TRAINING & RESEARCH

TRA RCH

TRA RCH

TRA RCH

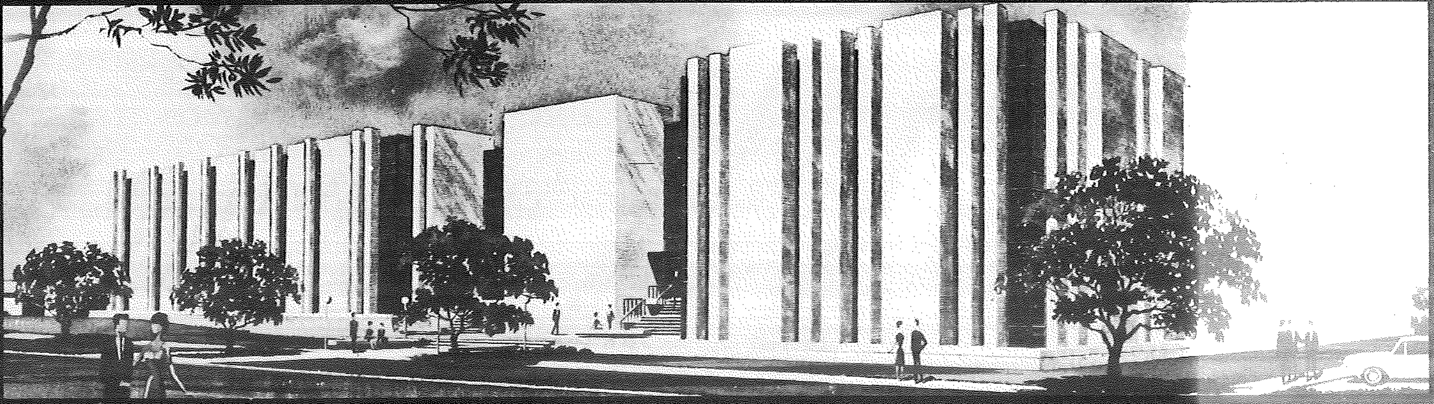
MULTINOMIAL SELECTION INDEX

by

W. B. Smith and D. M. Scott

TRAINING & RESEARCH

GRADUATE
INSTITUTE
OF
STATISTICS



TEXAS A&M UNIVERSITY · COLLEGE STATION

MULTINOMIAL SELECTION INDEX

by

W. B. Smith and D. M. Scott

Institute of Statistics
Texas A&M University

Technical Report #5
National Aeronautics and Space Administration
Research Grant NGR 44-001-097

April 1971

Multinomial Selection Index

by

W. B. Smith and D. M. Scott
Institute of Statistics
Texas A&M University
College Station Texas, U.S.A. 77843

1. Introduction

The linear selection index developed by H. F. Smith [1936] is a linear combination of the elements of the phenotypic observation vector X_j,

$$I_j = b'X_j \ , \tag{1}$$

where I_j denotes the composite index value associated with the j^{th} member of a population and b is an n-vector of unknown coefficients (weights). This index was conceived to aid in discriminating between selection programs among varieties of plants. Assuming X_j was distributed as a multivariate normal with covariance matrix P, Smith showed that the optimal choice of b (i.e., yielding greatest expected genetic advance) is

$$b = P^{-1}G\alpha \ , \tag{2}$$

where G is the genotypic covariance matrix and α is an n-vector of economic weights.

Since Smith's paper much research has been conducted on the linear index and its nonlinear competitors. Notable among these are Henderson [1963], Kempthorne and Nordskog [1959], Williams [1962], Hazel [1943] and VanVleck [1970]. See Williams for a thorough review.

Smith and Pfaffenberger [1970] considered index estimation using multivariate normal phenotypic observations, both full and partially complete vectors, assuming G and α are known, but P unknown. This procedure applied a technique of Hocking and Smith [1968] for estimating the parameters of a multivariate normal distribution in the presence of partial data. All data is used and several alternate methods are presented for indexing those individuals possessing partial records. A contrast between Henderson's techniques and that of Smith and Pfaffenberger (S-P in the sequel) is given.

