Confidence Interval Coverage Properties for Regression Estimators in Uni-Phase and Two-Phase Sampling

J.N.K. Rao, W. Jocelyn, and M.A. Hidiroglou

Confidence intervals based on the normal approximation are widely used in the design-based approach. Hansen, Madow and Tepping (1983) noted that design-based intervals are inferentially satisfactory despite failures of assumed models. Dorfman (1994) studied confidence interval coverage associated with the sample linear regression estimator under two-phase random sampling using standard and new variance estimators, and concluded that the contention of Hansen et al. is not tenable. In this article we provide reasons for the poor performance under model failures and practical solutions to improve the coverage probability.

1. Introduction

Validity of normal approximation based confidence intervals on the population mean, Υ, has been studied, both theoretically and through simulations, for simple random sampling (SRS). Madow (1948) and Hájek (1960) gave conditions under which the design-based distribution of the sample mean, ȳ, tends to normality. Cochran (1977, p. 42) gave a working rule for the minimum sample size, n, necessary for the normal approximation to hold for the standardized variable \( Z = (\bar{y} - \bar{Y})/\sigma(\bar{y}) \), where \( \sigma^2(\bar{y}) = (1/n - 1/N)S^2 \) is the variance of \( \bar{y} \), \( N \) is the population size, \( S^2 = N\sigma^2/(N-1) \) and \( \sigma^2 = \Sigma(y_i - \bar{Y})^2/N \) is the population variance. For populations positively skewed, Cochran’s rule is given by \( n > 25 \gamma^2 \), where \( \gamma = \Sigma(y_i - \bar{Y})^3/(N\sigma^3) \) is the skewness coefficient (\( \gamma = 0 \) for normal populations). Dalén (1986) used the rule \( n > K_{1-\alpha} \gamma_1^2 \), where \( K_{1-\alpha} \) depends on the nominal coverage probability \( 1 - \alpha \) and \( \gamma_1 = \Sigma|y_i - \bar{Y}|^3/(N\sigma^3) \). His empirical results supported the use of Student’s \( t \)-approximation over the normal approximation for smaller \( n \). For nominal \( \alpha = 0.95 \), he recommended \( K_{1-0.95} = 20 \) provided \( \gamma_1 < 3 \). Sugden, Smith, and Jones (2000) extended Cochran’s rule to the studentized variable \( t = (\bar{y} - \bar{Y})/s(\bar{y}) \), where \( s^2(\bar{y}) = (1/n - 1/N)s_1^2 \) is the estimated variance of \( \bar{y} \), and \( s_1^2 \) is the sample variance. Smith’s rule is given by \( n > 28 + 25\gamma^2 \). Note that \( t \) is used in practice because \( Z \) depends on the unknown population variance \( S^2 \). Sugden et al. (2000) also noted that design-based inference strongly depends on the validity of the normal approximation.

Hansen, Madow, and Tepping (1983) demonstrated that model-dependent confidence intervals can perform poorly under moderate departures from the assumed model. For

---

1 J.N.K. Rao is Professor Emeritus, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6. W. Jocelyn is Senior Methodologist and M.A. Hidiroglou is Assistant Director, Business Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada.

Acknowledgments: J.N.K. Rao’s research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors are thankful to a referee for insightful comments and constructive suggestions.
this purpose, they constructed a synthetic population, resembling business populations
with positive skewness, using a model misspecification that may not be detectable
through tests of significance for sample sizes as large as 400. By simulating stratified
random samples from this population using equal allocation, they demonstrated that in
accordance with design-based normal theory, two-sided confidence intervals on \( \bar{Y} \)
perform well in terms of coverage as the sample size increases. On the other hand,
the design-based coverage probability of confidence intervals based on the misspecified
model is substantially smaller than the nominal level \( \alpha = 0.95 \) and it becomes worse as
the sample size increases. Based on these results, Hansen, Madow, and Tepping (1983, p. 791) concluded that (i) “probability-sampling methods when carefully applied with
reasonably large samples, provide protection against failures of assumed models . . .”
and (ii) “. . . with reasonably large samples the inferences should not depend on the
model.” Brewer and Särndal (1983) made a comment similar to (i) and (ii), but some-
what stronger than (i): “probability sampling methods are robust by definition; since
they do not appeal to a model, there is no need to discuss what happens under model
breakdown.” We refer the reader to Valliant, Dorfman, and Royall (2000, pp. 87–90)
for a critical discussion of Hansen, Madow, and Tepping’s (1983) approach to detecting
model misspecification.

Dorfman (1994) compared several variance estimators of the simple linear regres-
sion estimator of \( \bar{Y} \) for two-phase simple random sampling. His simulation study indi-
cated that the resulting normal theory two-sided confidence intervals have poor
coverage properties when the underlying model generating the population is grossly
misspecified. Dorfman (1994, p. 139) concluded that the contention of Hansen,
Madow, and Tepping (1983) is not tenable: “The results on coverage of the regres-
sion estimator under a quadratic model . . . dramatically call this contention into
question.”

In this article, we study two-phase sampling and provide reasons for the poor per-
formance of design-based normal theory intervals, even for moderately large second-phase
samples, when the underlying model is grossly misspecified. We also propose practical
solutions to improve the coverage probability. The case of simple random sampling is
also studied, both theoretically and through simulation.

Remaining sections of this article are organized as follows. Section 2 provides some
theoretical insights based on Edgeworth expansions under simple random sampling
(SRS). Section 3 presents simulation results for simple random sampling. Section 4
studies the case of general uni-phase sampling, using design weights. Two-phase
random sampling is studied in Section 5. Some simulation results for two-phase random
sampling are presented in Section 6. Finally, some summary remarks are given in
Section 7.

