Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Calibration alternatives to poststratification for doubly classified data

by Ted Chang

June 2012

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 31-41
Statistics Canada, Catalogue No. 12-001-X

Calibration alternatives to poststratification for doubly classified data

Ted Chang [1]

Abstract

We consider alternatives to poststratification for doubly classified data in which at least one of the two-way cells is too small to allow the poststratification based upon this double classification. In our study data set, the expected count in the smallest cell is 0.36. One approach is simply to collapse cells. This is likely, however, to destroy the double classification structure. Our alternative approaches allows one to maintain the original double classification of the data. The approaches are based upon the calibration study by Chang and Kott (2008). We choose weight adjustments dependent upon the marginal classifications (but not full cross classification) to minimize an objective function of the differences between the population counts of the two way cells and their sample estimates. In the terminology of Chang and Kott (2008), if the row and column classifications have *I* and *J* cells respectively, this results in *IJ benchmark* variables and $I + J - 1$ *model* variables. We study the performance of these estimators by constructing simulation simple random samples from the 2005 Quarterly Census of Employment and Wages which is maintained by the Bureau of Labor Statistics. We use the double classification of state and industry group. In our study, the calibration approaches introduced an asymptotically trivial bias, but reduced the MSE, compared to the unbiased estimator, by as much as 20% for a small sample.

Key Words: Calibration; Poststratification; Prediction model.

1. Introduction

Suppose we have a population $\mathcal{U}$ which is doubly stratified by two categorical variables whose indices are denoted $(i, j), i = 1, \ldots, I, j = 1, \ldots, J$ and write $\mathcal{U}_{ij}$ for the (i, j)-stratum. If a simple random sample $\mathcal{S}$ of size n is taken and if y denotes the variable of interest a natural estimator for the total $T_y = \sum_{k \in \mathcal{U}} y_k$ is the poststratified estimator

$$\hat{t}_{y\text{PS}} = \sum_{i,\, j} N_{ij}\, \bar{y}_{ij} \qquad (1)$$

where N_{ij} is the size of $\mathcal{U}_{ij}$ and $\bar{y}_{ij}$ is the sample mean of y over $\mathcal{S} \cap \mathcal{U}_{ij}$. This estimator is widely used as long as all the sample sizes n_{ij} of $\mathcal{S} \cap \mathcal{U}_{ij}$ are reasonably large.

What to do if some of the n_{ij} are small, or even zero?

The standard approach would be to collapse some of the cells until all the n_{ij} are big enough. However such a collapsing might not be possible in a way that maintains the double classification scheme: that is the indices j might depend upon i.

The poststratified estimator $\hat{t}_{y\text{PS}}$ is a special case of a calibration estimator. Define for each $k \in \mathcal{U}$ the $I \times J$ vector variable $\mathbf{x}_k = (x_{11k}, \ldots, x_{IJk})^T$ where $x_{ijk} = 1$ if $k \in \mathcal{U}_{ij}$ and $x_{ijk} = 0$ otherwise. The population total $T_{\mathbf{x}}$ of $\mathbf{x}$ is $(N_{11}, \ldots, N_{IJ})^T$ and letting $d_k = N/n$ be the sampling weight and $\beta = (N^{-1} n_{11}^{-1} N_{11} n, \ldots, N^{-1} n_{IJ}^{-1} N_{IJ} n)^T$

$$\hat{t}_{y\text{PS}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta)\, y_k$$

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta)\, \mathbf{x}_k.$$

These equations establish that if the *benchmark* variables $\mathbf{x}$ are used, then $\hat{t}_{y\text{PS}}$ is the resulting calibrated estimator of T_y.

Chang and Kott (2008) derived the asymptotic properties of a calibrated estimate of the form

$$\hat{t}_{y,\, \mathbf{z}\, f\, \mathbf{V}} = \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \hat{\beta})\, y_k \qquad (2)$$

where $\hat{\beta}$ minimizes an objective function of the form

$$Q(\beta) = \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k \right)^T \mathbf{V}^{-1} \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k \right). \qquad (3)$$

In equations (2) and (3), $\mathbf{z}$ is a vector of *model* variables whose length Q is at most the length P of the *benchmark* variables $\mathbf{x}$, f is a positive real valued function which Chang and Kott (2008) calls the *back link* function, and $\mathbf{V}$ is some positive definite symmetric $P \times P$ matrix. $\mathbf{V}$ is allowed to depend upon β as would occur if $\mathbf{V}(\beta)$ is some measurement of the variability of $\sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k$.

In Chang and Kott (2008), the realized sample $\mathcal{S}$ is the respondents from an original sample with sampling weights d_k. The respondent sample $\mathcal{S}$ is assumed to be a Poisson subsample of the original sample with Poisson probabilities $f(\mathbf{z}_k^T \beta_0)^{-1}$, for some β_0. The asymptotic formulas derived there were under an asymptotic framework for this quasi-randomization (design based) model. We use the term *quasi-randomization* to remind ourselves that the assumed Poisson response mechanism is actually model based.

It should be noted that the use of calibration to correct for nonresponse goes back to Fuller, Loughin and Baker (1994), at least when $\mathbf{z} = \mathbf{x}$ and $f(\eta) = 1 + \eta$.

1. Ted Chang, University of Virginia, Department of Statistics, Charlottesville, VA, U.S.A. E-mail: tcc8v@virginia.edu.

We propose to use the Chang and Kott (2008) methodology with $\mathbf{x}$ remaining as indicator variables for the complete $I \times J$ cross classification but letting $\mathbf{z}$ be a vector of $I + J - 1$ indicator variables for the marginal classifications. In other words, we propose to rebalance the sample to come as close as possible, in the sense of minimizing (3), to the correct cell proportions in the complete cross classification, but requiring the rebalancing weights to depend only upon the marginal classifications.

