DISCUSSION PAPER

MAXIMUM LIKELIHOOD ESTIMATION OF THE
TRUNCATED AND CENSORED NORMAL REGRESSION MODELS

by

Michael J. Hartley and Eric V. Swanson

November, 1985

Development Research Department
Economics and Research Staff
World Bank

The World Bank does not accept responsibility for the views expressed herein which are those of the author(s) and should not be attributed to the World Bank or to its affiliated organizations. The findings, interpretations, and conclusions are the results of research supported by the Bank; they do not necessarily represent official policy of the Bank. The designations employed, the presentation of material, and any maps used in this document are solely for the convenience of the reader and do not imply the expression of any opinion whatsoever on the part of the World Bank or its affiliates concerning the legal status of any country, territory, city, area, or of its authorities, or concerning the delimitations of its boundaries, or national affiliation.
MAXIMUM LIKELIHOOD ESTIMATION OF THE
TRUNCATED AND CENSORED NORMAL REGRESSION MODELS

by

Michael J. Hartley*

and

Eric V. Swanson**

November, 1985

This paper is a revision of our earlier discussion paper
(Hartley and Swanson (1980)) and was 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 of the
Retention of Literacy and Numeracy (RPO 671-55), financed by the World
Bank. A draft final report (Hartley and Swanson (1984)) has been
completed.

*  Econometric Advisor/Economist, Development Research Department.

**  Systems Analyst, Economic Analysis and Projections Department.
Abstract

Researchers frequently encounter data sets in which the observable values of the dependent variable(s) are limited to a finite range. When no information is available on data outside the observable range, the sample is said to be "truncated." When the number of data points outside the observable range is known but the actual values are unknown, the sample is said to be "censored." Application of ordinary least squares to regression models of limited dependent variables results in biased and inconsistent parameter estimates.

This paper provides a general treatment of the problem of estimating the parameters of a regression model in the presence of censoring and truncation. The maximum likelihood estimator (MLE) and its covariance matrix for both the censored and truncated models are presented. Since the likelihood equations are nonlinear, solutions must be obtained by iterative methods. Four computer algorithms for obtaining the MLE are given and compared. In addition, a method of modifying the algorithms so as to improve their rate of convergence is given.

The other major contribution of this paper is to present a class of (weakly) consistent initial estimators for various types of censored and truncated regression models. The discussion of the initial estimator treats the case of both fixed and variable limits on the dependent variable. A method for obtaining consistent and positive estimates of the model variance is described. An "improved" initial variance estimator is also presented. The resulting parameter estimates can be used as the starting point for a "one-step," strongly consistent estimator that has the same asymptotic distribution as the MLE.

Finally, Monte Carlo experiments are used to compare the consistent initial estimator with the convergent MLE and to provide guidelines for the selection of an algorithm.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>II. Truncated and Censored Normal Regression Models</td>
<td>5</td>
</tr>
<tr>
<td>1. The Underlying Model</td>
<td>5</td>
</tr>
<tr>
<td>2. Truncated and Censored Samples</td>
<td>6</td>
</tr>
<tr>
<td>3. Assumptions</td>
<td>8</td>
</tr>
<tr>
<td>4. The Log-Likelihood Function and Its Derivatives</td>
<td>10</td>
</tr>
<tr>
<td>5. The Conditional Moments of the Latent Dependent Variable</td>
<td>13</td>
</tr>
<tr>
<td>III. Consistent Initial Estimators</td>
<td>15</td>
</tr>
<tr>
<td>1. A Taxonomy of Model Types</td>
<td>16</td>
</tr>
<tr>
<td>2. The Use of Conditional Moments and a Class of Instrumental Variable Estimators</td>
<td>20</td>
</tr>
<tr>
<td>Type 1</td>
<td>22</td>
</tr>
<tr>
<td>Type 2</td>
<td>25</td>
</tr>
<tr>
<td>3. Improvements to the Initial Variance Estimator</td>
<td>27</td>
</tr>
<tr>
<td>IV. Asymptotic Distribution Theory</td>
<td>29</td>
</tr>
<tr>
<td>V. Calculation of the MLE</td>
<td>32</td>
</tr>
<tr>
<td>1. The Newton-Raphson Method</td>
<td>34</td>
</tr>
<tr>
<td>2. The Method of Scoring</td>
<td>35</td>
</tr>
<tr>
<td>3. The Gauss-Newton Method</td>
<td>35</td>
</tr>
<tr>
<td>4. The Expectation-Maximization Method</td>
<td>36</td>
</tr>
<tr>
<td>1. The Doubly-Censored Model</td>
<td>39</td>
</tr>
<tr>
<td>2. The Doubly-Truncated Model</td>
<td>45</td>
</tr>
<tr>
<td>3. The Canonical Form</td>
<td>49</td>
</tr>
<tr>
<td>5. Computation of the Improved Initial Variance Estimator</td>
<td>51</td>
</tr>
<tr>
<td>6. Modified Algorithms</td>
<td>52</td>
</tr>
<tr>
<td>VI. An Illustrative Example</td>
<td>57</td>
</tr>
<tr>
<td>VII. Summary and Conclusions</td>
<td>61</td>
</tr>
</tbody>
</table>

1/ We are indebted to Aludia Oropesa and Miriam Bailey for their excellent word-processing.
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 (ML) 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 Maximum Likelihood Estimator (MLE) for the singly-truncated regression model, and Olsen (1980) has generalized the Pearson-Lee method-of-moments estimator to the regression problem. An excellent survey of this burgeoning econometric literature has been provided by Amemiya (1984), and Maddala (1983) devotes a chapter to the problem in his insightful book on limited dependent and qualitative variable models.

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), Maddala (1983) and Amemiya (1984)), 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

---

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.
particular tolerance level, as determined by an external measuring device, are
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 (Hartley and
Swanson (1985) for illustrative purposes. There, 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/or censored linear regression models, including those
for the analysis of variance and covariance, the theoretical literature is
still deficient in several respects:

(a) A consistent initial estimator is (to our knowledge) 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 MLE'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 a set of data generated from Monte Carlo Experiments. In a companion paper (Hartley and Swanson (1985)), we shall apply certain of the various methods developed here to estimate "learning" and dropout "retention curves" for an extensive sample of primary school-age children in Egypt.
II. 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,"

\[
(2.1) \quad y_i^* = x_i' \beta_0 + \epsilon_i, \quad -\infty < y_i^* < \infty,
\]

where \( x_i \) is a \( K \)-vector of observations on the independent variables, \( \beta_0 \) is a
\( K \)-vector of corresponding unknown constants and \( \epsilon_i \) is a normally and
independently distributed disturbance,

\[
(2.2) \quad \epsilon_i \sim \text{n.i.d.}(0, \sigma^2_0).
\]

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

\[
(2.3) \quad f_{\hat{y}}(y_i^*) = \frac{1}{\sqrt{2\pi} \sigma_0} \exp \left( -\frac{1}{2} \frac{1}{\sigma_0^2} (y_i^* - x_i' \beta_0)^2 \right),
\]

with associated distribution function,

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

\(^1\) The subscript "o" denotes evaluation at the true parameter values.
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 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

\[
(2.5) \quad -\infty \leq z_{i1} \leq y_i \leq z_{i2} \leq +\infty ,
\]

with the proviso that

\[
(2.6) \quad z_{i1} \leq z_{i2} .
\]

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,

\[
(2.8) \quad g_{i0}(y_i) = \begin{cases} 
\frac{f_{i0}(y_i)}{[F_{i0}(z_{i2}) - F_{i0}(z_{i1})]} , & \text{if} \quad z_{i1} < y_i < z_{i2} \\
0 , & \text{otherwise} .
\end{cases}
\]

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
$F_{10}(z_{i2}) = 1$ in (2.8). Similarly, if $z_{i1} = -\infty$ and $z_{i2} < +\infty$, we have the right-truncated model in which $g_{10}(y_i)$ of (2.8) simplifies via $F_{10}(z_{i1}) = 0$ and $0 < F_{10}(z_{i2}) < 1$. The general case of a bilaterally- or doubly-truncated regression model arises when both $z_{i1} > -\infty$ and $z_{i2} < +\infty$, whence $0 < F_{10}(z_{i1}) < F_{10}(z_{i2}) < 1$. Thus, in all three 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:

$$y_i = \begin{cases} 
    z_{i1}, & \text{if } y_i^* \leq z_{i1} \\
    y_i^*, & \text{if } z_{i1} < y_i^* < z_{i2} \\
    z_{i2}, & \text{if } y_i^* \geq z_{i2}
\end{cases}$$

(2.9)

This is the situation in censored samples in which $y_i^*$ is observed as $y_i$ within the open interval, $(z_{i1}, z_{i2})$, but, whenever $z_{i1}$ and/or $z_{i2}$ are finite, the knowledge that $y_i^*$ was less than or equal to $z_{i1}$ (greater than or equal to $z_{i2}$) is indicated by $y_i$ being observed as $z_{i1}$ ($z_{i2}$). Again, a left-censored normal regression (or Tobit) model obtains when $z_{i1} > -\infty$ and $z_{i2} = +\infty$; a right-censored model obtains if $z_{i1} = -\infty$ and $z_{i2} < +\infty$; and a bilaterally-censored normal regression model follows when both $z_{i1} > -\infty$ and $z_{i2} < +\infty$. In the last of these, the density of $y_i$ is defined by:

$$h_{10}(y_i) = \begin{cases} 
    F_{10}(z_{i1}), & \text{if } y_i = z_{i1} \\
    f_{10}(y_i), & \text{if } z_{i1} < y_i < z_{i2} \\
    1-F_{10}(z_{i2}), & \text{if } y_i = z_{i2}
\end{cases}$$

(2.10)
Thus, in this censored example, \( y_i \) 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 \( y_i^* \) are never observed. Rather, the fact that \( y_i^* \) was negative (positive) is indicated by the binary variable,

\[
y_i = \begin{cases} 
0 & \text{if } y_i^* \leq 0 \\
1 & \text{if } y_i^* > 0 
\end{cases}
\]

where \( y_i \) now has the density function,

\[
h_i(y_i) = \begin{cases} 
F_{i0}(0) & \text{if } y_i = 0 \\
1 - F_{i0}(0) & \text{if } y_i = 1 
\end{cases}
\]

and \( \sigma_0^2 \) 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 \( \hat{\theta} = \begin{bmatrix} \hat{\beta} \\ \cdots \\ \hat{\sigma}^2 \end{bmatrix} \) denote an arbitrary point in the parameter space, \( \Theta \), and let \( \hat{\theta}_0 = \begin{bmatrix} \hat{\beta}_0 \\ \cdots \\ \hat{\sigma}_0^2 \end{bmatrix} \) denote the true value of...
\( \Theta \). Then \( \Theta \) is compact, excludes the region, \( \sigma^2 < 0 \), and contains and open neighborhood, \( \Theta_0 \), of \( \Theta_0 \).

