A Comparison of the Accuracy of Design-Based and Model-Based Estimation in the Presence of "All or Nothing" Data

Hamid Ashtiani, Wendy Rotz
Grant Thornton LLP, 1250 Connecticut Ave NW #400, Washington, DC 20035

Abstract
Model-based estimation has been used for decades now. It can produce much more accurate estimates than classical design-based approaches in a variety of settings. Of particular interest is tax settings with a goal of estimating a dependent variable, y, such as a qualifying amount or taxable amount, from a model built from an independent variable, x, that typically is some type of cost or expenditure. The independent variable tends to be highly skewed with many low values and fewer larger values; it typically fits a gamma distribution.

Classical regression assumes model residuals from sampled values are normally distributed around the regression line. However, this is not a requirement of model-based estimation.

In many tax settings, y=x with probability p, and y=0 with probability 1-p.

This paper reviews the theoretical foundation of applying model-based estimation in this common tax setting and provides simulations demonstrating its efficacy in comparison to common design-based alternatives used in tax — the Mean Per Unit (MPU) or Horvitz-Thompson estimator, and the difference estimator (DIFF).

Model-based methods were found to be superior to the design approaches in nearly all settings.

Key Words: Tax, All or Nothing, Model-Based Estimation, Design-Based Estimation, CV, MSE, Confidence Interval Coverage, Simulation

1. Introduction

Many applications in tax require estimation of qualifying expenses for a credit deduction, where the expenses (the auxiliary variable, x) have a highly skewed gamma distribution, and the qualifying expenses (the dependent variable, y) is either equal to its auxiliary variable, x, with probability p, or is zero with probability 1-p. That is,

$$y = \begin{cases} x, & with\ probability\ p \\ 0, & with\ probability\ 1-p \end{cases}$$

where p is a constant, but unknown, value between 0 and 1.

The accuracy of estimating total values using design-based and model-based estimation for "all or nothing" data is studied in this paper for this common tax setting.

The Internal Revenue Service (IRS) issued Revenue Procedure 2011-42 (and many revenue procedures and field directives) advising their Exam Teams how to review taxpayer conducted samples that estimate values for federal tax purposes. Revenue Procedure 2011-42 suggests four classical estimators a taxpayer might use; however, taxpayers are not restricted to these four

methods.[1] Model-based estimation is not one of the four methods, but it has been used successfully in tax applications for three decades now. The IRS Revenue Procedure 2011-42 estimation methods are all design-based: i) the Mean Per Unit (MPU), 2) the difference (DIFF) estimator, iii) combined ratio, and iv) combined regression. In this paper, the accuracy of model-model based estimation was compared to the alternatives that would be used out of IRS Revenue Procedure 2011-42. However, focus was restricted to just the MPU and DIFF estimators because the other two required additional assumptions[2] for use that were not uniformly met in all of the simulation scenarios.

2. Research Questions

In this research, the following questions were researched either through theoretical derivations or analysis of simulation results.

1. Are there any theoretical reasons that would preclude the use of model-based estimation?
2. How accurate is it compared to design-based methods?
3. Is there evidence of bias?
4. How is the confidence interval coverage?
5. Is its efficacy impacted by low or high values of p?

Therefore, this study began with revisiting the theoretical foundation of a model-based estimation but with the added assumption that it is applied to all or nothing data. Then simulations were conducted to compare the accuracy of model-based estimates to two most common alternatives in the same tax settings - the MPU and DIFF estimators - which are design-based estimators. Confidence interval coverage was analyzed in the simulations, along with considerations of potential bias. The simulations used a range of low to high values for p to understand the behavior of the method across a span of qualifying percentages.

3. Revisit Model-Based Estimation Theory in the Context of All or Nothing Data

First, there is a refresher on model-based estimation, making all theoretical adaptations as appropriate to the setting. Following the textbook, *Sampling Design and Analysis* by Sharon Lohr[3], the statistical foundation of model-based estimation was revisited in the context of all or nothing data.

3.1 A Line through the Origin

A line through the origin was modeled because when there are no expenses, nothing would qualify, and therefore, a model through the origin makes the most sense in this tax setting. The calculations below closely follow the nomenclature and derivations by Lohr, only making necessary changes to accommodate the special case of its application to all or nothing data.

