

CONFIDENCE INTERVAL COVERAGE PROPERTIES FOR REGRESSION ESTIMATORS IN TWO-PHASE SAMPLING

J.N.K.Rao, W. Jocelyn, and M.A. Hidioglou¹

ABSTRACT

Confidence interval coverage associated with the simple regression estimator under two-phase sampling has been empirically investigated via a simulation by Dorfman (1994) to study the following contention of the design-based theory: "Methods that happen to incorporate a model are inferentially satisfactory despite failure of the model". He considered a number of variance estimators to obtain normal-theory based confidence intervals. They included Cochran's (1977) original variance estimators (exact and approximate), and a model-based modification. Dorfman concluded from his simulation that the contention of the design-based theory is not tenable. We replicated Dorfman's study to investigate the coverage properties in detail. Our conclusions differ from those arrived at by Dorfman because we recognize that skewness plays an important role in terms of confidence interval coverage under the design-based framework.

KEY WORDS: Finite population, Sample Size, Simulation, Skewness, Stratification.

RÉSUMÉ

La couverture de l'intervalle de confiance associée à l'estimateur de régression simple étant donné un plan de sondage à deux phases a été empiriquement étudiée par Dorfman(1994). Il a investigué l'assertion suivante de la théorie classique d'échantillonnage basée sur le plan: « Les méthodes incorporant un modèle restent quand même satisfaisantes du point de vue de l'inférence même si le modèle est non-valide ». Il a considéré plusieurs estimateurs de variance afin d'obtenir des intervalles de confiance basés sur la théorie de la loi normale. Parmi les estimateurs de variance considérés on retrouve entre autre les estimateurs de variance (exact et approximatif) suggérés par Cochran(1977) ainsi qu'une modification de ces estimateurs basés sur un argument modéliste. Dorfman a conclu à partir des résultats de la simulation qu'il a effectuée, que l'assertion mentionnée auparavant était indéfendable. Nous avons repris l'étude de Dorfman et avons étudié un peu plus en détail les propriétés de couverture. Nos conclusions diffèrent de celles de Dorfman parce que nous avons mis en lumière le rôle important joué par l'asymétrie lorsqu'on calcule la couverture des intervalles de confiance selon le plan.

MOTS-CLÉS : Population finie, taille de l'échantillon, simulation, asymétrie, stratification.

1. INTRODUCTION

Dorfman (1994) compared several variance estimators of the simple linear regression estimator of a mean for a two-phase design via simulation. His study indicated that when the regression model is not well specified, the resulting normal theory confidence intervals for the mean do not have good coverage properties. Dorfman therefore concluded that Hansen and Tepping (1990)'s contention "Methods that happen to

incorporate a model are inferentially satisfactory despite failure of the model" is incorrect. In this paper, we repeat Dorfman's simulations and provide reasons for the poor performance of design-based methods. We also propose simple solutions to improve the coverage.

The paper is organized as follows. We give a general overview of the problem along with some notation in Section 2. we describe the simulation in Section 3.

¹ J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada; W. Jocelyn, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada; M.A. Hidioglou, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada.

Results of this simulation are found in Section 4. A model-assisted approach is proposed in Section 5.

2. VARIANCE ESTIMATORS FOR TWO-PHASE SAMPLING

In two-phase sampling, a first-phase large sample (say s_1) of size n_1 is first drawn from the universe U . This first-phase sample provides auxiliary data (x) that is not expensive to collect. From the first-phase sample s_1 , a second-phase sample s_2 of size n_2 is drawn, and the variable of interest y is observed from this sample. The second-phase data $\{(y_i, x_i), i \in s_2\}$ is typically more expensive to collect than the first-phase information $\{x_i, i \in s_1\}$. We assume that the sampling at both phases uses simple random sampling without replacement.

