VARIANCE ESTIMATION OF LINEAR REGRESSION COEFFICIENTS IN COMPLEX SAMPLING SITUATION.

Dominique Caron, Wellesley Hospital and Paul N. Corey, University of Toronto
Dominique Caron, 160 Wellesley St E. Room 650 T.W. Toronto, Ontario M4Y 1J3, CANADA

KEY WORDS : Linear Regression, Complex Sampling Taylorization, Bootstrap, Simulation

INTRODUCTION

Surveys are now one of the main tools used by governmental agencies and private pollsters to collect information on a wide range of topics. For example, each month Statistics Canada turns to the population to evaluate the unemployment rate. In 1977-78, the Canada Health Survey (CHS) was done by Health and Welfare to study the health status of the Canadian population. (Health and Welfare Canada [1981]).

Survey data must be handled with care, and an awareness of exactly how the data were collected is crucial before a proper analysis can be conducted.

Most of the national and other large surveys use a complex sample design. This is mainly due to the large costs involved in using simpler sampling methods such as simple random sampling. Thus, a reasonable sampling design is often found to be a combination of stratification, clustering and simple random sampling.

Different methods of variance estimation have been developed for complex survey designs. The purpose of this study is to compare some of the techniques suggested in the literature for the estimation of the variance of the regression coefficients in a multiple linear regression model.

METHODS

The multiple linear regression equation may be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{Y}$ is the n x 1 matrix of observations (dependent variable);
$\mathbf{X}$ is the n x p matrix of independent variables;
$\boldsymbol{\beta}$ is the p x 1 vector of regression parameters;
$\boldsymbol{\epsilon}$ is the n x 1 vector of error.

Along with the equation, the following are defined:
$\mathbf{W}$ is the n x n diagonal matrix of weights;
and $\mathrm{Var}(\boldsymbol{\epsilon}) = \mathbf{V}\sigma^2$.

In the case of unweighted regression, the weight matrix and the variance matrix $\mathbf{V}$ are simply the identity matrix $\mathbf{I}$. In the case of simple regression, p = 2 and the first variable of the $\mathbf{X}$ matrix is a column of "1's" for the intercept.

Five different methods of estimating the variance of the estimator of $\boldsymbol{\beta}$ will be investigated in this study.

The first formula considered is **classical weighted least squares estimation (WLS 1)**. This choice was made because of the popularity of the method in studies in which a more complex sampling scheme than simple random sampling was used.

The WLS 1 estimate of the vector of coefficients is found in Draper and Smith [1981] to be

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{Y}$$

and its variance is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\sigma^2.$$

This estimator of the variance is exact only in the situation where

$$\mathrm{Var}(\mathbf{Y}) = \mathbf{W}^{-1}\sigma^2.$$

The second formula is a **modification to weighted least squares** found in Nathan [1981] **(WLS 2)**. In this case, the assumptions are that

i) $\mathrm{Var}(\mathbf{Y}) = \mathbf{I}\sigma^2$
and ii) the diagonal matrix of sampling probabilities $\boldsymbol{\pi}$ is available. $\mathbf{W}$, defined as $\mathbf{W} = \boldsymbol{\pi}^{-1}$, is the matrix of weights.

Nathan proposed the same estimator of $\boldsymbol{\beta}$ as above

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{Y}$$

with variance given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{W}\mathbf{W}\mathbf{X}\,(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\,\sigma^2$$

It is important to note here that s^2, the estimate of σ^2, is the estimate of the variance of $\mathbf{Y}$, and not of the weighted $\mathbf{Y}$. So s^2 is not the mean square error term from weighted regression, but the mean sum of squares of the error of the unweighted regression of $\mathbf{Y}$ on $\mathbf{X}$.