**Assumption 2:** The regressors, \( \{x_i\} \), are bounded and have an empirical distribution function, \( P_N (x) = \frac{j}{N} \) (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 \( P \)) 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:** 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 \( i \),

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

\[1/\] 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.
2.4. The Log-Likelihood Function and Its Derivatives:¹/  

Let \( S^N \) denote the observation index set, \( \{1, 2, \ldots, N\} \), and define the index subsets, \( S^N_j = \{ i \in S^N : y_i^* \in Y_j \} \), \( j = 1, 2, 3 \), where

\[
\begin{align*}
\text{\( Y_1 \)} & \quad \text{if} \quad - \infty < y_i^* \leq z_{11} \quad \text{i.e.,} \quad y_i = z_{11} \\
(2.13) \quad y_i^* & \quad \begin{cases} 
Y_2 & \text{if} \quad z_{11} < y_i^* < z_{12} \quad \text{i.e.,} \quad y_i = y_i^* \\
Y_3 & \text{if} \quad z_{12} \leq y_i^* < + \infty \quad \text{i.e.,} \quad y_i = z_{12} 
\end{cases}
\end{align*}
\]

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

For any function, \( q(y_i^*; x_i, \theta) \), we shall use the shorthand notation:

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

Then, in the case of truncated samples, the log-likelihood function is defined by

\[
(2.15) \quad \log L^T_N(\theta) = \sum_{i \in S_2} \log f_i - \sum_{i \in S_2} \log [F_i - F_{i1}] ,
\]

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

¹/ Henceforth, we shall discuss the general case of the bilaterally-or doubly-truncated and censored normal regression models—except in section III, where methodological differences between the singly- and doubly-truncated/censored cases require separate discussion.
\[ (2.16) \quad \log L_N^C(\theta) = \sum_{i \in S_1} \log f_{i1} + \sum_{i \in S_2} \log f_i + \sum_{i \in S_3} \log (1 - F_{i2}), \]

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

\[ (2.18a) \quad u_i = y_i^* - x_i \hat{\beta} \]

and

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

Then, the first partials of the log-likelihood function in the doubly-truncated model are given by:

\[ (2.19a) \quad \frac{\partial \log L_T}{\partial \hat{\beta}} = \frac{1}{\sigma^2} \sum_{i \in S_2} \left( u_i + \sigma^2 (g_{i2} - g_{i1}) \right) \cdot x_i \]

and

\[ (2.19b) \quad \frac{\partial \log L_T}{\partial \sigma^2} = \frac{1}{2\sigma^4} \sum_{i \in S_2} \left( u_i^2 - \sigma^2 + \sigma^2 (u_{i2} g_{i2} - u_{i1} g_{i1}) \right) , \]

whereas the second partials are given by:
\[
(2.20a) \quad \frac{\sigma^2 \log L}{\sigma \sigma_0} = -\frac{1}{\sigma^2} \cdot \epsilon_{i \in S_2} \left[ (1 - (u_{12} g_{12} - u_{11} g_{11})) \cdot (g_{12} - g_{11})^2 \right] \cdot x_i x_i^T \cdot x_i^T \\
(2.20b) \quad \frac{\sigma^2 \log L}{\sigma \sigma_2} = -\frac{1}{\sigma^4} \cdot \epsilon_{i \in S_2} \left[ \frac{1}{2} \cdot (u_{12}^2 - \sigma^2) g_{12} - (u_{11}^2 - \sigma^2) g_{11} \right] \\
- \frac{\sigma^2}{2} \cdot (g_{12} - g_{11}) \cdot (u_{12} g_{12} - u_{11} g_{11}) \cdot x_i \\
\text{and} \\
(2.20c) \quad \frac{\sigma^2 \log L}{(\sigma \sigma_2)^2} = -\frac{1}{\sigma^6} \cdot \epsilon_{i \in S_2} \left[ \frac{1}{2} \cdot (u_{12}^3 - u_{11}^3) g_{12} - \frac{3 \sigma^2}{4} \cdot (u_{12} g_{12} - u_{11} g_{11}) \right] \\
- \frac{1}{4} \cdot (u_{12}^3 - u_{11}^3) \cdot (g_{12} - g_{11}) \cdot x_i \\
\]

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

\[
(2.21a) \quad \frac{\partial \log L}{\partial \sigma} = \frac{1}{\sigma^2} \cdot \epsilon_{i \in S_1} \left[ \frac{f_{i1}}{F_{i1}} \cdot x_i + \frac{f_{i2}}{F_{i1}} \cdot x_i \cdot x_i^T \right] \\
\text{and} \\
(2.21b) \quad \frac{\partial \log L}{\partial \sigma_2} = \frac{1}{\sigma^4} \cdot \epsilon_{i \in S_1} \left[ \frac{f_{i1}}{F_{i1}} \cdot x_i + \frac{f_{i2}}{F_{i1}} \cdot x_i \cdot x_i^T \right] \\
\text{whereas the second partials are given by:} \\
(2.22a) \quad \frac{\partial^2 \log L}{\partial \sigma \partial \sigma_0} = -\frac{1}{\sigma^2} \cdot \epsilon_{i \in S_1} \left[ (u_{11} + \sigma^2) \cdot \frac{f_{i1}}{F_{i1}} \cdot \frac{f_{i1}}{F_{i1}} \cdot x_i \cdot x_i^T \right] \\
+ \epsilon_{i \in S_2} \left[ (-u_{12} + \sigma^2) \cdot \frac{f_{i2}}{F_{i1}} \cdot \frac{f_{i2}}{F_{i1}} \cdot x_i \cdot x_i^T \right] \\
(2.22b) \quad \frac{\partial^2 \log L}{\partial \sigma \partial \sigma_2} = -\frac{1}{\sigma^4} \cdot \epsilon_{i \in S_1} \left[ \frac{1}{2} \cdot \sigma^2 \cdot \frac{f_{i1}}{F_{i1}} \cdot \frac{f_{i1}}{F_{i1}} \cdot x_i \cdot x_i^T \right] \\
+ \epsilon_{i \in S_2} \left[ \frac{1}{2} \cdot \sigma^2 \cdot \frac{f_{i1}}{F_{i1}} \cdot \frac{f_{i1}}{F_{i1}} \cdot x_i \cdot x_i^T \right]
\]
and

\[
(2.22c) \quad \frac{\sigma^2 \log \mathbb{E}_C}{(\sigma^2)^2} = \frac{1}{\sigma^2} \cdot \left\{ -\frac{1}{4} \sum_{i \in S_1} \sigma^2 \cdot u_{1i} \cdot \frac{f_{i1}}{F_{i1}} \cdot (3 - \frac{1}{\sigma^2} \cdot u_{1i}^2 - u_{1i} \cdot \frac{f_{i1}}{F_{i1}}) + \sum_{i \in S_2} \sigma^2 \cdot u_{1i}^2 \cdot \frac{F_{i1}}{1 - F_{i1}} \cdot (-3 + \frac{1}{\sigma^2} \cdot u_{1i}^2 - u_{1i} \cdot \frac{f_{i1}}{F_{i1}}) \right\},
\]

respectively.

2.5. The Conditional Moments of the Latent Dependent Variable:

Let the conditional density function of \( y_i^* \), given \( y_i^* \in Y_i \), be denoted by:

\[
(2.23) \quad u_{ij}(y_i^*) = \begin{cases} 
\frac{f_i(y_i^*)}{F_{i1}}, & \text{if } y_i^* \in Y_1 \ (j = 1) \\
\frac{f_i(y_i^*)}{F_{i2} - F_{i1}}, & \text{if } y_i^* \in Y_2 \ (j = 2) \\
\frac{f_i(y_i^*)}{1 - F_{i2}}, & \text{if } y_i^* \in Y_3 \ (j = 3)
\end{cases}
\]

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

\[
(2.24) \quad \mu_{ij}^{(r)} = \mathbb{E}[y_i^{*r} | y_i^* \in Y_i] = \int_{Y_i} y_i^{*r} \cdot u_{ij}(y_i^*) \, dy_i^*.
\]

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

\begin{align}
\mu_{ij}^{(r)} &= \mu_{ij}^{(r-1)} \cdot x_{i}^{j} + (r-1) \cdot \mu_{ij}^{(r-2)} \cdot \sigma^{2} \cdot \frac{z_{i2}^{r-1} - F_{i1}}{1 - F_{i2}}, & \text{if } j = 1 \\
\mu_{ij}^{(r)} &= \mu_{ij}^{(r-1)} \cdot x_{i}^{j} + (r-1) \cdot \mu_{ij}^{(r-2)} \cdot \sigma^{2} \cdot \frac{z_{i2}^{r-1} - z_{i1}^{r-1} F_{i1}}{F_{i2} - F_{i1}}, & \text{if } j = 2 \\
\mu_{ij}^{(r)} &= \mu_{ij}^{(r-1)} \cdot x_{i}^{j} + (r-1) \cdot \mu_{ij}^{(r-2)} \cdot \sigma^{2} \cdot \frac{z_{i2}^{r-1}}{1 - F_{i2}}, & \text{if } j = 3
\end{align}

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

\begin{equation}
\log L_{N}^{C}(\theta) = \sum_{i \in S_{1}^{N}} \log F_{i}(0) + \sum_{i \in S_{2}^{N} + S_{3}^{N}} \log[1 - F_{i}(0)].
\end{equation}

Here, clearly, \( F_{i1} = F_{i2} = F_{i}(0) \) and the first and second partials of the log-likelihood with respect to \( \theta \) 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 \( z_{i1}^{r-1} \) as one when \( r = 1 \).

\footnote{Subsequent to our earlier draft (Hartley and Swanson (1980)), Professor Lung-Fei Lee provided us with an elegant proof of our autoregressive relation between the conditional moments of a normal density (2.25).}
III. Consistent Initial Estimators:

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

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

are not feasible. Thus, the MLE, \( \hat{\theta}^M \), must be found by iterative methods, and reliable starting parameter values reduce the overall computational cost. Our purpose, here, is to present a general methodology, which may be applied, \text{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.\(^1\)

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 expressions for the first and second conditional moments of the dependent variable, given that the actual value of the latent variable, \( y^*_i \), belongs to \( Y_2 \), so that \( y^*_i \) is

\(^1\) We conjecture that our procedures (involving suitable linear transformations of the elements of the sequence of certain higher order conditional moments, in combinations 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.
observed as \( y_1 \). This results in a model which is linear in the parameters of interest, \( \beta_0 \) and \( \sigma^2 \), 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 \((y_i, x_i)\) pairs for which the observed \(y_i\)-value is equal to the underlying \(y_i^*\). In the general formulation of the doubly-truncated/censored model, this would include all data such that \(z_{i1} < y_i^* < z_{i2}\), i.e., all \(i \in S_2\). Thus, for the censored models, in calculating initial estimates, we ignore observations, \(y_i\), equal to the limit points, \(z_{i1}\) or \(z_{i2}\) 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,

\[
(3.2) \quad y_i^* = x_i' \beta_0 + \epsilon_i, \quad z_{i1} < y_i^* < z_{i2}, \ i \in S_2.
\]

---

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 and any initial value, say \( \beta = 0 \).
We shall consider the cases in which $y^*_i$ is constrained to lie within three types of intervals:

(i) Left-truncated: $z_{i1} > -\infty$ and $z_{i2} = +\infty$

(ii) Right-truncated: $z_{i1} = -\infty$ and $z_{i2} < +\infty$

(iii) Doubly-truncated: $z_{i1} > -\infty$ and $z_{i2} < +\infty$

when $y^*_i$ is observed as $y_i$. If the finite limit points, $z_{i1}$ and/or $z_{i2}$ are the same for all $i \in S^N_2$, we shall refer to such a case as a "fixed" truncation point (with $z_{i1} = z_1$ and/or $z_{i2} = z_2$). 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:

$$\begin{pmatrix} y_i' \\ x_i \end{pmatrix} = \begin{pmatrix} 1 & x_i' \\ \beta_0 \\ \beta_2 \end{pmatrix} = \beta_1 + x_i^{*} \beta_2.$$  

(3.3)

In Table 3.1 we exhibit each of the 16 possible cases. For each case, we note the type of truncation (col. (1)), the assumptions on $z_{i1}$ and $z_{i2}$ (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:

$$y^*_i = x_i^{**} \delta_0 + \epsilon_i^{**}, \quad 0 < y_i^{**} < z_i^{**}, \quad i \in S^N_2,$$