The Chang and Kott (2008) framework applies in the presence of nonresponse (and/or noncoverage) if $f(\mathbf{z}_k^T \beta_0)^{-1}$ is the response (or combined response and coverage) probability. We note that poststratification, a special case of calibration, is often used for the purpose of nonresponse/ noncoverage correction. In our test example below, there is no nonresponse or noncoverage to correct for, and hence, the Chang and Kott (2008) framework applies with $\beta_0 = \mathbf{0}$ for any f with $f(0) = 1$. In other words, if the calibration is used solely for the purposes of sample rebalancing, we can use Chang and Kott (2008) with almost any f. But if we are trying to correct for nonresponse and/or noncoverage, stronger assumptions are required.

It should be noted that raking is simply the calibrated estimate using the $I + J - 1$ indicator variables of the marginal classifications as both benchmark and model variables and using $f(\eta) = e^{\eta}$. Thus we will also explore the use of this back link function.

Section 2 gives the precise formulas for the estimators we will use in this study. Chang and Kott (2008) can be applied to derive sample based variance estimators and these derivations are given in the Appendix.

In Section 3, we give the results on an empirical study using the 2005 first quarter Quarterly Census of Employment and Wages, collected by the Bureau of Labor Statistics. We will restrict ourselves to the five states which we will denote by A, B, C, D, E and to five industry groupings denoted by 1, 2, 3, 4, 5. We will not further identify either the states or the industry groupings to prevent identification of the outlier in the discussion below. This population has 283,725 firms. From this population we will take Monte Carlo simple random samples of size $n =$ 200, 1,000, 5,000 and use the double classification of state and industry group.

It should be noted that 0.18% of the population has the double classification of state E and industry grouping 5. Thus when $n = 200$, the expected sample size in this cell is 0.36 and poststratification using the double classification is out of the question.

Kott and Chang (2010) derives the properties of $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ using a model based framework. The models considered there do not apply with our selection of $\mathbf{x}$ and $\mathbf{z}$ variables. However, motivated by their approach, we examine in Section 4 the behavior of the estimator $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ defined by equation (2), under highly simplified assumptions, including that $f(\eta) = 1 + \eta$. This leads in Section 5 to the choice of a new weight matrix $\mathbf{V}^{-1}$ for use in (3). We then continue with our empirical exploration using this new estimator.

2. Mathematical formulas

In this section we list the formulas used in this study. They are all special cases of formulas in Chang and Kott (2008). We assume that a simple random sample of size n is taken from a population of size N and we use $\mathcal{S}$ and r to denote the respondents from that sample and the size of $\mathcal{S}$. We assume that the calibration weight function has a β_0 such that $f(\mathbf{z}^T \beta_0)^{-1}$ is the response probability for an element with model variables $\mathbf{z}$. In particular, and without loss of generality, if there is no nonresponse problem, we assume $f(0) = 1$.

The same formulas work with noncoverage, in which case $f(\mathbf{z}^T\beta_0)^{-1}$ is the combined response/coverage probability.

We denote N_{ij}, $\mathcal{S}_{ij}$, and r_{ij} to be the population size, respondent sample, and respondent sample size in classification (i, j). Although N_{ij} is assumed known, our methodology does not require the knowledge of the row and column classifications of nonrespondents.

We define $N_{i\cdot} = \sum_j N_{ij}$ and analogously define $N_{\cdot j}$.

We will use estimators for a total T_y of the form

$$\hat{t}_y = \frac{N}{n} \sum_i \sum_j \sum_{k \in \mathcal{S}_{ij}} w_{ij}\, y_k \qquad (4)$$

where the adjustment weights w_{ij} are defined as below. These are all special cases of equations (2) and (3) when we use $\mathbf{V} = \mathbf{I}$.

The *calibrated margins* estimator uses $f(\eta) = 1 + \eta$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. In this case $T_{\mathbf{x}}$ is a vector of $N_{i\cdot}$ and $N_{\cdot j}$. The adjustment weights $f(\mathbf{z}_k^T \beta)$ have the form $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ when $\mathbf{z}$ is the vector of indicator variables for membership in the i^{th} and j^{th} row and column classifications respectively. Since the number of equations (the dimension of $\mathbf{x}$) equals the number of unknowns (the dimension of $\hat{\beta}$), we expect to be able to solve the equations

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k \qquad (5)$$

exactly. Thus $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the linear equations of rank $I + J - 1$

$$N_{i\cdot} = \frac{N}{n} \sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij}$$

$$N_{\cdot j} = \frac{N}{n} \sum_i (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})\, r_{ij},$$

which easily follows from (5).

The *calibrated cell counts* estimator uses $f(\eta) = 1 + \eta$ and defines $\mathbf{x}$ to be the IJ indicator variables for the complete cross classification and $\mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. In this case $T_{\mathbf{x}}$ is a vector of N_{ij} and, since $\mathbf{V} = \mathbf{I}$, the adjustment weights $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2 .$$

The *raking* estimator uses $f(\eta) = e^{\eta}$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights are $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ where $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the $I + J$ equations

$$N_{i\cdot} = \frac{N}{n} \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$

$$N_{\cdot j} = \frac{N}{n} \sum_i \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}.$$

Since $\sum_i N_{i\cdot} = N = \sum_j N_{\cdot j}$, these $I + J$ equations yield only $I + J - 1$ constraints. It should be noted, however, that if a constant c is added to each $\hat{\beta}_{i\cdot}$ and subtracted from each $\hat{\beta}_{\cdot j}$, the w_{ij} are not changed.

The *exponential calibrated cell counts* estimator uses $f(\eta) = e^{\eta}$ and defines $\mathbf{x}$ to be the IJ indicator variables for the complete cross classification and $\mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2 .$$

Chang and Kott (2008) give formulas for sample based estimation of the variance of $\hat{t}_y$. In the appendix, we apply these formulas to the four estimators above.

3. Empirical study

The population we use here is the data from the 2005 first quarter Quarterly Census of Employment and Wages (QCEW), restricted to five states and five industry groupings. The QCEW is compiled from mandatory reports to state employment offices and hence is virtually a census and the data we used is the complete QCEW for these five states and five industry groupings. This population has N = 283,725 firms, divided as in Table 1.

