THE IMPACT OF GROUPING COARSENESS IN ALTERNATIVE GROUPED-DATA REGRESSION MODELS

by

T. A. Cameron*

Department of Economics

University of California, Los Angeles

UCLA Dept. of Economics
Working Paper #381
February, 1985
Revised August, 1985

*The author would like to thank Edward E. Leamer, Jay Stewart and Kenneth J. White for helpful suggestions and comments.

THE IMPACT OF GROUPING COARSENESS IN ALTERNATIVE GROUPED-DATA REGRESSION MODELS

by

T. A. Cameron*

Department of Economics

University of California, Los Angeles

February, 1985

Revised August, 1985

Abstract

In regression applications, it is not uncommon to find that the data for the desired dependent variable have been grouped or rounded. Other investigators have explored maximum-likelihood-based methods for fitting these models. This paper is a lateral extension, using Monte Carlo experiments and an empirical illustration to demonstrate (i) that the coarseness of the grouping will affect the need for special estimation methods, and (ii) that departures from the conventional linear Normal regression model will increase the distortion in parameter estimates due to the use of interval midpoints as an approximation. Researchers using the familiar log-linear Normal specification with grouped data should be advised to proceed with caution.

*The author would like to thank Edward E. Leamer, Jay Stewart and Kenneth J. White for helpful suggestions and comments.

I. Introduction

In regression applications, it is not uncommon to find that the data available for the dependent variable are less than ideal. The data may be grouped (perhaps for confidentiality), rounded upward or downward, or rounded conventionally to the nearest whole number. Furthermore, the rounding interval is sometimes large relative to the range of the data. The most typical recourse in these situations is to adopt the midpoint of the relevant interval as a proxy for the mean of the variable over that interval. However, this is at best only an approximation -- sometimes satisfactory, but at other times poor enough that it would be preferable to recognize the rounding explicitly in the regression process. This paper is a lateral extension of previous work in the area. Two issues are addressed: (a) the consequences of grouping or rounding in non-normal regression models, especially models which assume highly skewed distributions, and (b) the behavior of parameter estimates as the rounding or grouping of the dependent variable becomes increasingly coarse.

The layout of this paper is as follows. Section II examines the related literature and identifies the innovations of the present work. Section III specifies a general linear grouped-data regression model and then specializes this model to particular distributions. Section IV does the same for the log-linear model. Section V describes the implementation of the optimization algorithms. A set of simulation experiments is described in Section VI with particular emphasis on the more-common linear normal and log-linear normal regression models. Section VII demonstrates the impact of the correction in a log-linear Normal model of long-distance telephone call durations, and Section VIII summarizes.

II. Proximity to the Literature

Before proceeding, it is important to note that this is the usual textbook example of errors in variables. The grouping intervals are completely exogenously determined. The explanatory variables are assumed to be uncorrelated with the error term. Furthermore, in the conventional case with random measurement error in the dependent variable only, we would expect an unbiased estimator of the slope (for example). This is not the case here. I am addressing a fundamentally different (but pervasive) data problem.

The type of correction strategy explored in this paper was proposed first by Hasselblad, Stead and Galke (1980).[1] In their example, the dependent variable is the quantity of lead found in blood samples; explanatory variables are year, race and age group. The maintained hypothesis is that the original blood lead levels come from a lognormal distribution, so that the logarithm of this variable is normally distributed around a mean value which depends on time and several dummy variables. They find "...some justification for our uneasiness with the simple use of midpoints...." The present paper emphasizes that the consequences of using midpoints depend on the particular application. Depending on the specification, on the underlying error distribution, as well as on the coarseness of the grouping, the midpoint technique can yield a better or worse approximation to the true underlying parameter values.

A portion of the procedure to be described here is related to work by Stewart (1983), who chooses to focus on cost-saving OLS-based approximations to the preferred maximum likelihood estimation techniques. He undertakes a set of simulations, but uses only ten grouping intervals, devoting primary attention to the influence of different assumptions about the regressors. He considers true error distributions which violate the underlying assumption of normality, but does not work with fundamentally non-normal regression models.

Recognizing that computing technology is making algorithm complexity less of an issue in the choice of estimation technique, I opt for the theoretically-preferred maximum likelihood approach.

Burridge (1981), has also worked with grouped-data regression estimation. He examines maximum likelihood estimation where the underlying unobserved dependent variable has either a normal distribution or an extreme value distribution. Burridge's emphasis, however, lies with the use of special reparameterizations of the likelihood function which render it concave, thereby ensuring the existence of a global maximum. Simulations focus upon the relative convergence speeds of different algorithms.

Going beyond Burridge's extension to the extreme-value distribution I choose to explore in greater detail a common empirical consideration for economic models: many dependent variables may take on only non-negative values. To assume that the variable is distributed lognormally (so that its logarithm follows a normal distribution) is but one of many possible assumptions. The Generalized Gamma family of distributions has been explored by Cameron and White (1985) for both linear and log-linear regression models. A subset of these models is adapted here for grouped dependent variables. Aside from the Normal (N) distribution, I have limited the selection to distributions with closed forms for their cumulative density functions: the Exponential (E) and the Weibull (W).

III. Linear Regression Models

In this section, and in Section IV, the details of the optimization algorithms are provided. Subsection A in each case gives the general form of the algorithm, relevant to whatever specific underlying distribution might be assumed. While these formulas are not the focal point of the paper, they are provided to emphasize the adaptability of grouped-data strategies to a whole

range of distributional assumptions.

A. The General Case

For these linear models, the range of the true dependent variable, $\tilde{t}$, is $-\infty < \tilde{t} < +\infty$ for the N model and $0 < \tilde{t} < +\infty$ for the E and W models. For the following description we will assume that $\tilde{t}$ is rounded within a known interval. Also assume that the observed datum, t, takes on the value of the midpoint of the interval of $\tilde{t}$ which it represents (i.e., $t = t_m$ if $t_\ell < \tilde{t} < t_u$). In the general case, we have:

$$\Pr(t = t_m) = F(t_u) - F(t_\ell), \tag{1}$$

where, $F(.)$ is the c.d.f. for $\tilde{t}$. We assume that the distribution of $\tilde{t}$ is conditional upon a vector of p explanatory variables, x. For a random sample of n observations, the log-likelihood function will be:

$$\ell = \sum_{i=1}^{n} \log[F(t_{ui}) - F(t_{\ell i})] \tag{2}$$

In a linear regression model, the conditional distribution function will have a mean of $x'\beta$ and (usually) at least one separate shape parameter, c. Optimization of the likelihood function in (2) requires the derivatives of this function with respect to the unknown parameters (β,c). With the following simplifying notation:

$$F_i = F(t_{ui}) - F(t_{\ell i})$$

$$F_{(\theta)i} = \partial F_i / \partial\theta \quad \theta = \beta_1,\ldots,\beta_p,c$$

$$F_{(\theta\gamma)i} = \partial^2 F/\partial\theta\partial\gamma \quad \theta,\gamma = \beta_1,\ldots,\beta_p,c,$$

these derivatives can be stated generally as:

$$\frac{\partial\ell}{\partial\theta} = \sum_{i=1}^{n} F_{(\theta)i}/F_i \tag{3a}$$

$$\frac{\partial^2 \ell}{\partial\theta\partial\gamma} = \sum_{i=1}^{n} \frac{F_i F_{(\theta\gamma)i} - F_{(\theta)i} F_{(\gamma)i}}{(F_i)^2} \quad \theta,\gamma = \beta_1,\ldots,\beta_p,c. \tag{3b}$$

B. Specialization to Specific Distributions

a. Normal

For Normal errors, we have the parameter vector $(\beta_1,\ldots,\beta_p,\sigma)$.[2] The components of the likelihood function and its gradient will be as follows. (The components for the Hessian have been relegated to Appendix 1.) To simplify, let $z_{ui} = (t_{ui} - x_i'\beta)/\sigma$ and $z_{\ell i} = (t_{\ell i} - x_i'\beta)/\sigma$. Then:

$$F_i = \Phi(z_{ui}) - \Phi(z_{\ell i}) \tag{4}$$

$$F_{(\beta_r)i} = \frac{-x_{ir}}{\sigma}\,[\phi(z_{ui}) - \phi(z_{\ell i})] \quad r = 1,\ldots,p$$

$$F_{(\sigma)i} = -\frac{1}{\sigma}\,[z_{ui}\,\phi(z_{ui}) - z_{\ell i}\,\phi(z_{\ell i})]$$

b. Exponential

For an Exponentially distributed dependent variable, we need only estimate the parameters $(\beta_1,\ldots,\beta_p)$. There is no separate shape parameter. The components required for the likelihood function and its gradient are as follows. (Let $v_i^* = v_i/x_i'\beta$, for any variable v.)

