SSC Annual Meeting, June 1997
Proceedings of the Survey Research Section

A (PARTIALLY) MODEL-BASED LOOK AT JACKKNIFE VARIANCE ESTIMATION WITH TWO-PHASE SAMPLES

Phillip S. Kott[1]

ABSTRACT

This paper focuses on a design-consistent regression estimator in which the "auxiliaries" are estimated from a stratified cluster sample and the regression coefficients from an arbitrary subsample of the original sample. The reweighted expansion estimator described in Stukel and Kott (1997) is an example of such an estimator. Assuming that the target variable is a linear function of the auxiliaries plus an error term, asymptotic properties for both this estimator and the jackknife estimator of its mean squared error are developed. These theoretical results are used to explain some of Stukel and Kott's empirical findings, which in turn shed light on the asymptotic underpinnings of the theoretical results.

KEY WORDS: Asymptotic; Bias; Double expansion estimator; Primary sampling unit; Reweighted expansion estimator.

1. INTRODUCTION

This paper will focus on a two-phase design-consistent regression estimator for the mean of a single target variable computed in the following manner: first, a vector of covariate means is estimated from a *stratified cluster sample*; then, a vector of regression coefficients relating the target variable to the covariates at the element level is estimated from an *arbitrary* subsample of the first-phase sample. The estimator has the same form as the conventional design-consistent regression estimator for a population mean of a target variable except that a first-phase estimated mean for the vector of covariates is used in place of the population mean for a vector of auxiliary variables.

We will assume a model in which the target variable is a linear function of the covariates plus an error term. The two-phase regression estimator and an expression for its variance under a combined randomization (for the first phase) and model (for the second phase) framework is discussed in Section 2. A jackknife estimator for this variance is analyzed in Section 3. Section 4 concentrates on the special case of a two-phase regression estimator in projection form. Some empirical results for the weighted expansion estimator partially published in Stukel and Kott (1997) are reviewed in light of this analysis. Other results from the same empirical study are discussed in Section 5, which extends the theoretical treatment in earlier sections to ratios of two-phase regression estimators. Section 6 provides a more general discussion.

2. NOTATION FOR THE TWO-PHASE REGRESSION ESTIMATOR

We need quite a bit of notation. Let i denote an element in the population of interest, j a first-phase primary sampling unit (PSU), and h (= 1, ..., H) a first-phase stratum. There are n_h sampled PSU's in stratum h and n sampled PSU's overall. The population size (in number of elements) is M, while the second-phase sample size is m. Let S(hj) denote the set of elements in PSU j of stratum h. Let S denote the entire first-phase sample of elements (an element is in the first-phase sample if it is contained within a PSU in the first-phase sample).

Let w_{1i} denote the first-phase expansion factor for i; that is the inverse of the first-phase selection probability of the PSU containing i. Let w_{2i} denote the second-phase expansion factor for i: when element i is in the second-phase sample, w_{2i} is the inverse of the conditional probability of selecting i for that phase (conditioned on the first-phase sample); when i in not in the second-phase sample, $w_{2i} = 0$.

The two-phase regression estimator for the population mean $y_0 = \sum^M y_i/M$ we will be addressing here has the form:

$$t = \sum_S w_{1i}w_{2i}y_i/M + (\sum_S w_{1i}\mathbf{x}_i/M - \sum_S w_{1i}w_{2i}\mathbf{x}_i/M)\mathbf{b}, \qquad (1)$$

where $\mathbf{x}_i$ denotes a row vector of K covariates,

$$\mathbf{b} = (\sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'\mathbf{x}_i/M)^{-1} \sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'y_i/M, \qquad (2)$$

and the c_i are arbitrary constant.

In many applications, M is unknown. Consequently, t in equation (1) is not a practical estimator for a population mean. Observe, however, that Mt *is* a practical estimator for a population total. The two estimators have parallel properties. They have identical relative biases and identical relative mean squared errors. We focus here on t to simplify the asymptotics.

If the c_i in equation (2) have the form $c_i = 1/(\mathbf{x}_i\lambda)$, where λ is a column vector, then t can be put in projection form:

$$t_{PROJ} = (\sum_S w_{1i}\mathbf{x}_i/M)\mathbf{b}$$

since

$$\begin{aligned}(\sum_S w_{1i}w_{2i}\mathbf{x}_i/M)\mathbf{b} &= (\sum_S w_{1i}w_{2i}\mathbf{x}_i/M)(\sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'\mathbf{x}_i/M)^{-1} \sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'y_i/M \\ &= (\sum_S [c_i\lambda'\mathbf{x}_i']w_{1i}w_{2i}\mathbf{x}_i/M)(\sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'\mathbf{x}_i/M)^{-1} \sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'y_i/M \\ &= \lambda'(\sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'y_i/M = \sum_S w_{1i}w_{2i}y_i/M.\end{aligned}$$

The reweighted expansion estimator in Kott ans Stukel (1997) is in projection form as is the more general two-phase regression estimator discussed later in that paper.

We assume the y_i and $\mathbf{x}_i$ are related by the following model:

$$y_i = \mathbf{x}_i\beta + e_i, \qquad (3)$$

where $E(e_i \mid \{\mathbf{x}_k\})$, and $E(e_ie_g \mid \{\mathbf{x}_k\}) = 0$ for i and g from different PSU's, while $E(e_ie_g \mid \{\mathbf{x}_k\})$ is bounded other-wise. This structure allows the elemental errors – the e_i – within the same PSU to be correlated in an arbitrary manner. It should be noted, however that if the second-phase of sampling uses a clustered selection process, and second-phase clusters cut across first-phase PSU's, elements from the same second-phase cluster but different PSU's are assumed not to have correlated errors.