This paper considers the linear selection index as described by (1) with b chosen as in (2) assuming both full and partial records are available and that the phenotypic vectors follow a multinomial distribution. Thus, this index deviates both from the assumption that the phenotypic covariance matrix is known and from the assumption of normality.

An estimation procedure similar to that of Hocking and Oxspring [1971] is discussed and certain simulation studies are presented to support the claimed optimality properties. In addition, the S-P multivariate normal technique is applied to multinomial data for comparison with the multinomial estimation procedure.

2. Estimation Procedure

Consider a phenotypic observation vector $X' = (X_1 \ldots X_k)$ which is distributed multinomially with known parameter M and unknown parameters $\theta' = (\theta_1 \ldots \theta_k)$. That is,

$$P(X = x_j,\ j = 1, \ldots, k) = \left(\frac{M!}{\prod_{j=1}^{k+1} x_j!}\right) \prod_{j=1}^{k+1} \theta_j^{x_j} , \qquad (3)$$

where $x_{k+1} = M - \sum_{j=1}^{k} x_j$ and $\theta_{k+1} = 1 - \sum_{j=1}^{k} \theta_j$. We desire to index each vector from a population distributed as (3); however, some of these vectors have missing elements (recall that any marginal distribution from a multinomial is again multinomial in form). As in Smith and Pfaffenberger [1970] all information, both full and partial vectors, is utilized in estimating $\theta_1, \ldots, \theta_k$ and thus to estimating each individual's index, assuming G and α known.

Following the outline of Hocking and Smith [1968] group the data vectors by which elements are missing, estimate within each group the available θ_j, and then optimally combine these estimates. For example, consider a population of size N where n_1 individuals have recorded all elements of the phenotypic observation vector while n_2 individuals have only the first (renumbering if necessary) $\ell < k$ elements recorded, $n_1 + n_2 = N$. Thus, from the full data group each parameter θ_j can be estimated unbiasedly by ${}_1\hat{\theta}_j$, $j = 1, \ldots, k$, whereas from the second group (partial data) only $\theta_1, \ldots, \theta_\ell$ can be estimated by ${}_2\hat{\theta}_j$, $j = 1, \ldots, \ell$. In each case, the usual maximum likelihood estimates are used. Combining these estimates as in Hocking and Smith yields

$$\tilde{\theta}_j = {}_1\hat{\theta}_j + \sum_{r=1}^{\ell} a_{rj}({}_1\hat{\theta}_r - {}_2\hat{\theta}_r), \; j = 1, \ldots, k \; . \qquad (4)$$

Note that $A_j' = (a_{1j}, \ldots, a_{\ell j})$ is chosen to minimize the variance of $\hat{\theta}_j$, $j = 1, \ldots, k$. If A_j' does not depend on the parameters θ', then $\tilde{\theta}_j$ is unbiased and minimum variance. In general, $\tilde{\theta}_j$ is consistent, asymptotically unbiased and asymptotically efficient when full data estimates of θ' are used in A_j'.

A general formulation for A_j can be given. Let V be the covariance matrix of $(X_1, \ldots, X_k)$. Thus,

$$V = \text{Diag}\ (\theta) - \theta\theta' \; .$$

Then the covariance matrix of ${}_1\hat{\theta}'$ is given by V/n_1M and for ${}_2\hat{\theta}'$ by D_2VD_2'/n_2M, where $D_2 = (I_\ell \vdots 0)$ and I_ℓ is an identity matrix of order ℓ. Thus,

$$A = \begin{bmatrix} A_1' \\ \cdots \\ \cdots \\ A_k' \end{bmatrix} = - \frac{n_2}{n_1 + n_2} (D_2VD_2')^{-1} D_2V .$$

Note that $D_2\ \theta = {}_2\theta$, where ${}_2\theta' = (\theta_1, \ldots, \theta_\ell)$.

