ARPA ORDER NO.: 189-1
6D20 Human Resources Research

R-1685-ARPA
June 1976

Analytic Methods for Adjusting Subjective Rating Schemes

Richard V. L. Cooper and Gary R. Nelson

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

The research described in this Report was sponsored by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
6D20 Human Resources Research

R-1685-ARPA
June 1976

Analytic Methods for Adjusting Subjective Rating Schemes

Richard V. L. Cooper and Gary R. Nelson

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

Published by The Rand Corporation

PREFACE

This report was prepared as part of Rand's DoD Training and Manpower Management Program, sponsored by the Human Resources Research Office of the Defense Advanced Research Projects Agency (ARPA). With manpower issues assuming an ever greater importance in defense planning and budgeting, the purpose of this research program is to develop broad strategies and specific solutions for dealing with present and future military manpower problems. This includes the development of new research methodologies for examining broad classes of manpower problems, as well as specific problem-oriented research. In addition to providing analysis of current and future manpower issues, it is hoped that this research program will contribute to a better general understanding of the manpower problems confronting the Department of Defense.

This report presents a methodology for using supervisory evaluations of military personnel in models of manpower performance. Although the measurement of performance is crucial to many manpower models, frequently the only measures available are those obtained from supervisors. Past research has shown, however, that such ratings may be subject to biases, perhaps unintentional, making it difficult to determine the extent to which the ratings reflect "true" performance or the supervisor's own implicit rating scale.

This report provides a way of correcting for these biases. In particular, since researchers may want to assess the contribution of various factors to individual performance--often through the use of multiple regression models--it is necessary to have a method of adjusting the subjective measure of performance. The resulting approach to the problem--the multi-scale model--suggests that supervisory evaluations of individuals are subject to two types of biases. The first is the familiar location bias--that is, some supervisors may grade "easy" while other grade "hard." The second is a scale effect--that is, some supervisors may exaggerate differences among individuals while others may minimize these differences. The generic name for the methodology presented here--*the multi-scale model*--derives from the latter bias.

This research was motivated by previous Rand research under the DoD Training and Manpower Management Program that was concerned with measuring the cost of on-the-job training for first-term enlisted personnel. Indeed, the basic idea was first sketched out in Robert M. Gay, *Estimaing the Cost of On-the-Job Training in Military Occupations: A Methodology and Pilot Study*, The Rand Corporation, R-1351-ARPA, April 1974. It was decided to extend the brief discussion of the model contained there, both because the use of supervisory ratings is important to manpower planners and researchers in general and because on-going research at Rand dealing with first-term enlisted personnel performance requires such a model.

This report presents the model in the context of an extension to the classical regression model. The report is technical in nature and assumes that the reader has a good understanding of standard econometric theory.

Finally, although the methodology presented here was originally developed to deal with the supervisory ratings problem, it may be applicable to a number of other econometric problems, such as seasonal adjustment and other cases in which data fall into natural groupings.

SUMMARY

Subjective evaluations of individual performance, such as those provided by supervisors, may be subject to certain kinds of biases. Yet subjective evaluations are a common and frequently the only source of information about a person's performance and can therefore be an important element in the application of manpower policy. Furthermore, the development and application of appropriate manpower policies may depend on measuring the effects of specific variables on individual performance; it is therefore important to correct for biases in subjective evaluations of individual performance.

This report is concerned with the development of statistical and econometric techniques for correcting for biases in models of individual performance. The approach developed here is a variant of the classical linear regression model. Specifically, it is proposed that supervisory ratings may be subject to two types of bias. The *location* bias results when supervisors systematically overestimate or underestimate individual performance. The *scale* bias results when supervisors exaggerate or minimize differences among the individuals rated. This latter effect gives rise to the name of the model developed here--the *multi-scale model*. Finally, the multi-scale estimators are applied to the problem noted in an earlier Rand report about estimating the cost of on-the-job training in the military. Indeed, that problem was the genesis of the multi-scale approach and illustrates the value of the multi-scale model. Although the model was developed to deal with subjective supervisory ratings, the multi-scale model may be applicable to a wide variety of other estimation problems where observations can naturally be categorized into specific subgroups.

Several specific multi-scale estimating techniques are developed, including equal total variance, equal residual variance, maximum likelihood, and least squares. These differ primarily in the way the scale parameters are estimated. Asymptotic results are derived for each of the four techniques. However, because of the difficulty in deriving small sample properties analytically, Monte Carlo experiments were

conducted. The asymptotic and Monte Carlo results, taken together, suggest some practical guidelines for estimation of the multi-scale model. Maximum likelihood and equal residual variance techniques yield consistent parameter estimates. However, for small sample sizes and configurations for cases with large random errors, the equal total variance residual estimator is preferred.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of Roberta Jean Smith, whose computational assistance was invaluable, and of Gus Haggstrom and Bridger Mitchell, who provided many thoughtful and constructive comments on earlier drafts of this report. The authors accept responsibility for any errors or omissions that may remain.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGMENTS vii

Section
I. INTRODUCTION 1

II. THE MULTI-SCALE MODEL 3

III. ESTIMATION OF THE MULTI-SCALE MODEL 9
Estimation Strategies and Techniques 10
The Existence of Solutions and a Method of Computation 15

IV. PROPERTIES AND CHARACTERISTICS OF THE ESTIMATES 18
Consistency and Asymptotic Normality 18
Asymptotic Variance of $\hat{\theta} - \theta$ 24
Summary of Estimates 26

V. MONTE CARLO RESULTS 27
Monte Carlo Methodology 27
Estimator Distributions 31
Sample Size 32
Heterogeneous Explanatory Variables 39
The Scale Parameter 40
Summary 42

VI. AN APPLICATION OF THE MULTI-SCALE MODEL 45

VII. CONCLUDING REMARKS 48

Appendix
MAXIMUM LIKELIHOOD AND LEAST SQUARES ESTIMATES 51

REFERENCES 63

I. INTRODUCTION

Subjective evaluations are an important element of military manpower policy, both in the application of present policies and in the development of new policies. For example, measures of performance in the form of subjective ratings by an individual's superior play a crucial role in determining promotions and duty assignments and illustrate the *application* of military personnel policy. Similarly, the *development* of new personnel policies frequently depends on measuring the effects of specific factors on individual performance.

Although subjective evaluations are clearly an important input to manpower policy, these measures have certain inherent difficulties. In particular, they are likely to reflect the biases of those providing the ratings. In some instances, these biases may be deliberate and applied only selectively (e.g., because of personality conflict between the rater and ratee) and cannot therefore be properly controlled for. It is probably more common, however, for these biases to be unintentional and systematically applied, a result of the fact that raters may use different implicit rating scales or may perceive matters differently. Some raters may consistently grade "easy" or "tough."

This report develops a methodology--the multi-scale model and its corresponding estimators--for estimating the systematic biases inherent in the subjective measures (of such variables as individual performance) that are often used in the development and application of manpower policy. Specifically, it is argued that subjective measures of, say, individual performance may include two types of biases. The first, the *location* bias, is the familiar problem that occurs when some raters systematically overestimate and others systematically underestimate the "true" variable. The second, the *scale* bias, occurs when some raters exaggerate the differences among those who are rated while other raters minimize these differences.

The approach adopted here incorporates these biases into the traditional classical regression model. However, the presence of the

scale bias invalidates the standard estimating techniques, so that it becomes necessary to develop special multi-scale estimators. The practical importance of the multi-scale model and estimators is two-fold. The techniques provide a way of properly estimating such parameters of the underlying model as the effects of education, military training, mental aptitude, etc. on individual performance. The model also enables the analyst to construct "corrected" measures--adjusted for the inherent biases--of the subjectively estimated variables.

In the next section, we provide a brief discussion of the origin, structure, and applications of the multi-scale model. Section III discusses some basic issues in estimating the multi-scale model and suggests and derives five specific estimating techniques. Section IV examines the mathematical and statistical properties of the estimates. Since the small-sample superiority of any of the estimates cannot be proved, we have conducted an extensive series of Monte Carlo experiments involving the principal estimation techniques. The results of these experiments are reported in Section V. Section VI applies the multi-scale model to the supervisory rating problem discussed earlier. Section VII outlines possible extensions of the model.

II. THE MULTI-SCALE MODEL

The multi-scale model is a variant of the classical linear regression model in which the dependent variable is subjected to a linear transformation that varies from group to group. Thus, for the ith observation the observed dependent variable y is related to the "true" dependent variable z_i by

$$y_{ij} = \alpha_j + \delta_j z_{ij} \ . \tag{2.1}$$

The model is *multi-scale* in that the location parameter α_j and the scale parameter δ_j may take on different values in each of the J subsets into which the observations are partitioned. The value of the unobserved dependent variable is determined by the classical model

$$z_{ij} = X_{ij}\beta + \varepsilon_{ij} \ , \tag{2.2}$$

where X_{ij} is a vector of independent variables and the ε_{ij} a set of independent random variables with mean μ and variance σ^2. The problem is to estimate the three vectors of parameters, α, β, and δ. Inasmuch as only scale effects are being investigated, all values of δ_j are assumed to be strictly positive. The full model can be written

$$y_{ij} = \alpha_j + X_{ij}(\delta_j\beta) + (\delta_j\varepsilon_{ij}) \ . \tag{2.3}$$

Classical regression analysis has been extended to a number of cases in which the coefficients may vary in some fashion across subsets of observations. It is standard practice, for instance, to use separate intercept terms or separate coefficients for subsets of observations sharing some common attribute. Indeed, whole sets of procedures, known generally as analysis of covariance, have been devised for determining whether sets or subsets of coefficients differ among subsets of observations. (See Chou [1] and Johnston [2].) In a related area the pooling

of cross-section and time series data poses a number of estimation problems that have been exhaustively analyzed in the literature on error components models. (See Wallace and Hussain [3], Ballestra and Nerlove [4], Nerlove [5], [6], and Hsiao [7].) This report represents an extension of the literature on different coefficients and intercept terms to a case where the coefficient vector may differ by a scale factor among subsets of observations. The parameter vector $\gamma_j = \delta_j\beta$ in Eq. (2.3) may take on different values for each subset of observations j; however, unlike the case of pooling data, the parameter vectors differ by a scale factor rather than being identical or totally different. Because the δ_j is a multiplier of β as well as the ε_{ij}, the problem is more than a problem of heteroscedasticity. Hartley and Jayatillake [8] have, in fact, analyzed the case where the variance of the error term may differ by subset.

The problem is also more than a nonlinear regression problem, since in its conventional interpretation the nonlinear regression problem can be written as

$$y - g(x,\theta) = \varepsilon \ , \tag{2.4}$$

whereas (2.3) can only be written as

$$f(y,x,\theta) = \varepsilon \ . \tag{2.5}$$

The problem created by (2.3) is, strictly speaking, a multi-scale problem. We believe that the multi-scale model has considerable applicability in economics and the social sciences in analyzing data containing rating-scale phenomena, in analyzing pooled cross-section and time series data, and in analyzing time series data involving subannual observations.

Because to the best of our knowledge this model has not been analyzed previously, this report derives and discusses a variety of estimating techniques and suggests guidelines for using the various estimates. Although some guidance is obtained from asymptotic properties of the estimates, our recommendations are principally based on the results of a series of Monte Carlo experiments. We do not explore in any detail

the properties of the estimates or even the existence of estimates. Rather it is our desire to focus on the multi-scale model itself, on practical problems of one choice and computation of estimates, and on the application of the model to problems of statistics and econometrics. We have attempted to minimize the mathematical derivation and streamline whatever proofs are given. This last is at the expense of mathematical rigor but in keeping with the general nature of the report.