For our purposes, the target of estimation, y_0, is virtually identical to $\sum^M \mathbf{x}_i\beta/M = \mathbf{x}_0\beta$. As a result, *we will treat* $\mathbf{x}_0\beta$ *as the target of estimation from now on.*

The difference between t_{PROJ} and $\mathbf{x}_0\beta$ is

$$t_{PROJ} - \mathbf{x}_0\beta = (\mathbf{x}_1 - \mathbf{x}_0)\beta + \mathbf{x}_1(\mathbf{b} - \beta),$$

where $\mathbf{x}_1 = \sum_S w_{1i}\mathbf{x}_i/M$. The model mean squared error of t_{PROJ} is then

$$E_e[(t_{PROJ} - \mathbf{x}_0\beta)]^2 = [(\mathbf{x}_1 - \mathbf{x}_0)\beta]^2 + \mathbf{x}_1\mathbf{Var}_e(\mathbf{b})\mathbf{x}_1'.$$

The first term of this expression is the square of the model bias of t_{PROJ} . The expectation of this term with respect to randomization (i.e., its design expectation) is the randomization variance of $\mathbf{x}_1\beta$. For our purposes, then the randomization-model variance of t_{PROJ} is

$$\mathrm{Var}_{RM}(t_{PROJ}) = \mathrm{Var}_1(\mathbf{x}_1\beta) + \mathbf{x}_1\mathbf{Var}_e(\mathbf{b})\mathbf{x}_1',$$

where the subscript 1 denote randomization inference with respect to the first phase of sampling. The right hand side of this expression differs from the model expectation of the randomization mean squared of t (or, equivalently, the randomization expectation of the model mean squared error of t) in that the model variance component, $\mathbf{x}_1\mathbf{Var}_e(\mathbf{b})\mathbf{x}_1'$, is conditioned on the realized first-phase sample.

When t in equation (1) cannot be put in projection form, the situation is a bit messier. The model variance of t is $\mathbf{Var}_e\{e_2 + (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{b}\}$, where $e_2 = \sum_S w_{1i}w_{2i}e_i/M$, and $\mathbf{x}_2$ is defined analogously. The randomization component of the randomization-model variance of t is the same as that of t_{PROJ} (i.e., $\mathrm{Var}_1(\mathbf{x}_1\beta)$).

Let $\mathbf{x}_{1hj} = \sum_{S(hj)} w_{1i}\mathbf{x}_i/M$; $\mathbf{x}_{2hj} = \sum_{S(hj)} w_{1i}w_{2i}\mathbf{x}_i/M$; $y_{2hj} = \sum_{S(hj)} w_{1i}w_{2i}y_i/M$; $\mathbf{q}_{hj} = \sum_{S(hj)} c_i w_{1i}w_{2i}\mathbf{x}_i'y_i/M$; $e_{2hj} = \sum_{S(hj)} w_{1i}w_{2i}e_i/M$; $\mathbf{u}_{hj} = \sum_{S(hj)} c_i w_{1i}w_{2i}\mathbf{x}_i'e_i/M$; and $\mathbf{Z}_{hj} = \sum_{S(hj)} c_i w_{1i}w_{2i}\mathbf{x}_i'\mathbf{x}_i/M$. Just as $\mathbf{x}_2 = \sum^H \sum_j \mathbf{x}_{2hj}$, let $\mathbf{q} = \sum^H \sum_j \mathbf{q}_{hj}$ (since the subscript 2 is not needed for clarification, it has been suppressed). Define $\mathbf{u}$ and $\mathbf{Z}$ analogously.

Now $\mathbf{b} = \mathbf{Z}^{-1}\mathbf{q}$ and $\mathbf{Var}_e(\mathbf{b}) = \mathbf{Var}_e(\mathbf{Z}^{-1}\mathbf{u}) = \mathbf{Z}^{-1}E(\mathbf{u}\mathbf{u}')\mathbf{Z}^{-1} = \sum^H \sum_j \mathbf{Z}^{-1}E(\mathbf{u}_{hj}\mathbf{u}_{hj}')\mathbf{Z}^{-1} = \sum \sum \mathbf{Var}(\mathbf{Z}^{-1}\mathbf{u}_{hj})$. The model variance of t is thus

$$\begin{aligned}\mathrm{Var}_e(t) &= \sum \sum \mathrm{Var}_e (e_{2hj} + [\mathbf{x}_1 - \mathbf{x}_2]\mathbf{Z}^{-1}\mathbf{u}_{hj}) \\ &= \sum \sum \mathrm{Var}_e (a_{hj}),\end{aligned}$$

where $a_{hj} = e_{2hj} + [\mathbf{x}_1 - \mathbf{x}_2]\mathbf{Z}^{-1}\mathbf{u}_{hj}$. For t_{PROJ}, this collapses to $\sum \sum \mathbf{Var}_e (\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{u}_{hj})$.