2. Simple Random Sampling

In sample surveys, we are often interested in two-sided confidence intervals. Moreover,
both Hansen et al. (1983) and Dorfman (1994) considered two-sided intervals. We
therefore focus on two-sided intervals. In this section, we study the coverage probability
of normal theory two-sided intervals under simple random sampling.
2.1. Edgeworth expansions

Edgeworth expansions provide theoretical insights into the performance of normal theory intervals based on the $t$-variable, $t = (\bar{Y} - \bar{Y}) / s(\bar{Y})$. For simplicity, we assume that the sampling fraction, $n/N$, is negligible. Under some regularity conditions, we have the following Edgeworth expansion for the coverage error of the $(1 - \alpha)$-level normal theory interval on $\bar{Y}$, $I_{1-\alpha} = [\bar{Y} - z_{\alpha/2} s(\bar{Y}), \bar{Y} + z_{\alpha/2} s(\bar{Y})]$, where $z_{\alpha/2}$ is the upper $\alpha/2$-point of a $N(0, 1)$ variable:

$$CE = \Pr(\bar{Y} \in I_{1-\alpha}) - (1 - \alpha)$$

$$= \frac{2z_{\alpha/2}}{n} \left[ \frac{1}{12} \kappa(z_{\alpha/2}^2 - 3) - \frac{1}{18} \gamma^2(z_{\alpha/2}^4 + 2z_{\alpha/2}^2 - 3) - \frac{1}{4} (z_{\alpha/2}^2 + 3) \right] \phi(z_{\alpha/2}) \quad (2.1)$$

where $\phi(\cdot)$ is the probability density function of a $N(0, 1)$ variable and $\kappa = \Sigma(Y_i - \bar{Y})^4/(N\sigma^4) - 3$ is the kurtosis coefficient; see Hall (1992, Chapter 2).

Suppose $\alpha = 0.95$ so that $z_{\alpha/2} = 2$. In this case, it follows from (2.1) that the actual coverage probability will be smaller than the nominal level if

$$\kappa < 7(2\gamma^2 + 3) \quad (2.2)$$

That is, a large skewness coefficient (not necessarily positive) can lead to coverage probability substantially smaller than $1 - \alpha = 0.95$. Note that the coverage error $CE$ is of the order $n^{-1}$. If $\kappa < 21$, then (2.2) is satisfied for any $\gamma$.

In this article, we have focused on two-sided normal theory intervals, following Dorfman (1994), but it is also of interest to study one-sided normal theory intervals. In the latter case, we show that the effect of positive skewness is more pronounced relative to two-sided intervals.

For the one-sided lower interval $I_{1-\alpha}^* = [\bar{Y} - z_{\alpha} s(\bar{Y}), \infty)$ on $\bar{Y}$, the Edgeworth expansion will contain terms of order $n^{-1/2}$, unlike (2.1):

$$CE_1 = \Pr(\bar{Y} \in I_{1-\alpha}^*) - (1 - \alpha) = n^{-1/2} \gamma \frac{\gamma^2}{6} (2z_{\alpha}^2 + 1) \phi(z_{\alpha}) \quad (2.3)$$

It follows from (2.3) that the coverage error $CE_1$ is of order $n^{-1/2}$ and that the coverage probability of the one-sided lower interval $I_{1-\alpha}^*$ will be larger than the nominal level $1 - \alpha$ if the skewness coefficient, $\gamma$, is positive. On the other hand, for the one-sided upper interval $I_{1-\alpha}^{**} = (-\infty, \bar{Y} + z_{\alpha} s(\bar{Y})]$ on $\bar{Y}$, the Edgeworth expansion for the coverage error is given by

$$CE_2 = \Pr(\bar{Y} \in I_{1-\alpha}^{**}) - (1 - \alpha) = -n^{-1/2} \gamma \frac{\gamma^2}{6} (2z_{\alpha}^2 + 1) \phi(z_{\alpha}) \quad (2.4)$$

It follows from (2.4) that the coverage probability of the upper interval will be smaller than the nominal level $1 - \alpha$ if the skewness coefficient, $\gamma$, is positive. The above results are in agreement with the remarks in Cochran (1977, p. 41).

The above results are also valid under the model-based approach, assuming that $y_1, \ldots, y_N$ are independent and identically distributed (iid) variables generated from an infinite superpopulation with mean $\mu$ and variance $\sigma^2$. The coverage error (CE), given by (2.1), now refers to the model-based distribution of the sample mean $\bar{Y}$, and $\gamma$ and $\kappa$ are the skewness and kurtosis coefficients of the superpopulation. Note that for the case of normal superpopulation, we have $\gamma = \kappa = 0$, and (2.2) is satisfied. Here the $t$-variable
has a student-t distribution and we are approximating this distribution by a normal distribution. It is well-known that the normal approximation to student-t leads to coverage probability smaller than the nominal level 1 - α, and that the coverage error (CE) is close to zero for n larger than 30.

2.2. Linear regression estimator

We now study the performance of coverage probabilities associated with the linear regression estimator of $\hat{Y}$. Suppose $x$ is an auxiliary variable with known population mean $\overline{X}$ and correlated with $y$. A simple linear regression estimator of $\hat{Y}$ is given by $\hat{y}_i = \hat{Y} + b(\overline{X} - \overline{x})$, where $b$ is the sample regression coefficient and $\overline{x}$ is the sample mean of $x$. A normal theory interval on $\hat{Y}$, using $\hat{y}_i$, is based on the pivotal quantity

$$t_r = (\hat{y}_r - \overline{Y})/s(\hat{y}_r) \tag{2.5}$$

where $s^2(\hat{y}_r) = (1/n - 1/N)s^2$ and $s^2$ is the sample variance of the residuals $e_i = y_i - \hat{Y} - b(x_i - \overline{x})$.