The empirical problem that led us to estimate the multi-scale model arose in a study of on-the-job training in the Air Force.[1] Productivity indices were created for individual airmen on the basis of quantitative but somewhat subjective information provided by the airman's supervisor.[2] Multiple observations were available from individual supervisors. The parameters α_j and δ_j in (2.1) reflect the fact that each supervisor apparently used a different rating scale. Moreover, these differences were reflected in the mean scores (α_j) and in the standard deviations (δ_j) of the subsamples. Figure 1 plots the cost of on-the-job training (OJT) for individual airmen grouped under the 12 supervisors in the sample. The standard deviations of OJT costs range from \$214 to \$4297 across the 12 supervisors. Since supervisors typically oversee small numbers of individuals, statistical analysis of the data is impossible unless data from different supervisors are pooled together. Consequently, it was necessary to combine a rating-scale model (2.1) with a behavioral model (2.2). We would expect that statistical inference involving any variables containing rating-scale phenomena would give rise to the multi-scale model. Similar applications could relate to personnel evaluations, classroom performance, or other situations where personal ratings might be used.

Economists have only recently and then infrequently come to use data based on ratings. As a result, the most useful application of the multi-scale model for economists may be in the area of pooled cross-section and time series data. Pooled data usually involve combining

---

[1] See Gay [9] and Gay and Nelson [10].

[2] The cost of OJT for each airman was estimated as the difference between the airman's productivity, which was provided by his supervisor, and his wages.

time series data on firms, states, or countries. If the dependent variable is in the form of an aggregate quantity, such as fuel consumption or liquid asset balances, there will be huge scale differences among the different units combined in the analysis. The theoretical model may express the quantity demanded in the ith market at time t as a function of prices and income distribution parameters in the market:

$$q_{it} = \beta_o + \sum_{k=1}^{K} \beta_k p_{itk} + \beta_{K+1}\mu_{it} + \beta_{K+2}\sigma^2_{it} + \varepsilon_{it} , \qquad (2.6)$$

where q_{it} = quantity demanded in the ith market in period t,
p_{itk} = price of commodity k in the ith market in period t,
μ_{it} = mean income of potential buyers in the ith market in period t, and
σ^2_{it} = variance of income of potential buyers in the ith market in period t.

If the analyst is a good theorist, he may be able to specify *a priori* the relevant variables, the functional form, the distribution of ε_{it}, and whatever dynamic properties the demand may exhibit. The scaling problem is often handled by using another variable either to scale q_{it} by defining a new variable q_{it}/x_{it} or to use it in a weighted regression.[1] Only rarely, if ever, is there theoretical justification for the choice of such a variable.

The application of the multi-scale model could eliminate the need to specify an artificial scaling variable x_{it} by estimating the appropriate scale for each firm or state as well as the parameter vector β. Alternatively, one might use dependent variables of the form $\Delta q/q$ or ln q to take care of the scaling. But unless such a functional relationship is suggested by theoretical considerations, these measures may be just as artificial as the choice of x_{it}. Consequently, the multi-scale model may be a good substitute for several conventional practices

[1] To be sure, economists have paid considerable attention to the properties of ε_{it}, particularly $E\varepsilon\varepsilon'$, in estimating behavioral equations from pooled data.

in pooled data where scale problems exist. Moreover, even where scale problems are only suspected, the multi-scale model can be used to test for the presence of scale effects. This would provide another tool in the kit to determine to what extent sets of coefficients are the same or different. The multi-scale model permits coefficients to differ but remain proportional across sets of observations.

A third application of the multi-scale model would be in demand or supply models estimated from quarterly or monthly data where there is a strong annual cycle. If it is expected that the parameters β and the variance of the error are also subject to the cycle, the multi-scale model can be used. This could be written

$$q_{jt} = \alpha_j + \sum \delta_i M_i (p_{jt}\beta + \varepsilon_{jt}) \ , \qquad (2.7)$$

where j is the month, t is the year, and M_i is a dummy variable for month i. Examples of data subject to strong seasonal fluctuations would include grain sales, military enlistments, heating fuel consumption, and number of new entrants to the labor market. With strong seasonal fluctuations, better estimates of β can often be made by using annual data instead of monthly or quarterly data. The absence of a long time series of data or the presence of structural changes in the market often make the use of subannual data necessary. Moreover, the seasonal pattern itself is often of interest. The estimates $\hat{\alpha}$ and $\hat{\delta}$ from (2.7) can be used, in fact, to construct seasonally adjusted variables.

There are probably other applications for the multi-scale model, such as in estimating age-earnings profiles where the effects of education and ability differ by age; however, applications for time series and pooled cross-section and time series data would seem to be the most likely uses for economists.

III. ESTIMATION OF THE MULTI-SCALE MODEL

The assumptions of the multi-scale model basically reflect the assumptions of the classical normal linear model. In particular, we assume

$$y_{ij} = \alpha_j + \delta_j(X_{ij}\beta + \varepsilon_{ij}) , \qquad \begin{matrix} i = 1, \ldots, T_j , \\ j = 1, \ldots, J \end{matrix} \qquad (3.1)$$

where the ε_{ij} are independent random variables each distributed $N(0,\sigma^2)$. $X_{ij} = (X_{ij1}, \ldots, X_{ijK})$ is a vector of known constants of dimension K. The T cases or observations ($T = \sum T_j$) are partitioned into J subsets, as indicated. The vectors $\alpha = (\alpha_1, \ldots, \alpha_J)$, $\delta = (\delta_1, \ldots, \delta_J)$, and $\beta = (\beta_1, \ldots, \beta_K)$ are fixed unknown parameters.[1] The only restriction is that $\delta_j > 0$ for all j.

Equation (3.1) does not constitute a complete model. In particular, estimates of the form $\delta^* = k\hat{\delta}$, $\beta^* = \hat{\beta}/k$ and $\sigma^{*2} = \hat{\sigma}^2/k^2$ are observationally equivalent as k varies. To identify these parameters, the multi-scale model requires an additional condition on the set of parameter vectors. It seems most natural to place some restriction on the vector δ, and we deal with the strictly separable function

$$G(\delta) = G_1(\delta_1) + \ldots + G_J(\delta_J) = 0 \qquad (3.2)$$

as a basis for identifying δ, β, σ^2. Assuming that the geometric mean of the δ_j's overall observation is 1 leads to

$$\sum T_j \ln \delta_j = 0 . \qquad (3.3)$$

---

[1] Hsiao [7] has analyzed the error components problem as a random coefficients model. The treatment of α and δ as random variables may prove to be a fruitful approach; however, in this report α and δ are fixed parameters.

In the case of the supervisory ratings application, e.g., (3.3) requires that the supervisors provide on average unbiased ratings, in the geometric sense of the term "on average."

The choice of constraint is a matter of some importance. For instance, either changing the weights T_j or introducing the "bias" by setting $\sum T_j \ln \delta_j = a$, $a \neq 0$ will change both the absolute and relative values of the parameters β, δ, σ^2. Hence the identifying restriction is in every sense an integral part of the model.

A similar identification problem may arise between the location parameter and the intercept term β_0, if one exists. We assume, however, that the equation $z_{ij} = X_{ij}\beta$ goes through the origin.[1] There is no possible confounding of intercept terms in this case. The identification of intercept terms is usually less important to the analyst than the identification of coefficients.

ESTIMATION STRATEGIES AND TECHNIQUES

The analyst may choose any of a variety of strategies in estimating the behavioral parameters β of the multi-scale model. He may choose to estimate α, β, and δ together by applying maximum likelihood (ML) estimation, least squares (LS) estimation, or some other technique producing simultaneous estimates of all parameters. Such a strategy invariably requires iterative methods of estimation and possibly requires

---

[1] Note that it is necessary to estimate *both* the α_js and β_0 if one desires a "corrected" measure of the dependent variable, since

$$z_i = (y_{ij} - \alpha_j)/\delta_j \ ,$$

so that

$$\hat{z}_i = (y_{ij} - \hat{\alpha}_j)/\hat{\delta}_j \ .$$

A plausible assumption would be that α_j equals zero on average (where, in this instance, we mean arithmetically "on average"), so that the side condition for α_j becomes

$$\sum T_i \alpha_j = 0 \ .$$

In the remainder of this report, we will assume, without loss of generality, that β_0 equals zero.

the construction of special software packages. A simpler strategy would be to try to adjust y_{ij} for the effects of α_j and δ_j *prior to* estimating β. One method of doing this is to standardize the y_{ij} in different subsets for means and variances and then to regress the adjusted variable $\hat{z}$ on X by means of ordinary least squares (OLS). The apparent advantage is in the costs of estimation, and the apparent sacrifice is in not using information on X and β in developing estimates of α and δ. A third strategy, of course, would be to ignore δ entirely and estimate the parameters α and β using OLS.[1] We have already assumed that the scale parameters are distributed around 1.0. This last strategy is the one implicitly adopted in pooled regressions where the multi-scale model is not used. In summary, the three estimation strategies are

I. ML, LS estimates of α, β, δ
II. OLS estimates of δ with "adjusted" y
III. OLS estimates of α, β

Although strategies II and III may have the meager appearance of straw men, there is no guarantee that adopting the more elaborate approach of strategy I uniformly produces the best results. Figure 2 shows the strategy and technique producing minimum mean-square error in estimates of β in one series of Monte Carlo experiments conducted for this study. Individual experiments differ according to R^2 (the coefficient of determination) and the variance of ln δ in the multi-scale model. Perhaps surprisingly each strategy offers a region of superiority. Where all δ values are near unity, strategy III is superior in that it is better to ignore δ than try to estimate it. Where the R^2 is small (less than .30) and where the X_{ij} have similar distributions for different values of j, strategy II is superior. There apparently is little error due to standardizing values of y in this homogeneous case. However, where the X_{ij} have quite dissimilar distributions

[1] Actually, the third strategy consists of OLS with dummy variables for α (since a fourth strategy could be to ignore both the α and β).

across subsets (not shown), this region of superiority is sharply reduced. Since strategy II ignores the effects of X and β in estimating δ, its value depends crucially on the homogeneity of the X_{ij} across subsets. In the remaining areas of Fig. 2, strategy I provides better estimates of β.

We derive and evaluate five estimating techniques for the multi-scale model: (1) OLS; (2) the technique for adjusting the dependent variable before using OLS, referred to here as the equal total variance (ETV) technique; (3) ML estimation; (4) LS estimation; and (5) another simultaneous technique that determines δ such that there is equal residual variance (ERV) across subsets.[1] The first four have been mentioned previously; the last technique is heuristically determined based on the expectation that, if δ is controlled for, the variance of the residuals should be approximately equal across subsets of observations.

Each of the five techniques provides four sets of conditions, which can be associated with the parameters α, β, δ, and σ^2. If treatment of degrees of freedom is standardized,[2] the conditions for α, β, and σ^2 are identical for each of the five estimating techniques. The details for ML and LS estimates are provided in the appendix. Thus, as in the case with OLS, the intercept terms α can be estimated after the other parameters, because the following equation defines the estimator of δ_j for all methods.

$$\hat{\alpha}_j = \bar{y}_j - \hat{\delta}_j \bar{X}_j \hat{\beta} \ , \tag{3.4}$$

where $\bar{X}$ is the mean for subgroup j. Hereafter, in fact, we eliminate α from the model by redefining y_{ij} and X_{ij} as $y_{ij} - \bar{y}_j$ and $X_{ij} - \bar{X}_j$. Thus, the multi-scale model is

$$y_{ij} = \delta_j (X_{ij}\beta + \varepsilon_{ij}) \ . \tag{3.5}$$

---

[1] Thus, techniques (3), (4), and (5) belong to broad strategy I outlined earlier.

[2] Maximum likelihood estimates do not provide for adjustments for degrees of freedom. Here, as in other applications, we use ML estimates for degrees of freedom.

The condition for $\hat{\beta}$ is

$$\hat{\beta} = (X'X)^{-1} X'\hat{z} \ , \tag{3.6}$$

where $\hat{z}_{ij} = y_{ij}/\hat{\delta}_j$. Thus in all cases the condition for β is simply the OLS regression of $\hat{z}$ on X.

Finally, in each case we choose

$$\hat{\sigma}^2 = \frac{1}{T - 2J - K} \sum_j \sum_i \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right)^2 \ , \tag{3.7}$$

where 2J + K represent the total number of parameters in α, β, and δ.