The randomization component of the randomization-model variance of t is

$$\mathrm{Var}_1(\ \mathbf{x}_1\beta) = \sum \sum \mathrm{Var}(\mathbf{x}_{1hj}\beta).$$

We will assume either, 1, the first-phase sample was drawn with *replacement but that the population, strata definitions, and design are such that an element can almost never be selected more than once, or, 2, the first-phase sample was drawn* without *replacement but that the population, strata definitions, and sample design are such that using the with-replacement variance estimator has an ignorably small bias.*

If β were known, the standard with-replacement variance estimator for $\mathrm{Var}_1(\mathbf{x}_1\beta)$ is

$$\mathrm{var}_1(\mathbf{x}_1\beta) = \sum^H (n_n/[n_h - 1])\ [\ \sum_j (\mathbf{x}_{1hj}\beta)^2 - (\ \sum_j \mathbf{x}_{1hj}\beta)^2/n_h].$$

We will call

$$v_I(t) = \mathrm{var}_1(\mathbf{x}_1\beta) + \sum^H \sum_j \mathrm{Var}_e(a_{hj}) \qquad (4)$$

the *ideal* estimator for the randomization-model variance of t; that is

$$\mathrm{Var}_{RM}(t) = \mathrm{Var}_1(\mathbf{x}_1\beta) + \sum^H \sum_j \mathrm{Var}_e(a_{hj}).$$

3. THE JACKKNIFE VARIANCE ESTIMATOR FOR t

Let $f_{(hj)} = f - (n_n/[n_h - 1])(\ f_{hj} - \sum_g f_{hg}/n_h)$, where f has a linear form such as y_2, $\mathbf{x}_1$, $\mathbf{x}_2$, e, $\mathbf{q}$, $\mathbf{u}$, or $\mathbf{Z}$. The expression $f_{(hj)}$ is called the hj'th jackknife replicate of f.

Now t can be rendered as $y_2 + (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{q}$. The hj'th jackknife replicate of t is

$$t_{(hj)} = y_{2(hj)} + (\mathbf{x}_{1(hj)} - \mathbf{x}_{2(hj)})\mathbf{Z}^{-1}{}_{(hj)}\mathbf{q}_{(hj)}.$$

The jackknife variance estimator for t is

$$v_J(t) = \sum^H [(n_h - 1)/n_h] \sum_j (t_{(hj)} - t)^2.$$

To evaluate this variance estimator, we need first evaluate the differences $t_{(hj)} - t$:

$$\begin{aligned} t_{(hj)} - t &= \mathbf{x}_{2(hj)}\beta + e_{2(hj)} + (\mathbf{x}_{1(hj)} - \mathbf{x}_{2(hj)})\mathbf{Z}_{(hj)}{}^{-1}\mathbf{q}_{(hj)} \\ &\quad - x_2\beta - e_2 - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{q} \\ &= (e_{2(hj)} - e_2) + (\mathbf{x}_{1(hj)} - \mathbf{x}_1)\beta \qquad (5) \\ &\quad + (\mathbf{x}_{1(hj)} - \mathbf{x}_{2(hj)})\mathbf{Z}_{(hj)}{}^{-1}\mathbf{u}_{(hj)} - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{u} \end{aligned}$$

We need some asymptotics to handle the **Z**-inverse terms. *We will assume that all the linear expressions like f_{hj} are $O_p(1/n)$, while f itself is no more than $O_P(1)$.* In fact, e and **u** are $\mathbf{O}_P(1/\surd n)$, since each are the sum of n independent $\mathbf{O}_P(1/n)$ terms.

Let $f^{(hj)} = f - f_{(hj)} = (n_n/[n_h - 1])(\ f_{hj} - \sum_g f_{hg}/n_h)$, so that

$$\begin{aligned} \mathbf{Z}_{(hj)}{}^{-1} &= [\mathbf{Z} - \mathbf{Z}^{(hj)}]^{-1} = [\mathbf{Z}\ (\mathbf{I} - \mathbf{Z}^{-1}\mathbf{Z}^{(hj)})]^{-1} \\ &= (\mathbf{I} + \mathbf{Z}^{-1}\mathbf{Z}^{(hj)} + \mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1} + O_p(1/n^3). \end{aligned}$$

Plugging $\mathbf{Z}_{(hj)}{}^{-1}$ and the definition of $f^{(hj)}$ into the left hand side of equation (5) yields:

$$\begin{aligned} t_{(hj)} - t &= - \mathbf{x}_1{}^{(hj)}\beta - e_2{}^{(hj)} - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{u}^{(hj)} \\ &\quad - (\mathbf{x}_1{}^{(hj)} - \mathbf{x}_2{}^{(hj)})\mathbf{Z}^{-1}(\mathbf{u} - \mathbf{u}^{(hj)}) \\ &\quad + (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\mathbf{Z}^{-1}[\mathbf{u} - \mathbf{u}^{(hj)}] + O_p(n^{-5/2}) \end{aligned}$$

$$\begin{aligned} &= - \mathbf{x}_1{}^{(hj)}\beta - e_2{}^{(hj)} - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{u}^{(hj)} \qquad (6) \\ &\quad - \{(\mathbf{x}_1{}^{(hj)} - \mathbf{x}_2{}^{(hj)}) - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\}\mathbf{Z}^{-1}[\mathbf{u} - \mathbf{u}^{(hj)}] \\ &\quad + O_p(n^{-5/2}). \end{aligned}$$

Dropping terms of order $0_P(n^{-3/2})$, we have

$$\begin{aligned} t_{(hj)} - t &\approx - \mathbf{x}_1{}^{(hj)}\beta - e_2{}^{(hj)} - (\mathbf{x}_1 - \mathbf{x}_2)\mathbf{Z}^{-1}\mathbf{u}^{(hj)} \\ &= - \mathbf{x}_1{}^{(hj)}\beta - a^{(hj)} \\ &= - \{(n_n/[n_h - 1])(\ \mathbf{x}_{1hj}\beta - \sum_g x_{1hg}\beta/n_h) \\ &\quad + (n_n/[n_h - 1])(\ a_{hj} - \sum_g a_{hg}/n_h)\}. \qquad (7) \end{aligned}$$