Suppose that the underlying model generating the finite population is given by $y_i = \alpha + \beta x_i + e_i$, $i = 1, \ldots, N$, where the $e_i$’s are iid errors with mean 0 and variance $\sigma^2$. The errors $e_i$ will have a small value of skewness if the skewness of $y_i$’s depends solely on that of $x_i$’s. If the assumed model holds, then noting that $\overline{y}_r - \overline{Y} = (\overline{y} - \overline{Y}) + \beta(\overline{X} - \overline{x}) = \overline{e} - \overline{e}_N$, we have

$$t_r = (\overline{e} - \overline{e}_N)/s(\overline{e}) \tag{2.6}$$

where $\overline{e}$ is the sample mean of $e_i$’s, $\overline{e}_N = 0$ is the population mean of $e_i$’s and $s^2(\overline{e})$ is the estimate of the variance of $\overline{e}$. The approximation (2.6) shows that the design-based coverage error associated with $t_r$ will be small if the assumed model is (approximately) valid, regardless of the skewness in $y_i$’s (and $x_i$’s).

On the other hand, suppose the true model is a quadratic regression model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i^*$ with $z_i = x_i^2$ and iid errors $e_i^*$. In this case, the sample regression coefficient $b = \Sigma(x_i - \overline{x})y_i/\Sigma(x_i - \overline{x})^2$ converges in probability to $\beta_1^* = \beta_1 + a\beta_2$, where $a = \Sigma(x_i - \overline{x})z_i/\Sigma(x_i - \overline{x})^2$. Also, the numerator of $t_r$ is approximately equal to $-a\beta_2(\overline{x} - \overline{X}) + \beta_2(\overline{z} - \overline{Z}) + (\overline{e}^* - \overline{e}^*_N)$ while the denominator is approximately equal to the estimated variance of the mean $\overline{e}^* = \Sigma e^*_i/n$, where $e^*_i = -a\beta_2(x_i - \overline{x}) + \beta_2(z_i - \overline{z}) + (e^*_i - \overline{e}^*)$. The negative term, $-a\beta_2(\overline{x} - \overline{X})$, in the numerator reduces the skewness effect of the middle term, $\beta_2(\overline{z} - \overline{Z})$, but the middle term is likely to dominate the numerator since the $z_i$’s are typically far larger than the $x_i$’s. As a result, large skewness in $z_i$’s leads to a coverage probability substantially smaller than the nominal $(1 - \alpha)$-level if the $\beta_2$-coefficient is significantly large.

Royall and Cumberland (1985, Section 3) conducted an empirical study on real populations. Their results show that poor coverage is due to high correlation between numerator and denominator induced by large skewness. In their study, the latter correlation was as high as 0.80. Our theoretical result above is in agreement with the empirical finding of Royall and Cumberland.

The above observations suggest that a proper “model-assisted” approach is needed for reducing the coverage errors of design-based confidence intervals. Suppose the population $x_i$-values are all known so that both $\overline{X}$ and $\overline{Z}$ are known. We can first perform suitable
model diagnostics (e.g., residual analysis) on the sample data \( \{(y_i, x_i), i = 1, \ldots, n\} \) to identify the underlying model as approximately a quadratic regression model, provided the model misspecification is significant enough to be detectable, as in Dorfman’s (1994) simulation study. Then we use the multiple regression estimator 
\[
\bar{y}_{mr} = \bar{y} + b_1(x - \bar{x}) + b_2(z - \bar{z}),
\]
where \( b_1 \) and \( b_2 \) are the sample regression coefficients when \( y \) is regressed on \( x, \) \( x \) and \( z \). It follows that
\[
t_{mr} = (\bar{y}_{mr} - \bar{Y})/s(\bar{y}_{mr}) = (\bar{\varepsilon}^* - \varepsilon_N^*)/s(\bar{\varepsilon}^*)
\]
so that the coverage error associated with the pivotal quantity \( t_{mr} \) will be small regardless of the skewness in \( y_i \)'s and \( x_i \)'s. Note, however, that \( \bar{y}_{mr} \) cannot be implemented if only the population mean \( \bar{X} \) is known.

Under simple random sampling, both model-based and model-assisted approaches lead to the same pivotal quantity. But for general uniphase sampling this is not necessarily true because the model-assisted approach employs generalized regression estimators depending on design weights, unlike the model-based approach; see Section 4.

3. Simulation Results: SRS

We present some simulation results on the coverage probabilities associated with \( t_r \) by generating synthetic populations that obey the linear regression model or the quadratic regression model.

3.1. Generation of populations

Positively skewed synthetic populations, each of size \( N = 500 \), were generated using the same algorithms given by Dorfman (1994), using two of the models of that paper for generating the variable of interest \( y \) given the auxiliary variable \( x \):

(i) Linear regression model given by
\[
y_i = x_i + \varepsilon_i,
\]
where the \( \varepsilon_i \) are iid \( N(0, \sigma^2 = 0.04 \) or 0.16).

(ii) Quadratic regression model
\[
y_i = 8x_i^2 + \varepsilon_i^*,
\]
where the \( \varepsilon_i^* \) are iid \( N(0, \sigma^2 = 0.04 \) or 0.16).

In both cases (i) and (ii), we generated the \( x_i \)'s from a standard lognormal distribution with first and second moments given by \( \mu_1 = \sqrt{e}/2 \) and \( \mu_2 = (e^2 - e)/4 \), where \( e \) is the exponential constant. Model 1 represents the ideal case for the simple linear regression estimator \( \bar{y}_r \), whereas Model 2 represents an unfavourable case.

3.2. Sampling of the populations

We used two different approaches to sampling from the populations: (i) Generate a population, then draw a simple random sample of specified size \( n \), and repeat the whole process 1,000 times. (ii) Draw a single population and then draw 1,000 simple random samples. We repeated process (ii) 100 times to make it more comparable to (i). Dorfman (1994) used process (i) for his simulation study except that a two-phase random sample is drawn. Note that process (ii) is the standard repeated sampling framework employed in the design-based theory. It may be of interest to compare the two processes with respect to coverage probabilities.
Table 1. Skewness and kurtosis of population residuals $E_i$

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i = x_i + \varepsilon_i$</td>
<td>Variable</td>
<td>0.023</td>
<td>-0.056</td>
</tr>
<tr>
<td></td>
<td>Fixed</td>
<td>0.022</td>
<td>-0.057</td>
</tr>
<tr>
<td>$y_i = 8x_i^2 + \varepsilon_i^*$</td>
<td>Variable</td>
<td>6.57</td>
<td>96.79</td>
</tr>
<tr>
<td></td>
<td>Fixed</td>
<td>6.44</td>
<td>96.14</td>
</tr>
</tbody>
</table>

We refer to the first process as the variable population method because a different population is generated after each sample selection. The second method is referred to as the fixed population method because a single population is generated and repeated samples are drawn from this population. The variable population method is a “hybrid” design-based approach because it also uses repeated sampling, unlike the model-based approach based on the model distribution conditional on the sample. For example, Cochran (1977, p. 12) says: “Similarly, when a single sample is taken from each of a series of different populations, about 95% of the 95% confidence statements are correct.”