The estimator of interest is the simple *regression estimator* of the population mean \bar{Y} :

$$\bar{y}_{tr} = \bar{y}_2 + b_2(\bar{x}_1 - \bar{x}_2), \quad (2.1)$$

where \bar{y}_2 and \bar{x}_2 are the means for s_2 , \bar{x}_1 is the mean for s_1 and $b_2 = \hat{S}_{2xy} / \hat{S}_{2x}^2$ is the least squares regression coefficient of for x_i based on s_2 , with $\hat{S}_{2xy} = \frac{\sum_{s_2} (x_i - \bar{x}_2)(y_i - \bar{y}_2)}{(n_2 - 1)}$ and $\hat{S}_{2x}^2 = \frac{\sum_{s_2} (x_i - \bar{x}_2)^2}{(n_2 - 1)}$ denoting the estimators of the population covariance S_{xy} and the population variance S_x^2 . A number of estimators have been proposed to estimate the variance of the regression estimator \bar{y}_{tr} . Cochran (1977) used

$$\begin{aligned} v_{std}^{(1)} &= \left(\frac{1}{n_2} - \frac{1}{n_1} \right) \hat{S}_{2e}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{2y}^2 \\ &= \left(\frac{1}{n_2} - \frac{1}{N} \right) \hat{S}_{2e}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) b_2^2 \hat{S}_{2x}^2 \end{aligned} \quad (2.2)$$

where \hat{S}_{2e}^2 and \hat{S}_{2y}^2 are respectively the sample variances of the least squared residuals $e_i = (y_i - \bar{y}) - b_2(x_i - \bar{x}_2)$ and y_i computed from s_2 . We refer to this form of the two-phase variance

estimator as the *standard* estimator. This linearization variance estimator of \bar{y}_{tr} is design-consistent.

Cochran also proposed a hybrid version of the variance estimator (2.2), assuming super population model of the form $y_i = \alpha + \beta x_i + \varepsilon_i$, where for given x 's, the ε_i 's are i.i.d. $(0, \sigma_\varepsilon^2)$, and α, β , and σ_ε^2 are the superpopulation parameters. Under this model

$$\begin{aligned} E\{(\bar{y}_{tr} - \bar{Y})^2 | x\} &= \left(\frac{1}{n_2} - \frac{1}{N} \right) \sigma_\varepsilon^2 \\ &+ \beta^2 (\bar{x}_1 - \bar{X})^2 + \frac{\sigma_\varepsilon^2 (\bar{x}_1 - \bar{x}_2)^2}{(n_2 - 1) \hat{S}_{2x}^2} \end{aligned} \quad (2.3)$$

Noting that $E(\bar{x}_1 - \bar{X})^2 = \left(\frac{1}{n_1} - \frac{1}{N} \right) S_x^2$, under repeated sampling of s_1 , $(\bar{x}_1 - \bar{X})^2$ can be estimated by $\left(\frac{1}{n_1} - \frac{1}{N} \right) \hat{S}_{2x}^2$, σ_ε^2 by \hat{S}_{2e}^2 , and β^2 by \hat{b}_2^2 . Assuming that the second phase sample size is small and that the terms in n_2^{-1} are not negligible relative to 1, Cochran's alternative (1977) variance estimator (*hybrid*) is given by

$$v_{hyb}^{(1)} = v_{std}^{(1)} + \frac{\hat{S}_{2e}^2 (\bar{x}_1 - \bar{x}_2)^2}{(n_2 - 1) \hat{S}_{2x}^2} \quad (2.4)$$

The previous variance estimators are computed strictly using the second-phase sample data. Dorfman (1994) considered estimating some of the components using first-phase data. In particular, noting that $\hat{S}_{2e}^2 = S_{2y}^2 - b_2^2 \hat{S}_{2x}^2$, we can use \hat{S}_{1x}^2 instead of \hat{S}_{2x}^2 , where \hat{S}_{1x}^2 is the estimated variance of the x -variable from the first-phase sample s_1 . The resulting variance estimator referred to as the *full* estimator, is given by

