Biometrika (1967), **54**, 1 *and* 2, *p.* 167
Printed in Great Britain

Estimation of the probability of an event as a function of several independent variables

BY STROTHER H. WALKER† AND DAVID B. DUNCAN
Johns Hopkins University

SUMMARY

A method for estimating the probability of occurrence of an event from dichotomous or polychotomous data is developed, using a recursive approach. The method in the dichotomous case is applied to the data of a 10-year prospective study of coronary disease. Other areas of application are briefly indicated.

1. INTRODUCTION

The purpose of this paper is to develop a method for estimating from dichotomous (quantal) or polychotomous data, the probability of occurrence of an event as a function of a relatively large number of independent variables. A key feature of the method is a recursive approach based on Kalman's work (Kalman, 1960 and unpublished report) in linear dynamic filtering and prediction, derivable also from the work of Swerling (1959), which provides an example of many other possible uses of recursive techniques in non-linear estimation and in related areas.

The problem that motivated the investigation is a central one in the epidemiology of coronary heart disease, and it will be used to fix ideas and illustrate the method. Some indication of the range of applications will be given in the conclusion.

2. THE EPIDEMIOLOGICAL PROBLEM: PREVIOUS METHODS OF ANALYSIS

Clinical and epidemiological studies had by 1950 identified a long list of factors as possibly, or probably, associated with the occurrence of coronary heart disease. To gather data on such associations, four large, long-term prospective studies were established in the United States between 1948 and 1957. These may be identified briefly, in the order of initiation, as the Framingham (Dawber, Meadors & Moore, 1951), Los Angeles (Chapman *et al.* 1957), Albany (Hilleboe, James & Doyle, 1954) and Chicago (Paul *et al.* 1963) studies. These studies have now followed samples of from 1800 men, in round numbers, to 5000 men and women, for up to 10 years. Examinations are given annually or biennially to the study participants and their status with respect to coronary disease is recorded. The references cited describe the methods and objectives of these studies in detail and review the previous literature on factors suspected of association with occurrence of the disease.

The method of analysis generally used in these studies has been simply tabulation and cross-tabulation of incidence on one, two, or rarely, three of the factors under study, by sex and by age interval. Results obtained thereby have been extensively reported, e.g.

† Now at the University of Colorado Medical Center, Denver, Colorado.

for Framingham, by Kannel *et al.* (1961, 1962, 1964), Kannel (1964) and Kagan *et al.* (1963); for Los Angeles, by Chapman & Massey (1964); for Albany, by Doyle *et al.* (1959); and for Chicago, by Paul *et al.* (1963).

The limitations of this approach in a problem involving so many variables have been recognized. Cornfield (1962) pointed them out, concluding that 'The use of a mathematical model which summarizes the observations in a small number of disposable parameters seems to offer the only present hope of obtaining quantitative answers to questions of interest'.

Cornfield and his co-workers (Cornfield, Gordon & Smith, 1961) had shown that under certain assumptions the probability of occurrence of the disease can be represented as a logistic function whose argument is a linear function of the independent variables, provided that both coronary and non-coronary populations are characterized by multivariate normal frequency functions in the independent variables. Estimates of the linear coefficients are then obtained as simple functions of the estimated parameters of the underlying normal distributions. Cornfield (1962) carried out such a fit for $\log_{10}$ cholesterol and $\log_{10}$ (systolic blood pressure -75) in the data of the Framingham study.

Previously Gertler and associates (Gertler *et al.* 1959), in collaboration with the Mathematics Center at New York University, had applied discriminant analysis to essentially the same problem, in a relatively small sample of 61 coronary cases and 135 normal controls. This is equivalent to fitting a hyperplane to the zero-one observations corresponding to occurrence or non-occurrence of the disease.

3. THE DICHOTOMOUS MODEL

Our own approach has grown out of ideas developed in a series of unpublished reports for the U.S. Navy; see Duncan & Rhodes (1952). It also follows naturally from closely similar independent ideas developed by Cox (1958, 1966).

We assume a sample of N individuals free of coronary heart disease, measured or categorized with respect to each of s independent variables, followed for a period of years thereafter, and identified as having developed the disease by the end of that period, or not. In the dichotomous model, so-called because it assumes two possible states of the individual subject, the observed values of the dependent variable $p_n (n = 1, \ldots, N)$ will be one or zero, corresponding to occurrence or non-occurrence of the disease in the nth individual within the given period. We regard the observations as realizations of the individual probability of disease within the given time period.

Designate the vector of independent variables for the nth individual in the sample as $\mathbf{x}_n$, where

$$\mathbf{x}_n' = (x_{n0}, x_{n1}, \ldots, x_{ns}).$$

The 'dummy' regressor $x_{n0} \equiv 1\,(n = 1, \ldots, N)$ is included to provide for estimation of an intercept. We will also define the $N \times (s+1)$ matrix of independent variables for the sample as

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1s} \\ x_{20} & x_{21} & \cdots & x_{2s} \\ & & \cdots & \\ x_{N0} & x_{N1} & \cdots & x_{Ns} \end{bmatrix}.$$

It is clear that a linear model for the expected values of the observations, $P_n = E(p_n)$, would not be reasonable. If it be assumed that the probability P (suppressing the subscript

n) is some function of the independent variables and of a vector $\boldsymbol{\theta}$ of the unknown parameters, or $P = g(x_1, \ldots, x_s; \theta_1, \ldots, \theta_h)$, then the biological facts suggest that P will be near zero over a certain part of the domain of g, near one over another part, and will increase from near zero to near one over an intermediate part of the domain.