3.3. Coverage probabilities

Table 1 reports the skewness and kurtosis coefficients of the population residuals $E_i = (y_i - \bar{Y}) - B(x_i - \bar{X})$ where $B$ is the population regression coefficient. Values reported in Table 1 are averages over the generated populations in the two cases. As expected, both skewness and kurtosis of $E_i$’s are close to 0 under the linear regression model, whereas both are large under the quadratic regression model.

Table 2 reports the correlation between $\bar{y}_r$ and $s(\bar{y}_r)$ calculated from the simulated samples, each of size $n = 40$. Simulated coverage probabilities of the two-sided normal confidence interval $I_{r,0.95} = [\bar{y}_r - 2s(\bar{y}_r), \bar{y}_r + 2s(\bar{y}_r)]$ corresponding to nominal level $1 - \alpha = 0.95$ are also reported.

For the populations generated by the linear regression model, Table 2 shows that $\text{corr}(\bar{y}_r, s(\bar{y}_r))$ is close to 0 and that the coverage probability is larger than or equal to 0.95. On the other hand, for the population generated by the quadratic regression model, $\text{corr}(\bar{y}_r, s(\bar{y}_r)) = 0.5$ and the coverage probability is significantly smaller than the nominal 0.95: ranging from 0.85 to 0.89. Performance of the two methods (fixed and variable) is

Table 2. Corr($\bar{y}_r, s(\bar{y}_r)$) and coverage probabilities (%) of the confidence interval [$\bar{y}_r - 2s(\bar{y}_r), \bar{y}_r + 2s(\bar{y}_r)$]; nominal level = 0.95

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>$\sigma^2$</th>
<th>Correlation</th>
<th>Coverage probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.04</td>
<td>-0.01</td>
<td>95.2</td>
</tr>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.16</td>
<td>0.05</td>
<td>98.7</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.04</td>
<td>0.55</td>
<td>85.0</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.16</td>
<td>0.50</td>
<td>87.2</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.04</td>
<td>-0.10</td>
<td>95.0</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.16</td>
<td>-0.08</td>
<td>98.5</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.04</td>
<td>0.54</td>
<td>88.1</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.16</td>
<td>0.53</td>
<td>89.3</td>
</tr>
</tbody>
</table>
similar, but the coverage probability is slightly larger in the fixed population case (quadratic model): 0.88 vs 0.85 ($\sigma^2 = 0.04$) and 0.89 vs 0.87 ($\sigma^2 = 0.16$).

4. General Uni-phase sampling

For general uni-phase sampling, the model-assisted approach uses design-weighted linear regression estimators motivated by working linear regression models (see e.g., Särndal et al. 1992). Suppose that the working model is given by $y_i = \alpha + \beta x_i + e_i$, where the $e_i$’s are iid errors with mean 0 and variance $\sigma^2$. Then, the design-weighted (or “generalized”) linear regression estimator of $\bar{Y}$ is given by $\bar{Y}_{rw} = \bar{y}_w + b_w(\bar{X} - \bar{x}_w)$, where $\bar{y}_w = \sum_s w_i y_i / \sum_s w_i$, $\bar{x}_w = \sum_s w_i x_i / \sum_s w_i$, $w_i$ is the design weight taken as the inverse of the inclusion probability $\pi_i$ attached to unit $i$, and $\sum_s$ denotes summation over units $i$ in the sample $s$. Further, $b_w$ is the design-weighted estimator of the population regression coefficient $B$: $b_w = (\bar{u}_w - \bar{y}_w \bar{x}_w) / (\bar{z}_w - \bar{x}_w^2)$, where $\bar{u}_w = \sum_s w_i u_i / \sum_s w_i$ and $\bar{z}_w = \sum_s w_i z_i / \sum_s w_i$ with $u_i = y_i x_i$ and $z_i = x_i^2$.

The generalized regression estimator $\bar{Y}_{rw}$ is design-consistent for $\bar{Y}$ as well as model-unbiased under the working model, i.e., $E_m(\bar{Y}_{rw} - \bar{Y}) = 0$, where $E_m$ denotes model expectation. Further, $t_{rw} = \frac{\bar{Y}_{rw} - \bar{Y}}{s(\bar{w})} = \frac{\bar{y}_w - \bar{y}_N}{s(\bar{w})}$ (4.1)

where $\bar{y}_w$ is the weighted mean of the sample residuals $e_i = y_i - \bar{y}_w - b_w(x_i - \bar{x}_w)$, and $s(\bar{w})$ is a design-consistent estimator of the design variance of $\bar{Y}_{rw}$.

Under repeated sampling, $\bar{y}_w - \bar{y}_N$ is asymptotically normal with mean zero and design variance $V(\bar{w})$, and $s^2(\bar{w})/V(\bar{w})$ converges in design probability to 1. As a result, the design-based coverage error associated with the pivotal quantity $t_{rw}$ will be small if the assumed model is (approximately) valid, regardless of the skewness in $y_i$’s (and $x_i$’s), provided the skewness of $y_i$’s solely depends on that of $x_i$’s.