$$\begin{aligned} v_{full}^{(2)} &= \left(\frac{1}{n_2} - \frac{1}{n_1} \right) \tilde{S}_{2e}^2 + \left(\frac{1}{n_2} - \frac{1}{N} \right) \hat{S}_{2y}^2 \\ &= \left(\frac{1}{n_2} - \frac{1}{N} \right) \hat{S}_{2e}^2 + \left(\frac{1}{n_1} - \frac{1}{N} \right) b_2^2 \hat{S}_{1x}^2 \end{aligned} \quad (2.5)$$

where $\tilde{S}_{2e}^2 = S_{2y}^2 - b_2^2 \hat{S}_{2x}^2$. The rationale for introducing (2.5) is that it should be more precise than

(2.4) since it uses all the x - information. The hybrid version of the full estimator, *full-hybrid*, is given by:

$$v_{hybfull}^{(2)} = v_{full}^{(2)} + \frac{\hat{S}_{2e}^2 (\bar{x}_1 - \bar{x}_2)^2}{(n_2 - 1)\hat{S}_{2x}^2}. \quad (2.6)$$

Sitter (1997) obtained an estimator similar to the "full-hybrid" estimator, via the linearized jack-knife approach. Normal theory $(1 - \alpha)$ level confidence intervals on \bar{Y} are given by $(\bar{y}_{lr} - z_{\alpha/2}v^{1/2}, (\bar{y}_{lr} + z_{\alpha/2}v^{1/2}))$, where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the $N(0,1)$ variable and v denotes a variance estimator.

3. DESCRIPTION OF THE SIMULATION

3.1 Generation of the Populations

The populations were generated using the same algorithms given in Dorfman (1994). We chose two of the models that Dorfman suggested for generating the variable of interest y given the auxiliary variable x :

- (i) A linear model given by $y_i = x_i + \varepsilon_i$, where $E(\varepsilon_i | x_i) = 0$, $E(\varepsilon_i^2 | x_i) = \sigma^2 = 0.04$ or 0.16 , $E(\varepsilon_i \varepsilon_j | x_i, x_j) = 0$, $i \neq j$, and the errors ε_i are normally distributed.
- (ii) A quadratic model given by $y_i = \beta x_i^2 + \varepsilon_i$ with $\beta = 4$ and 8 , the errors distributed as in the first model.

Model 1 represents the ideal case for the linear regression estimator, whereas model 2 represents the unfavourable case. The x_i 's were generated from a lognormal distribution with first and second moments respectively given by $\mu_1 = \sqrt{e}/2$ and $\mu_2 = \frac{(e^2 - e)}{4}$

where e is the exponential constant.

3.2 Sampling from the Generated Populations

There are several possibilities for selecting the population, the first and second-phase samples. Two such possibilities are to: (i) Draw a population, and then draw the first and second phase samples, and

repeat the whole process 1000 times; (ii) draw a single population and then draw 1000 first and second-phase samples. We repeated the process (ii) one hundred times to make it more compatible with (i). Dorfmann (1994) used the process (i) for his simulation study. Note that process (ii) is the standard repeated sampling standard for the design-based theory. It is of interest to compare the two processes to see if they lead to different conclusions in terms of confidence interval coverage.

We refer to the first procedure as the *variable-population method* because the population is redrawn after each two-phase sample selection. The second method is referred to as the *fixed-population method* because a single population is drawn, and all subsequent draws of the first and second-phase samples are based on this population.

Normal theory confidence intervals associated with each variance estimate were computed for each realisation of two-phase sampling. The resulting confidence interval coverage for nominal level of 95% are given in Tables 2 and 4, for the *fixed-population method*, while those for the *variable-population method* are given in Tables 3 and 5 respectively.

4. SIMULATION RESULTS

4.1 Confidence Internal Coverage

Skewness values of y_i and the least-squares population residuals $E_i = (y_i - \bar{Y}) - (S_{xy}/S_x^2)(x_i - \bar{X})$, were calculated using a measure of skewness given in Cochran (1977, P. 42). The results reported in Table 1 are the averages over the selected populations.