So that

$$\begin{aligned} E_e[(t_{(hj)} - t)^2] &\approx \\ &[n_h/(n_h - 1)]^2\{(\mathbf{x}_{1hj}\beta - \sum_g x_{1hg}\beta/n_h)^2 + \\ &\quad [(1 - 2/n_h)\mathrm{Var}_e(a_{hj}) + \sum_g \mathrm{Var}_e(a_{hg})/n_h]\}, \end{aligned}$$

and

$$\begin{aligned} E_e[v_J(t)] &= \sum^H [(n_h - 1)/n_h] \sum_j E_e[\ (t_{(hj)} - t)^2] \\ &\approx \mathrm{var}_1(\mathbf{x}_1\beta) + \sum^H \sum_j \mathrm{Var}_e(a_{hj}). \end{aligned}$$

This last near equality tells us that the jackknife is a good estimator for the randomization-model variance of t discussed in the last section. In fact, incorporating some of the higher order terms dropped from equation (6), we can conclude that

$$\begin{aligned} E_e[v_J(t)] &= \mathrm{var}_1(\mathbf{x}_1\beta) + \sum^H \sum_j \mathrm{Var}_e(a_{hj}) + O(1/n^2) \\ &= v_I(t) + O(1/n^2), \end{aligned}$$

where $v_I(t)$ is the ideal variance estimator defined in equation (4). If we make the additional mild assumptions that the sampling design and population are such that $v_I(t)$ and $\mathrm{Var}_{RM}(t)$ are O(1/n), then the *relative* bias of the jackknife relative to the "gold standard" of the ideal variance estimator is O(1/n).

4. THE JACKKNIFE FOR THE TWO-PHASE REGRESSION ESTIMATOR IN PROJECTION FORM

When the two-phase regression estimator can be put in projection form; that is, when $c_i = 1/(\mathbf{x}_i\lambda)$ for some vector λ, equation (6) collapses to

$$\begin{aligned} t_{(hj)PROJ} - t_{PROJ} &= -\mathbf{x}_1^{(hj)}\beta - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{u}^{(hj)} \\ &\quad - (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1}[\mathbf{u} - \mathbf{u}^{(hj)}] \\ &\quad + O_p(n^{-5/2}). \end{aligned} \tag{6'}$$

This is because $\lambda' c_i \mathbf{x}_i' = 1$, so that $\mathbf{x}_2\mathbf{Z}^{-1} = \lambda'$, $\lambda'\mathbf{u}_{hj} = e_{2hj}$, and $e_2^{(hj)} = \mathbf{x}_2\mathbf{Z}^{-1}\mathbf{u}^{(hj)}$. In addition, $\lambda'\mathbf{Z}_{hj} = \mathbf{x}_{2hj}$, so that $\mathbf{x}_2^{(hj)} = \mathbf{x}_2\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}$.

Denote $E(\mathbf{u}_{hj}\mathbf{u}_{hj}')$ by $\mathbf{V}_{hj}$. Recall that the $\mathbf{u}_{hj}$ are independent and $\mathbf{u}^{(hj)} = [n_h/(n_h - 1)](\mathbf{u}_{hj} - \sum_g \mathbf{u}_{hg}/n_h]$, so that $E(\mathbf{u}^{(hj)}\mathbf{u}) = [n_h/(n_h - 1)](\mathbf{V}_{hj} - \sum_g \mathbf{V}_{hg}/n_h)$ and $E(\mathbf{u}^{(hj)}\mathbf{u}^{(hj)\prime}) = [n_h/(n_h - 1)]^2 (\mathbf{V}_{hj}[1 - 2/n_h] + \sum_g \mathbf{V}_{hg}/n_h^2)$. From (6'), we have

$$\begin{aligned} E_e[(t_{(hj)PROJ} - t_{PROJ})^2] &= (\mathbf{x}_1^{(hj)}\beta)^2 + \mathrm{Var}_e(\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{u}^{(hj)}) \\ &\quad + 2(\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1} \\ &\qquad E_e([\mathbf{u} - \mathbf{u}^{(hj)}]\mathbf{u}^{(hj)\prime})\mathbf{Z}^{-1}\mathbf{x}_1' \\ &\quad + (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1} \\ &\qquad E_e(\mathbf{u}\mathbf{u}')\mathbf{Z}^{-1}(\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})' \\ &\quad + O(n^{-7/2}) \end{aligned}$$

$$\begin{aligned} &\approx (\mathbf{x}_1^{(hj)}\beta)^2 + \mathrm{Var}_e(\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{u}^{(hj)}) \\ &\quad + 2(\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1}[n_h/(n_h - 1)^2] \\ &\qquad (\mathbf{V}_{hj} - \textstyle\sum_g \mathbf{V}_{hg}/n_h)\mathbf{Z}^{-1}\mathbf{x}_1' \\ &\quad + (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1}(\textstyle\sum\sum \mathbf{V}_{fg})\mathbf{Z}^{-1} \\ &\qquad (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})'. \end{aligned}$$

Letting $\mathbf{V}$ denote $\sum\sum \mathbf{V}_{hj}$ and $\mathbf{V}^{(hj)}$ denote $[n_h/(n_h - 1)](\mathbf{V}_{hj} - \sum_g \mathbf{V}_{hg}/n_h)$, we can express the expected value of the jackknife variance estimator for t_{PROJ} as

$$\begin{aligned} E_e(v_{J[PROJ]}) &\approx \\ &\mathrm{var}_1(\mathbf{x}_1\beta) + \mathrm{Var}_e(\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{u}) \\ &+ 2\textstyle\sum (1/n_h) \sum (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1}\mathbf{V}^{(hj)}\mathbf{Z}^{-1}\mathbf{x}_1' \\ &+ \textstyle\sum\sum ([n_h - 1]/n_h)(\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})\mathbf{Z}^{-1}\mathbf{V}\mathbf{Z}^{-1} \\ &\qquad (\mathbf{x}_1^{(hj)} - \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)})'. \end{aligned} \tag{8}$$

The asymptotic bias of the jackknife is captured by the last two lines on the left hand side of equation (8).

