The Annals of Statistics
1974, Vol. 2, No. 3, 558-561

CLASSIFICATION AND ESTIMATION OF SEVERAL MULTIPLE REGRESSIONS

BY IVOR FRANCIS AND SAMPRIT CHATTERJEE

Cornell University and New York University

Two problems, classifying an individual into one of several populations and estimating the regression in that population, are simultaneously treated as one problem. This can be viewed as a problem of a missing observation on a categorical variable. When all variables are jointly distributed multivariate normal, the maximum likelihood solution is the intuitively appealing one: classify the individual using the usual likelihood ratio procedure, then estimate the regression using the observations from the selected population.

1. Introduction. The classification problem is usually treated as two separate problems: the estimation of the parameters of the several populations followed by the classification of a new object into one of these populations. The two problems are often solved separately, each solution invoking different optimality criteria, so that the properties of the combined procedure are unclear.

In this paper we consider the following problem. Each of k populations has a different regression of a dependent variable on several independent variables, all of which are jointly distributed multivariate normal. We have measurements on all variables for a random sample of individuals from each population. For a new individual we have measurements on the independent variables, and we require an estimate of the dependent variable, but we do not know to which population this individual belongs. Alternatively, this problem could be viewed as a missing observation problem in which the dependent variable and all but one of the independent variables are multivariate normal, but where the last independent variable is of the categorical type, specifying group membership, and where one of those categorical observations is missing. In a review of the literature on missing values, Afifi and Elashoff (1966) remark that all writers on missing value problems who use the method of maximum likelihood consider only the case in which all variables are multinormal.

We approach this problem here as one single estimation problem, the simultaneous estimation of the regressions and of a membership parameter θ_i, which has the value 1 if the new individual belongs to the ith population, and the value 0 otherwise, a device used by Hartley and Rao (1968). The maximum likelihood solution is found to be the natural and intuitively appealing one.

2. Likelihood. Let Y denote the dependent variable and $\mathbf{X}$ the $(p \times 1)$ vector of independent variables. Suppose, in the ith population, $i = 1, 2, \cdots, k$, the

Received November 1972; revised July 1973.
AMS 1970 *subject classifications.* Primary 62H30; Secondary 62J05.
Key words and phrases. Classification, regression, missing observations.

distribution of Y conditional on $\mathbf{X} = \mathbf{x}$ is normal, $Y \sim N(\mathbf{x}'\boldsymbol{\beta}_i, \sigma^2)$, and the distribution of $\mathbf{X}$ is multivariate normal, $\mathbf{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\beta}_i$ and $\boldsymbol{\mu}_i$ are the $(p \times 1)$ vectors of regression coefficients and means respectively, and $\boldsymbol{\Sigma}$ is the $(p \times p)$ covariance matrix of full rank common to all populations. Let the unknown value of the dependent variable for the new individual be denoted by y_0 and the known values for the independent variables by $\mathbf{x}_0$. Suppose there are n_i observations from the ith population, and let $\mathbf{y}_i$ denote the $(n_i \times 1)$ vector of observations on Y, and $\mathbf{x}_i = (\mathbf{x}_{i1}\mathbf{x}_{i2} \cdots \mathbf{x}_{in_i})$ the $(p \times n_i)$ matrix of observations on $\mathbf{X}$ in the ith population. Then the likelihood function $L = L(y_0, \sigma^2, \boldsymbol{\Sigma}, \theta_i, \boldsymbol{\beta}_i, \boldsymbol{\mu}_i) =$ constant $\times\ L_1 \times L_2 \times L_3 \times L_4$, where

$$
\begin{aligned}
L_1 &= \sigma^{-n} \exp\left[-(2\sigma^2)^{-1} \sum (\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_i)'(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta}_i)\right], \\
L_2 &= \sigma^{-1} \exp\left[-(2\sigma^2)^{-1}(y_0 - \sum \theta_i \mathbf{x}_0'\boldsymbol{\beta}_i)^2\right], \\
L_3 &= |\boldsymbol{\Sigma}|^{-n/2} \exp\left[-2^{-1} \sum \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_{ij} - \boldsymbol{\mu}_i)\right], \\
L_4 &= |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-2^{-1} \sum \theta_i(\mathbf{x}_0 - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_0 - \boldsymbol{\mu}_i)\right],
\end{aligned}
$$

where all summations are from $i = 1$ to k, unless otherwise indicated.

3. Conditional maximum likelihood estimates. Let $\hat{y}_0$, $\hat{\sigma}^2$, $\hat{\boldsymbol{\Sigma}}^{(s)}$, $\hat{\boldsymbol{\beta}}_i$, $\hat{\boldsymbol{\mu}}_i$, $i = 1, 2, \cdots, k$, denote the conditional maximum likelihood estimates of the respective parameters given $\theta_s = 1$. In succession we set equal to zero the partial derivatives of $\log L$ with respect to (i) y_0, (ii) $\boldsymbol{\beta}_i$, $i \neq s$, (iii) $\boldsymbol{\beta}_s$, (iv) σ^2, and solve, giving

$$
\begin{aligned}
&\text{(i)} && \hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}_s, \\
&\text{(ii)} && \mathbf{x}_i\mathbf{x}_i'\hat{\boldsymbol{\beta}}_i = \mathbf{x}_i\mathbf{y}_i, && i \neq s, \\
&\text{(iii)} && (\mathbf{x}_s\mathbf{x}_s' + \mathbf{x}_0\mathbf{x}_0')\hat{\boldsymbol{\beta}}_s = \mathbf{x}_s\mathbf{y}_s + \mathbf{x}_0\hat{y}_0.
\end{aligned}
$$