The response variables y are total employment and total (quarterly) wages. For these variables T_y = 2,981,364 for total employment and T_y = 2,334,400 (in tens of thousands of dollars) for total wages. In this study, we took 10,000 samples of sizes n = 200, 1,000, 5,000. For each of the 4 estimators, we report the estimated bias, standard error, and root mean square error. We also report square root of the mean of the estimated variances using the first term of equation (15). For purposes of comparison, we report the theoretical and empirical values for the unweighted estimator $N/n \sum_{k \in \mathcal{S}} y_k$. These results are reported in Table 2 for total employment and Table 3 for total wages.

For sample size n = 5,000, the expected sample size in the smallest cell (state E and industry group 5) is 9.07. While this might be a little small for poststratification, the probability that this cell has a sample size less than 2, the minimum size necessary for variance estimation, is 0.0011. In our simulations 9 runs had a cell with sample size less than 2. For this sample size, we also report the empirical behavior of poststratified estimator, excluding the 9 problem cases, using the variance estimate (7.6.5) of Särndal, Swensson and Wretman (1992) and its theoretical behavior using the variance approximation given in (7.6.6) of Särndal *et al.* (1992).

Table 1
Business entities by state and industry group

	industry group					
	1	**2**	**3**	**4**	**5**	**sum**
A	5,986	5,548	7,712	3,969	1,299	24,514
	(2.11%)	(1.96%)	(2.72%)	(1.40%)	(0.46%)	(8.64%)
B	18,782	31,572	22,012	4,982	4,504	81,852
	(6.62%)	(11.13%)	(7.76%)	(1.76%)	(1.59%)	(28.85%)
C	13,518	13,099	17,837	5,610	3,001	53,065
	(4.76%)	(4.62%)	(6.29%)	(1.98%)	(1.06%)	(18.70%)
D	30,428	36,017	32,541	10,963	5,399	115,348
	(10.72%)	(12.69%)	(11.47%)	(3.86%)	(1.90%)	(40.65%)
E	2,225	2,020	3,110	1,076	515	8,946
	(0.78%)	(0.71%)	(1.10%)	(0.38%)	(0.18%)	(3.15%)
sum	70,939	88,256	83,212	26,600	14,718	283,725
	(25.00%)	(31.11%)	(29.33%)	(9.38%)	(5.19%)	

Table 2
Empirical comparison of 4 estimators of total employment

estimator	bias	st. err.	rt. MSE	rt. est. var.
	n = 200			
unweighted (theoretical)	0	1,113,220		
unweighted (empirical)	-1,280	1,068,944	1,068,945	1,059,463
cal. margins	-1,394	1,105,201	1,105,201	1,048,873
cal. cell cts.	-218,751	1,008,436	1,031,889	975,140
raking	-462	1,103,172	1,103,172	1,041,490
exp. cal. cell cts.	-227,578	1,000,154	1,025,719	962,153
	n = 1,000			
unweighted (theoretical)	0	497,144		
unweighted (empirical)	-5,435	505,941	505,970	501,144
cal. margins	-6,212	506,239	506,277	498,946
cal. cell cts.	-56,118	493,611	496,790	488,222
raking	-4,854	507,938	507,961	499,237
exp. cal. cell cts.	-58,891	492,939	496,445	487,281
	n = 5,000			
unweighted (theoretical)	0	220,751		
unweighted (empirical)	1,516	224,088	224,093	222,034
poststr. (theoretical)	0	220,315		
poststr. (empirical, 9 cases excluded)	1,234	223,225	223,228	221,094
cal. margins	1,649	223,091	223,098	220,833
cal. cell cts.	-8,606	222,170	222,337	220,347
raking	3,632	236,355	236,383	220,606
exp. cal. cell cts.	-10,643	223,472	223,725	220,207

Table 3
Empirical comparison of 4 estimators of total wages (tens of thousands of dollars)

estimator	bias	st. err.	rt. MSE	rt. est. var.
	n = 200			
unweighted (theoretical)	0	1,682,571		
unweighted (empirical)	-11,119	1,551,186	1,551,226	1,543,483
cal. margins	-11,474	1,582,383	1,582,425	1,510,413
cal. cell cts.	-214,323	1,451,931	1,467,664	1,413,411
raking	-11,220	1,579,842	1,579,882	1,501,170
exp. cal. cell cts.	-221,435	1,438,810	1,455,750	1,393,246
	n = 1,000			
unweighted (theoretical)	0	751,406		
unweighted (empirical)	-2,911	772,495	772,501	768,878
cal. margins	-4,372	776,955	776,968	768,869
cal. cell cts.	-51,649	756,201	757,963	751,384
raking	-4,684	778,302	778,316	769,428
exp. cal. cell cts.	-54,305	754,963	756,913	749,832
	n = 5,000			
unweighted (theoretical)	0	333,654		
unweighted (empirical)	2,678	336,057	336,068	337,239
poststr. (theoretical)	0	333,765		
poststr. (empirical, 9 cases excluded)	1,802	335,271	335,276	336,192
cal. margins	2,510	334,910	334,920	336,064
cal. cell cts.	-7,149	333,560	333,637	335,006
raking	-4,679	339,074	339,106	335,230
exp. cal. cell cts.	-9,251	334,365	334,493	334,755

The response variables, total employment and total wages, are strongly skewed right. There is one firm (in state C and industry group 4) whose total employment is more than double the total employment of the next largest firm and many hundreds times the mean employment of the remaining firms. We repeat this study using a population with this firm removed. The results are presented in Tables 4 and 5. In practice with this population, the sampling would normally sample this firm with certainty (a *self representing unit*) and samples constructed from the

remaining firms. Thus Tables 4 and 5 are perhaps more indicative of the relative performance of these estimators in actual practice.

The samples used for Tables 4 and 5 are identical to those used for Tables 2 and 3 except that if the outlier was included in the sample, it was replaced by a new observation from the population. This was done to improve the comparability of the results of Tables 4 and 5 with those of Tables 2 and 3.