For many applications, $\mathbf{V}$ will be roughly equal to a multiple of $\mathbf{Z}$. Observe that both $\mathbf{x}_1^{(hj)}$ and $\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}$ have randomization expectations of (asymptotically) zero, but that the former is often likely to be a good less variable because it is based on the entire first-phase sample. Consequently, in many applications the contribution to the asymptotic bias of the jackknife from first-phase stratum h – roughly proportional to $([n_h - 3]/n_h)\ \mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\mathbf{Z}^{-1}\mathbf{x}_1$ – will be negative (positive) when n_h is less (greater) than 3.

One popular example of the two-phase regression estimator in projection form is the reweighted expansion estimator for a population *total* explored in Stukel and Kott (1997). For this estimator $\mathbf{x}_i$ is a vector of group membership indicators, where the groups are mutually exclusive, and all the c_i are equal to 1. Consequently, the components of $M\mathbf{x}_1$ and $M\mathbf{x}_2$ are estimators of the group population totals based on the first- and second-phase samples respectively, $\mathbf{Z}$ is a diagonal matrix with the same values of $\mathbf{x}_2$, and $\mathbf{b}$ is a vector of estimates of the group y-means based on the second-phase sample.

The reweighted expansion estimator in Stukel and Kott had the form $Mt = M\mathbf{x}_1\mathbf{b}$. They also inves-tigated the double expansion estimator, which had the form $M\mathbf{x}_2\mathbf{b}$. Diane Stukel performed a simulation in which, first, a with-replacement, stratified, simple random sample of area clusters (PSU's) was drawn, then all the individuals from the sampled clustered were restratified into 5 age groups, and a without-replacement stratified, simple random second-phase sample of individuals was drawn. Two PSU's were sampled from each of the 18 first-phase strata, while second-phase stratum sample sizes ran from 5 to 50 individuals. Four thousand (4, 000) simulations were conducted for each sample size. For comparison purposes, completely-enumerated first-phase sample were also simulated. More details on the data set and simulations are provided in Stukel and Kott (1997).

The results of this empirical analysis for the variable "total employment" are presented in Table 1. For the jackknife of the double expansion estimator, the replicate $t_{(hj)}$ was set equal to $y_{2(hj)}$. A less successful alternative formulation discussed in Stukel and Kott is not presented here. Note that when the entire first-phase sample is enumerated the reweighted and double expansion estimators are the same.

The reweighted expansion estimator appears to have a trivial relative bias which increases in absolute value as the second-phase sample size decreases. The double expansion estimator is unbiased. The tiny relative biases for this estimator in the Table are due to the finite nature of the simulations.

The reweighted expansion estimator is modestly more efficient (has less mean squared error) that the corresponding double expansion estimator. The efficiency gain appears to increase as the second-phase sample size decreases.

The jackknife variance estimates for the reweighted expansion estimators have small negative biases, which were anticipated by the theory developed in this section. These biases tend to increase in absolute value as the second-phase sample size decreases. In theory, they should be an asymptotic function of the first-stage sample size only. In practice, the increasing relative variability of the $\mathbf{Z}_{hj}$ within first-phase strata – and thus $\mathbf{x}_1\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}\mathbf{Z}^{-1}\mathbf{Z}^{(hj)}$ $\mathbf{Z}^{-1}\mathbf{x}_1'$ (see equation (8) and the discussion following it) – may be the determining factor. Note that the components of the diagonal matrix $Mn_h\mathbf{Z}_{hj}$ are estimators of the number of individuals within first-phase stratum h from a particular age group (second-phase stratum) based on the second-phase sample in PSU j of stratum h. The smaller the second-phase sample size, the more variable the $\mathbf{Z}_{hj}$ within strata given fixed n_h.

The jackknife variance estimates for the double expansion estimators have a strong upward bias. Am explanation for this can be found in Kott (1990), where using the standard two-*stage*, with-replacement variance estimator – equivalent to the jackknife in this case – is shown to be biased upward.

Table 2 presents the results of another set of simulations conducted by Stukel using the same data, first-phase strata, and second-phase sample design as in Stukel and Kott, but with 70 first-phase sample PSU's (out of 220). In these new simulations, 8 of the strata have 4 or more sampled PSU's. The other 10 again have 2.

The absolute relative biases of the reweighted expansion estimator is a bit larger in the new simulations but are still small. Surprisingly, the efficiencies of the estimators go down (except , of course, when the whole first-phase sample is enumerated).

The efficiency gains from using the reweighted over the double expansion estimator are more pronounced in the new simulations. a reasonable explanation for this is that the precision of the estimator $\mathbf{x}_1$ is greatly improved by adding PSU's, while the precision of $\mathbf{x}_2$ – based a second-phase sample that has not increased in size – is not.