If in addition, there is a third data group of n_3 multinomial vectors with parameters M and $D_3\theta$, D_3 is a s x k unitary matrix of ones and zeros, then new estimates would combine $\tilde{\theta}$ with $\{{}_3\hat{\theta}_j\}$, the estimates from this third group. In such a case, in matrix notation

$$\tilde{\tilde{\theta}} = \tilde{\theta} + B'(D_3\tilde{\theta} - {}_3\hat{\theta}) .$$

Note that D_3 makes $\tilde{\theta}$ conformable to ${}_3\hat{\theta}$. B is chosen to minimize the variance of $\tilde{\tilde{\theta}}$ and satisfies

$$[D_3(W + V/n_3M)D_3']B = D_3W ,$$

where W is the covariance matrix for the combined first two data groups. That is,

$$W = [(V/n_1M)^{-1} + D_2'(D_2VD_2')/n_2M)D_2]^{-1}$$

$$= [n_1 + n_2\ VD_2'(D_2VD_2')D_2]^{-1}\ V/M .$$

To estimate θ' when more than three groups of data are available, continue in the fashion outlined above, producing at each stage the estimated asymptotic covariance matrix of the combined estimate for use at the next stage.

Now to achieve an estimate based on all data the final estimate of θ' is used for estimation of the covariance matrix P. That is, in the case above $\tilde{\tilde{\theta}}_j$ is substituted for θ_j in the formulation of P yielding a matrix $\tilde{\tilde{P}}$. Then set

$$\tilde{\tilde{b}} = \tilde{\tilde{P}}^{-1} G\alpha .$$

With $\tilde{\tilde{b}}$ an index for each phenotypic vector can be given; in those cases of missing data the final estimate for the mean of that element is substituted to produce a full vector to index. In general, the procedure yields consistent and asymptotically efficient estimates. Note that estimated phenotypic means, variances and covariances are available upon termination.

Example 1: Let $k = 3$, $\ell = 2$, as above. Then,

$$\tilde{\theta} = \hat{\theta} + A(D_2\hat{\theta} - {}_2\hat{\theta}) .$$

In this case $D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$ and

$$A = \frac{n_2}{n_1 + n_2} \begin{bmatrix} -1 & 0 & \theta_3/(1 - \theta_1 - \theta_2) \\ 0 & -1 & \theta_3/(1 - \theta_1 - \theta_2) \end{bmatrix} .$$

Thus, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are just the weighted sums of ${}_1\hat{\theta}_1$, ${}_2\hat{\theta}_1$ and ${}_1\hat{\theta}_2$, ${}_2\hat{\theta}_2$, respectively, and are unbiased and minimum variance. Note, however, that the coefficients for $\tilde{\theta}_3$ depend on θ', thus $\tilde{\theta}_3$ will be a consistent, asymptotic efficient estimate when ${}_1\hat{\theta}_j$ is substituted for θ_j, $j = 1, 2, 3$.

Example 2: Consider the same situation as in Example 1, but with a third data group, s = 1. Then

$$\tilde{\tilde{\theta}} = \tilde{\theta} + B(\tilde{\theta}_1 - {}_3\hat{\theta}_1) ,$$

where

$$D_3 = (1 \quad 0 \quad 0)$$

and

$$B = - \left[\left(\frac{1}{n_1 + n_2} + \frac{1}{n_3}\right) \theta_1(1 - \theta_1) \right]^{-1} \begin{bmatrix} \theta_1(1 - \theta_1)/(n_1 + n_2) \\ \theta_1\theta_2/(n_1 + n_2) \\ \theta_1\theta_3/(n_1 + n_2) \end{bmatrix}$$

$$= \frac{-n_3}{(n_1 + n_2 + n_3)} \begin{bmatrix} 1 \\ \theta_2/(1 - \theta_1) \\ \theta_3/(1 - \theta_1) \end{bmatrix} .$$

Again the coefficient depends on the parameters to be estimated, but substitution by the "best" previous estimates (i.e., $\tilde{\theta}$) yield consistent and asymptotically efficient estimates.