Table 4
Empirical comparison of 4 estimators of total employment: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
		n = 200		
unweighted (theoretical)	0	950,688		
unweighted (empirical)	5,395	975,617	975,632	965,448
cal. margins	5,777	1,019,583	1,019,599	963,314
cal. cell cts.	-211,568	909,070	933,365	877,343
raking	6,688	1,018,383	1,018,405	956,867
exp. cal. cell cts.	-217,810	902,756	928,660	868,797
		n = 1,000		
unweighted (theoretical)	0	424,552		
unweighted (empirical)	-8,393	422,116	422,199	414,019
cal. margins	-9,430	418,153	418,259	408,577
cal. cell cts.	-58,808	408,391	412,603	399,961
raking	-8,135	419,938	420,016	408,611
exp. cal. cell cts.	-61,014	407,780	412,320	399,311
		n = 5,000		
unweighted (theoretical)	0	188,517		
unweighted (empirical)	702	191,631	191,632	188,089
poststr. (theoretical)	0	187,691		
poststr. (empirical, 9 cases excluded)	563	190,854	190,855	187,180
cal. margins	820	190,662	190,664	186,664
cal. cell cts.	-9,376	189,884	190,115	186,202
raking	2,933	205,924	205,944	186,618
exp. cal. cell cts.	-9,922	189,813	190,072	186,140

Table 5
Empirical comparison of 4 estimators of total wages: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
		n = 200		
unweighted (theoretical)	0	1,330,930		
unweighted (empirical)	711	1,341,900	1,341,901	1,334,556
cal. margins	1,256	1,387,484	1,387,485	1,318,285
cal. cell cts.	-201,575	1,225,852	1,242,314	1,194,071
raking	1,473	1,386,978	1,386,979	1,311,353
exp. cal. cell cts.	-206,956	1,217,881	1,235,340	1,184,166
		n = 1,000		
unweighted (theoretical)	0	594,370		
unweighted (empirical)	-8,169	587,775	587,832	582,524
cal. margins	-10,093	583,606	583,693	576,251
cal. cell cts.	-56,429	569,158	571,948	563,022
raking	-10,529	584,532	584,626	576,282
exp. cal. cell cts.	-58,435	568,277	571,273	562,061
		n = 5,000		
unweighted (theoretical)	0	263,923		
unweighted (empirical)	1,185	266,779	266,782	264,110
poststr. (theoretical)	0	263,339		
poststr. (empirical, 9 cases excluded)	566	265,973	265,973	263,210
cal. margins	991	265,449	265,451	262,556
cal. cell cts.	-8,565	264,126	264,265	261,483
raking	-6,008	271,535	271,602	262,021
exp. cal. cell cts.	-9,070	264,038	264,194	261,394

Examining Tables 2 and 3, we see that the $P > Q$ methods, that is those that calibrate the cross classified cell counts using calibration weights which depend upon the marginal classifications, are clearly more biased than the other techniques. However the biases of these estimators relative to their standard deviations decrease with increasing sample sizes. We will show in the next section, that under a highly simplified model, the bias has order n^{-1} and the standard deviation has order $n^{-1/2}$. Consider, for example, the results for the "calibrated cell counts" estimator in Table 2. In this case, the bias divided by the standard error is 0.217, 0.114, 0.039 for $n = 200, 1{,}000, 5{,}000$ respectively. For these values of n, the values of $n^{-1/2}$ are 0.0707, 0.0316, 0.0141 and it appears that the former series of three ratios is approximately 3 times the latter series.

It also appears that the exponential back link function f performs slightly better than the linear choice for f. Computationally the former is much more expensive than the latter. We also notice that as the sample sizes increase, the estimators' performances appear to converge. This is to be expected: because there is no nonresponse, as $n \to \infty$, $\hat{\beta} \to 0$, so that the adjustment weights $w = f(\mathbf{z}^T \hat{\beta}) \to 1$.

Comparing the linear calibrated cell counts estimator to the empirical values of the unweighted estimator, the former is approximately 7.3% more efficient in MSE when $n = 200$ for total employment and 11.7% more efficient for total wages. (This means, for example, that the empirical MSE of the unweighted estimator is 1.117 times the empirical MSE for the linear calibrated cell counts estimator when estimating total wages.) For the exponential calibrated cell counts estimator, the improvement in efficiency relative to the empirical MSE of the unweighted estimator is 8.6% for total employment and 13.5% for total wages. Comparison to the theoretical values for the unweighted estimator would be more favorable to the calibrated cell counts estimators, but we will use the empirical results for the unweighted estimator as the various estimators have all used the same Monte Carlo samples. The calibrated cell counts estimator and exponential calibrated cell counts estimator still have an advantage in MSE over the unbiased estimator at sample size $n = 1{,}000$.

When the single extreme outlier is removed, leaving 283,724 remaining elements of the population, the calibrated cell count estimators have somewhat better performance relative to the unweighted estimator. For $n = 200$, the linear calibrated cell count estimator offers a 9.3% improvement in efficiency for total employment and a 16.7% improvement for total wages. The comparable ratios for the exponential calibrated cell count estimator are 10.4% for total employment and 18.0% for total wages.

Finally, the variance estimator in equation (15) has a slight downward bias.

4. Model based bias and variance of calibrated estimators

Kott and Chang (2010) derived the asymptotic properties of $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ under a different, model-based, probability structure. In that paper $\mathcal{S}$ is a sample selected with selection probabilities d_k^{-1} so that nonresponse is not an issue in $\mathcal{S}$. Rather, if P the number of benchmark variables $\mathbf{x}$ equals Q the number of model variables $\mathbf{z}$, Kott and Chang (2010) assume a *prediction* model

$$y_k = \mathbf{x}_k^T \theta + \varepsilon_k, k \in \mathcal{U}. \quad (6)$$

Here θ is a unknown fixed vector, ε_k are model independent errors subject to

$$E(\varepsilon_k \mid \mathbf{z}_j, I_j, j \in \mathcal{U}) = 0, \quad (7)$$

and I_k is a random variable defined by $I_k = 1$ if $k \in \mathcal{S}$ and $I_k = 0$ otherwise.

