DISCUSSION PAPER

The Measurement of Learning and Retention Curves for Basic Skills in Egyptian Primary Education I: An Application of Censored Analysis of Variance

Maximum Likelihood Estimation of the Truncated and Censored Normal Regression Models

by

Michael J. Hartley

and

Eric V. Swanson

Development Research Department
Economics and Research Staff
World Bank

The views presented here are those of the author, and they should not be interpreted as reflecting those of the World Bank
The Measurement of Learning and Retention Curves for Basic Skills in Egyptian Primary Education I: An Application of Censored Analysis of Variance

by

Michael J. Hartley*

and

Eric V. Swanson**

World Bank
August 1980

This paper has been completed as part of a joint World Bank/National Center for Educational Research (WB/NCER) research project entitled The International Study for the Retention of Literacy and Numeracy (RPO 671-55) financed by The World Bank. The authors are indebted to their colleagues and collaborators Messrs. I. Serageldin, M. Wilson, R. Durstine, E.M. Abdel-Mawgood, J. Issa, S. Jarrar, W. Haddad and S. Heyneman for their encouragement and thoughtful comments during the course of this project, to Messrs. B. B. King, E. Stoutjesdijck and M. Selowsky for permitting the considerable time input that has been required and to our collaborators at NCER, Dr. Youssef Khalil and Messrs. N. Hamshary and S. Saleh for their painstaking work in development of the test instruments and organization of the field work and data editing. Finally, Prof.s. I. Werdelin, M. Chaibderaine, A. El-Koussy, D. Cole, S. Ibrahim, N. Postlathwaite, C. Hoyt and D. Poirier served as consultants during earlier phases of the work, whilst Ms. V. Bencivenga provided excellent research assistance.

* Economist, Development Economics Department (DED)

** East and North Africa Projects (EMP) and Development Economics Department (DED)
I. Introduction:

The loss of literacy and numeracy skills by school leavers is a major administrative and pedagogical problem for educational systems in developing countries. Systems characterized by high dropout rates are said to incur "educational wastage," implying that resources expended on students who drop out are entirely wasted if the student lapses into a permanent state of illiteracy. The studies of Chabderaine (1978) and Prodito and Kapoor (1975) are examples of attempts to measure the causes and consequences of "educational wastage". A common prescription is that students should be obliged to remain in school until a certain minimum or "threshold level" of literacy and numeracy has been attained. Students with skill levels at or above such a threshold, it is hypothesized, will not revert to illiteracy or innumeracy.

In the Arab Republic of Egypt (ARE) concern for the teaching of basic skills has led to extensive discussion of educational reform in the primary grades. Among the recommendations made in a recent working paper on educational reform issued by the Ministry of Education (MOE) are the revision of school curricula, the upgrading of teacher competencies, and construction of new elementary school buildings. Steps have also been taken to extend the period of compulsory schooling from the sixth to the eighth grade.

In order to evaluate the extent of "wastage" in the Egyptian school system and to establish the existence of threshold effects, we have been

---

1/ The working paper was published in 1979 in Arabic under the title "Educational Reform and Innovation Working Paper." An English translation of the key points was prepared for us by one of our colleagues, M.E. Abdel-Mawgoud.
involved in a multi-year research effort to measure comparative skill-levels of primary school students (in-schoolers) and school leavers (dropouts) in Egypt. In this paper we report on the results of an analysis of variance (ANOVA) approach to the estimation of skill-specific "learning" and "retention" curves based on a single cross-section sample of approximately 4400 in-schoolers and dropouts. Within the ANOVA framework, we are also able to measure differences in achievement levels attributable to sex and urban/rural location of Egyptian children, as well as control for school-specific effects.

Since there were no suitable standardized tests of basic skills available in Arabic, measurement of learning and retention curves (covering in-schoolers and dropouts from grades three to six) required the development of a battery of new test instruments. Literacy was measured by reading and writing tests, and numeracy by tests on a simple arithmetic operations, problem solving and geometry. Finally, tests to measure "verbal and non-verbal intelligence" were designed.

A major problem in the design of such tests is to develop a set of items for a particular skill which is:

(1) representative of both the range of student abilities and the evolution of school curricula over grades;

(2) applicable to both in-schoolers and dropouts; and

\[\text{\footnotesize 1/}\]

The test instruments were prepared by a committee of Egyptian educators under the direction of Dr. A. A.H. El-Koussy and Dr. Ingvar Werdelin, both consultants to the project. Where possible, existing test instruments were reviewed and culled for suitable items.
(3) produces sufficient variation in test scores across the samples of testees to afford legitimate comparisons between students.

At the same time the length of tests must be restricted, with obvious implications for the time and money costs of administration. Satisfying these conditions with newly designed test instruments clearly requires careful piloting of alternative item choices prior to actual implementation. In the present case, the principal difficulty was to select test items with difficulty levels appropriate not only to in-schoolers over the range of grades (3 to 6), but also to dropouts from each of those grades. Failure to satisfy this criterion is evidenced by the "large" proportion of "zero" and, to a lesser extent, "perfect" scores on certain of the tests—see Table 3.3 below. In short, the problem encountered is that the "width" of certain of the test instruments—particularly writing skills—has been found to be too "narrow" for the resulting range of abilities. 1/

While it would have been desirable to have anticipated this problem ex ante (in the piloting phase of the study), we are, regrettably, in the position of having to adjust for the presence of "narrow band" test instruments ex post—by development of statistical methods appropriate to modeling test scores whose distribution exhibits an accumulation of probability mass at each of the two extremes. Indeed, it is probably fair to claim that virtually all test instruments are susceptible to this problem, whereby "zero scores" represent student abilities "below" that of the easiest test item and "perfect scores" represent abilities "above" the difficulty

1/ It should, perhaps, be noted that certain skills may be of an inherently "either/or" type prior to the onset of learning—i.e., one may have to "learn how to read, etc." before beginning to improve one's reading skills, and similarly with writing.
level of the hardest item — see, e.g., Wright and Stone (1979).

Our purpose in the present paper is to estimate the skill-levels of in-schoolers and dropouts within each cell — a "cell" being characterized by the \( (j, k) \) pair, where \( j \) denotes "grade-last-attended" and \( k \) denotes "years-out-of-school" — while controlling for sex, location and school-specific effects. The "natural" specification of this model for test scores is in terms of the standard \( n \)-way ANOVA model with unequal numbers of observations per cell (or its equivalent linear regression model representation). However, in order to perform tests of hypotheses within this framework it is customary to assume that the error terms follow a normal distribution — see, e.g., Graybill (1961) — implying that the test scores have an infinite range \((-\infty, +\infty)\). In our case, the domain is finite, and restricted to the range from zero to the maximum possible number of points on any test. In view of this, and the large proportion of test scores at the extremes, the use of standard ANOVA computer packages is clearly inappropriate — i.e., results in biased and inconsistent parameter estimates and inappropriate tests of hypotheses.

Accordingly, the major methodological contribution of this paper is to postulate an ANOVA model in which the resulting test scores follow a bilaterally censored normal distribution, with an accumulation of probability mass at the two extremes and a continuous distribution in between. A theoretical treatment of this problem — including the statistical properties of, and algorithms to obtain parameter estimates has been given elsewhere (see Hartley and Swanson (1980)). Our purpose, here, is to apply this methodology to the measurement to literacy/numeracy learning and retention curves.
The plan of this paper is as follows: In section 2 we define the concepts of learning and retention curves for basic skills and develop a simple linear model for their measurement. Section 3 describes the cross-section sample and test instruments employed. In section 4 we present the basic results from the use of the bilateral censored regression approach, and contrast these with the standard OLS ANOVA estimates. Some concluding remarks, tentative policy conclusions and caveats are reserved for section 5.
2. **Measurement of Learning and Retention Curves:**

2.1. **Definition of Individual Learning and Retention Curves:**

By a learning or a retention curve, we mean the locus of points which traces a particular individual's skill-levels over time. Thus a "learning curve" shall refer to the sequence of particular skill-levels associated with a given individual during the years he/she is in school, whilst "retention curves" shall refer to the skill-level sequence beginning with the last year in school and followed by the succession of years during which the individual is considered a school-leaver. Hence, the skill-level associated with the last school-year is both the last element in the learning curve sequence and the base-year reference point for an individual's retention curve.

More formally, let \( s=1,2,\ldots, S \) refer to the set of "skills" of interest and let \( t=t_1, t_2, \ldots, t_i, \ldots, T_i \) refer to the sequence of school-years on which measurements may be taken for individual \( i \), where

- \( t_1 \) = the first relevant school-year for individual \( i \) (the larger of the year of enrollment in grade 3 and the first year of observation on individual \( i \))
- \( t_i \) = the school-year of last attendance for individual \( i \), and
- \( T_i \) = the last school-year of observation on individual \( i \).

Note that the number of annual observations on each individual, i.e., \( T_i-t_i+1 \), for individual \( i \), may then be defined as

\[
(2.1) \quad \{y_{i,s}(t) : t = t_1, \ldots, t_i, \ldots, T_i\}
\]

for skill \( s \), where the pair \((j_i(t), k_i(t))\) refers to the "cell" associated with individual \( i \) in year \( t \),
\( j_i(t) \) = grade last attended by individual \( i \) as of year \( t \), and 
\( k_i(t) \) = number of years out of school for individual \( i \) as of year \( t \).

In the present study our cross-section sample restricts \( j \) to the grades 3, 4, 5 or 6 and \( k \) to the years 0, 1, \ldots, 4, where \( k<0 \) refers to an "in-schooler" and \( k>0 \) to a "dropout".

An individual's learning curve may therefore be defined as the skill sub-sequence,
\[
(y^*_s(t), j_i(t), k_i(t), 0 : t = t_i, \ldots, T_i)
\]
where, by definition, \( k_i(t) = 0 \) for \( t = t_i, \ldots, T_i \). Similarly, an individual's retention curve is defined by the sub-sequence,
\[
(y^*_s(t), j_i(t), k_i(t), t = t_i, \ldots, T_i)
\]
where \( j_i(t) = j_i(t_i) \) for \( t = t_i, \ldots, T_i \). In (2.3) we shall refer to the skill-level associated with the year-last-in-school, \( y^*_s(t_i), j_i(t_i), k_i(t_i) \), as the base-year reference level for individual \( i \)'s retention curve for skill \( s \), and note that the retention curve will refer to the grade last attended, \( j_i(t_i) \).

2.2. **ANOVA Models for Measurement of Learning and Retention Curves:**

The present paper addresses the problem of measuring the learning and retention curves for each skill on the basis of a single cross-section for the school-year 1978/79. From the sample of data described in Section 2.1, let \( t \in I = \{ i : t_i < \tau_i \leq T_i \} \),

\[
I = \{ i : t_i < \tau_i \leq T_i \}.
\]
Let $N_\tau$ denote the number of individuals within $I_\tau$, renumbered as $i=1, 2, \ldots, N_\tau$.

Then the available cross-section sample consists of

\[(2.5) \quad \{y^*_i(s), j^*_i(\tau), k^*_i(\tau) : i = 1, \ldots, N_\tau\},\]

where the particular skill, $s$, measured on each individual within the sample, depends on the grade, $j^*_i(\tau)$, and in-schooler/dropout status, $k^*_i(\tau)$, in year $\tau$ (see section 3 below). Insofar as $\tau$ is constant within a cross-section sample, we shall omit reference to it in subsequent discussion.

We shall adopt the following regression model for each particular skill, $s=1, \ldots, S$,

\[(2.6) \quad y^*_i(s) = \mu_i(s) + \epsilon_i(s),\]

where $j=j^*_i$, $k=k^*_i$, $\mu_{ijk}^{(s)}$ denotes the "regression function" and $\epsilon_{ijk}^{(s)}$ is a normal disturbance,

\[(2.7) \quad \epsilon_{ijk}^{(s)} \sim \text{n.i.d.}(0, \sigma^2(s)).\]

Our interest resides in estimation of the skill-level associated with each $(j,k)$ cell ($j=3,4,5,6$ and $k=0,1,2,3$ and 4) by "pooling" across all individuals within the particular cell in year $\tau$. The simplest specification of the regression function to be considered is given by

\[(2.8) \quad \mu_{ijk}^{(s)} = \mu_{jk}^{(s)}.\]

This is equivalent to a simple 2-way "saturated" ANOVA model in which all interactions are included and years-out-of-school are viewed as "treatments" which result in the skill-level, $y^*_i(s)$.

The above reparametrization, (2.8), is more useful for present purposes than the customary ANOVA form (in terms of a grand mean, main treatment effects and interactions).
Two simple extensions of the 2-way ANOVA model are also investigated.

Since, in addition to the test scores associated with each skill, data are also available on the sex of the individual and the school number — the first 30 being urban schools and the last 30 being rural — we may extend the model, (2.8), as follows: Either

\[ \mu_{ijk}^{(s)} = \mu_{jk}^{(s)} + \sigma_{(s)} \cdot \text{sex}_i + \gamma_{(s)} \cdot \text{com}_i, \]

or

\[ \mu_{ijk}^{(s)} = \mu_{jk}^{(s)} + \sigma_{(s)} \cdot \text{sex}_i + \sum_{l=1}^{\text{NS}-1} \delta_l^{(s)} \cdot \text{sch}_{l,i}, \]

where

\[ \text{sex}_i = \begin{cases} 1 & \text{if individual } i \text{ is male} \\ 0 & \text{if individual } i \text{ is female} \end{cases}, \]

\[ \text{com}_i = \begin{cases} 1 & \text{if individual } i \text{ resides in an urban community} \\ 0 & \text{if individual } i \text{ resides in a rural community} \end{cases}, \]

and, for \( l = 1, 2, \ldots, \text{NS}-1 \)

\[ \text{sch}_{l,i} = \begin{cases} 1 & \text{if individual is associated with school } l \\ 0 & \text{otherwise} \end{cases}, \]

where \( \text{NS}=50 \), the total number of sample schools. Models (2.9) and (2.10) are simply reparametrizations of the 4-way ANOVA model with complete interactions between the \( j \) and \( k \) treatments.

In model (2.8), \( \mu_{jk}^{(s)} \) may be interpreted as the mean skill-level associated with the \((j,k)\) cell. Common practice would suggest the usual ANOVA estimate of \( \mu_{jk}^{(s)} \) as the mean skill-level over all individuals found within the \((j,k)\) cell in year \( r \). We shall argue that in the present context — bilaterally censored test scores — such an estimator is biased and inconsistent.

In models (2.9) and (2.10), the parameters \( \mu_{jk}^{(s)} \) have the interpretation of mean skill-levels associated with each \((j,k)\) cell, but "corrected" for the effects of sex and community type (rural/urban) or school-specific effects. Thus, in both models \( \alpha_{(s)} \) measures the "importance of being male".
\( \gamma(s) \) measures the contribution to skill-levels of residing in an urban community in model (2.9), whereas the \( \{ \delta_l \} \) parameters provide school-effect contrasts - i.e., the contribution of factors associated with school \( l \) (relative to school NS) to skill-levels.

2.3 The Problem of Censoring:

In section 1 we have discussed the problem of censoring in the distribution of test scores associated with any skill, i.e., the fact that the range of possible test scores is finite and that the observed frequency distribution exhibits an accumulation of probability mass at the two extremes - zero and perfect scores.

To motivate subsequent discussion it may be useful to describe the process by which censoring takes place. \(^1\) Our view here is that each individual possesses or acquires a stock of latent (unobserved) basic skills. In order to measure those skills, individuals are subjected to a battery of test instruments - each containing a finite set of items of varying "inherent difficulty". An individual's score on a given test may be calculated as the sum of the number of correct item responses, or, in the case of multiple points per item, as the sum of points awarded (with its implicit weighting system). Let \( M(s) \) = maximum score possible on test \( s \) for \( s = 1,2,\ldots,S \). Then each individual's test score, \( \gamma_{ijk}^{(s)} \), must satisfy the inequality.

\[
0 \leq \gamma_{ijk}^{(s)} \leq M(s), \quad i = 1,2,\ldots,N_t.
\]

On this "measurement scale" the test instrument being employed is incapable of discriminating between underlying skill-levels below 0 and above \( M(s) \), in much the same way that a thermometer is unable to measure temperatures below, say, zero or above its maximum capacity. Thus, the events of "zero" and "perfect" scores

\(^1\) Our motivation is similar to that of Wright and Stone (1979) in their discussion of individual ability and item difficulty. In the present paper, however, we depart from their paradigm for the Rasch model in order to make full use of the observations on individuals with zero and perfect scores which constitute a large proportion of our total sample (see section 3) and would be discarded in the Rasch approach.
being registered do not convey actual measurements of skill-levels, but rather
the information that the individual's skill-level is less than or equal to 0 or
greater than or equal to $M^{(s)}$, respectively. In short, our view is that
each test instrument is incapable of measuring underlying skill-levels out-
side a finite range, $(0, M^{(s)})$.

To formalize these notions, it should be noted that we have drawn a
distinction between an individual's underlying or latent skill-level, $y^{(s)}_{ijk}$,
and his/her corresponding test score, $y_{ij}^{(s)}$, where, in contrast to (2.14), the
skill-level range is potentially infinite, i.e.,

$$2.15 \quad -\infty \leq y_{ijk}^{(s)} \leq +\infty.$$ 

Next, note that we may write each of the models, (2.6) combined with (2.8), (2.9)
or (2.10), as

$$2.16 \quad y_{ijk}^{*} = X_i \beta^{(s)} + e_{ijk}^{(s)} = \mu_{ijk}^{(s)} + \varepsilon_{ijk},$$

where $X_i$ is a $K \times 1$ vector of suitably chosen (0,1) "dummy variables" and $\beta^{(s)}$ is
a $K \times 1$ vector of parameters. \(^1\) Finally, the relationship between $y_{ijk}^{*}$ and $y_{ij}$
is defined by

$$2.17 \quad y_{ij}^{(s)} = \begin{cases} y_{ijk}^{*}, & \text{if } y_{ijk}^{*} \leq 0 \\ y_{ijk}, & \text{if } 0 < y_{ijk}^{*} < M^{(s)} \\ M^{(s)}, & \text{if } y_{ijk}^{*} \geq M^{(s)} \end{cases}$$

representing the fact that the test scores, $y_{ij}$, are doubly or bilaterally
censored at 0 and $M^{(s)}$. Thus, $y_{ijk}^{*}$, in the above model, is not observed over
its entire range, (2.15), but only within the bounds of the measurement test
instrument.

\(^1\) E.g., in the model (2.6) and (2.9) we have $K=22$ with

$$\beta^{(s)} = [\beta_3^{(s)}, \beta_4^{(s)}, \beta_5^{(s)}, \beta_6^{(s)}, \alpha^{(s)}, \gamma^{(s)}],$$

$$\mu_{ij}^{(s)} = [\mu_{i0}^{(s)}, \mu_{i1}^{(s)}, \mu_{i2}^{(s)}, \mu_{i3}^{(s)}, \mu_{i4}^{(s)}],$$

$$X_i = [d_{i13}^{(s)}, d_{i14}^{(s)}, d_{i15}^{(s)}, d_{i16}^{(s)}, \text{sex}_1, \text{com}_1],$$

$$d_{ij}^{(s)} = [d_{ij0}^{(s)}, d_{ij1}^{(s)}, d_{ij2}^{(s)}, d_{ij3}^{(s)}, d_{ij4}^{(s)}],$$

$$d_{ijk} = \begin{cases} 1 \text{ if individual } i \text{ is in cell } (j,k) \text{ in year } \tau \\ 0 \text{ otherwise} \end{cases}$$

for $s = 1, \ldots, S; \ i = 1, \ldots, N_t; \ j = 3, \ldots, 6; \ k = 0,1, \ldots, 4$. 
2.4 M. L. Estimation of the Bilaterally Censored Normal Regression Model:

We have argued that each of the ANOVA models can be parametrized (as in (2.16)) as a K-variate regression model, subject to bilateral censoring (as defined in (2.17)), with a normal and independent disturbance term, \( \varepsilon^{(s)}_{ijk} \) of (2.7). For each skill, \( s \), the problem is to estimate the K-vector, \( \beta^{(s)} \), of regression coefficients for skill-levels and the error variance, \( \sigma^2(s) \), based on a sample of \( N_t \) cross-section observations on test scores, \( y^{(s)}_{ijk} \), and associated dummy variable K-vector, \( \mathbf{x}_i^{1/2} \).

This problem was initially posed by Rosett and Nelson (1975), where the suggestion was to calculate the Maximum Likelihood (ML) Estimator of \( \beta \) and \( \sigma^2 \), an expression for the log-likelihood was obtained — to be maximized via the Newton-Raphson algorithm, and certain methods for obtaining initial parameter estimates of \( \beta \) and \( \sigma^2 \)—all of which are, unfortunately, inconsistent — were offered. The following properties for the left-censored regression or Tobit model were proved by Amemiya (1973), and can be shown to apply, mutatis mutandis, to the present case (see Hartley and Swanson, 1980):

1. Ordinary Least Squares (OLS) estimates are biased and inconsistent;
2. the ML estimator, defined as a root of the likelihood equations, is consistent and asymptotically normal (CAN) and asymptotically efficient with an asymptotic covariance matrix given by the inverse of the information matrix;
3. an instrumental variable (IV) estimator, utilizing only the observations between the limit points, can be derived, which is weakly consistent, but asymptotically inefficient (relative to the ML estimator); and
4. if the Newton-Raphson (N-R) algorithm is utilized, starting from a consistent initial estimator (such as the IV estimator in (3)), the first N-R iteration's parameter values are CAN, and have the same

---

1/ In the balance of this subsection we shall delete the superscript ‘\( s \)’ insofar as the same analysis applies to each skill/test.
asymptotic distribution as that of the convergent N-R iterate, viz., the MLE, though the numerical values will, of course, differ.

In addition to the N-R algorithm, Hartley and Swanson (1980) also examine the use of the Gauss-Newton (G-N) algorithm — see Hartley (1961) and Berndt, Hall, Hall and Hausman (1974), the Method of Scoring (MS) — see, e.g., Rao (1965), and the E-M algorithm — see Hartley (1958), Hartley (1976) and Dempster, Laird and Rubin (1977) — for evidence on the efficiency of computational methods to calculate the MLE.1/

In view of (2.7) and (2.16) we may write the density function for \( y^* \) as

\[
(2.18) \quad f_i(y^*_{ijk}) \equiv f(y^*_{ijk}; \xi_i) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left\{ -\frac{1}{2\sigma^2} (y^*_{ijk} - \mu_{ijk})^2 \right\},
\]

where

\[
(2.19) \quad \mu_{ijk} = \mu_{jk}(\xi_i) = \xi_i \beta
\]

denotes the regression function of \( y^*_{ijk} \). Thus, the observed dependent variable, \( y_{ijk} \), has a density defined by

\[
(2.20) \quad g_i(y_{ijk}) \equiv g(y_{ijk}; \xi_i) = \begin{cases} 
    f_i(y^*_{ijk}) & \text{if } 0 < y_{ijk} < M \\
    1 - f_i(M) & \text{if } y_{ijk} = M
\end{cases}
\]

1/ A computer program, BILATERAL, with the capability of dealing with all types of unilaterally/bilaterally and censored/truncated normal regression models has been written by Hartley with assistance from Swanson. It provides consistent initial IV estimators for all of the above types of models, calculates the MLE estimator via any of the G-N, N-R, MS and E-M algorithms, and produces the final log-likelihood score as well as the asymptotic covariance matrix of the MLE for purposes of performing the customary (asymptotic) t-tests and likelihood ratio tests of hypotheses. A copy of this FORTRAN program can be obtained from the authors at cost upon request.
While it is obvious that the latent skill-levels have mean and variance given by:

\[(2.21) \quad \mu_j = \mu_j(\xi_i) = \xi_i \beta \]

and

\[(2.22) \quad \text{Var}[y_{ijk}^*] = \sigma^2, \]

respectively, the effect of the censoring of \(y_{ijk}^*\) at the two extremes, 0 and \(M\), changes both of these measures. In particular, it can easily be shown that:\(^1\)

\[(2.23) \quad \mu_{ij} = \mu_{ij}(\xi_i) = \mu_j(\xi_i) \cdot [F_i(M) - F_i(0)] + M \cdot [1 - F_i(M)] - \sigma^2 \cdot [\xi_i - \mu_j(\xi_i)] \]

and

\[(2.24) \quad \text{Var}[y_{ij}] = -\sigma^2 \cdot \{M - \mu_j(\xi_i)\} \cdot [F_i(M) - F_i(0)] \]

\[-2\mu_j(\xi_i) \cdot [u_j(\xi_i)] \cdot [F_i(M) - F_i(0)] \]

\[-2 \cdot \{u_j(\xi_i) \cdot f_i(M) - f_i(0)\}] + [\sigma^2 \cdot (\mu_j(\xi_i) - \xi_i)] \cdot [F_i(M) - F_i(0)]

\[-(\mu_j(\xi_i) \cdot [F_i(M) - F_i(0)]) + M \cdot [1 - F_i(M)]

\[-\sigma^2 \cdot (\xi_i - \mu_j(\xi_i))^2 \]

Note that, in general,

\[(2.25) \quad \text{E}[y_{ij}] = \mu_j(\xi_i) \]

and

\[(2.26) \quad \text{Var}[y_{ij}] = \sigma^2 \]

\(^1\) We shall employ the notation,

\[F_i(x) = F(c; \xi_i) = \int^c_{-c} f_i(y_{ijk}^*) \cdot dy_{ijk}^* \]

to denote the distribution function for \(y_{ijk}^*\), a function of \(\xi_i\).
whenever the lower limit, \( L \) (here 0) and/or the upper limit, \( M \), are finite. In particular, it may easily be verified from (2.23) and (2.24) that

\[
\lim_{L \to -\infty} \frac{\mu_{jk}(x_i)}{M \to +\infty} = \mu_{jk}(x_i) = \frac{x_i^3}{3}
\]

which illustrates the fact that use of the sample mean of test scores (or OLS regression) to estimate the mean (or true regression function) for skill-levels will result in biased and inconsistent estimates. Further, the use of the sample variance (or average residual sum of squares) of test scores will underestimate the true variance in skill-levels, \( \sigma^2 \).