If the true model is a quadratic regression model, then we get results similar to those in Section 2.2 with the unweighted means replaced by the weighted means and $b$ by $b_w$. It then follows that large skewness in $z_i$’s leads to a coverage probability smaller than the nominal $(1 - \alpha)$-level if the pivotal quantity $t_{rw}$ is used and the $\beta_2$-coefficient is significantly large. Performing suitable model diagnostics, we may be able to identify the underlying model as approximately a quadratic regression model, provided the model misspecification is significant enough to be detectable. We can then use a design-weighted multiple regression estimator, $\bar{Y}_{mrw} = \bar{y}_w + b_{1w}(\bar{X} - \bar{x}_w) + b_{z,w}(\bar{Z} - \bar{z}_w)$. The pivotal quantity $t_{mrw}$ based on $\bar{Y}_{mrw}$ will lead to a result similar to (2.7), so that the coverage error associated with $t_{mrw}$ will be small regardless of the skewness in $y_i$’s (and $x_i$’s).

A model-based approach, based on the linear regression model, uses the pivotal quantity $t_e$ instead of $t_{rw}$, regardless of the design weights. The estimator $\bar{Y}_r$, however, is asymptotically design-based for $\bar{Y}$. As a result, the asymptotic mean of $\bar{y} - \bar{y}_N$ is not necessarily zero under repeated sampling. This affects the coverage error unlike (4.1), although the asymptotic mean of $\bar{y} - \bar{y}_N$ is likely to be small if the model holds.

As noted by Hansen, Madow, and Tepping (1983), moderate departures from the assumed model can lead to poor coverage of model-based confidence intervals under
repeated sampling, unlike the design-weighted intervals. In their study, the working model was the ratio model \( y_i = \beta x_i + e_i \) with unequal error variances \( \sigma^2 x_i \), while the synthetic population was generated from \( y_i = 0.4 + 0.25 x_i + e_i^* \) with error variances 0.0625 \( x_i^2 \). The model-based estimator under the working model is the ratio estimator \( \bar{y}_r = (\bar{y}/\bar{x})\bar{X} \).

The numerator of \( t_r \) is given by

\[
\bar{y}_r - \bar{y} = \alpha \left( \frac{\bar{X}}{\bar{x}} - 1 \right) + \left( \frac{\bar{X}}{\bar{x}} - \bar{e}_N \right) \tag{4.2}
\]

Hansen et al. (1983) employed stratified random sampling with disproportionate sample allocation. As a result, \( \bar{x} \) is heavily design-based for \( \bar{X} \) unlike the weighted mean \( \bar{x}_w \) used in the numerator of

\[
\bar{y}_{rw} - \bar{y} = \alpha \left( \frac{\bar{X}}{\bar{x}_w} - 1 \right) + \left( \frac{\bar{X}}{\bar{x}_w} - \bar{e}_N \right) \tag{4.3}
\]

where \( \bar{y}_{rw} \) is the design-weighted ratio estimator \( (\bar{y}_w/\bar{x}_w)\bar{X} \); for a stratified sample of \( n = 200 \) units they obtained \( \bar{x} = 14.644 \) and \( \bar{x}_w = 9.935 \) compared to \( \bar{X} = 9.965 \). The heavy bias in \( \alpha \) induced poor coverage for the model-based intervals under repeated sampling, despite the small intercept term, \( \alpha = 0.4 \), for the true model. On the other hand, the design-weighted intervals performed extremely well in terms of coverage since the asymptotic mean of \( \bar{y}_{rw} - \bar{y} \) is zero and \( \alpha \) is small.

5. Two-Phase Random Sampling

In two-phase random sampling, a large simple random sample of size \( n_1 \) is first drawn and auxiliary information \( \{x_i, i \in s_1\} \) is collected. From the first-phase sample, \( s_1 \), a simple random subsample, \( s_2 \), of size \( n_2 \) is drawn and the variable of interest, \( y_i \), is observed. The second-phase data \( \{y_i, i \in s_2\} \) is more expensive to collect than the first-phase information \( \{x_i, i \in s_1\} \).

A simple linear regression estimator of \( \bar{y} \) is given by \( \bar{y}_{2r} = \bar{y}_2 + b_2(\bar{x}_1 - \bar{x}_2) \), where \( b_2 \) is the sample regression coefficient based on the second-phase sample data \( \{(y_i, x_i), i \in s_2\} \). \( (\bar{y}_2, \bar{x}_2) \) are the second-phase sample means and \( \bar{x}_1 \) is the first-phase sample mean of \( x \). A number of variance estimators of \( \bar{y}_{2r} \) have been proposed in the literature. Cochran (1977, p. 343) used the variance estimator

\[
v_{std} = (1/n_2 - 1/N)s_{2e}^2 + (1/n_1 - 1/N)b_2^2 s_{2x}^2 \tag{5.1}
\]

where \( s_{2e}^2 \) is the sample variance of the residuals \( e_i = y_i - \bar{y}_2 - b_2(x_i - \bar{x}_2), i \in s_2 \) and \( s_{2x}^2 \) is the sample variance of \( x_i \)’s for \( i \in s_2 \). Cochran also proposed a hybrid version of (5.1) based on the sample linear regression model \( y_i = \alpha + \beta x_i + e_i, i = 1, \ldots, N \), where the \( e_i \)’s are iid errors with mean 0 and variance \( \sigma^2 \). It is given by

\[
v_{hyd} = v_{std} + [(\bar{x}_1 - \bar{x}_2)^2/s_{2x}^2]s_{2e}^2 \tag{5.2}
\]

The variance estimators (5.1) and (5.2) use only the second-phase sample data. Dorfman (1994) proposed alternatives to \( v_{std} \) and \( v_{hyd} \) that make full use of the first-phase sample data by replacing \( s_{2x}^2 \) by \( s_{1x}^2 \), the sample variance of \( x_i \)’s in the first-phase sample \( s_1 \):

\[
v_{std,f} = (1/n_2 - 1/N)s_{2e}^2 + (1/n_1 - 1/N)b_2^2 s_{1x}^2 \tag{5.3}
\]
and

$$v_{hyd,f} = v_{std,f} + \left(\bar{x}_1 - \bar{x}_2\right)^2 s_{2e}^2 / s_{2e}$$  \hspace{1cm} (5.4)

Sitter (1997) obtained a variance estimator similar to $v_{hyd,f}$ using jackknife linearization.

