Optimal Cutoff Sampling for the Annual Survey of Public Employment and Payroll

Brian Dumbacher[1], Carma Hogue[1]

[1]U.S. Census Bureau
4600 Silver Hill Road, Washington, DC 20233
Brian.Dumbacher@census.gov, Carma.Ray.Hogue@census.gov

Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract

The goal of cutoff sampling is to save cost, reduce respondent burden, and maintain accuracy of estimates by reducing the number of small units in sample. For the Annual Survey of Public Employment and Payroll, the Governments Division of the U.S. Census Bureau uses a modified version of cutoff sampling in which a subsample of units below the cutoff is selected. In this paper, we examine a numerical method based on minimizing the average of mean squared errors from linear regression models to find an optimal combination of cutoff and subsampling rate given a specified cost. Data from the 2002 and 2007 Censuses of Governments: Employment are used for this study.

Key Words: Cutoff sampling; Decision-based estimation; Linear regression; Mean squared error

1. Introduction

1.1 Survey Overview

The Annual Survey of Public Employment and Payroll (ASPEP) is conducted by the Governments Division of the U.S. Census Bureau to collect data on federal, state, and local government civilian employees and their gross payrolls. Key study variables for ASPEP include the total number of employees, total pay, and the number of full-time equivalent employees. Small area composite methodology is used to estimate local government totals for each combination of state and government function. Government functions are identified by item code, and a complete list of item codes is provided in Appendix A.

1.2 Sample Design

The sampling frame for ASPEP is a list of the 89,476 local governments identified during the 2007 Census of Governments and is updated annually with births (newly discovered governments), deaths (disincorporated governments), and mergers. Initial certainties are determined based on population size, school enrollment, and government function, and then a first-stage, stratified, probability-proportional-to-size sample (Särndal, Swensson, & Wretman, 1992, p. 90) is selected from the remaining local governments, where the strata are determined by the cross-classification of state and government type and the measure of size is total pay in 2007.

The different types of local government are counties, cities, townships, special districts, and independent school districts, but cities and townships are deemed similar enough to group together and are known collectively as subcounty governments. Counties and subcounties are known as general-purpose governments, and they tend to perform several of the functions listed in Appendix A. Special districts and school districts, on the other hand, are known as single-purpose governments and tend to perform one or a very limited number of functions. The purpose of school districts is education, while special districts may be cemetery districts, public utilities, transit authorities, etc. As such, their contribution to a single function like air transportation may be great, but they would have no contribution to other functions. Table 1 gives a breakdown of the local governments in 2007 by type. As you can see, there are many subcounties and special districts, but these units' shares of total employees and total pay are disproportionately small.

Table 1: Local governments in 2007

Government type	Number	%	Total employees	%	Total pay ($)	%
County	3,033	3.39	2,928,244	20.64	10,093,125,772	21.77
Subcounty	36,011	40.25	3,510,995	24.75	12,717,946,464	27.43
Special district	37,381	41.78	821,369	5.79	2,651,730,327	5.72
Independent school district	13,051	14.59	6,925,014	48.82	20,904,942,336	45.09
Total	89,476	100.00	14,185,622	100.00	46,367,744,899	100.00

Source: U.S. Census Bureau, 2007 Census of Governments: Employment

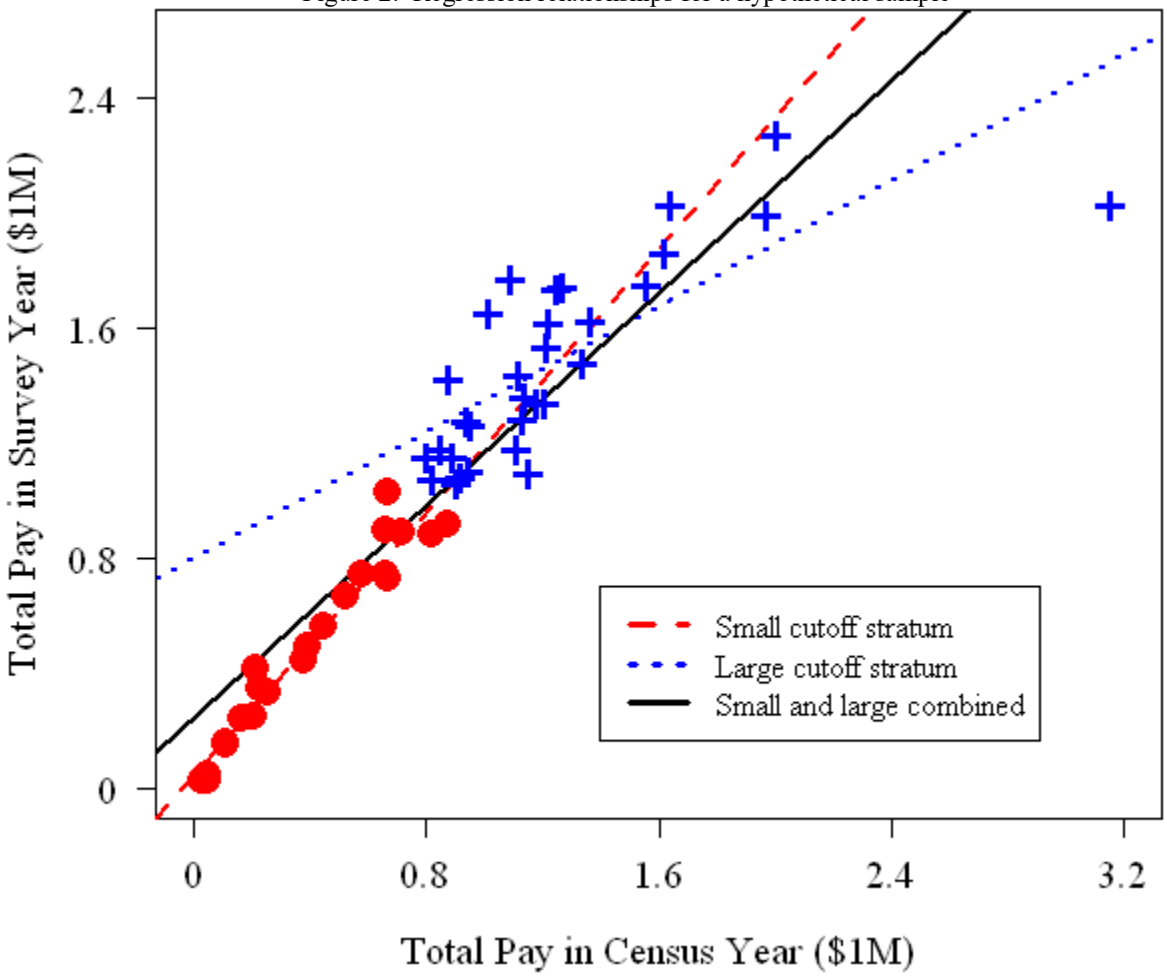
To reduce the number of non-contributory units in sample, the Governments Division uses cutoff sampling in the subcounty and special district strata to divide each first-stage sample into a small cutoff stratum and a large cutoff stratum (Barth, Cheng, & Hogue, 2009). For the 2009 sample design, the cutoffs were determined by the cumulative square root of the frequency method (Dalenius & Hodges, 1959). [See Cochran (1977, p. 130) for an overview and Appendix B for an illustration of this method.] Instead of completely ignoring the sample units in the small cutoff strata as is done in standard cutoff sampling, the Governments Division selects a subsample of them. This helps protect against bias that could result from small units becoming large units during intercensal years. A subsampling rate equal to 0.56 was used in all small cutoff strata, which resulted in a target reduction of 800 subcounty and special district units. This target was based on a rough cost-benefit analysis, and we have planned more optimal methods for the future when a new ASPEP sample based on the 2012 Census of Governments is selected.