We have portrayed these points in Figure 2.1, where the special case of a single (continuous) regressor, \( x_i \), is employed for simplicity. Thus, we have

\[
\mu_{jk}(x_i) = \beta_0 + \beta_1 x_i
\]

Here it may be seen that at the value, \( x_i = x_i^0 \), the mean skill-level is given by \( \mu_{jk}(x_i^0) = \beta_0 + \beta_1 x_i^0 \). The effect of censoring at 0 and \( M \), respectively, is to shift the mean score, \( \xi_{jk}(x_i^0) \), to the right, since \( (x_i = x_i^0) \) the proportion of left-censored observations, \( F_i(0) \), exceeds the proportion of right-censored observations, \( [1 - F_i(M)] \) - each denoted by the shaded areas under the density \( f_i(x_i^0) \), and accounted for as a fixed probability mass at the two extremes in the density of test scores, \( g_i(y_i; x_i^0) \). In the upper diagram we have also graphed the two expectations, \( E(y_i^1|x_i) \) of (2.21) and \( E(y_i|x_i) \) of (2.23), under the simple model, (2.29). The first of these is denoted by the solid line (—) and the latter, an inverted S-shape, by the dotted line (....), respectively. It will be noted that, since \( \beta \) is positive, \( \xi_{jk}(x_i) = E(y_i|x_i) \) is

\[1^1 A formal proof is given in Hartley and Swanson (1980).]
Figure 3.1: The Effects of Bilateral Censoring on Test Scores

The Distribution and Expectation of Test Scores

\[ E[y_i | x_i] = \tilde{y}_i(x_i) \]

\[ E[\tilde{y}_i | x_i] = \tilde{y}_i(x_i) \]
asymptotic to the line, \( y_1 = 0 \), as \( x_1 \to -\infty \). Thus, as a result of the censoring of scores, the observed pairs, \((y_1, x_1)\), in the sample would be expected to be scattered about the line, \( \xi_{jk}(x_1) \), rather than the linear regression function for skill-levels, \( \mu_{jk}(x_1) \) of (2.29). The estimation problem is therefore: How do we estimate \( \beta_0, \beta_1 \) and \( \sigma^2 \) from a set of observations around \( \xi_{jk}(x_1) \)? The solution resides in calculation of the MLE using the density, (2.20).

Let \( I^{(1)}_\tau \) denote the subset of observations on individuals with "zero" scores on a particular test, let \( I^{(2)}_\tau \) denote those individuals with scores in the open interval, \((0, M)\), and let \( I^{(3)}_\tau \) denote those with perfect scores. Then the log-likelihood function, based on (2.20), may be written as:

\[
(2.30) \quad \log L(\beta, \sigma^2) = \sum_{i \in I^{(1)}_\tau} \log F_i(0) + \sum_{i \in I^{(2)}_\tau} \log \xi_i(y_{i1k}) + \sum_{i \in I^{(3)}_\tau} \log (1 - F_i(M)),
\]

which is to be maximized with respect to the \( K \)-vector, \( \beta \), and the scalar, \( \sigma^2 \).

As noted earlier, Hartley and Swanson (1980) have provided a variety of procedures by which this may be done. Application to the censored ANOVA models for the Egyptian data set on in-schoolers and dropouts will be discussed in section 4.

---

1/ It may easily be verified that at the value \( x_1 = x_1^* \), where \( x_1^* \) is the \( \xi_{jk}^{-1}(0) \) (i.e., halfway between the zero and perfect scores), the conditional expectation \( \xi_{jk}(x_1^*) \) is identical to the expectation, \( \mu_{jk}(x_1^*) \), since at the value of \( x_1 \) "symmetric censoring" occurs.
3. The Cross Section Sample and Test Instruments:

In 1978 there were approximately 4,075,000 students attending grades 1 through 6 in 10,012 public elementary schools in Egypt. Attendance through the end of grade 6 is mandatory, but unenforced; there are proposals to increase the mandatory period to grade 8 and a significant number of children either never enroll in school or drop out before completing the sixth grade. Data on students who never enter the school system are not easily obtained. For students who do enroll, dropout rates are known only approximately. Aggregate data from the Ministry of Education (MOE) indicate that 23% of all males and 33% of all females who began elementary school in the year 1973-74 left permanently before completing the sixth grade. (MOE, 1974)

Our sample universe consists of two groups: (1) in-school students, i.e., those who were attending grades three through six in the 1978/79 school year and (2) dropouts, defined to be students who last attended one of the grades three through six, but dropped out of school sometime after September, 1974 and before November, 1978. Students who transferred to other schools and students who died were not considered to be "dropouts".

1/ We are in the process of collecting detailed enrollment and school-leaving data at each of our sample schools.

2/ These figures may somewhat overstate the dropout problem to the extent that a number of students attend school through the end of the sixth grade, but do not sit for the final (school leaving) exam. These students are treated as dropouts in the MOE data, although they presumably have received the same curriculum as their companions who have taken the final exam. Their decision not to take the exam may, however, reflect a self-evaluation of their own abilities. Understatement may arise as a result of a "paper promotions", whereby students who have in fact dropped out are carried forward on the school enrollment lists from one year to the next. Discussions with MOE/NCEER officials indicate that the latter effect is likely to dominate.
The sample design employed a two-stage sampling procedure. In the first stage thirty urban and thirty rural schools were selected from among the 10,012 schools in the population. The urban/rural classification was based on the official MOE definition. It distinguishes schools in the major cities of Cairo and Alexandria and the capitals of governorates from smaller cities, towns, and rural villages. In the second stage, in-school students were sampled at random. For urban schools the sampling rate was 50%, and for rural students it was 66-2/3%. Dropouts at each of the sample schools were identified from the lists of non-attending students, prepared by the schools for the MOE on the 15th of November of each year. Any student enrolled in the previous year and found on the list of non-attendants in the next year was considered to be a dropout of the previous year. All dropouts who were so identified, who were locatable and amenable to being tested were included in the sample. Table 3.1 gives a breakdown of the in-school sample by grade, sex, and community type. Table 3.2 gives similar data for the dropouts. The present study is based upon the complete dropout sample and a subsample of one third of the in-schoolers for whom, along with the dropouts, a supplementary questionnaire concerning family background and socio-economic characteristics was completed.

from other activities in sitting for tests. This may have led to a biased sample in that the "better" students may be presumed to be more likely to be employed and, if employed, command a higher wage. Such students are, therefore, less likely to find the honorarium attractive. Unfortunately, since such "absentees" were never tested, there is no obvious way to test for or correct for such a bias.

A sequel to the present paper will investigate the effect of these socioeconomic characteristics on literacy/numeracy retention curves, utilizing a bilaterally censored Analysis of Covariance (ANCOVA) model.
Table 3.1: IN-SCHOOL SAMPLE BY GRADE, SEX AND COMMUNITY TYPE

<table>
<thead>
<tr>
<th>Grade</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>531</td>
<td>506</td>
<td>1037</td>
<td>711</td>
<td>372</td>
<td>1083</td>
<td>1242</td>
<td>878</td>
<td>2120</td>
</tr>
<tr>
<td>4</td>
<td>644</td>
<td>521</td>
<td>1205</td>
<td>721</td>
<td>322</td>
<td>1043</td>
<td>1365</td>
<td>883</td>
<td>2248</td>
</tr>
<tr>
<td>5</td>
<td>507</td>
<td>483</td>
<td>990</td>
<td>680</td>
<td>266</td>
<td>946</td>
<td>1187</td>
<td>749</td>
<td>1936</td>
</tr>
<tr>
<td>6</td>
<td>430</td>
<td>424</td>
<td>854</td>
<td>479</td>
<td>239</td>
<td>718</td>
<td>909</td>
<td>663</td>
<td>1572</td>
</tr>
<tr>
<td>Total</td>
<td>2112</td>
<td>1974</td>
<td>4086</td>
<td>2591</td>
<td>1199</td>
<td>3790</td>
<td>4703</td>
<td>3173</td>
<td>7876</td>
</tr>
</tbody>
</table>

1/ A subsample of 1/3 of the in-schoolers was used in the subsequent analysis.
<table>
<thead>
<tr>
<th>Grade</th>
<th>Urban Male</th>
<th>Urban Female</th>
<th>Urban Total</th>
<th>Rural Male</th>
<th>Rural Female</th>
<th>Rural Total</th>
<th>Total Male</th>
<th>Total Female</th>
<th>Total Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>51</td>
<td>49</td>
<td>100</td>
<td>81</td>
<td>110</td>
<td>191</td>
<td>132</td>
<td>159</td>
<td>291</td>
</tr>
<tr>
<td>4</td>
<td>113</td>
<td>98</td>
<td>211</td>
<td>168</td>
<td>150</td>
<td>218</td>
<td>281</td>
<td>248</td>
<td>529</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>46</td>
<td>102</td>
<td>104</td>
<td>63</td>
<td>167</td>
<td>160</td>
<td>109</td>
<td>269</td>
</tr>
<tr>
<td>6</td>
<td>171</td>
<td>158</td>
<td>329</td>
<td>252</td>
<td>138</td>
<td>390</td>
<td>423</td>
<td>296</td>
<td>719</td>
</tr>
<tr>
<td>Total</td>
<td>391</td>
<td>351</td>
<td>742</td>
<td>605</td>
<td>461</td>
<td>1066</td>
<td>996</td>
<td>812</td>
<td>1808</td>
</tr>
</tbody>
</table>

1/ 19 students lacking complete test records were dropped from the subsequent analysis.
Literacy and numeracy skills were measured using a battery of nine tests which were developed and piloted in an earlier phase of this study. The literacy portion consisted of four exams. Two of these, Reading A and Reading 3, pertain to reading skills, and consist of multiple choice items. The remaining two, Writing A and Writing 3, require the child to write words, sentences, and, finally, an entire paragraph. The children's answers were graded on a varying scale. From 0 to 2 points were given for questions involving only a single word. For a large paragraph of 31 words the children could receive up to 31 points. The A exams were targeted at third and fourth grade students and the B exams at a fifth and sixth grade level of difficulty.

The numeracy portion consisted of three tests. On each test the students were given one point for each problem correctly solved. The Simple Operations test measures the ability to handle those elementary arithmetic operations (addition, subtraction, multiplication and division) which a child could be expected to acquire early in his school career. The Problem Solving test presents simple problems of the sort found in school arithmetic books, and covers subject matter appropriate to the fourth through sixth grade levels. The Geometry test was targeted at fifth and sixth graders, with problems involving the identification and manipulation of elementary geometric figures.

The two intelligence tests were intended to capture non-curriculum-dependent measures of the children's skills. The Verbal Intelligence test is based on a comprehensive intelligence test developed at Ain Shams University, Cairo, and consists of 10 items. The Non-Verbal Intelligence is composed of 15 multiple choice items - each consisting of five figures, where the testee
is asked to eliminate the one which does not belong.

Table 3.3 shows the number of students included in the present study who took each exam. Note that in-school students were given only the exams "appropriate" to their grade level, while dropouts received all nine exams.

Table 3.3 also shows the general pattern of censoring across tests and between the in-school and dropout samples. On the harder tests, the literacy 'B' tests, and the problem solving and geometry tests for example, the proportion of left-tailed censoring increases. On every test the number of dropouts receiving a zero score is greater than the number of in-schoolers with zero scores. A closer examination of the occurrence of censoring in each grade-year suggested a complementary pattern. On a given test the proportion of left (right) censored observations: (1) increases (decreases) with the number of years-out-of-school, holding grade-level-attended constant; and (2) decreases (increases) with the grade-level, holding the number-of-years-from-school constant. A systematic pattern of censoring must lead to systematic bias in the estimation of the learning and retention curves when the OLS estimator is employed. In the next section we shall examine the extent of this bias.
Table 3.3: NUMBERS TESTED AND EXTREME SCORES FOR NINE LITERACY/NUMERACY TESTS

<table>
<thead>
<tr>
<th>TEST</th>
<th>Maximum Score</th>
<th>Number of Testees</th>
<th>Zero Scores</th>
<th>Perfect Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dropouts</td>
<td>In-schoolers</td>
<td>Dropouts</td>
</tr>
<tr>
<td>Reading A</td>
<td>1/</td>
<td>10</td>
<td>1789</td>
<td>1450</td>
</tr>
<tr>
<td>2/ Reading B</td>
<td>10</td>
<td>1789</td>
<td>1156</td>
<td></td>
</tr>
<tr>
<td>Writing A</td>
<td>1/</td>
<td>10</td>
<td>1789</td>
<td>1450</td>
</tr>
<tr>
<td>2/ Writing B</td>
<td>10</td>
<td>1789</td>
<td>1156</td>
<td></td>
</tr>
<tr>
<td>Simple Operations</td>
<td>2/</td>
<td>28</td>
<td>1789</td>
<td>2606</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>2/</td>
<td>4</td>
<td>1789</td>
<td>1895</td>
</tr>
<tr>
<td>Geometry</td>
<td>2/</td>
<td>8</td>
<td>1789</td>
<td>1156</td>
</tr>
<tr>
<td>Verbal IA</td>
<td>0</td>
<td>1789</td>
<td>2606</td>
<td></td>
</tr>
<tr>
<td>Non-Verbal IA</td>
<td>5</td>
<td>1789</td>
<td>2606</td>
<td></td>
</tr>
</tbody>
</table>

1/ Given only to 3rd and 4th grade in-schoolers.

2/ Given only to 5th and 6th grade in-schoolers.

3/ Given only to 4th, 5th, and 6th grade in-schoolers.
4. Empirical Results:

In this section, as an illustration of the difference between the OLS and ML estimators of the learning and retention curves, we present results from applying both procedures to models (2.8) and (2.9). Since the bias in the OLS parameter estimates is considerable, we drop subsequent use of this method. We then attempt to measure the effects of differences in schools on students' latent skill-levels using consistent and asymptotically efficient estimates, obtained from a single N-R iteration from the consistent initial estimates. The coefficients of model (2.10) are only estimated for the three tests taken by all students: Simple Operations, Verbal Intelligence, and Non-Verbal Intelligence. Finally, we examine the portion of the retention curves based only on data for dropout cells \( k > 0 \).

4.1 Estimation of Cell Means:

Imposition of model (2.8), but using OLS to estimate the parameters of (2.6), is equivalent to the estimation of the cell means as simple averages of the measured test scores. This model excludes the partial effects of differences in sex and community or school, but is of interest to the extent that the sample means and variances are frequently used as summary statistics for test data.

Figure 4.1 presents the estimated learning and retention curves under (2.8) for the Verbal Intelligence test only. Because all students took this test, each grade-specific curve consists of five points: the first corresponding to the in-school cell associated with each grade level, and the next four to students one, two, three, and four years out of the same grade-level. The implied "learning curve" may be obtained from the
four in-school cells.

We have argued above that censoring of the latent skill-levels causes the mean of the measured test-scores to overestimate the mean of the underlying latent distribution when left-tailed censoring prevails, and to underestimate it when right-tailed censoring is dominant. The result is to "flatten" and "compress" the retention curves - as if squeezed by a "nut cracker" tool. Differences between cells of the same grade level, as well as differences between grade-levels, are erroneously reduced. In our sample, censoring occurs predominantly at zero scores: certain tests are apparently too hard for most dropouts, especially those in the lower grades. Consequently, as seen in Figure 4.1, the greatest bias occurs in the estimated mean skill levels for the lowest grades. Using the sample means of the test score, the third grade retention curve lies everywhere below but "close" to the fourth grade curve. Using the ML estimates, the difference between each curve increases considerably. The "flattening" of the OLS curves in Figure 4.1 is also apparent, although the dramatic drop in skill-levels from in-schoolers to dropouts is preserved.

4.2 Learning and Retention Curves Controlling for Sex and Community Effects:

In addition to data on individuals' cells, data on sex and location (urban/rural) are also available. For a host of cultural and attitudinal reasons, we would expect to find differences between the score of males and females and between the scores of city and country dwellers. The four-way ANOVA model, (2.9), permits the explicit measurements of such effects. The resulting learning and retention curves are exhibited in Figures 4.2 - 4.10.1/ 

1/ Since the linear regression now includes variables which measure the marginal effects of differences in sex and urban/rural location on all grade-year cells, the model is no longer "fully saturated" and the cell coefficients should not be interpreted as cell means, but rather as the partial effects of cell differences after accounting for sex and location. Since the value of the sex-variable is zero for females and the location variable is zero
PROBLEM SOLVING

<table>
<thead>
<tr>
<th>SEX</th>
<th>CTYPE</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;</td>
<td>ML</td>
<td>0.7693* 2.3992* 11.0012</td>
</tr>
<tr>
<td>( )</td>
<td>OLS</td>
<td>0.5869 1.7577 7.2763</td>
</tr>
</tbody>
</table>

Figure 4.7
<table>
<thead>
<tr>
<th>VERBAL INTELLIGENCE</th>
<th>SEX</th>
<th>CTYPE</th>
<th>VAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) - MLE</td>
<td>0.754</td>
<td>3.426</td>
<td>26.561</td>
</tr>
<tr>
<td>(b) - OLS</td>
<td>1.645</td>
<td>3.109</td>
<td>22.777</td>
</tr>
</tbody>
</table>

Figure 4.9
Because in-schoolers were not given all tests, certain in-school cell coefficients were inestimable. Estimable in-school cells are identified by \((j,0)\), \(j = 3, 4, 5, 6\), and, if connected, would represent learning curves. Each retention curve extends through the \((j,4)\) cell, corresponding to dropouts four years out of school. In these graphs, the initial points of each retention curve have been "staggered," so that cell-coefficients, \((j,k)\), have a constant sum, \(j+k\), for every vertical line, denoting (approximately) a common age-cohort group.\(^2\)

There are several common features of the plotted sets of curves. First, as in the case of model \((2.8)\), the OLS-estimated curves always lie on or above the ML estimates. Again, this reflects the predominance of left tailed censoring shown in Table 3.3. Second, the distance between the OLS and ML estimates generally "narrow" at higher grade-levels. This is due to the reduction in left-tailed censoring: the tests are, in effect, becoming better centered at higher grade levels. Third, the ordering of the curves is preserved under both OLS and ML, i.e., the correction of the "OLS bias" never causes a retention curve to cross its counterpart.

Contrary to the hypothesis of skill-attenuation or increasing "wastage," the plotted curves are neither "smooth" nor consistently negatively sloped. In fact, there appears to be an overall upward trend for third grade dropouts in all skills and for fourth grade dropouts in most after the first year out. Sixth graders, who begin at a much higher initial skill level, are the only group that shows a steady 'decline in skill-levels.

\(^1\) (continued)

\(^2\)

for rural schools, the estimated cell regression coefficients may be interpreted as a prediction of the mean skill-levels for rural females. The curves for males and urban dwellers are found by adding the appropriate coefficient to the cell coefficients. This amounts to a parallel vertical shift of each curve.

\(^2\) Variations in age across entrants to the first grade, as well as possible repetition of grades 2 and 4, disturb, somewhat, this interpretation.
The large differences between current in-schoolers and dropouts from the same grade after one year suggest that, if there is a legitimate loss of skills among dropouts, it occurs within one year. However, this conclusion must be treated with caution. Since we have restricted ourselves here to analysis of a cross-section sample, we have no information about the dropout decision-process and, with respect to the sample of in-schoolers, have no knowledge, at present, as to which of them will choose to continue in school as opposed to those who will, in fact, choose to drop out in the forthcoming year. It will be recalled from subsection 2.1 that a dropout's retention curve begins with the skill-level associated with the last year he/she was in school. Thus, skill-levels for in-schoolers who decide to continue as such should be excluded from the sample used to estimate a dropout retention curve. Without knowledge of the separation of the current set of in-schoolers into "would-be continuers" - which should be utilized in the sub-sample for estimation of the learning curve - and "would-be dropouts" - whose in-school scores would be utilized to estimate the base-year reference point for a dropout retention curve, the sharp drop in our estimated skill-levels during the first year of dropping out cannot be unequivocally ascribed to a lack of retention.

Further, to the extent that the level of a student's acquired skills while in school, at least in part, influences his/her decision to drop out, one would expect to find that the "poorer" students are more likely to be next year's dropouts, while the "better" students are more likely to continue in school. If this is correct — there is no way to verify the matter from a single cross-section, then the mean skill-level for the "would-be dropout" sub-sample of current in-schoolers — and, hence, the base-year reference
point for each grade's retention curve — is likely to be lower than the overall in-schoolers cell mean. Similarly, the mean skill-level of "would-be-continuers" would be expected to exceed the overall in-school cell mean. This issue has been termed the "self-selection problem", and the consequence is to bias all of the parameter estimates. Thus, the apparently sharp drop in estimated skill-levels during the first year of dropping out may be largely attributable to a self-selection bias in the retention curve estimates. Measurement of the true retention curves in the presence of self-selection bias requires longitudinal data on the historical sequence of decisions by students in order to make the dropout-decision, itself, endogenous to a skill-level model. We are currently collecting such data for in-schoolers and dropouts in our present sample. Meanwhile, we must regard both the estimated learning and retention curves as provisional, pending subsequent testing for self-selection issues.

Differences in skill levels attributable to sex are generally small, though significant.\(^1\) They are always positive, indicating that males possess slightly higher skill-levels across all grades. Differences attributable to urban/rural location, in contrast, are large and always favor urban students. This suggests that urban students may enjoy considerable advantages over rural students — both in terms of the educational resources available to them and the stimuli supplied by their environment to use and retain their skills.

4.3 Estimation of Individual School Effects:

Schools, even in a centrally-planned system (as in the ARE), do not supply a homogeneous mix of educational inputs. Ideally, we would like to have measurements of the quantity and quality of resources available at each

---

\(^1\) Significance at the 95% level is indicated by an asterisk (*) next to the ML coefficient.
school. Inclusion of such variables in the regression function would permit separation of the effect of school inputs from those due to school location. Lacking such information at present, we must, instead, posit a model of the form of (2.10), in which school-effects are captured by the use of "dummy variables". Since the identity of the school also determines its location, we must drop the location variable used in model (2.9). Furthermore, because there is a unique assignment of students to schools, we must exclude either one school dummy or one grade-year cell dummy from the regression to avoid singularities in estimation. We chose to exclude school number 60. The coefficients on the remaining school variables are interpreted as the marginal effect on skill-levels of being assigned to a particular school relative to school 60.

Table 4.1 shows the coefficients for each grade-year cell and the sex variable, estimated when model (2.10) is applied to the Simple Operations, Verbal Intelligence, and Non-Verbal Intelligence tests. ¹/ In general, the pattern of coefficient values for the grade-year cells suggests that the retention curves, net of school-effects, have very much the same shape as those derived under model (2.9). The third grade curves, after an initial drop, are upward sloping. The sixth grade curves are downward sloping or flat. All continue to show the characteristic fluctuations from cell to cell along a grade-specific curve, which were observed in the earlier specifications. ²/

¹/ The computational costs of estimating the 80 coefficients in the school effects model are many times higher than for the comparatively limited specifications of models (2.8) and (2.9). We have therefore restricted ourselves to just those tests which all students received. To minimize cost, we have employed the one-iteration N-R estimates from the consistent initial estimates, which then have the same asymptotic properties as those of the (convergent) M.L. estimates. [See Amemiya (1973) and Hartley and Swanson (1980)].