We modified the above estimator **(modification to Nathan's formula WLS 3)** and its variance for the situation in which

i) $\mathrm{Var}(\mathbf{Y}) = \mathbf{V}\sigma^2$
ii) the diagonal matrix of the sampling probabilities $\boldsymbol{\pi}$ is

available (i.e. $W = \pi^{-1}$ is the weights matrix).

Then,

$$\hat{\beta} = (X'WV^{-1}X)^{-1} X'WV^{-1}Y$$

and

$$Var(\hat{\beta}) = (X'WV^{-1}X)^{-1}X'WV^{-1}WX(X'WV^{-1}X)^{-1}\sigma^2.$$

In this situation, the s^2 estimate of σ^2 is the mean square error term from the regression in which the W matrix does not appear. That is, the s^2 estimate one gets from the regression of Y on X where $U = V^{-1}$ is the only weight matrix considered.

Another approach to the calculation of the variance involves expanding the function of the regression vector using Taylor's theorem **(Taylor's expansion TEX)**. For that purpose, it is found in Binder [1983] that one can write

$$T(\hat{\beta}) = (X'WX)\hat{\beta} - X'WY = 0.$$

Its Taylor approximation around $\hat{\beta} = \beta$ is

$$T(\hat{\beta}) = T(\beta) + \frac{T(\beta)}{\partial \beta} (\hat{\beta} - \beta) = 0$$

Isolating $\hat{\beta}$ and taking the variance conditional on X_i, one finds that

$$Var(\hat{\beta}|X_i) = (X'WX)^{-1}Var(T(\beta)|X_i)(X'WX)^{-1}.$$

In this approach, the problem is how to estimate $Var(T(\beta)|X_i)$.

$$T(\beta) = (X'WX)\beta - X'WY$$

$$= -(X'W)\epsilon$$

and so $Var(T(\beta)|X_i)$ can be estimated as the variance-covariance matrix found among the $x_i w_i e_i$'s. Variance matrices can be calculated following the survey design i.e. in each cluster or strata, $Var(T(\beta)|X_i)$ is estimates from the $x_i w_i e_i$'s. The combination of these different variance estimates is the overall estimate of $Var(T(\beta)|X_i)$ used in the formula of $Var(\hat{\beta}|X_i)$.

The final method studied is the bootstrap (Efron [1982]). This consists of sampling, with replacement, B "resamples" from the sample under investigation. All of the resamples are of the same size as the original sample. An estimate of β is then obtained from each of the B resamples. Finally, the estimator of the variance of β is evaluated using

$$Var(\hat{\beta}) = \Sigma \frac{(\hat{\beta}_i^* - \beta^*)^2}{B - 1}$$

where $\hat{\beta}_i^*$ is the estimator of β obtained in the i^{th} resample

and β^* is the mean of the B $\hat{\beta}_i^*$'s.

It is possible to incorporate the features of the sampling design used in this resampling process. This would involve sampling n_i elements from the n_i original observations within cluster i etc. In this manner, not only can the probabilities of selection be preserved, but the different cluster sizes can be kept as well.

The bootstrap technique was examined using two approaches to resampling. The first one **(bootstrap no design BND)** only considered the sampling weights. The second **(bootstrap with design BD)** reproduces the cluster structure of the sample in each resample as well as the selection probabilities. In both situations, B was taken to be equal to 200.

In summary, the five methods that are to be investigated are given in Table 1.