When $P > Q$, the model equation (6) must be replaced by

$$y_k = (\mathbf{A}_\infty \mathbf{x}_k)^T \theta + \varepsilon_k, k \in \mathcal{U} \quad (8)$$

for some limiting $Q \times P$ matrix $\mathbf{A}_\infty$ (which is defined in a suitable asymptotic framework, see Kott and Chang (2010)).

Thus when $\mathbf{x}$ represents indicator variables for the complete $I \times J$ cross classification, we have that $\mathbf{x}_k^T\theta$, for k in the $(i, j)^{\text{th}}$ classification, is the mean value of the response variable over the $(i, j)^{\text{th}}$ classification. Hence, by definition, $E(\varepsilon_k \mid \mathbf{x}_j, j \in \mathcal{U}) = 0$ and, since $\mathbf{z}$ is a function of $\mathbf{x}$, the model (6) and (7) automatically holds when the sampling (including nonresponse) is noninformative.

However, in our application of calibration, $P = IJ > Q = I + J - 1$ and the model equation (8) has no a priori reason to hold.

Motivated by Kott and Chang (2010) we examine the behavior of calibrated estimates under the following scenario:

1. The benchmark variables $\mathbf{x}$ are indicator variables for some partition of the population into classes $\mathcal{C}_r$. The model (6) automatically holds where the r^{th} component of θ is the population mean of $\mathcal{C}_r$. Let f_r denote the proportion of the population in $\mathcal{C}_r$ and $V_r = \text{Var}(\varepsilon_k \mid k \in \mathcal{C}_r)$. We shall also use the notation $\text{Var}(\mathbf{x}_k)$ for V_r when $k \in \mathcal{C}_r$.
2. The sample is a simple random sample of size n chosen *with replacement*.
3. The back link function $f(\eta)$ in the estimator $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ of equation (2) is $f(\eta) = 1 + \eta$.

Although these assumptions are unrealistic in practice, the main purpose of this section is to heuristically justify a choice, given in the next section, for the matrix $\mathbf{V}$. At this point, we no longer place any requirements on $\mathbf{z}$.

We note that in this situation $E(\varepsilon_k \mid \mathbf{x}_j, I_j, j \in \mathcal{U}) = 0$. Note that (7) will hold if the components of the model variables $\mathbf{z}$ are functions of $\mathbf{x}$, that is each component of $\mathbf{z}$ is constant on each class. However if $P > Q$, (8) will generally not hold. In any event, in this section we require neither (7) nor (8).

We let

$$\mu_{\mathbf{x}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j$$

$$\mu_{\mathbf{xz}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j \mathbf{z}_j^T$$

and the matrix $\mathbf{A}_\infty$ of equation (8) becomes

$$\mathbf{A}_\infty = \mathbf{V}^{-1}\mu_{\mathbf{xz}}.$$

Let $\hat{\mu}_{y,\mathbf{zV}} = N^{-1}\hat{t}_{y,\mathbf{zV}}$ where $\hat{t}_{y,\mathbf{zV}}$ is defined as in (2). We have suppressed the f in the notation $\hat{t}_{y,\mathbf{zV}}$ because, in this section, $f(\eta) = 1 + \eta$. Letting $\bar{y}_s$ and $\bar{\mathbf{x}}_s$ denote the indicated sample means and using Kott and Chang (2010)

$$\hat{\beta} = \left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j \mathbf{x}_j^T \mathbf{A}_\infty\right)^{-1}\left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j y_j\right) + O_p(n^{-1/2})$$

$$\begin{aligned}\hat{\mu}_{y,\mathbf{zV}} &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V}\mathbf{A}_\infty)^{-1}\left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j y_j\right) + O_p(n^{-1}) \\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V}\mathbf{A}_\infty)^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1}) \\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T \mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1}). \quad (9)\end{aligned}$$

If $\hat{\mu}_{y,\mathbf{zV}}$ is bounded, as would occur if $f(\eta) = 1 + \eta$ were modified for large η to prevent large calibration weight adjustments, we would have

$$\begin{aligned}\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}) &= \mathrm{E}(\bar{y}_s) + O(n^{-1}) = \mu_y + O(n^{-1}) \\ \mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s) &= \bar{\mathbf{x}}_s^T\theta + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T \mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta \\ &\quad + O(n^{-1}) \\ \mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] &= \frac{1}{n}\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\mathbf{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta \\ &\quad + o(n^{-1}),\end{aligned}$$

where $\mathbf{\Sigma}_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$ and

$$\mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}} = \mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T \mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T \mathbf{V}^{-1}. \quad (10)$$

Now

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s) &= \mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s) + o(n^{-1}) \\ &= \frac{1}{n^2}\sum_{j\in\mathcal{S}} V(\mathbf{x}_j) + o(n^{-1}) \\ \mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] &= \mathrm{E}[\mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] + o(n^{-1}) \\ &= \frac{1}{n}\sum_r f_r V_r + o(n^{-1}).\end{aligned}$$

It is easily seen that

$$\mathrm{Var}[\mathrm{E}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] = \frac{1}{n}\theta^T \mathbf{\Sigma}_{\mathbf{x}}\theta.$$

Since $\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) = \mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] + \mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)]$ and similarly for $\mathrm{Var}(\bar{y}_s)$, $\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) < \mathrm{Var}(\bar{y}_s)$ to terms $o(n^{-1})$ when

$$\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\mathbf{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta < \theta^T\mathbf{\Sigma}_{\mathbf{x}}\theta. \quad (11)$$

The derivation also establishes that the square bias has an asymptotically trivial contribution to the mean square error of $\hat{\mu}_{y,\mathbf{zV}}$.

5. A proposed new weight matrix $\mathbf{V}^{-1}$

In this section we return to our original benchmark $\mathbf{x}$ and model $\mathbf{z}$ variables. When $\mathbf{V} = \mathbf{I}$, the identity matrix, we see from (10) that $\mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{I}}^T\theta = \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{I}}\theta$ is the projection of θ onto the span of the columns of $\mu_{\mathbf{xz}}$. The left hand side of (11) will be zero if θ is in this column span.