²/ It is also worth noting that for the school-effect contrasts (portrayed in Table 4.2 with the exception of only 2 schools (#51 and #52) the pattern of measured effects associated with each school was confirmed by the appropriate NCER staff in school visits at the time of test administration.
Table 4.1: CELL-SPECIFIC EFFECTS FROM MODFL (2.10)

<table>
<thead>
<tr>
<th>Grade - Year Cell</th>
<th>Simple Operations</th>
<th>Verbal Intelligence</th>
<th>Non-Verbal Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>6.9290*</td>
<td>3.1574*</td>
<td>11.7389*</td>
</tr>
<tr>
<td>31</td>
<td>-4.1784*</td>
<td>-2.4830*</td>
<td>3.8824*</td>
</tr>
<tr>
<td>32</td>
<td>-2.9451*</td>
<td>-3.6425*</td>
<td>4.4700*</td>
</tr>
<tr>
<td>33</td>
<td>-3.6146*</td>
<td>-3.5826*</td>
<td>3.0853*</td>
</tr>
<tr>
<td>34</td>
<td>-1.4747</td>
<td>-1.2742</td>
<td>8.4698*</td>
</tr>
<tr>
<td>40</td>
<td>11.0438*</td>
<td>5.7023*</td>
<td>14.9669*</td>
</tr>
<tr>
<td>41</td>
<td>0.1232</td>
<td>-0.3772</td>
<td>8.9856*</td>
</tr>
<tr>
<td>42</td>
<td>1.5967*</td>
<td>0.8485</td>
<td>10.8540*</td>
</tr>
<tr>
<td>43</td>
<td>-0.3164</td>
<td>-0.6102</td>
<td>9.8360*</td>
</tr>
<tr>
<td>44</td>
<td>0.6743</td>
<td>1.3715*</td>
<td>12.6884*</td>
</tr>
<tr>
<td>50</td>
<td>15.9301*</td>
<td>8.5316*</td>
<td>17.4329*</td>
</tr>
<tr>
<td>51</td>
<td>4.1543*</td>
<td>2.9052*</td>
<td>10.9345*</td>
</tr>
<tr>
<td>52</td>
<td>4.6751*</td>
<td>3.6771*</td>
<td>13.2016*</td>
</tr>
<tr>
<td>53</td>
<td>3.5639*</td>
<td>3.4050*</td>
<td>14.0183*</td>
</tr>
<tr>
<td>54</td>
<td>3.3615*</td>
<td>2.9665*</td>
<td>14.8140*</td>
</tr>
<tr>
<td>60</td>
<td>19.8449*</td>
<td>11.4188*</td>
<td>19.3630*</td>
</tr>
<tr>
<td>61</td>
<td>9.2920*</td>
<td>6.8746*</td>
<td>16.2222*</td>
</tr>
<tr>
<td>62</td>
<td>8.7368*</td>
<td>7.2636*</td>
<td>16.6779*</td>
</tr>
<tr>
<td>63</td>
<td>6.6362*</td>
<td>3.8837*</td>
<td>13.3333*</td>
</tr>
<tr>
<td>64</td>
<td>6.6883*</td>
<td>6.4854*</td>
<td>16.4232*</td>
</tr>
</tbody>
</table>

Sex:

Variance:

* denotes a coefficient significantly different from zero (via the asymptotic t-test)
<table>
<thead>
<tr>
<th>Urban Schools:</th>
<th>Simple Operation</th>
<th>Verbal Intelligence</th>
<th>Non-Verbal Intelligence</th>
<th>Rural Schools:</th>
<th>Simple Operations</th>
<th>Verbal Intelligence</th>
<th>Non-Verbal Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tarek Ben Zeid</td>
<td>3.5828</td>
<td>5.4293*</td>
<td>4.5251*</td>
<td>31 Kafr Sendahoun</td>
<td>-0.3142</td>
<td>0.4753</td>
<td>2.6493*</td>
</tr>
<tr>
<td>Al-Suez City</td>
<td>0.8715</td>
<td>0.2108</td>
<td>-0.1879</td>
<td>32 Al-Akbar</td>
<td>5.6010*</td>
<td>0.6196*</td>
<td>2.6511</td>
</tr>
<tr>
<td>Al-Fatah</td>
<td>4.6891*</td>
<td>3.2351*</td>
<td>4.4695*</td>
<td>33 Al Bealy</td>
<td>-1.4991*</td>
<td>8.1426*</td>
<td>4.0942*</td>
</tr>
<tr>
<td>Hussein Fahmi</td>
<td>3.4772</td>
<td>5.3657*</td>
<td>2.7904*</td>
<td>34 Mouganla Beboat</td>
<td>-0.2580</td>
<td>-2.2963*</td>
<td>-3.4729*</td>
</tr>
<tr>
<td>Ahmed Maher</td>
<td>7.3374</td>
<td>7.8660*</td>
<td>6.2695*</td>
<td>35 Abou Taha</td>
<td>0.3607</td>
<td>0.0987</td>
<td>-3.0783*</td>
</tr>
<tr>
<td>Al-Naas</td>
<td>4.4036</td>
<td>3.5079*</td>
<td>1.0194*</td>
<td>36 Moustafo Abed El Nabi</td>
<td>1.7182</td>
<td>0.1265</td>
<td>-0.2637</td>
</tr>
<tr>
<td>Al-Shourafa</td>
<td>5.0098</td>
<td>5.7556*</td>
<td>2.9887*</td>
<td>37 Naser Kafr Naera</td>
<td>3.4401*</td>
<td>2.8161*</td>
<td>-1.8844*</td>
</tr>
<tr>
<td>Khaled Ben Walid</td>
<td>6.3343</td>
<td>4.6586*</td>
<td>3.1275*</td>
<td>38 Egbezi Al-Baz</td>
<td>0.7996</td>
<td>0.7326</td>
<td>-1.4257</td>
</tr>
<tr>
<td>Al-Shourafa</td>
<td>3.2351</td>
<td>4.0691*</td>
<td>4.2210*</td>
<td>39 Naser Al-Areed</td>
<td>-0.9415</td>
<td>0.3727</td>
<td>0.2005</td>
</tr>
<tr>
<td>Al-Naas</td>
<td>3.5748</td>
<td>3.6323*</td>
<td>4.3312*</td>
<td>40 NaBarret Faraskour</td>
<td>6.4448*</td>
<td>3.7912*</td>
<td>0.4334</td>
</tr>
<tr>
<td>Al-Nouria</td>
<td>3.3295</td>
<td>3.0114*</td>
<td>4.2607*</td>
<td>42 Sixthente of March</td>
<td>3.1346*</td>
<td>2.593*</td>
<td>0.1123</td>
</tr>
<tr>
<td>Al-Naas</td>
<td>4.0365</td>
<td>4.7425*</td>
<td>3.1053*</td>
<td>43 Al-Saheed Moustafo Kafr Hefna</td>
<td>1.4472</td>
<td>0.7136</td>
<td>0.3682</td>
</tr>
<tr>
<td>Amin Al-Rufee</td>
<td>1.1845</td>
<td>1.550*</td>
<td>3.2547*</td>
<td>44 Al-Rouda</td>
<td>-2.0748*</td>
<td>-3.5942*</td>
<td>-1.4477*</td>
</tr>
<tr>
<td>Dar Al Safa</td>
<td>5.1862</td>
<td>5.9043*</td>
<td>4.9471*</td>
<td>45 Abu Ayaad</td>
<td>2.2938*</td>
<td>2.1660*</td>
<td>2.1234*</td>
</tr>
<tr>
<td>Om Al-Moemenen</td>
<td>2.5274</td>
<td>2.2687*</td>
<td>5.3343*</td>
<td>46 Al Shaheed Gwad Hosn</td>
<td>3.3871*</td>
<td>3.5466*</td>
<td>3.256*</td>
</tr>
<tr>
<td>Al-Shaheed Zakaria Attia</td>
<td>-0.3925</td>
<td>-1.5582*</td>
<td>-2.9959*</td>
<td>47 Kafr Kandil</td>
<td>1.9673*</td>
<td>0.3466</td>
<td>1.7703</td>
</tr>
<tr>
<td>Sabah Sulem</td>
<td>2.4320</td>
<td>1.5013*</td>
<td>6.6511*</td>
<td>48 Nasser Mater Tares</td>
<td>-2.3919*</td>
<td>-1.9669*</td>
<td>1.7538</td>
</tr>
<tr>
<td>Al-Ahmadiah</td>
<td>6.8468</td>
<td>8.1154*</td>
<td>-1.2895</td>
<td>49 Al-Hassan</td>
<td>-1.5203*</td>
<td>2.0865*</td>
<td>-1.6433*</td>
</tr>
<tr>
<td>Faheem Girls School</td>
<td>2.1280</td>
<td>1.9268*</td>
<td>-0.3640</td>
<td>50 Henshat Al Dahab Al-Kebliah</td>
<td>-3.4269*</td>
<td>-0.3727</td>
<td>2.8702*</td>
</tr>
<tr>
<td>Kafar Shamak</td>
<td>3.2427</td>
<td>2.1360*</td>
<td>1.4466*</td>
<td>51 Tendah</td>
<td>-2.1319*</td>
<td>-0.3312</td>
<td>-2.634*</td>
</tr>
<tr>
<td>Meet Habeeb</td>
<td>4.0505</td>
<td>2.6118*</td>
<td>0.3449*</td>
<td>52 Der Al-Carnouze</td>
<td>-0.7441</td>
<td>0.5783</td>
<td>0.0273</td>
</tr>
<tr>
<td>Teda</td>
<td>0.6357</td>
<td>2.1895*</td>
<td>-0.1630</td>
<td>53 Haasarat Haggag Bany Hazer</td>
<td>-3.7804*</td>
<td>-3.2545*</td>
<td>-2.2220*</td>
</tr>
<tr>
<td>Al-Hedalh</td>
<td>3.0599</td>
<td>3.0565*</td>
<td>1.3985*</td>
<td>54 Koom Matay</td>
<td>-0.3519*</td>
<td>0.3917</td>
<td>1.3255</td>
</tr>
<tr>
<td>Al-Gomhoriah</td>
<td>1.0739</td>
<td>0.9052</td>
<td>2.1113*</td>
<td>55 Al-Satamoun Tema</td>
<td>-1.8366*</td>
<td>-0.2349</td>
<td>-1.4223*</td>
</tr>
<tr>
<td>Kafar Al Sheekh Shehaca</td>
<td>-2.0099</td>
<td>-1.6814*</td>
<td>3.5586*</td>
<td>56 Bany Massel</td>
<td>-2.6199*</td>
<td>0.2052</td>
<td>-0.9294</td>
</tr>
<tr>
<td>Moaassat Meet Bera</td>
<td>0.0322</td>
<td>-2.6507*</td>
<td>-1.6738*</td>
<td>57 Nagg Gd Al-Karim</td>
<td>-1.7411</td>
<td>-0.2914</td>
<td>-0.9032</td>
</tr>
<tr>
<td>Begerem Girls School</td>
<td>1.0920</td>
<td>0.3653</td>
<td>-0.4321</td>
<td>58 Al Dabadla</td>
<td>-5.8670*</td>
<td>-1.5146*</td>
<td>-6.9670*</td>
</tr>
<tr>
<td>New Shebeen Al-Kanater</td>
<td>2.0790</td>
<td>1.9800*</td>
<td>2.8425*</td>
<td>59 Nagg Al-Gameh</td>
<td>-6.9083*</td>
<td>-2.6374*</td>
<td>-2.0388*</td>
</tr>
<tr>
<td>Re: Al-Houria</td>
<td>4.2304</td>
<td>2.6708*</td>
<td>6.0088*</td>
<td>60 Taha Hussein</td>
<td>xxx</td>
<td>xxx</td>
<td>xxx</td>
</tr>
</tbody>
</table>

XXX Excluded from regression.

* denotes a coefficient significantly different from zero (via the asymptotic t-test).
4.4 Estimation of the Pure Dropout Retention Curve

In sub-section 4.2 we pointed out that a possible interdependence between skill-levels and the dropout decision may lead to a self-selection bias in the estimated base levels of the retention curves. In effect, the in-school cells contain non-homogeneous individuals: those who will dropout in the next year and those who will remain in school. Assuming that continuing students are, on the average, more skilled than would-be dropouts, we argued that the estimates of the base year coefficients are too high and exaggerate the loss in skills suffered by a first year dropout. Imposition of model (2.9) in the pooled sample of in-schoolers and dropouts also forces the coefficients on the sex and community-type (urban/rural location) variables and the estimated variance parameter to be identical for in-schoolers and dropouts.

If the advantage to a dropout, conferred by residing in an urban community, is primarily the stimulus to continue using the skills which he/she receives from his environment, and if the advantage of having once been enrolled in an urban, rather than a rural school, is of relatively little importance, then, if we estimate the "pure dropout retention curve", i.e. using only the dropout sample with associated sex and community-type variables, we should expect an increase in the coefficient for community-type. Table 4.3 summarizes the results of the "pure dropout retention curve" estimates.

As may be seen by examining Figures 4.11 through 4.19, the skill and grade-specific patterns, which were observed earlier, remain - apart from a translation of the vertical axis. There is still evidence of the attenuation of skills among sixth graders - particularly literacy skills - and small increases in skill levels among third and fourth graders. The
<table>
<thead>
<tr>
<th>Grade - Year Cell:</th>
<th>Reading Tests</th>
<th>Writing Tests</th>
<th>Simple Operations</th>
<th>Problem Solving</th>
<th>Geometry</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,1</td>
<td>.0410</td>
<td>-2.6982*</td>
<td>-41.8260*</td>
<td>-37.4892*</td>
<td>-6.5805*</td>
<td>-5.0922*</td>
</tr>
<tr>
<td>3,2</td>
<td>.0803</td>
<td>-0.1371</td>
<td>-34.4095*</td>
<td>-33.9059*</td>
<td>-6.0838*</td>
<td>-4.2289*</td>
</tr>
<tr>
<td>3,3</td>
<td>.538*</td>
<td>-0.9247</td>
<td>-37.2784*</td>
<td>-35.3847*</td>
<td>-6.0589*</td>
<td>-4.5012*</td>
</tr>
<tr>
<td>3,4</td>
<td>.520*</td>
<td>1.4925*</td>
<td>-25.0244*</td>
<td>-32.2150*</td>
<td>-3.5207*</td>
<td>-3.2492*</td>
</tr>
<tr>
<td>4,1</td>
<td>.922*</td>
<td>1.3967*</td>
<td>-16.3625*</td>
<td>-21.3772*</td>
<td>-1.3486*</td>
<td>-1.9997*</td>
</tr>
<tr>
<td>4,2</td>
<td>.897*</td>
<td>2.6622*</td>
<td>-15.6763*</td>
<td>-13.6466*</td>
<td>0.0674</td>
<td>-1.2017*</td>
</tr>
<tr>
<td>4,3</td>
<td>.341*</td>
<td>1.2792</td>
<td>-18.3428*</td>
<td>-19.5999*</td>
<td>-1.4030</td>
<td>-2.0499*</td>
</tr>
<tr>
<td>4,4</td>
<td>.685*</td>
<td>3.5496*</td>
<td>-14.7366*</td>
<td>-17.3208*</td>
<td>-0.7490*</td>
<td>-2.1216*</td>
</tr>
<tr>
<td>5,1</td>
<td>.910*</td>
<td>4.0813*</td>
<td>-8.9430*</td>
<td>-9.3339*</td>
<td>1.8555*</td>
<td>-1.2490*</td>
</tr>
<tr>
<td>5,2</td>
<td>.894*</td>
<td>3.3561*</td>
<td>-6.9539*</td>
<td>-9.9501*</td>
<td>1.9110*</td>
<td>-0.3765</td>
</tr>
<tr>
<td>5,3</td>
<td>.418*</td>
<td>6.0090*</td>
<td>-2.7376*</td>
<td>-4.1099*</td>
<td>2.3076*</td>
<td>-0.1026</td>
</tr>
<tr>
<td>5,4</td>
<td>.719*</td>
<td>4.4925*</td>
<td>-6.9549*</td>
<td>-10.7695*</td>
<td>2.4473*</td>
<td>-1.1711*</td>
</tr>
<tr>
<td>6,1</td>
<td>.249*</td>
<td>9.5472*</td>
<td>18.4615*</td>
<td>13.5962*</td>
<td>7.5210*</td>
<td>0.7417*</td>
</tr>
<tr>
<td>6,2</td>
<td>.102*</td>
<td>8.4033*</td>
<td>12.8348*</td>
<td>7.9850*</td>
<td>6.8282*</td>
<td>0.9291*</td>
</tr>
<tr>
<td>6,3</td>
<td>.005*</td>
<td>7.9284*</td>
<td>4.9343*</td>
<td>2.9804*</td>
<td>4.4097*</td>
<td>0.6028*</td>
</tr>
<tr>
<td>6,4</td>
<td>.970*</td>
<td>8.7238*</td>
<td>7.8667*</td>
<td>4.6583*</td>
<td>5.4715*</td>
<td>0.8578*</td>
</tr>
</tbody>
</table>

Sex: 0.512* 0.1823 2.8326* 1.6487 1.4939* 0.9986* 0.2777* 1.0733 1.3412*

Community Type: 7.414* 6.1012 24.1446* 16.8836* 6.0546* 2.8381* 2.3880* 4.7236 5.6100*

Variance: 60.73 42.378 684.56 583.09 49.977 10.449 7.4682 29.776 51.693

* denotes coefficient significantly different from zero (via an asymptotic t-test).
conjecture that the urban environment has a beneficial effect on dropouts is supported: the coefficients on the community-type variable are higher in every test than in the combined in-school/dropout sample. The sex coefficient does not exhibit a significant change.
Test 5: Simple Operation

Figure 4.15
5. Conclusions

We have shown that in an ANOVA model for test scores censoring leads to biased and inconsistent estimates of the parameters. The extent and direction of this bias depends upon the test "width", the "centering" of the test and upon the parameters of the underlying model of the individual's latent skill-levels. Because test cannot be properly centered for all "treatment groups" and can be made "wide enough" only at considerable increases in time and expense of administration we suspect that the problem of censoring is a general one, and of practical importance to most applications in educational measurement.

In the present case, the use of biased OLS parameter estimates causes an erroneous "flattening" and "compression" of the "true" learning and skill-retention curves. Although the relative positions of the grade-specific curves is retained, the existence of skill attenuation and threshold-level effects is masked by the use of biased estimates. The consistent ML estimates yield "steeper" curves, and exhibit larger differences between grade-levels.

Policy recommendations based on the results of the Egyptian literacy and numeracy retention study so far must, of necessity, be tentative. The available test instruments in our cross-section sample have a limited set of exogenous variables - i.e., sex and community-type/school number. Clearly more information is required concerning the social, economic, and educational determinants of skill-acquisition before we can identify from these significant determinants those variables subject to manipulation by policy makers. Further, time series data on individuals, i.e., a longitudinal sample, would be required-both in order to investigate the determinants of the dropout decision, itself, (and consequent self-selection bias in all grade-specific skill-level
estimates) and to permit specification of an individual error component in
the disturbance term of the skill-level model. The latter would permit a
decomposition of the error variance, $\sigma^2$, into the portion attributable to
individual-specific, but unmeasured, determinants of skill-levels and the
residual variance portion.

We have already identified substantial differences in skill-levels
across urban versus rural locations. This may be of some interest to policy
makers as they consider the type and location of new investment in educational
facilities. If differences between urban and rural students are due primarily
for difference in the quality of the schools and teaching staff found in such
communities, the remedy is fairly obvious. If, however, as we have conjectured,
some of the differences are environmental in origin, it will be far more
difficult to achieve educational parity between regions.

The differences between skill-levels attributable to individual
schools are both a problem for and a possible source of information to the
ARE as it undertakes its program of educational reform. The fact that some
schools apparently provide their students with the equivalent of a several
grade-level advantage over others suggests that models of successful
educational innovations may exist within the present system. A case by case
review of the facilities, teachers' qualifications, and educational methods
employed at each of the 60 sample schools may suggest plausible planning
goals. Needless to say, with additional data on school, teacher and community
variables, our statistical methods should permit an assessment of their
relative and quantitative significance.
It is interesting to note that, while differences between males and females are generally significant (in a statistical sense), they are relatively small. Formal equality for women has existed in Egypt since the 1920's, but considerable differences in their family and social roles are apparent to even a casual observer. Female dropout rates are higher than for males, and they are more likely to work in the home than in the outside economy. Nevertheless, the school system seems to be meeting at least some of their need for basic skills.

Finally, a few caveats. We have demonstrated the "correct" ANOVA estimates in the case of normal censored test-score data, but this not a panacea. A considerable amount of information is lost when the range of observation is artificially limited. Careful piloting of tests and adjustment of the instruments to match the anticipated range of skill-levels to be measured is clearly a preferred ex ante strategy. We note also that the restricted set of exogenous variables, available to us at the present time, limits the range of policy implications that may be derived from this study and, as in the case of the school effects model, obliges us to use rather unparsimonious model specifications. The expected skill-levels attributable to merely knowing a given grade-year cell without knowledge of the many associated factors which determine an individual's skill level is unsatisfactory. More extensive data collection on the cross-section sample is now being carried out. Finally, it is clear that in modeling a process characterized by decisions made over time a longitudinal sample design is of more than academic importance. We are, at present, simply unable to disentangle the
dropout decision from the process of skill acquisition. Follow up data on each dropout and a subset of in-schoolers in our sample will permit us to report further results on this problem at a later date.
References


MAXIMUM LIKELIHOOD ESTIMATION OF THE TRUNCATED AND CENSORED NORMAL REGRESSION MODELS

by

Michael J. Hartley *

and

Eric V. Swanson**

Development Economics Department
World Bank

November 1980

This paper has been completed as part of a joint World Bank/National Center for Educational Research, Ministry of Education, Egypt (WB/NCER) research project entitled The International Study for the Retention of Literacy and Numeracy (RP0 671-55), financed by the World Bank. The authors are indebted to their colleagues and collaborators, Messrs. I. Serageldin, M. Wilson, E.M. Abdel-Hawgood, R. Durstine, J. Issa, S. Jarar, J. Socknat, W. Haddad and S. Heynen for their encouragement and thoughtful comments during the course of this project; to Messrs. B.B. King, E. Scoutjesdijk and M. Selowsky for permitting the considerable time input that has been required; and to our collaborators at NCER, Dr. Youssef Khalil Youssef and Messrs. M. Hamshary and S. Saleh for their painstaking work in development of the test instruments, organization of the field work and editing of the data. Finally, Professors I. Wergedlin, M. Chalbderaine, A. El-Koussy, D. Cole, S. Ibrahim, N. Postlethwaite, C. Hoyt, R. Wolf and D. Poirier served as consultants during earlier phases of the work, whilst Ms. V. Bencivenga provided excellent research assistance.

* Economist, Development Economics Department, World Bank

** Researcher, Development Economics Department and Technical Assistance and Special Studies, Europe, Middle East and North Africa Region, World Bank
MAXIMUM LIKELIHOOD ESTIMATION OF THE TRUNCATED
AND CENSORED NORMAL REGRESSION MODELS

By
Michael J. Hartley
and
Eric V. Swanson

I. Introduction

Frequently in applied work one encounters a situation in which
observations on a dependent variable with potentially infinite range are only
recorded within a specific interval. In some cases, no information is
available on values outside the admissible range (truncated samples). In
other instances, the number of observations outside the specified range is
known, but their actual values are unknown (censored samples). In the case
where the underlying population distribution is normal, there is already an
extensive literature on the estimation of the mean and variance from both
truncated and censored samples. In the case of singly-truncated normal
samples with a known truncation point, the literature dates back to Pearson
and Lee (1908) and Fisher (1931), with subsequent contributions by Hald (1949)
and Cohen (1949, 1950). The corresponding censored normal sample case has
been studied by Stevens (1937), Cochran (1946), Hald (1949) and Cohen
(1950). Both the doubly-truncated and doubly-censored normal sample cases
have been treated by Stevens (1937) and Cohen (1949).

While the above contributions have treated, inter alia, Maximum
Likelihood (M.L.) estimation in the context of truncated or censored samples
from a normal population with constant mean and variance, more recent work has
placed the problem in a regression setting. Tobin (1958), Amemiya (1973) and
Hartley (1976) have examined the case of a linear regression model in which the dependent variable has a left (right)-censored normal distribution, and Rosett and Nelson (1975) have generalized the analysis to a doubly-censored normal regression model. More recently, Hausman and Wise (1977) have provided a discussion of the M.L.E. for the singly-truncated regression model, and Olsen (1980) has generalized the Pearson-Lee method-of-moments estimator to the regression problem.

Within the social sciences there have been numerous models in which the observations on the dependent variable within a regression model follow a truncated or censored normal distribution. Perhaps the richest harvest of examples is to be found within microeconomic models of individual behavior (see, e.g., Deaton and Muellbauer (1980)), where censored regression models have been applied to the study of labor supply, consumption of goods and services, expenditures on durable goods, etc.—all of which are left-censored at zero. A right-truncated normal regression model has been applied by Hausman and Wise (1977) to study the effects of age, education, intelligence and unionization on the earnings of poor families in the context of a social experiment.

Censoring and truncation frequently occur when a variable of interest must be measured or calibrated by means of a measurement instrument or gauge. For example, in quality control of the output of an industrial process, a truncated sample arises when "defectives" above or below a particular tolerance level, as determined by an external measuring device, are

---

1/ Rosett and Nelson (1975) refer to this situation as a "two limit Probit model." This suggests an affinity with the standard Probit model (see, e.g., Finney (1952)), which obtains as a special case of bilateral censoring with both limits points being constant and identical (say zero) for each observation.
discarded. Similarly, a thermometer is only capable of measuring temperatures over a pre-specified closed interval, say \([t_1, t_2]\). Readings of \(t_1(t_2)\) simply convey the information that the "true" temperature was less (greater) than or equal to \(t_1(t_2)\). The same may be said of achievement tests used to measure educational attainment—an example which we shall utilize subsequently for illustrative purposes. Zero or perfect scores may be viewed as indicative of the fact that, on the implied scale, a student's latent ability level was below or above that associated with the minimum or maximum number of possible points for the test instrument. This type of bilateral censoring (in the distribution of test scores) is likely to be prevalent when the test covers too "narrow" a range of item difficulties as compared to the range of student abilities within the sample. A doubly-censored regression model arises if one wishes to "explain" variation in test scores on the basis of various individual, school, teacher, family and community characteristics.

In spite of the great variety of important applications for various types of truncated and censored linear regression models, including those for the analysis of variance and covariance, the theoretical literature is still deficient in several respects:

(a) An initial consistent estimator is only available for the singly-truncated/censored case (Amemiya (1973)).

(b) The asymptotic distribution of the MLE has only been recorded for the case of the singly-censored normal dependent variable (Amemiya (1973)).

(c) The algorithms proposed for calculation of the MLE for various types of truncated/censored normal regression problems include the Newton-Raphson (Amemiya (1973)) and Gauss-Newton (Hausman and Wise (1977)) algorithms. Alternative approaches include applying the so-called
E-M algorithm (see Dempster, Laird and Rubin (1977), Hartley (1958) and Hartley (1976)), or the Method of Scoring (see, e.g., Rao (1965)) to such problems.