Exact determination of the function g does not seem possible, and in view of the known complexity of the underlying relations, such a function is not likely to be useful for estimation. In the light of present medical knowledge a reasonable assumption is that P follows a symmetric sigmoid curve; according to the model in an unpublished report by Duncan or Cox (1958) we take the argument as the linear function $\mathbf{x}'\boldsymbol{\beta}$ of the independent variables and of a vector $\boldsymbol{\beta}$ of $s+1$ unknown coefficients; we then seek an estimate of $\boldsymbol{\beta}$ which will provide a satisfactory fit of the observations p to such a symmetric sigmoid curve, which is itself adequately represented by the logistic function

$$P = (1 + e^{-\mathbf{x}'\boldsymbol{\beta}})^{-1}. \tag{3·1}$$

Restriction to a symmetric curve is not essential. A skew curve, involving the estimation of an additional parameter, could be fitted by the methods that follow.

By analogy with probit analysis, an alternative choice for P would be the normal function

$$\frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} e^{-\frac{1}{2}z^2} dz;$$

but the logistic is more tractable mathematically and better suited to computer operations. Even in the case of bioassay, where the model assumptions (e.g. normal distribution of tolerances in the population) are more restrictive, Finney (1964, p. 460) finds that '... the logistic distribution is scarcely distinguishable from the normal between response rates of 0·01 and 0·99, ... the choice [may be] made primarily on the score of convenience'.

4. Estimation of the parameters in the dichotomous case

In general, we will approach estimating through a least-squares argument using estimated weights. Specifically, the model assumes that the nth individual in the sample acquires the disease within the follow-up period of the study with probability P_n, or fails to do so with probability $Q_n = 1 - P_n$. Writing

$$P_n = f(\mathbf{x}_n, \boldsymbol{\beta}) = \{1 + \exp(-\mathbf{x}_n'\boldsymbol{\beta})\}^{-1}$$

in conformity with (3·1), we can start from the representation

$$p_n = f(\mathbf{x}_n, \boldsymbol{\beta}) + \epsilon_n,$$

$$E(\epsilon_n) = 0, \; V(\epsilon_n) = P_n Q_n \quad (n = 1, \ldots, N). \tag{4·1}$$

Expanding in a Taylor series around some initial guessed value of $\boldsymbol{\beta}$, $\overline{\boldsymbol{\beta}}$, and writing $\mathbf{f}'(\mathbf{x}_n, \overline{\boldsymbol{\beta}})$ for $\partial \mathbf{f}(\mathbf{x}_n, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$ at $\boldsymbol{\beta} = \overline{\boldsymbol{\beta}}$, we obtain

$$p_n \doteq f(\mathbf{x}_n, \overline{\boldsymbol{\beta}}) + \mathbf{f}'(\mathbf{x}_n, \overline{\boldsymbol{\beta}})(\boldsymbol{\beta} - \overline{\boldsymbol{\beta}}) + \epsilon_n,$$

or

$$y_n^* \doteq \mathbf{f}'(\mathbf{x}_n, \overline{\boldsymbol{\beta}})\boldsymbol{\beta} + \epsilon_n, \tag{4·2}$$

where the 'working observation'

$$y_n^* = p_n - f(\mathbf{x}_n, \overline{\boldsymbol{\beta}}) + \mathbf{f}'(\mathbf{x}_n, \overline{\boldsymbol{\beta}})\overline{\boldsymbol{\beta}}. \tag{4·3}$$

Noting that the vector of derivatives is

$$\mathbf{f}'(\mathbf{x}_n, \bar{\boldsymbol{\beta}}) = \bar{P}_n \bar{Q}_n \mathbf{x}_n', \quad \text{where} \quad \bar{P}_n = \{1 + \exp(-\mathbf{x}_n' \bar{\boldsymbol{\beta}})\}^{-1}, \quad \bar{Q}_n = 1 - \bar{P}_n,$$

defining $\mathbf{x}_n^* = \bar{P}_n \bar{Q}_n \mathbf{x}_n$ and writing $\mathbf{X}^*$ for the matrix having $\mathbf{x}_n^{*\prime}$ as its nth row, the system (4·2) can be rewritten in the usual linear form as

$$\mathbf{y}^* \doteq \mathbf{X}^* \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4·4}$$

where $\mathbf{y}^*$ and $\boldsymbol{\epsilon}$ are the $N \times 1$ vectors of the y_n^* and ϵ_n.

Equation (4·4) is in a form suited to the application of weighted iterative non-linear least-squares procedures, using a diagonal weight matrix W determined as the inverse of the variance matrix of the vector $\boldsymbol{\epsilon}$. The weights, as well as the derivatives, must be estimated from the data.

The normal equations are

$$\mathbf{X}^{*\prime} \mathbf{W} \mathbf{X}^* \boldsymbol{\beta} = \mathbf{X}^{*\prime} \mathbf{W} \mathbf{y}^*. \tag{4·5}$$

But, from a unique property of the logistic function (Garwood, 1941) $\mathbf{X}^* = \mathbf{W}^{-1}\mathbf{X}$ so that $\mathbf{X}^{*\prime}\mathbf{W}\mathbf{X}^* = \mathbf{X}'\mathbf{W}^{-1}\mathbf{X}$ and (4·5) may be rewritten as

$$\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}^{-1}\mathbf{y}, \tag{4·6}$$

where $\mathbf{y}$ is the vector of rescaled working observations,

$$\mathbf{y} = \mathbf{W}\mathbf{y}^*. \tag{4·7}$$

The normal equations may thus be regarded as based on the working observations y_n^* with weights $W_n = 1/(\bar{P}_n \bar{Q}_n)$ from (4·5), or as based on the rescaled observations y_n with variances $1/(\bar{P}_n \bar{Q}_n)$ and hence with weights $\bar{P}_n \bar{Q}_n$ (4·6). The equations (4·5) and (4·6), as is well known, are identical with those which would be obtained by the method of maximum likelihood.