The only differences among the five estimating techniques are the conditions associated with estimates of δ. These are listed below, beginning with ML, LS, and ERV, the three simultaneous techniques. The parameter $\hat{\lambda}$ in ML and LS is a Lagrange multiplier attached to the side condition. The results are for the general form of the side condition (3.2). In particular, for each j = 1, ..., J,

$$\text{ML:} \quad \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} + \hat{\lambda} \frac{\partial G}{\partial \delta_j} \frac{\hat{\delta}_j \hat{\sigma}^2}{T_j - 1} = \hat{\sigma}^2 \tag{3.8.1}$$

$$\text{LS:} \quad \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} - \frac{1}{2} \hat{\lambda} \frac{\partial G}{\partial \delta_j} \frac{\hat{\delta}_j}{T_j - 1} = 0 \tag{3.8.2}$$

$$\text{ERV:} \quad \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right)^2 = \hat{\sigma}^2 \tag{3.8.3}$$

$$\text{ETV:} \quad \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} \right)^2 = V, \text{ a constant} \tag{3.8.4}$$

$$\text{OLS:} \quad \hat{\delta} = 1 \ . \tag{3.8.5}$$

if the data matrix obeys some quite reasonable conditions.[1] More generally there are 2^J solutions to the system of ML equations. Only one solution will have all values of the $\delta_j > 0$. The 2^J solutions may be thought of as J pairs of values for $\hat{\delta}_j$, $A_j + |B_j|$ and $A_j - |B_j|$, where $|B_j| > A_j$. There are 2^J combinations but only one where $\hat{\delta}_j > 0$.

The necessary and sufficient conditions for this result can be simply stated. Let Y represent the $T \times J$ matrix, which places values of Y_{ij} in separate columns according to subgroup. Define $Q = Y'(I - X(X'X)^{-1}X')Y$, where $M = I - X(X'X)^{-1}X'$ is an idempotent matrix. Then the necessary and sufficient condition for a unique positive solution for the ML equation is that

$$\det |Q| \neq 0$$

or that Q be of full rank (rank = J). This condition will not hold if (1) the columns (variables) are not linearly independent; (2) there is a perfect fit between Y_{ij} and X_{ij} for all observations in any subgroup; (3) there is no variation in Y_{ij} in any subgroup; and (4) there is only one observation in any subgroup. Assuming the existence of a solution in the ML case requires (1) a proper specification of the variables and (2) the elimination of any subgroup satisfying any of the conditions (2) - (4).

To compute the estimates of the multi-scale model, we have developed an iterative approach that converges rather quickly to a set of parameters satisfying (3.4), (3.5), (3.6), (3.7) and one of the conditions (3.8.1)--(3.8.3). This is an approximate solution to the system of equations, but as indicated, only one of several possible solutions where the system is quadratic. Negative roots of the quadratic have been eliminated since these produce negative estimates of δ_j. This computation procedure has been applied in literally thousands of regressions in the Monte Carlo experiment and in no case did it yield unreasonable or outlandish results.

---

[1] This result was provided by Gus C. Haggstrom.

The estimating procedure can be described in six steps beginning with raw data, not grouped by subset. Documentation of this program has been provided by Smith [11]. The steps are as follows:

1. Raw data are ordered by subgroup. Sample means are calculated by subset for the dependent and independent variates, and these variates are then expressed in terms of deviations from the subgroup means. (This permits the vector of intercept terms to be estimated after all other parameters have been estimated.)
2. Initial trial values of the $\hat{\delta}_j$ are obtained from the standard deviations of y_{ij} in each subgroup. These are normalized to conform to the logarithmic constraint (3.3). An "adjusted" dependent variable is found by dividing y_{ij} by the estimate of δ_j. (This is simply the ETV procedure. If ETV estimates are desired, it is necessary only to calculate the OLS based on the adjusted dependent variable.)
3. Initial estimates of $\hat{\beta}$ are obtained by regressing the adjusted dependent variable on the independent variates.
4. Given the estimates of β, new estimates of δ are obtained (and the Lagrange multiplier λ where applicable). A gradient search technique (Newton's method) is used to find the appropriate Lagrange multiplier. Acceptable accuracy can usually be found within about five iterations. Given the proper λ, the values of δ can be calculated directly from the J equations involving δ.
5. Steps (3) and (4) are repeated until the values of $\hat{\beta}$ and $\hat{\delta}$ converge. A criterion is used that the maximum change in any δ_j, which is the most sensitive parameter, must be less than .001. Usually, fewer than four iterations are required.
6. Given values of $\hat{\beta}$ and $\hat{\delta}$, estimates of α can be calculated.

Data processing has been performed on the IBM 370/158. The average cpu time per estimate has been 4 seconds with 250 total observations, 50 subgroups, and two independent variables for the ML technique. Average cpu for the ETV technique (basically a single regression) is 1.1 seconds.

IV. PROPERTIES AND CHARACTERISTICS OF THE ESTIMATES

Ideally we would like to be able to derive the small-sample distributions of the five estimating techniques and base our choice of estimates principally on these theoretically determined distributions. Only large-sample properties and distributions of the estimates can be examined; consequently, the ultimate basis for our choice of estimates will be the Monte Carlo experiments of Section V. An analysis of the statistical properties and mathematical characteristics of the estimates contributes to an understanding of the estimation problem and provides more guidance in the choice of estimates.

This section is devoted to two topics: (1) the consistency and asymptotic normality of the estimates and (2) asymptotic variance of the estimates.

CONSISTENCY AND ASYMPTOTIC NORMALITY

The characteristics of the multi-scale model do not lend themselves to a mathematical analysis of small-sample properties. Neither unbiasedness nor minimum variance, for instance, can be demonstrated for finite sample sizes. We must restrict ourselves to the asymptotic properties. In the multi-scale problem, where observations are grouped into subsets, the question of consistency is complicated considerably by the fact that sample size can be increased by increasing the number of observations per subset (T_j) or by increasing the number of subsets J, or both. When sample size is increased by increasing the number of subjects J, however, the number of parameters to be estimated also increases--two new parameters for each new subset.

In general, we note that if ξ is a parameter vector of m elements, then $\hat{\xi}$ is said to be a consistent estimator of ξ if

$$\operatorname*{plim}_{n/m \to \infty} \hat{\xi} = \xi$$

$$\text{LS:}\quad \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} = \frac{1}{2}\lambda^* \ . \tag{4.2.2}$$

Here λ and λ^* are Lagrange multipliers corresponding to condition (3.3).

Since $\hat{\sigma}^2$ is the same for all j, (4.2.1) and (4.2.2) yield identical solutions. The appendix demonstrates that for (3.8.1) and (3.8.2) together with (3.4), (3.6), and (3.7) to yield identical estimates of α, β, δ, and σ^2, then

$$G(\delta) = c_o \sum T_j \lim \delta_j + C_1 = 0 \ , \tag{4.3}$$

where c_o and C_1 are arbitrary constants. A corollary to this is that where (4.3) does not hold, ML and LS estimates will be different for at least some values of y and X. Moreover, this difference is generally independent of sample size, so that LS estimates are different from ML estimates at all sample sizes even in the limiting case and, therefore, are inconsistent.

ETV Estimates

Under ETV estimates of $\hat{\beta}$ are the OLS estimates from a regression of y (adjusted for variance) on X. In particular,

$$\hat{\beta} = \sum_{j=1}^{J} \frac{1}{\hat{\delta}_j} (X'X)^{-1} X_j' y_j \ , \tag{4.4.1}$$

where X_j is the $T_j \times K$ matrix of independent variables for subgroup j.

Under any normalization rule, the ratio $\hat{\delta}_j/\hat{\delta}_h$ from (3.6.2) is

$$\frac{\hat{\delta}_j}{\hat{\delta}_h} = \left[\frac{\sum_j y_{ij}^2/(T_j - 1)}{\sum_h y_{ih}^2/(T_h - 1)} \right]^{\frac{1}{2}} . \tag{4.5}$$

Taking probability limits as $T_j \to \infty$ and $T_h \to \infty$ yields

$$\text{plim } \frac{\hat{\delta}_j}{\hat{\delta}_h} = \frac{\delta_j}{\delta_h}\left(\frac{\sigma^2 + \omega_j}{\sigma^2 + \omega_h}\right)^{\frac{1}{2}}, \tag{4.6}$$

where ω_j is the lim $1/(T - 1)(\beta' X_j' X_j \beta)$ (the "explained variance") and is assumed to exist and be non-zero. The consistency of the ratio $\hat{\delta}_j/\hat{\delta}_h$ depends on having the same explained variance in each subset. Thus, if the rows of X_j can be viewed as coming from a distribution that is more homogeneous than the distribution of X_h, then ω_j will be smaller than ω_h and plim $(\hat{\delta}_j/\hat{\delta}_h)$ will be too small. This result should be quite intuitive. Under ETV the value of $\hat{\delta}_j$ is determined without any information on X_j. The technique attributes to δ_j any variation regardless of source. If $X_1'X_1 = X_2'X_2 = \ldots = X_J'X_J$ this would seem to be a perfectly appropriate technique for any sample size.

The exact degree of inconsistency in $\hat{\beta}$ (defined as plim $\hat{\beta}/\hat{\beta}$) can be determined where there are only two subsets of observations. We assume further that

$$\lim \frac{1}{T_1} X'X_1 = \rho \lim \frac{1}{T} X_2'X_2 , \tag{4.7}$$

where ρ is the ratio of the variance of the two subsets. The probability limit $\hat{\beta}$ as $T_1 = T_2 \to \infty$ is

$$\text{plim } \hat{\beta} = \beta \left\{ \frac{1}{1 + \rho}\left[\frac{(1 + \rho) - r(1 - \rho)}{(1 + \rho) + r(1 - \rho)}\right]^{\frac{1}{4}} + \frac{\rho}{1 + \rho}\left[\frac{(1 + \rho) + r(1 - \rho)}{(1 + \rho) - r(1 - \rho)}\right]^{\frac{1}{4}} \right\} . \tag{4.8}$$

The parameter r is the limiting value of the R^2 of the model.[1] The limiting values of plim $\hat{\beta}/\beta$ as ρ and r approach extreme values amplifies on the results of (4.5). In particular

$$\lim_{\rho \to 1} \text{plim } \hat{\beta}/\beta = 1 \tag{4.9.1}$$

$$\lim_{r \to 0} \text{plim } \hat{\beta}/\beta = 1 \tag{4.9.2}$$

$$\lim_{\rho \to \infty} \text{plim } \hat{\beta}/\beta = \left(\frac{1 - r}{1 + r}\right)^{\frac{1}{4}} \tag{4.9.3}$$

$$\lim_{r \to 1} \text{plim } \hat{\beta}/\beta = \frac{\rho^{\frac{1}{4}} + \rho^{\frac{3}{4}}}{1 + \rho} . \tag{4.9.4}$$

As the degree of heterogeneity or as R^2 diminishes, the degree of inconsistency in $\hat{\beta}$ also diminishes. The fact that ETV works best (in the limit) for models with a small R^2 is because σ^2 tends to swamp the values $\beta' X_1' X_1 \beta$ and $\beta' X_2' X_2 \beta$ (see (4.5)). This therefore reduces the error from ignoring the explanatory variables in estimating δ.

ERV Estimates

The most important point to make about ERV estimates is that they approach ML estimates as all $T_j \to \infty$. Equality of ERV and ML estimates at all values of y and X require that there be no difference between (3.8.1) and (3.8.3). This requires

---

[1] In particular,

$$r = \frac{\frac{1}{2}(\omega_1^2 + \omega_2^2)}{\sigma^2 + \frac{1}{2}(\omega_1^2 + \omega_2^2)} .$$

$$\frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} - \frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right)^2$$

$$+ \lambda \frac{\partial G}{\partial \delta_j} \frac{\hat{\delta}_j}{T_j - 1} \sigma^2 = 0 , \qquad j = 1, \ldots, J . \tag{4.10}$$

This yields

$$\frac{1}{T_j - 1} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) X_{ij}\hat{\beta} + \lambda \frac{\partial G}{\partial \delta_j} \frac{\hat{\delta}_j}{T_j - 1} \sigma^2 = 0 . \tag{4.11}$$

As $T_j \to \infty$ for all j, the second term will tend to zero. The first term is zero if $\hat{\beta}$ is the same as the least squares estimate of β based only on observations in the jth subset.[1] $\hat{\beta}$, in fact, is the least squares estimate based on all observations. However, as $T_j \to \infty$ for all j, the coefficient $\hat{\beta}_j$ estimated from the jth subset approaches $\hat{\beta}$. Thus in the limit (4.10) is satisfied, and ERV and ML estimates are equivalent.