---

[1] When taxpayers use methods beyond the Revenue Procedure 2011-42, in §4.05(4), Exam Teams are instructed to elevate the review of the statistical work to one of the IRS Statistical Coordinators who have more background and training in statistical methods.

[2] For the combined ratio and combined regression estimators, per Revenue Procedure 2011-42, the IRS restrictions are 1) the minimum non-certainty stratum sample size is 30, 2) total sample size is 100, 3) the Coefficient of Variation (CV) of both the dependent and independent variables must 15% or less and 4) for the combined ratio estimate, the dependent and independent variable must have the same sign.

[3] Sharon L. Lohr, PhD is widely published in the area of statistical sampling and methods for education, public policy, law, and crime. She is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute. She received the Gertrude M. Cox, Morris Hansen, and Deming Awards. She was formerly a Dean's Distinguished Professor of Statistics at Arizona State University as well as a Vice President at Westat. She is now an independent statistical consultant.

A linear relation between expenses and qualifying expenses is assumed. Furthermore, the relationship is considered to pass through the origin and assumed to have random noise, ε, as well as heteroscedastic variance - a function of x to be specified later.

3.2 Nomenclature and Assumptions

Below are the nomenclature and standard assumptions used by Lohr.

Let:

i = 1,2,3 ... represent the i^{th} item in the population,
n = the sample size,
N = the population count,
S = the set of units selected in the sample, and
S' = the set of units not selected in the sample.

Assume:

$X > 0 \ \forall \ i = 1,2, \dots, N$ and are all known for every item in the population,
y_i are known in the sample $\forall \ i = 1,2, \dots, n$,
$Y_1, Y_2, \dots, Y_N$ are independent, and
$Y_i = \beta x_i + \ \varepsilon_i$ where $E_M(\varepsilon_i) = 0$.[4]

Lohr explains that under the model,

$T_y = \sum_{i=1}^{N} Y_i$ is a random variable, [5]
$t_y = \sum_{i=1}^{N} y_i$ is the total population value of interest and is just one realization of the random variable T_y,
x_i are constants $\forall \ i$,
β is a constant, while $\hat{\beta}$ (see below) is a random variable, and
$\hat{\beta}$ is independent of $\sum_{i \notin S} Y_i$ since $\hat{\beta}$ is based solely on sampled records.

So far there has been no departure from Lohr's textbook. However, now two more assumptions are introduced to accommodate the all or nothing data. Assume:

$$Y_i = \begin{cases} x_i & with\ probability\ p \\ 0 & with\ probability\ 1-p \end{cases}$$

and

$$V_M(\varepsilon_i) = \ f(x_i)\ \sigma^2 \text{ where } f(x_i) \text{ is some function of } x_i.^{6}$$

3.3 A closer look at $\mathbf{E_M(\varepsilon_i)}$

Refer to the graphical representation of the model in Figure 3.3.1 below. The plot of the data is split with p percent of the data falling on the line $y_i = x_i$ and $(1-p)$ percent of the data falling on the line $y_i = 0$. The regression line, $y_i = \beta x_i$, falls somewhere between the two lines. The error terms are $\varepsilon_i = x_i \ - \beta x_i$ for p percent of the population and $\varepsilon_i = 0 - \beta x_i$ for $(1-p)$ percent for the remaining population. Therefore,

$$E_M(\varepsilon_i) = p\ (x_i \ - \beta x_i) + (1-p)(\ 0 - \beta x_i) = x_i(p - \beta)$$

---

[4] The expected value under the model assumptions is denoted E_M.

[5] Lohr points out this is the main distinction between design and model-based estimation. In design-based estimation, based on probability theory, T_y is considered fixed — albeit unknown and the random variables are the sample indicators. It is a very different paradigm from the traditional design-based estimation that relies upon probability theory. Model based point estimates are often consistent with design-based estimates (sometimes exactly equal depending on the model). However, the variance is very different, and the model determines the variance.

[6] Here, V_M is the variance under the model assumptions and departs from Lohr's derivation. In her text, $f(x_i)$=x_i.

and since $E_M(\varepsilon_i) = 0$, and $x_i > 0 \; \forall \; i$, it follows that $p = \beta$. Thus, the slop of the regression line is the proportion, p, of expenses that qualify.