It may be noted that the confidence interval coverage associated with \bar{y}_{lr} depends on the skewness of the residuals E_i , and not on the skewness of y_i . It is therefore evident from Table 1 that the normal theory confidence intervals should perform well for the linear model $y_i = x_i + \varepsilon_i$, even though the skewness of y_i is quite high. On the other hand, if the true model is the quadratic model $y_i = 8x_i^2 + \varepsilon_i$, then the performance of confidence intervals is likely to be poor in terms of coverage because the skewness of the residuals E_i is very high (>6).

Table 1: Skewness of residuals for the linear and the quadratic models

Model	Population	Skewness of y_i	Skewness of E_i
$y_i = x_i + \varepsilon_i$	Fixed	3.290	0.022
$y_i = x_i + \varepsilon_i$	Variable	3.690	0.023
$y_i = 8 x_i^2 + \varepsilon_i$	Fixed	10.98	6.44
$y_i = 8 x_i^2 + \varepsilon_i$	Variable	10.95	6.57

Tables 2 through 5 summarise the results of the simulation using quartiles. These quartiles were constructed using the skewness E_i of values for the generated populations. The idea is to assess the effect of skewness on the performance of confidence intervals. We generated samples of different sizes during the simulation, but only report on results for $n_1 = 20$ and $n_2 = 40$, for brevity.

We next discuss in detail the results of the simulation. Table 2 shows the realised coverage obtained for a 95% nominal coverage for the linear model. These results closely parallel those obtained by Dorfman (1994). It should be noted that the difference between the full estimators and the non-full estimators in the confidence interval coverage seems to be significant. The full estimators have better coverage. Although we do not provide the supporting tables in this paper, we

noticed that the difference in coverage between the full and non-full estimators is more significant when the sample size is small. We also noticed that the coverage slowly decreases for all the estimators from the first to the third quartile. All in all, the methods perform quite well as the second-phase sample, n_2 increases.

Table 3 reports results for the variable population case. The conclusions obtained from Table 3 are similar to those given in Table 2. However, the coverage seems to be consistently lower when compared to the *fixed-population* case, regardless of the variance estimator considered. Also, the difference in coverage between a full estimator and a non-full estimator is slightly more pronounced than it was for the *fixed-population* case.

Table 2 (Fixed Population): Actual coverage of nominal 95% confidence intervals associated with variance of four estimators: population y_i generated from the model : $y_i = x_i + \varepsilon_i$

σ^2	Percentile	Standard	Hybrid	Full	Full hybrid
.04	25	92.7	92.8	94.2	94.2
.04	50	91.8	91.8	93.4	93.4
.04	75	89.1	89.1	92.3	92.4
.16	25	91.1	91.1	93.6	93.6
.16	50	90.9	90.9	93.2	93.2
.16	75	88.9	88.9	92.5	92.5

Table 3 (Variable Population) :Actual coverage of nominal 95% confidence intervals associated with variance of four estimators: population y_i generated from the model : $y_i = x_i + \varepsilon_i$

σ^2	Percentile	Standard	Hybrid	Full	Full hybrid
.04	25	89.7	89.7	93.4	93.4
.04	50	88.9	88.9	93.2	93.2
.04	75	88.0	88.0	91.9	91.9
.16	25	89.5	89.5	93.2	93.2
.16	50	87.8	87.8	92.9	92.9
.16	75	87.5	87.5	92.3	92.3

Table 4 (Fixed Population): Actual coverage of nominal 95% confidence intervals associated with four variance estimators: population y_i generated from the model $y_i = 8x_i^2 + \varepsilon_i$

σ^2	Percentile	Standard	Hybrid	Full	Full hybrid
.04	25	91.1	91.1	91.4	91.4
.04	50	88.1	88.1	88.5	88.5
.04	75	79.0	79.0	80.0	80.0
.16	25	90.2	90.2	90.7	90.7
.16	50	87.4	87.4	88.2	88.2
.16	75	78.5	78.5	79.2	79.2