It is the purpose of the present paper to collect the existing results within the literature and to fill in the above gaps. In particular, we shall provide a class of (weakly) consistent initial estimators for various types of truncated/censored normal regression models and provide explicit formulas for the asymptotic covariance matrices of the MLE's in the singly- and doubly-truncated, as well as the doubly-censored, regression models—thereby extending the results of Amemiya (1973). Further, we shall develop the requisite formulas for application of the Newton-Raphson (N-R), Method-of Scoring (M-S), Gauss-Newton (G-N) and Expectation-Maximization (E-M) algorithms to actually calculate the M.L.E.'s for the general doubly-truncated/censored case. Finally, we shall provide some tentative guidelines for selection of algorithms on the basis of evidence obtained from the application of the doubly-censored normal regression model to estimate the skill-retention "curves" of dropouts from primary school in Egypt. 1/

---

1/ This has arisen in the context of the aforementioned multi-year World Bank Research Project, "The International Study for the Retention of Literacy and Numeracy," conducted by the World Bank in collaboration with the National Center for Educational Research, Ministry of Education, Egypt.
2. Truncated and Censored Normal Regression Models:

2.1. The Underlying Model:

We shall assume that the observed sample of data on a truncated or censored dependent variable, \( y_i \), has been generated by the "latent model,"

\[
(2.1) \quad y_i^* = x_i' \beta_o + \varepsilon_i, \quad -\infty < y_i^* < \infty,
\]

where \( x_i \) is a K-vector of observations on the independent variables, \( \beta_o \) is a K-vector of corresponding unknown constants and \( \varepsilon_i \) is a normal and independently distributed disturbance,

\[
(2.2) \quad \varepsilon_i \sim n.i.d.(0, \sigma_o^2).
\]

It follows that the underlying dependent variable, \( y_i^* \), has the normal density,\(^1\)

\[
(2.3) \quad f_{i0}(y_i^*) = \frac{1}{\sqrt{2\pi} \sigma_o} \cdot \exp\left(\frac{-1}{2 \sigma_o^2} \cdot (y_i^* - x_i' \beta_o)^2\right),
\]

with associated distribution function,

\[
(2.4) \quad F_{i0}(a) = \int_{-\infty}^{a} f_{i0}(y_i^*) \, dy_i^*.
\]

2.2. Truncated And Censored Samples:

Truncated and censored samples arise when the underlying dependent variable, \( y_i^* \), is not observed over its entire range, \((-\infty, +\infty)\). Rather, for

\(^1\) The subscript "o" denotes evaluation at the true parameter values.
each observation, there exist known lower and/or upper limit points, \( z_{i1} \) and \( z_{i2} \), such that for each \( i \), the observed values for the dependent variable, \( y_i \), satisfy
\[
(2.5) \quad -\infty < z_{i1} < y_i < z_{i2} < +\infty,
\]
with the proviso
\[
(2.6) \quad z_{i1} < z_{i2}.
\]

In the case of truncated samples, the relationship between \( y_i \) and \( y_i^* \) is given by:
\[
(2.7) \quad y_i = y_i^* \quad \text{if} \quad z_{i1} < y_i^* < z_{i2},
\]
whereas the \( y_i^* \) values outside the range, \((z_{i1}, z_{i2})\), and the associated \( x_i \)-vectors are not observed. Accordingly, \( y_i \) has the density function,
\[
(2.8) \quad g_{i0}(y_i) = \begin{cases} 
\frac{f_{i0}(y_i)}{[F_{i0}(z_{i2}) - F_{i0}(z_{i1})]}, & \text{if } z_{i1} < y_i < z_{i2} \\
0, & \text{otherwise}
\end{cases}
\]
Three cases are commonly of interest. If \( z_{i1} > -\infty \) and \( z_{i2} = +\infty \) for each \( i \), we have the left-truncated normal regression model with \( 0 < F_{i0}(z_{i1}) < 1 \) and \( F_{i0}(z_{i2}) = 1 \) in (2.8). Similarly, if \( z_{i1} = -\infty \) and \( z_{i2} < +\infty \), we have the right-truncated model in which \( g_{i0}(y_i) \) of (2.8) simplifies via \( F_{i0}(z_{i1}) = 0 \) and \( 0 < F_{i0}(z_{i2}) < 1 \). The general case of a bilaterally- or doubly-truncated regression model arises when both \( z_{i1} > -\infty \) and \( z_{i2} < +\infty \), whence
\[
0 < F_{i0}(z_{i1}) < F_{i0}(z_{i2}) < 1.
\]
Thus, in all these cases the density for a truncated sample is continuous over the range of observation, \((z_{i1}, z_{i2})\).

In other applications one finds an accumulation of probability mass at either (or both) of the limit points, whereby the observations on \( y_i \) are related to the latent variable, \( y_i^* \), by:
(2.9) \[ y_1 = \begin{cases} z_{11} & \text{if } y_1^* < z_{11} \\ y_1^* & \text{if } z_{11} < y_1^* < z_{12} \\ z_{12} & \text{if } y_1^* > z_{12} \end{cases} \]

This is the situation in censored samples in which \( y_1^* \) is observed as \( y_1 \) within the open interval, \((z_{11}, z_{12})\), but, whenever \( z_{11} \) and/or \( z_{12} \) are finite, the knowledge that \( y_1^* \) was less than or equal to \( z_{11} \) (greater than or equal to \( z_{12} \)) is indicated by \( y_1 \) being observed as \( z_{11} \) (\( z_{12} \)). Again, a left censored normal regression (or Tobit) model obtains when \( z_{11} > -\infty \) and \( z_{12} = +\infty \); a right-censored model obtains if \( z_{11} = -\infty \) and \( z_{12} < +\infty \); and a bilaterally-censored normal regression model follows when both \( z_{11} > -\infty \) and \( z_{12} < +\infty \). In the last of these, the density of \( y_1 \) is defined by:

(2.10) \[ h_{10}(y_1) = \begin{cases} F_{10}(z_{11}) & \text{if } y_1 = z_{11} \\ f_{10}(y_1) & \text{if } z_{11} < y_1 < z_{12} \\ 1 - F_{10}(z_{12}) & \text{if } y_1 = z_{12} \end{cases} \]

Thus, in this censored example, \( y_1 \) has a non-null probability mass at each of the finite limit points and a continuous density between both limit points.

One special case of the bilaterally-censored model is the Probit model, in which \( z_{11} = z_{12} = 0 \). In this case the actual values of \( y_1^* \) are never observed. Rather, the fact that \( y_1^* \) was negative (positive) is indicated by the binary variable,

\[ y_1 = \begin{cases} 0 & \text{if } y_1^* < 0 \\ 1 & \text{if } y_1^* > 0 \end{cases} \]
where $y_i$ now has the density function,

$$h_{f_Y}(y_i) = \begin{cases} F_{\theta}(0), & \text{if } y_i = 0 \\ 1-F_{\theta}(0), & \text{if } y_i = 1 \end{cases}$$

and $\sigma^2_\theta$ is set to unity, since it is not identifiable.

Mixed truncation/censoring cases with multiple intervals are, of course, also possible, but do not fall explicitly within the bilateral model. Treatment of such cases, however, should be obvious from our present discussion.

2.3. Assumptions:

Following Amemiya's (1973) discussion of the singly-censored normal regression model, we shall make the following assumptions for the doubly-truncated/censored case:

**Assumption 1:** Let $\theta = \left[ \begin{array}{c} \beta \\ \sigma^2 \end{array} \right]$ denote an arbitrary point in the parameter space, $\Theta$, and let $\hat{\theta} = \left[ \begin{array}{c} \hat{\beta} \\ \hat{\sigma}^2 \end{array} \right]$ denote the true value of $\theta$. Then $\Theta$ is compact, excludes the region, $\sigma^2 < 0$, and contains and open neighborhood, $\Theta_0$, of $\hat{\theta}_0$.

**Assumption 2:** The regressors, $\{x_i\}$, are bounded and have an empirical distribution function, $P_N(x) = \frac{1}{N}$ (where $j$ is the number of points $x_1, x_2, \ldots, x_N$ less than or equal to $x$) which converges to a distribution function (say $F$) as $N \to \infty$.

**Assumption 3:** The moment matrix,

$$M_{xx} = \lim_{N \to \infty} \frac{1}{N} \left[ \sum_{i=1}^{N} x_i x_i' \right],$$

is positive definite.
Assumption 4.1/ The limit point difference, \( z_{i2} - z_{i1} \), is uniformly bounded from below by zero, i.e., there exists a \( \delta > 0 \) such that for all \( i \),

\[
(2.12) \quad (z_{i2} - z_{i1}) > \delta > 0 .
\]

2.4. The Log-Likelihood Function and Its Derivatives: 2/

Let \( S_N \) denote the observation index set, \{1, 2, ..., N\}, and define the index subsets, \( S^N_j = \{ i \in S^N : y^*_1 \in \gamma_j \} \), \( j = 1, 2, 3 \), where

\[
(2.13) \quad \begin{align*}
    y^*_1 & \in Y_1 \text{ if } -\infty < y^*_1 < z_{i1} , \text{ i.e., } y^*_1 = z_{i1} \\
    y^*_1 & \in Y_2 \text{ if } z_{i1} < y^*_1 < z_{i2} , \text{ i.e., } y^*_1 = y^*_1 \\
    y^*_1 & \in Y_3 \text{ if } z_{i2} < y^*_1 < +\infty , \text{ i.e., } y^*_1 = z_{i2} .
\end{align*}
\]

Let \( S^N_j \) have \( N_j > 0 \) elements, \( j = 1, 2, 3 \), with \( N_1 + N_2 + N_3 = N \).

For any function, \( q(y^*_1; x_1, \theta) \), we shall use the shorthand notation:

\[
(2.14) \quad \begin{align*}
    q_i & = q(y^*_1; x_1, \theta) , \\
    q_{ij} & = q(z_{ij}; x_1, \theta) , \quad j = 1, 2, \\
    q_{io} & = q(y^*_1; x_1, \theta_0) , \\
    q_{ijo} & = q(z_{ij}; x_1, \theta_0) , \quad j = 1, 2.
\end{align*}
\]

1/ Assumption 4 is not required in either the singly (left or right) truncated or the censored regression models considered by Amemiya (1973). It is also not required in the case of the Probit model, since, in this case, the terms, \( F_{10}(z_{i2}) - F_{10}(z_{i1}) \), never appear in relevant expressions.

2/ Henceforth, we shall discuss the general case of the bilaterally- or doubly-truncated and censored normal regression models—except in section 3, where methodological differences between the singly- and doubly-truncated/censored cases require separate discussion.
Then, in the case of truncated samples, the log-likelihood function is defined by

\[(2.15) \quad \log L_T^N(\theta) = \sum_{i \in S_2} \log f_i - \sum_{i \in S_2} \log \left[F_{i2} - F_{i1}\right],\]

whereas in censored samples the log-likelihood is defined by:

\[(2.16) \quad \log L_C^N(\theta) = \sum_{i \in S_1} \log F_{i1} + \sum_{i \in S_2} \log f_i + \sum_{i \in S_3} \log \left[1 - F_{i2}\right],\]

and \(N\) denotes the sample size. The Maximum Likelihood Estimator (M.L.E.), \(\hat{\theta}_M\), \(M = T, C\), is implicitly defined as the value of \(\theta\) such that:

\[(2.17) \quad \log L_N^M(\hat{\theta}_M) = \sup_{\theta \in \Theta} \{ \log L_N^M(\theta) \}, \quad M = T, C.\]

In section 5 we shall examine various methods to calculate \(\hat{\theta}_M\). It will be noted, there, that these algorithms require the first (and, in some cases, second) partial derivatives of the appropriate density/log-likelihood function.

Let us define:

\[(2.18a) \quad u_i = y_i - x_i \theta\]

and

\[(2.18b) \quad u_{ij} = z_{ij} - x_i \theta, \quad j = 1, 2.\]

Then, the first partials of the log-likelihood function in the doubly-truncated model are given by:
\[
(2.19a) \quad \frac{\partial \log L^T}{\partial \beta} = \frac{1}{\sigma^2} \cdot \sum_{i \in S_2} \left( u_i + 2 \sigma^2 (g_{12} - g_{11}) \right) \cdot x_i
\]

and

\[
(2.19b) \quad \frac{\partial \log L^T}{\partial \sigma^2} = \frac{1}{2 \sigma^4} \cdot \sum_{i \in S_2} \left( u_i^2 - \sigma^2 + \sigma^2 (u_{12} g_{12} - u_{11} g_{11}) \right)
\]

whereas the second partials are given by:

\[
(2.20a) \quad \frac{\partial^2 \log L^T}{\partial \beta \partial \beta^\prime} = -\frac{1}{\sigma^2} \cdot \sum_{i \in S_2} \left( 1 - (u_{12} g_{12} - u_{11} g_{11}) - \sigma^2 (g_{12} - g_{11})^2 \right) \cdot x_i x_i^\prime
\]

\[
(2.20b) \quad \frac{\partial^2 \log L^T}{\partial \beta \partial \sigma^2} = -\frac{1}{\sigma^4} \cdot \sum_{i \in S_2} \left( u_i - \frac{1}{2} \cdot \left( (u_{12} - \sigma^2) g_{12} - (u_{11} - \sigma^2) g_{11} \right) \right)
\]

\[-\frac{\sigma^2}{2} \cdot (g_{12} - g_{11}) \cdot (u_{12} g_{12} - u_{11} g_{11}) \cdot x_i
\]

and

\[
(2.20c) \quad \frac{\partial^2 \log L^T}{\partial \sigma^2 \partial \sigma^2} = -\frac{1}{\sigma^6} \cdot \sum_{i \in S_2} \left( u_i^2 - \frac{\sigma^2}{2} + \frac{3 \sigma^2}{4} \cdot (u_{12} g_{12} - u_{11} g_{11}) \right)
\]

\[-\frac{1}{4} \cdot (u_{12} g_{12} - u_{11} g_{11}) - \frac{\sigma^2}{4} \cdot (u_{12} g_{12} - u_{11} g_{11}) \cdot x_i
\]

In contrast, the first partials in the doubly-censored model are given by:

\[
(2.21a) \quad \frac{\partial \log L^C}{\partial \beta} = \frac{1}{\sigma^2} \cdot \left( - \sum_{i \in S_1} \sigma^2 \cdot \frac{f_{11} x_i}{F_{11}} + \sum_{i \in S_2} u_i x_i + \sum_{i \in S_3} \frac{\sigma^2}{1 - F_{12}} \cdot x_i \right)
\]

and
(2.21b) \[ \frac{3 \log L}{\partial \sigma^2} = \frac{1}{2 \sigma^4} \left( - \sum_{i \in S_1} \sigma^2 \cdot u_{i11} \cdot \frac{f_{11}}{F_{11}} + \sum_{i \in S_2} \left( u_{i1}^2 - \sigma^2 \right) \cdot \frac{f_{12}}{1 - F_{12}} \right) \],

whereas the second partials are given by:

(2.22a) \[ \frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} = - \frac{1}{\sigma^2} \left( \sum_{i \in S_1} \left( u_{i1}^2 + \sigma^2 \cdot \frac{f_{11}}{F_{11}} \cdot \frac{f_{11}}{F_{11}} \cdot x_i^2 \right) + \sum_{i \in S_2} \left( -u_{i12} + \sigma^2 \cdot \frac{f_{12}}{1 - F_{12}} \right) \cdot \frac{f_{12}}{1 - F_{12}} \cdot x_i^2 \right), \]

(2.22b) \[ \frac{\partial^2 \log L}{\partial \beta \partial \sigma^2} = - \frac{1}{\sigma^4} \left( \frac{1}{2 \cdot i \in S_1} \Sigma \left( \sigma^2 \cdot u_{i11} \cdot \frac{f_{11}}{F_{11}} - \sigma^2 + u_{i11}^2 \right) \cdot \frac{f_{11}}{F_{11}} \cdot x_i^2 \right) + \sum_{i \in S_2} \left( \sigma^2 \cdot u_{i12} \cdot \frac{f_{12}}{1 - F_{12}} + \sigma^2 - u_{i12}^2 \right) \cdot \frac{f_{12}}{1 - F_{12}} \cdot x_i^2 \right) \]

and

(2.22c) \[ \frac{\partial^2 \log L}{(\partial \sigma^2)^2} = - \frac{1}{\sigma^6} \left( - \frac{1}{4 \cdot i \in S_1} \Sigma \sigma^2 \cdot u_{i11} \cdot \frac{f_{11}}{F_{11}} \cdot \left( 3 - \frac{1}{\sigma^2} \cdot u_{i11}^2 - \frac{f_{11}}{F_{11}} \right) \right) \]

\[ + \sum_{i \in S_2} \left( u_{i1}^2 - \frac{1}{2} \cdot \sigma^2 \right) - \frac{1}{x} \sum_{i \in S_1} \sigma^2 \cdot u_{i12} \cdot \frac{f_{12}}{1 - F_{12}} \cdot \left( -3 + \frac{1}{2} \cdot u_{i12}^2 - \frac{f_{12}}{1 - F_{12}} \right) \right), \]

respectively.

2.5. The Conditional Moments of the Latent Dependent Variable:

Let the conditional density function of \( y_{1i}^* \), given \( y_{1i} \), be denoted by:
(2.23) \( u_{ij}(y_1^*) = \begin{cases} \frac{\varepsilon_i(y_1^*)}{F_{il}}, & \text{if } y_1^* \in Y_1 \ (j=1) \\ \frac{\varepsilon_i(y_1^*)}{F_{i2}-F_{i1}}, & \text{if } y_1^* \in Y_2 \ (j=2) \\ \frac{\varepsilon_i(y_1^*)}{1-F_{i2}}, & \text{if } y_1^* \in Y_3 \ (j=3) \end{cases} \)

Then the \( r \)th conditional moment of \( y_1^* \), given \( y_j^* \), \( j=1,2 \) or \( 3 \), is defined as:

(2.24) \( \mu_{ij}^{(r)} = E[y_1^{*r} | y_j^*] = \int y_1^{*r} \cdot u_{ij}(y_1^*) \, dy_1^* \).

It may easily be verified that the conditional moments of \( y_1^* \) for arbitrary \( r \) satisfy the following recursion formulas:

(2.25) \( \mu_{ij}^{(r)} = \mu_{ij}^{(r-1)} \cdot x_i^* + (r-1) \cdot \mu_{ij}^{(r-2)} \cdot \sigma^2 \begin{cases} -\sigma^2 \cdot \frac{z_{i1} f_{i1}}{F_{i1}}, & \text{if } j=1 \\ -\sigma^2 \cdot \frac{z_{i2} f_{i2} - z_{i1} f_{i1}}{F_{i2}-F_{i1}}, & \text{if } j=2 \\ +\sigma^2 \cdot \frac{z_{i2} f_{i2}}{1-F_i}, & \text{if } j=3 \end{cases} \)

---

1/ To our knowledge, such autoregressive relations between the conditional moments of a normal density have not appeared in the literature.
for \( r=1,2,\ldots \), where \( \mu_{ij}^{(0)} = 1 \) and \( \mu_{ij}^{(-1)} = 0 \). From Assumptions 2 and 4, it follows that \( F_{i1}, F_{i2} - F_{i1} \) and \( 1-F_{i2} \) are uniformly bounded away from zero, and thus all conditional moments, \( \mu_{ij}^{(r)} \), are uniformly bounded in \( i \) for any \( r=1,2,\ldots \) and \( j=1,2 \) or 3.

Finally, we note that in the Probit model, there are no observations in \( S_2^N \), so that the log-likelihood is the special case of (2.16) in which \( \sigma^2 = 1, z_{i1} = z_{i2} = 0 \) and

\[
(2.26) \quad \log L_N^C(\beta) = \sum_{i \in S_1^N} \log F_i(0) + \sum_{i \in S_3^N} \log [1 - F_i(0)] .
\]

Here, clearly, \( F_{i1} = F_{i2} = F_i(0) \) and the first and second partials of the log-likelihood with respect to \( \beta \) are given, mutatis mutandis, by (2.21a) and (2.22a), respectively. Also (2.25) holds for any \( r=1,2,\ldots \) with \( j=1 \) (for \( i \in S_1^N \)) or \( j=3 \) (for \( i \in S_3^N \)) provided we define \( z_{ij}^{r-1} \) as one when \( r=1 \).
3. **Consistent Initial Estimators:**

It may easily be verified from inspection of the expressions for the first partials of the log-likelihood functions, \( \log L_N^T \) or \( \log L_N^C \), given in (2.19a) - (2.19b) or (2.21a) - (2.21b), respectively, that analytic solutions to the \((K+1)\) - equation system,

\[
\frac{\partial \log L_N^M}{\partial \theta} = 0, \quad M=T, C,
\]

are not feasible. Thus, the MLE, \( \hat{\theta}^M \), must be found by iterative methods, and we shall require reliable starting parameter values. Our purpose, here, is to present a general methodology, which may be applied, *mutatis mutandis*, to obtain consistent initial estimates, with a positive initial variance estimator, for any type of truncated/censored normal regression problem, though our discussion in this section will be restricted to a taxonomy of model specifications contained within the bilateral model form.

We shall follow the basic approach of Amemiya (1973), in his discussion of Instrumental Variable (IV) estimators for the singly-truncated/censored (Tobit-type) model. Amemiya's ingenious idea was to make use of approximations for the first and second conditional moments of the dependent variable, given that the actual value of the latent variable, \( y_1^* \).

---

1/ We conjecture that our procedures (involving suitable linear transformations of the elements of the sequence of certain higher order conditional moments, in combination with Amemiya's (1973) Instrumental Variable approach) also will have fruitful applications to the problems of obtaining consistent initial estimators in many of the other "limited dependent variable" types of models. This point will be pursued elsewhere.
belongs to $Y^*$, so that $y^*$ is observed as $y_1$. This results in a model which is linear in the parameters of interest, $\beta_0$ and $\sigma_0^2$, and Amemiya shows that, under general conditions, his IV estimator is weakly consistent. We shall extend these ideas to exhibit an entire class of IV estimators, which can be applied to all types of truncated/censored normal regression models by making use of the general recursion formulas for conditional moments provided in (2.25), but adapted to the present problem.

3.1. A Taxonomy of Model Types:

The methodology surrounding development of a class of consistent initial estimators utilizes the sample data on all $(y_1, x_1)$ pairs for which the observed $y_1$-value is equal to the underlying $y^*_1$. In the general formulation of the doubly-truncated/censored model, this would include all data such that $z_{11} < y^*_1 < z_{12}$, i.e., all $i \in S^N_2$. Thus, for the censored models, in calculating initial estimates, we ignore observations, $y_1$, equal to the limit points, $z_{11}$ or $z_{12}$ of (2.9). In contrast, all observations are used in the truncated case.

It will be convenient to analyze a taxonomy of special cases which emerges from the "observed" part of the doubly-truncated/censored normal regression model, i.e.,

$$
(3.2) \quad y^*_1 = x_1' \beta_0 + \varepsilon_1, \quad z_{11} < y^*_1 < z_{12}, \quad i \in S^N_2.
$$

We shall consider the cases in which $y^*_1$ is constrained to lie within three types of intervals:

\[1/\] Note that this precludes use of the IV approach to obtain consistent initial estimators for the Probit model. However, since the log-likelihood is globally concave for any $\beta$, all of the customary algorithms in the Probit case (see sections 5.1 to 5.4) are guaranteed to converge from any initial value, say $\beta = 0$. 
(i) Left-truncated: \( z_{11} > -\infty \) and \( z_{12} = -\infty \)
(ii) Right-truncated: \( z_{11} = -\infty \) and \( z_{12} < +\infty \)
(iii) Doubly-truncated: \( z_{11} > -\infty \) and \( z_{12} < +\infty \)

when \( y_i^* \) is observed as \( y_i \). If the finite limit points, \( z_{11} \) and/or \( z_{12} \) are the same for all \( i \in S_2^N \), we shall refer to such a case as a "fixed" truncation point (with \( z_{11} = z_1 \) and/or \( z_{12} = z_2 \)). Cases in which the upper and/or lower limit points vary from observation to observation are called "variable" limits. Finally, as will be seen, the procedures to be employed vary according to whether or not the regression function, \( x_i \beta_0 \), contains a constant term. If so, we denote this as:

\[
(3.3) \quad x_i \beta_0 = \begin{bmatrix} 1 & x_i^* \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_2 \end{bmatrix} = \beta_0 + x_i^* \beta_2 \quad \cdot
\]

In Table 3.1 we exhibit each of the 15 possible cases. For each case, we note the type of truncation (col. (1)), the assumptions on \( z_{11} \) and \( z_{12} \) (cols. (2) and (3)) and whether or not a constant term is present (col. (4)). Without loss of generality, each of the special cases within model (3.2) may be transformed into a model of the form:

\[
(3.4) \quad y_i^{**} = x_i^{**} \delta_0 + \varepsilon_i^{**}, \quad 0 < y_i^{**} < z_i^{**}, \quad i \in S_2^N,
\]

where the appropriate transformation and the definitions of \( x_i^{**}, \delta_0, z_i^{**} \)...
<table>
<thead>
<tr>
<th>Truncation Type:</th>
<th>( x_{11} )</th>
<th>( x_{12} )</th>
<th>Constant</th>
<th>( y_1 ) **</th>
<th>( z_1 ) **</th>
<th>( \delta_0 ) **</th>
<th>Linear Restrictions on ( \delta_0 )</th>
<th>(9) Order of ( x_1 ) **</th>
<th>(10) ( z_1 ) **</th>
<th>(11) ( c_1 ) **</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Fixed Left</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Fixed Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Fixed Double</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Fixed Left and Variable Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Variable Left and Fixed Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Variable Left</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Variable Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Variable Double</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>Yes</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Fixed Left</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Fixed Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Fixed Double</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Fixed Left and Variable Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Variable Left and Fixed Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Variable Left</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Variable Right</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Variable Double</td>
<td>( x_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>( y_1 - z_1 )</td>
<td>( x_1 )</td>
<td>No</td>
<td>K ( \leftrightarrow ) ( \epsilon_1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and $\varepsilon_{a}^*$ are given in cols. (5), (6), (7), (10) and (11), respectively.1/ 

Finally, whether or not $\delta_{0}$ has a linear restriction (e.g., in some cases the first element of $\delta_{0}$ is, by construction, unity) and the order of the $K^*$ vector, $x_{1}^*$, (either a $K$ or $K+1$ element vector) are given in cols. (8) and (9). For example, in the most general case of double-truncation in which both the upper and lower limits may vary ($\varepsilon_{a}^*$), we employ the transformation,

$$ y_{1}^* = y_{1} - z_{11} = -z_{11} \cdot 1 + x_{1}^* \delta_{0} + \varepsilon_{1} 

= x_{1}^* \delta_{0} + \varepsilon_{1}^* , \quad 0 < y_{1}^* < z_{11}^* ,$$

where

$$ x_{1}^* = \begin{bmatrix} -z_{11} & x_{1}^* \end{bmatrix} ,$$

$$ \delta_{0}^* = \begin{bmatrix} 1 & \beta_{0}^* \end{bmatrix} .$$

1/ When fixed truncation or censoring at zero occurs—either $z_{11} = 0$ or $z_{12} = 0$ for all $i$—a singularity in the matrix, $[ \sum_{i \in S_{2}} K_{i} x_{i}^* ]$, forces certain changes. In these cases:

Case 9 may be treated as Case 1 with $z_{1} = 0$.