TABLE 1

SUMMARY OF FORMULAE

	HYPOTHESES	ESTIMATOR OF β	ESTIMATOR OF VAR($\hat{\beta}$)
WLS 1	$W = \pi^{-1}$ $Var(Y) = V\sigma^2$	$(X'WX)^{-1} X'WY$	$(X'WX)^{-1} s^2$
WLS 2	$W = \pi^{-1}$ $Var(Y) = I\sigma^2$	$(X'WX)^{-1} X'WY$	$(X'WX)^{-1} X'WWX (X'WX)^{-1} s^2$
WLS 3	$W = \pi^{-1}$ $Var(Y) = V\sigma^2$	$(X'WV^{-1}X)^{-1} X'WV^{-1}Y$	$(X'WV^{-1}X)^{-1} X'WV^{-1}WX (X'WV^{-1}X)^{-1} s^2$
TEX	$T(\hat{\beta}) = (X'WX)\hat{\beta} - X'WY = 0$ around $\beta = \beta$	$(X'WX)^{-1} X'WY$	$(X'WX)^{-1} G (X'WX)^{-1}$
BND	Sampling does not follow sample design	$(X'WX)^{-1} X'WY$	$\frac{\Sigma (\hat{\beta}_i^* - \bar{\beta}^*)^2}{299}$
BD	Sampling does follow sample design	$(X'WX)^{-1} X'WY$	$\frac{\Sigma (\hat{\beta}_i' - \bar{\beta}')^2}{299}$

where X is the independent matrix;

Y is the dependent matrix;

W is the diagonal matrix of the inverse of the probabilities of selection;

V and s^2 are the diagonal matrix and the constant that compose the estimated variance of Y; such that $Var(Y) = V s^2$;

G is a matrix that estimates the variance of $X'W\epsilon \mid X_i$

$\hat{\beta}_i^*$ and $\hat{\beta}_i'$ are the estimates of β obtained from the i^{th} resample.

and $\bar{\beta}^*$ and $\bar{\beta}'$ are the average value of the respective 300 $\hat{\beta}_i$'s

SIMULATED POPULATIONS

The creation of populations with definite characteristics and the repeated random sampling from them is a powerful tool in the comparison of different methods of estimation of parameters and their variances. Since all the characteristics of the simulated populations are known, it is possible to assess under which circumstances each of the approaches is most effective and robust.

The model used in the simulation was a linear model that related blood pressure to age and body fatness. The regression parameters used in the model were chosen to approximate the actual values found in studies using the results of the Canada Health Survey.

In a large population like that of Canada, the variability in blood pressure may be due to many different factors. An important factor is due to intra-individual variation, that is, the variation in response at different times in the same person. This variability is referred to as "person error". A second important factor is the inter-individual variability. This is observed in that, under the same conditions, people from the east coast of Canada may be very different than people from the Prairies but not so different than other people from the Maritimes. This is referred to as "cluster error". Another way of understanding cluster error is to consider how members of a family tend to have more similarities than do unrelated people.

Four combinations of person error and cluster error are created. Table 2 summarizes the combinations of error used to define the four population structures.

TABLE 2

DIFFERENT COMBINATIONS OF ERROR (CLUSTER AND PERSON) USED TO DEFINE THE FOUR POPULATIONS STRUCTURES

Small cluster error		Large cluster error	
Homogeneous person error A	Heterogeneous person error B	Homogeneous person error C	Heterogeneous person error D
$C \sim N(0,20)$	$C \sim N(0,20)$	$C \sim N(0,80)$	$C \sim N(0,80)$
$P \sim N(0,80)$	$P \sim N(0,v_i)$	$P \sim N(0,20)$	$P \sim N(0,v_i)$
	$v_i=59.5+0.5*Age$		$v_i=-0.5+0.5*Age$
	$\bar{v}_i = 80$		$\bar{v}_i = 20$

where C stands for cluster error and P stands for person error

The equation of v_i reproduces the slope found in the Canada Health Survey for the variance of blood pressure on age. The coefficient for body fatness was non significantly different then zero in the Canadian population.

Each of these combinations of error are added to a basic diastolic blood pressure value. The basic blood pressure regression is

$$\text{Blood Pressure} = 58.5 + 0.19 \times \text{Age} + 0.38 \times \text{Body Fatness}$$

where Age follows a normal distribution with mean 41 and variance 316 ($\sim$ N(41,316)) and Body Fatness $\sim$ N(25,19.9). So a woman aged 30 and with body fatness 25 would have a basic blood pressure of 73.7. Values of cluster and pure error would be simulated for her following the definitions found in Table 2. Finally, her observation on population A would be 73.7 + the cluster error simulated following a N(0,20) + the pure error simulated following a N(0,80). In population B, her observation would be 73.7 + the same cluster error as simulated in population A + the pure error simulated following a N(0,74.5). Etc.