Proceeding from (4·6), an estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ is now available in the form

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y}. \tag{4·8}$$

Iterative solution of (4·8) by the usual method, essentially the Newton–Raphson method, would depend critically on the quality of initial estimates $\bar{P}_n$. With several independent variables these estimates are more difficult to obtain than, for instance, in bioassay, where only simple regression is involved. Each observation will, in general, lie at a different point in s-dimensional space. Standard initial estimates are one's and zeros, based on single occurrences, and, as such, cannot be used to start the iterations.

Useful solutions to this problem are ones proposed by Duncan & Rhodes (1952) and by Cox (1966). Initial estimates are obtained by fitting a discriminant function. Convergence in the initial iterations is accelerated by the use of centre-interval approximations to the sigmoid function being fitted.

The recent recursive methods due to Swerling (1959) and Kalman (1960), provide an alternative and often more useful approach, permitting the estimates to be 'updated' at the addition of each new item $(p_n, \mathbf{x}_n')$ of data. We now develop this in §5.

5. Derivation of the recursive estimation procedure

Let $\mathbf{b}_n$ be the estimate of $\boldsymbol{\beta}$ based on the first n observations. Then

$$\mathbf{V}_n = \mathbf{V}(\mathbf{b}_n) = (\mathbf{X}_n^{*\prime}\mathbf{W}_n\mathbf{X}_n^*)^{-1} = (\mathbf{X}_n'\mathbf{W}_n^{-1}\mathbf{X}_n)^{-1}, \tag{5·1}$$

where $\mathbf{X}_n^*$, $\mathbf{W}_n$, $\mathbf{X}_n$ and later $\mathbf{y}_n^*$ are comprised of the first n observations.

Letting $w_n = 1/(\hat{P}_n\hat{Q}_n)$ (the form of the estimate $\hat{P}_n$ will be specified below at (5·6)), we have

$$\begin{aligned}\mathbf{V}_n &= (\mathbf{V}_{n-1}^{-1} + \mathbf{x}_n^* w_n \mathbf{x}_n^{*\prime})^{-1} \\ &= \mathbf{V}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n^*(\mathbf{x}_n^{*\prime}\mathbf{V}_{n-1}\mathbf{x}_n^* + w_n^{-1})^{-1}\mathbf{x}_n^{*\prime}\mathbf{V}_{n-1},\end{aligned} \tag{5·2}$$

where $\mathbf{x}_n^* = \hat{P}_n\hat{Q}_n\mathbf{x}_n$ and $\mathbf{x}_n$ is the vector of values of the independent variables for the nth observation. The first step is obvious and the second, yielding a recursive expression for $\mathbf{V}_n$ in terms of $\mathbf{V}_{n-1}$, is proved by multiplication to obtain the identity matrix.

A similar recursive expression for $\mathbf{b}_n$ is given by

$$\mathbf{b}_n = \mathbf{b}_{n-1} + \mathbf{V}_{n-1}\mathbf{x}_n^* d_n^{-1}(y_n^* - \mathbf{x}_n^{*\prime}\mathbf{b}_{n-1}), \tag{5·3}$$

where

$$d_n = \mathbf{x}_n^{*\prime}\mathbf{V}_{n-1}\mathbf{x}_n^* + w_n^{-1}.$$

This may be obtained as follows:

$$\begin{aligned}\mathbf{b}_n &= \mathbf{V}_n(X_n^{*\prime}\mathbf{W}_n\mathbf{y}_n^*) \\ &= (\mathbf{V}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n^* d_n^{-1}\mathbf{x}_n^{*\prime}\mathbf{V}_{n-1})(\mathbf{X}_{n-1}^{*\prime}\mathbf{W}_{n-1}\mathbf{y}_{n-1}^* + \mathbf{x}_n^* w_n y_n^*) \\ &= \mathbf{b}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n^* d_n^{-1}\mathbf{x}_n^{*\prime}\mathbf{b}_{n-1} + \mathbf{V}_{n-1}\mathbf{x}_n^* d_n^{-1}(d_n w_n - \mathbf{x}_n^{*\prime}\mathbf{V}_{n-1}\mathbf{x}_n^* w_n)\, y_n^*\end{aligned}$$

which is the result (5·3) since $d_n w_n - \mathbf{x}_n^{*\prime}\mathbf{V}_{n-1}\mathbf{x}_n^* w_n = 1$.

The recursive equations (5·2) and (5·3) are simple to program for a computer, and can be made still more convenient by rewriting in terms of the original data. Recalling that $\mathbf{X}^* = \mathbf{W}^{-1}\mathbf{X}$ and noting that, on evaluation of the derivatives in terms of which it was defined at (4·3) (with $\mathbf{b}_{n-1}$ used as the estimate of $\boldsymbol{\beta}$) $y_n^* = p_n - \hat{P}_n + \mathbf{x}_n^{*\prime}\mathbf{b}_{n-1}$, we obtain:

$$\mathbf{V}_n = \mathbf{V}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n c_n \mathbf{x}_n'\mathbf{V}_{n-1}, \tag{5·4}$$

where

$$c_n = (w_n + \mathbf{x}_n'\mathbf{V}_{n-1}\mathbf{x}_n)^{-1},$$

and

$$\mathbf{b}_n = \mathbf{b}_{n-1} + \mathbf{V}_{n-1}\mathbf{x}_n c_n w_n (p_n - \hat{P}_n). \tag{5·5}$$

The term $\hat{P}_n$ is the estimate of P_n based on $\mathbf{b}_{n-1}$ given by

$$\hat{P}_n = \hat{P}_{n|n-1} = \{1 + \exp(-\mathbf{x}_n'\mathbf{b}_{n-1})\}^{-1}. \tag{5·6}$$

It is of interest to note that $\mathbf{V}_n$, usually thought of as the inverse of an $(s+1)\times(s+1)$ matrix, is obtained in (5·4) by application of a simple correction to $\mathbf{V}_{n-1}$.