A practical problem in applying ERV is that in models with small sample sizes and a high R^2, the probability of obtaining solutions with imaginary components is substantial. In cases where this occurred, we adopted the procedure (certainly unsound) of setting the imaginary component equal to zero. This poses a serious practical drawback to use of ERV.

ASYMPTOTIC VARIANCE OF $\hat{\theta} - \theta$

The literature on asymptotic properties of ML estimates suggests that the asymptotic variance of $\hat{\theta} - \theta$ is

$$\lim_{T_j \to \infty} \text{Var } (\hat{\theta} - \theta) = \frac{1}{T} \Gamma^{-1}(\theta) , \qquad \text{for all } j , \tag{4.12.1}$$

---

[1] That is, $\sum_j \left(\frac{Y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) X_{ij}\hat{\beta} = 0$ if $\hat{\beta} = (X_j'X_j)\, X_j' \frac{y_j}{\hat{\delta}_j}$.

where

$$\Gamma(\theta) = E\left(-\frac{\partial^2 \log L}{\partial\theta^2}\right). \tag{4.12.2}$$

This suggests an estimator for the variance of the estimates of the multi-scale model. Under this estimate it can be shown that variance-covariance of the behavioral parameters β will not be equal to

$$\frac{\hat{\sigma}^2}{T}(X'X)^{-1},$$

unless

$$\frac{\partial^2 \log L}{\partial\delta\partial\beta} = 0. \tag{4.13}$$

The appendix shows that (4.13) holds if and only if $X_j'X_j$ is equal for every subset j. Thus, the "t-statistics" for $\hat{\beta}$ from the classical normal model will hold asymptotically for the multi-scale model if and only if the moment matrix $X_j'X_j$ is the same for every subset.

If the independent variables do not have the same dispersion matrix, a correction to $\hat{\sigma}^2/T\ (X'X)^{-1}$ is required. This will be given by calculation of $\Gamma^{-1}(\theta)$. In the specific case of one variable and two subsets we have calculated the specific asymptotic variance of β:

$$\text{Var}\ (\hat{\beta} - \beta) = \frac{\hat{\sigma}^2}{T\sum x^2}\left(\frac{2 + \beta^2\sigma_x^2}{2 + \beta^2\sigma_{x_1}^2\sigma_{x_2}^2/\sigma_x^2}\right),$$

where $\sigma_{x_1}^2$, $\sigma_{x_2}^2$, and σ_x^2 are the "variances" of x from the two subsets and the total sample. The value of the term in parentheses attains a value of 1.0 where $\sigma_x^2 = \sigma_{x_1}^2 = \sigma_{x_2}^2$, but otherwise is greater than unity. Hence the asymptotic variance of $\hat{\beta}$ is at least as great as the variance of $\hat{\beta}$ in the classical normal model.

SUMMARY OF ESTIMATES

We have presented some of the mathematical characteristics and statistical properties of ML, LS, ERV and ETV in this section and in the appendix. Table 1 briefly recapitulates the most important features of all five methods of estimating the multi-scale model. The results of this section suggest, if anything, that ML estimates display the fewest bad features if one ignores computational costs. Nevertheless, there are many cases where ML estimates have impressive asymptotic properties but are not the best estimates in small-sample situations. In the absence of any specific guidance on small-sample properties, our method has been to rely on Monte Carlo experiments to determine the superior estimating techniques. This is the subject of the following section.

Table 1

FEATURES OF MULTI-SCALE ESTIMATES

Feature	ML[a]	LS	ERV	ETV	OLS
Type of solution	Iterative	Iterative	Iterative	Noniterative	Noniterative
Invariance of $\hat{\beta}$ to (3.2)	Not invariant	Not invariant	Invariant	Invariant	Invariant
Values of $\hat{\delta}$[b]	Real Positive	Real May be negative[c]	May be complex Positive if real	Real Positive	N.A.
Consistency	Consistent	Inconsistent[c]	Consistent	Inconsistent[d]	Inconsistent

[a] Adjusted for degrees of freedom.

[b] Actually, solutions to normal equations for $\hat{\delta}$, provided solutions exist.

[c] Consistent and positive if (3.3) holds.

[d] Consistent if plim $1/T_j$ $(\beta' X_j' X_j \beta)$ is equal across subsets.

V. MONTE CARLO RESULTS

Specification of an estimator's distribution is an important aspect of the development of any econometric estimating procedure. Such information assumes a special importance in the present study insofar as several alternative estimators have been developed for estimating the parameters of the multi-scale model. Therefore, knowledge of the statistical properties of these alternatives is important not only for establishing the statistical reliability of any given estimate, but also for selecting the *appropriate* multi-scale estimator under different sample conditions. Indeed, it was stated at the outset that no single estimator is dominant over the entire range of possibilities. Instead, the appropriateness of any of the multi-scale estimators depends, among other things, upon the sample size, the signal-to-noise ratio, and the degree to which the model is multi-scale.

Other than for consistency, the statistical properties of the alternative estimators cannot be derived analytically. We must therefore resort to numerical approximations through the setup of Monte Carlo experiments to obtain the distributional properties of the multi-scale estimators. A description of the experimental approach and the results from these experiments is given below.[1]

MONTE CARLO METHODOLOGY

Since we were able to establish only the consistency of the multi-scale estimators on an analytical basis, we have had to resort to Monte Carlo experimentation to determine other statistical properties of the estimators. The strategy used in these experiments is, for the most part, dictated by the results derived previously. For example, sample size may be increased either by increasing the number of subgroups or by increasing the number of observations per subgroup, and the effects of these two alternatives may be considerably different. Therefore,

---

[1] For a more complete description of the Monte Carlo methodology and results, see Cooper [17].

sample composition, as well as sample size, must be an explicit part of the experimental design. Before dealing with these specifics, however, we first outline the general approach used in the experiments.

Given the basic multi-scale model,

$$y_{ij} = \alpha_j + \delta_j(X_i\beta + \varepsilon_i) ,$$

and specified values for the α, δ, and β vectors and for the X matrix, we conducted Monte Carlo experiments by simulating values for the ε vector. The elements of the ε vector were drawn from a normal population and the number of cases run for each experiment was determined according to the number required to yield a stable representation of the parameter distributions.[1]

A number of experiments were conducted for different formulations of the multi-scale model, as indicated below.

Model Specification

Two specifications of the model were tested, one with two explanatory variables and one with five explanatory variables. Only the two variable version is reported here.[2] The parameter values were: $\beta_1 = 1.0$ and $\beta_2 = 2.0$. The two explanatory variables were uncorrelated.

Sample Size

As noted previously, sample size in the multi-scale model has two

---

[1]That is, a concern in conducting Monte Carlo experiments is how many cases must be run before the estimated distributions of the parameter estimates "reasonably reflect" the true distributions. Although a precise reflection would require an infinite number of cases, such an approach is, of course, not feasible. Instead, the procedure was to run 200 cases for one of the experiments, with a summary printed every 20 cases. These summary statistics were then examined to determine where the estimated distributions began to stabilize--that is, where the addition of another 20 cases did not appreciably change the estimates of the distributions. For medium to large samples, the number of cases required was 20; for very small samples, the number of cases required was 100. The number of cases is reported with the results. For a more complete description of the approach, see Cooper [17].

[2]The five variable results are reported in Cooper [17]. They yield essentially the same results as the two explanatory variable specification.

dimensions: the number of observations per subgroup and the number of subgroups. Accordingly, experiments were conducted for a number of different sample sizes and configurations. Complete results are reported for one "larger" sample experiment: five subgroups with 50 observations per subgroup--the so-called 5 × 50 sample. Detailed results are also reported for one "small" sample situation--five subgroups with five observations per subgroup (denoted the 5 × 5 sample)--and for one "medium" sample configuration--50 subgroups with five observations per subgroup (denoted the 50 × 5 sample). Summary results are also reported for the following sample sizes (where the first number shows the number of subgroups and the second shows the number of observations per subgroup): 10 × 5, 25 × 5, 100 × 5, and 25 × 10. Finally, experiments were conducted for one sample configuration where the number of observations per subgroup varied: 50 subgroups with an average of five observations per subgroup (as few as three and as many as 10), denoted the 50 × 5 (var) sample.

Explanatory Variables

The explanatory variables were chosen such that the correlation between the two was zero.[1] Two sets of experiments were conducted with regard to the explanatory variables. In the first, the explanatory variables were drawn from homogeneous populations--that is, the explanatory variables for each subgroup came from the same population. In the second, the explanatory variables were drawn from heterogeneous populations--that is, the populations from which the explanatory variables were drawn differed by subgroup. In half of the subgroups, the standard deviation of the underlying population for each of the explanatory variables was twice that for the other half of the subgroups.[2] These two sets of experiments are referred to as the homogeneous and heterogeneous cases, respectively.

---

[1]Although the explanatory variables were drawn from two populations with zero correlation, the sample correlation for the actual variables used was 0.08.

[2]To clarify the procedure, the two explanatory variables for the homogeneous case were each drawn from a normal population with mean zero and standard deviation of 10. Once the particular set of homo-

Signal-to-Noise Ratio

The signal-to-noise ratio is based on the "true" model, as given previously in Eq. (2.2), rather than on the observed model given in Eq. (2.3). Three different signal-to-noise ratios were tested, yielding true R^2s of 0.1, 0.5, and 0.9.

α and δ

One α vector was used for all the experiments; the elements of the α vector were generated from a uniform distribution.[1] Four different δ vectors were used in the experiments. Each was generated from a log-normal distribution and normalized such that the geometric mean equaled one. The only difference in these δ vectors is the set of parameters describing its corresponding normal distribution. In each vector, the mean of its corresponding normal was zero; for δ_1, the standard deviation was 0.1; for δ_2, it was 0.25; for δ_3, it was 0.5; and for δ_4, it was 1.0. This yielded four δ vectors, where the geometric mean of each was one, but where the variances were 0.04, 0.33, 3.57, and 411.1.[2]

It will be shown later that the multi-scale estimators are unaffected by the δ vector (so long as the geometric mean equals one).[3]

---

geneous explanatory variables was drawn, it was used for the remainder of the experiments (that is, the Monte Carlo experiments were not conducted for "random" explanatory variables). In the heterogeneous case, for half of the subgroups, x_1 and x_2 were the same as for the homogeneous case (i.e., drawn from normal populations with mean zero and a standard deviation of 10). For the other half of the subgroups, x_1 and x_2 were each drawn from normal populations with mean zero and standard deviation of 20. Again, once the basic set of heterogeneous explanatory variables was drawn, it was used for the remainder of the heterogeneous experiments.

[1]Since each of the estimating techniques estimates the δ vector by merely subtracting out the subgroup means, the distribution of the elements of δ does not affect the estimation. A uniform distribution was chosen for convenience only.

[2]Again, note that one δ vector was used for any given set of experiments. The term "variance" is not meant to imply randomness nor that the δ elements were redrawn.

[3]For convenience, the particular δ vector used for the multi-scale estimators was the one where the variance equals 3.57.

However, the variability of the elements of δ determines when it is worthwhile to use one of the multi-scale estimators rather than just relying on ordinary least squares.

Estimators

As shown previously, least squares yields identical results to maximum likelihood when the δ vector is normalized such that the geometric mean of the elements equals one; otherwise it yields inconsistent estimates. Therefore, Monte Carlo results are presented only for three multi-scale estimators: maximum likelihood (MLE), equal residual variance (ERV), and equal total variance (ETV). As a basis for comparison, ordinary least squares with dummy variables for the subgroup intercepts (OLS-DV) was also used.