(3.4)

where the appropriate transformation and the definitions of $x_i^{**}, \delta_0, z_i^{**}$
Table 3.1: Transformations of Original Model Prior To Application of Instrumental Variables

| Original Model: | $y^* = x^*_1 \beta_0^* + \epsilon_1^*$, $z_{11}^* \leq y^* \leq z_{12}^*$, $\beta_0^* = [\beta_{10}^* \beta_{20}^*]$ |
| Transformed Model: | $y^{**} = x^{**}_1 \beta_0^{**} + \epsilon_1^{**}$, $0 \leq y^{**} \leq \epsilon_1^{**}$ |

<table>
<thead>
<tr>
<th>Truncation Type</th>
<th>$z_{11}$</th>
<th>$z_{12}$</th>
<th>Constant</th>
<th>$y^{**}_1$</th>
<th>$x^{**}_1$</th>
<th>$\epsilon_0^{**}$</th>
<th>Linear Restrictions</th>
<th>Order of Restrictions</th>
<th>$X^{**}$</th>
<th>$z_1$</th>
<th>$\epsilon_1^{**}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Fixed Left</td>
<td>$z_1 &lt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_1 [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>No</td>
<td>K</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Fixed Right</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_2^* - y_1^* [1:y_2^*]$</td>
<td>$[z_{12} - z_{12}^* \beta_{20}^*]$</td>
<td>No</td>
<td>K</td>
<td>$\epsilon_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Fixed Double</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_1 [1:y_1^*]$</td>
<td>$[z_{12} - z_{12}^* \beta_{20}^*]$</td>
<td>No</td>
<td>K</td>
<td>$z_2 - z_1$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Fixed Left and Variable Right</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_1 [1:y_1^*]$</td>
<td>$[z_{12} - z_{12}^* \beta_{20}^*]$</td>
<td>No</td>
<td>K</td>
<td>$z_{12} - z_1$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Variable Left and Fixed Right</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $z_2 - y_1^* [1:y_2^*]$</td>
<td>$[z_{22} - z_{12}^* \beta_{20}^*]$</td>
<td>No</td>
<td>K</td>
<td>$z_{22} - z_{11}$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Variable Left</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Variable Right</td>
<td>$z_2 &gt; -$</td>
<td>$z_1 &lt; -$</td>
<td>Yes, $y_2^* - y_1^* [1:y_2^*]$</td>
<td>$[z_{22} - z_{22}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Variable Double</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{22} - z_{22}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$z_{12} - z_{11}$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Fixed Left</td>
<td>$z_2 &lt; -$</td>
<td>$z_1 &lt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Fixed Right</td>
<td>$z_2 &gt; -$</td>
<td>$z_1 &lt; -$</td>
<td>Yes, $y_2^* - y_1^* [1:y_2^*]$</td>
<td>$[z_{22} - z_{22}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Fixed Double</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$z_{22} - z_{11}$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Fixed Left and Variable Right</td>
<td>$z_1 &gt; -$</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$z_{22} - z_{11}$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Variable Left and Fixed Right</td>
<td>$z_2 &lt; -$</td>
<td>Yes, $z_2 - y_1^* [1:y_2^*]$</td>
<td>$[z_{22} - z_{22}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$z_{22} - z_{11}$</td>
<td>$\epsilon_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Variable Left</td>
<td>$z_1 &gt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Variable Right</td>
<td>$z_2 &gt; -$</td>
<td>Yes, $z_{12} - y_1^* [1:y_2^*]$</td>
<td>$[z_{12} - z_{12}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Variable Double</td>
<td>$z_2 &gt; -$</td>
<td>Yes, $y_1^* - z_{11} [1:y_1^*]$</td>
<td>$[z_{11} - z_{11}^* \beta_{20}^*]$</td>
<td>Yes</td>
<td>K+1</td>
<td>$z_{12} - z_{11}$</td>
<td>$\epsilon_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and $\varepsilon_i^{**}$ 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, $x_i^{**}$, (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 (#16), we employ the transformation,

$$y_i^{**} = y_i^* - z_{i1} = -z_{i1} \cdot 1 + x_i^t \delta_0 + \varepsilon_i$$

$$\equiv x_i^{**,1} \delta_0 + \varepsilon_i^{**}, \quad 0 < y_i^{**} < z_i^{**},$$

where

---

\(^1\) When fixed truncation or censoring at zero occurs -- either $z_{i1} = 0$ or $z_{i2} = 0$ for all $i$ -- a singularity in the matrix $[ E \; x_i \; x_i^t ]_{i \in S_2}$ 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 12 may be treated as Case 4 with $z_1 = 0$,
Case 13 may be treated as Case 5 with $z_2 = 0$, and
Case 11 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 $z_{j,j}, j = 1$ or 2, into $\delta_0$ to avoid the singularity.
\[ x_i^{**'} = [-z_{i1} : x_i'] \] ,

\[ \delta_0' = [1 : g_i'] \] ,

\[ \epsilon_i^{**} = \epsilon_i \]

and

\[ z_i^{**} = (z_{i2} - z_{i1}) < \infty . \]

It follows in all cases that, provided suitable restrictions are imposed on \( \delta_0' \) when required by \( \delta_0 \), consistent estimates of the elements of \( \delta_0' \) may be directly recovered from consistent estimates of \( \delta_0 \). Hence, the problem reduces to one of estimating \( \delta_0' \) and \( \sigma_0^2 \) in model (3.4), with

\[ (3.5) \quad \epsilon_i^{**} \sim \text{n.i.d.} \left(0, \sigma_0^2\right) . \]

Thus, to summarize, in (3.4) \( \delta_0' \) is a \( K^{**} \)-vector (with either \( K^{**} = K \) or \( K+1 \)), which may (cases #1 to #5) or may not (cases #6 to #10) be subject to a linear restriction, \( \delta_{10} = 1 \) with \( \delta_0 = \delta_{10} \). Finally, either

\[ z_i^{**} = \infty \text{ (Type 1)} \] or \[ z_i^{**} < \infty \text{ (Type 2)} . \]

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

The conditional density of \( y_i^{**} \) in (3.4), given \( 0 < y_i^{**} < z_i^{**} \), is defined by:
\[ g_{i0}(y_i^{**}) = \frac{f_{i0}^{**}(y_i^{**})}{[F_{i0}^{**}(z_i^{**}) - F_{i0}^{**}(0)]}, \quad 0 < y_i^{**} < z_i^{**}, \]

where

\[ f_{i0}^{**}(y_i^{**}) = \frac{1}{\sqrt{2\pi}\sigma_0} \cdot \exp \left\{ -\frac{1}{2\sigma_0^2} (y_i^{**} - \frac{z_i^{**}'}{z_i^{**}})^2 \right\}. \]

Thus, in this section, letting

\[ u_{i20}^{(r)} = E_{0} [y_i^{**r} | 0 < y_i^{**} < z_i^{**}] = \int_{0}^{z_i^{**}} y_i^{**r} g_{i0}^{**}(y_i^{**}) \, dy_i^{**}, \]

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:

\[ u_{i20}^{(r)} = u_{i20}^{(r-1)} \cdot \frac{x_i^{**}}{\sigma_0} + (r-1) \cdot \frac{\sigma_0^2}{\sigma_0^2} \cdot \frac{z_i^{**} f_{i0}^{**}(z_i^{**})}{[F_{i0}^{**}(z_i^{**}) - F_{i0}^{**}(0)]} \]

for \( z_i^{**} = \infty \) and \( r=1,2,... \). 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 \( z_i^{**} = \infty \), clearly \( f_{i0}^{**}(z_i^{**}) = 0 \) and, as in Amemiya (1973), equation (3.8) then reduces to:

\[ u_{i20}^{(r)} = u_{i20}^{(r-1)} \cdot \frac{x_i^{**}}{\sigma_0} + (r-1) \cdot \frac{\sigma_0^2}{\sigma_0^2}, \quad r=1,2,... \]

which is linear in the \((K^{**} + 1)\)-vector,

\[ \gamma_0 \equiv \left[ \begin{array}{c} \delta_0 \\ \sigma_0 \\ \vdots \\ \sigma_0 \\ \sigma_0^2 \end{array} \right], \]
for all \( r \). Thus, we may consider IV estimators based on (3.9) for cases \#1, \#2, 
\#6, \#7, \#9, \#10, \#14 and \#15 (in which \( z_i^{**} = \infty \)) and estimators based on (3.8) 
for the remaining cases of Table 1.

**Type 1 (Cases where \( z_i^{**} = \infty \)):**

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

\[
(3.11) \quad y_i^{**r} = (y_i^{**r-1} \cdot x_i^{**})' \cdot \delta_o + ((r-1)y_i^{**r-2}) \cdot \sigma_o^2 + \eta_i^{(r)}, \text{ for } i \in S_2^N,
\]

where

\[
(3.12) \quad \eta_i^{(r)} = \left[ y_i^{**r} - \mu_i^{(r)} \right] - \left[ y_i^{**(r-1)} - \mu_i^{(r-1)} \right] \cdot x_i^{**} \cdot \delta_o
\]

\[
- \left[ y_i^{**(r-2)} - \mu_i^{(r-2)} \right] \cdot \sigma_o^2.
\]

Note that

\[
(3.13) \quad E_{\delta} \eta_i^{(r)} = 0,
\]

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

\[
(3.14) \quad y_i^{**} = x_i^{**} \cdot \left[ \sum_{i \in S_2^N} x_i^{**} \cdot x_i^{**} \right]^{-1} \cdot \left[ \sum_{i \in S_2^N} x_i^{**} \cdot y_i^{**} \right],
\]

provided that \( \left[ \sum_{i \in S_2^N} x_i^{**} \cdot x_i^{**} \right] \) is positive definite. Then, for any \( r \geq 2 \),
using the \((K^{**} + 1)\) - vector, \([y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2}]\), as an
instrument for \([y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2}]\), define the class of
Unrestricted Instrumental Variable (UIV) estimators (see, e.g., Theil (1971)):

\[
(3.15) \quad \hat{y}_o = \left[ \begin{array}{c}
\hat{\delta}_o \\
\hat{\sigma}_o^2(t)
\end{array} \right] = \left[ \sum_{i \in S_2} \left[ y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2} \right] \cdot [y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2}] \right]^{-1}
\]

\[
\cdot \left[ \sum_{i \in S_2} \left[ y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2} \right] \cdot y_{i}^{**r} \right], \quad r = 2, 3, \ldots .
\]

This will immediately result in (weakly) consistent estimators for \(\delta_o\) (and
hence \(\sigma_o^2\)) as well as \(\sigma_o^2\) in cases \#1 and \#2 by arguments similar to those in

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

\[
(3.16a) \quad \hat{w}(r) = \left[ y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2} \right]_{N_2 \times K^{**}}
\]

\[
(3.16b) \quad w(r) = \left[ y_{i}^{**r-1} \cdot x_{i}^{**'} : (r-2) y_{i}^{**r-2} \right]_{N_2 \times K^{**}}
\]

\[
(3.16c) \quad \chi^{**}(r) = [y_{i}^{**r}]_{N_2 \times 1}
\]

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

\[
(3.17) \quad \hat{y}_o = \left[ \begin{array}{c}
\hat{\delta}_o \\
\hat{\sigma}_o^2(t)
\end{array} \right] = [\hat{w}(r)' \hat{w}(r)]^{-1} \cdot [\hat{w}(r)' \chi^{**}(r)]
\]

Suppose, however, that \(\delta_{10} = 1\), i.e.,
(3.18) \[ q' \gamma_0 = 1 \]

where \( q' = [10...0] \). Then, we must proceed to the Restricted Instrumental Variable (RIV) estimator,