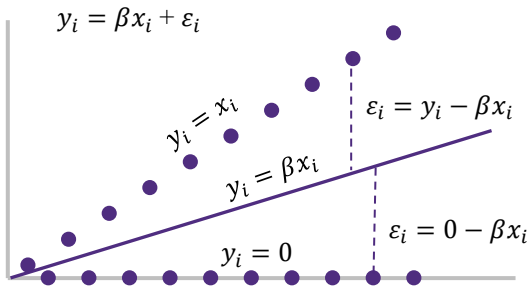

Figure 3.3.1: Illustration of Model with All of Nothing Data

Note the similarities of the plots in Figure 3.3.2 below. The first plot shows the regression line and two of the residuals when $p = 90\%$. Flipping each qualifying case to non-qualifying as well as each non-qualifying to qualifying, produces the next plot where $p = 10\%$. Note that the lengths of the corresponding residuals are the same. Therefore, the accuracy of model-estimators, which heavily depends on the size of the residuals, should be similar when p is close to zero or when p is close to one.

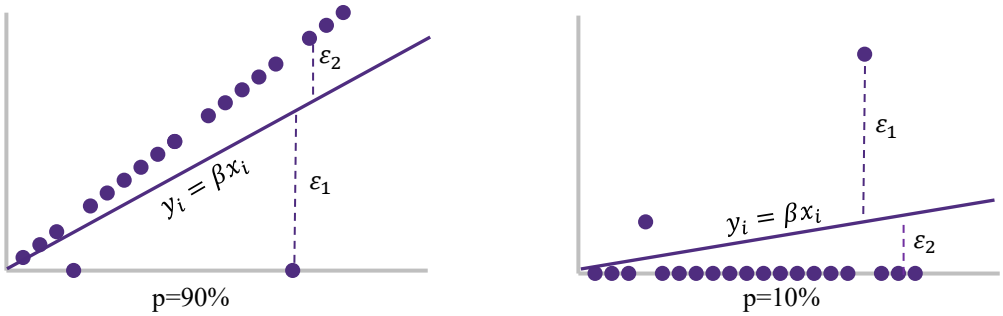

Figure 3.3.2: Mirror Images of Residuals from p=90% and p=10%

3.4 Heteroscedasticity of All or Nothing Data

Note that the $|\varepsilon_i|$ clearly increases with x_i, therefore, it is unsurprising that $\text{Var}(\varepsilon_i)$ increases with x_i as well. The calculations are as follows:

$$\begin{aligned} V_M(\varepsilon_i) &= E_M(\varepsilon_i^2) - E_M^2(\varepsilon_i) = E_M(y_i - \beta x_i)^2 - 0 \\ &= p(x_i - \beta x_i)^2 + (1-p)(\beta x_i)^2 \\ &= x_i^2 p(1-p). \end{aligned}$$

Therefore, take $f(x) = {x_i}^2$, and it follows that $\sigma^2 = p(1-p)$ so that

$$V_M(\varepsilon_i) = x_i^2 p(1-p) = x_i^2 \sigma^2.$$

Lohr's derivations used the variance assumption $V_M(\varepsilon_i) = x_i \sigma^2$, therefore, the model-based estimate and its variance require their own derivations in the presence of all or nothing data. The derivations closely follow Lohr's work with the main exception being the assumption of the variance structure.

3.5 Total, T_y and $\widehat{\beta}$ with All or Nothing Data

When estimating the total qualifying amount for a population, t_y, the model-based approach uses the values found in the sample, y_i, to predict the values for the units not sampled. This gives:

$$t_y = \sum_{i\epsilon S} y_i + \sum_{i\notin S} y_i = \sum_{i\epsilon S} y_i + \sum_{i\notin S} \hat{\beta} x_i$$

Find $\hat{\beta}$ to minimize the sum of the weighted squares in the errors: $\sum_{i\epsilon S} x_i^{-2} (y_i - \beta x_i)^2$.

Set the first derivative to 0 and solve:

$$\frac{\partial}{\partial \beta}\left[\sum_{i\epsilon S} x_i^{-2}(y_i - \beta x_i)^2\right] = 0$$

$$-2\sum_{i\epsilon S} x_i^{-2}(y_i - \beta x_i)\, x_i = 0$$

$$\sum_{i\epsilon S} y_i\, x_i^{-1} = n\beta.$$

Hence, take $\hat{\beta} = (1/n)\sum_{i\epsilon S} y_i / x_i$.