The foregoing constitutes the basis of the Monte Carlo experiments. These experiments are addressed to two principal questions: (1) Which is the preferred multi-scale estimator under alternative sample and model configuration? and (2) When is the multi-scale approach to be preferred to least squares with dummy variables for the intercepts?

ESTIMATOR DISTRIBUTIONS

As noted previously, the small sample distributions for the multi-scale estimators cannot be derived analytically. Since the Monte Carlo approach just outlined suggests that we examine the multi-scale estimators under a variety of conditions, it is desirable to simplify these comparisons as much as possible. In this regard, a useful first step is to obtain the functional form of the distributions so that the comparisons can be made in terms of the "sufficient statistics" for the distributions. Our concern will be with the distribution of $\hat{\beta}$, since α and δ can, for the most part, be regarded as nuisance parameters. Thus, while $\hat{\beta}$ is known to be asymptotically normal, we must rely on Monte Carlo experiments to demonstrate the small sample distributions.

To generate the distributions of $\hat{\beta}$, 1000 cases were run on the 10×5 sample, with homogeneous explanatory variables, and an R^2 of

0.5.[1] The results from these experiments, which are given in Table 2, suggest that the MLE and ETV estimators for β are approximately normally distributed. The areas shown for both parameters of both estimators closely approximate the theoretical normal distribution.[2]

The above finding has the important implication that the multi-scale estimators can be described fully by their means and variances. The problem of comparing the estimators is correspondingly simplified. With respect to choosing among the estimators, the criterion that will be used is that of the minimum mean squared error (i.e., the sum of variance and the bias squared).

SAMPLE SIZE

The two dimensions of sample size in the multi-scale model raise a potentially important distinction for the composition of the sample, since the number of parameters to be estimated equals $2J + K$, where J is the number of subgroups and K is the number of explanatory variables.[3] Therefore, the more subgroups there are, the more parameters there are to estimate such that, for a given number of observations, there are fewer degrees of freedom and the ratio of observations to parameters declines. In the discussion below, detailed results are presented for the 5×50 sample and for the 50×5 sample. Summary statistics are then reported for (1) increasing sample size by adding subgroups and (2) the effect of sample composition holding sample size constant. These experiments are all based on homogeneous explanatory variables.

The 5×50 Sample

The Monte Carlo results for five subgroups with 50 observations

---

[1] Although 20 to 100 cases are sufficient to yield reasonably accurate estimates of the means and variances of the distributions, 1000 cases were required to reflect the entire shape of the distributions.

[2] Also shown in Table 2 is a computer generated normal distribution, based on 1000 cases. This shows how the results from even 1000 cases can deviate modestly from the theoretical distribution.

[3] Note that since each subgroup has its own intercept term, there is no general constant term.

Table 2

DISTRIBUTIONS OF THE MLE AND ETV ESTIMATORS AND THE NORMAL VARIABLE[a]

	Normal		MLE		ETV	
	Actual[b]	Computer-Generated[c]	$\tilde{B}_1 - \bar{B}_2$	$\tilde{B}_2 - \bar{B}_2$	$\tilde{B}_1 - \bar{B}_1$	$\tilde{B}_2 - \bar{B}_2$
	(1)	(2)	(3)	(4)	(5)	(6)
Mean	0	-0.025	0.983	1.975	1.109	1.455
Std. dev.	1	1.019	0.117	0.165	0.115	0.103
$\hat{\beta}$			Areas			
$< -3\delta$	0.0013	0.0	0.002	0.002	0.005	0.003
-3δ to -2δ	0.0214	0.019	0.023	0.029	0.020	0.023
-2δ to -1δ	0.1360	0.147	0.129	0.123	0.121	0.122
-1δ to 0	0.3413	0.332	0.347	0.357	0.352	0.354
0 to $+1\delta$	0.3413	0.346	0.350	0.335	0.341	0.340
$+1\delta$ to $+2\delta$	0.136	0.134	0.123	0.127	0.140	0.137
$+2\delta$ to $+3\delta$	0.0214	0.021	0.026	0.026	0.021	0.020
$> +3\delta$	0.013	0.001	0.0	0.001	0.0	0.010

[a]Based on 1000 cases.

[b]Actual probability distribution for a standardized normal variable.

[c]Probability distribution for a standardized normal variable as from the random normal variable generator on the computer.

per subgroup are given in Table 3. These show the intuitively appealing result that all three multi-scale estimators have essentially the same properties in the large sample. This is to be expected for two reasons. First, all three estimators were shown to be consistent when the explanatory variables are drawn from homogeneous populations. Since consistency is defined in terms of increasing sample size holding the number of subgroups constant, and since the 5×50 sample would be considered "large" by most measures, we would expect the means of the distributions for all three estimates to be approximately the same.

Table 3

MONTE CARLO RESULTS FOR THE 5 × 50 SAMPLE[a]

(Five subgroups with 50 observations per subgroup)

		β_1[c]			β_2[c]		
R^2	Estimator[b]	$\bar{\beta}_1$	SV_1	MSE_1	$\bar{\beta}_2$	SV_2	MSE_2
.1	MLE	1.099	0.193	0.203	1.957	0.401	0.403
	ERV	1.108	0.197	0.209	1.971	0.405	0.406
	ETV	1.094	0.189	0.198	1.945	0.398	0.401
.5	MLE	0.991	0.011	0.011	1.995	0.025	0.025
	ERV	0.991	0.011	0.011	1.987	0.025	0.025
	ETV	0.989	0.011	0.011	1.987	0.025	0.025
.9	MLE	0.989	0.003	0.004	2.010	0.002	0.002
	ERV	0.999	0.003	0.004	2.010	0.002	0.002
	ETV	0.993	0.004	0.004	2.001	0.002	0.002

NOTE: Explanatory variables drawn from a homogeneous population.

[a]The basic model is given as $y_{ij} = \alpha_j + \delta_j (\beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + u_i)$. The results are based on 20 cases and homogeneous explanatory variables.

[b]Maximum likelihood (MLE), equal residual variance (ERV), and equal total variance (ETV).

[c]$\beta_1 = 1.0$; $\beta_2 = 2.0$. $\bar{\beta}_i$ refers to the mean for the experiments; SV_i refers to the variance; and MSE_i refers to the mean squared error.

Second, consider the methods of estimation. The ETV estimator explicitly assumes that differences in the within-subgroup variances of the dependent variable are due exclusively to differences in the scale parameters. Since this assumption is in fact correct when the xs are drawn from homogeneous populations and when the number of observations per subgroup is large enough to avoid small sample problems, the ETV estimator provides consistent estimates. Moreover, since the MLE and ERV estimators are different from ETV only so long as differences in within-subgroup variances of the dependent variable are partly

always equal to or less than that for either MLE or ERV. In fact, this always holds regardless of sample size, explanatory variables, or model specification.[1]

Second, althouth ETV always has the smallest variance, it it not always the preferred estimator on the basis of the mean squared error criterion. For low R^2s, ETV does have a smaller mean squared error than either MLE or ERV. However, MLE has a much smaller mean square than the ETV for high R^2s, because although the explanatory variables were drawn from homogeneous populations, the variables themselves will almost necessarily be heterogeneous when there are as few as five observations in a subgroup. Therefore, ETV will be biased in the small sample, where small sample refers to the number of observations per subgroup, even though it is consistent, because ETV attributes all differences in within-subgroup variances in the dependent variable to the scale parameter when, in fact, some of it is due to differences in the variation of the explanatory variables. MLE and ERV, however, explicitly take such differences in within-subgroup explanatory variable variations into account, thus leading not only to consistent estimates, but to estimates that are also unbiased.

Although ETV has the smallest variance, it is sufficiently biased when there are few observations per subgroup that its mean squared error is larger than MLE and ERV for high R^2s. Moreover, ETV does relatively worse as the true R^2 of the model increases for two reasons: (1) the bias in ETV increases as R^2 increases and (2) for higher R^2s, bias plays a relatively more important role in the mean squared error criterion (since the variance decreases as R^2 increases).

Finally, MLE is generally preferred to ERV since the MLE variance tends to be much less than that for ERV. Both will be unbiased, though.

The 5×50 and 50×5 samples yield three important conclusions. First, the three multi-scale estimators all yield approximately the same results when there are many observations per subgroup. Second, when there are few observations per subgroup, ETV has the smallest variance

---

[1] We conjecture that this result occurs because ETV uses a less complicated procedure for estimating the δ vector.

of the estimators. Third, though, ETV is biased when there are few observations per subgroup. Although this bias is not sufficient to counteract the savings in variance for low R^2s, it more than offsets this savings for high R^2s--the result being that ETV is preferred for low R^2s and MLE for high R^2s.

Sample Size

The effects of sample size on the parameter distributions are a necessarily important question. Of particular importance is the effect on the parameter estimates when sample size is increased through the addition of new subgroups, since this will often be the only way of increasing the sample in situations where the multi-scale model is applicable. In the standard linear regression model, the effect of sample size can be solved analytically: (1) the least squares estimator is unbiased, regardless of sample size, and (2) the variance of the least squares estimator is proportional to sample size. In the multi-scale model, however, the result is less clear, for every time a new subgroup is added, two more parameters are also added.

Summary results, showing the marginal distributions of $\tilde{\beta}$ for different sample sizes, holding the number of observations per subgroup constant, are reported in Table 5.[1] These show the perhaps surprising result that increasing the sample through the addition of subgroups reduces the variance of $\tilde{\beta}$ almost in proportion to the number of observations, as with the standard linear regression model, even though each additional subgroup adds two parameters.[2] That is, although the addition of more parameters somewhat reduces the benefit of the additional observations, this reduction is modest. This is an important

[1] For simplicity in presentation, Table 5 reports the sum of mean squared errors for $\tilde{\beta}_1$ and $\tilde{\beta}_2$, rather than the separate mean squared errors. Note that we can perform this simple addition since both x_1 and x_2 are uncorrelated.

[2] This holds for MLE and ERV since both are unbiased. It is also approximately true for ETV at $R^2 = 0.1$, since the bias for ETV is small. For higher R^2s, however, this does not hold for ETV, since the bias in the ETV estimator at higher R^2s does not fall as more subgroups are added.

Table 5

SUMMARY MONTE CARLO RESULTS FOR HOMOGENEOUS EXPLANATORY VARIABLES: MEAN SQUARED ERROR AS A FUNCTION OF SAMPLE SIZE

		$MSE_1 + MSE_2$ for Sample Size[b]				
R2	Estimator[a]	5 × 5	10 × 5	25 × 5	50 × 5	100 × 5
0.1	MLE	4.670	3.485	1.268	0.744	0.336
	ERV	9.493	5.220	1.718	1.087	0.621
	ETV	3.411	2.860	1.093	0.680	0.331
0.5	MLE	0.535	0.406	0.179	0.085	0.040
	ERV	1.008	0.563	0.326	0.144	0.108
	ETV	0.501	0.558	0.271	0.219	0.196
0.9	MLE	0.093	0.041	0.013	0.009	0.007
	ERV	0.132	0.044	0.020	0.015	0.007
	ETV	0.360	0.283	0.178	0.179	0.201

[a] See note b, Table 3.

[b] Mean squared error for $\tilde{\beta}_1$ plus mean squared error for $\tilde{\beta}_2$. Results for 5 × 5 and 10 × 5 samples are based on 100 cases, results for 25 × 5, 50 × 5, and 100 × 5 samples based on 20 cases.

practical result since, as noted above, the only means of adding more observations in situations where the multi-scale model is the appropriate specification may be through the addition of more subgroups.