1.3 Small Area Estimation Methodology

Estimates of local government totals are calculated for each combination of state and item code using small area composite methodology (Tran & Cheng, 2012). Each composite estimate is a weighted average of the direct Horvitz-Thompson estimate and a synthetic estimate. This synthetic estimate equals the product of a decision-based regression estimate of the state total and a proportion for the item code within the state. The term "decision-based" refers to statistical hypothesis tests that are carried out to determine whether the regression relationships in the small and large cutoff strata are similar enough that the strata can be combined for estimation purposes.

Figure 2 is a scatterplot of total pay in a survey year versus total pay in the most recent Census year for a hypothetical sample after the cutoff is determined and subsampling is performed. Separate linear regressions are fitted in the small and large cutoff strata using sample data, and then a statistical hypothesis test of the equality of the regression slopes is carried out. If the null hypothesis is rejected, then the cutoff strata are kept separate. If the null hypothesis is not rejected, then the cutoff strata are combined and a new regression is fitted to all the units. Whichever regression is decided on is then applied to Census data to estimate the state total. This is a simplified description of the process as robust regression and auxiliary data are used to handle outliers and to strengthen poor fitting models. Also, the variable total pay is used in this example, but in production, the decision-based methodology would be applied to full-time pay and part-time pay separately. For a much more detailed description of decision-based estimation, see Cheng, Corcoran, Barth, and Hogue (2009).

Figure 2: Regression relationships for a hypothetical sample



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

1.4 Objective

This study continues the research by Corcoran and Cheng (2010) on finding an optimal combination of cutoff and subsampling rate. In their paper, they investigated a numerical method of minimizing the sum of unweighted mean squared errors (MSEs) from linear regression models fitted separately to units from the small and large cutoff strata. The rationale is the following: a reasonable linear relationship between sample data and data from the most recent Census of Governments in each cutoff stratum would improve the efficiency of decision-based estimation. Corcoran and Cheng found that if they minimized the sum of MSEs with respect to the cutoff and subsampling rate, the algorithm would tend to keep 100 percent of the small units. To address this, they suggested introducing a penalty term that accounts for the cost associated with small units. We continue their work by considering different measures of MSE and by conducting a larger simulation that incorporates sampling weights. Given a first-stage sample, our objective is to find a combination of cutoff and subsampling rate that minimizes some measure of MSE subject to a cost constraint.

2. Methodology

2.1 Notation

s	First-stage sample
c	Cutoff
p	Subsampling rate
$n_1(c)$	Size of the small cutoff stratum
$n_1^*(c,p)$	Size of the subsample selected from the small cutoff stratum $= round(n_1(c) \times p)$
$n_2(c)$	Size of the large cutoff stratum
y_i	Value of total pay 2007 for unit i
$\hat{y}_i$	Predicted value of total pay 2007 for unit i based on a weighted regression of total pay 2007 on total pay 2002
w_i	Sampling weight for unit i
B	Number of simulated subsamples for a certain combination of c and p
b	Index for simulated subsamples
s_b	Subsample selected from s
$MSE_b^1(c,p)$	Weighted MSE for the small cutoff stratum $= \frac{1}{n_1^*-2}\sum_{i=1}^{n_1^*} w_i(y_i-\hat{y}_i)^2$
$MSE_b^2(c)$	Weighted MSE for the large cutoff stratum $= \frac{1}{n_2-2}\sum_{i=1}^{n_2} w_i(y_i-\hat{y}_i)^2$
$MSE_b^{Simple}(c,p)$	Simple average MSE $= \frac{1}{2}[MSE_b^1(c,p) + MSE_b^2(c)]$
$\overline{MSE}^{Simple}(c,p)$	Average $MSE_b^{Simple}(c,p)$ $= \frac{1}{B}\sum_{b=1}^{B} MSE_b^{Simple}(c,p)$
$MSE_b^{Pooled}(c,p)$	Pooled average MSE $= \frac{1}{n_1^*+n_2-4}[(n_1^*-2)MSE_b^1(c,p) + (n_2-2)MSE_b^2(c)]$
$\overline{MSE}^{Pooled}(c,p)$	Average $MSE_b^{Pooled}(c,p)$ $= \frac{1}{B}\sum_{b=1}^{B} MSE_b^{Pooled}(c,p)$
$COST_i$	Cost of unit i
$COST_b(c,p)$	Cost of subsample s_b $= \sum_{i\in s_b} COST_i$
$\overline{COST}(c,p)$	Average cost $= \frac{1}{B}\sum_{b=1}^{B} COST_b(c,p)$
C_0	Cost constraint

2.2 Measures of Mean Squared Error

We consider two measures of MSE. The simple average MSE, MSE^{Simple}, is just the unweighted average of the MSEs from the small and large cutoff strata. The pooled average MSE, MSE^{Pooled}, is the sample size-weighted average and is appropriate when the MSEs from the small and large cutoff strata are estimating the same variance component. Because of heteroskedasticity in Census data, MSE^{Pooled} may not apply theoretically, but we would like to see how it performs anyway.

2.3 Measure of Cost

It would be ideal to measure $COST_i$ as a function of inputs such as dollar amount to conduct a case and estimated response propensity. This response propensity could itself be a function of y_i, state, government type, and other covariates. However, because of a lack of cost information currently, we could not come up with an adequate measure. As data become available in the future, the cost measure will be re-examined. Instead, we model cost as some constant times sample size. The optimization problem then becomes finding the optimal combination of cutoff and subsampling rate subject to a given sample size. This model is simplistic and ignores anecdotal evidence that smaller units require more nonresponse follow-up.