The relative biases of the jackknife for the reweighted expansion estimator remain small but now are not all negative. This may be due to those strata with more than 3 sampled PSU's.

Table 1. Estimating Total Employment With a Two-PSU-Per-Stratum Design

Second-Phase Stratum	*Reweighted Expansion Estimator*			*Double Expansion Estimator*		
Sample Size	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife
All	0.04	100	0.94	0.04	100	0.94
50	0.14	183	-0.99	0.16	186	46.4
20	-0.30	323	-2.51	-0.01	360	68.2
10	-0.29	549	-5.81	0.03	632	78.2
5	-0.56	1002	-5.13	0.12	1171	86.2

Table 2. Estimating Total Employment With a Variable-PSU-Per-Stratum Design

Second-Phase Stratum	*Reweighted Expansion Estimator*			*Double Expansion Estimator*		
Sample Size	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife
All	0.11	59	-1.94	0.11	59	-1.94
50	-0.15	186	-4.49	0.09	287	31.9
20	-1.12	366	-6.56	-0.34	620	41.1
10	-1.51	622	-6.03	-0.06	1197	41.6
5	-2.35	1088	3.57	0.14	2319	44.7

All results based on 4,000 simulations.

* Scaled so that *original* estimator based on the full first-phase sample has Variance (MSE) equal to 100.

There does not appear to be reductions in relative bias for a given second-phase sample size from increasing number of PSU's. One reason for this may be the reduced first-phase variance, which – all other things being equal – increases the potential contribution to *relative* bias from second-phase variance estimation. Another possibility is that the $\mathbf{Z}_{hj}$ have become more variable as the same second phase sample size is distributed over move PSU's.

Finally, the jackknifes for the double expansion estimator have reduced, but still unacceptably high, relative biases.

5. THE RATIO OF TWO-PHASE REGRESION ESTIMATORS

Let us adopt the framework of Section 2 for two variables of interest, $y_i^{[1]}$ and $y_i^{[2]}$. In particular for k = 1 or 2, let $y_0^{[k]} = \sum^M y_i^{[k]}/M$,

$$t^{[k]} = \sum_S w_{1i}w_{2i}y_i^{[k]}/M + \\ (\sum_S w_{1i}\mathbf{x}_i/M - \sum_S w_{1i}w_{2i}\mathbf{x}_i/M)\mathbf{b}^{[k]},$$

and

$$\mathbf{b}^{[k]} = (\sum_S c_i w_{1i}w_{2i}\mathbf{x}_i'\mathbf{x}_i/M)^{-1} \sum_S c_i w_{1i}w_{2i}\mathbf{x}_i' y_i^{[k]}/M.$$

We assume that each $y_i^{[k]}$ is related to $\mathbf{x}_i$ through the model:

$$y_i^{[k]} = \mathbf{x}_i\beta^{[k]} + e_i^{[k]},$$

where the $e_i^{[k]}$ have mean zero and are uncorrelated across PSU's and bounded within PSU's. *It is not necessary for $e_i^{[1]}$ and $e_i^{[2]}$ to be uncorrelated.*

We are interested in the properties of $r = t^{[1]}/t^{[2]}$ as an estimator for $y_0^{[1]}/y_0^{[2]}$, which for our purposes is indistinguishable from $R = \mathbf{x}_0\beta^{[1]}/\mathbf{x}_0\beta^{[2]}$. Define y_i as $y_i^{[1]} - Ry_i^{[2]}$, and define e_i analogously. Defining other terms from Section 2 as before based on y_i, e_i, $y_i^{[2]}$, and $e_i^{[2]}$ above, we have

$$r - R = (t^{[2]})^{-1}[\mathbf{x}_1(\beta^{[1]} - R\beta^{[2]}) + \sum\sum a_{hj}]$$

$$= (t^{[2]})^{-1}[(\mathbf{x}_1 - \mathbf{x}_0)(\beta^{[1]} - R\beta^{[2]}) + \sum\sum a_{hj}].$$

Under the asymptotic assumptions analogous to those in Section 3, we have

$$t^{[2]}/\mathbf{x}_1\beta^{[2]} = 1 - (\mathbf{x}_1\beta^{[2]})^{-1} \sum\sum a_{hj}^{[2]} + 0_p(1/n).$$

So that

$$E_e[(r - R)^2] = (\mathbf{x}_1\beta^{[2]})^{-2}\{[(\mathbf{x}_1 - \mathbf{x}_0)(\beta^{[1]} - R\beta^{[2]})]^2 \\ + \sum\sum \mathrm{Var}_e a_{hj}\}[1 + 0(1/n)]. \qquad (9)$$

If the population and first-phase sampling design are such that

$$\mathrm{Var}_1\{(\mathbf{x}_1\beta^{[2]})^{-1}[(\mathbf{x}_1 - \mathbf{x}_0)(\beta^{[1]} - R\beta^{[2]})]\} = \\ \mathrm{Var}_1\{(\mathbf{x}_0\beta^{[2]})^{-1}[(\mathbf{x}_1 - \mathbf{x}_0)(\beta^{[1]} - R\beta^{[2]})]\} \\ [1 + 0(1/n)], \quad (10)$$

then a reasonable expression for randomization-model mean squared error of r is

$$MSE_{RM}(r) = \{(\mathbf{x}_0\beta^{[2]})^{-2}\mathrm{Var}_1[\mathbf{x}_1(\beta^{[1]} - R\beta^{[2]})] \\ + (\mathbf{x}_1\beta^{[2]})^{-2}\sum\sum \mathrm{Var}_e a_{hj}\}[1 + 0(1/n)].$$