$$F_i = \exp(-t^*_{\ell i}) - \exp(-t^*_{ui}) \tag{5}$$

$$F_{(\beta_r)i} = x^*_{ir}[t^*_{\ell i}\exp(-t^*_{\ell i}) - t^*_{ui}\exp(-t^*_{ui})] \quad r = 1,\ldots,p$$

c. Weibull

For a Weibull-distributed dependent variable, we must estimate both $(\beta_1,\ldots,\beta_p)$ and the shape parameter, c. Recall that the simple Weibull probability density function is given by:

$$f(\tilde{t}) = \frac{c\tilde{t}^{c-1}}{b^c}\exp[-(\tilde{t}/b)^c], \quad 0 < \tilde{t} < +\infty.$$

As described in Cameron and White (1985), the Weibull linear regression model begins with a simple Weibull density and substitutes for the scale parameter, b, the expression $x'\beta/\Gamma(\frac{c+1}{c})$. To simplify, set $c^* = (c+1)/c$ and again let $v_i^* = v_i/x_i'\beta$ for any variable v. Also, let $T_{\ell i} = [t_{\ell i}^* \Gamma(c^*)]$, $T_{ui} = [t_{ui}^* \Gamma(c^*)]$, and let $\Gamma'(k) = \partial\Gamma(k)/\partial k$. Then:

$$F_i = \exp(-T_{\ell i}^c) - \exp(-T_{ui}^c) \tag{6}$$

$$F_{(\beta_r)i} = cx_{ir}^* [T_{\ell i}^c \exp(-T_{\ell i}^c) - T_{ui}^c \exp(-T_{ui}^c)] \quad r = 1,\ldots,p$$

$$F_{(c)i} = [t_{\ell i}^* \frac{\Gamma'(c^*)}{c} T_{\ell i}^{(c-1)} - T_{\ell i}^c \log T_{\ell i}] \exp(-T_{\ell i}^c) -$$

$$[t_{ui}^* \frac{\Gamma'(c^*)}{c} T_{ui}^{(c-1)} - T_{ui}^c \log T_{ui}] \exp(-T_{ui}^c).$$

IV. Log-Linear Regression Models

A. The General Case

Lawless (1982) describes the derivation of log-linear regression models under several non-normal error distributions. We assume that the actual dependent variable, $\tilde{t}$, is distributed around its expected value, $\exp(x'\beta)$, according to a density function $g(\tilde{t}|x)$. In log-linear models, we define the dependent variable as $\tilde{y} = \log(\tilde{t})$, which has a density function given by $f(\tilde{y}|x) = \exp(\tilde{y}) \cdot g(\exp(\tilde{y})|x)$.[3] When $\tilde{t}$ is rounded to the midpoint of an interval with bounds t_ℓ and t_u, the observed values of t and y are thus:

$$t = t_m, \; y = \log(t_m) \quad \text{if} \quad t_\ell < \tilde{t} < t_u \tag{7}$$

A further transformation, $z = (y - x'\beta)/\sigma$ yields the error term for the log-linear regression model. Now we may argue that:

$$\Pr(t=t_m) = \Pr[(-x'\beta + \log(t_\ell))/\sigma < z < (-x'\beta + \log(t_u))/\sigma] \tag{8}$$

To simplify, let $z_\ell = (-x'\beta + \log(t_\ell))/\sigma$ be the lower limit of the interval

wherein the error term z may lie and let $z_u = (-x'\beta + \log(t_u))/\sigma$ be the upper limit of this interval. Then:

$$\Pr(t=t_m) = F(z_u) - F(z_\ell) \tag{9}$$

As before, the underlying parameters β and σ are determined by the method of maximum likelihood. To simplify the following exposition, we will use the abbreviations given below (where $f(.)$ is an arbitrary p.d.f., and $F(.)$ is the corresponding c.d.f.):

$$P_i = F(z_{ui}) - F(z_{\ell i})$$

$$p_i = f(z_{ui}) - f(z_{\ell i})$$

$$Q_i = z_{ui}f(z_{ui}) - z_{\ell i}f(z_{\ell i})$$

$$q_i = z_{ui}f'(z_{ui}) - z_{\ell i}f'(z_{\ell i})$$

$$R_i = z_{ui}^2 f'(z_{ui}) - z_{\ell i}^2 f'(z_{\ell i})$$

$$\rho_i = p_i/P_i$$

$$\omega_i = Q_i/P_i$$

The log likelihood function for n independent observations on t and x is:

$$\ell = \sum_{i=1}^{n} \log P_i \tag{10}$$

The elements of the gradient vector are given by:

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{i=1}^{n} -\frac{x_{ir}}{\sigma}\rho_i \qquad r = 1,\ldots,p \tag{11a}$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_{i=1}^{n} -\frac{\omega_i}{\sigma} \tag{11b}$$

and the elements of the Hessian are:

$$\frac{\partial^2 \ell}{\partial \beta_r \partial \beta_s} = \sum_{i=1}^{n} \frac{x_{ir}x_{is}}{\sigma^2} [\rho_i(1-\rho_i)] \quad r,s = 1,\ldots,p \tag{11c}$$

$$\frac{\partial^2 \ell}{\partial \beta_r \partial \sigma} = \sum_{i=1}^{n} \frac{x_{ir}}{\sigma^2} [\rho_i + q_i/P_i - \rho_i \omega_i] \quad r = 1,\ldots,p \tag{11d}$$

$$\frac{\partial^2 \ell}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{1}{\sigma^2} [2\omega_i + R_i/P_i - \omega_i^2] \tag{11e}$$

B. Specialization to Specific Distributions

a. Lognormal

If the conditional distribution of $\tilde{t}$ is lognormal, then z has a standard normal distribution.[4] To obtain estimates for both β and σ, the required components of equations (10) and (11) for the normal distribution (for $z = z_u, z_\ell$) are:

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(\frac{-1}{2} t^2)dt \tag{12}$$

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-1}{2} z^2)$$

$$f'(z) = -z\phi(z) = \frac{-z}{\sqrt{2\pi}} \exp(\frac{-1}{2} z^2)$$

b. Exponential

If the conditional distribution of $\tilde{t}$ is exponential, z follows a standard extreme value distribution (see Hastings and Peacock, 1974) and σ is constrained to unity, so we need estimate only β. We will use only the derivatives (11a) and (11c). The necessary functions are (for $z = z_u, z_\ell$):

$$F(z) = 1 - \exp(-\exp z) \tag{13}$$

$$f(z) = \exp(z - \exp z)$$

$$f'(z) = [1 - \exp z] \exp(z - \exp z).$$

c. Weibull

If the conditional distribution of $\tilde{t}$ is Weibull, then z follows a the same standard extreme value distribution but σ is no longer constrained to unity and must now be estimated, so we will use the full set of derivatives, (11a) through (11e), but the same set of components, (13).

V. Estimation Techniques

Maximum likelihood techniques are utilized to determine which values of the parameters in question maximize the joint probability of occurrence of the data in the sample. While a variety of algorithms are available for generating maximum likelihood estimates, I have found it expedient to utilize the GQOPT computer program with its Fortran-based subroutines. While this package offers the programmer the option of numerically computed derivatives, this choice can make the program quite slow and therefore very expensive. Where the researcher is confronted with a binding constraint on computing funds, the use of analytical first and second derivatives can reduce costs. Hasselblad, Stead, and Galke (1980) utilize the so-called EM algorithm developed by Dempster, Laird, and Rubin (1977) for their blood lead study. A variant of this technique is also used by Stewart (1983), and explored by Burridge (1982). This algorithm is also a cost-saving measure, but it has the disadvantage of failing to provide an asymptotic variance-covariance matrix for the estimated parameters. Without this information, hypothesis testing regarding these parameters is not feasible unless the point estimates for the parameters are combined directly with the data to generate either the estimated Hessian matrix, or alternatively the estimated gradient-outer-product matrix (Berndt, Hall, Hall, and Hausman, 1974). (Either of these will approximate the Information Matrix under appropriate assumptions.) In either case, the formulas for the analytic derivatives are required, which justifies their full

presentation in this paper.

VI. Simulation Experiments

The matrix of possible simulations which could be undertaken is very large. Obviously, it is only practical to present a select subset of these. In particular, the grouping processes considered here involve (i) only equal intervals of the dependent variable, and (ii) only simple regressions (although the algorithms have been designed for multiple regression). The same set of values was used for the (deterministic) explanatory variable in all of the following models. Two hundred values of x were drawn from a one-time sampling from a uniform $(0,1)$ distribution. The "true" parameters of the regression relationship vary across models, but for each of the one hundred Monte Carlo samples used in each model, the values of the dependent variable are constructed as random drawings from the appropriate distribution (having a conditional mean of $x'\beta$).