The recursive method appears to offer advantages both in more rapid convergence and in easier starting. Experience to date has borne this out.

In the usual methods the whole sample is used, with weights determined by initial estimates, to obtain the first iterative estimate; with the new weights thus determined the whole sample is again employed, and so on. Under recursion the estimate of $\boldsymbol{\beta}$ and hence of the weights is changed with each observation, which allows convergence *within* the span of the usual iteration. The number of computations per iteration is larger for the recursive method but the increase is surprisingly small in view of the additional output available.

The extra time required per iteration can vary from none at all up to 50 or 100 % depending on the data read-in technique and on the dimensions of the problem. In the examples to be presented convergence is essentially accomplished within the first iteration, a second being used only for minor adjustment and final confirmation.

Thanks to rapid convergence within iterations, the need for good initial estimates is considerably relaxed. This fact, together with another Bayes-like feature of the more general Kalman methods, can be used to solve the starting problem. Any rough initial estimate $\mathbf{b}_0$ can be ascribed a prior variance matrix $\mathbf{V}_0$, the two providing a starting point for the recursive process based on actual observations. Now letting $\mathbf{b}_k^*$ and $\mathbf{V}_k^*$ denote the recursive estimates based on $\mathbf{b}_0$, $\mathbf{V}_0$ and the first k items of data, the contributions of $\mathbf{b}_0$ and $\mathbf{V}_0$ to $\mathbf{b}_k^*$ and $\mathbf{V}_k^*$ can be removed, for all practical purposes, by taking

$$\mathbf{V}_k = (\mathbf{V}_k^{*-1} - \mathbf{V}_0^{-1})^{-1} \tag{5·7}$$

and

$$\mathbf{b}_k = \mathbf{V}_k(\mathbf{V}_k^{*-1}\mathbf{b}_k^* - \mathbf{V}_0^{-1}\mathbf{b}_0). \tag{5·8}$$

Equations (5·7) and (5·8) follow from the fact that $\mathbf{b}_k^*$ is effectively the weighted combination of $\mathbf{b}_0$ and $\mathbf{b}_k$ with weights $\mathbf{V}_0^{-1}$ and $\mathbf{V}_k^{-1}$, where $\mathbf{b}_k$ and its variance $\mathbf{V}_k$ are based on the data alone. Finally, the recursive process continues to completion through the remaining $n-k$ items of data, starting with $\mathbf{b}_k$ and $\mathbf{V}_k$.

On consideration it is evident that the successful performance of the recursions can be affected by the order of presentation of the individual records. This is not a serious problem. However, any special ordering of the records which would introduce obvious autocorrelation between successive records should be avoided.

6. THE TRICHOTOMOUS MODEL

In situations to which our model is adapted, it often happens that more than two states of the subjects are reported in the data. For example, in the Framingham study of coronary heart disease, the data permit classification of the subjects as having suffered myocardial infarction (MI), angina pectoris (AP), or as being free of coronary heart disease ($\overline{\text{CHD}}$). It is desirable to use this additional information.

One hypothesis treats MI and AP as distinct, although presumably correlated, disease entities. Another treats MI as a more, and AP as a less severe form of coronary heart disease. There is evidential support for each of these hypotheses. A model for the second will be developed in this section.

The discussion of the polychotomous model for a vector independent variable by Duncan (unpublished report) is the basis of our treatment here. The polychotomous model in bioassay has been discussed by Aitchison & Silvey (1951), Ashford (1959), Gurland, Lee & Dahm (1960) and Cox (1966). Gurland *et al.* worked out the problem for the bioassay situation with scalar independent variable, using the minimum logit chi-square method.

We will define the observations for a given individual in the sample as follows:

$$\begin{aligned}
\text{observed proportion MI} = p_1 &= 0, \text{ MI not present} \\
&= 1, \text{ MI present,} \\
\text{observed proportion AP} = p_2 &= 0, \text{ AP not present} \\
&= 1, \text{ AP present,} \\
\text{observed proportion } \overline{\text{CHD}} = p_3 &= 1 - p_1 - p_2.
\end{aligned}$$

Now we may put $$P_2 = E(p_2),$$

$$P_1 = E(p_1) = f_1 = f(\alpha_1, \boldsymbol{\beta}, \mathbf{x}) = \{1 + \exp(-\alpha_1 - \mathbf{x}'\boldsymbol{\beta})\}^{-1} \tag{6·1}$$

and $$E(p_1 + p_2) = P_1 + P_2 = f_2 = f(\alpha_2, \boldsymbol{\beta}, \mathbf{x}) = \{1 + \exp(-\alpha_2 - \mathbf{x}'\boldsymbol{\beta})\}^{-1}. \tag{6·2}$$

Also $$P_3 = 1 - P_1 - P_2$$

and $$Q_i = 1 - P_i \quad (i = 1, 2, 3).$$

Considered separately (6·1) and (6·2) involve just the same assumptions as those discussed above for the dichotomous case. Considered jointly they involve the further assumption that the state of an individual described by the vector $\mathbf{x}$, which is sufficient to entail the more severe form MI, is certainly sufficient to entail the less severe form AP. If MI and AP are in reality grades of severity of coronary disease, this assumption will hold at least approximately. If on the other hand these are distinct, even though closely related diseases, it is not likely to hold.