Substituting from (i) into (iii) gives $\mathbf{x}_s\mathbf{x}_s'\hat{\boldsymbol{\beta}}_s = \mathbf{x}_s\mathbf{y}_s$. Hence

$$
\hat{\boldsymbol{\beta}}_i = (\mathbf{x}_i\mathbf{x}_i')^{-1}(\mathbf{x}_i\mathbf{y}_i), \qquad \text{for all } i = 1, 2, \cdots, k.
$$

$$
\text{(iv)} \qquad \hat{\sigma}^2 = (n+1)^{-1}\left[\sum (\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_i)'(\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_i) + (\hat{y}_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}}_s)^2\right].
$$

Substituting from (i) into (iv) gives

$$
\hat{\sigma}^2 = (n+1)^{-1} \sum (\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_i)'(\mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_i).
$$

In the expression for the likelihood function only the terms L_3 and L_4 contain $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$. Hence we can find the conditional estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ given $\theta_s = 1$ by maximizing $L_3 \times L_4$, after putting $\theta_s = 1$. But if $\theta_s = 1$, all other $\theta_i = 0$, $i \neq s$, and we simply have a situation where there are n_i observations from $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i \neq s$, and $n_s + 1$ from $N(\boldsymbol{\mu}_s, \boldsymbol{\Sigma})$. Thus the conditional estimates $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}^{(s)}$ can be obtained from standard multivariate theory (Anderson (1958), page 248):

$$
\hat{\boldsymbol{\Sigma}}^{(s)} = (n+1)^{-1}\mathbf{A}^{(s)},
$$

where

$$\mathbf{A}^{(s)} = \sum_{i=1;i\neq s}^{k} \mathbf{A}_i + \mathbf{A}_{s+} ,$$

$$\mathbf{A}_i = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)' ,$$

$$\mathbf{A}_{i+} = \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i+})(\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i+})' + (\mathbf{x}_0 - \bar{\mathbf{x}}_{i+})(\mathbf{x}_0 - \bar{\mathbf{x}}_{i+})' ,$$

$$\bar{\mathbf{x}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{x}_{ij} , \qquad \text{and}$$

$$\bar{\mathbf{x}}_{i+} = (n_i + 1)^{-1}[\sum_{j=1}^{n_i} \mathbf{x}_{ij} + \mathbf{x}_0] .$$

Finally

$$\hat{\boldsymbol{\mu}}_i = \bar{\mathbf{x}}_i , \quad i \neq s \qquad \text{and} \qquad \hat{\boldsymbol{\mu}}_s = \bar{\mathbf{x}}_{s+} .$$

It can be shown (Anderson (1958), page 141) that

$$\mathbf{A}_{i+} - \mathbf{A}_i = n_i(n_i + 1)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_i)(\mathbf{x}_0 - \bar{\mathbf{x}}_i)' .$$

Hence, if we let

$$\mathbf{C} = \sum_{i=1}^{k} \mathbf{A}_i ,$$

then

$$\begin{aligned} \mathbf{A}^{(s)} &= \mathbf{C} + \mathbf{A}_{s+} - \mathbf{A}_s \\ &= \mathbf{C} + n_s(n_s + 1)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_s)(\mathbf{x}_0 - \bar{\mathbf{x}}_s)' . \end{aligned}$$

4. Maximum likelihood estimate of s. When all these conditional estimates are substituted into the expression for the likelihood function, L_1 and L_2 become independent of s, and so the conditional maximum of the likelihood, (Anderson (1958), page 248) is

$$\text{constant} \times |\hat{\boldsymbol{\Sigma}}^{(s)}|^{-(n+1)/2} .$$

Thus the maximum likelihood estimate of s is the value that maximizes this conditional maximum, or the value that minimizes $|\hat{\boldsymbol{\Sigma}}^{(s)}|$ or equivalently $|\mathbf{A}^{(s)}|$. But, provided $\mathbf{C}$ is of full rank,

$$|\mathbf{A}^{(s)}| = |\mathbf{C}||1 + n_s(n_s + 1)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_s)'\mathbf{C}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_s)| ,$$

and so we want the value of s that minimizes the distance $n_s(n_s + 1)^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_s)' \times \mathbf{C}^{-1}(\mathbf{x}_0 - \bar{\mathbf{x}}_s)$, which is the usual likelihood ratio criterion for classification, and which becomes the commonly used linear discriminant function when all n_i are equal.

5. Summary. As far as the missing observation problem is concerned, the estimates of all the $\boldsymbol{\beta}_i$ do not depend on the observation with the missing value, and the estimates of σ^2, $\boldsymbol{\Sigma}$, and the $\boldsymbol{\mu}_i$ depend on this observation by an amount that decreases as all the n_i increase. (This can be compared with results cited by Afifi and Elashoff (1966), page 601.)

The estimation of y for this new observation consists of first classifying the individual by the usual likelihood ratio criterion, then predicting y by the usual multiple regression estimated using the sample from the selected population. This is the commonly used, and therefore intuitively appealing, solution (consider medical diagnosis and treatment), but its properties other than its maximum likelihood optimality, for example its mean squared error, should be evaluated relative to those of alternative procedures.

REFERENCES

[1] AFIFI, A. A. and ELASHOFF, R. M. (1966). Missing observations in multivariate statistics I. Review of the literature. *J. Amer. Statist. Assoc.* **61** 595–604.

[2] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistics.* Wiley, New York.

[3] HARTLEY, H. O. and RAO, J. N. K. (1968). Classification and estimation in analysis of variance problems. *Rev. Inst. Internat. Statist.* **36** 141–147.

STATISTICS CENTER
CORNELL UNIVERSITY
ITHACA, NEW YORK 14850

QUANTITATIVE ANALYSIS AREA
NEW YORK UNIVERSITY
68 TRINITY PLACE
NEW YORK, NEW YORK 10006