A comprehensive set of tables detailing the numerical Monte Carlo simulation results is available from the author. Here, for clarity, approximate graphical summaries will be employed. For some perspective on the severity of the rounding process, scatter diagrams of the untransformed dependent variable plotted against x for the first 100 observations are occasionally provided.

Linear Normal Regression Models

We begin by examining the relative performance of ordinary normal linear regression using interval midpoints, versus the grouped dependent variable model. The "true" model is $\tilde{t} = 1.0 + 2.0x$, and the standard deviation of the error distribution varies between .10 and 2.5.

When the true standard deviation is relatively small (on the order of 0.5 or less) the mean slope and intercept estimates are extremely close, provided

the rounding interval is kept to less than (roughly) three times this standard deviation. In this case, the only apparent shortcoming of the OLS midpoint method is that the standard deviation is systematically overestimated. This tendency can be seen in Table 1. As the rounding intervals widen, the midpoint method overestimates the true standard deviation by an increasingly greater margin. This implies that use of the OLS midpoint method could result in failure to reject null hypotheses when there actually is sufficient resolution in the data to reject the hypothesis in question.

For larger values of the true standard deviation (on the order of 0.75 and larger), both methods, when applied to grouped data, yield biased estimates of the true intercept and slope. In particular, the slope becomes biased downwards and the intercept biased upwards; the degree of bias also increases approximately linearly with the width of the rounding interval. This is not surprising. In the limit (as the intervals become more and more coarse) the entire sample range of the dependent variable falls into one interval, the fitted slope goes to zero, the intercept becomes the mean of that interval, and there is no remaining information about the variance. (It is just this obscurity about the variance that limits the sizes of the rounding intervals actually explored.)

The slope and intercept parameters (in the models with larger underlying standard deviations), the OLS midpoint method yields virtually identical point estimates to those obtained (at the expense of considerably more effort) from the theoretically more-valid grouped data maximum likelihood method. There is a slight divergence of the point estimates at extreme (unrealistic) degrees of coarseness, but it seems justifiable to consider the two sets of estimates to be identical for practical purposes. It is worth noting that both sets of estimates for the standard deviation begin to underestimate, rather than

TABLE 1

Mean Point Estimates (and Mean Asymptotic Standard Errors)

For the Error Standard Deviation Parameter, σ

(Linear Normal Regression, 100 samples, $n = 200$, $t = 1.0 + 2.0x + \sigma\varepsilon$)

True error st. dev., σ	Rounding Interval Width	OLS Midpoint Method	MLE Grouped Data Method
0.10	0.10	0.1042 (0.0052)	0.1001 (0.0054)
	0.20	0.1153 (0.0057)	0.0997 (0.0065)
	0.30	0.1291 (0.0064)	0.0995 (0.0076)
0.25	0.25	0.2610 (0.0131)	0.2508 (0.0136)
	0.50	0.2879 (0.0144)	0.2486 (0.0163)
	0.75	0.3323 (0.0166)	0.2512 (0.0202)
0.50	0.50	0.5190 (0.0260)	0.4984 (0.0270)
	1.00	0.5745 (0.0287)	0.4956 (0.0328)
	1.50	0.6692 (0.0335)	0.4885 (0.0418)

overestimate, the true parameter value as the true standard deviation increases beyond 0.75.[7] In any event, the similarity of the estimates at smaller rounding interval widths should be good news for those who do not wish to incur the fixed costs of setting up the more elaborate technique.

Regarding specification testing, it should be noted that the midpoint and grouped data techniques are not nested models, and do not even model the same "data", so there are no straightforward statistical criteria for evaluating the importance of differences in the maximized values of the log-likelihood values. Bear in mind that the log-likelihood for the midpoint model reflects the joint density, not for the actual data, but for an artificial construct: the midpoint of each interval. It is not surprising that the maximized log-likelihood value for the grouped data model should get progressively larger as the grouping intervals widen, since the task of the model becomes easier when it is only necessary to predict "residence" of the dependent variable within a wider interval.

Log-linear Lognormal Regression

Here we begin to see the potential advantages of the grouped data techniques. When equal intervals in the raw dependent variable get transformed into unequal intervals in the logarithms of the dependent variable, the distinction between the alternative estimation methods becomes quite marked. Figure 1 illustrates the differences in a regression model where the true relationship (selected arbitrarily) is $y = \log t = 0.0 + 1.0\ x$. The figure consists of three horizontal panels, each describing the Monte Carlos results for one underlying model. In each panel, the three graphs describe the results for each of the three parameters: intercept (β_o), slope (β_1), and standard deviation (σ). The horizontal axis in all cases describes the width of the rounding interval for the dependent variable (anchored at zero).[8] (An

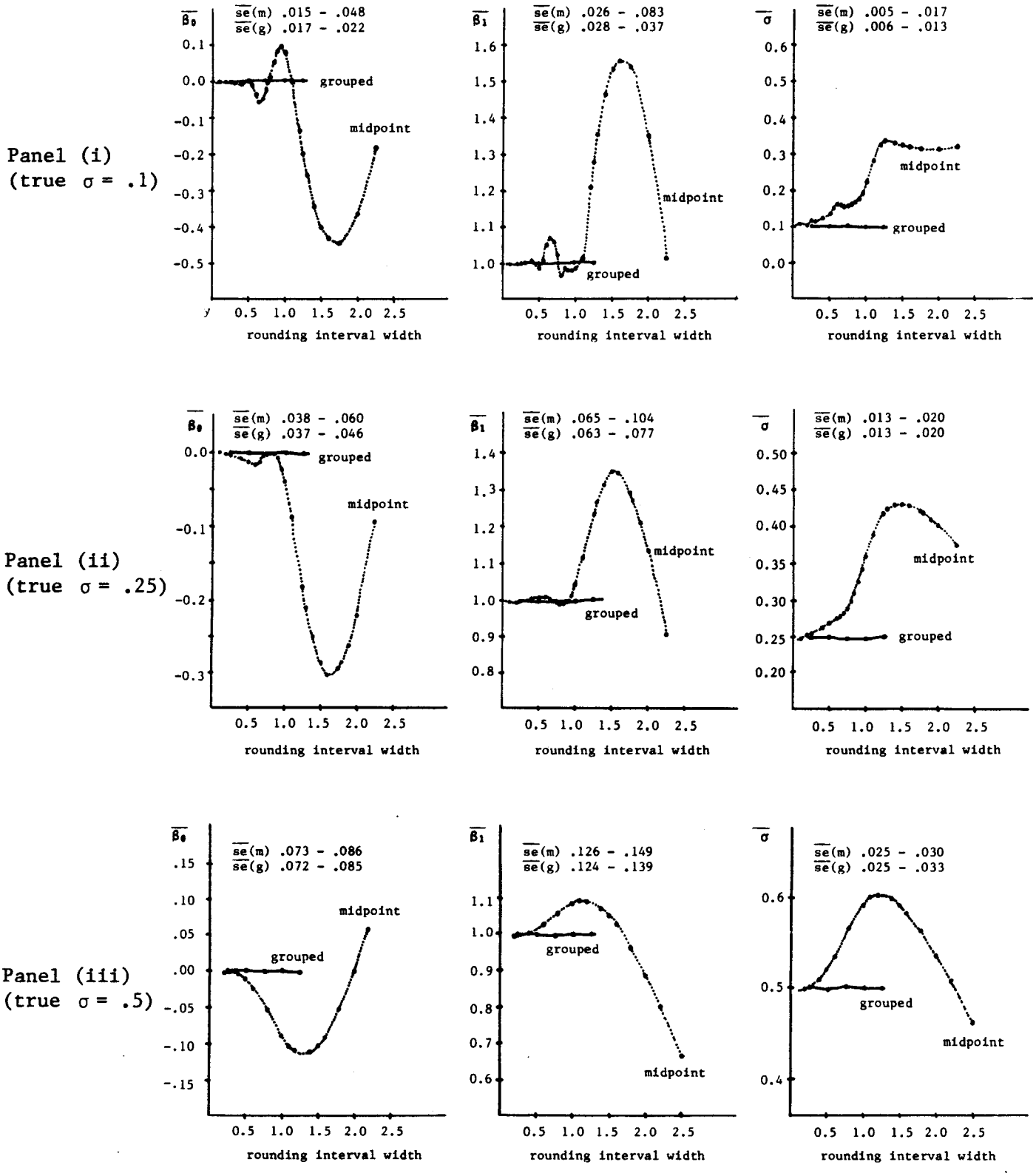

Figure 1 - Mean point estimates, log-linear Normal models, 100 Monte Carlo samples, n = 200; grouped data method versus midpoint technique; identical intercepts and slopes, varying standard deviations. Mean asymptotic standard errors = "$\overline{se}$(m)", "$\overline{se}$(g)".