(3.19) \[ \hat{\gamma}^{(r)} = \hat{\gamma}^{(r)} + [\hat{W}(r)' \hat{W}(r)]^{-1} \cdot q' [q' \hat{W}(r)' \hat{W}(r)]^{-1} \cdot \hat{\gamma}^{(r)} \cdot (1-\delta_{10}) \]

Note that in (3.19),

(3.20) \[ q' [\hat{W}(r)' \hat{W}(r)]^{-1} q = w_{11}(r) \]

where \( w_{ij} \) denotes the \((i,j)\) element of \([\hat{W}(r)' \hat{W}(r)]^{-1}\), so that (3.19) reduces to:

(3.21) \[ \hat{\gamma}^{(r)} = \hat{\gamma}^{(r)} + d^{(r)} \]

where

(3.22) \[ d^{(r)'} = \frac{(1-\delta_{10})}{w_{11}} \cdot [w_{11}^2 w_{21} \ldots w_{K*1}^2] \]

It is therefore evident that:

(3.23) \[ \delta_{10} = \delta_{10} + (1-\delta_{10}) = 1 \]

and the restriction is imposed on the estimates, \( \hat{\gamma}^{(r)} \gamma_0 \), for any \( r \geq 2 \).
Type 2 (Cases where \( z_i^{**} < \infty \)):

In cases where \( z_i^{**} < \infty \), we have \( f_i^{**}(z_i^{**}) > 0 \) and the last term on the RHS of (3.8) does not vanish. Hence, the approach of Amemiya with \( r=2 \) must be modified. Note that from (3.8) we may write:

\[
(3.24) \quad \left( \mu_i^{(r-1)} \right) - \left( \mu_i^{(r-2)} \right) = \left( \mu_i^{(r-1)} - \mu_i^{(r-2)} \right) \cdot \frac{x_i^{**'}}{\delta_o} + \left( \mu_i^{(r-2)} - \mu_i^{(r-3)} \right) \cdot \sigma_0^2,
\]

which is, again, linear in \( \delta_o \) and \( \sigma_0^2 \) and eliminates the last term on the RHS of (3.8). Thus, for any \( r \geq 3 \),

\[
(3.25) \quad \left( y_i^{**r} - z_i^{**r-1} \right) = \left( y_i^{**r-1} - y_i^{**r-2} \right) \cdot \frac{x_i^{**'}}{\delta_o} + \left( y_i^{**r-2} - y_i^{**r-3} \right) \cdot \sigma_0^2 + \eta_i^{(r)}, \text{ \( i \in S_2 \)},
\]

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

\[
(3.26) \quad \eta_i^{(r)} = \left[ y_i^{**r} - \mu_i^{(r)} \right] - \left( z_i^{**} + x_i^{**'} \delta_o \right) \cdot \left[ y_i^{**r-1} - \mu_i^{(r-1)} \right]
+ \left( z_i^{**} x_i^{**'} \delta_o - (r-1)\sigma_o^2 \right) \cdot \left[ y_i^{**r-2} - \mu_i^{(r-2)} \right] + (r-2)\sigma_o^2 \cdot \left[ y_i^{**r-3} - \mu_i^{(r-3)} \right]
\]

and, by convention, \( y_i^{**0} = \mu_i^{(0)} = 1 \). Clearly, \( E_o \eta_i^{(r)} = 0 \) and \( E_o \eta_i^{(r)2} \) is bounded.

In the cases where \( \delta_{10} \) 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) \[ \hat{\omega}(r) = [(y_i^{**-r-2} - z_i y_i^{**-r-2}) \cdot \hat{x}_i \cdot (r-1)y_i^{**-r-2} - (r-2)z_i y_i^{**-r-3} ] \]

(3.27b) \[ \hat{\omega}(r) = [(y_i^{**-r-1} - z_i y_i^{**-r-2}) \cdot \hat{x}_i \cdot (r-1)y_i^{**-r-2} - (r-2)z_i y_i^{**-r-3} ] \]

(3.27c) \[ \hat{\gamma}(r) = [(y_i^{**-r} - z_i y_i^{**-r-1}) ] \]

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 \( \gamma_o \) (and, hence, \( \delta_o \) and \( \sigma_o^2 \)) may be derived from the relation:

(3.28) \[ (\mu_{i20}^{(r)} - z_i^{**} \mu_{i20}^{(r-s)}) = (\mu_{i20}^{(r-1)} - z_i^{**} \mu_{i20}^{(r-s-1)}) \cdot \hat{x}_i \cdot \frac{\delta_o}{\delta_o} \]

\[ + ((r-1)\mu_{i20}^{(r-2)} - (r-s-1)\mu_{i20}^{(r-s-2)}) \cdot \sigma_o^2 \]

for \( r=2,3,4,\ldots \) and \( s=0,1,2,\ldots,r-2 \) in Type 1 problems, and \( r=3,4,\ldots \) and \( s=1,2,\ldots,r-2 \) in Type 2 problems. Any feasible choice of \( r \) and \( s \) in our formulation preserves the linear relation in \( \delta_o \) and \( \sigma_o^2 \), while eliminating the last term on the RHS of (3.28). Thus we may define our general, weakly consistent, initial estimator of \( \gamma_o \) as:

(3.29a) \[ \hat{\gamma}_o = \hat{\gamma}_o^{(r,s)} \]

\[ \frac{\delta_o}{\delta_o} \]

\[ \frac{\sigma_o^2}{\sigma_o^2} \]

if Type 1

\[ \frac{\gamma_o}{\gamma_o} \]

\[ \frac{\gamma_o}{\gamma_o} \]

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, \( \hat{\theta}_0 \), which we shall represent as:

\[
\hat{\theta}(r,s) \equiv \frac{\hat{\beta}(r,s)}{\hat{\sigma}^2(r,s)} \ldots.
\]

\[
(3.29b) \quad \hat{\sigma}_0(r,s) \equiv \frac{\hat{\sigma}(r,s)}{\hat{\sigma}_0^2(r,s)} \ldots.
\]

3.3. Improvements to the Initial Variance Estimator:

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

\[
(3.30a) \quad \hat{\sigma}_0^2(r,s) = \begin{cases} \hat{\sigma}_0^2(r,s) & \text{if } \hat{\sigma}_0^2(r,s) > 0 \\ \omega^2(r,s) & \text{if } \hat{\sigma}_0^2(r,s) \leq 0 \end{cases},
\]

where

\[
(3.30b) \quad \omega^2(r,s) \equiv \frac{1}{N_2} \sum_{i \in S_2} (y_i^* - x_i^* \hat{\beta}(r,s))^2.
\]

\(^1\) In practice, negative values of \( \hat{\sigma}_0^2(r,s) \) for any \( (r,s) \) choice are quite common—see the discussion of computational experience in section VI below and Hartley and Swanson (1985). Such values clearly violate Assumption 1, i.e., lie outside of \( \theta \), and thus are infeasible.

\(^2\) In some cases, it may even be preferable to employ a censored variance estimator in which \( \hat{\sigma}_0^2(r,s) \) is observed as \( \hat{\sigma}_0^2(r,s) \) when the latter exceeds some \( \eta > 0 \) and is \( \omega^2(r,s) \) otherwise.
By construction, $\hat{\sigma}_o^{-2}(r,s)$ is always positive and its weak consistency follows from the fact that $\hat{\sigma}_o^{-2}(r,s) + \sigma_o^2 > 0$ as $N \to \infty$. \footnote{We are indebted to Professor Marcello Pagano for this point.}

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 occasion, either fail to converge or converge rather slowly--see Section VI below and Hartley and Swanson (1985). A plausible solution to this difficulty is the selective improvement of the adjusted variance estimator, conditional on the consistent initial coefficient estimates $\hat{\beta}_o (\equiv \hat{\beta}_o)$, 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:

$$(3.31) \quad \hat{\sigma}_o^{-2}(r,s) = \sup_{\sigma^2 > 0} \{ \log I_N(\sigma^2 | \hat{\beta}_o), M=T,C \}.$$ 

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

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.

IV. 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.,

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

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,

\[
\Sigma^M(\theta^0) = \lim_{N \to \infty} \mathbb{E}^0 \left[ \frac{\partial^2 \log L^M_N(\theta^0)}{\partial \theta \partial \theta'} \right]^{-1}
\]

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 \( E^{M(g_o)}^{-1} \) has the general form,

\[
E^{M(g_o)}^{-1} = \lim_{N \to \infty} \left[ \begin{array}{cccc}
N & a_i^M & x_i & x_i' \\
i=1 & b_i^M & x_i & x_i' \\
i=1 & N & c_i & x_i \\
i=1
\end{array} \right], \quad M=T,C,
\]

where, using the definition \( g_{ij0} = g_{ij}(z_{ij}), \ j=1,2, \) we find that \( T \)

\[ a_i^T = -\frac{1}{2 \sigma_o} \cdot \left( 1 - (u_{i20} g_{i20} - u_{i10} g_{i10}) - \sigma_o^2 (g_{i20} - g_{i10})^2 \right) \]

(4.3a)

\[ b_i^T = \frac{1}{2 \sigma_o^4} \cdot \left( [(u_{i20}^2 + \sigma_o^2)g_{i20} - (u_{i10}^2 + \sigma_o^2)g_{i10}] 
\]

\[ + \sigma_o^2 (g_{i20} - g_{i10})(u_{i20} g_{i20} - u_{i10} g_{i10}) \right) \]

(4.3b)

\[ c_i^T = -\frac{1}{4 \sigma_o^6} \cdot \left( 2 \sigma_o^2 - (u_{i20}^3 g_{i20} - u_{i10}^3 g_{i10}) - \sigma_o^2 (u_{i20} g_{i20} - u_{i10} g_{i10}) 
\]

\[ - \sigma_o^2 (u_{i20} g_{i20} - u_{i10} g_{i10})^2 \right) \]

(4.3c)

and

\[ a_i^C = -\frac{1}{2 \sigma_o^2} \cdot \left( \sigma_o^2 \cdot \left( \frac{f_{i20}^2}{(1-F_{i20})} + \frac{f_{i10}^2}{F_{i10}} \right) - (u_{i20} f_{i20} - u_{i10} f_{i10}) 
\]

(4.4a)

\[ + (F_{i20} - F_{i10}) \right) \]
\[(4.4b) \quad b_i^C = - \frac{1}{2C^2} \cdot \left( \frac{u_{i20} f_{i20}^2}{1-F_{i20}} + \frac{u_{ilo} f_{ilo}^2}{F_{ilo}} \right) - \sigma_o^2 \cdot (f_{i20} - f_{ilo}) \]

\[ - (u_{i20}^2 f_{i20} - u_{ilo}^2) \]

\[(4.4c) \quad c_i^C = - \frac{1}{4\sigma_o^4} \left( 3 \cdot (u_{i20} f_{i20} - u_{ilo} f_{ilo}) - \frac{1}{2} \cdot (u_{i20}^3 f_{i20} - u_{ilo}^3 f_{ilo}) \right) \]

\[ + \frac{u_{i20}^2 f_{i20}^2}{1-F_{i20}} + \frac{u_{ilo}^2 f_{ilo}^2}{F_{ilo}} + 2(f_{i20} - f_{ilo}) \]

\[- 4 \cdot (u_{i20} f_{i20} - u_{ilo} f_{ilo}) \] .