Case 10 may be treated as Case 2 with $z_{2} = 0$.

Case 11 may be treated as Case 4 with $z_{1} = 0$.

Case 12 may be treated as Case 5 with $z_{2} = 0$, and

Case 13 may be treated as Case 3 if $z_{1} = 0$ and by Case 11 if $z_{2} = 0$.

—all of which involve either changing the transformation employed or absorbing $z_{j}$, $j=1$ or 2, into $\delta_{0}$ to avoid the singularity.
\[ z_1^{**} = (z_{12} - z_{11}) < \infty. \]

It follows in all cases that, provided suitable restrictions are imposed on \( \delta \), when required by \( \delta \), consistent estimates of the elements of \( \delta \) may be directly recovered from consistent estimates of \( \delta \). Hence, the problem reduces to one of estimating \( \delta \) and \( \sigma^2 \) in model (3.4), with

\[ f_2^{**} \sim \text{n.i.d.} \ (0, \sigma^2). \]

Thus, to summarize, in (3.4) \( \delta \) is a \( K^{**} \)-vector (with either \( K^{**} = K \) or \( K+1 \)), which may (cases #1 to #5) or may not (cases #6 to #16) be subject to a linear restriction, \( \delta_{10} = 1 \) with \( \delta \equiv \begin{bmatrix} \delta_{10} \\ \vdots \\ \delta_{20} \end{bmatrix} \). Finally, either \( z_1^{**} = \infty \) (Type 1) or \( z_1^{**} < \infty \) (Type 2).

3.2. The Use of Conditional Moments and a Class of Instrumental Variable Estimators:

The conditional density of \( y^{**} \) in (3.4), given \( 0 < y_{1}^{**} < z_{1}^{**} \), is defined by:

\[ f_{10}^{**} (y_1) = \frac{f_{10}^{**} (y_1)}{[F_{10}^{**} (z_1) - F_{10}^{**} (0)]}, \quad 0 < y_{1}^{**} < z_{1}^{**}. \]

where

\[ f_{10}^{**} (y_1) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp \left\{ - \frac{1}{2\sigma_0^2} (y_{1}^{**} - \frac{y_{1}^{**}}{\delta_0})^2 \right\}. \]

Thus, in this section, letting
(3.7)  \[ \mu_{120}^{(r)} = E_0 \left[ y_{1}^{**} \mid 0 < y_{1}^{**} < z_{1}^{**} \right] = \int_{0}^{z_{1}^{**}} y_{1}^{**} g_{10}^{**} (y_{1}^{**}) \ dy_{1}^{**} \]

where \( E_0 \) denotes the expectation relative to the true parameter point, \( \theta_0 \), and, following the recursion formulas of (2.25), subject to a zero lower limit as a result of the transformation to (3.4), we have:

(3.8)  \[ \mu_{120}^{(r)} = \mu_{120}^{(r-1)} \cdot x_{1}^{**} \cdot \frac{z_{1}^{**} - \delta_0}{\sigma_0} + (r-1) \cdot \mu_{120}^{(r-2)} \cdot \frac{\frac{z_{1}^{**} - \delta_0}{\sigma_0} - \frac{z_{1}^{**} - \delta_0}{\sigma_0}}{[F_{10}^{**}(z_{1}^{**}) - F_{10}^{**}(0)]} \]

for \( z_{1}^{**} = 0 \) and \( r=1,2,\ldots \). The existence of a fixed zero lower limit after the transformation is crucial to the method. In the particular cases of Table 3.1 in which \( z_{1}^{**} = 0 \), clearly \( F_{10}^{**}(z_{1}^{**}) = 0 \) and, as in Amemiya (1973), equation (3.8) then reduces to:

(3.9)  \[ \mu_{120}^{(r)} = \mu_{120}^{(r-1)} \cdot x_{1}^{**} \cdot \frac{z_{1}^{**} - \delta_0}{\sigma_0} + (r-1) \cdot \mu_{120}^{(r-2)} \cdot \frac{\frac{z_{1}^{**} - \delta_0}{\sigma_0} - \frac{z_{1}^{**} - \delta_0}{\sigma_0}}{[F_{10}^{**}(z_{1}^{**}) - F_{10}^{**}(0)]} \]

which is linear in the \((K+1)\)-vector,

(3.10)  \[ y_{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \delta_0 \\ \cdots \\ \delta_0 \\ \sigma_0 \end{bmatrix} \]

for all \( r \). Thus, we may consider IV estimators based on (3.9) for cases \( \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9, \theta_{10}, \theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{16}, \theta_{17}, \theta_{18}, \theta_{19}, \) and the remaining cases of Table 3.1:

Type \( \mathbb{R} \) (Cases where \( z_{1}^{**} = 0 \)):

For any value of \( r > 2 \), equation (3.9) may be rewritten as:

(3.11)  \[ y_{1}^{**} = (y_{1}^{**} - 1 \cdot x_{1}^{**} \cdot \delta_0 + ((r-1)y_{1}^{**} - 2) \cdot \delta_0 + \mu_{1}^{(r)} \cdot t \in S_{1}^{N} \]
where

\[(3.12) \quad \eta_1^{(r)} = [\gamma_1^{**} - u_{120}] - [\gamma_1^{**(r-1)} - u_{120}^{(r-1)}] \cdot \delta \cdot \delta_0 \]

\[- (r-1) \cdot [\gamma_1^{**(r-2)} - u_{120}^{(r-2)}] \cdot \delta \cdot \delta_0 \]

Note that

\[(3.13) \quad \epsilon_0^{(r)} \eta_1^{(r)} = 0 , \]

and, by Assumptions 1 and 2, \( \epsilon_0^{(r)} \eta_1^{(r)} \) is uniformly bounded—results that are required for the asymptotic distribution theory of section 4. Then, following Amemiya (1973), for all \( i \in S_2 \), define the "predicted values",

\[(3.14) \quad \hat{y}_1^{**} = \hat{x}_1^{**} \cdot \left[ \sum_{i \in S_2} \hat{x}_1^{**} \hat{x}_1^{**} \right]^{-1} \cdot \left[ \sum_{i \in S_2} \hat{x}_1^{**} \hat{y}_1 \right] , \]

provided that \( \left[ \sum_{i \in S_2} \hat{x}_1^{**} \hat{x}_1^{**} \right] \) is positive definite. Then, for any \( r > 2 \),

using the \((K^{**} + 1)\) - vector, \( [\hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \cdot (r-1) \hat{y}_1^{**r-2}] \), as an instrument for \( [\hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \cdot (r-1) \hat{y}_1^{**r-2}] \), define the class of

Unrestricted Instrumental Variable (UIV) estimators (see, e.g., Theil (1971)):

\[(3.15) \quad Y_o^{(r)} = \left[ \begin{array}{c} \hat{y}_o^{(r)} \\ \cdots \\ \hat{y}_o^{(r)} \\ \sigma_o^{(r)} \end{array} \right] = \left[ \begin{array}{c} \sum_{i \in S_2} \hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \\ \cdots \\ \sum_{i \in S_2} \hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \\ \sum_{i \in S_2} \hat{y}_1^{**r-2} \cdot \hat{y}_1 \end{array} \right] \cdot \left[ \begin{array}{c} \hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \\ \cdots \\ \hat{y}_1^{**r-1} \cdot \hat{x}_1^{**} \\ \hat{y}_1^{**r-2} \cdot \hat{y}_1 \end{array} \right]^{-1} \]
This will immediately result in (weakly) consistent estimators for \( \delta_o \) (and hence \( B_o \)) as well as \( \sigma_o^2 \) in cases \#1 and \#2 by arguments similar to those in Amemiya (1973).

If, however, a linear restriction, \( \delta_{10} = 1 \), must be imposed on \( \delta_o \), further calculations are required. In this case, define:

\[
(3.16a) \quad W(r) = [ y_{1}^{**r-1} x_{1}^{**r-2} : (r-1) y_{1}^{**r-2} ]^{T} N_2 \times K^{
\n}

(3.16b) \quad W(r) = [ y_{1}^{**r-1} x_{1}^{**r-2} : (r-1) y_{1}^{**r-2} ]^{T} N_2 \times K^{
\n}

(3.16c) \quad \gamma^{**}(r) = [ y_{1}^{**r} ]^{T} N_2 \times 1^{
\n}

Then, the UIV estimator of order \( r \) for \( y_o \) may be rewritten as:

\[
(3.17) \quad \hat{\gamma}(r) \equiv \left[ \begin{array}{c}
\hat{\gamma}(r) \\
\delta_o \\
\vdots \\
\sigma_o(r)
\end{array} \right]^{-1} = [W(r)]^{-1} \cdot [W(r) \gamma^{**}(r)]^{-1}.
\]

Suppose, however, that \( \delta_{10} = 1 \), i.e.,

\[
(3.18) \quad q^\prime Y_o = 1,
\]

where \( q^\prime = [10 \ldots 0] \). Then, we must proceed to the Restricted Instrumental

\[
\text{variable (KIV) estimator,}
\]

\[
(3.19) \quad \hat{\gamma}(r) = \gamma_o + [W(r) W(r)^{-1}] q [a (W(r) W(r)^{-1} q]^{-1}. (1-\delta_{10}^r)
\]

Note that in (3.19),

\[
(3.20) \quad q^\prime (W(r) W(r)^{-1} q = W_{11}(r),
\]
where $w^{ij}$ denotes the $(i, j)$ element of $\left[ \hat{w}(r), w(r) \right]^{-1}$, so that (3.19) reduces to:

$$
(3.21) \quad \hat{\gamma}_r = \gamma_0 + d(r)
$$

where

$$
(3.22) \quad d(r) = \frac{(1 - \delta(r))}{\omega^{11}} \cdot [w^{11} w^{21} \ldots w^{K*1}]
$$

It is therefore evident that:

$$
(3.23) \quad \hat{\gamma}_{10} = \delta_{10} + (1 - \delta_{10}) = 1
$$

and the restriction is imposed on the estimates, $\gamma_0$, for any $r > 2$.

**Type 2 (Cases where $z^{**}_1 < \infty$):**

In cases where $z^{**}_1 < \infty$, we have $f^{**}_1(z^{**}_1) > 0$ and the last term on the RHS of (3.8) does not vanish. Hence, the approach of Amemiya with $r=2$ must be modified. Since there is no more

$$
(3.24) \quad (\mu^{(r)}_{120} - z^{**}_1 \mu^{(r-1)}_{120}) = (\mu^{(r-1)}_{120} - z^{**}_1 \mu^{(r-2)}_{120}) \cdot \frac{x_i^{**}}{\delta \sigma^2}
$$

$$
+ ((r-1) \mu^{(r-2)}_{120} - (r-2) \mu^{(r-3)}_{120}) \cdot \sigma^2
$$
which is, again, linear in \( \delta_0 \) and \( \sigma_0^2 \) and eliminates the last term on the RHS of (3.8). Thus, for any \( r \geq 3 \),

\[
(3.25) \quad (y_{1}^{**r} - z_1^{**} y_{1}^{**r-1}) = [(y_{1}^{**r-1} - z_1^{**} y_{1}^{**r-2}) \cdot x_1^{**}'] \cdot \delta_0 \\
+ [(r-1) \cdot y_{1}^{**} y_{1}^{**r-2} - (r-2) z_1^{**} y_{1}^{**}(r-3)] \cdot \sigma_0^2 + \eta_i^{(r)}, \ i \in S_2,
\]

where now the disturbance, \( \eta_i^{(r)} \), is defined by:

\[
(3.26) \quad \eta_i^{(r)} = [y_{1}^{**r} - \mu_i^{(r)}] - (z_1^{**} + x_1^{**'}) \cdot \delta_0 \cdot [y_{1}^{**}(r-1) - \mu_i^{(r-1)}] \\
+ (z_1^{**} x_1^{**'} - \mu_i^{(r-2)}) \cdot (r-1) \sigma_0^2 + (r-2) \sigma_0^2 \cdot [y_{1}^{**}(r-3) - \mu_i^{(r-3)}]
\]

and, by convention, \( y_{1}^{**0} = \mu_1^{(0)} = 1 \). Clearly, \( E_0 \eta_i^{(r)} = 0 \) and \( \eta_i^{(r)} \) is bounded.

In the cases where \( \delta_1 \) is unrestricted (i.e., cases \#3, \#4 and \#5 of Table 1), by arguments similar to Amemiya (1973), the UIV estimator, (3.17), in which now

\[
(3.27a) \quad \hat{y}_i^{(r)} = [(y_{1}^{**r-2} - z_1^{**} y_{1}^{**r-3}) \cdot x_1^{**} \cdot (r-1) y_{1}^{**r-2} - (r-2) z_1^{**} y_{1}^{**r-3}]
\]

\[
(3.27b) \quad \hat{y}_i^{(r)} = [(y_{1}^{**r-1} - z_1^{**} y_{1}^{**r-2}) \cdot x_1^{**} \cdot (r-1) y_{1}^{**r-2} - (r-2) z_1^{**} y_{1}^{**r-3}]
\]

\[
(3.27c) \quad \hat{y}_i^{(r)} = [(y_{1}^{**r} - z_1^{**} y_{1}^{**r-1})]
\]

will be (weakly) consistent. Similarly, to impose the restriction \( \delta_{10} = 1 \),

(in cases \#8, \#11, \#12, \#13 and \#16) one may utilize the RIV estimator, (3.21)
and (3.22), where now definitions (3.27a) - (3.27c) replace (3.16a) - (3.16c), respectively.

Finally, we note that a general class of IV estimators for $\gamma_0$ (and, hence, $\delta_0$ and $\sigma^2_0$) may be derived from the relation:

$$
(3.28) \quad (\mu_{120}^{(r)} - z_1^{**s}\mu_{120}^{(r-s)}) = (\mu_{120}^{(r-1)} - z_1^{**s}(r-s-1)) \cdot x_1^{**} \frac{\delta}{\sigma_0}
$$

$$
+ ((r-1)\mu_{120}^{(r-2)} - (r-s-1)\mu_{120}^{(r-s-2)}) \cdot \frac{\sigma^2}{\sigma_0}
$$

for $r=2,3,4,...$ and $s=0,1,2,...,r-2$ in Type 1 problems, and $r=3,4,...$ and $s=1,2,...,r-2$ in Type 2 problems. Any feasible choice of $r$ and $s$ in our formulation preserves the linear relation in $\frac{\delta}{\sigma_0}$ and $\frac{\sigma^2}{\sigma_0}$, while eliminating the last term on the RHS of (3.9). Thus we may define our general, weakly consistent, initial estimator of $\gamma_0$ as:

$$
(3.29a) \quad \gamma_0 \equiv \begin{bmatrix} \gamma(r,s) \\ \delta_0 \\ \sigma^2_0 \\ \gamma_0 \end{bmatrix} = \begin{cases} 
\gamma(r,s) \\ \delta_0 \\ \sigma^2_0 \\ \gamma_0 
\end{cases}, \text{ if Type 1} \\
\gamma(r,s) \\ \delta_0 \\ \sigma^2_0 \\ \gamma_0, \text{ if Type 2}
$$

which, via the inverse of the appropriate transformation to (3.4), and appeal to Slutsky's Theorem, yields weakly consistent estimators of the original parameter vector, $\gamma_0$, which we shall represent as:

$$
(3.29b) \quad \theta(r,s) \equiv \begin{bmatrix} \gamma(r,s) \\ \delta_0 \\ \sigma^2_0 \\ \gamma_0 
\end{bmatrix}
$$

3.3. Improvements to the Initial Variance Estimator:

Although the IV estimator, $\sigma^2_0$, converges in probability as $N \to \infty$ to the true parameter value, $\sigma^2_0 > 0$, it is estimated (via the IV approach) as a
regression coefficient and may, therefore, take on negative values in finite samples.1/ In order to avoid the inadmissible values of \( \hat{\sigma}_o^2 \), we define the adjusted variance estimator:2/ 

\[
(3.30a) \quad \hat{\sigma}_o^2 (r,s) = \begin{cases} 
\frac{\hat{\sigma}_o^2 (r,s)}{\sigma_o^2} & \text{if } \frac{\hat{\sigma}_o^2 (r,s)}{\sigma_o^2} > 0 \\
\omega^2 (r,s) & \text{if } \frac{\hat{\sigma}_o^2 (r,s)}{\sigma_o^2} < 0
\end{cases}
\]

where

\[
(3.30b) \quad \omega^2 (r,s) = \frac{1}{N_2} \sum_{i \in S_2} \left( y_i - \frac{x_i}{\hat{\sigma}_0} \right)^2
\]

By construction, \( \hat{\sigma}_o^2 (r,s) \) is always positive and its weak consistency follows from the fact that \( \frac{\hat{\sigma}_o^2 (r,s)}{\sigma_o^2} + \sigma_o^2 > 0 \) as \( N \to \infty \).

Although the use of \( \hat{\sigma}_o^2 (r,s) \) guarantees that the initial variance estimator will always take on an admissible value, the estimator resulting from a particular choice of \( r \) and \( s \) in a finite sample still may not be "reliable," in the sense that, while positive, it may lie "far" from the true parameter value. In our experience, very small values of \( \hat{\sigma}_o^2 (r,s) \) are frequently encountered, with the result that subsequent ML algorithms may, on

1/ In practice, negative values of \( \hat{\sigma}_o^2 (r,s) \) for any \((r,s)\) choice are quite common—see the discussion of computational experience in section 6.2 below. Such values clearly violate Assumption I, i.e., lie outside \( \sigma > 0 \), and thus are infeasible.

2/ In some cases, it may even be preferable to employ a censored variance estimator in which \( \hat{\sigma}_o^2 (r,s) \) is observed as \( \hat{\sigma}_o^2 (r,s) \) when the latter exceeds some \( \eta > 0 \) and is \( \omega^2 (r,s) \) otherwise.

3/ We are indebted to Professor Marcello Pagano for this point.
occasion, either fail to converge from them at all or converge rather slowly—see Section 6.1 below. A plausible solution to this difficulty is the selective improvement of the adjusted variance estimator, conditional on the consistent initial coefficient estimates \( \hat{\theta}_0(\hat{z}_0) \), utilizing the likelihood function defined in (2.15) for truncated samples and (2.16) for censored samples. The conditional ML initial variance estimator is then defined by:

\[
(3.31) \quad \sigma^2_{o}(r,s) = \sup_{\sigma^2 > 0} \left\{ \log L^N(\sigma^2|\hat{\theta}_0), \quad M=T,C \right\}.
\]

Consistency and positivity of \( \sigma^2_{o}(r,s) \) in (3.31) follow from Assumption 1 and the consistency of \( \hat{\theta}_o(r,s) \). We also note that \( \sigma^2_{o}(r,s) \) should be asymptotically more efficient than either \( \sigma^2_{o}(r,s) \) or \( \sigma^2_{o}(r,s) \).

Further, in censored samples, since we are now able to utilize all of the information contained in the sample—including limit-point observations, this should result in a further gain in asymptotic efficiency relative to the use of the truncated sample. Various methods for calculation of the conditional ML initial variance estimator are discussed in Section 5.5 below.
4. Asymptotic Distribution Theory:

For the case of a singly-censored normal regression model, Amemiya (1973) has provided formal proofs that under Assumptions 1, 2 and 3:

(a) a root of the likelihood equations, \( \hat{\theta}^C \), is strongly consistent,
(b) \( \hat{\theta}^C \) is asymptotically normal, i.e.,

\[
\sqrt{N} (\hat{\theta}^C - \theta_0) \sim N \left( 0, \Sigma^C(\theta_0) \right),
\]

where \( \Sigma^C(\theta_0) \) is positive definite.

Our purpose here is to extend these results to the cases of singly- and doubly-truncated and doubly-censored normal linear regression models.

Strong consistency of the MLE (defined as a root of the likelihood equations, (3.1)) follows by a proof analogous to Amemiya's, provided we invoke, in addition, Assumption 4. Similarly, the asymptotic normality follows from Assumptions 1-4. It remains, therefore, to record the form of the asymptotic covariance matrices,

\[
(4.1) \quad \Sigma^{M}(\theta_0) = \lim_{N \to \infty} \mathbb{E} \left[ \frac{\partial^2 \log L_N(\theta)}{\partial \theta \partial \theta'} \right]^{-1},
\]

for use in practical applications of the doubly-truncated (M=T) and doubly-censored (M=C) cases.

It may be verified, from equations (2.20a) - (2.20c) and (2.22a) - (2.22c), that \( \Sigma^{M}(\theta_0)^{-1} \) has the general form,

\[
(4.2) \quad \Sigma^{M}(\theta_0)^{-1} = \lim_{N \to \infty} \mathbb{E} \begin{bmatrix}
\sum_{i=1}^{N} a_i^M X_i X_i' \\
\sum_{i=1}^{N} b_i^M X_i \\
\sum_{i=1}^{N} c_i^M X_i \\
\end{bmatrix}, \quad M=T,C,
\]
where, using the definition $g_{i j o} \equiv g_{i o}(z_{i j})$, $j=1, 2$, we find that $a^T_1$, $b^T_1$ and $c^T_1$ are the scalar constants:

\[(4.3a) \quad a^T_1 = - \frac{1}{\sigma^2_o} \cdot \left(1 - \left(u_{i20} g_{i10} - u_{i10} g_{i10}ight)^2\right)\]

\[(4.3b) \quad b^T_1 = \frac{1}{\frac{1}{2} \sigma^2_o} \cdot \left\{ \left(u_{i20}^2 + \sigma^2_o g_{i20} - \left(u_{i10}^2 + \sigma^2_o g_{i10}\right)\right) + \sigma^2_o (g_{i10} - g_{i10}) (u_{i20} g_{i20} - u_{i10} g_{i10})\right\}\]

\[(4.3c) \quad c^T_1 = - \frac{1}{2} \sigma^2_o \cdot \left\{ (u_{i20}^2 - \sigma^2_o g_{i20} - u_{i10}^2 g_{i10}) - \sigma^2_o (u_{i20} g_{i20} - u_{i10} g_{i10})\right\}\]

and

\[(4.4a) \quad a^C_1 = - \frac{1}{\sigma^2_o} \left\{ \sigma^2_o \cdot \left(\frac{\sigma^2_{i20}}{1 - F_{i20}} + \frac{\sigma^2_{i10}}{F_{i10}}\right) - \left(u_{i20} f_{i20} - u_{i10} f_{i10}\right)\right\}\]

\[(4.4b) \quad b^C_1 = \frac{1}{1 - \sigma^2_o} \left\{ \left(\sigma^2_{i20} + \sigma^2_{i10}\right) - \left(u_{i20}^2 + u_{i10}^2\right)\right\}\]

\[(4.4c) \quad c^C_1 = - \sigma^2_o \left\{ \left(u_{i20}^2 f_{i20} - u_{i10}^2 f_{i10}\right)\right\}\]
\begin{equation}
(4.4c) \quad c_l^i = - \frac{1}{4 \sigma_0^4} \left\{ \frac{3}{8} (u_{120} f_{120} - u_{110} f_{110}) - \frac{1}{\sigma_0^2} (u_{120} f_{120} - u_{110} f_{110}) \right. \\
\left. + \frac{u_{120} f_{120}^2}{1-F_{120}} + \frac{u_{110} f_{110}^2}{F_{110}} + 2 (F_{120} - F_{110}) \right\}.
\end{equation}

Results for the singly-truncated/censored cases may be obtained from the above by setting \( f_{110} = 0 \) and \( F_{110} = \left\{ \begin{array}{ll}
0 & \text{if } j = 1 \\
1 & \text{if } j = 2
\end{array} \right. \)
whenever \( z_{ij} = \left\{ \begin{array}{ll}
-\infty & , j = 1 \\
+\infty & , j = 2
\end{array} \right. \).

For all cases, in finite samples, \( E^M(\theta) \) may be consistently estimated by removing the limit sign from (4.2) and evaluating \( a_i^M, b_i^M \) and \( c_i^M \) at the

\begin{equation}
M = T, C.
\end{equation}

Similarly, Amemiya (1973) establishes in the singly-censored normal case ... that under general assumptions:

(c) the "second round" estimator, \( \theta_1^M \) (see equation (5.3) below), from a single Newton-Raphson iteration, starting from any weakly consistent initial estimate, is strongly consistent and has the same asymptotic distribution as the MLE.

It is easy to show that the same applies to all other truncated/censored cases discussed in this paper. In addition, (c) also applies to the Method of Scoring and Gauss-Newton algorithms (see Berndt, Hall, Hall and Hausman (1974)).
In the special case of the Probit model, setting $\sigma^2 = 1$ and

$$u_{i2o} = u_{i1o} = x_i \tilde{\beta}_0,$$

we have $\mathbb{E}^C(\tilde{\beta}_0)^{-1} = \lim_{N \to \infty} \frac{N}{N} \sum_{i=1}^{N} a_i^C \cdot x_i x_i^T$, in which

$$a_i^C = \frac{f_{1o}(0)}{F_{1o}(0) \cdot [1 - F_{1o}(0)]}.$$
5. Calculation of the M.L.E.:

In this and the subsequent section we consider four methods for the calculation of the M.L.E., \( \hat{\theta}^M \), of (2.17) for either truncated (M=T) or censored (M=C) samples. These are as follows:

(i) Newton-Raphson

(ii) Method of Scoring

(iii) Gauss-Newton

(iv) Expectation-Maximization

For each method we treat only the doubly-truncated/censored case, though algorithms for other model specifications can be readily obtained from the present discussion.