For simplicity, we will write $\mu_{\mathbf{xz}}$ as a singular matrix, of rank $I + J - 1$, with one row for each possible double classification cell (i, j) and one column for each row classification i and each column classification j. Thus, the $(i, j)^{\text{th}}$ row of $\mu_{\mathbf{xz}}$ has $f_{ij} = N_{ij}/N$ in the columns corresponding to i and j and zero elsewhere. Thus θ will be in the column span of $\mu_{\mathbf{xz}}$ if and only if for each i and j

$$\frac{\theta_{ij}}{f_{ij}} = \alpha_i + \beta_j \quad (12)$$

for some α_i and β_j. In other words, the θ_{ij}/f_{ij} satisfy a two way ANOVA model, without interaction, in the column and row classifications.

Recalling that θ_{ij} represents the mean value of the variable of interest y in the $(i, j)^{\text{th}}$ cell, (12) does not appear to be a very promising approximation to the truth. A more likely approximation would be the usual two way ANOVA model

$$\theta_{ij} = \alpha_i + \beta_j. \quad (13)$$

Suppose we change variables $\tilde{\mathbf{x}} = \mathbf{C}\mathbf{x}$ for some diagonal matrix $\mathbf{C}$. Note that the rows and columns of $\mathbf{C}$ are doubly indexed by (i, j) and we will let c_{ij} denote the diagonal entry in the $(i, j)^{\text{th}}$ row and column. Let $\tilde{\theta} = \tilde{\mathbf{C}}^{-1}\theta$ so that model (6) can be rewritten as

$$y_k = \tilde{\mathbf{x}}_k^T\tilde{\theta} + \varepsilon_k.$$

Now the matrix $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ has $c_{ij}f_{ij}$ in the $(i, j)^{\text{th}}$ row and the columns corresponding to i and j. Now $\tilde{\theta}$ will be in the column span of $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ if and only if

$$c_{ij}^{-1}\theta_{ij} = \tilde{\theta}_{ij} = c_{ij} f_{ij}(\alpha_i + \beta_j).$$

Thus (13) is equivalent to $c_{ij} = f_{ij}^{-1/2}$. It is easily checked that

$$\tilde{\theta}^T(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}',\mathbf{I}}})\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}',\mathbf{I}}})\tilde{\theta} = \theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{x}\mathbf{z}',\mathbf{V}}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{x}\mathbf{z}',\mathbf{V}}})^T\theta$$

when $\mathbf{V} = \mathbf{C}^{-2}$. We thus propose using the diagonal matrix $\mathbf{V}_o$ whose diagonal entries are f_{ij}.

With this choice of $\mathbf{V}_o$, equation (9) suggests the estimator for simple random sampling

$$\hat{t}_{y,\mathbf{z}\mathbf{V}_o} = N\bar{y}_s + (T_{\mathbf{x}} - N\bar{x}_s)^T \mathbf{V}_{\mathbf{o}}^{-1}\hat{\mu}_{\mathbf{xz}}(\hat{\mu}_{\mathbf{xz}}^T\mathbf{V}_o^{-1}\hat{\mu}_{\mathbf{xz}})^{-1}\left(\frac{1}{n}\sum_{k\in\mathcal{S}}\mathbf{z}_k y_k\right) \quad (14)$$

where $\hat{\mu}_{\mathbf{xz}} = n^{-1}\sum_{k\in\mathcal{S}}\mathbf{x}_k\mathbf{z}_k^T$. In our case both $\mu_{\mathbf{xz}}$ and $\mu_{\mathbf{x}}$ are known from the N_{ij}, but in the spirit of ratio estimation it is preferable to use $\hat{\mu}_{\mathbf{xz}}$ in place of $\mu_{\mathbf{xz}}$. This heuristic observation has been demonstrated using simulations (not shown) with the QCEW population.

We shall call the estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ of equation (14) the *weighted calibrated cell counts* estimator.

Simulations with artificial response variables y, also not shown, demonstrate that when the model (13) holds, then weighted calibrated cell counts estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ performs markedly better than the other estimators considered here. Table 6 gives statistics for the estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ for the populations and variables studied in Tables 2 - 5.

Comparing to Tables 2 - 5, we see that in all cases $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has the highest bias but the lowest MSE of the estimators considered. For $n = 200$ and the full population, $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has a 14.8% gain in efficiency (as measured by MSE) relative to the empirical results for the unbiased estimator when estimating total employment and a 21.1% efficiency gain when estimating total wages. For $n = 200$ and the population with a single extreme outlier deleted, the corresponding gains are 14.2% and 21.7% for total employment and total wages respectively.

The Associate Editor suggested that we compare our estimators to a poststratified estimator using collapsed cells to avoid the problem of empty cells in the sample. We explored this question for sample size $n = 200$ where it is most likely that empty cells will occur. We constructed 14 poststrata. Nine of these poststrata are the nine largest cells in the original data. The other 5 poststrata are A1 and A2; A3, A5, and B4; A4, B5, and C4; C5 and D4; and all cells from state E together with D5. After these combinations, the 5 combined poststrata had sizes that ranged between 4.07% and 5.06% of the population and the 9 retained original cells had sizes in the range of 4.62% to 11.47%.

Table 6
Empirical statistics for $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ of equation (14)

n	bias	st. err.	rt. MSE	rt. est. var.
		Full population - total employment		
200	-244,749	967,066	997,556	923,492
1,000	-64,839	490,758	495,023	483,550
5,000	-10,767	221,702	221,964	219,408
		Full population - total wages		
200	-242,528	1,388,489	1,409,511	1,333,793
1,000	-62,091	752,603	755,160	744,315
5,000	-9,821	332,682	332,827	333,782
		Population with outlier deleted - total employment		
200	-236,812	881,844	913,088	842,191
1,000	-67,468	405,215	410,793	396,105
5,000	-11,482	189,501	189,848	185,483
		Population with outlier deleted - total wages		
200	-228,441	1,194,922	1,216,562	1,151,417
1,000	-66,765	565,008	568,939	557,676
5,000	-11,138	263,699	263,934	260,768

Unfortunately, the author no longer has access to the QCEW data base. Besides the cell counts in Table 1, the author has only the means, standard deviations, and maximum values by cell. The author constructed a pseudo population using the squares of randomly generated gamma variables. The square gamma variables were constructed to have the same cell means and standard deviations as the cell means and standard deviations in the original data. After doing this, the square gamma variables were rounded upwards to integer values. For these pseudo populations, $T_y = 3{,}149{,}491$ for employment and 2,305,273, in tens of thousands of dollars, for wages.