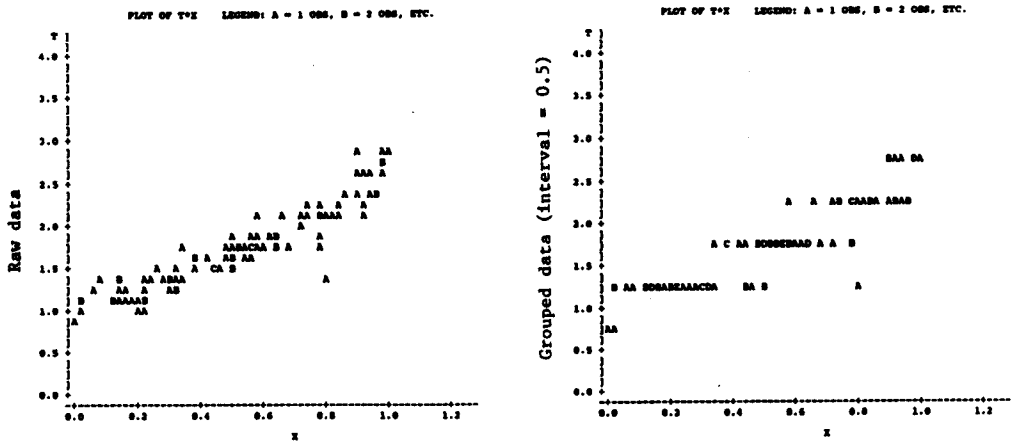

Figure 1a - First 100 observations for log-linear Normal model
(intercept = 0.0, slope = 1.0, std. dev. = 0.10)

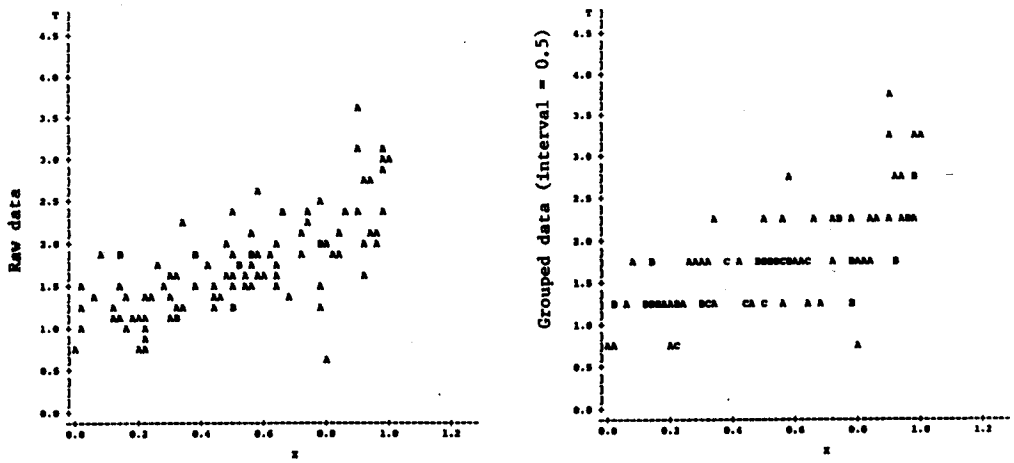

Figure 1b - First 100 observations for log-linear Normal model
(intercept = 0.0, slope = 1.0, std. dev. = 0.25)

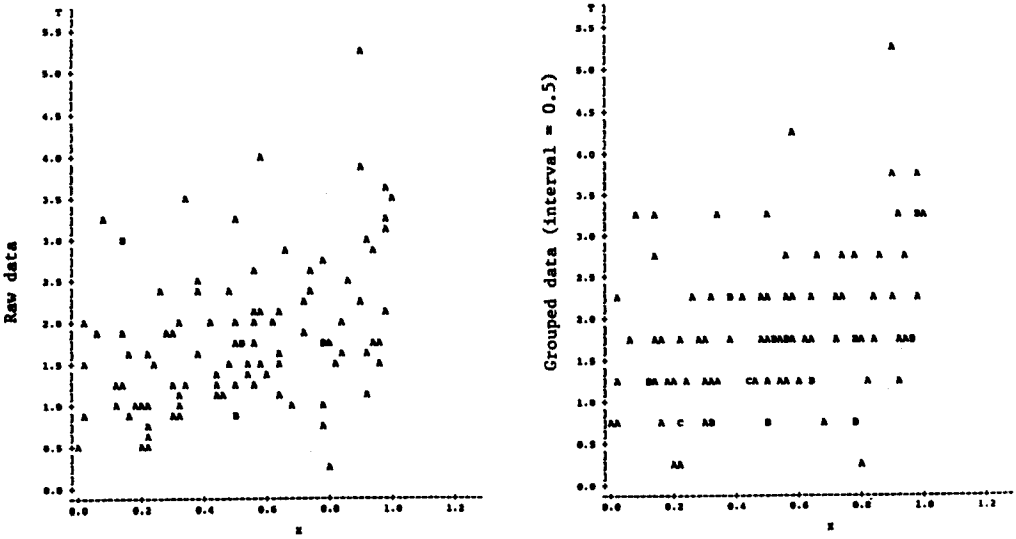

Figure 1c - First 100 observations for log-linear Normal model
(intercept = 0.0, slope = 1.0, std. dev. = 0.50)

interval of width zero implies no grouping.) The vertical dimension gives: (a) the mean of the 100 Monte Carlo estimates by the midpoint method; (b) the mean of the 100 Monte Carlo estimates by the grouped data method; and implicitly, (c) the "true" parameter value used to generate the samples. To avoid clutter in the graphs, standard error bounds at each grouping coarseness are not depicted. Instead, the approximate range of values for the asymptotic standard error point estimates (across all grouping intervals) is stated separately (as "se(m)" for the midpoint method and as "se(g)" for the grouped data estimates). Furthermore, as the rounding intervals widen, there is eventually too little information remaining in the rounded data to identify the dispersion in the dependent variable, and the grouped data estimates become unattainable after a certain point. Midpoint estimates remain feasible even when the grouping coarseness becomes unreasonable.

In Figure 1, panel (i) describes the effects of increasingly coarse rounding intervals in a model where the true standard deviation is 0.10. (To illustrate, Figure 1a shows the first 100 observations on t and x for this model, and for comparison, also gives the scatter diagram when t is rounded to the midpoint of an interval of width 0.50.) We see that for rounding intervals up to about 0.5, the midpoint method seems to yield relatively good approximations to the grouped data estimates. For coarser groupings, however, the midpoint method begins to oscillate away from the grouped data estimates.

Panel (ii) describes the same basic model, but this time with a true standard deviation of 0.25. (Figure 1b shows the corresponding scatter diagrams.) Here the correspondence between the two methods is reasonably good for rounding interval widths up to about three standard deviations (i.e., 0.75), but the midpoint method estimates diverge rapidly for larger interval widths.

Panel (iii), along with Figure 1c, describes the situation for the same regression, but with a true standard deviation of 0.50. The divergence between estimates using the two techniques is now quite marked. Note that the estimates for the grouped data technique are consistently very close to the true parameter values, but the midpoint method estimates diverge rapidly and markedly from the true values. The midpoint estimates, interestingly, are biased first in one direction, and then swing back to become biased in the opposite direction, with the bias in the second direction apparently becoming monotonically larger with increasing interval width.

The important result, then, is that -- especially in log-linear models -- the parameter estimates from the midpoint method can be markedly biased and the grouped data method tends to produce much more accurate point estimates. Furthermore, the direction of bias in the midpoint method seems to vary with the width of the rounding interval, so that it is not even possible, for the log-linear model, to claim that the incorrect midpoint method estimates represent either an upper or lower bound on the true parameter values. In light of this, the grouped data method would seem to be much preferred.

Non-Normal Regression Models

Detailed assessments of the consequences of grouping in the linear and log-linear Exponential and Weibull models are presented in Appendix 2. In summary, it appears that the linear and log-linear Exponential Regression Models yield biased estimates in small samples, even with no grouping in the dependent variable. Conditional on this fundamental bias, however, the grouped data estimates exhibit consistently better conformity with the parameter estimates obtained using ungrouped data. The Weibull regression estimates also show small-sample bias. While the superiority of the grouped data MLE slope and intercept estimates is less obvious for these models, the

shape parameter for the conditional distribution is much more accurately estimated by the grouped data technique than by the midpoint method. See Appendix 2 for diagrammatic summaries.