Table 5 (Variable Population): Actual coverage of nominal 95% confidence intervals associated with four variance estimators: population y_i generated from the model $y_i = 8x_i^2 + \varepsilon_i$

σ^2	Percentile	Standard	Hybrid	Full	Full hybrid
.04	25	88.0	88.0	88.1	88.1
.04	50	84.1	84.1	84.6	84.6
.04	75	77.2	77.2	77.9	77.9
.16	25	87.5	87.5	87.9	87.9
.16	50	83.8	83.8	83.9	83.9
.16	75	75.6	75.6	77.2	77.2

Table 4 reports the results for the quadratic model $y_i = 8x_i^2 + \varepsilon_i$ and the fixed population case. The results in Table 4 are comparable to those provided in Table 3 of Dorfman (1994). We note that the coverage is rather poor starting from the third quartile. However, as the sample size increases, there is significant improvement in the coverage. For instance, the coverage for the full estimators for the fourth quartile increased from 60% for a $n_2 = 20$ and $n_1 = 40$ to 76% for a $n_2 = 80$ and $n_1 = 160$. Once more, the full estimators appear to have better coverage properties than the non-full ones. For the linear model, the coverage slowly decreases from the first to the third quartile. This is not the case for the quadratic model. The coverage decreases rapidly from the first to the third quartile under the quadratic model.

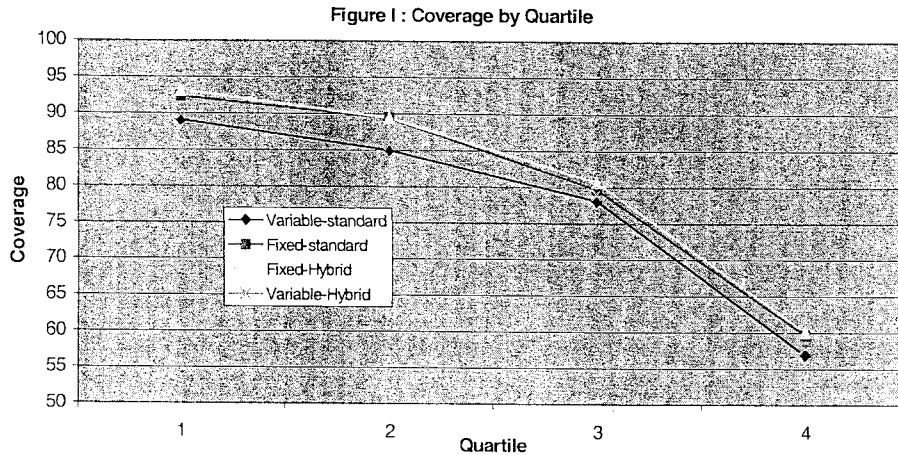
Table 5 is the *variable-population* version of Table 4. Again we observe that the coverage values are consistently lower than those obtained for the *fixed-population* case. For instance, for the third quartile, the variable-population case yields coverage ranging from 75% to 77% while the corresponding fixed case coverage are all above 78%.

4.2 Impact of Skewness

Next, we looked at the impact of skewness for the two

procedures for generating the population given that a quadratic model was used. We drew 4000 populations and 4000 samples of size $n_1 = 80$ and $n_2 = 40$ using Dorfman's scheme. For each of the populations we computed the skewness of x_i . We then divided the populations and samples by quartile based on the skewness of x_i . Note that the skewness of residuals E_i increases with the skewness of x_i under the quadratic model.

Figure 1 shows the coverage associated with the standard and the full-hybrid variance estimators by quartile. We have not shown the results for the other two variance estimators because the coverage for the hybrid variance estimator is similar to the standard variance estimator and those for the full variance estimator are similar to those for the full-hybrid variance estimator. Note that the coverage is over 90% for all the variance estimators for the first quartile and slowly decreases to 60% for the fourth quartile. This implies that if we are "unlucky" and that the 1000 selected populations belong predominantly to the fourth quartile, the coverage rates will not be close to the nominal 95% coverage. On the other hand, if the skewness of the generated x -populations is similar to those close to the first quartile, the coverage will be close to the 95% nominal coverage.