Let \( \hat{\theta}_2 \equiv \left[ \begin{array}{c} \hat{\theta}_2^1 \\ \hat{\theta}_2^2 \\ \hat{\theta}_2^3 \end{array} \right] \) denote the value of \( \theta \in \Theta \) in iteration \( \ell, \ell = 0, 1, 2, \ldots \)

For any function, \( q_1 = q(y_1^*; x_1, \Theta) \), let \( q_{12} = q(y_1^*; x_1, \hat{\theta}_2) \). Let \( E_2 \) denote the unconditional expectation operator relative to \( \hat{\theta}_2 \) and \( E_{12} \) denote the corresponding conditional expectation operator, given \( y_1^* \in Y_1 \).

Thus,

\[
(5.1a) \quad E_2 q_1 = \int_{-\infty}^{\infty} q(y_1^*; x_1, \Theta) \cdot f_{12} (y_1^*) \, dy_1^*
\]

and, using (2.22)) with the conditional density, we define

\[
(5.1b) \quad E_{12} q_1 = \int_{Y_1} q(y_1^*; x_1, \Theta) \cdot u_{12} (y_1^*) \, dy_1^*
\]

so that, symbolically, we have the customary relation,

\[
(5.2) \quad E_2 q_1 = F_{12} \cdot E_{12} q_1 + [F_{12} - F_{11}] \cdot E_{22} q_1 + [1 - F_{12}] \cdot E_{32} q_1
\]
Each of our algorithms may be represented in **canonical form** by the following **recursion formula**:

\[
\hat{\theta}_k^{m+1} = \hat{\theta}_k^m + \lambda_k^m \cdot A_k^m \cdot b_k^m,
\]

where $\lambda_k^m$ is a nonnegative scalar, $0 < \lambda_k^m < \infty$, which, in an "unmodified" algorithm, is equal to unity and, in a "modified" algorithm, may vary from iteration to iteration; $A_k^m$ is a $(K+1) \times (K+1)$ symmetric matrix and $b_k^m$ is a $(K+1)$-element vector. The superscript, $m=1, 2, 3,$ or 4, refers to the particular algorithm in subsection 5.m below.1/

**5.1. The Newton-Raphson Method:**

Amemiya (1973) proposes the use of the Newton-Raphson (N-R) algorithm for singly-censored (Tobit) models. In the present (doubly-truncated/censored) case this requires:

\[
A_k^1 = -\left[ \frac{3^2 \log L_N^M (\hat{\theta}_k^m)}{\partial \hat{\theta} \partial \hat{\theta}'} \right]^{-1}, \quad M = T, C,
\]

\[
b_k^1 = \left[ \frac{3 \log L_N^M (\hat{\theta}_k^m)}{\partial \hat{\theta}} \right], \quad M = T, C
\]

and $\lambda_k^1 = 1$, if in the unmodified form. Expressions for the elements of $A_k^1$ and $b_k^1$ have already been given in equations (2.19a) - (2.19b) and (2.20a) - (2.20c) for $M=T$, and by equations (2.21a) - (2.21b) and (2.22a) - (2.22c) when $M=C$.

---

1/ Each of these algorithms may be directly applied to the **Probit** ($M=C$) model, where $\sigma_k^2 \equiv 1$ for all $k$, $A_k^m = \hat{A}_k^m$ and $z_{11} = z_{12} = 0$. 

As noted by Amemiya (1973), the N-R algorithm has the advantage that under general conditions "... if the initial estimate, $\hat{\theta}_0$, is consistent and $\sqrt{N}(\hat{\theta}_0 - \theta_0)$ has a proper limit distribution, the second round estimate, $\hat{\theta}_1$, has the same asymptotic distribution as a consistent root of the normal equations". The disadvantage of the (unmodified) N-R method is that there is no guarantee that the sequence, $\{\hat{\theta}_k : k=0,1,2,\ldots\}$, will ever converge---much less to the root of (3.1) corresponding to the global maximum of the log-likelihood function.\(^1\)

5.2. The Method of Scoring:

The Method of Scoring (M-S), see, e.g., Rao (1965), is identical to the N-R method, except that the matrix of second partials of the log-likelihood in the latter is replaced by its expectation (relative to $\hat{\theta}_k$).

Thus, noting (4.1), (4.2), (4.3a) -- (4.3c) and (4.4a) -- (4.6c), we may define:

$$
(5.5a) \quad A_k^2 = - \left[ E_{\hat{\theta}} \frac{\partial^2 \log L^N(\theta)}{\partial \theta \partial \theta'} \right]^{-1}
$$

whereas,

$$
(5.5b) \quad b_k^2 = b_k^1
$$

and, if unmodified, $\lambda_k^2 = 1$. The unmodified M-S algorithm shares both the advantage and the disadvantage of the N-R method previously noted.

\(^1\) In practical applications---see section 6---the lack of convergence of the N-R algorithm, in our experience, is not uncommon, particularly in heavily censored samples.
5.3. The Gauss-Newton Method:

The Gauss-Newton (G-N) method employs:

\[ A_2^3 = \left[ \sum_{i=1}^{N} \frac{\partial \log P_i^M(y_i; \hat{\theta}_2)}{\partial \theta}, \frac{\partial \log P_i^M(y_i; \hat{\theta}_2)}{\partial \theta'} \right]^{-1} \]

and, once again,

\[ b_2^3 = b_2^1, \]

where \( P_i^M \) is the appropriate density function,

\[ P_i^M(y_i; \theta) = \begin{cases} g_i(y_i; \theta), & \text{if } M=T \\ h_i(y_i; \theta), & \text{if } M=C \end{cases} \]

and (in the unmodified case) \( \lambda_2^3 = 1 \). Since \( \frac{\partial \log L_N^M(\theta)}{\partial \theta} \) are given in (2.19a) and (2.19b) or (2.21a) and (2.21b),

respectively, except that \( \Sigma_{i=1}^N \) must be removed from the RHS.

Methods for choosing \( \lambda_2^3 \) in the modified G-N algorithm, which guarantee convergence to a local maximum of \( \log L_N^M(\theta) \), have been given—see subsection 5.6 below—by Hartley (1961) and Berndt, Hall, Hall and Hausman (1974). Further, if the likelihood function satisfies the usual "regularity conditions", the fact that, if \( \hat{\theta}_2 + \theta \), then

\[ \lim_{\lambda_2 \to \infty} A_2^3 - \left[ \frac{\partial^2 \log L_N^M(\theta)}{\partial \theta \partial \theta'} \right]^{-1} = \lim_{\lambda_2 \to \infty} A_2^2 \]
(Berndt, Hall, Hall and Hausman (1974)), permits either $A_2$, $A_2^2$ or $A_2^3$ evaluated at $\hat{\theta}^M = \lim_{k \to \infty} \hat{\theta}_k$, to serve as the asymptotic covariance matrix of $\sqrt{N} (\hat{\theta}^M - \theta_0^\circ)$, $M = T, C$. Our practice is to use $A_2^2$ for calculating the asymptotic covariance matrices of all of the methods, $m = 1, 2, 3$ and 4.

### 5.4. The Expectation-Maximization Method:

The so-called Expectation-Maximization (E-M) algorithm was originally proposed by Hartley (1958) to calculate the MLE for "grouped data" problems. It was subsequently extended to a general class of ML estimation problems involving various types of "incomplete data" by Dempster, Laird and Rubin (1977). A discussion of its application to the Tobit (singly-censored) and Probit normal regression models is given in Hartley (1976). Further, a general discussion of the use of the E-M algorithm in censored and truncated (non-regression) problems, as well as some general properties of the method, are also given in Dempster, Laird and Rubin (1977).

Both the doubly-truncated and doubly-censored normal regression models derive from the same underlying regression model for $y_1^*$ given by (2.1) and (2.2). Conceptually, consider a given set of $N$ values for the independent variables, $(x_{1i} : i = 1, \ldots, N)$, and, for each observation on $x_{1i}$, let the corresponding $y_{i}^*$ - value be generated by adding a random drawing $\varepsilon_i$ from the density $n(0, \sigma_0^2)$ to the true regression function $x_{1i}^T \beta_0$. We may write the log-likelihood function in the "complete data" case for a sample of size $N$, $\{y_{1i}^* , x_{1i}^*\}$, as:

$$
\log L_N^* (\theta ; y^*) = \sum_{i=1}^{N} \log f_i (y_{1i}^*) = \sum_{j=1}^{3} \sum_{i \in S_j}^{N} \log f_i (y_{1i}^*)
$$
where \( S_j^N \) has been defined above equation (2.13) and contains \( N_j \) observations, \( j=1,2 \) and 3. Clearly, if the \( \{ y_i^* \} \) were observed over their entire range, \( (-\infty, +\infty) \), then the ML estimates of \( \beta_0 \) and \( \sigma_0^2 \) for the "complete data" case would be obtained from the standard formulas:

\[
(5.10a) \quad \hat{\beta} = \left[ X' X \right]^{-1} \cdot \left[ X' \ y^* \right]
\]

and

\[
(5.10b) \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} s_i^2; \quad s_i^2 = (y_i^* - x_i' \hat{\beta})^2,
\]

where \( X = (x_i') \) and \( y^* = (y_i^*) \).

In the case of truncated samples from the "complete" index set, \( S^N \), representing \( N \) independent drawings from \( f_{10}(y_i^*) \), only the pairs, \( \{(y_i^*, x_i^*) : i \in S_2^N\} \), are actually available; and all sample information on the remaining drawings, \( \{(y_i^*, x_i^*) : i \in S_1^N \text{ or } i \in S_3^N\} \), may be viewed as having been discarded. Indeed, since only the \( N_2 \) observations in \( S_2^N \) are known, even the size, \( N \), of the original sample—much less the number of "missing" observations, \( N_1 \) or \( N_3 \)—is unavailable.

The censored case is somewhat different. Here, not only are the actual data pairs, \( \{(y_i^*, x_i^*) : i \in S_2^N\} \), known, but also the values of the regressors, \( \{(x_i^* : i \in S_j^N, j=1,3\} \), as well as the range of the unknown \( y_i^* \) values — \( y_i^* \leq z_{11} \) (if \( i \in S_1^N \)) or \( y_i^* > z_{12} \) (if \( i \in S_3^N \))—and, hence, the number, \( N_j \), of missing observations in each index subset, \( S_j^N, j=1,3 \)—are known. This additional information, relative to the truncated case, has considerable computational and asymptotic advantages.
Thus, each of these situations represents a particular type of "incomplete data" sample. In such instances, the E-M algorithm involves the iterative replacement of all "unobserved" values of \( y_i \) by their suitable conditional expectations, evaluated relative to the parameter values of the previous iteration (the Expectation step), and then use of the "complete data" formulas to obtain updates of the parameter estimates (the Maximization step). In many applications these two steps may be combined.

We shall now outline the actual computational steps required to implement the E-M algorithm on the doubly-truncated and doubly-censored normal regression models, as well as provide motivation for the method and sketch a proof of convergence in each case.

5.4.1. The Doubly-Censored Model:

We begin by defining the \( r^{th} \) conditional "central" moments, of \( y_i \), given \( y_i \varepsilon Y_j \) (and relative to an arbitrary \( \theta \)), by

\[
(5.11) \quad \xi_{ij}^{(r)} = E \left[ (y_i - x_i \theta)^r | y_i \varepsilon Y_j \right],
\]

for \( j = 1, 2 \) and 3. Then, using (2.25), with \( r = 1 \) and 2, we have:

\[
(5.12) \quad \xi_{ij}^{(1)} = \mu^{(1)}_{ij} - x_i \theta = \begin{cases} 
\frac{f_{11}}{F_{11}}, & \text{if } j=1 \\
- \sigma^2 \cdot \frac{f_{12} - f_{11}}{F_{12} - F_{11}}, & \text{if } j=2 \\
\sigma^2 \cdot \frac{f_{12}}{1 - F_{12}}, & \text{if } j=3
\end{cases}
\]

and
\[
(5.13) \quad \xi_{ij}^{(2)} = \left\{ \begin{array}{ll}
\sigma^2 \cdot (1 - \frac{u_{ij}^2 f_{ij}}{F_{ij}}), & \text{if } j = 1 \\
\sigma^2 \cdot (1 - \frac{u_{ij}^2 f_{ij} - u_{ij}^2 f_{ij}}{F_{ij} - F_{ij}}), & \text{if } j = 2 \\
\sigma^2 \cdot (1 + \frac{u_{ij}^2 f_{ij}}{1 - F_{ij}}), & \text{if } j = 3,
\end{array} \right.
\]

respectively. Note, as a check, that with \( q_1 = (y_1^* - x_1^* \beta)^y \), equation (5.2) implies that the corresponding first two unconditional moments are given by \( \xi^{(1)}_i = 0 \) and \( \xi^{(2)}_i = \sigma^2 \), respectively.

Using the definitions of the conditional densities, (2.23), the log-likelihood function, \( \log L_N^C(\theta) \) of (2.16), may be written as:

\[
(5.14) \quad \log L_N^C(\theta) = \sum_{i \in S_1} \int_{-\infty}^{Z_{i1}} \{ \log f_1(y_i^*) - \log u_{i1}(y_i^*) \} \cdot u_{i1}(y_i^*) dy_i^*

+ \sum_{i \in S_2} \log f_1(y_i^*)

+ \sum_{i \in S_3} \int_{Z_{i12}}^{\infty} \{ \log f_1(y_i^*) - \log u_{i3}(y_i^*) \} \cdot u_{i3}(y_i^*) dy_i^*

= \sum_{i \in S_1} E_1 \{ \log f_1 - \log u_{i1} \} + \sum_{i \in S_2} \log f_1 + \sum_{i \in S_3} E_3 \{ \log f_1 - \log u_{i3} \}.
\]

Then the formulas for the E-M algorithm can be viewed as being obtained by maximizing the "pseudo log-likelihood".
(5.15) \[ \Lambda^C(\hat{\theta}_1 | \hat{\theta}_2) = \sum_{i \in S_1} E_{1i} \{ \log \gamma_i - \log u_{i1} \} + \sum_{i \in S_2} \log f_i \]
\[ + \sum_{i \in S_3} E_{3i} \{ \log f_i - \log u_{i3} \}, \]

with respect to \( \hat{\theta}_2 \), obtained by fixing \( u_{i1} \) and \( u_{i3} \) in (5.14) at \( \hat{\theta}_2 \), while leaving \( \hat{\theta}_1 \) present within the \( f_i \) free to vary. Thus, we seek solutions, \( \hat{\theta}_{2+1} \), defined for \( M=C \), implicitly via:

(5.16) \[ \Lambda^M(\hat{\theta}_{2+1} | \hat{\theta}_2) = \sup_{\theta \in \Theta} \{ \Lambda^M(\theta | \hat{\theta}_2) \}. \]

It may easily be verified that the resulting "pseudo likelihood equations" are given by:

(5.17a) \[ \frac{\Delta^C(\theta | \hat{\theta}_2)}{\Delta \theta} = \frac{1}{\sigma^2} \left\{ \sum_{i \in S_1} \int_{-\infty}^{\infty} (y_i^* - x_i^* \hat{\theta}) \cdot u_{i1i}(y_i^*) dy_i^* x_i^* \right\} + \sum_{i \in S_2} \int_{-\infty}^{\infty} (y_i^* - x_i^* \hat{\theta}) \cdot u_{i2i}(y_i^*) dy_i^* x_i^* = 0 \]

and

(5.17b) \[ \frac{\Delta^C(\theta | \hat{\theta}_2)}{\Delta \sigma^2} = \frac{1}{2\sigma^4} \left\{ \sum_{i \in S_1} \int_{-\infty}^{\infty} [(y_i^* - x_i^* \hat{\theta})^2 - \sigma^2] \cdot u_{i1i}(y_i^*) dy_i^* \right\} + \sum_{i \in S_2} [(y_i^* - x_i^* \hat{\theta})^2 - \sigma^2] + \sum_{i \in S_3} [(y_i^* - x_i^* \hat{\theta})^2 - \sigma^2] \cdot u_{i3i}(y_i^*) dy_i^* = 0 \]
Then, using equations (5.11) - (5.13), the E-M algorithm is obtained by setting \( \hat{\theta}_{k+1} \) equal to the unique solution to the (K+1)-equation system, (5.17a) - (5.17b), and is defined by:

\[
(5.18a) \quad \hat{\theta}_{k+1} = [X'X]^{-1} \cdot [X'Y] = \mathbb{R}_{Y}^{C} ; \quad R = [X'X]^{-1} \cdot X
\]

and

\[
(5.18b) \quad \sigma_{k+1}^2 = \frac{1}{N} \sum_{i=1}^{N} s_{i\ell}^{2C}
\]

where

\[
(5.19) \quad Y_{\ell}^{C} = [y_{i\ell}^{C}] , \quad y_{i\ell}^{C} = \begin{cases} u_{i1\ell}^{(1)} , & \text{if } i \in S_1 \\ y_{i\ell}^{*} , & \text{if } i \in S_2 \\ u_{i13\ell}^{(1)} , & \text{if } i \in S_3 \\ \end{cases}
\]

and

\[
(5.20) \quad s_{i\ell}^{2C} = \begin{cases} \xi_{11\ell}^{(2)} - 2[y_{i\ell}^{C} - \hat{\theta}_{k+1} - \hat{\theta}_{k}] \cdot [y_{i\ell}^{C} - \hat{\theta}_{k+1} - \hat{\theta}_{k}]^2 , & i \in S_1 \\ \xi_{13\ell}^{(2)} - 2[y_{i\ell}^{C} - \hat{\theta}_{k+1} - \hat{\theta}_{k}] \cdot [y_{i\ell}^{C} - \hat{\theta}_{k+1} - \hat{\theta}_{k}]^2 , & i \in S_3 \\ \end{cases}
\]

Several points are worth noting. First, it may easily be verified that, if \( \hat{\theta}_{k+1} = \hat{\theta}_{k} = \hat{\theta} \) and \( \mathbb{R}=C \), then both

\[1/ \quad \text{Note that for the Probit model, equation (5.18a) is sufficient to define the EM algorithm.} \]
\[ \lambda^M(\theta; \theta^*) = \log L_N^M(\theta) \]

and

\[ \frac{\partial \lambda^M(\theta; \theta^*)}{\partial \theta} = \frac{\partial \log L_N^M(\theta)}{\partial \theta} \]

consider the sequence of solutions to the "pseudo-likelihood equations," (5.22a and 5.17b), defined by \( \hat{\theta}_l; l=0,1,2, \ldots \), converges, then it converges to a solution to the original likelihood equations, (3.1) with \( M=C \). Second, by minimizing equations (5.16) and (5.21a) with \( \theta = \hat{\theta}_2 \), we have for \( M=C \) and \( \tilde{N} \),

\[ \log L_N^M(\hat{\theta}_2) = \Lambda^M(\hat{\theta}_2; \hat{\theta}_2) < \Lambda^M(\hat{\theta}_{l+1}; \hat{\theta}_l) \]

implying

\[ \Lambda^M(\hat{\theta}_{l+1}; \hat{\theta}_l) < \Lambda^M(\hat{\theta}_{l+1}; \hat{\theta}_{l+1}) = L_N^M(\hat{\theta}_{l+1}) \]

The inequality (5.22b), holds by virtue of the fact that \( \hat{\theta}_{l+1} \) and \( \hat{\theta}_l \) are precisely the solutions to the calculus-of-variations problem of maximizing the "pseudo log-likelihood function,"

\[ F_{11}^M(\hat{\theta}_2, \hat{\theta}_{l+1}) = \Sigma_{i \epsilon S_1} E_1(\log f_{1, l+1}^* - \log u_{11}) + \Sigma_{i \epsilon S_2} \log f_{1, l+1}^* + \Sigma_{i \epsilon S_3} E_3(\log f_{1, l+1}^* - \log u_{12}) \]

subject to the \( N_j \)-vector of functionals, \( v_{ij} = [u_{ij}(y_i)] \), \( j=1,3 \), subject to the so-called "isoperimetric conditions,"

\[ \int_{y_{ij}} u_{ij}(y_i)^* dy_i^* = 1 \quad , \quad j=1,3 \]
Thus, (5.22a) and (5.22b), together, imply that both \( \{ \log L_N^{M}(\hat{\theta}_k) \} \) and \( \{ \Lambda_N(\hat{\theta}_{k+1} | \hat{\theta}_k) \} \) are monotone increasing sequences, and uniformly bounded from above (due to Assumptions 2 and 4). Also, by (5.21a), \( \log L_N^{M}(\theta) \) and \( \Lambda(\theta | \theta) \) have all stationary points in common. Finally, since \( \log L_N^{M}(\theta) \to -\infty \) as \( \| \theta \| \to \infty \), \( \sigma^2 \to 0 \) or \( \sigma^2 \to -\infty \), it follows that the sequence of E-M parameter points, \( \{ \hat{\theta}_k : k = 0, 1, \ldots \} \), always remains within a bounded space. Thus, \( \{ \Lambda_N(\hat{\theta}_{k+1} | \hat{\theta}_k) \} \) must converge to a unique limit, which, in turn, implies that \( \{ \hat{\theta}_k \} \) has at least one point of accumulation and that hence, a subsequence, \( \bar{j} \), converges to that limit point,

\[
\lim_{k \to \infty} \hat{\theta}_k^{(1)} = \hat{\theta}^M,
\]

where \( \hat{\theta}^M \) is a solution to the original likelihood equations associated with a (local) maximum. Hence, convergence to, at least, a local maximum is guaranteed.

At a computational level, we note that \( \hat{\beta}_{k+1} \) is determined as a least squares regression of \( y_2^C \) on \( X \), where \( y_2^C \) is defined as the actual \( y_2^* \)-value (when \( i \in S_2^N \)), and by the expected value of \( y_2^* \), given knowledge of the \( X_2^* \)-value and the information that \( y_2^* \in Y \), evaluated relative to \( \hat{\theta}_1 \), for \( i \in S_2^N \), \( j = 1 \) or \( 3 \). The latter elements of the vector, \( y_2^C \), must therefore be updated each iteration, whereas the matrix \( R \) is invariant with \( \ell \). Hence, once \( \hat{\theta}_k \) has been initially calculated, further matrix inversion in iterations, \( \ell = 1, 2, \ldots \), is avoided, and \( \hat{\beta}_{k+1} \) is simply calculated as \( R_{yx}^{-1} \).

\[1/ \]

In applications where the number of regressors, \( K \), is large (see section 6), avoiding repeated inversion of the \( X \times K \) matrix, \( [X'X] \), saves substantial computation time. Thus, even though the number of E-M iterations required for convergence may be larger than that required by the N-R, N-S and G-N methods, the reduction in "average time per iteration" may lead to dominance of the E-M algorithm over its competitors in terms of total computation time.
Turning now to completion of the algorithm via $\hat{\sigma}^2_{k+1}$ of (5.18b), with the $(s_{1k})$ of (5.10), we again note the similarity of the E-M formula with that of (5.10b) for $\hat{\sigma}^2$ in the "complete data" case. Clearly, $\hat{\sigma}^2_{k+1}$ consists of the sum of squared residuals, $(y_1^* - \mathbf{x}' \hat{\theta}_k)^2$, for all $i \in S_2$, and the expected value of such squared residuals, given $y_1^* \epsilon Y_j^*$, relative to $\hat{\theta}_k$, for all $i \in S_j$, $j=1$ or 3. Hence, $\hat{\sigma}^2_k$ is always positive for any $k$—a property not shared with unmodified versions of the N-R, M-S and G-N algorithms.

5.4.2. The Doubly-Truncated Model:

In contrast, for the case of the doubly-truncated model, it is suggestive to rewrite the log-likelihood function, $\log L_N^T(\theta)$, as follows:

\begin{equation}
\log L_N^T(\theta) = \sum_{i \in S_2} \left( \log f_1(y_1^*) - \frac{z_{12}}{Z_{11}} \left[ \log f_1(y_1^*) - \log u_{12}(y_1^*) \right] - \frac{1}{u_{12}}(y_1^*)dy_1^* \right) = \sum_{i \in S_2} \left( \log f_1 - E_2[\log f_1 - \log u_{12}] \right).
\end{equation}

Thus, by analogy with the censored case, the E-M algorithm is obtained by maximizing:

\begin{equation}
\Lambda^T(\hat{\theta} | \hat{\theta}_k) = \sum_{i \in S_2} \left( \log f_1 - E_2[\log f_1 - \log u_{12}] \right)
\end{equation}

with respect to $\hat{\theta}$, holding the functions $(u_{12})$ of (5.14) fixed at $\hat{\theta}_k$.