A normal approximation interval on $\bar{y}$ using $\bar{y}_{2r}$ is based on the pivotal quantity

$$t_{2r} = \left(\bar{y}_{2r} - \bar{y}\right) / s(\bar{y}_{2r})$$  \hspace{1cm} (5.5)

where $s^2(\bar{y}_{2r})$ denotes a variance estimator of $\bar{y}_{2r}$. We now examine the accuracy of the normal approximation along the lines of Section 2.2. Suppose that the underlying model generating the population is given by the simple linear regression model.

If the assumed model holds and $s^2(\bar{y}_{2r}) = v_{std,f}$, then

$$t_{2r} = \frac{\left(\bar{x}_2 - \bar{y}\right) + \beta(\bar{x}_1 - \bar{X})}{\left[\frac{1}{n_2 - 1} / \left(\frac{1}{n_1 - 1} \beta^2 s_{1x}^2\right)\right]^{1/2}}$$  \hspace{1cm} (5.6)

where $\bar{e}_2$ and $s_{2e}$ are respectively the sample mean and the sample variance of $e_i^*$ for $i \in s_2$. It follows from (5.6) that the coverage error associated with the pivotal quantity $t_{2r}$ will be affected by the skewness in $x_i^*$ in contrast to the case of (single-phase) simple random sample (compare (5.6) to (2.6)). As a result, the numerator of (5.6) will be positively correlated with the denominator, which in turn leads to a coverage probability smaller than the nominal $(1 - \alpha)$-level if the $x_i^*$ is positively skewed. However, the skewness effect is dampened because $\bar{x}_1$ and $s_{1x}^2$ are based on the first-phase sample of size $n_1$ which is large relative to the second-phase sample of size $n_2$.

Suppose now that the underlying true model is a quadratic model $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i^*$ with $z_i = x_i^2$ and iid errors $e_i^*$. Following the argument in Section 2.2, the numerator of $t_{2r}$ is approximately equal to $-a \beta_2 (\bar{x}_2 - \bar{X}) + \beta_2 (\bar{z}_2 - \bar{Z}) + (\beta_1 + a \beta_2) (\bar{x}_1 - \bar{X})$ while the denominator is approximately equal to the formula obtained by replacing $s_{2e}^2$ by $s_{2e}^2$, $s_{1x}^2$ by $s_{1x}^2$, and $\beta_2 s_{1x}^2$ by $\beta_2 s_{1x}^2$, where $e_i^* = -a \beta_2 (x_i - \bar{X}) + \beta_2 (z_i - \bar{Z}) + (e_i^* - \bar{e}_2^*)$. Therefore, positive skewness in the $z_i$'s induces significant positive correlation between the numerator and the denominator through the positive correlation between $\bar{z}_2$ and $s_{2e}^2$. Note that the latter correlation is based on the smaller second-phase sample, unlike the case of (5.6). As a result, the effect of a quadratic model on the coverage error associated with $t_{2r}$ will be more pronounced.

The above observations suggest that a model-assisted approach is needed for reducing the coverage error in two-phase sampling. Such an approach is feasible for two-phase sampling because the first-phase sample $x$-values are all known. (Note that for single-phase sampling only $X$ may be known.) Following Section 2.2, we can first perform suitable model diagnostics (e.g., residual analysis) to identify the working model as a quadratic regression model, provided the model misspecification is significant enough to be detectable, as in Dorfman’s (1994) simulation study. Then we use the multiple linear regression estimator $\bar{y}_{2mr} = b_{22} (\bar{z}_2 - \bar{Z}) + b_{22} (\bar{z}_1 - \bar{Z})$, where $b_{12}$ and $b_{22}$ are the sample regression coefficients obtained from the second-phase sample when $y$ is regressed
on 1, x and z. It now follows that

\[ t_{2mr} = (\bar{y}_{2mr} - \bar{Y}) s_{2mr} \]

\[ = \left( \frac{1}{n_{2} - 1} - \frac{1}{N} \right) \left( s_{x_{1}}^{2} + \frac{1}{n_{1}} \left( \beta_{1} s_{x_{z}}^{2} + \beta_{2}^{2} s_{1z}^{2} + 2 \beta_{1} \beta_{2} s_{1xz} \right) \right)^{1/2} \]  

(5.7)

where \( s_{1xz} \) is the first-phase sample covariance between x and z. In (5.7) we used the analogue of \( v_{std, f} \) (Equation (5.3)) for the multiple linear regression estimator \( \bar{y}_{mr} \). Noting that the \( z_{i} \)'s will be more skewed than the \( x_{i} \)'s, the coverage error associated with (5.7) for the quadratic model will be larger than the coverage error associated with \( t_{2r} \), given by (5.6) for the linear model \( y_{i} = \alpha + \beta x_{i} + e_{i} \). However, the skewness effect is dampened because \( (\bar{x}_{1}, \bar{z}_{1}) \) and \( (s_{x_{1}^{2}}, s_{y_{1}^{2}}, s_{1xz}) \) are based on the larger first-phase sample.

For general two-phase sampling, the model-assisted approach uses design-weighted linear regression estimators motivated by working linear models. Results of Section 5 can be extended to general two-phase sampling and the conclusions will be similar to those for \( t_{2r} \) and \( t_{2mr} \) under two-phase random sampling. We omit details for simplicity.

6. Simulation Results: Two-Phase Sampling

6.1. Sampling of the populations

Section 3.1 described the generation of synthetic populations based on a linear regression model \( y_{i} = x_{i} + e_{i} \) with errors \( e_{i} \sim N(0, \sigma^2) \) and a quadratic regression model \( y_{i} = 8x_{i}^{2} + e_{i} \) with errors \( e_{i} \sim N(0, \sigma^2) \). The \( x_{i} \)'s were generated from a standard lognormal distribution.