2.4 Simulation

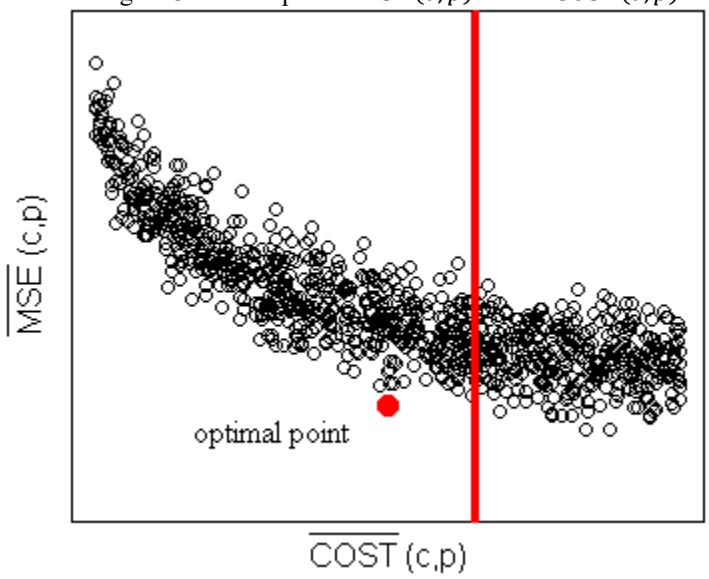
Given a first-stage sample and specified cost (specified sample size) C_0, the objective is to find c and p that minimize $\overline{MSE}(c,p)$ subject to

1. $\overline{COST}(c,p) \leq C_0$
2. $n_1^* \geq 15$
3. $n_2 \geq 15$
4. $n_1^* + n_2 \geq 40$

Condition 1 is the cost constraint, and conditions 2-4 make up the so-called 15-40 rule, which was used during the original sample selection to ensure enough data were available for the decision-based regression models. For a given first-stage sample, we iterate through all possible combinations of cutoff and subsampling rate (p = 0.05, 0.10, ... , 0.90, 0.95), and for each combination we select B subsamples from the small cutoff stratum. The final subsample for iteration b, denoted s_b, is the union of the subsample from the small cutoff stratum and the units from the large cutoff stratum. For each subsample, we calculate MSE_b^{Simple}, MSE_b^{Pooled}, and $COST_b$. Finally, we average these B values to obtain $\overline{MSE}^{Simple}(c,p)$, $\overline{MSE}^{Pooled}(c,p)$, and $\overline{COST}(c,p)$.

Figure 3 is a scatterplot of simulated values of $\overline{MSE}(c,p)$ versus $\overline{COST}(c,p)$ that shows what we expect to find. Each point represents a unique combination of cutoff and subsampling rate. As $\overline{COST}$ increases, $\overline{MSE}$ should decrease. The vertical red line represents the cost constraint, and the red point to the left of this line with the smallest $\overline{MSE}$ is the optimal point. The optimal cutoff and subsampling rate are the values corresponding to this optimal point.

Figure 3: Scatterplot of $\overline{MSE}(c,p)$ versus $\overline{COST}(c,p)$



Source: Simulated data for illustrative purposes only

3. Simulated Data

3.1 Description

We first test this numerical method on simulated data. Consider a first-stage sample defined by the following:

$$\begin{array}{lllll} x_1 & \cdots & x_{60} & \sim & Uniform(0,60) \\ x_{61} & \cdots & x_{100} & \sim & Uniform(60,100) \end{array} \quad , \quad \text{and}$$

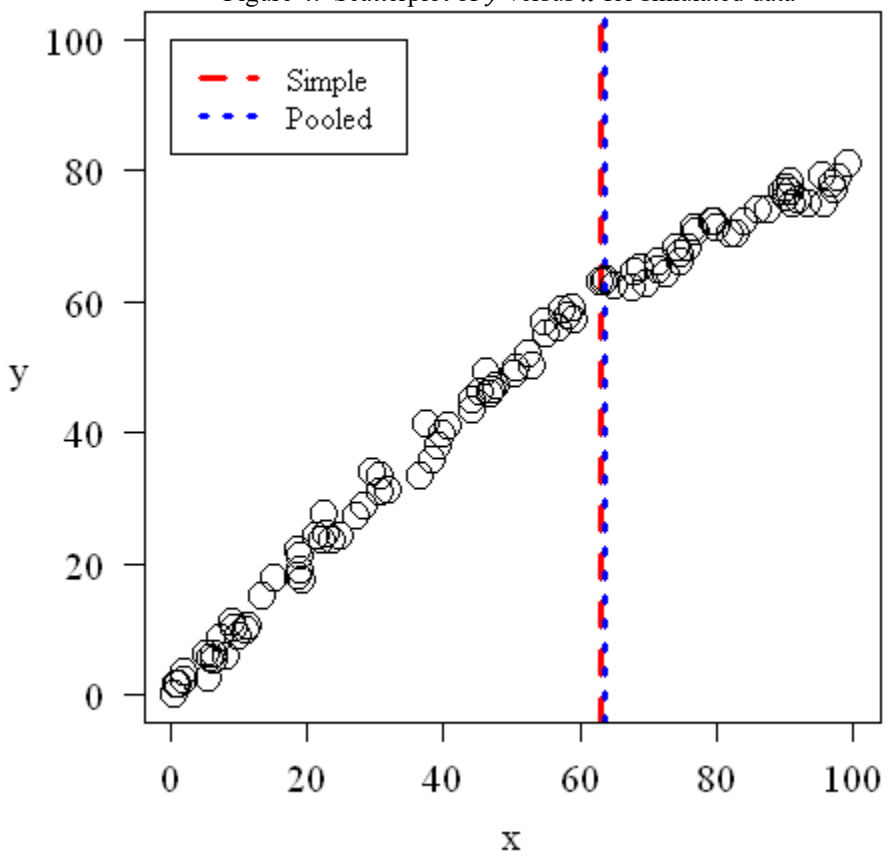
$$y_i = \begin{cases} x_i + \varepsilon_i & i = 1, \cdots, 60 \\ 0.5x_i + 30 + \varepsilon_i & i = 61, \cdots, 100 \end{cases} ,$$

where ε_i are independent $N(0, 1.5)$ random variables. Note that the linear relationship between x and y changes at $x = 60$. We iterate through all 100 possible cutoffs and 19 subsampling rates (p = 0.05, 0.10, ... , 0.90, 0.95). Altogether there are 1,900 (= 100 × 19) combinations of cutoff and subsampling rate. For each combination we select B = 50 subsamples.

3.2 Results

Figure 4 shows a scatterplot of y versus x for the simulated data. The vertical lines represent the optimal cutoffs when $C_0 = 100$, which is equivalent to there being no cost constraint. The variability in y does not change with x, so both measures of MSE give approximately the same optimal cutoff, which is located near the change point in the scatterplot.

Figure 4: Scatterplot of y versus x for simulated data



Source: Simulated data for illustrative purposes only

Table 5 gives the optimal combinations of c and p for different cost constraints C_0. As a reminder, the cost constraint requires that $n_1^* + n_2 \leq C_0$. As C_0 decreases, the optimal cutoff c stays fairly constant for both measures of MSE. However, MSE^{Pooled} gives a smaller optimal subsampling rate p for $C_0 \geq 70$.

Table 5: Optimal combinations of c and p for different C_0

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
100	63.12	0.45	27	39	2.55	63.70	0.25	16	38	2.47
90	63.12	0.45	27	39	2.55	63.70	0.25	16	38	2.47
80	63.12	0.45	27	39	2.55	63.70	0.25	16	38	2.47
70	63.12	0.45	27	39	2.55	63.70	0.25	16	38	2.47
60	63.70	0.25	16	38	2.60	63.70	0.25	16	38	2.47
50	67.95	0.25	16	34	2.86	67.95	0.25	16	34	2.67

4. Census of Governments: Employment Data

4.1 Description

Next, we apply our method to the 2002 and 2007 Censuses of Governments: Employment data from the following six strata:

- California special districts
- Illinois special districts
- Kentucky special districts
- New York subcounties
- Pennsylvania subcounties
- Wisconsin subcounties

California, Illinois, and Pennsylvania are large states with many governments, so it is important to examine how our numerical method perform for them. The strata for Kentucky, New York, and Wisconsin give us a variety of first-stage sample sizes and regression relationships between total pay in 2007 and total pay in 2002. As with the simulated data in Section 3, we select $B = 50$ subsamples for each combination of c and p.