Sample Composition

Finally, consider the effect of sample composition, holding the number of observations constant. It is clear from the results shown in Table 6 that the more subgroups there are (and, hence, the more parameters), the less precise are the estimates of β. Yet, with the exception of ETV, which becomes severely biased as subgroup size is reduced, the effects of sample composition are not as large as one might expect. Note further that most of the gain from increasing subgroup size, again with the exception of ETV, occurs when the subgroup

Table 6

SUMMARY MONTE CARLO RESULTS FOR HOMOGENEOUS EXPLANATORY VARIABLES: MEAN SQUARED ERROR AS A FUNCTION OF SAMPLE COMPOSITION

		$MSE_1 + MSE_2$ for Sample Size[b]			
R^2	Estimator[a]	5 × 50	25 × 10	50 × 5	50 × 5(var)
0.1	MLE	0.606	0.507	0.744	n.a.
	ERV	0.615	0.561	1.087	n.a.
	ETV	0.599	0.485	0.680	n.a.
0.5	MLE	0.036	0.039	0.085	0.089
	ERV	0.036	0.049	0.144	0.109
	ETV	0.036	0.074	0.219	0.250
0.9	MLE	0.006	0.006	0.009	0.017
	ERV	0.005	0.006	0.015	0.010
	ETV	0.006	0.047	0.179	0.220

[a]See note b, Table 3.

[b]See note b, Table 5.

size is increased from five to ten observations. The estimates are quite unaffected if the subgroup size is variable. In fact, ERV actually does better in the variable subgroup size sample than in the constant subgroup size sample, thus suggesting that it benefits more from the introduction of a few large subgroups than it is hurt by the presence of very small subgroups.

HETEROGENEOUS EXPLANATORY VARIABLES

When the explanatory variables are not homogeneous across all subgroups, ETV yields inconsistent estimates of the parameters. The reason is obvious. ETV attributes all differences in within-subgroup variances of the dependent variable to the scale parameter. However, when the explanatory variables themselves are heterogeneous across subgroups, this is not appropriate. MLE and ERV will still be consistent when the explanatory variables are heterogeneous.

Monte Carlo results for the 5 × 50 sample with heterogeneous explanatory variables are given in Table 7. They show the expected finding that ETV is biased and inconsistent when the explanatory variables are heterogeneous, but that MLE and ERV are both unbiased. As a result, both MLE and ERV are generally to be preferred to ETV. For very low R^2s, ETV is to be preferred in spite of its inconsistency, since in the 5 × 50 sample the smaller variance for ETV more than offsets its bias.

Table 7

MONTE CARLO RESULTS FOR HETEROGENEOUS EXPLANATORY VARIABLES: 5 × 50 SAMPLE

		β_1			β_2		
R^2	Technique	$\bar{\beta}_1$	SV_1	MSE_1	$\bar{\beta}_2$	SV_2	MSE_2
.1	MLE	1.175	0.120	0.150	1.944	0.423	0.427
	ERV	1.183	0.121	0.154	1.956	0.434	0.436
	ETV	1.091	0.109	0.117	1.833	0.350	0.378
.5	MLE	0.981	0.009	0.010	2.005	0.035	0.035
	ERV	0.983	0.010	0.010	2.011	0.034	0.034
	ETV	0.775	0.005	0.056	1.651	0.018	0.140
.9	MLE	0.992	0.004	0.004	2.004	0.003	0.003
	ERV	1.007	0.006	0.006	2.028	0.006	0.007
	ETV	0.678	0.002	0.105	1.470	0.001	0.282

The 50 × 5 sample shown in Table 8 yields largely the same results. ETV is very biased for medium to high R^2s, so that MLE and ERV are again generally preferred to ETV. As before, MLE yields better estimates than ERV. In general, these results illustrate the importance of heterogeneity in the explanatory variables.

THE SCALE PARAMETER

The scale parameter δ is clearly what distinguishes the multi-scale model from the classical linear regression model. Therefore, δ determines when it is appropriate to use one of the multi-scale

of the elements.[1] Finally, a fifth δ vector with all the elements set identical to one is used as a base case.

The Monte Carlo results from experiments conducted on these different δ vectors are shown in Table 9. They show the expected result that as the elements of δ become more widely dispersed--i.e., as the variance of δ, V_δ increases--OLS-DV yields poorer estimates. These results also suggest that there is a tradeoff--or efficiency frontier--that defines the appropriate estimator to use. For example, OLS-DV is clearly the appropriate estimator when all the δs are set identically to one. That is, there is no need to employ the multi-scale model when the true model is not in fact multi-scale.

As the δs diverge from unity, it begins to pay to use one of the multi-scale techniques. In particular, for low R^2s one will want to use ETV when the variance increases to somewhere between 0.0 and 0.04. For high R^2s (in the 50 × 5 sample), one will want to use MLE when the variance of δ get much larger than zero.

SUMMARY

The Monte Carlo results allow us to assess the performance of several competing estimators when the true model is multi-scale in nature. They show that no single estimator is dominant over the entire range of possible sample sizes and model specifications. Instead, the appropriateness of any single estimator depends on a number of factors: (1) sample size, (2) sample composition, (3) signal-to-noise ratio, (4) the degree of heterogeneity in the explanatory variables, and (5) the degree to which the multi-scale model is multi-scale. These can be combined to

---

[1] That is, for any given δ vector, say δ_1, the variance of the δs is given simply by

$$V_\delta = \frac{1}{J} \sum_{j=1}^{J} (\delta_{ij} - \bar{\delta}_1)^2 .$$

The δ vectors used in Table 9 were each generated from log normal distributions. The reason that V_δ in the 5 × 50 sample differs from V_δ in the 50 × 5 sample is simply that five δs were drawn from each vector in the first case, while 50 were drawn in the second.

Table 9

MONTE CARLO RESULTS: COMPARISONS OF MEAN SQUARED ERRORS FOR OLS-DV AND MULTI-SCALE ESTIMATORS

(Homogeneous explanatory variables)

		$MSE_1 + MSE_2$ for							
							Multi-Scale		
Sample	R^2	OLS-DV					MLE	ERV	ETV
5 × 50		$V_\delta = 0.0$	$V_\delta = 0.03$	$V_\delta = 0.24$	$V_\delta = 1.68$	$V_\delta = 35.3$			
	0.1	0.535	0.645	0.860	2.100	49.01	0.606	0.615	0.599
	0.5	0.032	0.037	0.095	1.160	41.51	0.036	0.036	0.036
	0.9	0.005	0.008	0.061	1.070	39.54	0.006	0.005	0.006
50 × 5		$V_\delta = 0.0$	$V_\delta = 0.04$	$V_\delta = 0.33$	$V_\delta = 3.57$	$V_\delta = 411$			
	0.1	0.674	0.717	1.028	5.587	552.8	0.744	1.087	0.680
	0.5	0.061	0.068	0.128	1.709	198.7	0.085	0.144	0.219
	0.9	0.007	0.011	0.050	1.561	195.8	0.009	0.015	0.179

yield regions when particular estimators are to be preferred, such as that shown earlier in Fig. 2 for the 50 × 5 sample.

The Monte Carlo results do enable us to make the following general statements.

- Ordinary least squares with dummy variables is, of course, appropriate when the model is not multi-scale. It is also preferred when the degree to which the model is multi-scale is very small (even though OLS is inconsistent).
- Maximum likelihood always does reasonably well. It is always consistent, and for medium to large R^2s it is the "best" estimator when there are few observations per subgroup.
- Equal residual variance, though always consistent, is never the preferred estimator. Though it generally does reasonably well, it sometimes yields estimates with larger variance, particularly with low R^2.

- Equal total variance sometimes does very well and sometimes very poorly. It always yields estimates with the least variance and is generally preferred for low R^2s. When the explanatory variables are drawn from heterogeneous populations, it yields inconsistent estimates--an inconsistency that becomes severe at medium to high R^2s. Even when the explanatory variables are drawn from homogeneous populations, the heterogeneity that occurs when subgroup sizes are small makes the bias more than large enough to offset any savings in variance.

These statements provide some practical guidelines for the application of the multi-scale estimators. To begin with, the within-subgroup standard deviation of the dependent variable should be calculated to determine whether the model appears to be multi-scale. If the within-subgroup standard deviations are normalized such that their geometric mean equals one, then the variance of these standard deviations can be determined. If this variance exceeds about 1.0 for very small samples or about 0.05 for large samples, then it probably pays to use the multi-scale model.

Second, the ETV estimates should be calculated. If the R^2 is small, then the ETV is probably the best estimator. If the R^2 is moderate to large, then the explanatory variables should be examined to determine whether they are homogeneous. If not, then the maximum likelihood estimates should be computed and used.

VI. AN APPLICATION OF THE MULTI-SCALE MODEL

The multi-scale model can be put into some perspective by applying it to an actual estimation problem. The example used here--Gay's problem of estimating the cost of on-the-job training for airmen--actually served as the genesis of the multi-scale model.

In trying to estimate the cost of on-the-job training for first-term aircraft maintenance specialists in the Air Force, Gay had to rely on supervisory estimates of the individual's productivity during his first term of duty. As shown in Table 10, in addition to the familiar problem of supervisory bias (i.e., the location effect, α), the variances of the dependent variable differed considerably according to subgroup, thus pointing toward the applicability of the multi-scale model.

Table 10

SUBGROUP MEANS AND STANDARD DEVIATIONS FOR GAY'S STUDY[a]

Subgroup	Number of Observations	Mean	Standard Deviation	$\hat{\delta}$ ETV	$\hat{\delta}$ MLE	$\hat{\delta}$ ERV
1	4	$ 3,461	$ 254	0.27	0.27	0.28
2	3	4,184	1,987	2.08	1.62	1.37
3	3	11,017	4,297	4.49	3.69	3.48
4	5	6,359	3,223	3.37	3.25	2.87
5	7	3,176	214	0.22	0.25	0.25
6	5	5,314	409	0.37	0.35	0.33
7	5	5,596	985	1.03	0.96	0.86
8	3	2,514	624	0.65	0.55	0.68
9	4	3,634	1,092	1.14	0.80	0.74
10	6	7,690	1,736	1.82	1.87	1.80
11	8	12,225	2,593	2.71	3.05	3.45
12	11	8,084	732	0.77	0.88	0.98
μ_δ				1.58	1.46	1.42
v_δ				1.69	1.41	1.33

[a]SOURCE: Robert M. Gay, *Estimating the Cost of On-the-Job Training in Military Occupations: A Methodology and Pilot Study*, The Rand Corporation, R-1351-ARPA, April 1974, p. 28.

As can be seen in Table 10, Gay's sample consisted of 64 observations spread across 12 subgroups. The variance of $\hat{\delta}$ implied by the ETV estimator (i.e., the simplest transformation of the data) is 1.69, which suggests that OLS-DV will not yield the best results. Therefore, the multi-scale model is probably appropriate.

Two primary specifications were considered:

$$COST_i = \beta_1 \cdot EXP_i + \beta_2 \cdot ED_i + \beta_3 \cdot APT_i + \beta_4 \cdot S_i + \beta_5 \cdot W_i + \varepsilon_i \tag{6.1}$$

and

$$COST_i = \beta_1 \cdot EXP_i + \beta_2 \cdot ED_i + \beta_3 \cdot APT_i + \beta_4 \cdot S_i + \beta_5 \cdot W_i + \beta_6 \cdot TECH_i + \varepsilon_i \tag{6.2}$$

where $COST_i$ = cost of on-the-job training for the ith individual,
EXP_i = years of possible civilian job experience,
ED_i = years of education,
APT_i = percentile score on mechanical aptitude test,
S_i = dummy variable for whether the individual is from the north (equals 1 if from the south and zero otherwise),
W_i = dummy variable for whether the individual is white (equals 1 if white and zero otherwise),
$TECH_i$ = percentile score on the performance test in technical training and,
ε_i = error term.

