strother H., and David B. Duncan
1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54:167-179.

HAMILTON, DAVID A., JR.

1974. Event probabilities estimated by regression, USDA Forest Service Res. Pap. INT-152, 18 p., illus. (Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.)

Describes an algorithm for fitting the relationship between a dichotomous dependent variable and a number of independent variables. The specific functional form utilized is the logistic function. The algorithm, proposed and programmed by Walker and Duncan (1967), has been generalized by adding a subroutine to handle data transformations and by introducing procedures to handle data sampled with unequal probability. A user's guide discusses makeup of program control cards, data transformation capabilities, program output, and program-generated errors.

KEYWORDS: nonlinear regression, dichotomous dependent variable, mortality modeling, computer algorithm, logistic function, probability estimation.

HAMILTON, DAVID A., JR.

1974. Event probabilities estimated by regression, USDA Forest Service Res. Pap. INT-152, 18 p., illus. (Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.)

Describes an algorithm for fitting the relationship between a dichotomous dependent variable and a number of independent variables. The specific functional form utilized is the logistic function. The algorithm, proposed and programmed by Walker and Duncan (1967), has been generalized by adding a subroutine to handle data transformations and by introducing procedures to handle data sampled with unequal probability. A user's guide discusses makeup of program control cards, data transformation capabilities, program output, and program-generated errors.

KEYWORDS: nonlinear regression, dichotomous dependent variable, mortality modeling, computer algorithm, logistic function, probability estimation.

HAMILTON, DAVID A., JR.

1974. Event probabilities estimated by regression, USDA Forest Service Res. Pap. INT-152, 18 p., illus. (Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.)

Describes an algorithm for fitting the relationship between a dichotomous dependent variable and a number of independent variables. The specific functional form utilized is the logistic function. The algorithm, proposed and programmed by Walker and Duncan (1967), has been generalized by adding a subroutine to handle data transformations and by introducing procedures to handle data sampled with unequal probability. A user's guide discusses makeup of program control cards, data transformation capabilities, program output, and program-generated errors.

KEYWORDS: nonlinear regression, dichotomous dependent variable, mortality modeling, computer algorithm, logistic function, probability estimation.

HAMILTON, DAVID A., JR.

1974. Event probabilities estimated by regression, USDA Forest Service Res. Pap. INT-152, 18 p., illus. (Intermountain Forest and Range Experiment Station, Ogden, Utah 84401.)

Describes an algorithm for fitting the relationship between a dichotomous dependent variable and a number of independent variables. The specific functional form utilized is the logistic function. The algorithm, proposed and programmed by Walker and Duncan (1967), has been generalized by adding a subroutine to handle data transformations and by introducing procedures to handle data sampled with unequal probability. A user's guide discusses makeup of program control cards, data transformation capabilities, program output, and program-generated errors.

KEYWORDS: nonlinear regression, dichotomous dependent variable, mortality modeling, computer algorithm, logistic function, probability estimation.

Headquarters for the Intermountain Forest and
Range Experiment Station are in Ogden, Utah.
Field Research Work Units are maintained in:

Boise, Idaho

Bozeman, Montana (in cooperation with
Montana State University)

Logan, Utah (in cooperation with Utah
State University)

Missoula, Montana (in cooperation with
University of Montana)

Moscow, Idaho (in cooperation with the
University of Idaho)

Provo, Utah (in cooperation with Brigham
Young University)

Reno, Nevada (in cooperation with the
University of Nevada)

U.S. DEPARTMENT OF THE INTERIOR

BUREAU OF LAND MANAGEMENT

WATER RESOURCES DIVISION