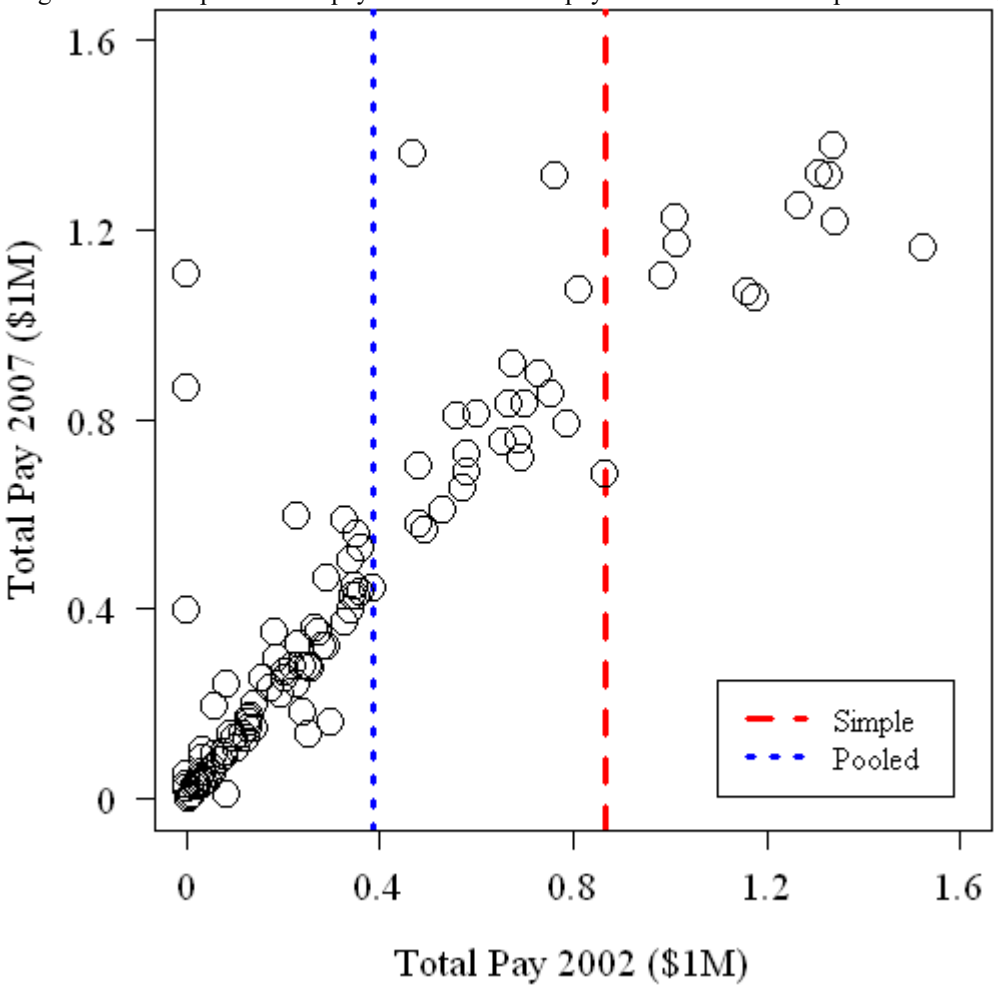
4.2 Results

Results for these six strata are below. For each stratum, we have a scatterplot and table just like the ones for the simulated data. The vertical lines in the scatterplot represent the cutoff points for $C_0 = n_1 + n_2$, the size of the first-stage sample s. This is equivalent to there being no cost constraint.

For both MSE^{Simple} and MSE^{Pooled}, as C_0 decreases and the cost constraint is tightened, the optimal cutoff c tends to increase and the optimal subsampling rate p tends to decrease. In many strata, MSE^{Pooled} gives less extreme cutoffs and more stable cutoffs as C_0 varies. Less extreme cutoffs allow for more sample units in the cutoff strata and could result in more reliable regression models in the decision-based methodology. Also, the optimal subsampling rates for MSE^{Pooled} tend to be smaller than the ones for MSE^{Simple}. This is desirable in terms of cost because this means fewer small units in sample. In general, both measures of MSE give similar optimal cutoffs and subsampling rates when there is no cost constraint ($C_0 = n_1 + n_2$). In this case, the optimal subsampling rate is close to 100 percent, which makes sense since you would want to take as large a subsample as possible to decrease MSE. This observation was also made by Corcoran and Cheng (2010). On an added note, using sampling weights, which are inversely proportional to total pay in 2007, could be accounting for heteroskedasticity in the Census data in the sense that units with a smaller value of total pay are given more weight just as in the standard ratio model (Särndal et al., 1992, p. 248). This could be making MSE^{Pooled} a more appropriate measure of MSE.

4.2.1 California Special Districts

Figure 6: Scatterplot of total pay 2007 versus total pay 2002 for California special districts (C_0 = 117)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

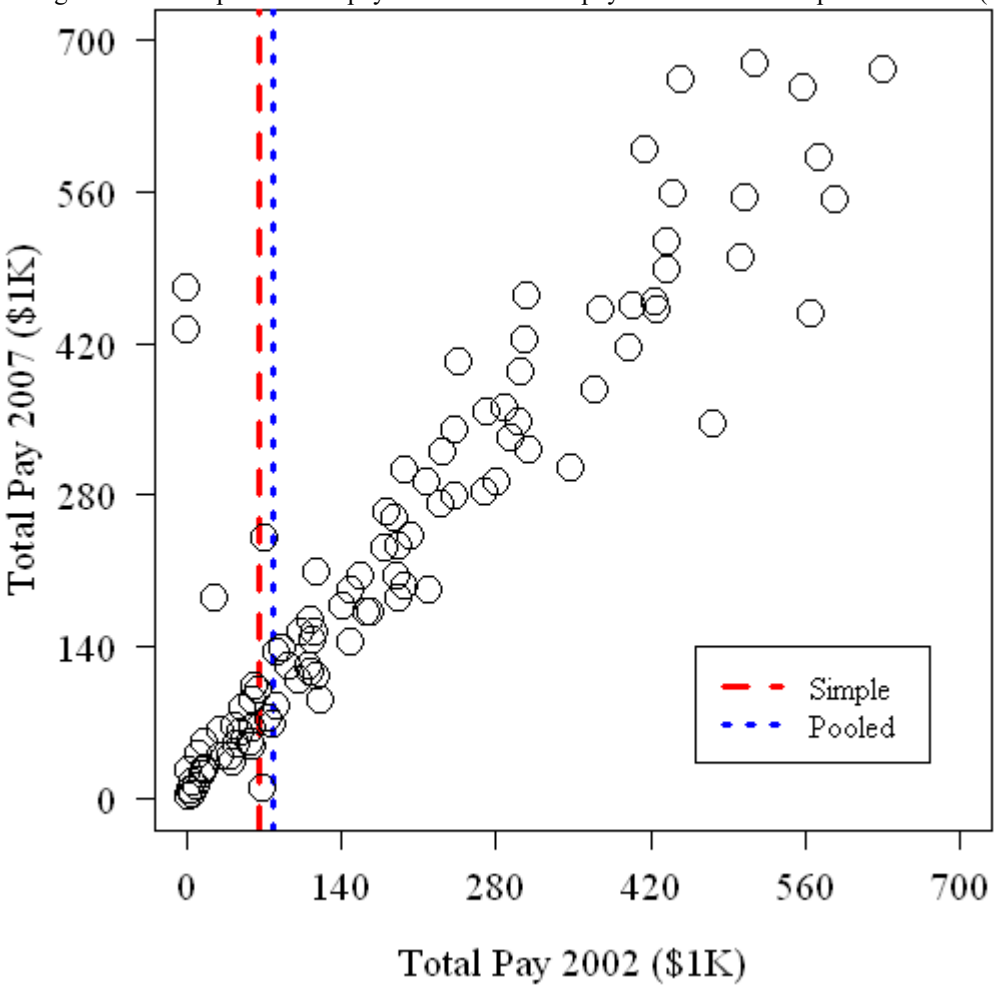
Table 7: Optimal combinations of c and p for different C_0 for California special districts