Example 3: Consider the same situation as in Example 1, but with a third data group containing information on X_2 and X_3.

$$\tilde{\tilde{\theta}} = \tilde{\theta} + C' \begin{bmatrix} \tilde{\theta}_2 - {}_3\hat{\theta}_2 \\ \\ \tilde{\theta}_3 - {}_3\hat{\theta}_3 \end{bmatrix} ,$$

where

$$D_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$C = -\frac{n_3}{n_1+n_2+n_3}\begin{bmatrix}\theta_2(1-\theta_2) & -\theta_2\theta_3 \\ -\theta_2\theta_3 & \theta_3(1-\theta_3)\end{bmatrix}^{-1}\begin{bmatrix}-\theta_1\theta_2 & \theta_2(1-\theta_2) & -\theta_2\theta_3 \\ -\theta_1\theta_3 & -\theta_2\theta_3 & \theta_3(1-\theta_3)\end{bmatrix}.$$

3. Simulation Studies

In the

In the following Monte Carlo simulation studies we set $\alpha' = (1 \quad 1 \quad 1)$ and

$$G = \begin{bmatrix} 2 & 0.75 & 2 \\ 0.75 & 3 & 1.5 \\ 2 & 1.5 & 4 \end{bmatrix}$$

with $\theta' = (.15, .25, .40)$ in each case.

Table 1 records a summary of simulation studies (200 runs) of Example 2 where $M_1 = M_2 = M_3 = 20$, $n_1 = 100$, $n_2 = 50$, $n_3 = 25$, and clearly indicates the greater precision achieved by using the partial data vectors. Table 2 summarizes a similar experiment with the same data configuration but in the presence of a much higher percentage (80%) of partial data vectors ($n_1 = 10$, $n_2 = 15$, $n_3 = 25$). Again reductions in the sample variance of the estimates are noted. In both cases the estimates of θ' are virtually unbiased but the estimabe of $b' = (b_1, b_2, b_3)$ are biased slightly upward. These examples, of course, are for a situation for which we have "nested" data, and in such situations Hocking and Oxspring [1971] have shown that this technique yields maximum likelihood estimates.

Tables 3 and 4 summarize simulation conducted on Example 3 but with two different sample sizes. Each of these tables again reflect the consistency of the estimate and the reduction in variances achieved by utilizing the partial data. Again there is some bias noted in the b term.

Table 5 summarizes an Example 2 experiment using the same parameters as those of Table 1. That is, 175 vectors $X' = (X_1 \ X_2 \ X_3)$ were generated to follow a multinomial distribution with parameters $M = 20$, and $\theta' = (\ \theta_1 \ \ \theta_2 \ \ \theta_3\ )$. Each of these vectors is indexed by equations (1) and (2) using the population values for θ'. The order resulting is called the "true" order. It is desired to compare this order with the order resulting from estimated indexes in several different cases. First, all 175 full data vectors are used to estimate θ' and P, thus yielding an estimated index for each. The correlation between the estimated ordering and the "true" ordering is given by the first entry of Table 5. Next a missing data situation is created by randomly selecting 75 vectors and deleting X_3 from 50 of them, and X_2 and X_3 from 25 of them. Thus now we have available 100 full vectors and 75 partial vectors of two types. The procedure of Section 2 is applied to estimate θ' and P (and thus b). The index order resulting is compared to the "true" ordering yielding the second entry of the table. Finally, the Smith-Pfaffenberger [1970] multivariate normal indexing technique is applied to the partial multinomial data and the final entry is the correlation between the resulting ordering and the true ordering. It should be noted that in the above case the partial data vectors were indexed by means of applying the estimated b vector to the partial vectors where the missing element is in turn estimated by its expected value in the multinomial case and by regression estimate in the multivariate normal case. Further explanation of the regression estimate in the multivariate normal case is given in the paper by Smith and Pfaffenberger. It should be noted that the estimate of the population mean is precise as indicated by the simulation of estimates of θ_j.