Use the known Y_i for sampled units and apply the model, $Y_i = \hat{\beta} x_i$, for non-sampled units to arrive at the model-based estimator:

$$\hat{T}_y = \sum_{i\epsilon S} Y_i + \hat{\beta} \sum_{i\notin S} x_i.$$

3.6 Consistency

Letting $r_i = y_i/x_i$, note that $\hat{\beta} = \bar{r}$ is the average ratio of y_i to x_i. Since $r_i = 1$ with probability p, and $r_i = 0$ with probability $1 - p$, $\hat{\beta}$ is also equivalent to the binomial estimate for p. While this makes intuitive sense, it should also be noted that $\bar{r}$ is an inconsistent[7] estimator of $R' = \sum_{i=1}^{N} Y_i / \sum_{i=1}^{N} X_i$.

3.7 Model-Unbiased Estimates

Note that $\hat{\beta}$ and $\hat{T}_y$ are model-unbiased estimates of β and T respectively:

$$E_M[\hat{\beta} - \beta] = E_M\left[\frac{1}{n}\sum_{i\epsilon S}\frac{Y_i}{x_i} - \beta\right] = E_M\left[\frac{1}{n}\sum_{i\epsilon S}\frac{\beta x_i + \varepsilon_i}{x_i} - \beta\right] = E_M\left[\beta + \frac{1}{n}\sum_{i\epsilon S}\frac{\varepsilon_i}{x_i} - \beta\right] = 0$$

and

$$E_M[\hat{T}_y - T] = E_M[\textstyle\sum_{i\epsilon S} Y_i + \hat{\beta}\sum_{i\notin S} x_i - \sum_{i\epsilon S} Y_i - \sum_{i\notin S} Y_i] = E_M[(\hat{\beta} - \beta)\sum_{i\notin S} x_i] = 0.$$

3.8 The Variance of $\hat{T}_y$

Note that:

$$V_M[\hat{T}_y - T] = V_M[\textstyle\sum_{i\epsilon S} Y_i + \hat{\beta}\sum_{i\notin S} x_i - \sum_{i\epsilon S} Y_i - \sum_{i\notin S} Y_i] = (\sum_{i\notin S} x_i)^2 V_M[\hat{\beta}] + \sum_{i\notin S} V_M[\hat{Y}_i].$$

Now,

$$V_M[Y_i] = V_M[(\beta x_i + \varepsilon_i)] = V_M[\varepsilon_i] = x_i^2\sigma^2.$$

Also,

---

[7] Consider a population with only two records, a \$1 expense that qualifies, and a \$1,000,000 expense that does not. Sampling the entire population results in $\bar{r} = 0.5$, while $R' = \frac{1}{1{,}000{,}001} = 0.000001$. Therefore, by definition, $\bar{r}$ is an inconsistent estimate of R' since a sample of the entire population does not result in the true population value R'.

$$V_M[\hat{\beta}] = V_M\left[\frac{\sum_{i\epsilon S} y_i/x_i}{n}\right] = \frac{1}{n^2}\sum_{i\epsilon S} x_i^{-2} V_M[y_i] = \frac{\sum_{i\epsilon S} x_i^{-2}x_i^2\sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

.

Hence

$$V_M[\hat{T}_y - T] = (\sum_{i\notin S} x_i)^2 \frac{\sigma^2}{n} + \sum_{i\notin S} x_i^2\ \sigma^2 = \left[(\sum_{i\notin S} x_i)^2\ /n + \sum_{i\notin S} x_i^2\right]\sigma^2.$$

Find σ^2 from the sum of weighted squares of the residuals.

$$E_M\left[\sum_{i\epsilon S} x_i^{-2}(y_i - \hat{\beta}x_i)^2\right] = E_M\left[\sum_{i\epsilon S}\left(\frac{y_i}{x_i} - \hat{\beta}\right)^2\right] = E_M\left[\sum_{i\epsilon S}\left(\frac{y_i}{x_i} - \frac{1}{n}\sum_{i\epsilon S}\frac{y_i}{x_i}\right)^2\right]$$

$$= E_M\left[\frac{1}{n}\sum_{i\epsilon S}\left(\sum_{i\epsilon S}\left(\frac{y_i}{x_i} - \frac{1}{n}\sum_{i\epsilon S}\frac{y_i}{x_i}\right)^2\right)\right]$$

$$= E_M\left[\frac{1}{n}\sum_{i\epsilon S}(n-1)\left(\frac{1}{n-1}\sum_{i\epsilon S}\left(\frac{y_i}{x_i} - \frac{1}{n}\sum_{i\epsilon S}\frac{y_i}{x_i}\right)^2\right)\right]$$