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
117	864,558	0.95	97	15	31,694,665,727	385,632	0.90	68	41	44,283,250,162
100	864,558	0.80	82	15	36,008,302,588	299,284	0.70	46	52	48,257,651,800
80	864,558	0.60	61	15	46,571,475,765	299,284	0.40	26	52	57,651,483,097
60	864,558	0.40	41	15	65,701,745,703	578,461	0.35	30	30	70,791,759,055
40	701,817	0.20	19	21	140,974,185,977	701,817	0.20	19	21	135,161,793,627

Figure 6 shows that the large cutoff stratum created by the simple MSE cutoff for C_0 = 117 contains fewer than the n_2 = 15 units reported in Table 7. Some units had missing values for total pay 2002, so they could not be used to fit the regressions and could not be plotted. However, we assigned these units to the large cutoff stratum based on their large values for total pay 2007. An assignment like this would have to be done in practice.

4.2.2 Illinois Special Districts

Figure 8: Scatterplot of total pay 2007 versus total pay 2002 for Illinois special districts (C_0 = 105)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

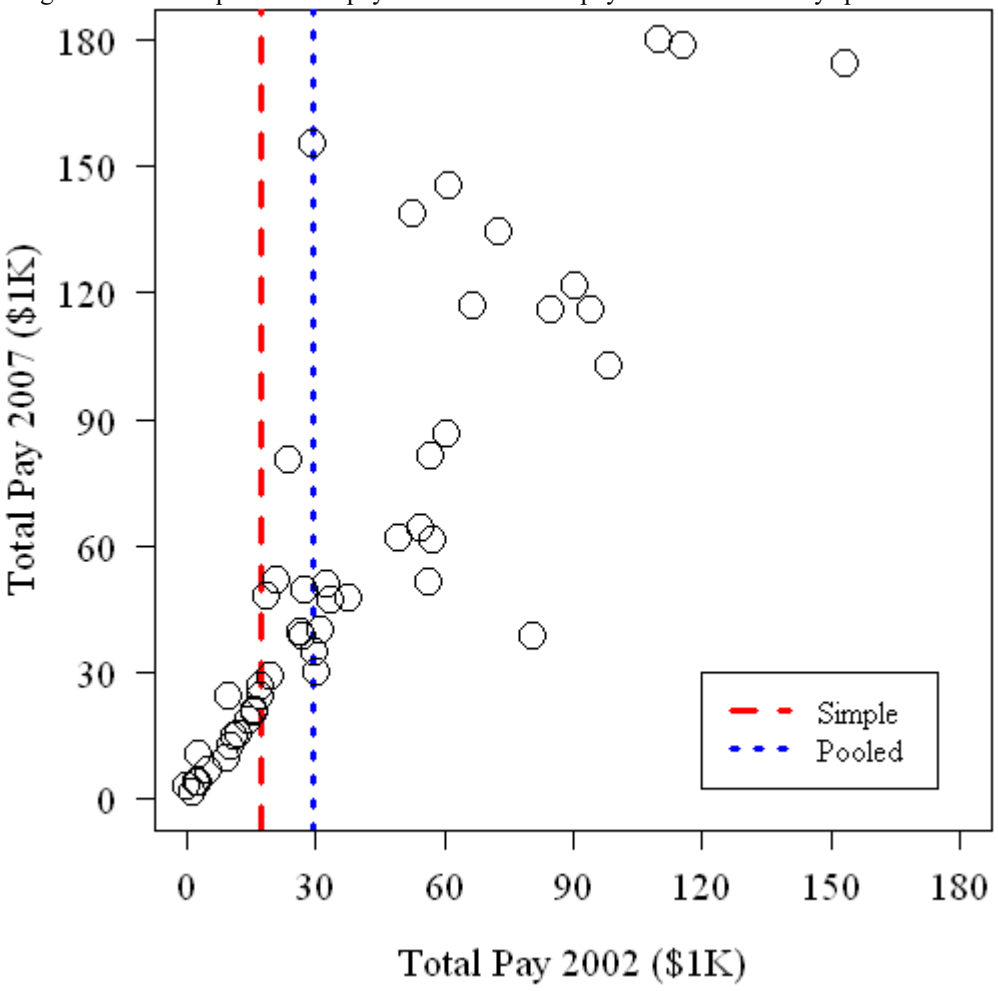
Table 9: Optimal combinations of c and p for different C_0 for Illinois special districts

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
105	64,790	0.95	25	79	9,819,988,031	78,028	0.95	29	75	9,672,177,612
90	243,133	0.75	50	39	11,783,057,827	78,028	0.50	15	75	10,297,195,604
80	400,778	0.70	59	21	13,388,688,940	122,215	0.45	20	60	12,047,518,976
70	281,294	0.50	36	34	14,787,571,381	162,614	0.30	15	54	12,707,754,395
60	404,622	0.45	38	20	16,949,757,609	230,577	0.30	19	41	16,367,532,747

For both measures of MSE, as C_0 decreases, the optimal cutoff increases and the optimal subsampling rate decreases. However, the optimal cutoff for MSE^{Pooled} increases more smoothly.