Table 1

Simulation of Example 2 ($n_1 = 100$, $n_2 = 500$, $n_3 = 25$; 200 runs)

Mean Estimate						
Parameter	$\theta_1 = .15$	$\theta_2 = .25$	$\theta_3 = .40$	$b_1 = 2383$	$b_2 = 2169$	$b_3 = 2124$
1st Data Group	.1502	.2507	.4005	2401.4	2185.0	2141.3
2nd Adjoined	.1496	.2515	.4004.	2402.9	2184.9	2141.9
3rd Adjoined	.1490	.2517	.4006	2404.4	2183.2	2140.2

Sample Variances of Estimates						
	θ_1	θ_2	θ_3	b_1	b_2	b_3
1st Data Group	$.7976 \times 10^{-4}$	$.9419 \times 10^{-4}$	1.0223×10^{-3}	7151.0	6115.8	6759.9
2nd Adjoined	.4006	.6386	.8862	6494.8	5665.9	5725.7
3rd Adjoined	.3705	.6372	.8860	6394.8	5585.3	5648.8

Table 2

Simulation of Example 2 ($n_1 = 10$, $n_2 = 15$, $n_3 = 25$; 500 runs)

Mean Estimate						
Parameter	$\theta_1 = .15$	$\theta_2 = .25$	$\theta_3 = .40$	$b_1 = 2383$	$b_2 = 2169$	$b_3 = 2124$
1st Data Group	.1503	.2515	.3994	2455.2	2276.2	2181.5
2nd Adjoined	.1495	.2515	.3999	2438.6	2215.2	2171.3
3rd Adjoined	.1493	.2516	.4001	2434.4	2213.9	2171.2

Sample Variances of Estimates						
	θ_1	θ_2	θ_3	b_1	b_2	b_3
1st Data Group	$.6766 \times 10^{-3}$	$.8813 \times 10^{-3}$	1.939×10^{-3}	83296.2	77727.3	77617.3
2nd Adjoined	.2816	.3781	.8791	72107.6	67039.3	60450.5
3rd Adjoined	.1325	.3557	.8606	70834.9	65625.6	59290.9

Table 3

Simulation of Example 2 ($n_1 = 50$, $n_2 = 50$, $n_3 = 50$; 100 runs)

			Mean Estimate			
Parameter	$\theta_1 = .15$	$\theta_2 = .25$	$\theta_3 = .40$	$b_1 = 2383.3$	$b_2 = 2170$	$b_3 = 2125$
1st Data Group	.1486	.2514	.3995	2395.1	2171.8	2128.7
2nd Adjoined	.1488	.2511	.3995	2391.9	2170.8	2127.8
3rd Adjoined	.1487	.2518	.3991	2391.7	2168.9	2127.6

Sample Variances of Estimates						
	θ_1	θ_2	θ_3	b_1	b_2	b_3
1st Data Group	1.063×10^{-4}	1.921×10^{-4}	1.654×10^{-4}	12,832.8	9546.9	10,163.6
2nd Adjoined	.555	.969	1.449	10,686.5	8569.4	9167.9
3rd Adjoined	.512	.689	1.042	7407.8	7809.4	7300.4

Table 4

Simulation of Example 2 ($n_1 = 10$, $n_2 = 15$, $n_3 = 25$; 100 runs)

Mean Estimate						
	$\theta_1 = .15$	$\theta_2 = .25$	$\theta_3 = .40$	$b_1 = 2383.3$	$b_2 = 2170$	$b_3 = 2125$
1st Data Group	.1506	.2514	.3992	2454.8	2227.1	2181.9
2nd Adjoined	.1496	.2512	.4000	2437.8	2215.1	2171.7
3rd Adjoined	.1495	.2515	.3998	2418.7	2195.2	2152.3