Using similar arguments, we can show that the jackknife variance estimator based on the jackknife replicates $r_{(hj)} = t_{(hj)}^{[1]}/t_{(hj)}^{[2]}$ has a model expectation (given a realized sample) asymptotically equal to

$$[\mathbf{x}_1\beta^{[2]}]^{-2}\mathrm{var}_1[\mathbf{x}_1(\beta^{[1]} - R\beta^{[2]})] + (\mathbf{x}_1\beta^{[2]})^{-2}\sum\sum \mathrm{Var}_e a_{hj},$$

which is itself asymptotically close to

$$[\mathbf{x}_0\beta^{[2]}]^{-2}\mathrm{var}_1[\mathbf{x}_1(\beta^{[1]} - R\beta^{[2]})] + (\mathbf{x}_1\beta^{[2]})^{-2}\sum\sum \mathrm{Var}_e a_{hj}.$$

Thus, the jackknife provide a reasonable estimator for the randomization-model variance of r.

Table 3 is based on the same set of simulations as Table 1 from the last section. Here, the employment rate, the ratio of total employment to the total number of individuals in the workforce, is the target of estimation.

Observe that the ratio of two double expansion estimators is usually no less efficient than the ratio of two reweighted expansion estimators. Moreover, the jackknife provides reasonable mean squared error estimates for both the double and reweighted expansion estimators. These two results have a simple explanation: $(\beta^{[1]} - R\beta^{[2]})$ in equations (9) and (10) must be close to $\mathbf{0}$; that is, the employment rate must be close to equal across second-phase strata. This (near) equality (virtually) removes the randomization component, which revolves around the estimation of $\mathbf{x}_0$ with either $\mathbf{x}_1$ or $\mathbf{x}_2$, from the mean squared errors of the two expansion estimators. As a result, the jackknife need only estimate the model variance of the two estimators, which it does reasonably well in both cases. Its negative bias for the double expansion estimator can be shown to have an explanation analogous to the reweighted expansion estimator for the employment total.

The absolute relative biases for the jackknife in Table 3 seem larger than those in Table 1. This may be due to the (near) disappearance of the randomization component of mean squared error increasing the potential for contribution to relative bias in variance estimation from the second phase of sampling.

Table 4 returns to the data set, sampling design, and target as Table 1. Now, however, two new estimators are considered. Both divide the first-phase strata into four post-strata. Two of these post-strata contain 2 strata, a third contains 4, and the last 10.

Within each post-stratum, the ratio of total employment to total number of individuals is estimated using the ratio of reweighted and double expansion estimators, respectively ($y_j^{[k]}$ is set to 0 when i is out of the post-stratum). Census counts for the four post-strata are used to weight the estimated ratios together. This produces estimators that are usually more efficient than analogous reweighted expansion estimators (except when the second-phase sample size is 5 elements per stratum).

The post-stratified estimators have more noticeable, although still small, relative biases. The post-stratified double expansion estimator is less efficient than the post-stratified reweighted expansion estimator and its jackknife variances have pronounced upward bias. It seems that within post-strata ($\beta^{[1]} - R\beta^{[2]}$) is not close to $\mathbf{0}$; that is, the ratio of total employment to total number of individuals varies across second-phase strata.

The absolute relative biases for the jackknife of the post-stratified reweighted expansion estimator are the largest we have seen for reweighted estimators. This is likely due to the small number of PSU's within some post-stratum ratios, which severely challenges a theoretical result based on asymptotics.

Stukel conducted analogous simulations to those summarized in Tables 3 and 4 based on 70 first-phase sample PSU's. They offer little additional insight, however, and will not been reviewed here.

Table 3. Estimating the Employment Rate With a Two-PSU-Per-Stratum Design

Second-Phase Stratum	*Reweighted Expansion Estimator*			*Double Expansion Estimator*		
Sample Size	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife
All	-0.09	100	2.08	-0.09	100	2.08
50	-0.09	314	-3.53	-0.08	314	-2.46
20	-0.31	663	-3.45	-0.27	662	-1.53
10	-0.19	1261	-7.09	-0.12	1251	-5.21
5	-0.26	2525	-6.55	-0.13	2516	-7.41

Table 4. Estimating Total Employment With a Two-PSU Per Design and a Post-stratified Estimator

Second-Phase Stratum	*Reweighted Expansion Estimator*			*Double Expansion Estimator*		
Sample Size	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife	% RelBias of Estimate	Scaled MSE*	RelBias of Jackknife
All	0.06	33	3.30	0.06	33	3.30
50	-0.08	117	4.88	-0.05	122	28.5
20	-0.93	273	6.42	-0.71	284	32.0
10	-1.96	522	12.03	-1.67	541	35.3
5	-4.44	1101	9.20	-3.98	1141	22.4

All results based on 4,000 simulations.

* Scaled so that *original* estimator based on the full first-phase sample has Variance (MSE) equal to 100.

6. DISCUSSION

Kott and Stukel (1997) shows that the jackknife provides an asymptotically unbiased estimator for the randomization mean squared error of the two-phase design-consistent regression estimator, at least in projection form, when, 1, the second phase of sampling is stratified, simple random sampling at the element level, and 2, the covariates in the regression estimator include indicator variables for all the second-phase strata.

By invoking the usual linear model, we saw here that even without those two restrictive conditions, the jackknife provides an asymptotically unbiased estimator of what we called the randomization-model variance; that is, the model variance of the two-phase design-consistent regression estimator plus the randomization expectation of its squared model bias. The lone non-asymptotic restriction needed is that the element errors be uncorrelated across PSU's.