The mathematical reflexion of this assumption is seen in the fact that $P_1 + P_2 \geqslant P_1$, which holds if and only if the 'slope' coefficient $\boldsymbol{\beta}$ is identical in (6·1) and (6·2), as is easily shown.

The definitions and assumptions given yield the model:

$$p_{1n} = f(\alpha_1, \boldsymbol{\beta}, \mathbf{x}) + \epsilon_{1n}, \tag{6·3}$$

$$p_{1n} + p_{2n} = f(\alpha_2, \boldsymbol{\beta}, \mathbf{x}) + \epsilon_{2n} \quad (n = 1, \ldots, N). \tag{6·4}$$

It is seen that the errors are correlated in pairs. Defining

$$\boldsymbol{\epsilon}_n = (\epsilon_{1n}, \epsilon_{2n})', \mathbf{p}_n = (p_{1n}, p_{2n})',$$

$$\mathbf{P}_n = (P_{1n}, P_{2n})',$$

and $$\mathbf{K} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix},$$

and noting that $\mathbf{p}_n$ may be taken as a trinomial variate (which corresponds to the assumption of binomial errors in the dichotomous case), so that

$$\mathbf{V}(\mathbf{p}_n) = \begin{bmatrix} P_{1n}Q_{1n} & -P_{1n}P_{2n} \\ -P_{1n}P_{2n} & P_{2n}Q_{2n} \end{bmatrix},$$

it follows that

$$\mathbf{V}(\boldsymbol{\epsilon}_n) = \mathbf{V}(\mathbf{K}\mathbf{p}_n) = \mathbf{K}\mathbf{V}(\mathbf{p}_n)\mathbf{K}' = \begin{bmatrix} P_{1n}Q_{1n} & P_{1n}P_{3n} \\ P_{1n}P_{3n} & P_{3n}Q_{3n} \end{bmatrix}. \tag{6·5}$$

Since now $$\det \mathbf{V}(\boldsymbol{\epsilon}_n) = P_{1n}P_{2n}P_{3n},$$

$$\mathbf{V}^{-1}(\boldsymbol{\epsilon}_n) = \frac{1}{P_{2n}} \begin{bmatrix} \dfrac{Q_{3n}}{P_{1n}} & -1 \\ -1 & \dfrac{Q_{1n}}{P_{3n}} \end{bmatrix}. \tag{6·6}$$

The model (6·3), (6·4) is now approximately linearized by Taylor series expansion of the functions f_1, f_2 around some guessed initial values $\bar{\alpha}_1$, $\bar{\alpha}_2$, $\bar{\boldsymbol{\beta}}$. Proceeding as in the dichotomous case, working variables are defined by

$$\left.\begin{aligned} y^*_{1n} &= p_{1n} - f(\bar{\alpha}_1, \bar{\boldsymbol{\beta}}, \mathbf{x}_n) + \frac{\partial f(\bar{\alpha}_1, \bar{\boldsymbol{\beta}}, \mathbf{x}_n)}{\partial \alpha_1}\bar{\alpha}_1 + \frac{\partial \mathbf{f}(\bar{\alpha}_1, \bar{\boldsymbol{\beta}}, \mathbf{x}_n)}{\partial \boldsymbol{\beta}}\bar{\boldsymbol{\beta}}, \\ y^*_{2n} &= p_{1n} + p_{2n} - f(\bar{\alpha}_2, \bar{\boldsymbol{\beta}}, \mathbf{x}_n) + \frac{\partial f(\bar{\alpha}_2, \bar{\boldsymbol{\beta}}, \mathbf{x}_n)}{\partial \alpha_2}\bar{\alpha}_2 + \frac{\partial \mathbf{f}(\bar{\alpha}_2, \bar{\boldsymbol{\beta}}, \mathbf{x}_n)}{\partial \boldsymbol{\beta}}\bar{\boldsymbol{\beta}}. \end{aligned}\right\} \tag{6·7}$$

Since

$$\frac{\partial f_1}{\partial \alpha_1} = P_{1n}Q_{1n}, \quad \frac{\partial f_2}{\partial \alpha_2} = P_{3n}Q_{3n},$$

$$\frac{\partial f_1}{\partial \beta_j} = P_{1n}Q_{1n}x_{jn}, \quad \frac{\partial f_2}{\partial \beta_j} = P_{3n}Q_{3n}x_{jn} \quad (j = 1, \ldots, s),$$

we obtain finally

$$\mathbf{y}^*_n = \begin{bmatrix} y^*_{1n} \\ y^*_{2n} \end{bmatrix} = \begin{bmatrix} p_{1n} - f(\bar{\alpha}_1, \bar{\boldsymbol{\beta}}, \mathbf{x}_n) + \bar{P}_{1n}\bar{Q}_{1n}\bar{\alpha}_1 + \bar{P}_{1n}\bar{Q}_{1n}\mathbf{x}'_n\bar{\boldsymbol{\beta}} \\ p_{1n} + p_{2n} - f(\bar{\alpha}_2, \bar{\boldsymbol{\beta}}, \mathbf{x}_n) + \bar{P}_{3n}\bar{Q}_{3n}\bar{\alpha}_2 + \bar{P}_{3n}\bar{Q}_{3n}\mathbf{x}'_n\bar{\boldsymbol{\beta}} \end{bmatrix} \quad (n = 1, \ldots, N). \tag{6·8}$$