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

For all cases, in finite samples, \( \Sigma^M(\theta) \) may be consistently estimated by removing the limit sign from \( (4.2) \) and evaluating \( a_u^M, b_i^M \) and \( c_i^M \) at the MLE, \( \hat{\theta}^M, M=T,C. \)

Finally, Amemiya (1973) establishes in the singly-censored normal case \( (M=C) \) that under general assumptions: (c) the "second-round" estimator, \( \hat{\theta}_1^1 \) (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 of that of 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 \( \sigma_o^2 = 1 \) and
\[
u_i \equiv u_{i0} = -x_i' \beta_o,
\]
we have
\[
\varepsilon C(\beta_o)^{-1} = -\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N a_i^C \cdot x_i x_i',
\]
in which
\[
a_i^C = \left( \frac{f_{i0}(0)}{F_{i0}(0) \cdot [1 - F_{i0}(0)]} \right).
\]

V. Calculation of the MLE:

In this and the subsequent section we consider four methods for the
calculation of the MLE, \( \hat{\theta}^M \), 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-Newton-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 \( \hat{\theta}_2^* = \left[ \begin{array}{c} \hat{\theta}_2^* \\ \hat{\sigma}_2^* \end{array} \right] \) denote the value of \( \theta \in \Theta \) in iteration \( k \), \( k=0,1,2,... \).

For any function, \( q_i \equiv q(y^*_i; x_i, \theta) \), let \( q_i^* \equiv q(y^*_i; x_i, \hat{\theta}_2^*) \). Let \( E_{\hat{\theta}_2^*} \) denote the
unconditional expectation operator relative to \( \hat{\theta}_2^* \) and \( E_{\hat{\theta}_2^*} \) denote the
corresponding conditional expectation operator, given \( y^*_i \in Y_i \).

Thus,

\[
E_{\hat{\theta}_2^*} q_i = \int_{-\infty}^{\infty} q(y^*_i; x_i, \theta) \cdot f_i^*(y^*_i) \, dy^*_i,
\]

and, using (2.23) with the conditional density, \( v_{ijh}(y^*_i) \), we define
\[ (5.1b) \quad E_{j \lambda} q_i = \int_{Y_j} q(y_i^*; x_i, \theta) \cdot u_{ij \lambda}(y_i^*) \, dy_i^* , \]

so that, symbolically, we have the customary relation,

\[ E_{\lambda} q_i = F_{i1\lambda} \cdot E_{1\lambda} q_i + [F_{i2\lambda} - F_{i1\lambda}] \cdot E_{2\lambda} q_i + [1 - F_{i2\lambda}] \cdot E_{3\lambda} q_i . \]

Each of our algorithms may be represented in canonical form by the following recursion formula:

\[ (5.3) \quad \hat{\theta}_{\lambda+1}^m = \hat{\theta}_{\lambda}^m + \lambda_{\lambda}^m \cdot A_{\lambda}^m \cdot \hat{b}_{\lambda}^m , \]

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

\(^1\) Each of these algorithms may be directly applied to the Probit \((M=C)\) model, where \( \hat{\sigma}_\lambda^2 = 1 \) for all \( \lambda \), \( \hat{\theta}_{\lambda}^m = \hat{\theta}_{\lambda}^m \) and \( z_{i1} = z_{i2} = 0 \).
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:

\[ A_{\theta}^{1} = - \left[ \frac{\partial^2 \log L_{N}^{M}(\hat{\theta}_{\theta})}{\partial \theta \partial \theta'} \right]^{-1}, \quad M = T, C, \]

\[ b_{\theta}^{1} = \frac{\partial \log L_{N}^{M}(\hat{\theta}_{\theta})}{\partial \theta}, \quad M = T, C \]

and \( \chi_{\theta}^{1} = 1 \), if in the unmodified form. Expressions for the elements of \( A_{\theta}^{1} \) and \( b_{\theta}^{1} \) 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 \( M = C \).

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, \( \{ \hat{\theta}_{\ell} : \ell=0,1,2,... \} \), will ever converge--much less to the root of (3.1) corresponding to the global maximum of the log-likelihood function.1/

---

1/ In practical applications--see section VI -- the lack of convergence of the N-R algorithm, in our experience, is not uncommon, particularly in heavily censored samples.
5.2. The Method of Scoring:

The Method of Scoring, 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 \( \hat{\theta} \)). Thus, noting (4.1), (4.2), (4.3a) - (4.3c) and (4.4a) - (4.4c), we may define:

\[
(5.5a) \quad A_\|^2 = - \left[ \frac{\partial^2 \log L_N^{M}(\theta)}{\partial \theta \, \partial \theta'} \right]^{-1}.
\]

whereas

\[
(5.5b) \quad b_\|^2 = b_\| \]

and, if unmodified, \( \chi^2_\|= 1 \). The unmodified M-S algorithm shares both the advantage and the disadvantage of the N-R method previously noted.

5.3. The Gauss-Newton Method:

The Gauss-Newton method employs:

\[
(5.6a) \quad A_\|^3 = \left[ \sum_{i=1}^{N} \frac{\partial \log p_i^{M}(y_i; \hat{\theta}_\|)}{\partial \theta} \cdot \frac{\partial \log p_i^{M}(y_i; \hat{\theta}_\|)}{\partial \theta'} \right]^{-1}
\]

and, once again,

\[
(5.6b) \quad b_\|^3 = b_\|^1 \]

where \( p_i^{M} \) is the appropriate density function,

\[
(5.7) \quad p_i^{M}(y_i; \theta) = \begin{cases} g_i(y_i; \theta), & \text{if } M=T \\ h_i(y_i; \theta), & \text{if } M=C \end{cases}
\]
and (in the unmodified case) $\lambda_3^3 = 1$. Since $\frac{2 \log L_N^M(\theta)}{2 \theta}$

$$\sum_{i=1}^{N} \frac{2 \log p_i^M(y_i; \theta)}{2 \theta}, \text{ the expressions required for } \frac{2 \log g_i(y_i; \theta)}{2 \theta} \text{ or } \frac{2 \log h_i(y_i; \theta)}{2 \theta}$$

are given in (2.19a) and (2.19b) or (2.21a) and (2.21b), respectively, except that the summation operator must be removed from the right hand side. Methods for choosing $\lambda_3^3$ 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 $\hat{\theta}_3^2 \theta_o$, then

$$\lim_{\lambda \to \infty} A_3^3 + \left( E_0 \frac{2 \log L_N^M(\theta)}{2 \theta} \right)^{-1} = \lim_{\lambda \to \infty} A_3^2$$

(Berndt, Hall, Hall and Hausman (1974)), permits either $A_3^1$, $A_3^2$ or $A_3^3$, evaluated at $\hat{\theta}_3^M = \lim_{\lambda \to \infty} \hat{\theta}_3$, to serve as the asymptotic covariance matrix of $
\sqrt{N} (\hat{\theta}_3^M - \theta_o )$, $M=T,C$. Our practice is to use $A_3^2$ 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 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 \( y^*_i \) given by (2.1) and (2.2). Conceptually, consider a given set of \( N \) values for the independent variables, \( \{x_i : i=1, \ldots, N\} \), and, for each observation on \( x_i \), let the corresponding \( y^*_i \) value be generated by adding a random drawing, \( \epsilon_i \), from the density \( n(0, \sigma^2) \) to the true regression function \( x'_i \beta \). We may write the log-likelihood function in the "complete data" case for a sample of size \( N \), \( \{y^*_i, x_i \} \), as:

\[
(5.9) \quad \log L^*_N(\beta; y^*) = \frac{N}{2} \log f_i(y^*_i) = \frac{3}{2} \sum_{j=1}^{3} \sum_{i \in S_j} \log f_i(y^*_i) ,
\]

where \( S_j \) 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 \( \beta \) and \( \sigma^2 \) 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

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

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 \( f_{10}(y^*_1) \), only the pairs, 
\( (y^*_i, x_i) : i \in S_2^N \), are actually available; and all sample information on the 
remaining drawings, \( (y^*_i, x_i) : i \in S_1^N \) 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 \( y^*_i \geq z_{12} \) (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 \( y_{i}^{*} \epsilon Y_{i} \) (and relative to an arbitrary \( \theta \)), by

\[
(5.11) \quad \xi_{ij}^{(r)} = \mathbb{E}[(y_{i}^{*} - x_{i}^{*} \theta)^{r} | y_{i}^{*} \epsilon Y_{i}]
\]

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

\[
(5.12) \quad \xi_{ij}^{(1)} = u_{ij}^{(1)} - x_{i}^{*} \theta = \begin{cases} 
- \sigma^2 \cdot \frac{f_{ii}^{(1)}}{F_{ii}} & \text{, if } j = 1 \\
- \sigma^2 \cdot \frac{f_{i2}^{(1)} - f_{i1}^{(1)}}{F_{i2} - F_{i1}} & \text{, if } j = 2 \\
\sigma^2 \cdot \frac{f_{i2}^{(1)}}{1 - F_{i2}} & \text{, if } j = 3
\end{cases}
\]

and

\[
(5.13) \quad \xi_{ij}^{(2)} = u_{ij}^{(2)} - 2(x_{i}^{*} \theta) \cdot u_{ij}^{(1)} + (x_{i}^{*} \theta)^2 = \begin{cases} 
\sigma^2 \cdot (1 - \frac{u_{i1} f_{i1}^{(1)}}{F_{i1}}) & \text{, if } j = 1 \\
\sigma^2 \cdot (1 - \frac{u_{i2} f_{i2}^{(1)} - u_{i1} f_{i1}^{(1)}}{F_{i2} - F_{i1}}) & \text{, if } j = 2 \\
\sigma^2 \cdot (1 + \frac{u_{i2} f_{i2}^{(1)}}{1 - F_{i2}}) & \text{, if } j = 3
\end{cases}
\]

respectively. Note, as a check, that with \( q_{i} = (y_{i}^{*} - x_{i}^{*} \theta)^{r} \), equation (5.2)
implies that the corresponding first two unconditional moments are given by 
\( \xi_1^{(1)} = 0 \) and \( \xi_1^{(2)} = \sigma^2 \), 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:

\[
(5.14) \quad \log L_N^C(\theta) = \sum_{i \in S_1} \int_{-\infty}^{y_{i1}^*} (\log f_i(y_{i1}^*) - \log u_{i1}(y_{i1}^*)) \cdot v_{i1}(y_{i1}^*) \, dy_i^* \\
+ \sum_{i \in S_2} \log f_i(y_{i1}^*) \\
+ \sum_{i \in S_3} \int_{y_{i12}^*}^{y_{i13}^*} (\log f_i(y_{i1}^*) - \log u_{i13}(y_{i1}^*)) \cdot v_{i13}(y_{i1}^*) \, dy_i^* \\
= \sum_{i \in S_1} E_1(\log f_i - \log u_{i1}) + \sum_{i \in S_2} \log f_i + \sum_{i \in S_3} E_3(\log f_i - \log u_{i13}).
\]

Then the formulas for the E-M algorithm can be viewed as being obtained by
maximizing the "pseudo log-likelihood",

\[
(5.15) \quad \Lambda^C(\theta | \hat{\theta}_2) = \sum_{i \in S_1} E_1(\log f_i - \log u_{i1}) + \sum_{i \in S_2} \log f_i \\
+ \sum_{i \in S_3} E_3(\log f_i - \log u_{i13}),
\]