The randomization-model variance of a two-phase design-consistent regression estimator is related to more conventional measures of accuracy in the following manner. The randomization expectation of the randomization-model variance is the randomization expectation of the model mean squared error or, equivalently, the model expectation of the randomization mean squared error.

The subject of the variance of the jackknife has not yet been addressed. Employing equation (7), the jackknife variance estimator for the two-phase regression estimator can be expressed as

$$\begin{aligned} v_J(t) &= \sum^H \sum_j \ (n_n/[n_h - 1])\{(\mathbf{x}_{1hj}\beta - \sum_g x_{1hg}\beta/n_h) + (a_{hj} - \sum_g a_{hg}/n_h)\}^2 \\ &= \sum^H \sum_j \ (n_n/[n_h - 1])(\mathbf{x}_{1hj}\beta - \sum_g x_{1hg}\beta/n_h)^2 \\ &\quad + \sum^H \sum_j \ (n_n/[n_h - 1])(a_{hj} - \sum_g a_{hg}/n_h)^2 \qquad (12) \\ &\quad + 2 \sum^H \sum_j \ (n_n/[n_h - 1])(\mathbf{x}_{1hj}\beta - \sum_g x_{1hg}\beta/n_h) (a_{hj} - \sum_g a_{hg}/n_h). \end{aligned}$$

Let up call the expressions in the last three lines of equation (12), a, B, and 2C, respectively . Even allowing the simplifying assumptions that E(AC) = E(BC) = Cov(a, B) = 0 and Var(C) = E(AB), we have $\mathrm{Var}[v_J(t)] = \mathrm{Var}(a) + \mathrm{Var}(B) + 6\mathrm{Var}(C)$. This expression is difficult to analyze without specifying a model for $\mathbf{x}_i$, which is beyond the scope of this analysis.

Recall that with the Stukel and Kott data, the randomization-model variance of employment rate estimator virtually had no randomization component. As a result, the variance of the jackknife is, for all intents and purposes, the variance of B. Since the relative variance of B tends to increase as the second-phase sample size decreases, it is not surprising that Stukel and Kott found that the coefficient of variation (CV) of the jackknife of the ratio of two reweighted expansion estimators increased as the second-phase sample size per stratum decreased. With 50 individuals per stratum the CV was 59.2%, with 20 individuals 65.7%, with 10 individuals 74.2%, and with 5 individuals per stratum 103.1%. The number are quite similar for the ratio of two double expansion estimators and for the set of simulations with 72 PSU's.

One surprising result in Stukel and Kott was the 78.4% CV for the jackknife when the first-phase sample was completely enumerated. This may be due to the impact on variance estimation of with-replacement first-phase sampling. If the same PSU was selected twice in a stratum, that stratum only contributed to the variance estimate when the first-phase sample was *not* completely enumerated. This is likely the cause of the relative instability of the jackknife for the fully-enumerated first-phase sample. When Stukel increased the number of PSU's to 72, the CV shrunk to 30.4% for the fully-enumerated first-phase sample, while staying in the same general neighorhood for the other sub-sample sizes.

This begs the question: if Stuckel's simulations contained double-hits of some PSU's, why do her results have any relevance to the analysis in the text, which assumed away such a possiblity? The reason they are relevant is because the likelihood of double-hits at the *element* level for the subsampling simulations is very small. Hence, the model-based portion of the analysis in the text applies as long as we assume independence (or near independence) of the element errors, the $e_{i,}$ across elements in the same PSU.

It is a simple matter to extend the analysis in the text to an estimator similar to t in equation (1) but with a more complex covariate mean estimator than $\mathbf{x}_1 = \sum_S w_{1i}\mathbf{x}_i/M$. For example, the estimator $\mathbf{x}_1$ may itself be a multi-phase design-consistent estimator. It may also incorporate auxiliary variables whose population means are known and do not have to be estimated. For example, a better use of the post-stratum population sizes in the Stukel and Kott data would have been in the estimation of M$\mathbf{x}_i$, the vector of age group population totals (i.e., second-phase strata) based on the first-phase sample.

The randomization variance of $\mathbf{x}_1\beta$, a key component of the randomization-model variance of t, may itself be replaced by a randomization-model variance or even by a model variance; for example, by assuming the M$\mathbf{x}_i$ are have a common mean within first-phase strata and are uncorrelated across PSU's.

REFERENCES

Isaki, C.T. and Fuller, W.A. (1982). Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89-96.

Kott, P.S. (1990). Variance Estimation when a First Phase Area Sample is Restratified. *Survey Methodology*, 99-104.

Kott, P.S. and Stukel (1997). Can the Jackknife Be Used With a Two-Phase Sample? Submitted to *Survey Methodology*.

Rao, J.N.K, and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation. *Biometrika*, 811-822.

Rust, Keith (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1, 381-397.

Sarndal, C.E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). Methodology of the Canadian Labour Force Survey: 1984-1990. *Statistics Canada publication*, Catalogue 71-526.

Stukel, D.M. and Boyer, R. (1992). Calibration Estimation: An Application to the Canadian Labour Force Survey. *Statistics Canada Branch Working Paper*, SSMD # 009E.

Stukel, D.M. and Kott, P.S. (1997). Jackknife Variance Estimation Under Two-Phase Sampling: An Empirical Investigation. *Statistics Canada Branch Working Paper*, HSMD - 97- # 009E.