The system (6·8) of $2N$ equations provides a multiple regression model, where $P_{rn}Q_{rn}$ are to be estimated from the data, and where the errors are correlated in pairs. We wish to solve it by the recursive method, weighting inversely as the estimated variances. To this end, it is convenient to write the system (6·8) as the matrix equation

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \mathbf{X}^* = \mathbf{A}\mathbf{X}, \tag{6·9}$$

which can be done on appropriate definition of the several factors. Let

$$\mathbf{y}^* = \begin{bmatrix} y^*_{11} \\ y^*_{21} \\ \cdots \\ y^*_{1N} \\ y^*_{2N} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \cdots \\ \beta_s \end{bmatrix},$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & x_{11} & x_{12} & \cdots & x_{1s} \\ 0 & 1 & x_{11} & x_{12} & \cdots & x_{1s} \\ & & & \cdots & & \\ 1 & 0 & x_{N1} & x_{N2} & \cdots & x_{Ns} \\ 0 & 1 & x_{N1} & x_{N2} & \cdots & x_{Ns} \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \cdots \\ \epsilon_{1N} \\ \epsilon_{2N} \end{bmatrix},$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & & & 0 \\ & \mathbf{A}_2 & & \\ & & \ddots & \\ 0 & & & \mathbf{A}_N \end{bmatrix}, \quad \text{where} \quad \mathbf{A}_n = \begin{bmatrix} \hat{P}_{1n}\hat{Q}_{1n} & 0 \\ 0 & \hat{P}_{3n}\hat{Q}_{3n} \end{bmatrix}.$$

We also note that

$$\mathbf{x}'_n = \begin{bmatrix} 1 & 0 & x_{n1} & \cdots & x_{ns} \\ 0 & 1 & x_{n1} & \cdots & x_{ns} \end{bmatrix}$$

and $\mathbf{x}^{*\prime}_n = \mathbf{A}_n\mathbf{x}'_n$. Then $\mathbf{X}^*$ is just the matrix having $\mathbf{x}^{*\prime}_n$ as its nth row.

The variance of the vector $\boldsymbol{\epsilon}$ can now conveniently be written as a diagonal matrix of 2×2 submatrices of the form (6·5) and hence the estimated weights for regression are given by the diagonal matrix $\mathbf{W}$ with the 2×2 elements:

$$\mathbf{W}_n = \mathbf{V}^{-1}(\boldsymbol{\epsilon}_n) = \frac{1}{\hat{P}_{2n}} \begin{bmatrix} \frac{\hat{Q}_{3n}}{\hat{P}_{1n}} & -1 \\ -1 & \frac{\hat{Q}_{1n}}{\hat{P}_{3n}} \end{bmatrix}.$$

Now the desired estimators are available through iterative or, as we propose, recursive solution to yield

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}^{*\prime}\mathbf{W}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{W}\mathbf{y}^*;$$

also

$$\mathbf{V}(\hat{\boldsymbol{\theta}}) = (\mathbf{X}^{*\prime}\mathbf{W}\mathbf{X}^*)^{-1} = (\mathbf{X}'\mathbf{A}\mathbf{W}\mathbf{A}\mathbf{X})^{-1}$$

since here $\mathbf{A} \neq \mathbf{W}^{-1}$.

Now the derivation of the recursive equations goes through in a series of steps formally identical with those used in §5, leading to:

$$\mathbf{V}_n = \mathbf{V}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n^* D_n^{-1}\mathbf{x}_n^{*\prime}\mathbf{V}_{n-1}$$

and

$$\mathbf{b}_n = \mathbf{b}_{n-1} + \mathbf{V}_{n-1}\mathbf{x}_n^* D_n^{-1}(\mathbf{y}_n^* - \mathbf{x}_n^{*\prime}\mathbf{b}_{n-1}),$$

where

$$D_n = \mathbf{x}_n^{*\prime}V_{n-1}\mathbf{x}_n^* + \mathbf{W}_n^{-1}.$$

On rewriting in terms of the data,

$$\mathbf{V}_n = \mathbf{V}_{n-1} - \mathbf{V}_{n-1}\mathbf{x}_n\mathbf{A}_n D_n^{-1}\mathbf{A}_n\mathbf{x}_n'\mathbf{V}_{n-1} \tag{6·10}$$

and

$$\mathbf{b}_n = \mathbf{b}_{n-1} + \mathbf{V}_{n-1}\mathbf{x}_n\mathbf{A}_n\mathbf{D}_n^{-1}\begin{bmatrix} p_{1n} - \hat{P}_{1n} \\ p_{1n} + p_{2n} - (\hat{P}_{1n} + \hat{P}_{2n}) \end{bmatrix}. \tag{6·11}$$

Here $\mathbf{V}_n$ is obtained by inverting the 2×2 matrix $\mathbf{D}_n = \mathbf{A}_n\mathbf{x}_n'\mathbf{V}_{n-1}\mathbf{x}_n\mathbf{A}_n + \mathbf{W}_n^{-1}$, the corresponding factor d_n in the dichotomous case having been a scalar.

The advantage of this model over the dichotomous one is in the additional information being used to estimate $\boldsymbol{\beta}$. Gurland *et al.* (1960) showed, as one would expect intuitively, that for the bioassay situation 'If the response...is polychotomous, it is more efficient to use this information explicitly...rather than pool certain outcomes in order to make the response dichotomous'.