The results are given in Table 11. From the results in Table 10, we know that the OLS-DV is not appropriate. Given the R^2 shown for Eq. (6.1) in Table 11, ETV yields probably the best estimates. However, the R^2 for version (6.2) is probably sufficient to warrant use of MLE. Moreover, there is also considerable heterogeneity in the explanatory variables among subgroups. This will tend to make ML

Table 11

ESTIMATES FOR GAY'S OJT MODEL: EQS. (6.1) AND (6.2)

Equation		EXP	ED	APT	S	W	TECH	R^2
(6.1)	OLS-DV	-95.79	-692.7	-35.6	287.7	294.4		0.17
		(199.6)	(270.6)	(17.75)	(504.3)	(534.8)		
	MLE	-28.29	-406.4	24.90	-65.78	483.2		0.26
		(104.2)	(141.3)	(9.26)	(263.2)	(279.2)		
	ERV	-6.45	-424.2	-25.16	-110.0	464.4		0.27
		(104.4)	(141.6)	(9.28)	(253.8)	(279.8)		
	ETV	-38.52	-364.6	-22.08	-15.41	467.1		0.21
		(105.2)	(142.7)	(9.36)	(265.9)	(282.0)		
(6.2)	OLS-DV	-129.3	-662.6	-19.81	205.5	550.0	-89.63	0.21
		(197.6)	(267.1)	(19.85)	(499.0)	(548.4)	(53.55)	
	MLE	-34.55	-379.5	-10.15	-127.8	724.3	-80.79	0.36
		(99.57)	(134.6)	(10.00)	(251.5)	(276.4)	(26.99)	
	ERV	52.01	-358.4	-5.82	-194.9	804.5	-112.7	0.42
		(104.5)	(141.3)	(10.50)	(263.9)	(290.0)	(28.32)	
	ETV	-62.80	-342.8	-10.67	-74.95	652.4	-64.95	0.28
		(101.9)	(137.7)	(10.23)	(257.3)	(282.7)	(27.61)	

estimates even more attractive. Finally, it is interesting to note that all three multi-scale estimators yield similar coefficient estimates, estimates that differ considerably from the OLS results.

The estimated standard errors appearing in Table 11 are derived from the esimate $\hat{\sigma}^2(X'X)^{-1}$ rather than from the more general information matrix that takes into account all parameter values. Thus, these estimated standard deviations are conditional on estimates of δ. The only difference between these standard errors and the OLS errors is that $\hat{\sigma}^2$ is calculated on the basis of T - 2J - K degrees of freedom rather than T - K - 1 degrees of freedom as in OLS. It is possible to calculate the more general variance-covariance matrix of the coefficients by calculating and storing the matrix values $X_j'X_j$ for each subgroup j during the processing of the data. However, the estimates of Var $(\hat{\beta} - \beta)$ are not a byproduct of obtaining $\hat{\beta}$ as in the case of OLS.

VII. CONCLUDING REMARKS

The multi-scale model offers a potentially valuable tool for analyzing manpower problems where the variables of concern are obtained by subjective measurement, such as supervisory ratings of individual performance.[1] Specifically, it is suggested that measures obtained by subjective evaluations, such as supervisory ratings, may include two types of biases: the location bias and the scale bias. Although the location bias can be handled in multiple regression models through familiar dummy variable techniques, the scale bias poses special estimation problems and necessitates the development of a multi-scale estimator.

Of the several multi-scale estimating procedures developed, three are found to be appropriate for real-world applications: (1) OLS (with dummy variables), (2) ETV, and (3) MLE. Although maximum likelihood is the only one of these three procedures always to yield consistent estimates, OLS and ETV may be more appropriate from an efficiency standpoint when consistency is less of a concern--e.g., in small sample situations. In particular, OLS is the appropriate technique in small samples when the true model is only "modestly" multi-scale; ETV is appropriate when there is a scale problem but the R^2 is small; and MLE is appropriate when there is a scale problem and when the R^2 is moderate to large.

The resulting estimates are useful in two respects. First, the multi-scale approach allows the analyst to estimate the parameters in multiple regression models when the dependent variable is subject to the scale transformation. These estimates can then be used to construct "corrected" estimates of the dependent variable.

Some of the limitations of the multi-scale approach and some possible directions for future research are that the multi-scale model, as we have structured it, is not necessarily appropriate for all problems in which the dependent variable is obtained through subjective

---

[1] As noted previously, the multi-scale model may be appropriate in other cases where the data fall into natural groupings.

evaluation. In particular, our formulation requires *cardinal* measures of the dependent variable, not the *ordinal* measures one often finds on supervisory evaluation forms. When cardinal measures are available, though, the multi-scale approach is probably worth investigating, as implied by our analysis of Gay's model. Second, like other techniques, the multi-scale approach is not valid for cases in which the measurement bias is selective. Indeed, the essence of the multi-scale approach rests in the notion that the measurement bias is consistent and systematic within subgroups.

Finally, the problem, as it has been structured here, allows for only one measured observation per "true" observation--e.g., one supervisory rating per individual. Sometimes, though, there may be several subjective evaluations (i.e., measured observations) for each true observation, such as several supervisors rating one individual.[1] It would therefore be desirable to extend the basic multi-scale framework to allow the multiple observation case.

---

[1]This is the case in Gay's current work.

Appendix

MAXIMUM LIKELIHOOD AND LEAST SQUARES ESTIMATES

This appendix derives estimates of the multi-scale model, assuming that the vector of location parameters α is zero. Thus,

$$y_{ij} = \delta_j(X_{ij}\beta + \varepsilon_{ij}) , \qquad \begin{matrix} i = 1, 1, \ldots, T_j \\ j = 1, 1, \ldots, J \end{matrix} \tag{A.1}$$

X_{ij1} may be a vector everywhere equal to unity. The side condition is a strictly separable function

$$G(\delta) = 0 \tag{A.2}$$

with a particular form

$$\sum T_j \ln \delta_j = 0 . \tag{A.3}$$

We write the parameter vector θ as

$$\theta = (\beta, \delta, \sigma^2) .$$

MAXIMUM LIKELIHOOD ESTIMATES

ML estimation utilizes the property that the ε_{ij} are normal independent deviates with $E\varepsilon = 0$ and $E\varepsilon\varepsilon' = \sigma^2 I$. The likelihood function can be written

$$L = \prod_{j=1}^{J} \prod_{i=1}^{T_j} \left(2\pi\sigma^2\delta_j^2\right)^{-\frac{1}{2}} \exp\left[- \frac{1}{2\sigma^2\delta_j^2} \left(y_{ij} - \delta_j X_{ij}\beta \right)^2 \right] \tag{A.4}$$

$$= (2\pi\sigma^2)^{-\frac{T}{2}} \prod_{j=1}^{J} \delta_j^{-T_j} \exp\left[- \frac{1}{2\sigma^2} \sum_j \sum_i \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right)^2 \right] \tag{A.5}$$

ML estimates are found through maximization of the logarithm of the likelihood function, adding a term with a Lagrangian multiplier to account for the side condition (A.2):

$$\log L = -\frac{T}{2} \ln 2\pi\sigma^2 - \sum_j T_j \ln \delta_j - \frac{1}{2\sigma^2} \sum_j \sum_i \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right)^2 + \lambda G(\delta) \ . \tag{A.6}$$

FIRST-ORDER CONDITIONS

The ML estimates $\hat{\theta}$ are solutions to the equations

$$\frac{\partial}{\partial \theta} \log L = 0 \ . \tag{A.7}$$

Partial differentiation yields four sets of equations.

$$\frac{\partial}{\partial \beta} \log L = \frac{1}{\sigma^2} \sum_j \sum_i \left(\frac{X'_{ij} y_{ij}}{\delta_j} - X'_{ij} (X_{ij}\beta)^2 \right) = 0 \tag{A.8.1}$$

$$\frac{\partial}{\partial \sigma^2} \log L = -\frac{T}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \sum_j \sum_i \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right)^2 = 0 \tag{A.8.2}$$

$$\frac{\partial}{\partial \delta_j} \log L = -T_j \frac{1}{\delta_j} + \frac{1}{\sigma^2} \sum_j \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right) \frac{y_{ij}}{\delta_j^2} + \lambda \frac{\partial G}{\partial \delta_j} = 0 \ ,$$

$$j = 1, \ldots, J \tag{A.8.3}$$

$$\frac{\partial}{\partial \lambda} \log L = G(\delta) = 0 \ . \tag{A.8.4}$$

The first two conditions yield ML estimates to the classical normal model:

$$\hat{\beta} = (X'X)^{-1} X'\hat{z} \tag{A.9}$$

and

$$\hat{\sigma}^2 = \frac{1}{T} \sum_j \sum_i \left(\hat{z}_{ij} - X_{ij}\hat{\beta} \right)^2 , \qquad (A.10)$$

where

$$\hat{z}_{ij} = y_{ij}/\hat{\delta}_j .$$

For condition (A.8.3), where the logarithmic form (A.3) of the side conditions is used, the equation may be written

$$\frac{1}{T_j} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} = (1 - \lambda)\hat{\sigma}^2 , \qquad j = 1, \ldots, J . \qquad (A.11)$$

Equation (A.11) is a quadratic equation in $1/\delta_j$. Only the positive root, however, satisfies the nonnegativity conditions for δ.

The necessary and sufficient condition for the existence of a unique solution to this system of equations with all $\hat{\delta}_j > 0$ is that

$$\det |Q| \neq 0$$

where $Q = Y'MY$, M is the $T \times T$ idempotent matrix $I - X(X'X)^{-1}X'$, and Y is the $T \times J$ matrix, which assigns the values of Y_{ij} to separate columns according to subgroup. Thus,

$$Y = \begin{bmatrix} Y_{11} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ Y_{T_1 1} & 0 & \ldots & 0 \\ 0 & Y_{12} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & Y_{T_2 2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & Y_{T_J J} \end{bmatrix} .$$

Some of the implications of this condition are discussed in Section III.

Properties of ML Estimates

Asymptotic properties for the multi-scale model can be defined as

$$T_j \to \infty \text{ for some or all } j ,$$
$$J \to \infty ,$$
$$T_j \to \infty \text{ and } J \to \infty .$$

The usual theorems for ML estimates apply only where the number of parameters is fixed or at least bounded. Since the number of parameters in θ is $J + K + 1$, it will remain fixed only as the number of subsets remains fixed. Hence we are able to define consistency only as

$$T_j \to \infty \text{ for all } j = 1, \ldots, J .$$

We conjecture that with certain restrictions on the data, ML estimates are consistent (in this sense), joint asymptotically normal, and asymptotically efficient.

The literature on the asymptotic properties of ML estimates (see LeCam [16]) suggests as an estimate of the dispersion matrix of the parameters in the limit

$$\lim \sqrt{T}\ (\hat{\theta} - \theta) = \Gamma^{-1}(\theta) , \qquad T \to \infty , \tag{A.12}$$

where

$$\Gamma(\theta) = E \frac{\partial^2 \log L}{\partial \theta^2} .$$

An estimator of the asymptotic distribution of the ML estimates can be calculated from the matrix of second partial derivatives. In particular

$$\Gamma^{-1}(\theta) = -\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}^{-1}, \qquad \text{(A.13)}$$

where

$$A_{11} = E\,\frac{\partial^2 \log L}{\partial \beta^2} = -\frac{1}{\sigma^2}\,(X'X) \qquad \text{(A.13.1)}$$

$$A'_{21} = A_{12} = E\,\frac{\partial^2 \log L}{\partial \delta \partial \beta} \qquad \text{(A.13.2)}$$

$$A_{22} = E\,\frac{\partial^2 \log L}{\partial \delta^2} \qquad \text{(A.13.3)}$$

$$A_{33} = E\,\frac{\partial^2 \log L}{\partial (\sigma^2)^2} \qquad \text{(A.13.4)}$$

and

$$A'_{31} = A_{13} = 0 \qquad \text{(A.13.5)}$$

$$A'_{32} = A_{23} = 0\ . \qquad \text{(A.13.6)}$$

The value $\hat{\sigma}^2(X'X)^{-1}$ will be recognized as the estimate of the variance-covariance of the β coefficients in the classical normal regression model. The need to adjust these estimates to take account of the presence of the vector δ depends on the inverse of the matrix $\Gamma(\theta)$. If $A'_{21} = A_{12} = 0$, then

$$\Gamma^{-1}(\theta) = -\begin{pmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A_{33} \end{pmatrix}^{-1} = -\begin{pmatrix} A_{11}^{-1} & 0 & 0 \\ 0 & A_{22}^{-1} & 0 \\ 0 & 0 & A_{33}^{-1} \end{pmatrix}.$$