4.2.3 Kentucky Special Districts

Figure 10: Scatterplot of total pay 2007 versus total pay 2002 for Kentucky special districts (C_0 = 56)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

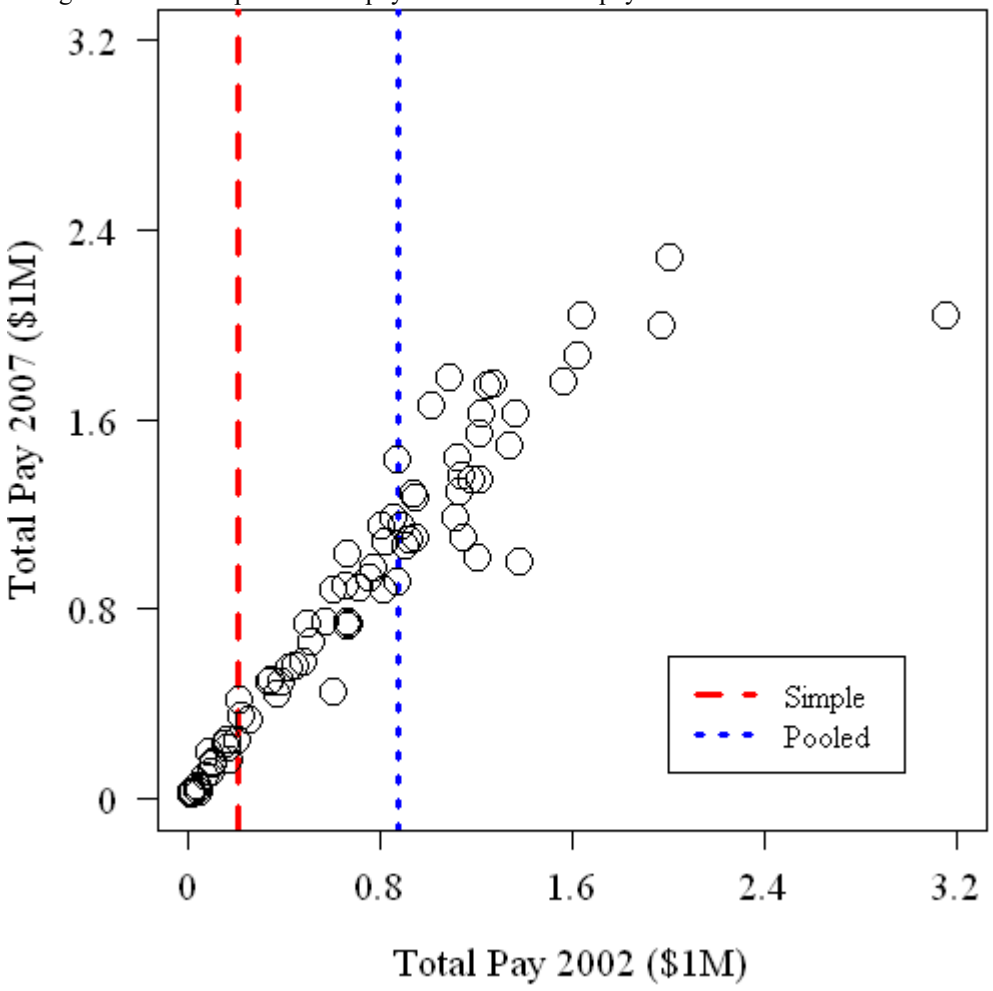
Table 11: Optimal combinations of c and p for different C_0 for Kentucky special districts

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
56	17,398	0.95	15	40	1,155,847,287	29,227	0.95	25	30	1,606,571,065
53	18,513	0.85	16	37	1,284,609,668	29,227	0.90	23	30	1,674,166,574
50	20,856	0.70	15	35	1,431,825,687	29,227	0.75	20	30	1,760,470,628
47	27,615	0.65	16	31	1,650,857,639	29,227	0.60	16	30	1,857,955,810
44	29,958	0.55	15	29	1,998,516,751	29,958	0.55	15	29	2,020,537,507

This is a small stratum with 56 units in the first-stage sample. As C_0 decreases, the optimal cutoff for MSE^{Simple} increases, but the optimal cutoff for MSE^{Pooled} stays fairly constant. The optimal subsampling rates for both measures of MSE are similar.

4.2.4 New York Subcounties

Figure 12: Scatterplot of total pay 2007 versus total pay 2002 for New York subcounties (C_0 = 75)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

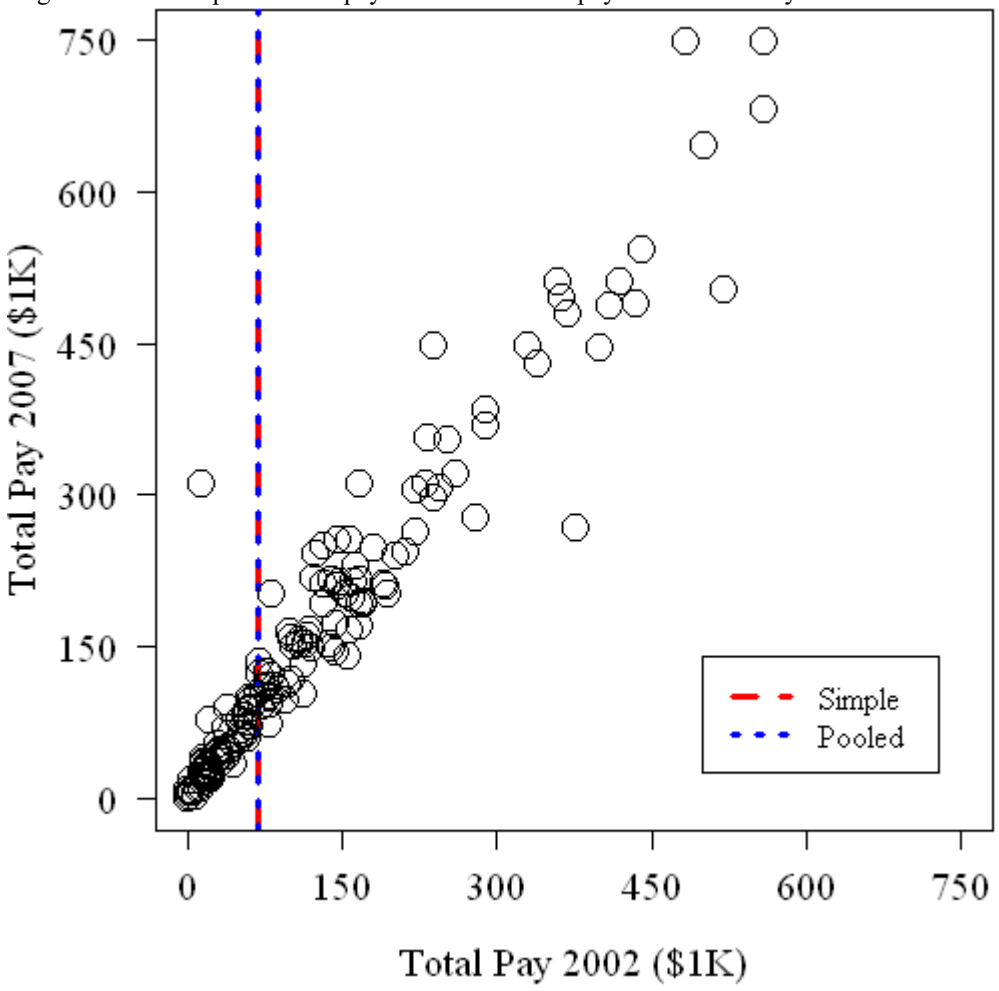
Table 13: Optimal combinations of c and p for different C_0 for New York subcounties

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
75	207,536	0.95	16	58	63,222,123,614	876,386	0.95	43	30	68,085,966,290
70	393,472	0.80	19	51	69,246,181,840	1,092,494	0.90	49	21	70,451,496,368
60	576,226	0.50	15	45	80,126,843,940	950,431	0.70	36	24	81,549,279,001
50	876,386	0.45	20	30	95,029,600,875	876,386	0.45	20	30	98,790,128,415
40	1,092,494	0.35	19	21	124,373,229,506	1,092,494	0.35	19	21	124,525,663,947

As C_0 decreases, the optimal cutoff for MSE^{Simple} increases, but the optimal cutoff for MSE^{Pooled} stays around $900,000. As with the other strata, as C_0 decreases, the optimal subsampling rate for both measures of MSE decreases.