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

\[
(5.16) \quad \Lambda^M(\hat{\theta}_{2+1} | \hat{\theta}_2) = \sup_{\theta \in \mathcal{O}} (\Lambda^M(\theta | \hat{\theta}_2)).
\]
It may easily be verified that the resulting "pseudo likelihood" equations are given by:

\[
(5.17a) \quad \frac{\partial \Lambda^C(\hat{\theta}_k|\hat{\theta}_k)}{\partial \theta} = \frac{1}{2\sigma} \left( \sum_{i \in S_1} \int_{-\infty}^{\infty} (y_i^* - x_i^\prime \hat{\theta}) \cdot u_{i1k}(y_i^*) \, dy_i^* \cdot x_i + \sum_{i \in S_2} (y_i^* - x_i^\prime \hat{\theta}) \cdot x_i \right) \\
+ \sum_{i \in S_3} \int_{0}^{\infty} (y_i^* - x_i^\prime \hat{\theta}) \cdot u_{i13k}(y_i^*) \, dy_i^* \cdot x_i \right) = 0
\]

and

\[
(5.17b) \quad \frac{\partial \Lambda^C(\hat{\theta}_k|\hat{\theta}_k)}{\partial \sigma^2} = \frac{1}{2\sigma^4} \left( \sum_{i \in S_1} \int_{-\infty}^{\infty} [(y_i^* - x_i^\prime \hat{\theta})^2 - \sigma^2] \cdot u_{i1k}(y_i^*) \, dy_i^* \\
+ \sum_{i \in S_2} [(y_i^* - x_i^\prime \hat{\theta})^2 - \sigma^2] \cdot u_{i3k}(y_i^*) \, dy_i^* \right) = 0 \\
+ \sum_{i \in S_3} \int_{0}^{\infty} [(y_i^* - x_i^\prime \hat{\theta})^2 - \sigma^2] \cdot u_{i13k}(y_i^*) \, dy_i^* \right) = 0
\]

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:\(^{1/}\)

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

and

\(^{1/}\) Note that for the Probit model, equation (5.18a) is sufficient to define the E-M algorithm.
\begin{equation}
\sigma^2_{\ell + 1} = \frac{1}{N} \sum_{i=1}^{N} s_{i\ell}^{2C},
\end{equation}

where

\begin{equation}
\gamma^C_{i\ell} = \left\{ \begin{array}{ll}
\mu_{i\ell}^{(1)}, & \text{if } i \in S_1^N \\
\gamma^*_{i}, & \text{if } i \in S_2^N \\
\mu_{i\ell}^{(1)}, & \text{if } i \in S_3^N 
\end{array} \right.
\end{equation}

and

\begin{equation}
s_{i\ell}^{2C} = \left\{ \begin{array}{ll}
\xi_{i\ell}^{(2)} - 2\left[ x_i^2 \left( \hat{\theta}_{\ell+1} - \hat{\theta}_{\ell} \right) \right] - \xi_{i\ell}^{(1)} + \left[ x_i^2 \left( \hat{\theta}_{\ell+1} - \hat{\theta}_{\ell} \right) \right]^2, & \text{if } i \in S_1^N \\
\left( \gamma^*_{i} - x_i \hat{\theta}_{\ell+1} \right)^2, & \text{if } i \in S_2^N \\
\xi_{i\ell}^{(2)} - 2\left[ x_i^2 \left( \hat{\theta}_{\ell+1} - \hat{\theta}_{\ell} \right) \right] - \xi_{i\ell}^{(1)} + \left[ x_i^2 \left( \hat{\theta}_{\ell+1} - \hat{\theta}_{\ell} \right) \right]^2, & \text{if } i \in S_3^N
\end{array} \right.
\end{equation}

Several points are worth noting. First, it may easily be verified that, if \( \theta_{\ell+1} = \theta_{\ell} = \theta \) and \( M=C \), then both

\begin{equation}
\Lambda^M(\theta | \theta) = \log L^M_N(\theta)
\end{equation}

and

\begin{equation}
\frac{\partial \Lambda^M(\theta | \theta)}{\partial \theta} = \frac{\partial \log L^M_N(\theta)}{\partial \theta}.
\end{equation}

Hence, if the sequence of solutions to the "pseudo-likelihood" equations, (5.17a) and 5.17b), defined by \( \left\{ \hat{\theta}_{\ell} : \ell = 0, 1, 2, \ldots \right\} \), converges, then it converges to a solution to the original likelihood equations, (3.1) with \( M=C \). Second, by combining equations (5.16) and (5.21a) with \( \theta = \hat{\theta}_{\ell} \), we have for \( M=C \) and for any \( \ell \),
(5.22a) \[ \log L_N^M(\hat{\theta}_k) = \Lambda_N^M(\hat{\theta}_k | \hat{\theta}_k) \leq \Lambda_N^M(\hat{\theta}_{k+1} | \hat{\theta}_k) \]

whereas

(5.22b) \[ \Lambda_N^M(\hat{\theta}_{k+1} | \hat{\theta}_k) \leq \Lambda_N^M(\hat{\theta}_{k+1} | \hat{\theta}_{k+1}) = L_N^M(\hat{\theta}_{k+1}) \]

The inequality, (5.22b), holds by virtue of the fact that \( u_{i1,k+1} \) and \( u_{i3,k+1} \) are precisely the solutions to the calculus-of-variations problem of maximizing the "pseudo log-likelihood" function,

(5.23a) \[ \Lambda_N^M(u_1, u_3 | \hat{\theta}_{k+1}) = \sum_{i \in S_1} E_1(\log f_i, \omega_{k+1} - \log u_{i1}) + \sum_{i \in S_2} \log f_i, \omega_{k+1} \]

\[ + \sum_{i \in S_3} E_3(\log f_i, \omega_{k+1} - \log u_{i2}) \]

with respect to the \( N_j \)-vector of functionals, \( u_j = [u_{ij}(y_i^*)] \), \( j=1,3 \), subject to the set of so-called "iso-perimetric conditions",

(5.23b) \[ \int_{Y_{ij}} u_{ij}(y_i^*) \, dy_i^* = 1 , \quad j = 1,3 \]

Thus, (5.22a) and (5.22b), together, imply that both \( \{\log L_N^M(\hat{\theta}_k)\} \) and \( \{\Lambda_N^M(\hat{\theta}_{k+1} | \hat{\theta}_k)\} \) 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 either \( |\theta| \to \infty \), \( \sigma^2 \to 0 \) or \( \sigma^2 \to \infty \), it follows that the sequence of E-M parameter points, \( \{\theta_k : k=0,1,\ldots\} \), always remains within a bounded space. Thus, \( \{\Lambda_N^M(\hat{\theta}_{k+1} | \hat{\theta}_k)\} \) 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,
\( \lim_{l \to \infty} \hat{\theta}(l) = \hat{\theta}^N \),

where \( \hat{\theta}^N \) 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 \( \hat{\theta}_{l+1} \) is determined as a least squares regression of \( y^C \) on \( X \), where \( y^C \) is defined as the actual \( y^* \) value (when \( i \in S_2^N \)), and by the expected value of \( y^* \), given knowledge of the \( x^*_i \) value and the information that \( y^*_i \in Y_j \), evaluated relative to \( \hat{\theta}_l \), for \( \hat{\theta}_j \in S_j^N \), \( j = 1 \) or 3. The latter elements of the vector, \( y^C \), must therefore be updated each iteration, whereas the matrix \( R \) is invariant with \( l \). Hence, once \( R \) has been initially calculated, further matrix inversion in iterations, \( l = 1, 2, \ldots \), is avoided, and \( \hat{\theta}_{l+1} \) is simply calculated as \( R_y^C \). 1/

Turning now to completion of the algorithm via \( \hat{\sigma}^2_{l+1} \) of (5.18b), with the \( \{s_{i,k}^2\} \) of (5.20), 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, \( \hat{\sigma}^2_{l+1} \) consists of the sum of squared residuals, \( (y^*_i - x^*_i \hat{\theta}_l)^2 \), for all \( i \in S_2^N \), and the expected value of such squared residuals, given \( y^*_i \in Y_j \), relative to \( \hat{\theta}_l \), for all \( i \in S_j^N \), \( j = 1 \) or 3. Hence, \( \hat{\sigma}^2 \) is always positive for any \( l \)—a property not shared with unmodified versions of the N-R, M-S and G-N algorithms.

---

1/ In applications where the number of regressors, \( K \), is large (see Hartley and Swanson (1985)), avoiding repeated inversion of the \( K \times K \) 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, M-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.
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:

\[
(5.25) \quad \log L_N^T(\theta) = \sum_{i \in S_2} \{ \log f_i(y_i^*) - \int_{y_i^*}^{\infty} \log f_i(y) - \log u_{i2}(y) \} \cdot u_{i2}(y) dy_i
\]

\[
= \sum_{i \in S_2} \{ \log f_i - E_2[\log f_i - \log u_{i2}] \}.
\]

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

\[
(5.26) \quad \Lambda^T(\theta | \hat{\theta}_x) = \sum_{i \in S_2} \{ \log f_i - E_2[\log f_i - \log u_{i2}] \}
\]

with respect to \( \theta \), holding the functions \( \{u_{i2}\} \) of (5.14) fixed at \( \hat{\theta}_x \). Again, the solution for \( \hat{\theta}_{x+1} \), which satisfies (5.16) with \( M=T \), is unique and defined by the iteratively-reweighted least squares algorithm:

\[
(5.27a) \quad \hat{\theta}_{x+1} = [X'(F_{2x} - F_{1x})^{-1}X]^{-1} \cdot [X'(F_{2x} - F_{1x})^{-1}X']^T
\]

and the expected residual sum of squares,

\[
(5.27b) \quad \hat{\sigma}_{x+1}^2 = \frac{1}{\sum_{i \in S_2} (F_{i2x} - F_{i1x})^{-1}} \cdot \sum_{i \in S_2} s_{i2x}^2
\]

where

\[
(5.28) \quad F_{2x} - F_{1x} = \text{Diag} [F_{i2x} - F_{i1x}],
\]

\[
(5.29) \quad Y_{x}^T = [y_{i1x}^T] = F_{i1x} \cdot \mu_{i1x}^{(1)} + [F_{2x} - F_{1x}] \cdot \gamma^x + [I - F_{2x}] \cdot \mu_{i3x}^{(1)},
\]

\[
\]
(5.30) \[ \mu_{jz}^{(1)} = [\mu_{ij}^{(1)}] = E_{jz} \gamma^* \text{, } j = 1 \text{ or } 3 \text{ ,} \]

and

(5.31) \[ s_{i\ell \ell}^{2T} = \frac{1}{F_{i2\ell} - F_{i1\ell}} \cdot \left[ \frac{1}{E_{i2\ell} - E_{i1\ell}} \cdot \left( y_i^* - x_i^* \hat{\theta}_{z+1} \right)^2 + \left( y_i^* - x_i^* \hat{\theta}_{z+1} \right)^2 \right] 

+ \left[ 1 - F_{i2\ell} \right] \cdot E_{3\ell} \left( y_i^* - x_i^* \hat{\theta}_{z+1} \right)^2 \}

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

(5.32) \[ s_{i\ell \ell}^{2T} = \left( y_i^* - x_i^* \hat{\theta}_{z+1} \right)^2 + \left( 1 - \frac{F_{i2\ell} - F_{i1\ell}}{F_{i2\ell} - F_{i1\ell}} \right) \cdot \left( \hat{\sigma}_z^2 + [x_i^*(\hat{\theta}_{z+1} - \hat{\theta}_z)] \right) \]

+ \left( \frac{\hat{\sigma}_z^2}{F_{i2\ell} - F_{i1\ell}} \right) \cdot \left( u_{i2\ell} f_{i2\ell} - u_{i1\ell} f_{i1\ell} \right) 

- 2 \left[ x_i^*(\hat{\theta}_{z+1} - \hat{\theta}_z) \right] \cdot \left( f_{i2\ell} - f_{i1\ell} \right) \}

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 \( M = T \).

Thus the E-M algorithm in the doubly-truncated case, by iteratively maximizing \( \Lambda_T(\theta | \hat{\theta}_z) \) with respect to \( \hat{\theta}_z \), and, defining the solution as \( \hat{\theta}_{z+1} \), yields a sequence, \( \{ \hat{\theta}_z : z = 0, 1, \ldots \} \), containing a subsequence which converges to a solution of (3.1) with \( M = T \) and yields a (local) maximum of \( L_N^T(\theta) \).