Variations in the variance of the pure error are introduced to allow a comparison of the methods in regards to heterogeneous variance among observations. The distribution of age and body fatness are those found in women in the Canada Health Survey. For both age and body fatness, a minimal value is fixed at 15 in the simulations.

Each of the four populations was used to create samples of size 50, made up of 10 clusters of 5 individuals. In order to study the effect of sample size on the methods, the same population structures were also used to simulate samples of size 100. In that situation, a sample was made up of 10 clusters of 10 individuals.

From each of the population structures, 300 samples of size 50 and 300 samples of size 100 were simulated.

RESULTS

To compare the different values of variance estimates of a given population, a "gold standard" is needed. The value should be as close as possible to the "true" value.

For each population, 300 estimates of the regression coefficients of blood pressure with age and body fatness as well as the intercept were obtained. The calculated variance among them was used as the gold standard. For populations B and D, there were, in fact, an extra 300 estimators of the slopes and intercept obtained from the modification to

Nathan's formula (WLS 3). So, for these two populations, two gold standards were calculated. The first one was used to compare the methods that consider the variance of blood pressure as constant over all values of age in their estimation, while the second one was used to compare WLS 3. The values of the gold standards for the variance of the age coefficients are found in Table 3 along with their mean value and coefficient of variation. No significant differences were found between estimates of the coefficients obtained from WLS 3 and from other methods.

TABLE 3

GOLD STANDARDS

MEAN, VARIANCE AND COEFFICIENT OF VARIATION OF THE ESTIMATES OF THE REGRESSION

COEFFICIENT FOR AGE.*

		A	B	C	D
N = 50	μ	.1919	.1925	.1910	.1916
	σ^2	.009108	.009359	.007835	.007941
	CV	.50	.50	.46	.47
	μ **		.1907		.1887
	σ^2 **		.009228		.007827
	CV**		.50		.47
N = 100	μ	.1904	.1909	.1866	.1876
	σ^2	.004545	.004688	.004523	.004726
	CV	.35	.36	.36	.37
	μ **		.1890		.1873
	σ^2 **		.004636		.004727
	CV**		.36		.37

* for the formula

Blood Pressure = 58.5 + 0.19 * Age + 0.38 * Body Fatness

** statistics of the estimates obtained using WLS 3 (Modification to Nathan's formula)

To summarize the results, coefficients of variation and "ratio tables" are presented for the estimation of the variance of the age coefficient. The coefficient of variation (CV) is defined as the ratio of the standard deviation to the mean value of an estimator. These are presented in Tables 4 and 5. For Tables 6 and 7 (the ratio tables), each entry is obtained as the ratio of the mean value of the estimator over its gold standard.

TABLE 4

COEFFICIENTS OF VARIATION OF VARIANCE ESTIMATORS FOR AGE

300 SAMPLES OF SIZE 50

A		B		C		D	
GOLD	.50	GOLD	.50	GOLD	.46	GOLD	.47
WLS 2	.31	WLS 2	.31	WLS 2	.53	WLS 2	.53
WLS 1	.34	WLS 1	.34	WLS 1	.58	WLS 1	.58
		WLS 3	.33			WLS 3	.55
BND	.51	BND	.54	BND	.72	BND	.71
TEX	.64	TEX	.64	TEX	.96	TEX	.92
BD	.67	BD	.67	BD	.97	BD	.95