Sample Variances of Estimates						
	θ_1	θ_2	θ_3	b_1	b_2	b_3
1st Data Group	$.700 \times 10^{-3}$	$.882 \times 10^{-3}$	1.215×10^{-3}	82,713.9	77,909.5	77,439.5
2nd Adjoined	.283	.380	.883	72,867.4	67,739.1	61,087.3
3rd Adjoined	.243	.194	.297	25,302.4	32,804.3	31,159.7

Table 5

Average Correlation With "True" Ranking (10 runs each) of Multinomial Data

Simulation of Example 2 ($n_1 = 100$, $n_2 = 50$, $n_3 = 25$)

Estimation Index Before Deletion	.7448	.7236
Multinomial Estimates	.7415	.7175
Smith-Pfaffenberger Multinomial Normal Estimates	.7356	.7151
	$\theta' = (.15, .25, .40)$	$\theta' = (.27, .08, .47)$

4. Conclusions

In this paper we develop a linear selection index using phenotypic observation vectors multinomially distributed and estimate the index value of each vector by estimating in an optimal, sequential fashion the parameters of the parent multinomial distribution. Moreover, the estimation procedure of Section 2 does not require that all data records be full (i.e., have no missing elements), but only that there exist full vectors. In addition, a vector with missing elements are indexed by multiplying b by that vector with its missing elements filled by the combined mean estimate (i.e., M $\tilde{\tilde{\theta}}_j$).

Thus, the index of Section 2 deviates from a "standard" index in that, first, we estimate b by estimating P, and, second, the parent phenotypic vector distribution is non-normal. We cite the results tabulated in Section 3 (Tables 1 to 4) as empirical indications of the procedure's properties, viz., consistency and asymptotic efficiency. For further theoretical justifications for a similar technique see Hocking and Oxspring [1971].

Note that in Table 5, the comparison of the ranking resulting from the S-P multivariate normal procedure and the Section 2 procedure indicates that use of an estimated index assuming a multivariate normal configuration does not lead to unwarranted results. Thus, the value of the Section 2 procedure would be in the slightly more precise ordering achieved and since during the indexing process both estimates of phenotypic means and covariance matrix are found.

Future problems include combining the multivariate normal and multinomial procedures to yield an index of vectors some of whose elements are continuously distributed, others discrete. In addition, nonlinear competitors for both estimate indices are being considered.

5. Acknowledgments.

This research was partially supported by the National Aeronautics and Space Administration, Research Grant NGR 44-001-097.

6. References

[1] Afifi, A. A. and Elashoff, R. M. (1967) "Missing Observations in Multivariate Statistics II. Point Estimation in Simple Linear Regression", Journal of The American Statistical Association, 62, 10-29.

[2] Hazel, L. N. (1943) "Genetic Basis for Selection Indices", Genetics, 28, 476-490.

[3] Henderson, C. R. (1963) "Selection Index and Expected Genetic Advance", NAS-NRC Publication 982, 141-163.

[4] Hocking, R. R. and Oxspring, H. H. (1971) "Maximum Likelihood Estimation with Incomplete Multinomial Data", Journal of the Americal Statistical Association 66, 65-70.

[5] Hocking, R. R. and Smith, W. B. (1968) "Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations," Journal of the American Statistical Association, 63, 159-173.

[6] Kempthorne, O. and Nordskog, A. W. (1959), "Restricted Selection Indices", Biometrics 15, 10-19.

[7] Smith, H. F. (1936) "A Discriminant Function for Plant Selection", Ann. Eugen Lond., 7, 240-250.

[8] Smith, W. B. and Pfaffenberger, R. C. (1970) "Selection Index Estimation from Partial Multivariate Normal Data", Biometrics 26, 625-639.

[9] VanVleck, L. D. (1970) "Index Selection for Direct and Maternal Genetic Components of Economic Traits", Biometrics 26, 477-484.

[10] Williams, J. S. (1962) "The Evaluation of a Selection Index", Biometrics, 18, 375-393.