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, \( i \in S_2^N \), which may be viewed as a random drawing from the (conditional) p.d.f., \( \psi_{i20}(y_i^*) = \frac{f_{i0}(y_i)}{F_{i20} - F_{i0}} \). Note that, relative to \( \hat{\theta}_0 \), \( y_i^* \) has the (conditional) expectation,
(5.33) \[ E_{20} y_i^* = \mu_{120} = x_i' \theta_o - \sigma_o^2 \cdot \frac{f_{120} - f_{110}}{f_{120} - f_{110}} \neq x_i' \theta_o. \]

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

\[
\Pr[y_i^* \in Y_j] = \begin{cases} 
F_{i10} & \text{if } j = 1 \\
\frac{[F_{120} - F_{i10}]}{[F_{120} - F_{i10}]} & \text{if } j = 2 \\
[1 - F_{120}] & \text{if } j = 3
\end{cases}
\]

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

\[
\hat{N}_{i0} = \begin{cases} 
\sum_{i \in S_2} \frac{F_{i10}}{[F_{120} - F_{i10}]} & \text{if } j = 1, \\
N_2 & \text{if } j = 2, \\
\sum_{i \in S_2} \frac{[1 - F_{120}]}{[F_{120} - F_{i10}]} & \text{if } j = 3.
\end{cases}
\]
With these preliminaries in hand, it is now easy to motivate the E-M algorithm. Suppose our current estimate of \( \hat{\theta} \) after \( \ell \) iterations is \( \hat{\theta}_\ell \). Then, we see that, insofar as the functionals, \( u_{ij\ell} \), \( j=1,2 \) and \( 3 \), are fixed when \( \hat{\theta}_\ell \) is given, the "pseudo log-likelihood," \( \Lambda^T(\hat{\theta}|\hat{\theta}_\ell) \) of (5.26), may be rewritten more instructively (using (5.2)) as:

\[
(5.37) \quad \Lambda^T(\hat{\theta}|\hat{\theta}_\ell) = \sum_{i \in \mathbb{S}_2^N} \frac{1}{F_{i2\ell} - F_{i1\ell}} \cdot \left( F_{i1\ell} \cdot E_{1\ell} \left( \log f_i - E_{1\ell} \log f_i \right) \right. \\
+ \left. \left[ F_{i2\ell} - F_{i1\ell} \right] \cdot \left( \log f_i - E_{2\ell} \log f_i \right) + \left[ 1 - F_{i1\ell} \right] \cdot E_{3\ell} \left( \log f_i - E_{3\ell} \log f_i \right) \right) \\
+ \sum_{i \in \mathbb{S}_2^N} E_{2\ell} \cdot u_{12\ell} 
\]

The last term on the RHS of (5.37) in 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 \( \hat{N}_2 = \sum_{i \in \mathbb{S}_2^N} \left( F_{i2\ell} - F_{i1\ell} \right)^{-1} \) "pseudo observations" on the deviations, \( \left[ \log f_i - E_{j\ell} \log f_i \right] \), each weighted by the probability that \( y_{i}^* \in Y_j \), and with the "incomplete" portions of the original log-likelihood (when \( j=1 \) or \( j=3 \)) replaced by their expectations, \( E_{j\ell} \left[ \log f_i - E_{j\ell} \log f_i \right] \), which are identically zero.

It follows that the value of \( \hat{\theta} \) which partially maximizes \( \Lambda^T(\hat{\theta}|\hat{\theta}_\ell) \) is obtained by the weighted least squares regression of the "pseudo dependent variable", \( y_{\ell}^T \), defined as the suitably-weighted average of the observed dependent variable, \( y_* \), within \( Y_2 \) and the conditional expectations of the missing values, \( u_{12\ell} \) and \( u_{32\ell} \), within \( Y_1 \) and \( Y_3 \), with "weights," \( F_{i2\ell} - F_{i1\ell} \) and \( 1 - F_{i2\ell} \), respectively -- all evaluated relative to \( \hat{\theta}_\ell \), and requiring weighting factors for each "pseudo-observation," \( (y_{i\ell}^T \cdot x_i) \), given by
\[(F_{i2\ell} - F_{1\ell})^{-1}. \text{ Since the requisite weighting matrix, } (F_{2\ell} - F_{1\ell})^{-1}, \text{ within}
[X'(F_{2\ell} - F_{1\ell})^{-1}X]^{-1}, \text{ requires updating in each iteration, a major computational advantage of the E-M algorithm, exhibited previously when } M = C, \text{ is not obtained when } M = T. \text{ Further, as one might expect, the variance, } \hat{\sigma}_{\ell}^2, \text{ is obtained as the weighted sum of squared "residuals," } (y_i^* - \hat{x}_{i-\ell+1})^2 -- \text{using the actual values, weighted by } [F_{i2\ell} - F_{1\ell}] = Pr[y_i^* \in Y_2], \text{ and the expected values,}
E_j(y_j^* - x_j^\hat{\theta}_{\ell+1})^2, \text{ weighted by } F_{1\ell} (j = 1) \text{ and } [1 - F_{i2\ell}] (j = 3), \text{ respectively -- all divided by the number of "pseudo-observations", } N_{\ell}. \text{ Again, } \hat{\sigma}_{\ell+1} > 0 \text{ for all } \ell = 0, 1, \ldots .

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:

\[
A^4_{\ell} = \begin{cases}
\begin{bmatrix}
X'X : 0 \\
\vdots \vdots \\
0 \vdots N
\end{bmatrix}^{-1}, & \text{if } M = C, \\
\begin{bmatrix}
X'(F_{2\ell} - F_{1\ell})^{-1}X : 0 \\
\vdots \vdots \\
0 \vdots N_{\ell}
\end{bmatrix}^{-1}, & \text{if } M = T,
\end{cases}
\]

and

\[
b^4_{\ell} = \begin{cases}
\begin{bmatrix}
X'(P_{\ell} \cdots N \cdots C \\
\cdots \cdots \cdots \cdots \cdots \\
\sum_{i=1}^{\Sigma C} q_{i\ell}
\end{bmatrix}, & \text{if } M = C, \\
\begin{bmatrix}
X'(F_{2\ell} - F_{1\ell})^{-1} P_{\ell}^T \\
\vdots \vdots \\
\sum_{i \in S_{\ell}^T} q_{i\ell}
\end{bmatrix}^T, & \text{if } M = T,
\end{cases}
\]
where

$$E^C_{i\&} = f_{i\&}^{\frac{K_i}{F_{i\&}}}$$

$$P^C_{i\&} = \begin{cases} c_2 \cdot \frac{f_{i\&}^{\frac{K_i}{F_{i\&}}}}{F_{i\&}}, & \text{if } i \in S_{1}^N, \\ (y_i - x_i^{1\&})^2, & \text{if } i \in S_{2}^N, \\ c_2 \cdot \frac{f_{i\&}^{\frac{K_i}{F_{i\&}}}}{1 - F_{i\&}}, & \text{if } i \in S_{3}^N, \end{cases}$$

$$E^T_{i\&} = [F_{i\&}^{\frac{K_i}{F_{i\&}}} - F_{i\&}^{\frac{K_i}{F_{i\&}}}](y_i^{*} - x_i^{\&}) + c_2 \cdot \left(\frac{F_{i\&}^{\frac{K_i}{F_{i\&}}} - F_{i\&}^{\frac{K_i}{F_{i\&}}}}{F_{i\&}^{\frac{K_i}{F_{i\&}}}}\right)$$

$$Q^C_{i\&} = \begin{cases} c_2 \cdot \frac{u_{i\&}^{\frac{K_i}{F_{i\&}}} f_{i\&}^{\frac{K_i}{F_{i\&}}}}{F_{i\&}^{\frac{K_i}{F_{i\&}}}} - 2[x_i^{1\&} (\hat{\beta}_{i+1} - \hat{\beta}_{i})] \cdot \xi_{i\&}^{(1)} + [x_i^{1\&} (\hat{\beta}_{i+1} - \hat{\beta}_{i})]^2, & \text{if } i \in S_{1}^N \\ (y_i^{*} - x_i^{1\&})^2 - c_2, & \text{if } i \in S_{2}^N \\ c_2 \cdot \frac{u_{i\&}^{\frac{K_i}{F_{i\&}}} f_{i\&}^{\frac{K_i}{F_{i\&}}}}{1 - F_{i\&}^{\frac{K_i}{F_{i\&}}}} - 2[x_i^{1\&} (\hat{\beta}_{i+1} - \hat{\beta}_{i})] \cdot \xi_{i\&}^{(1)} + [x_i^{1\&} (\hat{\beta}_{i+1} - \hat{\beta}_{i})]^2, & \text{if } i \in S_{3}^N \end{cases}$$

and

$$Q^T_{i\&} = s_{i\&}^{2\&} - \frac{\frac{2\&}{s_{i\&}^{2T}}}$$

in which $\hat{\beta}_{i+1}$, defined by (5.18a), is inserted into $q^C_{i\&}$ of (5.42) and $\hat{\beta}_{i+1}$, defined by (5.27a), is inserted into $s_{i\&}^{2\&}$ of (5.43) to eliminate the artificial dependence upon $\hat{\beta}_{i+1}$.
5.5. Computation of the Improved Initial Variance Estimator

In Section 3.3 above we have described an improved initial variance estimator \( \hat{\sigma}^2_o \), which was derived by maximizing the likelihood function \( L_N(\theta) \), conditional upon the consistent initial estimates of the regression coefficients, \( \hat{\beta}_o^{(r,s)} \). 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, \( A^m_\lambda \), and the \((K+1)\)-vector, \( b^m_\lambda \), as:

\[
A^m_\lambda = \begin{bmatrix}
A^m_{11} & \cdots & A^m_{12} \\
\vdots & \ddots & \vdots \\
A^m_{12} & \cdots & A^m_{22}
\end{bmatrix}
\quad \text{and} \quad
b^m_\lambda = \begin{bmatrix}
b^m_1 \\
\vdots \\
b^m_2
\end{bmatrix},
\]

where the scalars, \( a^m_{22} \) and \( b^m_{22} \), refer to the parameter, \( \hat{\sigma}^2_\lambda \).

Then the \((\lambda+1)\)-iterate of \( \hat{\sigma}^2_o \) is given by:

\[
\hat{\sigma}^2_{\lambda+1} = \hat{\sigma}^2_o + \lambda^{m}_\lambda \cdot a_{22}^m \cdot b_{22}^m,
\]

and the conditional ML initial variance estimator is defined by:

\[
\hat{\sigma}^2_o = \lim_{\lambda \to \infty} \hat{\sigma}^2_\lambda,
\]

where \( \lambda^m_\lambda \) 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,
provides excellent initial estimates of \( \hat{\theta}_0 \), and that the "cost of adjustment" in moving from \( \hat{\theta}_0 \) (of (3.29b)) to \( \hat{\tilde{\theta}}_0 \) (of (5.47)) is small.