VII. An Illustration: Long Distance Telephone Call Durations

To illustrate the impact of using ordinary least squares regression analysis with rounded data in a log-linear Normal model, I will make use of two datasets pertaining to long-distance telephone calls (utilized in Cameron and White (1985)). One shortcoming in the datasets is that the telephone utility in question rounds the duration of each call upwards to the next integer number of minutes. (This is a common practice in long-distance billing, although some smaller companies now bill in tenths of minutes.) The previous paper focuses on using the Generalized Gamma density function for the conditional distribution of call duration. The overall conclusion of the paper is that the Generalized Gamma distributions can be superior to the typical assumption of normality for the logarithm of the dependent variable. For both datasets, however, the log-linear Normal model yielded coefficients which were very close to the log-linear Generalized Gamma model (and markedly different than those for the simple Gamma, Weibull, or Exponential log-linear regression models). Consequently, I will use the log-linear Normal model to illustrate the consequences of ignoring the grouping in the data.

The first dataset consists of a stratified sample of 21,738 residential long distance telephone calls originating in the Canadian province of British Columbia on July 13, 1983. These calls were directed to Canada and the U.S. (except Alaska). A second (complete) sample consists of both business and residential calls to all overseas destinations, yielding a total of 4,934 calls. The dependent variable in both datasets is the log of duration in minutes. The explanatory variables common to both datasets include: marginal

per-minute rate in dollars, log of distance in miles, and a set of dummy variables taking on a value of one if a call is credit card or third party, person-to-person, or originating in the vast, sparsely-populated Northeast region. For the first sample, there are also dummy variables for the evening rate discount period and for the night discount period, as well as for collect calls. For the overseas sample, there is only a single (night) discount period, and a dummy variable is also included for business calls. Table 2 gives the parameter point estimates and the asymptotic standard errors obtained (i) the "midpoint" (in this case, "endpoint") method, and (ii) by the more-correct grouped data technique.

Scatter plots of the data for the original study reveal that there is considerable dispersion in the data. For the logarithms of the data on duration, the first sample has a marginal mean of 1.368 and a marginal standard deviation of 1.007. For the overseas sample, the corresponding mean is 1.635 and the standard deviation is .9484. Given the Monte Carlo evidence, this suggests that the parameter bias introduced by the rounding up to the nearest minute in the raw data could be quite significant. Table 2 bears out this hypothesis. It would seem that we are at a level of rounding coarseness comparable to that in Figure 1, panel (iii) beyond about 2.25. Explicit recognition of the rounding process in the estimation yields a smaller estimate for the intercept, larger estimates for the slope coefficients, and a larger estimate of the standard deviation.

For rate policies, telephone companies may be very interested in the change in the typical durations of long-distance calls as a result of adjustments to the marginal per minute cost of a long-distance call. For the first sample, rates do not appear to be a statistically significant determinant of duration in the log-linear model (although they are significant, with a

TABLE 2

Comparison of Estimates by OLS and by Grouped Data Technique

Variable	Canada and U.S. Destinations		Overseas Destinations	
	OLS Midpoint Method	MLE Grouped Data Method	OLS Midpoint Method	MLE Grouped Data Method
Intercept	.4747 (.0228)	.0040 (.0301)	2.251 (.754)	2.216 (.9991)
RATE	.0280 (.0894)	.1118 (.1163)	-.1184 (.0557)	-.1547 (.0706)
LOG(DIST)	.1482 (.0111)	.1756 (.0145)	-.0453 (.0954)	-.0521 (.1283)
EVENING	.4318 (.0198)	.5507 (.0257)		
NIGHT	.4171 (.0381)	.5539 (.0495)	.0794 (.0489)	.0829 (.0594)
BUSINESS			-.0780 (.0301)	-.0957 (.0360)
COLLECT	.1740 (.0266)	.2349 (.2451)		
CARD	-.0709 (.0284)	-.0710 (.0366)	.1572 (.0731)	.1819 (.0869)
PERSON	.0284 (.0726)	.0469 (.0923)	.4245 (.0706)	.5153 (.0834)
NORTHEAST	-.0731 (.0133)	-.0936 (.0172)	-.0344 (.0433)	.0458 (.0516)
σ	.9511	1.189 (.0068)	.9404	1.107 (.0126)

negative coefficient, in a linear Log-normal specification in the Cameron and White paper). The point estimates are more intuitively plausible for the log-linear Normal specification with the overseas data. In this second sample, the model which corrects for the rounding suggests that in response to a 10-cent per minute decrease in marginal per-minute rates, the mean duration of calls will increase by .015 percent, rather than by .011 percent. Thus the grouped data model uncovers an apparently greater elasticity of duration with respect to rates than would be implied by the OLS method using rounded data. Given the large number of calls involved per year, even seemingly modest differences in the elasticity estimates could mean a very large change in revenue predictions.

VIII. Implications and Conclusions

With a study of this nature, the ideal summary statement would include a strategy for diagnosis of a "rounding bias hazard" and guidelines for assessing the circumstances under which acknowledgement of this bias is called for. Unfortunately, the range of simulations has not been sufficiently exhaustive to allow such a concise summary. The primary purpose of this paper has been to demonstrate that there do exist conditions under which the grouped technique is very likely to be superior.

Clearly, if the true errors are normally distributed, there is negligible loss from using the midpoint technique instead of the grouped data regression method. Both appear to be about equally accurate (inaccurate). A researcher with strong convictions about the normality of the error distribution of the grouped variable around the true regression relationship could probably argue that no special techniques are necessary. However, one of the important findings of this paper is that when the density function for the true errors deviates from the familiar normal curve, the midpoint technique does not

perform as well. In particular, the popular log-linear normal model can yield seriously biased parameter estimates when the untransformed variable is grouped or rounded. The adverse consequences of using the midpoint approximation technique instead of the more-appropriate grouped data method are most important when the ungrouped data would exhibit a very diffuse relationship. I have illustrated the potential quantitative consequences of incorrect estimation using two samples of real data on the durations of long-distance telephone calls.

I have also offered formulas for estimation in other situations where the dependent variable is constrained to be strictly positive, so that an Exponential density, or perhaps a Weibull form is more appropriate. By making available a set of general formulas for these alternative models, I hope to facilitate the efforts of other researchers who may wish to explore the consequences of deficient data. While even ungrouped regressions yield biased point estimates in small samples in these alternative models, the grouped regression method can easily claim superiority in large samples, or if identical correction formulas can be applied to both sets of estimates.

In sum, when working with grouped or rounded data for a dependent variable, one must pay attention to two factors: (i) the coarseness of the grouping interval, and (ii) their priors regarding the true shape of the underlying conditional density of the dependent variable. But as usual, whether or not the bias in parameters introduced by the use of midpoint regression on grouped or rounded data actually matters in the "grand scheme of things" will depend on the precise application involved.

Footnotes

[1]Kreiger and Gastwirth (1984) address the issue of grouped data when accurate interpolation of different percentiles is required. In lieu of fitting a parametric curve, they introduce a method for deriving upper and lower bounds on percentiles assuming only that the underlying density function is unimodal and that the interval in which the mode lies is known. It must be emphasized, however, that this is a univariate technique; regression analysis is not involved. They conclude that data providers should be encouraged to report group means along with group frequencies. When this information is available, the spread between the percentile bounds can be dramatically reduced.

[2]Some of the MLE formulas in this subsection are similar to those derived in Section 2 of Stewart (1983). They are reproduced here for consistency with the general notation adopted in this paper.

[3]After the transformation, it is possible to express $\tilde{y}$ in "location-scale" form: $\tilde{y} = x'\beta + \sigma z$.

[4]While the symmetry of this distribution could be exploited, as in Tobit models, I leave the values as they are so that this version conforms to the general model.

[5]It follows from the probability integral transformation (DeGroot, 1975, pp. 127-29) that a single value from an exponentially-distributed random variable with mean $x_i'\beta$ can be generated from a value u_i drawn from a U(0,1) distribution by the transformation:

$$t_i = -(x_i'\beta)\log(u_i).$$

For the linear Weibull model, the corresponding transformation is:

$$t_i = \frac{(x_i'\beta)}{\Gamma(\frac{c+1}{c})} [-\log(u_i)].$$

[6]For the log-linear models, the means of the conditional distributions must be $\exp(x_i'\beta)$. Exponential samples for t_i can be generated from random drawings from a U(0,1) distribution through the transformation:

$$t_i = -\exp(x_i'\beta)\log(u_i)$$

Weibull samples can be generated by:

$$t_i = \frac{\exp(x_i'\beta)}{\Gamma(\frac{c+1}{c})} [-\log(u_i)]$$

[7]For a given slope and intercept, the bias is larger, the larger the underlying error variance. Similarly, for a given error variance, the bias is larger for smaller slopes. Other results have shown the same direction for the bias in the absolute magnitude of the coefficients when a negative slope is considered. In general, the regression line becomes "flatter" due to rounding.