We used (i) the variable population method and (ii) the fixed population method to draw two-phase random samples from the simulated populations, following Section 3.2. We report the results for \( n_{1} = 80 \) and \( n_{2} = 40 \).

6.2. Coverage probabilities

Table 3 reports \( \text{corr}(\bar{y}_{2r}, s(\bar{y}_{2r})) \), calculated from the simulated samples, for each of the four variance estimators (5.1)–(5.4). Table 4 presents the simulated coverage probabilities

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>( \sigma^2 )</th>
<th>Variance estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( v_{std} )</td>
<td>( v_{hyd} )</td>
</tr>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.04</td>
<td>0.56</td>
</tr>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.16</td>
<td>0.56</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.04</td>
<td>0.80</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.16</td>
<td>0.78</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.04</td>
<td>0.54</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.16</td>
<td>0.53</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.04</td>
<td>0.80</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.16</td>
<td>0.78</td>
</tr>
</tbody>
</table>
of the two-sided normal confidence interval $I_{2r,0.95} = [\bar{y}_{2r} - 2s(\bar{y}_{2r}), \bar{y}_{2r} + 2s(\bar{y}_{2r})]$ corresponding to the nominal level $1 - \alpha = 0.95$. Four coverage probabilities, corresponding to the four variance estimators, are reported for each setting.

For the populations generated by the linear regression model, Table 3 shows that the correlation between $\bar{y}_{2r}$ and $s(\bar{y}_{2r})$ is around 0.5 when we use $v_{\text{std}}$ or $v_{\text{hyd}}$, and increases to 0.75 when we use $v_{\text{std, f}}$ or $v_{\text{hyd, f}}$. This result is in contrast to simple random sampling where $\text{corr}(\bar{y}, s(\bar{y}))$ is close to zero. The representation of $t_{2r}$ given by (5.6) explains the reason for a significant $\text{corr}(\bar{y}_{2r}, s(\bar{y}_{2r}))$ when the $x_i$’s are positively skewed. For the populations generated by the quadratic regression model, the values of $\text{corr}(\bar{y}_{2r}, s(\bar{y}_{2r}))$ are significantly larger than those generated by the linear regression model. The correlation increases to about 0.8 when we use $v_{\text{std}}$ or $v_{\text{hyd}}$ and increases to about 0.85 when we use $v_{\text{std, f}}$ or $v_{\text{hyd, f}}$. As noted in Section 5, positive correlation between $\bar{y}_2$ and $s_{\bar{y}_2}$ contributes to the inflation of correlation.

As seen from Table 4, the normal intervals perform very poorly under the quadratic model, with coverage probability ranging from 0.56 to 0.67. On the other hand, the coverage probability is much better under the linear model, ranging from 0.88 to 0.94. Also, the coverage error associated with the hybrid variance estimator $v_{\text{hyd}}(v_{\text{hyd, f}})$ is smaller than the coverage error associated with the standard variance estimator $v_{\text{std}}(v_{\text{std, f}})$. Further, a slightly better coverage rate is achieved by using a full variance estimator $v_{\text{std, f}}(v_{\text{hyd, f}})$. We also observed that the difference in coverage rates between the full and the standard variance estimators becomes more pronounced (better coverage for the full estimator) as the second-phase sample size, $n_2$, decreases; supporting tables are not reported here. Our results for the variable population method closely parallel those reported by Dorfman (1994). Also, coverage performance is generally better for the fixed population method.

Results in Tables 3 and 4 are obtained by averaging over the generated populations. But the skewness of the residuals varies significantly across the generated populations. To account for this variation, we generated quartile ranges 0 to 25%, 25% to 50%, 50% to 75% and 75% to 100% based on the skewness values of the residuals. In the fixed population case, the ranges were based on the 100 skewness values, whereas in the variable population case the ranges were based on the 1000 skewness values. Tables 5 and 6 respectively provide average coverage rates within the quartile ranges for the variable and the

\footnotesize

<table>
<thead>
<tr>
<th>Method</th>
<th>Model</th>
<th>$\sigma^2$</th>
<th>$v_{\text{std}}$</th>
<th>$v_{\text{hyd}}$</th>
<th>$v_{\text{std, f}}$</th>
<th>$v_{\text{hyd, f}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.04</td>
<td>88.9</td>
<td>90.5</td>
<td>92.9</td>
<td>93.1</td>
</tr>
<tr>
<td>Variable</td>
<td>Linear</td>
<td>0.16</td>
<td>88.8</td>
<td>89.1</td>
<td>93.8</td>
<td>93.8</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.04</td>
<td>64.1</td>
<td>64.1</td>
<td>65.9</td>
<td>65.9</td>
</tr>
<tr>
<td>Variable</td>
<td>Quadratic</td>
<td>0.16</td>
<td>56.5</td>
<td>56.8</td>
<td>59.5</td>
<td>59.8</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.04</td>
<td>92.9</td>
<td>92.9</td>
<td>94.1</td>
<td>94.1</td>
</tr>
<tr>
<td>Fixed</td>
<td>Linear</td>
<td>0.16</td>
<td>87.7</td>
<td>88.1</td>
<td>92.4</td>
<td>92.5</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.04</td>
<td>69.7</td>
<td>69.9</td>
<td>66.7</td>
<td>66.7</td>
</tr>
<tr>
<td>Fixed</td>
<td>Quadratic</td>
<td>0.16</td>
<td>60.2</td>
<td>60.4</td>
<td>61.1</td>
<td>61.3</td>
</tr>
</tbody>
</table>
Table 5. Coverage probabilities (%) of the variable population method based on quartile ranges: nominal level = 0.95