Again, the solution for $\hat{\theta}_{k+1}$, which satisfies (5.16) with $M=T$, is unique and defined by the iteratively-reweighted least squares algorithm:

\begin{equation}
\hat{\theta}_{k+1} = \left[ X' \left( F_{2k} - F_{1k} \right)^{-1} X \right]^{-1} \cdot \left[ X' \left( F_{2k} - F_{1k} \right)^{-1} X \right]^T
\end{equation}
and the expected residual sum of squares,

\begin{equation}
\hat{\sigma}_{2+1}^2 = \frac{1}{\sum_{i \in \mathcal{S}'} (F_{i;2} - F_{i;1})^{-1} \cdot \sum_{i \in \mathcal{S}_2} s_{i;i}^2}
\end{equation}

where

\begin{equation}
F_{2;2} - F_{1;2} = \text{Diag}[F_{i;2} - F_{i;1}],
\end{equation}

\begin{equation}
\nu_T = [y_T] - F_{1;2} \cdot u_{1;2} + [F_{2;2} - F_{1;2}] \cdot \nu + [I - F_{2;2}] \cdot u_{3;2},
\end{equation}

\begin{equation}
\nu_j = [u_{j}] = E_{j;2} \nu, \quad j = 1 \text{ or } 3,
\end{equation}

and

\begin{equation}
s_{1;2}^2 = \frac{1}{F_{i;1} - F_{i;2}} \cdot (F_{1;1} \cdot E_{1;2}(y^*_1 - \hat{x}^*_1 \hat{B}_{2;1}^r)^2 + [F_{1;2} - F_{1;1}] \cdot (y^*_1 - \hat{x}^*_1 \hat{B}_{2;1}^r)^2
\end{equation}

\begin{equation}
+ (1 - F_{1;2}) \cdot E_{3;2}(y^*_1 - \hat{x}^*_1 \hat{B}_{2;1}^r)^2)
\end{equation}

which, using (5.1), (5.2) and (5.11) – (5.13), may be rewritten as:

\begin{equation}
s_{1;2}^2 = (y_1^* - \hat{x}^*_1 \hat{B}_{2;1}^r)^2 + \frac{1 - (F_{1;2} - F_{1;1})}{F_{1;2} - F_{1;1}} \cdot \left(\hat{\sigma}_2^2 + (\hat{x}_1^* (\hat{B}_{2;1}^r - \hat{B}_2)) \right)
\end{equation}

\begin{equation}
+ \frac{\hat{\sigma}_2^2}{F_{1;2} - F_{1;1} \cdot (u_{1;2} \hat{\ell}_{1;2} - u_{1;1} \hat{\ell}_{1;1})}
\end{equation}

\begin{equation}
- 2 \left[ \hat{x}_1^* \hat{B}_{2;1}^r \hat{B}_2 \right] \cdot (\hat{\ell}_{1;2} - \hat{\ell}_{1;1})
\end{equation}

By arguments similar to those in the doubly-censored case, it can be shown that (5.21a), (5.21b), (5.22a), (5.22b) and (5.24) also hold with \( N \rightarrow T \).

Thus the E-M algorithm in the doubly-truncated case, by iteratively maximizing \( \Lambda^T(\theta | \hat{\theta}_2) \) with respect to \( \theta \), and defining the solution as \( \hat{\theta}_{2+1} \),
yields a sequence, \( \{ \hat{y}_k : k=0,1, \ldots \} \), containing a subsequence which converges to a solution of (3.1) with \( N = 1 \) and yields a (local) maximum of \( \hat{L}^*_N(Y) \).

It remains to provide an interpretation for the E-M estimation method and to comment on the computations required. Consider the arbitrary \( i \)-th observation, \( i \in S_2 \), which may be viewed as a random drawing from the (conditional) p.d.f.,

\[
\pi_{i20}(y_1^*) = \frac{f_{10}(y_1^*)}{f_{i20} - f_{i10}}.
\]

Note that, relative to \( \theta_0 \),

\( y_1^* \) has the (conditional) expectation,

\[
(5.33) \quad E_{20} y_1^* = \mu_{i20} = x_1 \theta_0 + \sigma_0^2 f_{i20} - f_{i10} / f_{i20} - F_{i10}.
\]

Further, if we return to the conceptual experiment associated with the "complete data" model, then corresponding to the \( i \)-th regressor vector, \( x_1 \), the probability that an untruncated \( y_1^* \) will obtain within \( Y_j \), \( j=1,2 \) or 3, is given by:

\[
(5.34) \quad Pr(y_1^* \in Y_j) = \begin{cases} 
F_{i10} & \text{if } j = 1 \\
[F_{i20} - F_{i10}] & \text{if } j = 2 \\
[1 - F_{i20}]^* & \text{if } j = 3
\end{cases}
\]

for \( k=1,2, \ldots, N \). Since our truncated sample arises from discarding all pairs, \( (y_1^*, x_1) \), except for the \( N_2 \) observations satisfying \( y_1^* \in Y_2 \), if we knew the value of \( \theta_0 \), our estimate of the missing \( y_1^* \)-value within \( Y_j \) would be the conditional expectation of \( y_1^* \), given knowledge of \( x_1 \) and given that \( y_1^* \in Y_j \), obtained by evaluating (2.25) at \( \theta_0 \), for \( j=1 \) or 3. Further, the probability of this event is given by (5.34) — as \( F_{i10} \) if \( j=1 \) and as \( [1 - F_{i20}]^* \) if \( j=3 \)— whereas a particular observation, \( y_1^* \in Y_2 \), would occur with
probability \[ \{ F_{120} - F_{110} \} \]. Finally, since of all of the \( N_j \)s, only \( N_2 \) is known, the "missing" number of observations in \( \gamma_j \) associated with \( \{ x_i : i = 1, \ldots, N_2 \} \), would be estimated as:

\[
\hat{N}_{jo} = \begin{cases} 
\sum_{i \in S_2^N} \frac{F_{110}}{F_{120} - F_{110}} , & \text{if } j = 1, \\
N_2 & , \text{if } j = 2, \\
\sum_{i \in S_2^N} \frac{1 - F_{120}}{F_{120} - F_{110}} , & \text{if } j = 3.
\end{cases}
\]

(5.35)

With these preliminaries in hand, it is now easy to motivate the E-M algorithm. Suppose our current estimate of \( \hat{\theta}_2 \) after \( \ell \) iterations is \( \hat{\theta}_2^\ell \). Then, we see that, insofar as the functionals \( U_{1j}^\ell \), \( j = 1, 2 \) and 3, are fixed when \( \hat{\theta}_2^\ell \) is given, the "pseudo log-likelihood," \( \Lambda^{T}(\hat{\theta} | \hat{\theta}_2^\ell) \) of (5.26), may be rewritten more instructively (using (5.2)) as:

\[
(5.37) \quad \Lambda^{T}(\hat{\theta} | \hat{\theta}_2^\ell) = \sum_{i \in S_2^N} \frac{1}{F_{120} - F_{110}} \left\{ \frac{F_{112} F_{122} (\log f_1^\ell - E_{12} \log f_1^\ell)}{F_{122} - F_{112}} + [F_{122} - F_{112}] (\log f_1^\ell - E_{12} \log f_1^\ell) + [1 - F_{122}] E_{32} (\log f_1^\ell - E_{32} \log f_1^\ell) \right\}
\]

The last term on the RHS of (5.37) is independent of \( \hat{\theta} \) with fixed \( U_{12}^\ell \) and can therefore be ignored. The denominator has the effect of "blowing up" the sample size (via (5.36)) from \( N_2 \) to an estimated "complete sample" of \( \hat{N}_2 = \sum_{i \in S_2^N} (F_{122} - F_{112})^{-1} \) "pseudo observations" on the deviations,
(log f_{1j} - E_{j2} \log f_1), each weighted by the probability that y_1^* \in Y_1, and with the "incomplete" portions of the original log-likelihood (when j=1 or j=3) replaced by their expectations, E_{j2}(\log f_{1j} - E_{j2} \log f_1), which are identically zero.

It follows that the value of \theta which partially maximizes \Lambda(\theta | \theta_2) is obtained by the weighted least squares regression of the "pseudo dependent variable", \underline{x}_{12}^T, defined as the suitably weighted average of the observed dependent variable, \underline{x}_1, within Y_2 and the conditional expectations of the missing values, \mu_{12}^{(1)} and \mu_{32}^{(1)}, within Y_1 and Y_3, with "weights," (F_{122} - F_{112}) and (1 - F_{122}), respectively — all evaluated relative to \theta_2, and requiring weighting factors for each "pseudo-observation," (y_{1i}, \underline{x}_1), given by (F_{122} - F_{112})^{-1}. Since the requisite weighting matrix, (F_{22} - F_{1x})^{-1}, within [X'(F_{2x} - F_{1x})^{-1}X]^{-1}, requires updating in each iteration, a major computational advantage of the E-M algorithm, exhibited previously when M=C, is not obtained when M=T. Further, as one might expect, the variance, \hat{\sigma}^2_{k+1}, is obtained as the weighted sum of squared "residuals," \hat{y}_1^* - \hat{x}_1^* \hat{\beta}_{k+1}^2 --using the actual values, weighted by (F_{122} - F_{112}) = Pr[y_1^* \in Y_2], and the expected values, E_{j2}(y_{1j} - \hat{x}_{1j} \hat{\beta}_{k+1}^2), weighted by F_{112} (j=1) and (1 - F_{122}) (j=3), respectively — all divided by the number of "pseudo-observations", N_{k}. Again, \hat{\sigma}^2_{k+1} > 0 for

5.4.3. The Canonical Form:

To complete our discussion of the E-M algorithm, it remains to show that (5.18a) - (5.18b) and (5.27a) - (5.27b) can be represented in the form of equation (5.3). To this end note that:
\[ A_2^4 = \begin{cases} 
\left( X'X - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ N \end{bmatrix} \right)^{-1}, & \text{if } M = C, \\
X'(F_{2l}^\ell - F_{1l}^\ell)^{-1}X - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ N \end{bmatrix}, & \text{if } M = T, 
\end{cases} \]

and

\[ b_2^4 = \begin{cases} 
\left( X'X - \begin{bmatrix} C \\ \vdots \\ \Sigma_{i=1}^C q_{i\ell} \end{bmatrix} \right)^{-1}, & \text{if } M = C, \\
X'(F_{2l}^\ell - F_{1l}^\ell)^{-1}E_{2l}^\ell - \begin{bmatrix} \Sigma_{i \in S_2^N} q_{i\ell} \end{bmatrix}, & \text{if } M = T, 
\end{cases} \]

where

\[ E_{2l}^C = \{ p_{1l}^C \}, \quad p_{1l}^C = \begin{cases} 
\frac{-\sigma_{l}}{\hat{f}_{1l}^\ell}, & \text{if } i \in S_1^N, \\
\frac{\hat{f}_{1l}^\ell}{\sigma_{l}}, & \text{if } i \in S_2^C, \\
\frac{\hat{f}_{1l}^\ell}{1 - \hat{f}_{1l}^\ell}, & \text{if } i \in S_2^C, 
\end{cases} \]

\[ D_{2l}^T = \left[ F_{2l}^\ell - F_{1l}^\ell \right]^* (X^* - \hat{X}B_{\ell}) + \frac{A^2}{\sigma_{\ell}} \cdot \left( \hat{X}_{2l}^\ell - \hat{X}_{1l}^\ell \right) \]
\[
q_{12}^{C} = \begin{cases} 
-\sigma_{\hat{e}_{\hat{e}_{12}}}^2 \frac{u_{12} \hat{e}_{12}}{F_{12}} - 2[U_{1}(\hat{B}_{k+1} - \hat{B}_{k})]^2 \xi(1) + [X_{1}(\hat{B}_{k+1} - \hat{B}_{k})]^2, & \text{if } \hat{e}_{12} \in S_1^N \\
(y_{1} - \hat{x}_{12} \hat{B}_{k+1})^2 - \sigma_{\hat{e}_{\hat{e}_{12}}}^2, & \text{if } \hat{e}_{12} \in S_2^N \\
\sigma_{\hat{e}_{\hat{e}_{12}}}^2 \frac{u_{12} \hat{e}_{12}}{1 - F_{12}}, & \text{if } \hat{e}_{12} \in S_3^N 
\end{cases}
\]

and

\[
q_{12}^{T} = \frac{2\hat{e}_{12}}{s_{12}^T_{\hat{e}_{12}}} - \frac{\sigma_{\hat{e}_{\hat{e}_{12}}}^2}{[F_{12} - F_{11\hat{e}_{12}}]}
\]

in which \( \hat{B}_{k+1} \), defined by (5.18a), is inserted into \( q_{12}^{C} \) of (5.42) and \( \hat{B}_{k+1} \), defined by (5.27a), is inserted into \( s_{12}^{T} \) of (5.43) to eliminate the artificial dependence upon \( \hat{B}_{k+1} \).

5.5. Computation of the Improved Initial Variance Estimator

In Section 3.3 above we have described an improved initial variance estimator \( \sigma^{2}(r,s) \), which was derived by maximizing the likelihood function \( L_{N}(\theta) \), conditional upon the consistent initial estimates of the regression coefficients, \( \hat{\beta}(r,s) \). It remains to sketch the method of computation for each of the ML algorithms noted above.

Using the general recursion formula for each ML iterate given in 5.3, partition the \((K+1)\times(K+1)\) matrix, \( A_{k}^{m} \), and the \((K+1)\)-vector, \( b_{m}^{m} \), as:

\[
A_{k}^{m} = \begin{bmatrix} a_{12}^{m} & \ldots & a_{12}^{m} \\
\vdots & \ddots & \vdots \\
a_{12}^{m} & \ldots & a_{22}^{m} \\
\end{bmatrix} \quad \text{and} \quad b_{m}^{m} = \begin{bmatrix} b_{12}^{m} \\
\vdots \\
\vdots \\
b_{22}^{m} \\
\end{bmatrix}
\]

where the scalars, \( a_{12}^{m} \) and \( b_{22}^{m} \), refer to the parameter, \( \sigma_{\hat{e}_{\hat{e}_{12}}}^2 \).
Then the \((L+1)\) - iterate of \(\sigma^2_o(r,s)\) is given by:

\[
\sigma^2_{L+1} = \sigma^2_L + \lambda_L^m \cdot a_{22L} \cdot b_{2L}^m,
\]

and the conditional ML initial variance estimator is defined by:

\[
\sigma^2_o(r,s) = \lim_{L \to \infty} \sigma^2_L,
\]

where \(\lambda_L^m\) is an appropriately chosen scalar—equal to unity in the unmodified approach. Our experience (see sections 6.1 and 6.2) is that the vector,

\[
\sigma_o = \begin{bmatrix}
\sigma^2_o(r,s) \\
\sigma^2_o \\
\vdots \\
\sigma^2_o(r,s)
\end{bmatrix},
\]

provides excellent initial estimates of \(\sigma_o\), and that the "cost of adjustment" in moving from \(\lambda_o\) (of (3.29b)) to \(\lambda_o\) (of (5.47)) is small.

5.6. Modified Algorithms:

Our previous discussion has been restricted to the unmodified versions of the \(\lambda\)-\(\nu\), \(\lambda\)-\(\xi\), \(\xi\)-\(\nu\) and \(\xi\)-\(\nu\) algorithms, in which \(\lambda_L^m\) is fixed for \(L = 0, 1, \ldots\). In the present subsection, we consider modified versions of these algorithms, in which \(\lambda_L^m\) varies over the non-negative domain at each iteration. Such modifications are usually justified on the grounds that they either:

(a) convert an otherwise potentially divergent unmodified algorithm into one with guaranteed convergence to a stationary point, and/or

(b) improve the rate-of-convergence to a stationary point, relative to that of the unmodified form.
We have already noted that, of the unmodified versions of the four algorithms, only the E-M is guaranteed to converge to a stationary point of \( \log L_N^M \). We now examine procedures for modification of \( \lambda^m_\xi \) and the consequences with respect to (a) and/or (b), above.

Berndt, Hall, Hall and Hausman (1974) have provided an "existence theorem" for the choice of \( \lambda^m_\xi \), which applies to all "gradient-type" methods (i.e., methods, such as the N-R, M-S and G-N, in which \( \frac{d}{d \theta} \log L_N^M(\hat{\theta}_N) \)), provided that the following conditions are satisfied:

(i) \( L_N^M(\theta) \) is twice-differentiable, and defined over a compact, upper-contour parameter space, \( \Theta \),

(ii) \( A^m_\xi \) of (5.3) satisfies the restriction,

\[
\frac{\frac{\partial^2}{\partial \theta^2} A^m_\xi}{\frac{\partial^2}{\partial \theta^2} \frac{\partial}{\partial \theta}} > \alpha > 0,
\]

(5.48)

where \( \alpha \) is a pre-assigned, positive constant less than unity.

The BHNN modification is to define \( \lambda^m_\xi \), at each iteration, by the rule:

\[
\lambda^m_\xi = \begin{cases} 
\lambda^m_\xi, & \text{if } Q^m_\xi(1) < \delta \\
1, & \text{if } Q^m_\xi(1) > \delta 
\end{cases}
\]

(5.49)

where \( \delta \) is a pre-assigned constant in the interval, \( (0, \frac{\pi}{2}) \), \( Q^m_\xi(\lambda) \) is the criterion function,

\[
Q^m_\xi(\lambda) = \frac{\log L_N^M(\hat{\theta}^m_\xi + \lambda \cdot A^m_\xi \frac{\partial}{\partial \theta}) - \log L_N^M(\hat{\theta}^m_\xi)}{\lambda \cdot \frac{\partial^2}{\partial \theta^2} A^m_\xi \frac{\partial}{\partial \theta}}.
\]

(5.50)
and \( \lambda^* \) is a \( \lambda \)-value satisfying the inequality,

\[(5.51) \quad 0 < Q^m_{\lambda^*} < 1 - \delta. \]

They show that, under condition (i), a \( \lambda^* \) satisfying (5.51) will always exist, and that, under conditions (i) and (ii), the sequence \( \{R^m_{\lambda^*}\} \) of (5.3) converges to a stationary point of \( \log L_N^{\lambda}(p) \) at which \( \lim_{\lambda \to \infty} b^m_{\lambda^*} = 0 \). Finally, BHNNH show that these results always apply to the so-modified G-N algorithm, but not necessarily to the N-R or M-S algorithms. In the present case, since the negative of the matrix of second partials, \( A_{\lambda^*}^2 \), or its expectation, \( A_{\lambda^*}^2 \), cannot be guaranteed to be positive definite over the entire parameter space, \( \Theta \) (see Amemiya (1973)), guaranteed convergence cannot be claimed for the so-modified N-R and M-S algorithms.

While the BHNN results establish the existence of such a \( \lambda^* \) satisfying (5.51), they do not provide an explicit algorithm for either its feasible or optimal choice. Further, optimal choice of \( \lambda^m_{\lambda} \) at each iteration to maximize \( Q^m_{\lambda}(\lambda) \), while guaranteeing convergence, may impose an excessive computational burden (see Powell (1971)). Finally, all of our algorithms—whether modified or not—can only guarantee convergence to a stationary point. To insure against convergence to a local maximum, our development of simulation estimates (GUMS), and subsequent variance adjustments (given in section 3), provide protection in "asymptotic" samples.

A previous discussion of an operational method for modification of the G-N algorithm has been given by Hartley (1961). He proposes that \( \lambda^m_{\lambda} \) be chosen as the value, \( \lambda^* \), which maximizes the numerator of the BHNN criterion function over the unit-interval, i.e., instead of \( Q^m_{\lambda}(\lambda) \), consider maximization of:
\[ S^m_\lambda(\lambda) = \log N(\tilde{S}^m_\lambda + \lambda \cdot A^m_\lambda b^m_\lambda) \]

over \(0 < \lambda < 1\). In practice, however, Hartley recommends maximizing the parabolic approximation to \( S^m_\lambda(\lambda) \), supported at the points \( \lambda = 0, \frac{1}{2} \) and 1, respectively.

The Hartley modified G-N algorithm at iteration \( \ell \), therefore, consists of the following steps:

**Initialization Step (L.0):** Set \( j=1 \) and the support points \( \lambda_{\ell j1} = 0, \lambda_{\ell j2} = \frac{1}{2} \) and \( \lambda_{\ell j3} = 1 \).

**Step (L.1+j):** Evaluate \( S^m_\lambda(\lambda) \) at the three support points, \( \lambda = \lambda_{\ell j1}, \lambda_{\ell j2} \) and \( \lambda_{\ell j3} \), respectively, and solve for the coefficients,

\[ a^m_{\ell j} = [a_{\lambda j1}, a_{\lambda j2}, a_{\lambda j3}]^T, \]

such that the parabolic approximation,

\[ S^m_\lambda(\lambda) = a_{\lambda j1} + a_{\lambda j2} \lambda + a_{\lambda j3} \lambda^2, \]

is identical to \( S^m_\lambda(\lambda) \) of (5.52) at the three support points. Thus, letting

\[ S^m_{\lambda j} = [S^m_\lambda(\lambda_{\ell j1}), S^m_\lambda(\lambda_{\ell j2}), S^m_\lambda(\lambda_{\ell j3})]^T, \]

and

\[ D(\lambda_{\ell j}) = \begin{bmatrix} 1 & \lambda_{\ell j1} & \lambda_{\ell j1}^2 \\ 1 & \lambda_{\ell j2} & \lambda_{\ell j2}^2 \\ 1 & \lambda_{\ell j3} & \lambda_{\ell j3}^2 \end{bmatrix}, \]

the solution is defined by
(5.57) \[ a_{kj} = D(\lambda_{kj})^{-1} \cdot s_{kj} \cdot m. \]

**Step (\$j.2):** Evaluate \( S_{\lambda}^m(\lambda) \) at the value of \( \lambda \) maximizing the parabolic approximation, \( S_{\lambda}^{\ast m}(\lambda) \), i.e., at the value,

\[
(5.58) \quad \lambda_{kj}^{\ast} = \frac{-a_{kj2}}{2a_{kj3}},
\]

provided that \( a_{kj3} < 0 \) for a maximum.

**Step (\$j.3):** If \( S_{\lambda}^m(\lambda_{kj}^{\ast}) > S_{\lambda}^m(\lambda_{kj}^{\ast}) \), set \( \lambda_{kj}^{\ast} = \lambda_{kj}^{\ast} \). Otherwise, set \( \lambda_{kj}^{\ast k} \) equal to \( \lambda_{kj}^{\ast k}/2 \), \( k=1,2,3 \); set \( j \) to \( j+1 \); and repeat steps (\$j.1) - (\$j.3).

Hartley (1961) proves the convergence of the modified G-N algorithm, embodying (5.3) and Steps (\$j.1) - (\$j.3). Our procedure is to replace Step (\$j.3) above by the alternative:

**Step (\$j.3)':** Let \( j' = j+1 \) and define \( \lambda_{kj}^{\ast k}, k=1,2,3 \), by the following rule:

(a) \[ \text{if } \lambda_{kj}^{\ast} < \lambda \]

\[ \lambda_{kj}^{\ast 1} = \lambda_{kj}, \lambda_{kj}^{\ast 2} = \lambda_{kj}, \lambda_{kj}^{\ast 3} = \lambda_{kj} \]

(b) \[ \text{if } \lambda_{kj}^{\ast 1} < \lambda_{kj}^{\ast} < \lambda_{kj}^{\ast 2} \]

\[ \lambda_{kj}^{\ast 1} = \lambda_{kj}, \lambda_{kj}^{\ast 2} = \lambda_{kj}, \lambda_{kj}^{\ast 3} = \lambda_{kj} \]
\( \lambda_{x_j} = \lambda_{x_{j2}} \)

(c) if \( \lambda_{x_{j2}} < \lambda^*_x < \lambda_{x_{j3}} \)

\( \lambda_{x_j} = \lambda^*_x \)

\( \lambda_{x_{j1}} = \lambda_{x_{j2}} \)

\( \lambda_{x_{j2}} = \lambda_{x_{j3}} \)

\( \lambda_{x_{j3}} = \lambda^*_x \)

(d) if \( \lambda_{x_{j3}} < \lambda^*_x \)

\( \lambda_{x_{j1}} = \lambda_{x_{j2}} \)

\( \lambda_{x_{j2}} = \lambda_{x_{j3}} \)

\( \lambda_{x_{j3}} = \lambda^*_x \)

Then repeat steps (\textit{l}.j.1) - (\textit{l}.j.3) until either convergence obtains or a value of \( S^m_x(\lambda^*_x) < S^m_x(\lambda^*_{x,j-1}) \) occurs. If the latter obtains when \( j=1 \), repeat with \( \lambda_{x_k} = \lambda_{x_{jk}}/2 \), \( k=1,2,3 \).
6. **Illustrative Examples:**

In this section we examine certain computational problems which arise in the use of the Consistent Initial Estimator (CIE) of section 3 and subsequent determination of the Maximum Likelihood Estimator (MLE) by each of the four algorithms described in section 5. In section 6.1 we employ a generated data set, for which the true specification is known, to illustrate the variety of CIE's which result from alternative combinations of the moment orders, r and s, and the choice between the "unimproved" and the "improved" CIE. In section 6.2 we provide comparative computation times for each of the four algorithms using a bilaterally-censored and its corresponding truncated data sample. We take, as our example, a problem which we believe to be typical of those encountered by researchers working with sample survey data, involving both a large number of observations (N) and a large number of regressors (K). Such problems require considerable processing time—even on relatively modern computers. Hence, selection of a "computationally efficient" algorithm is of some importance. At most computer installations charges depend upon the amount of central memory occupied, the amount of central processing (CP) time required, and the amount of input-output (I-O) time used. Central memory requirements are essentially constant across each of our algorithms, and I-O costs are generally small relative to CP time. Therefore, in our comparison of algorithms, we shall focus upon the CP time-requirements.

6.1. **Choice of the Consistent Initial Estimator:**

In Table 6.1 we present alternative estimates of the "adjusted" CIE, defined in (3.30a), and the corresponding "improved" variance estimator of (3.31). Our example is based on a generated sample from the model:
\[
\begin{align*}
(6.1a) \quad y_i^* &= 30.0 + 0.3x_{1i} - 0.2x_{2i} - 1.4x_{3i} + 6.7x_{4i} + \varepsilon_i, \\
(6.1b) \quad \varepsilon_i &\sim \text{n.i.d. } (0, 0.64),
\end{align*}
\]

where our regressors have been generated as independent drawings from the distribution,

\[
\begin{bmatrix}
    x_{1i} \\
    x_{2i} \\
    x_{3i} \\
    x_{4i}
\end{bmatrix} \sim \text{n.i.d. } (0, \\
\begin{bmatrix}
    6.0 & 1.0 & -0.6 & 2.0 \\
    1.0 & 12.0 & -3.0 & 0.6 \\
    0.6 & -3.0 & 36.0 & -1.2 \\
    2.0 & 0.6 & -1.2 & 42.0
\end{bmatrix}).
\]

Censoring of \( y_i^* \) occurs at the fixed limits of 0 and 100. Of 500 observations, there were 30 at the lower limit; 348 between the limits; and 122 at the upper limit. In the typology of section 3, this example belongs in case 11: fixed bilateral censoring with a constant term.