4.2.5 Pennsylvania Subcounties

Figure 14: Scatterplot of total pay 2007 versus total pay 2002 for Pennsylvania subcounties (C_0 = 150)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

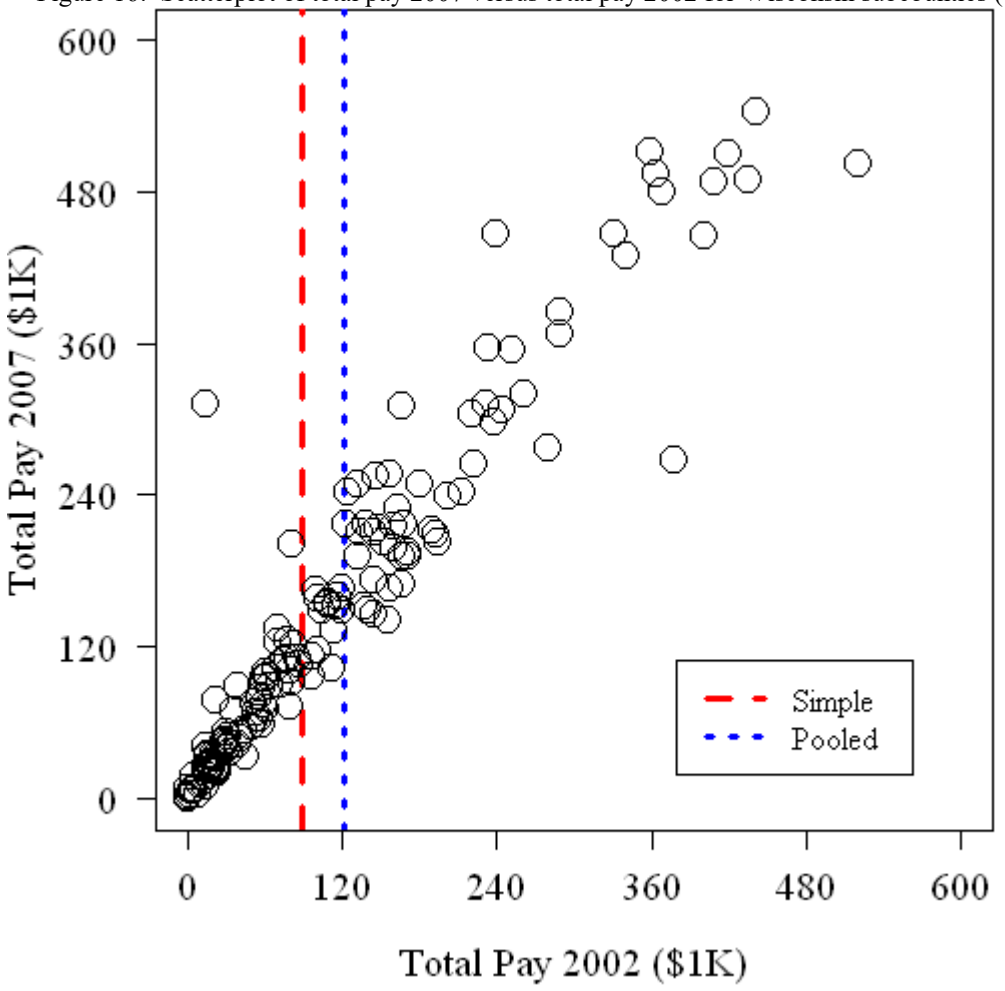
Table 15: Optimal combinations of c and p for different C_0 for Pennsylvania subcounties

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
150	69,003	0.95	59	88	6,389,218,799	69,003	0.95	59	88	6,427,537,409
130	100,988	0.75	60	70	7,216,883,985	32,616	0.45	18	110	7,057,067,333
110	145,309	0.60	60	50	8,199,606,208	63,041	0.35	21	89	7,941,222,906
90	163,037	0.45	49	41	10,299,842,681	96,329	0.20	15	73	8,906,481,479
70	171,291	0.30	35	34	12,416,360,291	171,291	0.30	35	34	12,471,354,735

This is a large stratum with 150 units in the first-stage sample. For both measures of MSE, as C_0 decreases, the optimal cutoff increases and the optimal subsampling rate decreases. However, the optimal cutoff for MSE^{Simple} increases more smoothly.

4.2.6 Wisconsin Subcounties

Figure 16: Scatterplot of total pay 2007 versus total pay 2002 for Wisconsin subcounties (C_0 = 93)



Source: U.S. Census Bureau, 2002 and 2007 Censuses of Governments: Employment

Table 17: Optimal combinations of c and p for different C_0 for Wisconsin subcounties

C_0	Simple					Pooled				
	c	p	n_1^*	n_2	$\overline{MSE}^{Simple}$	c	p	n_1^*	n_2	$\overline{MSE}^{Pooled}$
93	88,557	0.95	27	65	5,493,539,055	121,598	0.95	31	60	5,947,367,790
80	315,787	0.80	58	21	6,258,067,102	139,548	0.60	22	57	6,652,072,765
70	348,619	0.70	53	17	6,742,583,858	239,256	0.50	24	45	7,665,543,516
60	337,073	0.55	41	19	8,145,360,960	253,722	0.30	15	44	8,764,661,572
50	360,393	0.45	35	15	9,745,476,969	302,054	0.35	23	26	10,901,466,900

For both measures of MSE, as C_0 decreases, the optimal cutoff increases and the optimal subsampling rate decreases. However, the optimal cutoff for MSE^{Pooled} increases more smoothly. These trends are similar to those observed for Illinois special districts.

5. Limitations

Data from the 2012 Census of Governments: Employment were not available for this study. If 2012 data had been available, we would have regressed 2012 on 2007 instead of 2007 on 2002.

Accurate estimates of cost per case were not available. Because of a lack of cost information currently, we could not come up with an adequate cost measure. Modeling cost as a constant for every unit ignores anecdotal evidence that smaller units require more resources related to nonresponse follow-up. However, once a better cost model is created, perhaps one based on an estimated response propensity, it can be incorporated into the optimization algorithm easily.