$$= \frac{(n-1)}{n} E_M\left[\sum_{i\epsilon S} Var\left(\frac{y_i}{x_i}\right)\right] = \frac{(n-1)}{n} E_M\left[\sum_{i\epsilon S}\frac{1}{{x_i}^2}{x_i}^2\sigma^2\right]$$

$$= \frac{(n-1)}{n} n\sigma^2 = (n-1)\sigma^2$$

So, take

$$\sigma^2 \cong \hat{\sigma}^2 = \sum_{i\epsilon S} x_i^{-2}(y_i - \hat{\beta}x_i)^2/(n-1).$$

4. Simulations

4.1 Generating Populations

Since financial data is highly skewed, the gamma distribution using the rgamma function in the statistical software R was used to simulate expenses (the auxiliary variable, x) of three populations of different sizes: small ($\boldsymbol{N = 100}$), medium ($\boldsymbol{N = 1,000}$), and large ($\boldsymbol{N = 10,000}$). The resulting distributions are shown below in Figure 4.1.1. The smoothest distribution, of course, is the largest one with $\boldsymbol{N = 10,000}$.

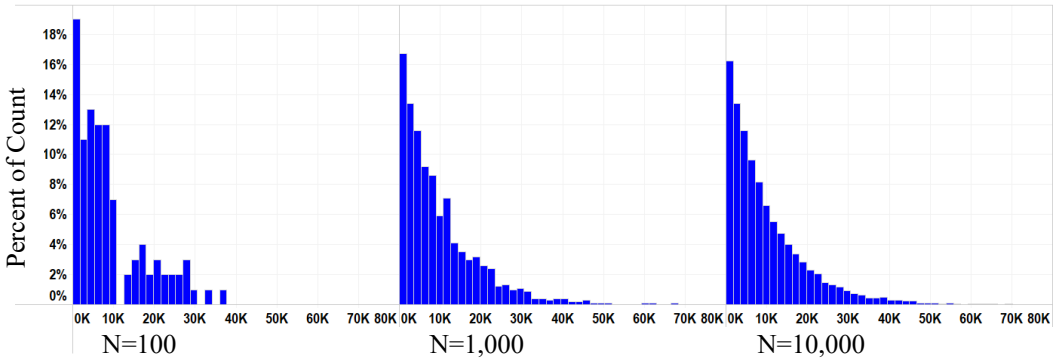

Figure 4.1.1: Distribution of Auxiliary Variable, X = Expenses

For each population, five dependent variables (qualifying amounts, y) were created corresponding to five levels of p: 10%, 25%, 50%, 75%, and 90%. To achieve this, for every x_i value in each of the three populations, five random numbers between zero and one were created for each level of p. For p =10% for example, the $y_{i,10\%}$ value was assigned equal to $x_i$with a 10% probability and it was assigned zero otherwise.

4.2 Sampling

After the populations were created, samples of three different sizes were drawn, $\boldsymbol{n = 15}$, $\boldsymbol{n = 30}$, and $\boldsymbol{n = 100}$. However, for the smallest population of $\boldsymbol{N = 100}$, only the two smallest sample sizes were drawn, as $\boldsymbol{n = 100}$ would have been a complete census in each draw and those results would have been uninformative. With two levels of $\boldsymbol{n}$ for the smallest population, three levels of $\boldsymbol{n}$ for the other populations, five levels of $\boldsymbol{p}$ for every population, and three total population sizes, altogether this made 40 scenarios in the various combinations of $\boldsymbol{N}, \boldsymbol{n},$ and $\boldsymbol{p}$.

For each of the 40 scenarios, 5,000 independent iterations of sample draws were performed.

4.3 Estimators

For each of the 5,000 samples from each of the 40 scenarios, the total qualifying amount, $\hat{Y}$, was estimated 3 ways as listed below.