A square gamma distribution was used because the gamma distribution is insufficiently right skewed. Even so, in almost all cells the largest value in the original population exceeded the largest value in the pseudo population. Of course without the original data, we cannot distinguish between right skew and a tendency to produce outliers.

10,000 Monte Carlo samples were constructed were taken for each sample size. The results are shown in Table 7. For the poststratified estimator, 5 of the samples of size 200 had an empty poststratum and these runs were excluded from the results in Table 7.

Table 7
Empirical comparison of 4 estimators

estimator	bias	st. err. total employment	rt. MSE	bias	st. err. total wages	rt. MSE
			n = 200			
unweighted	644	1,006,956	1,006,956	-9,970	1,481,450	1,481,483
poststratified	-5,387	1,026,266	1,026,280	-2,149	1,548,833	1,548,834
cal. cell cts.	-224,198	942,164	968,472	-203,531	1,377,823	1,392,775
wtd. cal. cell cts.	-248,937	919,419	952,523	-232,558	1,326,234	1,346,469
			n = 1,000			
unweighted	-3,317	445,676	445,687	1,544	679,148	679,150
poststratified	-2,967	448,218	448,228	1,672	685,370	685,372
cal. cell cts.	-54,311	436,821	440,185	-44,942	665,799	667,314
wtd. cal. cell cts.	-63,327	432,396	437,008	-54,913	660,726	663,004
			n = 5,000			
unweighted	2,466	206,249	206,264	-2,539	304,852	304,863
poststratified	2,108	205,661	205,672	-2,705	304,751	304,763
cal. cell cts.	-8,265	204,693	204,859	-12,096	303,231	304,472
wtd. cal. cell cts.	-10,551	204,080	204,352	-14,697	302,311	302,668

Evidently the poststratification did not help. Even though no poststratum had an expected count below eight, the actual poststrata had quite variable sizes. In addition, the cell populations are quite skewed so that the poststrata sample means are quite variable.

The other conclusions for the pseudo populations reflect the conclusions from the actual populations. In particular, when n = 200 and for the employment pseudo population, the weighted calibrated cell counts estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has an 11.8% gain in efficiency relative to the unbiased estimator. For the wages pseudo population and n = 200, the efficiency gain is 21.1%.

6. Concluding remarks

The use in (3) of weight matrices $\mathbf{V}(\beta)^{-1}$ which depend upon β has not been explored in this paper. Experimentation with the use of such a matrix was not encouraging. Computation time increased dramatically, and there were significant numbers of cases which failed to numerically converge, with no improvement in efficiency over the fixed $\mathbf{V}$ estimators considered here. Perhaps the authors did not try the right $\mathbf{V}(\beta)$.

Besides the exponential back link function, the authors tried the logistic back link $f(\eta) = (1 + e^{-\eta})^{-1}$. These runs also did not converge. On reflection, the reason is obvious: because in the simulations there was no nonresponse or noncoverage problems, the calibration weight adjustments $f(\mathbf{z}^T\beta) \to 1$ as $n \to \infty$. But 1 is not in the range of f. It should be noted that in Chang and Kott (2008) a logistic back link was used to correct for nonresponse.

Several obvious issues arise. For example, how would the results of this study change if a more complicated sampling design than simple random sampling were used, or if non response and/or non coverage occur and the calibration was used to correct for it. Falk (2010) considers these questions both theoretically and with further simulations using the QCEW population. Falk (2010) also considers non linear link functions.

There are obvious extensions to 3-way (and beyond) cross classified data. If I, J, K denote the number of cells in each of the 3 classifications, there are IJK fully classified cells whose totals can be used for benchmark $\mathbf{x}$ variables. There are $IJ + IK + JK - I - J - K + 1$ one-way and two-way marginal variables that can be used for model $\mathbf{z}$ variables. Clearly, one might not want to use the plethora of variables available.

In the context of linear calibration using the same $\mathbf{x}$ and $\mathbf{z}$ variables, several studies have been made on the choice of variables. Examples of such studies are Banker, Rathwell and Majkowski (1992), Silva and Skinner (1997), and Clark and Chambers (2008). The last paper remarks that too many variables can deteriorate the MSE of $\hat{T}_y$.

The alternatives to poststratification discussed here can be used in the presence of small and even empty cells. For example, in our simulations, the expected count in the state E, industry group 5, cell is 0.36 when n = 200. One might be tempted to collapse cells and use poststratification. Generally, however, it is not possible to do so and maintain the convenient doubly classified structure of the data. Our approaches, like poststratification, introduce weights for the purpose of sample balancing but avoid collapsing cells. These approaches generally increase bias but can offer substantial reductions in MSE.

Furthermore, in the presence of nonresponse or noncoverage, the inverse of the weight adjustments can be

considered, under a quasi-randomization model for the response or coverage, as estimated probabilities of response and/or coverage. In our calibration approaches, these probabilities are assumed to be a function of the row and column classifications. When cells are collapsed without maintaining the double classification, these probabilities are harder to interpret.

Acknowledgements

The author would like to thank Phil Kott and John Eltinge for several very interesting insights. The author also thanks Larry Huff and Ken Robertson of the Bureau of Labor Statistics for help in obtaining and understanding the data.

Appendix

Here we derive, using Chang and Kott (2008) equations (16) and (17), sample based variance estimators for the 4 estimators studied in Section 2.

Let

$$\hat{\mathbf{H}}_y = \frac{\partial \hat{t}_{y,\mathbf{z}f\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

Here $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ is defined in (2). $\hat{\mathbf{H}}_y$ is a row vector with one entry for each $\mathbf{z}$ variable. In out case, $\hat{\mathbf{H}}_y$ has $(I + J - 1)$ entries, one for each of the $I + J - 1$ linearly independent indicator variables for the row and column classifications.