[8]An alternative strategy might have been to formulate the ad hoc midpoint assignment in terms of the midpoint of the logarithms of the intervals (rather than using the logs of the midpoints). This technique has not yet been examined. In any event, it would require the ad hoc assignment of a value for the midpoint of the interval between log(0) (negative infinity) and the log of the upper limit of the first interval.

REFERENCES

Berndt, E.R., B. H. Hall, R. E. Hall, and J. A. Hausman (1974), "Estimation and Inference in Non-Linear Structural Models," Annals of Economic and Social Measurement, pp. 653-665.

Burridge, J. (1982) "A Note on Maximum Likelihood Estimation for Regression Models Using Grouped Data," Journal of the Royal Statistical Society, Series B, vol. 43, 41-45.

Cameron, T. A., and White, K. J. (1985) "Generalized Gamma Family Regression Models for Long Distance Telephone Call Durations," Discussion Paper #363. Department of Economics, University of California, Los Angeles.

DeGroot, Morris H. (1975) Probability and Statistics, Reading, MA: Addison-Wesley.

Dempster, Laird, and Rubin (1977) "Maximum Likelihood from Incomplete Data Via the EM Algorithm," Journal of the Royal Statistical Society, Series B. Vol. 38, pp. 1-22.

Hasselblad, V., Stead, A. G., and Galke, W. (1980) "Analysis of Coarsely Grouped Data from the Lognormal Distribution," Journal of the American Statistical Association, December, Vol. 75, No. 372, pp. 771-778.

Hastings, N.A.J., and J.B. Peacock (1974) Statistical Distributions, New York: John Wiley and Sons.

Kreiger, Abba M. and Joseph L. Gastwirth (1984) "Interpolation from Grouped Data With Unimodal Densities," Econometrica, 52, 419-26.

Lawless, J. F. (1982) Statistical Models and Methods for Lifetime Data, New York: John Wiley and Sons.

Stewart, Mark B. (1983) "On Least Squares Estimation When the Dependent Variable is Grouped," Review of Economic Studies, 50, 737-53.

APPENDIX 1 - Specialized Hessian Components

A. The Linear Normal Regression Model

$$F_{(\beta_r\beta_s)i} = \frac{x_{ir}x_{is}}{\sigma}\left[\phi'(z_{ui}) - \phi'(z_{\ell i})\right] \quad r,s = 1,\ldots,p.$$

$$F_{(\beta_r\sigma)i} = \frac{x_{ir}}{\sigma^2}\left[(\phi(z_{ui}) - \phi(z_{\ell i})) + (z_{ui}\phi(z_{ui}) - z_{\ell i}\phi(z_{\ell i}))\right] \quad r = 1,\ldots,p.$$

$$F_{(\sigma\sigma)i} = \frac{-1}{\sigma}\left[z_{ui}\phi'(z_{ui}) - z_{\ell i}\phi'(z_{\ell i})\right]$$

$$+ \frac{2}{\sigma^2}\left[z_{ui}\phi(z_{ui}) - z_{\ell i}\phi(z_{\ell i})\right]$$

B. The Linear Exponential Model

$$F_{(\beta_r\beta_s)i} = x^*_{ir}x^*_{is}\left[(t^*_{\ell i}-2)t^*_{\ell i}\exp(-t^*_{\ell i}) - (t^*_{ui}-2)t^*_{ui}\exp(-t^*_{ui})\right]$$

$$r,s = 1,\ldots,p$$

C. The Linear Weibull Model

$$F_{(\beta_r\beta_s)i} = -(1+c)cx^*_{ir}x^*_{is}\left[T^c_{\ell i}\exp(-T^c_{\ell i}) - T^c_{ui}\exp(-T^c_{ui})\right]$$

$$+ c^2x^*_{ir}x^*_{is}\left[T^{2c}_{\ell i} - T^{2c}_{ui}\right] \quad r,s = 1,\ldots,p.$$

$$F_{(\beta_r c)i} = x^*_{ir}\left[T^c_{\ell i}\exp(-T^c_{\ell i}) - T^c_{ui}\exp(-T^c_{ui})\right]$$

$$+ cx^*_{ir}\left[(1-T^c_{\ell i})\left(\frac{\partial T^c_{\ell i}}{\partial c}\right)\exp(-T^c_{\ell i}) - (1-T^c_{ui})\left(\frac{\partial T^c_{ui}}{\partial c}\right)\exp(-T^c_{ui})\right] \quad r = 1,\ldots,p$$

where $$\left(\frac{\partial T^c_{.i}}{\partial c}\right) = \left(T^{(c-1)}_{.i}\, t^*_{\ell i}\,\frac{\Gamma'(c^*)}{c} + T^c_{.i}\,\log T_{.i}\right)$$

Now let

$$R_{.i} = -t^*_{.i}\,\frac{\Gamma'(c^*)}{c^2}\,T^{(c-1)}_{.i}$$

$$S_{.i} = (\log T_{.i})\, T_{.i}^{(c-1)}$$

$$F_{(cc)i} = (\{[R_{\ell i} + \frac{t^*_{\ell i}}{c}(\frac{-\Gamma''(c^*)}{c^2} T_{\ell i}^{(c-1)} + \Gamma'(c^*)[(c-1)T_{\ell i}R_{\ell i} + S_{\ell i}])]$$

$$- [(cR_{\ell i} + S_{\ell i}) \log T_{\ell i} + R_{\ell i}]\} \exp(-T_{\ell i}^{c}) +$$

$$\{-[cR_{\ell i} + T_{\ell i}S_{\ell i}]\ [(R_{\ell i} - S_{\ell i}) \exp(-T_{\ell i}^{c})]\})$$

$$- (\{[R_{ui} + \frac{t^*_{ui}}{c}(\frac{-\Gamma''(c^*)}{c^2} T_{ui}^{(c-1)} + \Gamma'(c^*)[(c-1)T_{ui}R_{ui} + S_{ui}])]$$

$$- [(cR_{ui} + S_{ui}) \log T_{ui} + R_{ui}]\} \exp(-T_{ui}^{c}) +$$

$$\{-[cR_{ui} + T_{ui}S_{ui}]\ [(R_{ui} - S_{ui}) \exp(-T_{ui}^{c})]\})$$

APPENDIX 2 - Results for Models With Alternative Distributional Assumptions

I. Linear Models With Alternative Distributional Assumptions

A. Linear Exponential Regression

The relative performance of exponential regression compared to ordinary least squares will not be evaluated here. (See Cameron and White, 1985.)

Since an exponential density function is highly skewed, we would expect qualitatively different results than in the normal case. The mean of an exponential distribution completely characterizes its shape, so only intercept and slope parameters are involved in an exponential regression model. Figure A3.1 summarizes the consequences of grouping the dependent variable for three different underlying true regressions.

Observations on these results are:

a) Unlike the outcomes with the normal error terms, we see in Figure A3.1 that the point estimates for the grouped data regressions, while biased even with no grouping of the data (interval width = 0), correspond more closely with the true underlying parameters than do the midpoint regression estimates. While, as expected, the two methods yield virtually identical results at very narrow grouping intervals, the midpoint regression estimates depart quite rapidly from the grouped data results as the intervals widen.

b) The fact that the means of the sample point estimates are biased even without grouping seems to be a small-sample property of the MLE estimators in these models. For the exponential model with slope parameter 1.5, rerunning the ungrouped simulations with 2000 observations instead of just 200 yields a mean intercept estimate of .998 and a mean slope estimate of 1.498. Both values are much closer to the true parameters than in the small sample simulation.