5.6. Modified Algorithms:

Our previous discussion has been restricted to the unmodified versions of the N-R, M-S, G-N and E-M algorithms, in which \( \lambda^m_\ell = 1 \) for \( \ell = 0,1, \ldots \). In the present subsection, we consider modified versions of these algorithms, in which \( \lambda^m_\ell \) 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 \( \log L^M_N \). We now examine procedures for modification of \( \lambda^m_\ell \) 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 \( \lambda^m_\ell \), which applies to all "gradient-type" methods (i.e., methods, such as the N-R, M-S and G-N, in which \( b^m_\ell = \frac{\partial \log L^M_N(\hat{\theta}_\ell)}{\partial \theta} \)), provided that the following conditions are satisfied:
(i) $I_N^M(\hat{\theta})$ is twice-differentiable, and defined over a compact, upper-contour parameter space, $\Theta$,

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

\begin{equation}
\frac{b^{m'}_\lambda A^m_\lambda b^{m'}_\lambda}{b^m_\lambda b^m_\lambda} > \alpha > 0,
\end{equation}

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

The BHHH modification is to define $\lambda_{\lambda,\lambda}^m$, at each iteration, by the rule:

\begin{equation}
\lambda_{\lambda,\lambda}^m = \begin{cases} 
\lambda_{\lambda,\lambda}^m, & \text{if } Q_{\lambda,\lambda}^m(1) < \delta \\
1, & \text{if } Q_{\lambda,\lambda}^m(1) \geq \delta 
\end{cases}
\end{equation}

where $\delta$ is a pre-assigned constant in the interval, $(0,1/2)$, $Q_{\lambda,\lambda}^m(\lambda)$ is the criterion function,

\begin{equation}
Q_{\lambda,\lambda}^m(\lambda) = \frac{\log L_N^M(\hat{\theta}^m_\lambda + \lambda \cdot A^m_\lambda b^{m'}_\lambda) - \log L_N^M(\hat{\theta}^m_\lambda)}{\lambda \cdot b^{m'}_\lambda A^m_\lambda b^{m'}_\lambda} ,
\end{equation}

and $\lambda_{\lambda,\lambda}^m$ is a $\lambda$-value satisfying the inequality,

\begin{equation}
\delta \leq Q_{\lambda,\lambda}^m(\lambda_{\lambda,\lambda}^m) \leq 1 - \delta .
\end{equation}

They show that, under condition (i), a $\lambda_{\lambda,\lambda}^m$ satisfying (5.51) will always exist, and that, under conditions (i) and (ii), the sequence $\hat{\theta}^m_\lambda$ of (5.3) converges to a stationary point of $\log L_N^M(\hat{\theta})$ at which $\lim_{\lambda \to m} b^m_\lambda = 0$. Finally, BHHH 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, $A^1_\lambda$, or its expectation, $A^2_\lambda$, 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 BHHH results establish the existence of such a $\lambda^{*}_\lambda$ 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 consistent initial estimators,
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, $\lambda^{*\ast m}_\lambda$, which maximizes the numerator of the BHHH criterion function
over the unit-interval, i.e., instead of $Q^m_\lambda(\lambda)$, consider maximization of:

\[
(5.52) \quad S^m_\lambda(\lambda) = \log L^m_N(\hat{m}_\lambda + \lambda \cdot A^m_\lambda b^m_\lambda)
\]

over $0 \leq \lambda \leq 1$. 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 $k$, therefore, consists
of the following steps:

 Initialization Step $(k,0)$: Set $j=1$ and the support points
\[ \lambda_{j1} = 0, \lambda_{j2} = \frac{1}{2} \text{ and } \lambda_{j3} = 1. \]

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

\[ a_{j} = [a_{\lambda j1}, a_{\lambda j2}, a_{\lambda j3}]', \]

such that the parabolic approximation,

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

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

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

and

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

the solution is defined by

\[ a_j = D(\lambda_{j})^{-1} \cdot S^m_\ell. \]

**Step (\(\ell, j, 2\)):** Evaluate \(S^{\star m}_\ell(\lambda)\) at the value of \(\lambda\) maximizing the parabolic approximation, \(S^{\star m}_\ell(\lambda)\), i.e., at the value,
(5.58) \( \lambda_{kj}^* = \frac{-a_k j^2}{2 a_k j^3} \),

provided that \( a_k j^3 < 0 \) for a maximum.

**Step (5.j.3):** If \( S^m_k(\lambda^*_{kj}) > S^m_k(\lambda_{k1}) \), set \( \lambda^*_{kj} = \lambda_{kj}^* \). Otherwise, set \( \lambda_{kj} \) equal to \( \lambda_{kj} / 2 \), \( k = 1, 2, 3 \); set \( j \) to \( j+1 \); and repeat steps (5.j.1) - (5.j.3).

Hartley (1961) proves the convergence of the modified G-N algorithm, embodying (5.3) and Steps (5.j.1) - (5.j.3). Our procedure is to replace Step (5.j.3) above by the alternative:

**Step (5.j.3)':** Let \( j' = j+1 \) and define \( \lambda^*_{kj'}^k \), \( k = 1, 2, 3 \), by the following rule:

**Conditions:**

(a) \( \lambda^*_{kj} \leq \lambda_{kj1} \)

\[
\begin{align*}
\lambda^*_{kj1}' &= \lambda^*_{kj} \\
\lambda^*_{kj2}' &= \lambda^*_{kj1} \\
\lambda^*_{kj3}' &= \lambda^*_{kj2}
\end{align*}
\]

(b) \( \lambda_{kj1} < \lambda^*_{kj} \leq \lambda_{kj2} \)

\[
\begin{align*}
\lambda^*_{kj1}' &= \lambda^*_{kj} \\
\lambda^*_{kj2}' &= \lambda^*_{kj} \\
\lambda^*_{kj3}' &= \lambda^*_{kj2}
\end{align*}
\]

(c) \( \lambda_{kj2} < \lambda^*_{kj} \leq \lambda_{kj3} \)

\[
\begin{align*}
\lambda^*_{kj1}' &= \lambda^*_{kj2} \\
\lambda^*_{kj2}' &= \lambda^*_{kj} \\
\lambda^*_{kj3}' &= \lambda^*_{kj3}
\end{align*}
\]

(d) \( \lambda_{kj3} < \lambda^*_{kj} \)

\[
\begin{align*}
\lambda^*_{kj1}' &= \lambda^*_{kj2} \\
\lambda^*_{kj2}' &= \lambda^*_{kj3} \\
\lambda^*_{kj3}' &= \lambda^*_{kj}
\end{align*}
\]
Then repeat steps \((\ell, j, 1) - (\ell, j, 3)\) until either convergence obtains or a value of \(S^m(\lambda_{\ell j}^* ) \leq S^m(\lambda_{\ell j-1}^* )\) occurs. If the latter obtains when \(j=1\), repeat with \(\lambda_{\ell j} = \lambda_{\ell j}/2\), \(k=1, 2\) and 3.

VI. An Illustrative Example

In this section we provide a brief illustration of the four ML algorithms and the consistent initial estimator, using a small, bilaterally-censored regression model. Data were generated by Monte Carlo simulation from the model:

\[
(6.1a) \quad y_i^* = 40.0 + 0.3 \cdot x_{i1} - 0.2 \cdot x_{i2} - 1.4 \cdot x_{i3} + 6.7 \cdot x_{i4} + \varepsilon_i
\]

where

\[
(6.1b) \quad \varepsilon_i \sim \text{n.i.d.} (0, 64)
\]

and

\[
(6.1c) \quad X_i \equiv \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ x_{i4} \end{bmatrix} \sim \text{NID} \left( \begin{bmatrix} 6.00 & 1.00 & 0.60 & 2.00 \\ 0 & 1.00 & 12.00 & -3.00 & 0.60 \\ 0 & 0.60 & -3.00 & 36.00 & -1.20 \\ 0 & 2.00 & 0.60 & -1.20 & 42.00 \end{bmatrix} \right)
\]

Censoring of \(y_i^*\) occurs at the fixed limits of 0 and 100. We generated 100 samples of 500 observations each. In the resulting 50000 drawings on \(y_i^*\), there were 9139 cases at the lower bound and 4843 cases at the upper bound. In the typology of section 3, we have case #11 -- fixed bilateral censoring.
For each sample, consistent initial estimators were obtained using the recursive relation (3.28) with \( r=3 \) and \( s=1 \). Because past experience had indicated that the Newton-type algorithms were computationally sensitive to the choice of the initial variance estimator, the conditional ML initial variance estimator of equation (3.31) was used in every case. In only one sample out of the 100 did we encounter a negative variance estimate. This was replaced by the adjusted variance estimator of (3.30a). Starting from the initial, adjusted variance estimator, the conditional ML algorithm was allowed to iterate 10 times (or until convergence at the absolute level of \( 10^{-3} \) obtained). Results for the consistent initial estimator are shown in Table 6.1 below.

<table>
<thead>
<tr>
<th></th>
<th>Coef1</th>
<th>Coef2</th>
<th>Coef3</th>
<th>Coef4</th>
<th>Constant</th>
<th>Variance**</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Coefficients</td>
<td>-0.300</td>
<td>-0.420</td>
<td>1.400</td>
<td>6.700</td>
<td>40.000</td>
<td>64.000</td>
</tr>
<tr>
<td>Mean of Estimates</td>
<td>0.369</td>
<td>-0.234</td>
<td>-1.388</td>
<td>6.664</td>
<td>40.152</td>
<td>63.812</td>
</tr>
<tr>
<td>Standard Error of</td>
<td>0.233</td>
<td>0.140</td>
<td>0.106</td>
<td>0.350</td>
<td>0.707</td>
<td>7.327</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Seconds of Central Processing Unit time on a Burroughs B79000.

** All variance estimators were "improved" using the E-M algorithm with initial beta estimates. The initial variance estimate from one sample was negative and was replaced by an unrestricted OLS estimator, using the residuals from the consistent beta estimates and the observations between limit points.

Maximum likelihood estimates of the model parameters were calculated for each algorithm. Each run began from the consistent initial estimator and was allowed to iterate until convergence at an absolute level of \( 10^{-3} \). Computer time and the number of iterations for each algorithm are shown in Table 6.2.
Since each algorithm converged to the same estimates for each sample, only the average of 100 samples and the standard error of such estimates are shown. It is interesting to note that, while the ML estimator of the variance term is actually farther from the true value than was the (improved) consistent initial estimator, the variance of the ML estimators are, as expected, uniformly smaller.

The timings indicate that the algorithms which employ an exact form of the (weighted) moment matrix converge in fewer iterations than do the Gauss-Newton and E-M algorithms. Despite their greater computational burden per-iteration, they are also faster. In our experience, the Gauss-Newton is always dominated by the Newton-Raphson and Method-of-Scoring algorithms -- even when supplemented by routines to optimize the chosen step-size. Its practical utility is apparently restricted to problems for which exact second-partial derivatives are unavailable.

These results are typical of those that might be obtained in a practical application of the techniques. However, they do not demonstrate the comparative advantage of the E-M algorithm in problems with large numbers of regressors. In such cases the E-M achieves computational efficiency by avoiding the repeated inversion of the moment matrix, (X'X). For an empirical illustration of this point see our companion paper (Hartley and Swanson (1985)).
Table 6.2: Summary of Monte-Carlo Results for the Maximum Likelihood Estimator

(Based on 100 Samples of 500 Observations)

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time (sec)</th>
<th>Number of Iterations</th>
<th>Coef1</th>
<th>Coef2</th>
<th>Coef3</th>
<th>Coef4</th>
<th>Constant</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton-Raphson</td>
<td>3.8</td>
<td>4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.6)</td>
<td>(0.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gauss-Newton</td>
<td>8.2</td>
<td>9.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.8)</td>
<td>(2.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method of Scoring</td>
<td>5.5</td>
<td>4.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.8)</td>
<td>(0.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expectation-</td>
<td>25.3</td>
<td>24.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximization</td>
<td>(3.4)</td>
<td>(3.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Standard errors in brackets.
VII. 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 Amemiya'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 in Hartley and Swanson (1985) 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 of the E-M method will decrease. In truncated data sets, the E-M algorithm is 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 large 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. 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 the Probit) normal regression model and discussed explicitly in this paper, has been written, and will be made available at cost upon written request.
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.


Hartley, M.J. and E.V. Swanson (1984), Achievement and Wastage: An Analysis of the Retention of Basic Skills in Egyptian Primary Education, draft manuscript, Development Research Department, World Bank (June), pp. 1-546.