For the calibrated margins and calibrated cell counts estimators, $f(\eta) = 1 + \eta$. Define the constants s_{ij} and t_{ij} by

$$s_{ij} = \frac{N}{n} r_{ij}$$
$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} y_k.$$

Then a simple calculation shows that if a entry exists in $\hat{\mathbf{H}}_y$ for the i^{th} row classification, we place in that entry $\sum_j t_{ij}$. Similarly if a entry exists for the j^{th} column classification, we place in that entry $\sum_i t_{ij}$. Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding $\beta_{i\cdot}$ or $\beta_{\cdot j}$ is 0.

For the raking and exponential calibrated cell counts estimators, $f(\eta) = e^{\eta}$ and we can similarly calculate $\hat{\mathbf{H}}_y$ using instead

$$s_{ij} = \frac{N}{n} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$
$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) y_k.$$

Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding $\beta_{i\cdot}$ or $\beta_{\cdot j}$ is 1.

Analogously to (2), let

$$\hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}} = \sum_{k \in \mathcal{S}} d_k \, f(\mathbf{z}_k^T \hat{\beta}) \, x_k.$$

$\hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}}$ is a column vector with one entry for each $\mathbf{x}$ variable. Define the $\hat{\mathbf{H}}$ matrix to be

$$\hat{\mathbf{H}} = \frac{\partial \hat{t}_{\mathbf{x},\mathbf{z}f\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

$\hat{\mathbf{H}}$ is a matrix with one row for each $\mathbf{x}$ variable and one column for each $\mathbf{z}$ variable.

For the calibrated cell counts and exponential calibrated cell counts estimators the matrix $\hat{\mathbf{H}}$ has dimensions $IJ \times (I + J - 1)$. Each of the rows of $\hat{\mathbf{H}}$ corresponds to a pair (i, j) of row and column classifications. We place s_{ij} in the row corresponding to (i, j) and the columns corresponding to the i^{th} row classification and the j^{th} column classification (whenever these columns exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

For the calibrated margins and raking estimators the matrix $\hat{\mathbf{H}}$ has dimensions $(I + J - 1) \times (I + J - 1)$. If a row (and hence a column) of $\hat{\mathbf{H}}$ exists for the i^{th} row classification we put $\sum_j s_{ij}$ in the corresponding diagonal entry of $\hat{\mathbf{H}}$. Similarly, if a row and column exist for the j^{th} column classification, we put $\sum_i s_{ij}$ on the diagonal of $\hat{\mathbf{H}}$. We place s_{ij} in the entry whose row corresponds to the i^{th} row classification and whose column corresponds to j^{th} column classification (whenever these exist). We also place s_{ij} in the entry whose column corresponds to the i^{th} row classification and whose row corresponds to j^{th} column classification (again whenever these exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

Let $\mathbf{B} = \hat{\mathbf{H}}_{\mathbf{y}}^{\mathbf{T}} (\hat{\mathbf{H}}^{\mathbf{T}} \mathbf{V}^{-1} \hat{\mathbf{H}})^{-1} \hat{\mathbf{H}}^{\mathbf{T}} \mathbf{V}^{-1}$ where currently we are using an identity matrix for $\mathbf{V}$. $\mathbf{B}$ has dimensions $1 \times (I + J - 1)$ for the calibrated margins and raking estimators and $1 \times IJ$ for the calibrated cell counts and the exponential calibrated cell counts estimators. In the former cases, we will denote the entries of $\mathbf{B}$ by $b_{i\cdot}$ or $b_{\cdot j}$, and, for the single case when a column or row index does not correspond to one of the $I + J - 1$ independent indicator variables, we will set the corresponding b to zero. In the latter cases, we will denote the entries of $\mathbf{B}$ by b_{ij}. For $k \in \mathcal{S}_{ij}$, let $u_k = w_{ij}(y_k - b_{i\cdot} - b_{\cdot j})$ for the calibrated margins and raking estimators and $u_k = w_{ij}(y_k - b_{ij})$ for the calibrated cell counts and exponential calibrated cell counts estimators.

Essentially Chang and Kott (2008) showed that, asymptotically, the calibrated estimator has the same form as a regression estimator of the form Särndal *et al.* (1992)

equation (6.6.1) where the above **B** plays the role of **B** in (6.6.1) and the sampling weights d_k are replaced by $d_k f(\mathbf{z}_k^T \hat{\beta})$. For non replacement designs, they propose to estimate the variance of $\hat{t}_{y,\mathbf{z}f\mathbf{V}}$ using the analogous changes to Särndal *et al.* (1992) equation (6.6.3).

For simple random sampling, and in the absence of nonresponse or noncoverage, the variance estimator works out to

$$\widehat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)s_u^2 \qquad (15)$$

where s_u^2 is the sample variance of the u_k.

In the presence of nonresponse, if one assumes that the respondents $\mathcal{S}$ are a Poisson sample from the original simple random sample with Poisson probabilities $f(\mathbf{z}^T\beta_0)^{-1}$, the variance estimator becomes

$$\hat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)s_u^2 + \frac{N}{n}\sum_i\sum_j(1 - w_{ij})\sum_{k\in\mathcal{S}_i j} u_k^2 \qquad (16)$$

where s_u^2 is the sample variance of the u_k. The same formula works for noncoverage where $f(\mathbf{z}^T\beta_0)^{-1}$ represents the combined coverage and response probability in a three stage model in which the covered universe is assumed to be a Poisson sample from the desired universe, the sample is a simple random sample from the covered universe, and the respondents are a Poisson sample from the original sample.

References

Banker, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 764-769.

Chang, T., and Kott, S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.

Clark, R., and Chambers, R.L. (2008). Adaptive Calibration for Prediction of Finite Population Totals. University of Wollongong (on line working paper).

Falk, G. (2010). *Calibration Adjustment for Nonresponse in Cross-Classified Data*, University of Virginia (dissertation).

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the Presence of Nonresponse with application to the 1987-1988 Nationalwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Kott, S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.

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

Silva, P.L.D.N., and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.