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma^2$</th>
<th>Quartile range (%)</th>
<th>Variance estimator</th>
<th>$\psi_{std}$</th>
<th>$\psi_{hyd}$</th>
<th>$\psi_{std,f}$</th>
<th>$\psi_{hyd,f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.04</td>
<td>0–25</td>
<td>90.8</td>
<td>90.8</td>
<td>94.2</td>
<td>94.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>89.7</td>
<td>89.8</td>
<td>93.8</td>
<td>93.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>88.9</td>
<td>88.9</td>
<td>92.4</td>
<td>92.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>87.8</td>
<td>87.8</td>
<td>90.8</td>
<td>90.9</td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>0.16</td>
<td>0–25</td>
<td>90.1</td>
<td>90.1</td>
<td>93.8</td>
<td>93.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>88.6</td>
<td>88.7</td>
<td>93.4</td>
<td>93.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>88.4</td>
<td>88.4</td>
<td>93.0</td>
<td>93.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>87.4</td>
<td>87.5</td>
<td>92.2</td>
<td>92.4</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.04</td>
<td>0–25</td>
<td>91.9</td>
<td>91.9</td>
<td>92.1</td>
<td>92.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>88.7</td>
<td>88.7</td>
<td>88.9</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>79.8</td>
<td>79.8</td>
<td>80.3</td>
<td>80.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>59.9</td>
<td>60.0</td>
<td>61.1</td>
<td>61.7</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.16</td>
<td>0–25</td>
<td>91.1</td>
<td>91.2</td>
<td>91.3</td>
<td>91.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>87.9</td>
<td>87.9</td>
<td>88.8</td>
<td>88.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>79.1</td>
<td>79.1</td>
<td>79.7</td>
<td>79.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>58.8</td>
<td>58.9</td>
<td>61.8</td>
<td>61.9</td>
<td></td>
</tr>
</tbody>
</table>

fixed population cases. For the linear model case, the coverage probability decreases slowly as the range increases from 0–25% to 75–100%: 0.94 to 0.91 for $\psi_{std,f}$. On the other hand, for the quadratic model case the coverage probability decreases rapidly: 0.92 to 0.61 for $\psi_{std,f}$. The above results suggest that the skewness size of the residuals $E_i$ has substantial impact on the coverage performance of normal intervals.

We now turn to the performance of the normal interval associated with the

Table 6. Coverage probabilities (%) of the fixed population method based on quartile ranges: nominal level = 0.95

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma^2$</th>
<th>Quartile range (%)</th>
<th>Variance estimator</th>
<th>$\psi_{std}$</th>
<th>$\psi_{hyd}$</th>
<th>$\psi_{std,f}$</th>
<th>$\psi_{hyd,f}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>0.04</td>
<td>0–25</td>
<td>93.6</td>
<td>93.7</td>
<td>94.4</td>
<td>94.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>91.9</td>
<td>91.9</td>
<td>93.7</td>
<td>93.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>89.5</td>
<td>89.5</td>
<td>92.6</td>
<td>92.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>88.3</td>
<td>88.4</td>
<td>91.2</td>
<td>91.6</td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>0.16</td>
<td>0–25</td>
<td>91.6</td>
<td>91.7</td>
<td>94.1</td>
<td>94.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>91.5</td>
<td>91.5</td>
<td>93.4</td>
<td>93.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>89.2</td>
<td>89.2</td>
<td>93.4</td>
<td>93.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>88.0</td>
<td>88.1</td>
<td>91.2</td>
<td>91.2</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.04</td>
<td>0–25</td>
<td>92.2</td>
<td>92.2</td>
<td>92.5</td>
<td>92.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>88.5</td>
<td>88.5</td>
<td>88.8</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>79.6</td>
<td>79.6</td>
<td>80.5</td>
<td>80.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>59.9</td>
<td>59.9</td>
<td>60.8</td>
<td>61.5</td>
<td></td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.16</td>
<td>0–25</td>
<td>90.9</td>
<td>90.9</td>
<td>91.5</td>
<td>91.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>25–50</td>
<td>87.8</td>
<td>87.8</td>
<td>88.8</td>
<td>88.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>50–75</td>
<td>79.1</td>
<td>79.1</td>
<td>79.9</td>
<td>79.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>75–100</td>
<td>59.1</td>
<td>59.1</td>
<td>61.6</td>
<td>61.9</td>
<td></td>
</tr>
</tbody>
</table>
model-assisted estimator \( \bar{z}_{2m,r} \). We found that the skewness values of the population residuals for the quadratic model are 0.072 and 0.076 respectively for the fixed and the variable population cases, as compared to 6.44 and 6.57 for the linear residuals reported in Table 1. As a result, the confidence interval associated with \( t_{2m} \) leads to much better coverage relative to the interval associated with \( t_{2r} \) for the quadratic model, as seen from Table 7: 0.91 for \( t_{2m} \) compared to 0.63 for \( t_{2r} \).

As noted before, a model-assisted approach is implementable for two-phase sampling because all the first-phase \( x \)-values are known. Gross violations of the underlying model associated with the simple linear regression estimator should be accounted for through a model-assisted approach. Then the design-based intervals associated with the model-assisted estimator will be inferentially satisfactory despite minor violations of the working model, especially as the sample size increases.

### 7. Summary Results

Our study highlights the fact that a large skewness in the linear residuals, \( E_j \), affects the design-based performance of normal approximation confidence intervals associated with the simple linear regression estimator in two-phase sampling. If the true underlying model that generated the population deviated significantly from the linear model, then the coverage performance of the intervals can be poor even for moderately large second-phase samples. A proper model-assisted approach can lead to residuals with small skewness and hence better coverage rates. We also observed that the traditional fixed population approach yields consistently better coverage rates than the variable population approach, although both approaches are asymptotically correct in the design-based framework.

For single-phase sampling, a model-assisted approach cannot be implemented if only the population mean \( \bar{X} \) is known, say, from external sources, because the regression estimator under the quadratic model depends on the population total of \( x_{i}^2 \)-values. It may be possible to use some other auxiliary variable (say from a recent census) related to \( x \) to construct a take all stratum and a take some stratum. Such a stratification of the population will reduce the skewness of the \( x \)'s (and hence of the residuals \( E_j \)) in the take some stratum under model failure and lead to better coverage performance of normal approximation intervals when only \( \bar{X} \) is known.

### 8. References


Received May 2001
Revised December 2002