The first row of Table 6.1 shows the MLE based on the full censored sample of 500 observations. The following 10 rows show the CIE for values of \( r=3,4,5,6 \), and \( s=1,\ldots,r-2 \). The last two columns show the value of the adjusted variance estimator and the improved variance estimator. In this example, the unadjusted variance estimator is positive for all specified choices of \( (r,s) \), and so, therefore, identical with the adjusted estimator.

By inspection, all choices of \( r \) and \( s \) yield plausible estimates of the regression coefficients. The adjusted variance estimate, \( \hat{\sigma}^2_{o(r,s)} \), is consistently less than both the true value and the MLE of \( \sigma^2_o \), but none provides any difficulty in subsequent calculation of the MLE. The improved variance estimates are virtually identical for all of our choices of \( r \) and \( s \), and all are "close" to the MLE, \( \hat{\sigma}^2_o \).
Table 6.1: Comparison of Consistent Initial Estimators

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>$\beta_4$</th>
<th>$\sigma^2$</th>
<th>$\tilde{\sigma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum Likelihood Estimator:</td>
<td>31.67</td>
<td>0.13</td>
<td>-0.21</td>
<td>-1.38</td>
<td>6.46</td>
<td>58.25</td>
<td></td>
</tr>
<tr>
<td>Consistent Initial Estimator:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r=3 s=1</td>
<td>31.88</td>
<td>0.16</td>
<td>-0.13</td>
<td>-1.37</td>
<td>6.12</td>
<td>39.97</td>
<td>57.68</td>
</tr>
<tr>
<td>r=4 s=1</td>
<td>32.91</td>
<td>0.12</td>
<td>-0.19</td>
<td>-1.32</td>
<td>6.18</td>
<td>46.09</td>
<td>57.56</td>
</tr>
<tr>
<td>r=4 s=2</td>
<td>32.21</td>
<td>0.15</td>
<td>-0.15</td>
<td>-1.35</td>
<td>6.16</td>
<td>43.53</td>
<td>57.55</td>
</tr>
<tr>
<td>r=5 s=1</td>
<td>34.01</td>
<td>0.05</td>
<td>-0.27</td>
<td>-1.30</td>
<td>6.25</td>
<td>47.98</td>
<td>57.82</td>
</tr>
<tr>
<td>r=5 s=2</td>
<td>33.34</td>
<td>0.09</td>
<td>-0.22</td>
<td>-1.31</td>
<td>6.21</td>
<td>46.70</td>
<td>57.61</td>
</tr>
<tr>
<td>r=5 s=3</td>
<td>32.49</td>
<td>0.13</td>
<td>-0.17</td>
<td>-1.34</td>
<td>6.18</td>
<td>44.88</td>
<td>57.50</td>
</tr>
<tr>
<td>r=6 s=1</td>
<td>34.89</td>
<td>-0.03</td>
<td>-0.37</td>
<td>-1.30</td>
<td>6.34</td>
<td>50.51</td>
<td>58.42</td>
</tr>
<tr>
<td>r=6 s=2</td>
<td>34.40</td>
<td>0.03</td>
<td>-0.31</td>
<td>-1.30</td>
<td>6.28</td>
<td>49.04</td>
<td>58.01</td>
</tr>
<tr>
<td>r=6 s=3</td>
<td>33.69</td>
<td>0.07</td>
<td>-0.25</td>
<td>1.31</td>
<td>6.24</td>
<td>47.48</td>
<td>57.69</td>
</tr>
<tr>
<td>r=6 s=4</td>
<td>32.75</td>
<td>0.12</td>
<td>-0.19</td>
<td>-1.34</td>
<td>6.20</td>
<td>45.75</td>
<td>57.49</td>
</tr>
</tbody>
</table>
Computation times, measured by the internal central processor clock on a Burroughs B7700 computer, were approximately 2 seconds for the unimproved CIE. Iterative improvement of the variance, employing the E-M version of the algorithm of (5.45), required an average of .7 seconds per iteration. Convergence, at a tolerance level of $10^{-4}$, required between 10 and 11 of these iterations. In practice, however, we have found that one or two iterations usually suffices to adequately approximate the conditional maximum.

In the present example, use of the improved variance estimator is not computationally efficient. A single iteration, using any one of the four ML estimators, required between 1.8 and 2.5 seconds to update the entire parameter vector. However, as the problem size increases, both in terms of $N$, the number of observations, and $K$, the number of regressors, the cost of a complete ML iteration, relative to a conditional ML iteration on $\sigma_o^2$ alone, will rise.

For example, in a regression model similar to the one described in section 6.2 below, employing $K=34$ regressors with $N=3781$ observations, the improvement of the CIE required an additional 115 seconds of CP time. Using the improved CIE, the time required to achieve convergence to the unconditional MLE, using the N-R algorithm, was reduced by 266 seconds. However, the use of improved CIE did not significantly reduce the total time required to update the entire parameter vector.

The selection of a particular value of $r$ and $s$ and the decision to employ the adjusted or the improved variance estimator is, in terms of computer time, a problem-specific issue. Of the ten CIE's in Table 6.1, the choice of $r=6$ and $s=1$ yields the "best" initial estimator (in the sense that the value of the log-likelihood, evaluated at that parameter point, is
Table 6.2: ML and OLS Estimators for Censored and Truncated ANCOVA models

<table>
<thead>
<tr>
<th>Regressors:</th>
<th>Censored Sample: (N=3781)</th>
<th>Truncated Sample: (N_2=3274)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLP</td>
<td>OLS</td>
</tr>
<tr>
<td>1. SEX</td>
<td>1.1909</td>
<td>0.9553</td>
</tr>
<tr>
<td></td>
<td>0.2169</td>
<td>0.1928</td>
</tr>
<tr>
<td>2. CTYPE</td>
<td>2.1891</td>
<td>1.7718</td>
</tr>
<tr>
<td></td>
<td>0.2975</td>
<td>0.2653</td>
</tr>
<tr>
<td></td>
<td>1.1558</td>
<td>1.0346</td>
</tr>
<tr>
<td></td>
<td>1.1557</td>
<td>1.0356</td>
</tr>
<tr>
<td>5. CELL 50</td>
<td>17.7285</td>
<td>18.0266</td>
</tr>
<tr>
<td></td>
<td>1.1567</td>
<td>1.0352</td>
</tr>
<tr>
<td></td>
<td>1.1686</td>
<td>1.0451</td>
</tr>
<tr>
<td>7. CELL 31</td>
<td>-1.7330</td>
<td>2.2558</td>
</tr>
<tr>
<td></td>
<td>1.5124</td>
<td>1.2505</td>
</tr>
<tr>
<td>8. CELL 32</td>
<td>-0.6774</td>
<td>3.5295</td>
</tr>
<tr>
<td></td>
<td>1.4352</td>
<td>1.2175</td>
</tr>
<tr>
<td>9. 'CELL 33</td>
<td>-0.8902</td>
<td>2.9047</td>
</tr>
<tr>
<td></td>
<td>1.5068</td>
<td>1.2667</td>
</tr>
<tr>
<td>10. CELL 34</td>
<td>1.3278</td>
<td>1.0247</td>
</tr>
<tr>
<td></td>
<td>1.4387</td>
<td>1.2511</td>
</tr>
<tr>
<td>11. CELL 41</td>
<td>2.7903</td>
<td>4.7809</td>
</tr>
<tr>
<td></td>
<td>1.2590</td>
<td>1.1135</td>
</tr>
<tr>
<td>12. CELL 42</td>
<td>4.0957</td>
<td>5.4110</td>
</tr>
<tr>
<td></td>
<td>1.3028</td>
<td>1.1580</td>
</tr>
<tr>
<td>13. CELL 43</td>
<td>2.8454</td>
<td>4.7320</td>
</tr>
<tr>
<td></td>
<td>1.3308</td>
<td>1.1728</td>
</tr>
<tr>
<td>14. CELL 44</td>
<td>3.2730</td>
<td>4.9405</td>
</tr>
<tr>
<td></td>
<td>1.3135</td>
<td>1.1607</td>
</tr>
<tr>
<td>15. CELL 51</td>
<td>5.7361</td>
<td>6.7391</td>
</tr>
<tr>
<td></td>
<td>1.4268</td>
<td>1.2712</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>16. CELL 52</td>
<td>5.3118</td>
<td>7.2500</td>
</tr>
<tr>
<td></td>
<td>1.4045</td>
<td>1.2550</td>
</tr>
<tr>
<td>17. CELL 53</td>
<td>6.4910</td>
<td>7.5636</td>
</tr>
<tr>
<td></td>
<td>1.4478</td>
<td>1.2888</td>
</tr>
<tr>
<td>18. CELL 54</td>
<td>5.6557</td>
<td>6.5455</td>
</tr>
<tr>
<td></td>
<td>1.4244</td>
<td>1.2666</td>
</tr>
<tr>
<td></td>
<td>1.2240</td>
<td>1.0984</td>
</tr>
<tr>
<td></td>
<td>1.2091</td>
<td>1.0838</td>
</tr>
<tr>
<td></td>
<td>1.2619</td>
<td>1.1307</td>
</tr>
<tr>
<td>22. CELL 64</td>
<td>9.3740</td>
<td>9.9342</td>
</tr>
<tr>
<td></td>
<td>1.2590</td>
<td>1.1290</td>
</tr>
<tr>
<td>23. FILL</td>
<td>-0.3006</td>
<td>-0.3424</td>
</tr>
<tr>
<td></td>
<td>0.2365</td>
<td>0.2115</td>
</tr>
<tr>
<td>24. MILL</td>
<td>-1.1135</td>
<td>-1.0443</td>
</tr>
<tr>
<td></td>
<td>0.3608</td>
<td>0.3242</td>
</tr>
<tr>
<td>25. SRES</td>
<td>-2.6582</td>
<td>-2.5396</td>
</tr>
<tr>
<td></td>
<td>0.9851</td>
<td>0.8826</td>
</tr>
<tr>
<td>26. MSTA</td>
<td>1.2045</td>
<td>1.1770</td>
</tr>
<tr>
<td></td>
<td>0.4006</td>
<td>0.3528</td>
</tr>
<tr>
<td>27. HAND</td>
<td>-3.7030</td>
<td>-2.7590</td>
</tr>
<tr>
<td></td>
<td>0.7185</td>
<td>0.6130</td>
</tr>
<tr>
<td>28. MAIL</td>
<td>-1.3704</td>
<td>-1.1583</td>
</tr>
<tr>
<td></td>
<td>0.2722</td>
<td>0.2432</td>
</tr>
<tr>
<td>29. ELEC</td>
<td>1.0019</td>
<td>0.8365</td>
</tr>
<tr>
<td></td>
<td>0.2690</td>
<td>0.2392</td>
</tr>
<tr>
<td>30. OWNR</td>
<td>-0.6331</td>
<td>-0.5957</td>
</tr>
<tr>
<td></td>
<td>0.2947</td>
<td>0.2639</td>
</tr>
<tr>
<td>31. RDIO</td>
<td>1.1300</td>
<td>0.9663</td>
</tr>
<tr>
<td></td>
<td>0.2951</td>
<td>0.2600</td>
</tr>
<tr>
<td>32. INCM</td>
<td>0.0071</td>
<td>0.0083</td>
</tr>
<tr>
<td></td>
<td>0.0041</td>
<td>0.0037</td>
</tr>
<tr>
<td></td>
<td>SCH</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>33</td>
<td>-1.1460</td>
<td>-0.9989</td>
</tr>
<tr>
<td></td>
<td>0.3039</td>
<td>0.2625</td>
</tr>
<tr>
<td>34</td>
<td>0.1613</td>
<td>0.1371</td>
</tr>
<tr>
<td></td>
<td>0.0721</td>
<td>0.0643</td>
</tr>
<tr>
<td>35</td>
<td>39.3824</td>
<td>32.7209</td>
</tr>
<tr>
<td></td>
<td>1.0177</td>
<td>0.5715</td>
</tr>
</tbody>
</table>
largest) for both the adjusted and the improved variance estimators. However, we would, in general, prefer lower-order choices of \((r,s)\), since they involve marginally lower computation times; exhibit less numerical round-off error; and should be asymptotically more efficient. It should also be clear that, while the estimator \(\hat{\sigma}^2(r,s)\) maximizes \(\log L^N_N(\sigma^2|\hat{\beta}_0)\), given the initial regression coefficient estimates \(\hat{\beta}_0\), it may still lie some distance from both the true parameter value and the MLE. In our experience, the improved variance estimator is of particular utility when, for given \(r\) and \(s\), the IVE is "small," relative to the MLE. However, the additional computational cost is often unwarranted when the IVE is "reliable." In these cases, it may be sufficient to only perform one or two iterations of the algorithm (5.45). Further, if the E-M algorithm \((m=4)\) is utilized, we can be sure that each iterate not only returns a positive estimate, but also improves the conditional log-likelihood.

6.2. Choice of the Maximum Likelihood Algorithm

Given an appropriate set of initial estimates for the parameter vector, the practitioner must next choose an ML algorithm. In our experience, of the unmodified algorithms, the N-R, G-N and M-S on occasion fail to converge, while the E-M (as noted in section 5.4) always converges, though in some cases, at a relatively slow rate.\(^1\) Further, all algorithms which do converge result in identical MLE's. Hence, in those cases where all algorithms converge, the preferred algorithm minimizes computation costs.

The examples cited below are taken from our continuing study of the

\(^1\) Perhaps the analogy of the race between the surefooted-but-slow tortoise and the erratic-but-speedy hare is useful.
retention of literacy and numeracy skills among primary school students and school leavers in Egypt. While these results are typical of our experience over a wide range of bilaterally-censored data sets, we should still emphasize that our particular outcomes are undoubtedly problem-specific.

A principal concern of the Egyptian-study has been to establish the shape of the representative student's skill-specific "learning curve" and the associated grade-specific "retention curves" of school leavers.\(^1\) The learning curve may be defined as the sequence of skill-levels achieved by a student as he/she advances through the sequence of primary school grades at annual intervals. The retention curve of a dropout is, analogously, the sequence of skill-levels retained by an individual—measured at annual intervals following a decision to leave primary school. In order to estimate these learning and retention curves from a cross-sectional sample, one approach is to cast the problem as a 2-way Analysis of Variance (ANOVA) model in which the "treatment effects" are the "grade-in-school last attended" and the number of "years-since-leaving-school." When the effects of other variables are also controlled for, the problem may be viewed as an Analysis of Covariance (ANCOVA) model. Both of these models may be represented in the linear regression framework, with appropriately-specified \((0,1)\) "dummy" variables to capture the treatment effects.

The observed dependent variable is the student's aggregate test score. However, because the test instruments measuring skill-levels span an inherently limited range of item difficulties, while the underlying skill-levels are assumed to vary according to a normal distribution, the appropriate

\(^1\) For a more complete discussion of the project's goals, the data collected and some empirical results, see Hartley and Swanson (1980). Our purpose, here, is only to illustrate the statistical methods developed in sections 2 to 4 and to examine certain computational aspects of section 5.
normal regression model has a bilaterally-censored dependent variable, with fixed limit points at zero and at the maximum-possible test score. If the censored character of the dependent variable is ignored, and the customary OLS estimates are employed for the ANOVA or ANCOVA models, biased parameter estimates result. Alternatively, some researchers may choose to discard all observations at the limit points, producing a corresponding truncated sample. The MLE of the parameters, using a truncated sample, will also be consistent, but less efficient than the MLE based on the full (censored) data set. OLS estimates, based on the truncated sample, are still biased.

In Table 6.3, below, we present results for a typical ANCOVA specification, applied to the results from our "Test 5," an instrument designed to measure skill in performing simple arithmetic operations. The sample consists of 3781 children for which test scores and limited family background data were available. On this test, censoring at the lower limit predominated: there were 408 "zero scores" and 79 "perfect scores." The estimated ANCOVA model involves 34 regression coefficients and the error variance.1/

The OLS bias using the full data sample is readily apparent in the results shown in Table 6.2 and is not reduced by employing the truncated data

1/ The 20 variables labelled CELLjk refer to the (j,k) levels of "treatments" applied to each child, where j=3,4,5,6 denotes the grade-last-attended, and k=0,1,2,3,4 denotes the number of years-lapsed-since-last-attending. k=0 denotes a student still in school. A CELL-variable takes on a value of one when a child is a member of that cell and is zero otherwise. SEX has a value of one for males and zero for females. CTYTYPE is one for urban schools and zero for rural schools. The other binary variables are FILL, father illiterate, MILL, mother illiterate, SRES, student lives at home; MSTTA, parents living together, HAND, child handicapped; MATL, house built of clay; ELEC, electricity in home; OWNR, family owns home; RDIO, radio in home. In each of these, the variable takes on a value of one for an affirmative response and zero otherwise. The remaining variables are INCNM, family income in Egyptian pounds per month; SCRA, numbers of schools attended; and ROOM, number of rooms in the family house.
set. The coefficients of the \(CELL_{jk} \) variables, which may be interpreted as points along the learning and retention curves, controlling for the observed family characteristics, are consistently overestimated by OLS. This reflects the predominance of left-tailed censoring or truncation in the sample. The degree of left-censoring, and, hence, the cell-bias, increases with \( k \) (the number of years since attending school) and decreases with \( j \) (grade-last-attended). As a consequence, the OLS-estimated retention curves give the erroneous impression of a threshold in the limiting values of the grade-specific retention curves (as \( k \rightarrow \infty \) for given \( j \)) between grades 5 and 6. The OLS estimates of the variances of the parameter estimates are shown below and are consistently smaller than the ML variances.

Comparative CP times for the four unmodified ML algorithms, applied to this problem, are shown in Table 6.3 for both the censored and corresponding truncated data sets. Estimation of the MLE, under each method, began with the improved CIE (using values of \( r=3 \) and \( s=1 \)) and iterated until convergence was obtained (at a \( 10^{-4} \) tolerance level). For the censored sample, the M-S algorithm achieves convergence in the fewest iterations. But, in terms of total computation time, the E-M algorithm is significantly faster than any other method. The E-M's comparative advantage lies in avoiding the repeated matrix inversion, otherwise required in updating the matrix, \( \Lambda_k \), defined in (5.4a), (5.5a), and (5.6a) for the N-R, G-N, and M-S algorithms, respectively, at each iteration. Because the \( [\mathbf{X}'\mathbf{X}]^{-1} \) matrix in (5.18a) must be formed once, the first iteration of E-M is slow—requiring, in this example, 111 seconds. Subsequent iterations, however, averaged only 43 seconds, where, during these steps, only \( y_{12}^c \) and \( s_{12}^c \) need to be recomputed.

In smaller problems—especially those involving fewer regressors, our experience suggests that the absolute computational advantage would lie with
### Table 6.3: Maximum Likelihood Algorithm Computation Times for Modelling a Simple Arithmetic Operations Test (K=34)

<table>
<thead>
<tr>
<th></th>
<th>Time for CIE: (seconds)</th>
<th>Time for MLE: (seconds)</th>
<th>No. of iterations to converge:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I. Censored Sample (N=3781):</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Newton-Raphson</td>
<td>185</td>
<td>685</td>
<td>7</td>
</tr>
<tr>
<td>B. Gauss-Newton</td>
<td>185</td>
<td>1702</td>
<td>17</td>
</tr>
<tr>
<td>C. Scoring</td>
<td>185</td>
<td>637</td>
<td>6</td>
</tr>
<tr>
<td>D. Expectation-Maximization</td>
<td>185</td>
<td>495</td>
<td>10</td>
</tr>
<tr>
<td><strong>II. Truncated Sample (N=3274):</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Newton-Raphson</td>
<td>175</td>
<td>740</td>
<td>8</td>
</tr>
<tr>
<td>B. Gauss-Newton</td>
<td>175</td>
<td>1636</td>
<td>18</td>
</tr>
<tr>
<td>C. Scoring</td>
<td>175</td>
<td>455</td>
<td>5</td>
</tr>
<tr>
<td>D. Expectation-Maximization</td>
<td>175</td>
<td>2302</td>
<td>*</td>
</tr>
</tbody>
</table>

* Had not converged after 23 iterations.
either the N-R or M-S algorithm.

When the corresponding truncated data set is employed, the ranking of the four algorithms changes dramatically. The comparative advantage of E-M disappears, since the weighted moment matrix of \((5.38)\) for M-T must both be constructed and inverted at each iteration. Furthermore, although the E-M improves the likelihood at each iteration, the algorithm had not yet converged after 23 iterations. Average-times-per-iteration for the other algorithms are somewhat reduced due to the smaller number of observations included. However, the N-R and G-N each require one additional iteration to converge, which increases their total processing time. The M-S algorithm, somewhat surprisingly, requires one less step to converge and is, therefore, easily the fastest of the four. Whether or not this result obtains in general requires further controlled sensitivity studies of each algorithm, using a variety of artificial data sets—a task which was beyond the scope of the present paper.

In the examples shown in Table 5.3, the G-N, M-S, and E-M algorithms all converged monotonically to a global maximum of the log-likelihood. The N-R, however, took a first step which actually decreased the value of the log-likelihood in both the censored and truncated samples—a not uncommon event. In general, convergence of the three gradient-type algorithms cannot be guaranteed, unless modified by an appropriate choice of \(\lambda^m\). Indeed, in the case of N-R, since the matrix of second partials of the log-likelihood cannot be assumed to be negative definite, except in a sufficiently small neighborhood of \(\theta\) (see Amemiya (1973)), convergence with a positive \(\lambda^1\)-choice cannot be assured. In our experience, the most difficult convergence problems occur in samples characterized by relatively heavy and asymmetrical censoring. For example, one of our skill-specific tests produced a sample of data for 2499 individuals, in which 741 received the minimum, and
the maximum score. Neither the E-M nor the unmodified N-R algorithms, applied to an ANCOVA model similar to that shown in Table 6.2, produced convergent estimates after 20 iterations. The modified N-R exceeded processing time limits after 9 iterations, but, at that point, was "closer" to the MLE, as measured by the value of the log-likelihood, than the unmodified N-R, and farther from the MLE than the (unmodified) E-M. In general, we have found that the computation of \( \lambda^m \) reduces the number of iterations required to achieve convergence, but often increases total computation time. Given such circumstances, we would prefer to rely on the E-M algorithm, especially for censored data sets, where it offers not only substantial time advantages over both modified and unmodified versions of N-R, C-N, and M-S, but also guarantees eventual convergence.

For problems involving much larger data sets (in terms of the number of regressors) than those described here, the "second-round" or "one-step" N-R, C-N or M-S estimator may be the best available alternative to calculation of the MLE. In our earlier empirical paper, for example, we report "one-step" results for three equations involving 80 regressors and approximately 4400 observations. Calculation of each complete ML iteration for these problems averaged 180 seconds. However, the one-step estimators for a given problem will,

Despite the equivalence in the asymptotic distributions of the two estimators, our experience in "small" samples, is that the numerical values of the "one-step" estimates differ considerably from the corresponding MLE, \( \hat{\theta}^m \), \( M = T, G \).

1/ See Hartley and Swanson (1980), Table 4.2. Separate timings for the unimproved CIE used are not available, but CP time is estimated to have averaged 180 seconds for the choice of the CIE defined by \( r=3 \) and \( s=1 \).
7. Summary and Conclusions:

We have examined the problem of obtaining MLE's of the parameters in a normal regression model from singly- and bilaterally-censored/truncated samples. The estimators obtained are strongly consistent, asymptotically efficient and asymptotically normally distributed. A consistent estimate of the asymptotic covariance matrix is also available. In addition, we have presented a class of weakly consistent initial estimators, which extends Anemiya's Instrumental Variables approach to the case of bilaterally-censored and truncated samples.

Four ML algorithms have been described: The Newton-Raphson, Gauss-Newton, Method of Scoring, and Expectation-Maximization methods. The appropriate choice of algorithm depends, predominantly, on the type of problem (censoring or truncation), the number of regressors, and, to a lesser extent, on the number of observations included. In our experience, problems employing a censored data set and involving a relatively large number of regressors (K>30) are most efficiently handled by the E-M method. Our illustrations involve problems with more than 3,000 observations and up to 80 regressors—dimensions that are increasingly being confronted in analysis-of-covariance models with sample survey data. In smaller samples, the comparative advantage dominated by all of the gradient methods. Of the three presented, we have found the Method of Scoring to be the fastest—both in terms of the number of iterations required to achieve convergence and the amount of central processing time required.

We have presented two methods of "modifying" the gradient-method algorithms. The modified algorithms will, in general, require fewer iterations but more processing time to achieve convergence. Use of the
modified algorithms is generally not recommended unless the corresponding
unmodified algorithm produces a decrease in the log-likelihood in its initial
step. For censored samples, the (unmodified) E-M algorithm will usually be
more efficient than any of the modified gradient algorithms.

In selection of the consistent initial estimator, our preference is
to use lower-order moments—both on grounds of efficiency and computational
accuracy. Since the initial variance estimator is frequently negative—even
in fairly large samples, we have provided an "adjusted" variance estimator,
which is positive and weakly consistent.$^{1/}$ We have also discussed the
selective "improvement" of the variance estimator, using a Conditional Maximum
Likelihood approach. The improved variance estimator is particularly useful
for samples which yield very small, but positive, initial variance estimators.

Finally, a general computer program, BILATERAL, which performs all
of the calculations and options within the bilaterally truncated/censored
(including Probit) normal regression model and discussed explicitly in this
paper, has been written, and will be made available at cost upon written
request.

$^{1/}$ Amemiya (1973) attributes the occurrence of negative initial IV variance
estimates to either "small" sample sizes or a misspecification of the
model. If one suspects the former cause, he suggests the use of Tobin's
(1958) inconsistent initial estimator. If this is also negative (or
imaginary), "...one could only guess the initial value... ."
Bibliography


Hald, A. (1949) "Maximum Likelihood Estimation of the Parameters of a Normal Distribution which is Truncated at a Known Point," Skandinavisk Aktuarietidskrift, 32, 119-134.