The same ideas readily extend to a recursive analysis with a k-chotomous response, $k > 3$. The errors are correlated in groups of $k-1$. Recursive expressions for $\mathbf{b}_n$ and $\mathbf{V}_n$ are obtained, similar to (6·10) and (6·11). In particular the matrix which must be inverted to obtain $\mathbf{V}_n$ is now $(k-1) \times (k-1)$.

7. Numerical illustration in the dichotomous case

To illustrate the method, a program to implement the dichotomous model on a digital computer was written and applied to the records of 5209 participants in the Framingham Study. Between 4 and 8 min. are required on the IBM 7094 computer to estimate any given set of parameters (up to 19, the largest number tried so far), using two iterations.

Table 1 shows the computational results for a particular subset of all the regressors available in the Framingham data (independent variables will be characterized as 'regressors' here, previous sections having made the sense of this term clear). The dependent or

Table 1. *Computational results for one set of regressors*

1*a*. Brief definition of regressors:

Intercept (INT);
Sex: coded 1 (male), 2 (female);
Age: in years and hundredths;
Height (HT): in inches and hundredths;
Systolic blood pressure (SYS): in mg. Hg;
Diastolic blood pressure (DIA): in mg. Hg;
Serum cholesterol (SCH): in mg. per ml.;
Electrocardiographic abnormalities (ECG): coded 0 (absent), 1 (present);
Framingham relative weight (FWT): weight (lb.) divided by median weight (lb.) of sex-height group, as a percentage;

Alcohol consumption (ALC) (oz./month):

Code	Oz. (men)	Oz. (women)
0	None	None
1	< 4	< 1
2	5–14	1–9
3	15–39	10–24
4	40–69	25–39
5	70–99	40–999
6	100–999	

1*b*. Estimated coefficients and associated values of t

j	Regressor	b_j	t_j	j	Regressor	b_j	t_j
0	INT	−5·3695	—	5	DIA	0·005493	0·81
1	Sex	−1·5883	−9·12	6	SCH	0·006631	5·41
2	Age	0·08095	10·15	7	ECG	0·8543	4·99
3	HT	−0·05279	−2·28	8	FWT	1·3586	3·77
4	SYS	0·009116	2·50	9	ALC	−0·05873	−1·60

1*c*. Analysis of variance (adjusted for the mean)

Source	D.F.	S.S.	M.S.	F
Regression	9	343·15	38·1277	41·65
Error	4661	4267·14	0·9155	—
(Theoretical error	∞	4661·00	1·00)	—
Total	4670	4610·29	—	—

response variable, as in all cases discussed in this section, is coronary heart disease, coded zero if absent and one if present, as determined by well-defined diagnostic criteria approximately 10 years from the date of determination of the regressor values for the given individual.

In addition to the estimated regression coefficients b_j, Table 1 shows the values $t_j = b_j/s_{b_j}$ of their sizes relative to their standard deviations. The overall analysis of variance is also exhibited in the table. With the sample size involved this may be interpreted as though based on a normally distributed dependent variable, with an expected error mean square of unity. The evidence of overall regression ($F = 41{\cdot}65$) is strong as is that for several of the individual regressors.

Table 2 shows t and F values for nine other similar analyses with different combinations of regressors in the Framingham data. The structure of the regression equations in each case provides information of considerable interest about the disease and the factors with which the risk of its occurrence is associated and lays a foundation for more detailed investigations.

TABLE 2. *Values of t and F for nine different combinations of regressors*

2*a*. Brief definition of regressor not defined in Table 1*a*:

Cigarettes smoked (CIG); coded

0 if none
1 if < 1 pack/day
2 if 1 pack/day
3 if > 1 pack/day

2*b*. Values of *t* and *F*:

Regressor	*t*								
Sex	−9·85	−10·44	−10·68	−9·91	−10·95	−9·47	−9·66	−9·12	−6·21
Age	13·45	12·68	11·56	11·06	11·54	10·79	10·79	10·15	9·21
HT	—	—	—	−2·78	—	−2·85	−2·64	−2·28	−1·83
SYS	—	—	2·94	2·78	2·66	2·04	1·68	2·50	0·55
DIA	8·12	7·32	2·03	1·16	1·41	2·30	1·55	0·81	2·04
SCH	—	6·05	6·01	6·02	6·17	6·01	6·13	5·41	3·85
ECG	—	—	—	—	—	5·91	5·60	4·99	4·69
FWT	—	—	—	—	3·48	—	3·43	3·77	3·69
ALC	—	—	—	—	—	—	—	−1·60	−2·14
CIG	—	—	—	—	—	—	—	—	5·44
*N**	5195	5074	5074	5068	5036	5068	5036	4671	2687
F	133·23	100·87	81·10	68·10	65·38	62·73	52·47	41·65	24·39

* Sample size after removal of observations with missing values of one or more regressors.

8. CONCLUSION; RANGE OF APPLICATIONS

In this paper we have proposed a method of data analysis two features of which seem worth stressing. The first is the utility of a symmetric sigmoid transform, such as the logit, in analyzing the dependence of zero-one data on a relatively large number of independent variables (Cox, 1966). By effecting a workable linearization such a transform submits the zero-one data not only to the techniques of multiple regression, but to its extensions, such as analysis of variance and analysis of covariance.

The second is the effectiveness of the recursive technique in estimating the regression coefficients. This is a method widely applicable in non-linear estimation. By allowing convergence within each of the usual iterations, the method is less dependent on initial approximations and converges more rapidly.