Thus, the usual estimators of the variance-covariance matrix will be provided by $\Gamma^{-1}(\theta)$ if and only if

$$A_{12} = \lim_{T_j \to \infty} \frac{\partial^2 \log L}{\partial \delta \partial \beta} = 0 \ . \tag{A.14}$$

In this case lim $\Gamma^{-1}(\theta)$ is a block diagonal matrix, and the var $\sqrt{T}$ $(\hat{\beta} - \beta)$ is merely $-\sigma^2 A_{11}^{-1}$. In general, condition (A.14) will not hold. Suppose the model (A.1) and (A.3) is transformed to replace δ_1 with its value as determined by the side condition:

$$\delta_1 = \left(\delta_2^{-T_2} \ \delta_3^{-T_3} \ \ldots \ \delta_J^{-T_J} \right)^{\frac{1}{T_1}} \ . \tag{A.15}$$

To simplify the problem further, we have assumed that all subsets contain n observations (T = nJ). The expected value of the partial derivative is

$$E \frac{\partial^2 \log L}{\partial \delta_j \partial \beta} = - \frac{1}{\sigma^2_{\delta_j}} (X_1'X_1\beta - X_j'X_j\beta) \ , \tag{A.16}$$

for all j. The only way for these values to be zero for all parameter values is for

$$X_1'X_1 = \ldots = X_J'X_J \ . \tag{A.17}$$

Hence if the raw moment matrices tend to equality in the limit, then

$$E \frac{\partial^2 \log L}{\partial \delta \partial \beta} = 0 \ , \tag{A.18}$$

and the asymptotic variance of $\hat{\beta}$ is merely

$$\frac{\sigma^2}{T} (X'X)^{-1} .$$

Thus, the "t-scores" in the regression program would require no special adjustment (except for degrees of freedom) in this special case.[1]

The addition of the vector δ to the classical linear model should in general increase the variance of $\hat{\beta}$. However, this cannot be demonstrated here for the multivariate case. The asymptotic variance of $\hat{\beta}$ is

$$\text{Var } (\hat{\beta} - \beta) = \frac{\sigma^2}{T} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} , \qquad (A.19)$$

according to the rule for invariance of a partitioned matrix. A_{11}, A_{12}, and A_{21} are as previously defined. A_{22} represents

$$E\left[- \frac{\partial^2 \log L}{\partial \delta^2} \right] .$$

The diagonal elements of this matrix are

$$\frac{1}{\sigma^2 \delta_j^2} \left(4n\sigma^2 + \beta' X_1' X_1 \beta + \beta' X_j' X_j \beta \right)$$

and the off-diagonal elements are

$$\frac{1}{\sigma^2 \delta_j \delta_h} \left(2n\sigma^2 - \beta' X_1' X_1 \beta \right) .$$

The adjustment to be made to the asymptotic variance of $\hat{\beta}$ can be shown exactly for the case with one behavioral parameter in β and two subsets. Using side condition (A.3) to replace δ_1, δ_2 with a single

[1] In all estimates, we reduced the total number of degrees of freedom by 2J + K to take account of the number of parameters in α, β, δ.

value of δ results in the following matrix for

$$E\left[-\frac{\partial^2 \log L}{\partial\theta^2}\right],$$

where again both subsets are of size n:

$$\lim_{T\to\infty} \text{Var}\ (\hat{\theta} - \theta) = \frac{\sigma^2}{T}\begin{pmatrix} \Sigma x^2 & \frac{\beta}{\delta}(\Sigma_1 x_1^2 - \Sigma_2 x_2^2) & 0 \\ \frac{\beta}{\delta}(\Sigma_1 x_1^2 - \Sigma_2 x_2^2) & \frac{4n + \beta^2 \Sigma x^2}{\delta^2} & 0 \\ 0 & 0 & \frac{n}{\sigma^2} \end{pmatrix}^{-1}. \tag{A.20}$$

It can be shown that the limiting value of the variance of $\hat{\beta}$ is

$$\frac{\sigma^2}{T\Sigma x^2}\left(\frac{2 + \beta^2 \sigma x^2}{2 + \beta^2 \sigma_{x1}^2 \sigma_{x2}^2 / \sigma_x^2}\right). \tag{A.21}$$

The expression before the parenthesis is the Var $\hat{\beta}$ in the classical normal model. The expression in parentheses takes on a minimum value 1 where

$$\frac{\Sigma_1 x_1^2}{T/2} = \frac{\Sigma_2 x_2^2}{T/2}.$$

Where the variances of subsets 1 and 2 differ, the Var $\hat{\beta}$ takes on larger values. For $\beta = 1$ and

$$\frac{\Sigma_1 x_1^2}{T/2} = 9\,\frac{\Sigma_2 x_2^2}{T/2},$$

$\lim_{T\to\infty}$ Var $\hat{\beta}$ is nearly twice as great as where the subset variances are equal.

LEAST SQUARES ESTIMATES

LS estimation minimizes the sum of squared residuals subject to the side condition in the vector δ. In particular,

$$S^2 = \sum_j \sum_i \epsilon_{ij}^2 = \sum_j \sum_i \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right)^2 + \lambda G(\delta) \ . \qquad \text{(A.22)}$$

The first-order conditions with respect to β and δ are

$$\frac{\partial S^2}{\partial \beta} = -2 \sum \sum \left(X_{ij} \frac{y_{ij}}{\delta_j} - X_{ij}' X_{ij} \beta \right) \qquad \text{(A.23.1)}$$

and

$$\frac{\partial S^2}{\partial \delta_j} = -2 \sum_j \left(\frac{y_{ij}}{\delta_j} - X_{ij}\beta \right) \frac{y_{ij}}{\delta_j^2} + \lambda \frac{\partial G(\delta)}{\partial \delta_j} \ . \qquad \text{(A.23.2)}$$

(A.23.1) differs from the same condition for ML estimates (A.8.1) only by a constant $(-1/2\sigma^2)$, and hence yields the same conditional estimates for $\hat{\beta}$.

Under side condition (A.3) LS and ML estimates of β and δ are identical. The first-order conditions for LS and ML are

$$\text{ML:} \quad \frac{1}{T_j} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} = (1 - \lambda)\hat{\sigma}^2 \qquad \text{(A.24.1)}$$

$$\text{LS:} \quad \frac{1}{T_j} \sum_j \left(\frac{y_{ij}}{\delta_j^*} - X_{ij}\beta^* \right) \frac{y_{ij}}{\delta_j^*} = \frac{1}{2} \lambda^* \ ,$$

$$j = 1, \ \ldots, \ J \ . \qquad \text{(A.24.2)}$$

Inasmuch as $\hat{\sigma}^2$ is constant for all j, then (A.24.1) and (A.24.2) differ only by a constant and, thus,

$$\lambda^* = 2(1 - \hat{\lambda})\hat{\sigma}^2 \ . \tag{A.25}$$

Moreover $\beta^* = \hat{\beta}$ and $\delta^* = \hat{\delta}$. Asymptotically, LS estimates have all the properties of ML estimates plus

$$\text{plim } \lambda^* = 2\sigma^2 \ . \tag{A.26}$$

It can be proved that for ML and LS estimates to be equivalent the side condition must be in the identical form of (A.3). Take any side condition $G(\delta) = 0$ that is strictly separable in the δ_j. The ML condition for $\hat{\delta}_j$ is

$$\frac{1}{T_j} \sum_j \left(\frac{y_{ij}}{\hat{\delta}_j} - X_{ij}\hat{\beta} \right) \frac{y_{ij}}{\hat{\delta}_j} = \left(1 - \hat{\lambda} \frac{\hat{\delta}_j}{T_j} \frac{\partial G}{\partial \hat{\delta}_j} \right) \hat{\sigma}^2 \ . \tag{A.27}$$

For LS the condition is

$$\frac{1}{T_j} \sum_j \left(\frac{y_{ij}}{\delta_j^*} - X_{ij}\beta^* \right) \frac{y_{ij}}{\delta_j^*} = \frac{1}{2} \lambda \frac{\delta_j^*}{T_j} \frac{\partial G}{\partial \delta_j^*} \ . \tag{A.28}$$

The conditions are equivalent so that $\hat{\beta} = \beta^*$ and $\hat{\delta} = \delta^*$ under some weak conditions if and only if the right sides of (A.27) and (A.28) are the same for every subset and do not depend on any element of $\hat{\beta}$ or $\hat{\delta}$. Thus, at most we could have

$$\frac{\hat{\delta}_j}{T_j} \frac{\partial G}{\partial \hat{\delta}_j} = f(G) \ . \tag{A.29}$$

But since $G = 0$, then $f(G) = c_o$, a constant. Since G is strictly separable into $G_1(\delta_1) + \ldots + G_J(\delta_J)$, $\partial G/\partial \delta_j$ can be written as G_j'. Thus,

$$G_j' \frac{\hat{\delta}_j}{T_j} = c_o \ . \tag{A.30}$$

Simple integration yields

$$G_j = c_o T_j \ln \hat{\delta}_j + C_1 \ , \tag{A.31}$$

where both c_o and C_1 are arbitrary constants. Thus, for LS and ML to be equivalent, we must have

$$G = c_o \sum_{j=1}^{J} T_j \ln \hat{\delta}_j + C_1 \ . \tag{A.32}$$

In (A.3) $c_o = 1$ and $C_1 = 0$.

For a function G not satisfying (A.30) either for small samples or in the limit, then LS and ML estimates are not equal in the limit and LS yields estimates that in general are inconsistent.

REFERENCES

1. Gregory C. Chow, "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, Vol. 28, no. 3, pp. 591-605, July 1960.

2. J. Johnston, *Econometric Methods*, McGraw-Hill, New York, 1963.

3. T. D. Wallace and A. Hussain, "The Use of Error Components Models in Combining Cross Section with Time Series Data," *Econometrica*, Vol. 37, no. 1, pp. 55-72, January 1969.

4. P. Ballestra and M. Nerlove, "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model," *Econometrica*, Vol. 34, no. 3, pp. 585-612, July 1966.

5. M. Nerlove, "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, Vol. 39, no. 2, pp. 359-382, March 1971.

6. M. Nerlove, "A Note on Error Components Models," *Econometrica*, Vol. 39, no. 2, pp. 383-396, March 1971.

7. C. Hsiao, "Some Estimation Methods for a Random Coefficient Model," *Econometrica*, Vol. 43, no. 2, pp. 305-326, March 1975.

8. H. O. Hartley and K.S.E. Jayatillake, "Estimation for Linear Models with Unequal Variances," *Journal of the American Statistical Association*, Vol. 68, no. 341, pp. 189-192, March 1973.

9. R. M. Gay, *Estimating the Cost of On-the-Job Training in Military Occupations: A Methodology and Pilot Study*, The Rand Corporation, R-1351-ARPA, April 1974.

10. R. M. Gay and G. R. Nelson, "Cost and Efficiency in Military Specialty Training," The Rand Corporation, P-5160, January 1974.

11. Roberta J. Smith, *Analytical Methods for Adjusting Subjective Rating Schemes: Computer Algorithms*, The Rand Corporation, R-1685/2-ARPA, forthcoming.

12. M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, Hafner, New York, 1961, p. 61.

13. A. Zellner, *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York, 1971, p. 114.

14. B. Hoadley, "Asymptotic Properties of Maximum Likelihood Estimators for the Independent Not Identically Distributed Case," *The Annals of Mathematical Statistics*, Vol. 42, no. 6, pp. 1977-1991, 1971.

15. R. A. Bradley and J. J. Gart, "The Asymptotic Properties of ML Estimators when Sampling from Associated Populations," *Biometrika*, Vol. 49, nos. 1 and 2, p. 205, 1962.

16. L. LeCam, "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates," *University of California Publications in Statistics*, Vol. 1, no. 11, 1953, pp. 277-329.

17. R.V.L. Cooper, *Analytical Methods for Adjusting Subjective Rating Schemes: Monte Carlo Results*, The Rand Corporation, R-1685/1-ARPA, forthcoming.