TABLE 5

COEFFICIENTS OF VARIATION OF VARIANCE ESTIMATORS FOR AGE

300 SAMPLES OF SIZE 100

A		B		C		D	
GOLD	.35	GOLD	.36	GOLD	.36	GOLD	.37
WLS 2	.23	WLS 2	.23	WLS 2	.42	WLS 2	.41
		WLS 1	.24			WLS 3	.44
WLS 1	.24	WLS 3	.25	WLS 1	.45	WLS 1	.45
BND	.42	BND	.43	BD	.60	BND	.63
BD	.44	BD	.44	BND	.63	BD	.63
TEX	.68	TEX	.66	TEX	.80	TEX	.79

In order to compare the different results within themselves and to the gold standards, a number of statistics were used. First, ANOVA was performed to test the null hypothesis

H_0 : WLS1 = WLS2 = WLS3 = TEX = BND = BD

for each of the variance components. In all cases the null hypothesis was rejected. Scheffe's multiple comparisons of the means was used to determine which pairs of means are significantly different. Lastly, t-tests were performed to compare each of the estimates to the gold standards.

Taking a closer look at the tables, the following observations were made.

WLS 2, WLS 3 and WLS 1 displayed the smallest variation between estimates of variance. (Tables 4 and 5)

The order in which the different ratios appeared was almost always the same with WLS 2, followed by WLS 3 (where appropriate) and BND. When the sample sizes were 50 WLS 1 performed better than TEX and BD. Alternatively, when sample sizes were 100, BD performed better than TEX and WLS 1. (Tables 6 and 7)

TABLE 6

RATIO OF VARIANCE OF ESTIMATOR TO "TRUE" VARIANCE

300 SAMPLES OF SIZE 50

A	B	C	D
GOLD .009108	GOLD .009359	GOLD .007835	GOLD .007941
WLS 2 .99*	WLS 2 .97*	WLS 2 1.09	WLS 2 1.08*
BND .91	WLS 3 .96	BND .97	WLS 3 1.07*
WLS 1 .86	BND .89	WLS 1 .94	BND .97*
TEX .85	TEX .84	BD .89	WLS 1 .92*
BD .80	WLS 1 .83	TEX .86	BD .88*
	BD .78		TEX .85

* non significantly different then the Gold Standard

TABLE 7

RATIO OF VARIANCE OF ESTIMATOR TO "TRUE" VARIANCE

300 SAMPLES OF SIZE 100

A	B	C	D
GOLD .004545	GOLD .004688	GOLD .004523	GOLD .004726
WLS 2 .94	WLS 2 .92	WLS 2 .88	WLS 2 .85
BND .87	WLS 3 .87	BND .78	WLS 3 .80
BD .82	BND .86	BD .78	BND .76
TEX .82	TEX .81	WLS 1 .75	BD .75
WLS 1 .80	BD .80	TEX .70	WLS 1 .72
	WLS 1 .78		TEX .69

Few significant differences appeared when comparing the means of the estimates among themselves. In all cases, when differences were detected, WLS 2 and/or WLS 3 were involved. Although most of the estimates were significantly different from their gold standard, it appeared that WLS 2 constituted the majority among those who showed no differences when the sample sizes are 50. For samples of size 100, no differences were found.

Looking at the effect of homogeneous variance of blood pressure (populations A and C) by comparison to heterogeneity over age (populations B and D), one found only negligeable effects. The ratios of the estimates to the gold standards showed a slight increase with sample sizes of 50, while they showed a decrease with samples of size 100. In the great majority of the cases, the fluctuations were in the magnitude of ±3%.

Another issue is that of changes in cluster and person error. The gold CVs did not change as a function of person or cluster error while the estimates CVs are larger for large cluster error. In the ratio tables of size 50 (Table 6), most estimates increased, getting closer to the gold standard or over-estimating it with large cluster error. When sample size was 100, (Table 7) most estimates decreased with important cluster error.