We only considered total pay when fitting regressions. We did this because total pay was used as the measure of size for the current sampling design. Total pay is a good measure of size because it is strongly correlated with the other study variables. Using another measure of size such as total employees to fit the regression models might result in different optimal cutoffs and subsampling rates.

6. Recommendations

The measure MSE^{Simple} may be more appropriate for Census data, but MSE^{Pooled} appears to give less extreme cutoffs. Less extreme cutoffs seem ideal as they allow for more sample units in the cutoff strata and could result in more reliable regression models in the decision-based methodology. Our next step in the research is to compare the measures of MSE in terms of the accuracy of the decision-based estimates of state totals. We should also compare these results from those obtained from the current cumulative square root of the frequency method. There are also geometric cutoffs (Gunning & Horgan, 2004) and Lavallée and Hidiroglou cutoffs (Lavallée & Hidiroglou, 1988) that would be interesting to include in the analysis. On an added note, the optimal subsampling rates for MSE^{Pooled} tend to be smaller than the ones for MSE^{Simple}. This is desirable in terms of cost because this means fewer small units in sample.

Future cutoff sampling research could involve coming up with an optimal rule for determining which first-stage samples should be divided into small and large cutoff strata. The current 15-40 rule has intuitive appeal but was not optimally chosen. A related project could involve using numerical and graphical means to determine how many cutoff strata should be created. ASPEP currently uses two, but it may be more appropriate in certain strata to use more.

References

Barth, J., Cheng, Y., & Hogue, C. (2009). Reducing the Public Employment Survey Sample Size. *JSM*.

Cheng, Y., Corcoran, C., Barth, J., & Hogue, C. (2009). An Estimation Procedure for the New Employment Survey Design. Governments Division Report Series, Research Report #2009-9. Available online at <www.census.gov/prod/2009pubs/govsrr2009-9.pdf>

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

Corcoran, C. & Cheng, Y. (2010). Alternative Sample Approach for the Annual Survey of Public Employment and Payroll. Governments Division Report Series, Research Report #2010-5. Available online at <www2.census.gov/govs/pubs/2010pubs/govsrr2010-05.pdf>

Dalenius, T. & Hodges, J.L. (1959). Minimum Variance Stratification. *Journal of the American Statistical Association*, Vol. 54, pp. 88-101.

Gunning, P. & Horgan, J.M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, Vol. 30, pp. 159-166.

Lavallée, P. & Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, Vol. 14, pp. 33-43.

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

Tran, B. & Cheng, Y. (2012). Application of Small Area Estimation for Annual Survey of Employment and Payroll. *JSM*.

Appendix A: Item Codes

Table A-1: Item code descriptions and levels of applicability

Item Code	Description	Federal	State	Local
001	Air transportation	✓	✓	✓
002	Space research and technology	✓		
005	Corrections	✓	✓	✓
006	National defense and international relations	✓		
014	Postal service	✓		
022	Social insurance administration	✓	✓	DC
023	Financial administration	✓	✓	✓
924	Fire			✓
024	Firefighters			✓
124	Other employees			✓
025	Judicial and legal	✓	✓	✓
928	Education		✓	✓
912	Elementary and secondary education		✓	✓
012	Instructional employees		✓	✓
112	Other employees		✓	✓
918	Higher education		✓	✓
018	Instructional employees		✓	✓
016	Other employees		✓	✓
021	Other education	✓	✓	
029	Other government administration	✓	✓	✓
032	Health	✓	✓	✓
040	Hospitals	✓	✓	✓
044	Highways	✓	✓	✓
050	Housing and community development	✓	✓	✓
052	Libraries	✓	✓	✓
059	Natural resources	✓	✓	✓
061	Parks and recreation	✓	✓	✓
962	Police	✓	✓	✓
062	Officers		✓	✓
162	Other employees		✓	✓
079	Public welfare	✓	✓	✓
080	Sewerage		✓	✓
081	Solid waste management		✓	✓
087	Water transport and terminals	✓	✓	✓
089	All other and unallocable	✓	✓	✓
090	State liquor stores		✓	
091	Water supply		✓	✓
092	Electric power		✓	✓
093	Gas supply		✓	✓
094	Transit		✓	✓

Notes: Item code 090 is misleading in its current classification. Local governments do operate liquor stores, and data from them are coded under item code 089. Data for federal police are coded under the umbrella item code 962 because the data are not detailed enough to be broken down into item codes 062 and 162. Also, the only local government to which item code 022 applies is the District of Columbia (DC).

Appendix B: Cumulative Square Root of the Frequency Method

This appendix describes the cumulative square root of the frequency method, as used in the 2009 design of the Annual Survey of Public Employment and Payroll. Consider a stratum in which a first-stage probability-proportional-to-size sample has been selected with measure of size equal to total pay in 2007 (TotalPay07). The following steps are used to determine the cutoff:

1. Denote by min(TotalPay07) and max(TotalPay07) the minimum and maximum values of TotalPay07 in the first-stage sample, respectively
2. Let $I =$ (min(TotalPay07), max(TotalPay07))
3. Partition I into 20 subintervals of equal width
4. Denote by I_i the i^{th} subinterval, $1 \leq i \leq 20$
5. Denote by N_i the number of sample units in subinterval I_i
6. Calculate $C_i = \sum_{j=1}^{i} \sqrt{N_j}$
7. Find the smallest integer k, $1 \leq k \leq 20$, such that $C_k \geq C_{20}/2$
8. Set the cutoff equal to the upper end point of subinterval I_k

Example

Consider a stratum with 117 first-stage sample units. The information needed to determine the cutoff is summarized in Table B-1. This table has 20 rows, one for each subinterval. The columns Lower and Upper give the lower and upper end points of each subinterval. The smallest integer k such that $C_k \geq C_{20}/2 = 22.367$ is $k = 7$. Therefore, we set the cutoff equal to the upper end point of subinterval I_7, which equals 483,185. The small cutoff stratum consists of the 74 units in the first-stage sample with TotalPay07 ≤ 483,185, and the large cutoff stratum consists of the 43 units in the first-stage sample with TotalPay07 > 483,185.

Table B-1: Information needed to determine the cutoff

i	I_i		N_i	C_i	$C_i \geq C_{20}/2$
	Lower	Upper			
1	2,000	70,741	15	3.873	
2	70,741	139,482	13	7.479	
3	139,482	208,222	12	10.943	
4	208,222	276,963	10	14.105	
5	276,963	345,704	10	17.267	
6	345,704	414,445	7	19.913	
7	414,445	483,185	7	22.559	✓
8	483,185	551,926	4	24.559	✓
9	551,926	620,667	4	26.559	✓
10	620,667	689,408	3	28.291	✓
11	689,408	758,148	7	30.936	✓
12	758,148	826,889	4	32.936	✓
13	826,889	895,630	4	34.936	✓
14	895,630	964,371	1	35.936	✓
15	964,371	1,033,111	0	35.936	✓
16	1,033,111	1,101,852	4	37.936	✓
17	1,101,852	1,170,593	3	39.669	✓
18	1,170,593	1,239,334	2	41.083	✓
19	1,239,334	1,308,074	2	42.497	✓
20	1,308,074	1,376,815	5	44.733	✓