We developed the method especially for analysis of the data of large, long-term prospective studies in epidemiology, but there are indications that it will have a broader range of applications. Much of the data in modern biomedical research is of a zero-one or polychotomous character, involving relatively large numbers of independent variables, and the method is usually applicable in such cases. It is now being applied in several other areas, by investigators using our program or variants of it. These include: estimation of probabilities from meteorological data; a marketing problem, involving estimation of the probability of a purchase, given the values of the independent variables characterizing the customers; and a politico-social problem, involving the probability of certain actions by a citizen, given his social, political and economic status. Application to a military problem where the data are hits, near-misses and misses against certain types of targets, is being considered.

The bulk of this work was done while the first author was an N.I.H. Special Fellow in the Biostatistics Department, Johns Hopkins School of Hygiene and Public Health, forming part of a dissertation written under the second author's guidance. Programming and data processing support were supplied by Dr S. A. Talbot, then head of the Biomedical Engineering Program, Johns Hopkins Medical School, under a N.I.H. General Medical Science Grant, particular acknowledgement being due to Mr David Frederick for his skilful contributions. Finally, we acknowledge the generosity of Dr Thomas Dawber and Jerome Cornfield, who made a large part of the Framingham Study data available for this investigation.

REFERENCES

AITCHISON, J. & SILVEY, S. D. (1951). The generalization of probit analysis to the case of multiple responses. *Biometrika* **44**, 131–40.

ASHFORD, J. R. (1959). An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics* **15**, 573–81.

CHAPMAN, J. M., GOERKE, L. S., DIXON, W., LOVELAND, D. B. & PHILLIPS, E. (1957). The clinical status of a population group in Los Angeles under observation for two to three years. *Am. J. Publ. Hlth.* **47**, 33–42.

CHAPMAN, J. M. & MASSEY, F. J. (1964). The interrelationship of serum cholesterol, hypertension, body weight, and risk of coronary disease. *J. Chron. Dis.* **17**, 933–49.

CORNFIELD, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Fedn. Proc.* **21**, no. 4, Pt. II, suppl. no. 11, 58–61.

CORNFIELD, J., GORDON, T. & SMITH, W. W. (1961). Quantal response curves for experimentally uncontrolled variables. *Bull. Inst. Int. Statist.* **38**, 97–115.

COX, D. R. (1958). The regression analysis of binary sequences. *J. R. Statist. Soc.* B **20**, 215–32.

COX, D. R. (1966). Some procedures connected with the logistic qualitative response curve. *Research papers in statistics: essays in honour of J. Neyman's 70th birthday* (ed. by F. N. David). London: Wiley.

DAWBER, T. R., MEADORS, G. F. & MOORE, F. E., Jr. (1951). Epidemiological approaches to heart disease: the Framingham study. *Am. J. Publ. Hlth.* **41**, 279–86.

DOYLE, J. T., HESLIN, A. S., HILLEBOE, H. E. & FORMEL, P. F. (1959). Early diagnosis of ischemic heart disease. *New Engl. J. Med.* **261**, 1096–101.

DUNCAN, D. B. & RHODES, R. D. (1952). Multiple regression with a quantal response. (Abstract). *Ann. Math. Statist.* **23**, 300.

FINNEY, D. J. (1964). *Statistical Method in Biological Assay*. 2nd Ed. London: Griffin.

GARWOOD, F. (1941). The application of maximum likelihood to dosage mortality curves. *Biometrika* **32**, 46–58.

GERTLER, M. M., WOODBURY, M. A., GOTTSCH, L. G., WHITE, P. D. & RUSK, H. A. (1959). The candidate for coronary heat disease; discriminating power of biochemical, hereditary and anthropometric measurements. *J. Am. Med. Ass.* **170**, 149–52.

GURLAND, J., LEE, J. & DAHM, P. A. (1960). Polychotomous quantal response in biological assay. *Biometrics* **16**, 382–98.

HILLBOE, H. E., JAMES, G. & DOYLE, J. T. (1954). Cardiovascular Health Unit: I. Project design for public health research. *Am. J. Publ. Hlth* **44**, 851–63.

KAGAN, A., KANNEL, W. B., DAWBER, R. T. & REVOTSKIE, H. (1963). The coronary profile. *Ann. N.Y. Acad. Sci.* **97**, 883–94.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Bas. Engng.* (ASME Trans.) **82**D, 35–45.

KANNEL, W. B. (1964). Cigarette smoking and coronary heart disease. *Ann. Intern. Med.* **60**, 1103–6.

KANNEL, W. B., DAWBER, T. R., FRIEDMAN, G. D., GLENNON, W. E. & MCNAMARA, P. (1964). Risk factors in coronary heart disease. *Ann. Intern. Med.* **61**, 888–99.

KANNEL, W. B., DAWBER, T. R., KAGAN, A. & REVOTSKIE, N. (1962). Epidemiology of coronary heart disease. *Geriatrics* **17**, 675–90.

KANNEL, W. B., DAWBER, T. R., KAGAN, A., REVOTSKIE, N. & STOKES, J. (1961). Factors of risk in the development of coronary heart disease—six-year follow-up experience. *Ann. Intern. Med.* **55**, 33–50.

PAUL, O., KEPPER, M. H., PHELAN, W. H., DUPERTUIS, G. W., MACMILLAN, A., MCKEAN, H. & HEEBOK, P. (1963). A longitudinal study of coronary heart disease. *Circulation* **28**, 20–31.

SWERLING, P. (1959). First order error propagation in a stagewise smoothing procedure for satellite observations. *J. Astronaut. Sci.* **6**, 46–52.

[*Received June* 1966. *Revised January* 1967]