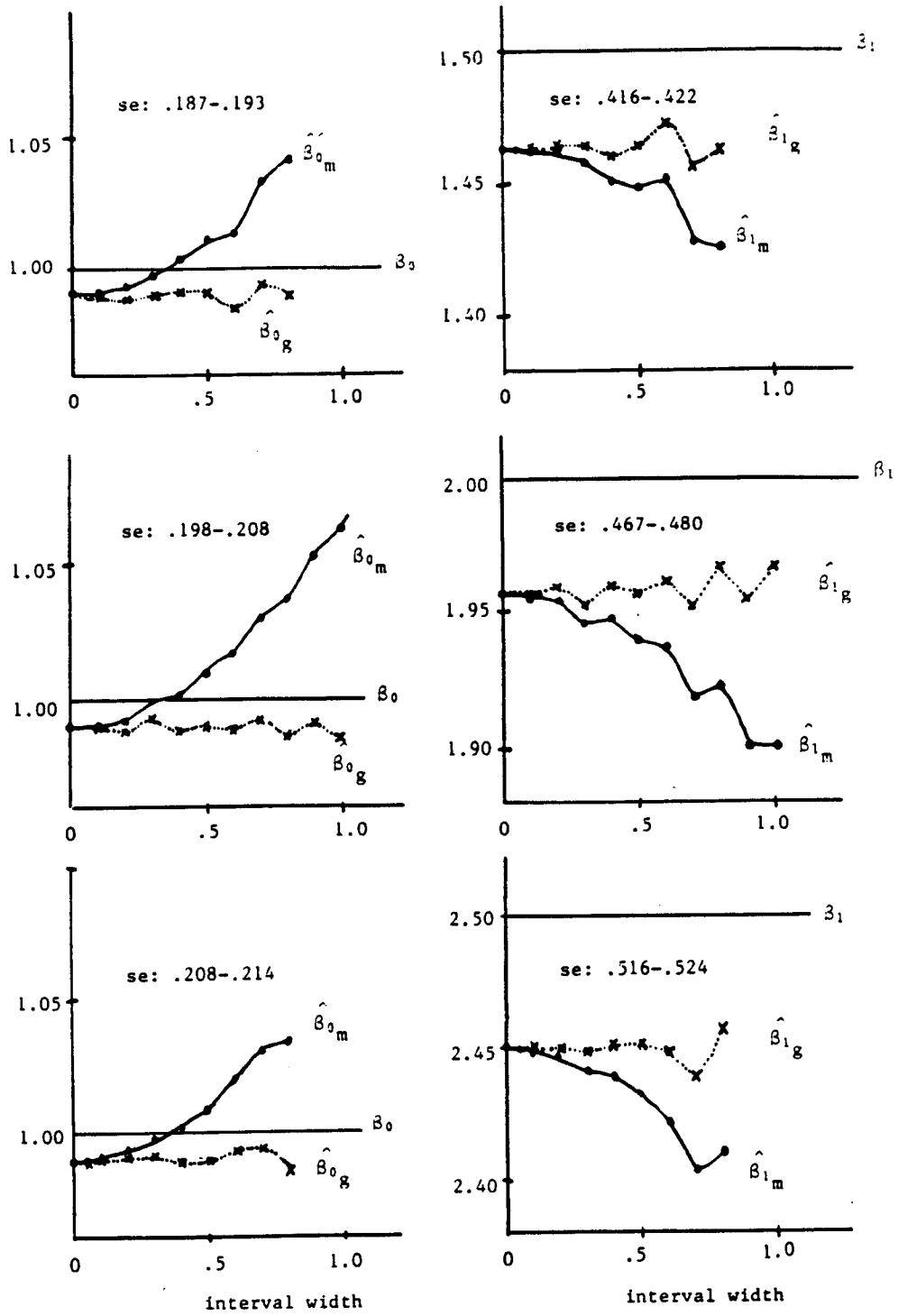

Figure A3.1 - Mean point estimates, linear Exponential models, 100 Monte Carlo samples, n=200; grouped data method (g) versus midpoint technique (m); identical intercepts, varying slopes. Horizontal line = true parameter value, "se" = range of mean asymptotic standard errors across all interval widths and both estimation methods.

Since the likelihood function for this model is a highly non-linear function of the unknown parameters, finding a simple analytic formula for the bias is likely to be a non-trivial task. However, if such a formula were available, and if it could be applied to both the grouped data and midpoint estimates, the result would be extremely accurate estimates from the grouped data method and rapidly unsatisfactory results from the midpoint technique.

B. Linear Weibull Regression

The linear Weibull regression model is simply a generalization of the exponential regression model which incorporates a "shape" parameter c. (When $c = 1$, we have the linear exponential model.) One interesting aspect of the grouped data problem is the extent to which grouping will bias the model choice for or against an exponential specification, when the underlying errors are Weibull-distributed.

Figure A3.2 shows the simulation results for one Weibull linear regression model.

a) Note first of all that the Weibull regression algorithm also seems to yield biased estimates of the intercept and slope parameters in small samples, even with no grouping.

b) For these graphs, bear in mind that the entire range of the ungrouped dependent variable is approximately 0 to 10, so that an interval width of 7 units, is rather extreme and would probably not be relevant to any sensible empirical application. A researcher might, however, attempt a regression model when the grouping intervals are up to a width of 2 or 3. In this context, the remaining observations are made.

c) The grouped data technique performs marginally better in estimating the slope parameter, but marginally worse for the intercept parameter. For the shape parameter, however, the grouped data technique performs much better,

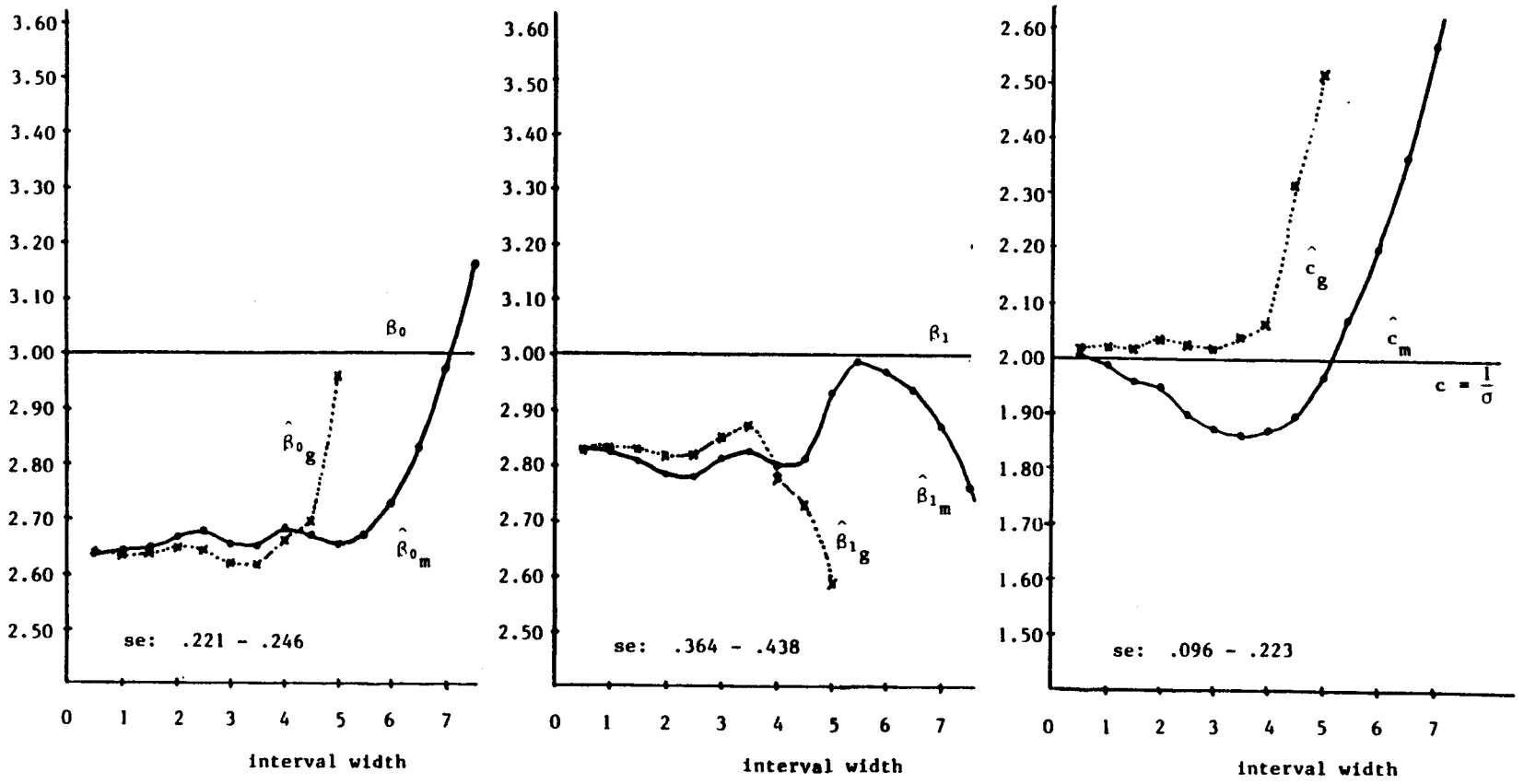

Figure A3.2 - Mean point estimates, a linear Weibull model, 100 Monte Carlo samples, n=200; grouped data method (g) versus midpoint technique (m); Horizontal line = true parameter value, "se" = range of mean asymptotic standard errors across all relevant interval widths and both estimation methods.

yielding estimates very close to the true shape parameter of 2.00. The midpoint technique, though, rapidly begins to underestimate the shape parameter, suggesting an error distribution that is initially closer to exponential. After the rounding becomes extremely (unreasonably) coarse, both methods begin to overstate drastically the magnitude of the shape parameter. (Incidentally, a shape parameter of 3.0 yields an almost-symmetric Weibull density function.)

d) The asymptotic standard error magnitudes given in Figure A3.2 pertain only to rounding intervals up to 4. Beyond this level, the standard errors of the estimates quickly become exceedingly large.

e) In sum, since economists are generally more interested in the slope estimates than in the intercept estimates, the grouped data technique would seem to be marginally preferred when errors are Weibull-distributed, especially if some correction for small-sample bias can be made.