When the sample size was doubled, the estimated variances were, as expected, smaller. That is, the gold standards and its estimates were smaller. The ranking of the different ratios did not change much except for BD, which improved over WLS 1 with increasing sample size. (Tables 6 and 7)

DISCUSSION

The order in which the different ratios appear is fairly consistent throughout the study. One of the main features is the improvement of BD by comparison to WLS 1 when passing from samples of size 50 to 100. This seems to indicate that, in the simulated populations, WLS 1 is adversely affected by increasing sample size while BD is not.

The homogeneity/heterogeneity of the person error term has very little effect on the results, even when person error has the greatest influence on the total variance of blood pressure. The gold standards do not change with the presence or absence of homogeneity. All the methods seem to react in the same way to this situation. It is interesting to note that WLS 3 - created especially for this situation - does not perform any better than WLS 2 from which it is not significantly different. This might be because the variability (heterogeneity) of the variance of blood pressure was not large enough. Another explanation can simply be that the methods are robust for this change.

The effect of increasing cluster error over person error is a bit puzzling to evaluate. While it is clear that the CVs are larger for large cluster error, the results from the ratio tables are not consistent over samples of size 50 and 100. Even if it seemed after examination of the results for size 50 that some trend could be found, it does not hold when applied to samples of size 100. One also notes that the gold standards decrease with growing cluster effect. This suggests that, with important cluster effect, there is less variability in the estimation of the slopes than with important person error. This lowering of the gold standards and of the estimates of the variance is observed in the CVs where the average of the estimates of the variances are used in the denominator.

Increasing the sample size from 50 to 100, one observes, as expected, that

the variance of the estimators of the coefficients are diminished. When comparing the ratios in population A of size 50 to population A of size 100 etc, no trend is apparent, although in most cases, ratios increase with increasing sample size. The effect of sample size is also found in the CV tables where the coefficients decrease with large sample sizes. This means that the estimation of the variance of the regression coefficients is done with more precision - or less variability - when more observations are present.

CONCLUSION

Considerations for the estimation of the variance of regression coefficients are given to different methods. From the simulations of size 50, WLS 2 - followed by WLS 3 - performs best, having the ratio closest to the gold standard and being, most often, significantly different from the other methods. When the samples are of size 100, one witnesses an improvement of BD, bringing the WLS 2-WLS 3-BND-BD group closer together, with TEX an WLS 1 following behind. From these results it seems that WLS 2-WLS 3 and/or BND-BD evaluate the most reliable estimates of the variance when the sample size is relatively small and survey design is not too complex.

Further work needs to be done to properly make a distinction between WLS 2-WLS 3 and BND-BD and assess what are the conditions that bring about significant differences between them. In order to achieve this, more simulations can be done where more conditions are imposed on the populations - such as fixed correlations between age and body fatness, or varying cluster size within a sample etc. Just as important is the question of what happens when the design is of increasing complexity. TEX very well could perform better in a more complex setting. One would also hope to see WLS 1 show its limitations in such circumstances. To begin to answer this question, simulations can be done with increasing complexity in their sampling design and also greater sample size.

BIBLIOGRAPHY AND REFERENCES

1. Binder B.A.: On the variances of asymptotically normal estimators from complex survey. International Statistical Reviews 51:279-292, 1983

2. Draper N.R, Smith H: Applied regression analysis. ed. John Wiley & Sons Inc, 1981

3. Efron B: The jackknife, the bootstrap and other resampling plans. ed. Philadelphia: Society for Industrial and Applied Mathematics, 1982

4. Fuller W.A: Regression analysis for sample surveys. Sankhya 37:117-132, 1975

5. Health & Welfare and Statistics Canada: The health of Canadians. Report of the Canada Health Survey. ed. Ottawa: Ministry of supply and Services Canada Catalogue 82-538E, 1981

6. Nathan G: Notes on inference based on data from complex sample designs. Survey Methodology 7:109-129, 1981

The authors would like to give special thanks to Mr. Tony Panzarella from the Princess Margaret Hospital for kindly reviewing this paper.