4.3 Stratification

Although the coverage improved as the sample size n_2 gets larger, it still falls short of the 95% nominal coverage for the quadratic model. This implies that the coverage rates would have been better, had we properly designed our two-phase sample. Recall from Table 1 that the residuals E_i , when the true model is quadratic, are highly skewed. Cochran (1977, P. 42) and Dalen (1986) recommend simple rules to determine the sample size for such skewed populations. Applying their rules we found that we should take a full census of the population considered or use at the very least, a much larger sample size.

The skewness of the residuals E_i can be reduced by stratifying the population. Note that under the quadratic model, the skewness of E_i increases with that of x_i . Table 6 shows results for a very simple stratification scheme that decreases the skewness significantly. We divided each finite population into two distinct strata: a take-all (certainty) and a take-some (non-certainty) stratum. The take-all stratum was made up of units with large x_i selected with certainty, and we label the sample size in that stratum as n_{2a} in Table 6. The remaining sample was chosen, using

simple random sampling without replacement, from the take-some stratum. Table 6 shows that the coverage has improved dramatically in comparison to the corresponding one (the third quartile) given in Table 5.

The take-all and the take some boundary used the x -values. However, in the two-phase sampling context, we do not know the population x -values. Therefore, it is necessary to use some other variable z (say from a census) related to x to construct the strata.

5. MODEL-ASSISTED APPROACH

When the least squares residuals have a large skewness, as in the case of the quadratic model, the confidence intervals associated with the two-phase regression estimator \bar{y}_{lr} can perform poorly in terms of coverage. By inspecting the scatter plot of $(y_i, x_i; i \in s_2)$ and the residual plot of $(e_i, x_i; i \in s_2)$, it is possible to construct a model-assisted estimator based on a non-linear model. This estimator will remain design consistent and will lead to residuals with significantly lower skewness, and hence better coverage of the confidence intervals.

Table 6 (Stratified Population): Actual coverage of nominal 95% confidence intervals associated with variance of four estimators: population y_i generated from the model $y_i = \beta x_i^2 + \varepsilon_i$

β	σ^2	$n_2 (n_{2a})$	Standard	Hybrid	Full	Full-hybrid
4	.04	20 (10)	83.5	83.5	84.1	84.1
4	.04	40 (20)	86.9	86.8	88.1	88.2
8	.16	20 (10)	83.2	83.2	83.9	83.9
8	.16	40 (20)	86.7	86.7	87.9	87.9

6. CONCLUDING REMARKS

Our study highlights the fact that the skewness in the least squares residuals E_i affects the coverage of the normal theory confidence intervals associated with the linear regression estimator. If the true underlying model that generated the population deviated significantly from the linear model, then the coverage can be poor even in moderate size samples, although asymptotically correct in the design-based framework. A model-assisted estimator based on inspecting the scatter plots of $(y_i, x_i; i \in s_2)$ and $(e_i, x_i; i \in s_2)$ can lead to residuals with significantly lower skewness and hence better confidence intervals coverage.

We also observed that the traditional *fixed-population* approach yields consistently better coverage than the *variable-population* approach.

REFERENCES

Cochran W.G. (1977). *Sampling Techniques* (3rd ed.), New York.: Wiley.

Dorfman, A.H. (1994). A note on variance estimation for the regression estimator in double sampling. *Journal of the American Statistical Association*, 89, 137-140.

Dalen, J (1986). Sampling from finite populations: actual coverage probabilities for confidence intervals on the population mean. *Journal of Official Statistics*, 2, 13-24.

Hansen, M.H., and Tepping, B.J. (1990) Regression estimates in federal welfare quality control programs. *Journal of the American Statistical Association*, 85, 856-864.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787