II. Log-Linear Models With Alternative Distributional Assumptions

A. Log-linear Exponential Regression

The density function for the logarithm of an exponentially distributed random variable has a standard extreme-value density function. Again, we use interval midpoints of the untransformed data as proxies for the true values of the dependent variable. These midpoint estimates are compared to the results for the corresponding grouped data regression in Figure A3.3. Bear in mind that these parameters pertain to an underlying relationship of the form: $t = \exp(x'\beta)$.

a) The configurations of the point estimates are reminiscent of those in Figure A3.1. The point estimates are biased, but again, there is strong suspicion that this is merely a small-sample phenomenon.

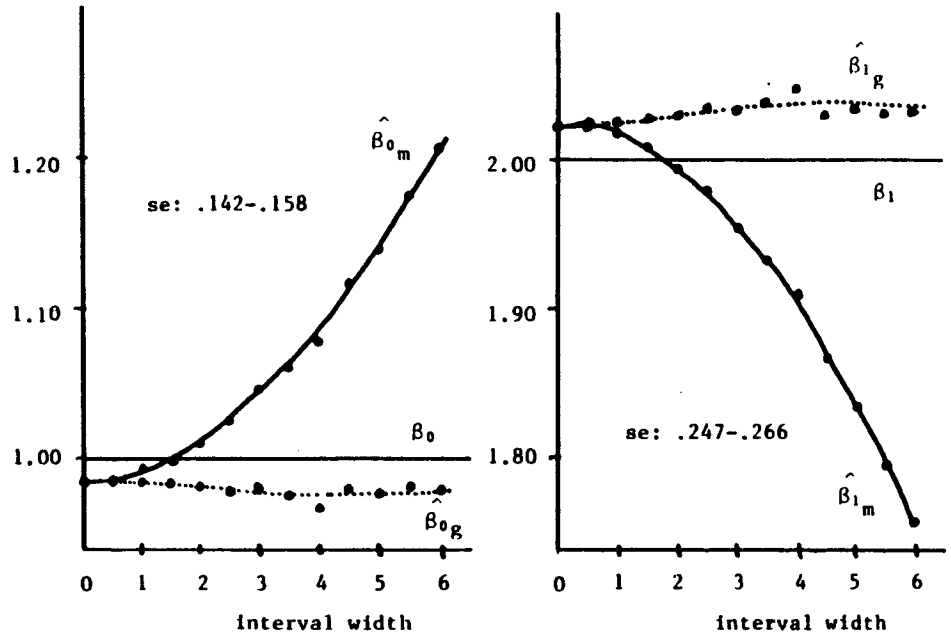

Figure A3.3 - Mean point estimates, a log-linear Exponential model, 100 Monte Carlo samples, n=200; grouped data method (g) versus midpoint technique (m); Horizontal line = true parameter value, "se" = range of mean asymptotic standard errors across all relevant interval widths and both estimation methods.

b) If a formula to correct for the underlying bias in the parameters could be derived, correcting both types of estimates by the same factor would again result in much-superior estimates by the grouped method than by midpoints.

B. Log-linear Weibull Regression

For log-linear specifications, the Weibull model is again a generalization of the exponential regression model. To facilitate estimation, the c parameter is transformed into $\sigma = 1/c$ (the slope and intercept parameters reflect the linear relationship which results after taking the log of the dependent variable). Results are depicted in Figure A3.4, from which we make the following observations:

a) With no grouping of the data, this regression model yields an intercept estimate which is biased slightly downward, and a slope estimate with a slight upward bias. With an appropriate correction for these sample-size biases, the grouped technique will again be superior.

b) For the shape parameter, the grouped data technique gives remarkably constant (if again slightly biased) point estimates. The midpoint method, however, exhibits a behavior pattern similar to the results for the c parameter in the linear model. (Note that this graph shows $\sigma = 1/c$.) The midpoint estimate of σ is first biased towards 1 (exponential errors), but then returns to shoot well beyond the true value, becoming very small with wider rounding intervals.

c) For rounding intervals beyond 5, some of the 100 random samples resulted in data sets which, due to groupings too large relative to the variance in the errors, did not allow the algorithm for the grouped data model to converge.

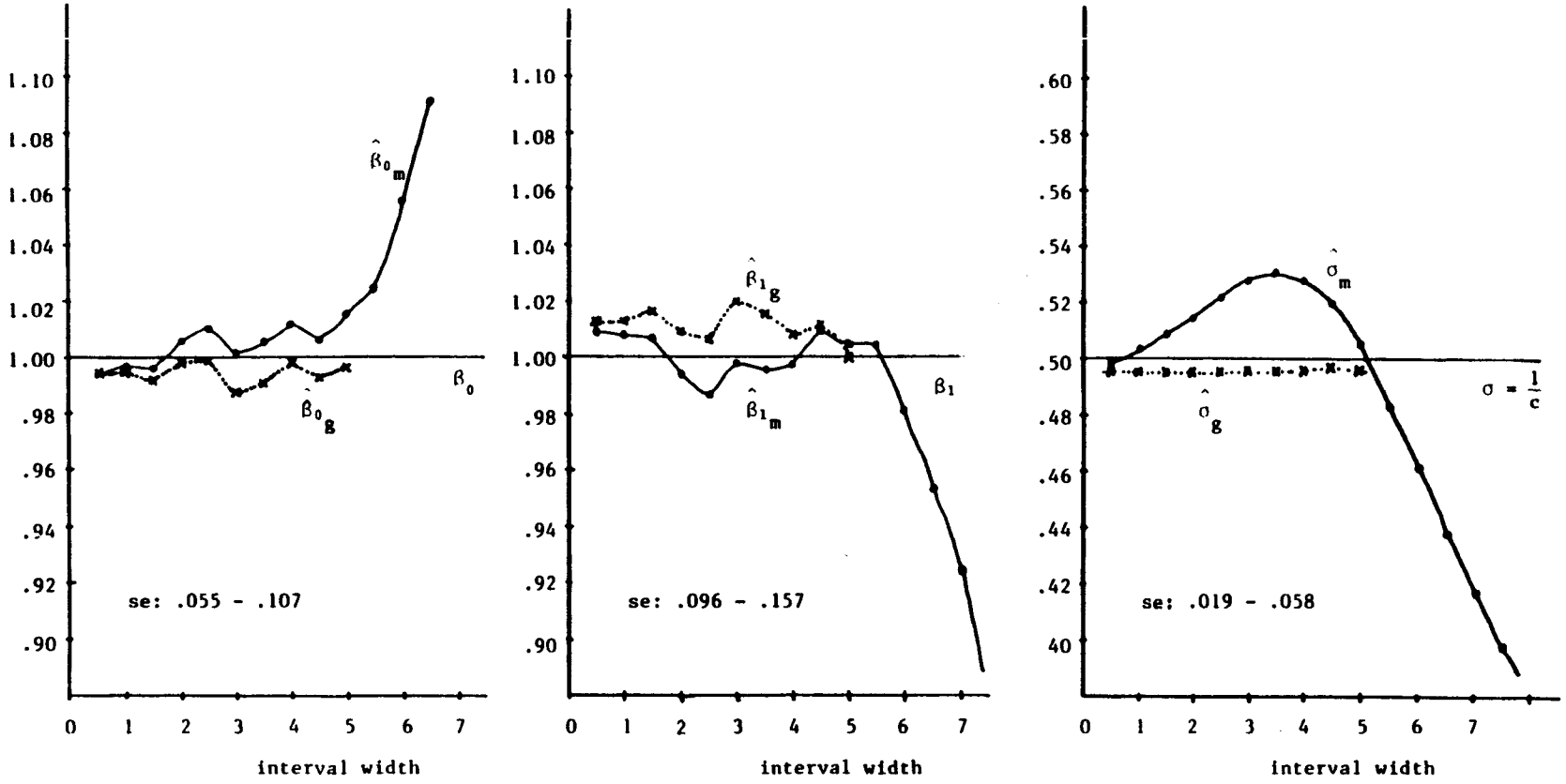

Figure A3.4 - Mean point estimates, a log-linear Weibull model, 100 Monte Carlo samples, n=200; grouped data method (g) versus midpoint technique (m); Horizontal line = true parameter value, "se" = range of mean asymptotic standard errors across all relevant interval widths and both estimation methods.