1) The Difference Estimator (DIFF). This is a design-based approach and listed in IRS Revenue Procedure 2011-42. Rather than estimating the total of qualifying directly, the difference estimator, estimates the expense amount that does not qualify, and then subtracts that from the known total expense amount, $T_x = \sum_{i=1}^{N} x_i$. The estimate is calculated by:

$$\hat{Y}_{DIFF} = T_x - \mathrm{N}\sum_{i=1}^{n}(x_i - y_i)/n.$$

2) The Mean Per Unit (MPU) estimator. This is the Horvitz-Thompson estimator, which is another design-based approach listed in IRS Revenue Procedure 2011-42. Its formula is given by:

$$\hat{Y}_{MPU} = N\sum_{i=1}^{n}\frac{y_i}{n}\,.$$

3) Weighted Model-Based (MOD). This is the model-based estimator derived in Section 3 above; it is calculated from:

$$\hat{Y}_{MOD} = \hat{T}_y = \sum_{i\in S} Y_i \;\; + \hat{\beta}\sum_{i\notin S} x_i\,.$$

Note each of the three estimators was calculated on every sample drawn.

4.4 Efficacy Metrics

Once all 5,000 iterations of sample selection and estimation via the three methods above were completed for a scenario, the accuracies of the three estimators were compared by scenario using the metrics listed below.

1) Average Standard Error ($\overline{SE}$)

The average standard error is given by:

$$\overline{SE} = \frac{1}{5{,}000}\sum_{j=1}^{5{,}000} SE_j,$$

where for the j[th] iteration in a scenario,

$$SE_{j,DIFF} = \frac{N(N-n)}{n}\frac{\sum_{i=1}^{n}(d_{i,j}-\overline{d_j})^2}{n-1}, \text{ where}$$

$$d_{i,j} = x_{i,j} - y_{i,j}, \text{ and}$$

$$\overline{d_j} = \sum_{i=1}^{n} d_{i,j}/n;$$

$$SE_{j,MPU} = \frac{N(N-n)}{n}\frac{\sum_{i=1}^{n}(y_{i,j}-\overline{y_j})^2}{n-1}, \text{ where}$$

$$\overline{y_j} = \sum_{i=1}^{n} y_{i,j}/n; \text{ and}$$

$$SE_{j,mod} = [(\sum_{i\notin S} x_i)^2/n + \sum_{i\notin S} x_i^2]\,\sigma^2 \text{ where}$$

$$\sigma^2 \cong \hat{\sigma}^2 = \sum_{i\epsilon S} x_i^{-2}(y_i - \hat{\beta}x_i)^2/(n-1), \text{ and}$$

$$\hat{\beta} = (1/n)\sum_{i\epsilon S} y_i/x_i.$$

Guidance in IRS Revenue Procedure 2011-42 states that when there are multiple estimators that are appropriate to use, taxpayers should use the most accurate one, where accuracy is defined by each estimate's standard error.

2) Average Coefficient of Variation (CV)

The average coefficient of variation is given by:

$$\overline{CV} = \frac{1}{5{,}000}\sum_{j=1}^{5{,}000} \frac{SE_j}{\hat{Y}_j}.$$

The average CV was included in the analyses because it closely relates to the relative precision, RP, defined as the margin of error[8] divided by the estimated amount. The IRS focuses on RP when discerning whether estimates have met an acceptable level of accuracy.[9] However, unlike RP, the

[8] The margin of error for a desired level of confidence probability is the result of multiplying the standard error by a critical value corresponding to the confidence level under an appropriate assumption for the distribution of the estimate – such as the normal or student's t distribution.

[9] IRS Revenue Procedure 2011-42 allows taxpayers to use their estimated values as is when the RP is less than 10%. When the RP exceeds 15%, the IRS still allows the taxpayer to use an estimate, albeit a biased one. Instead of their point estimates, taxpayers are allowed to use the least advantageous bound (to the taxpayer) of a 95% one-sided confidence interval for the point estimate. So, for example if the taxpayer is estimating a credit or a deduction (which are to their advantage) and the RP is 20%, there is a 20% reduction of their estimated value. If estimating an amount owed with 20% RP, there is a 20% increase. Therefore, the taxpayer suffers a penalty for imprecision. There is a sliding scale bias from zero to a 15% adjustment of the estimate when the relative precision is between 10% and 15%.

This approach allows the IRS to avoid dictating minimal sample sizes in their guidance and instead establishes an expectation of accuracy as measured by RP.

When the IRS conducts their own samples in audit, the Internal Revenue Manual IRM 4.473, calls for the same approach, with the exception that they give the taxpayer the benefit of the imprecision if their estimated adjustment has a poor RP, unless they have worked out some other mutually agreed upon arrangement with the taxpayer or the taxpayer has been exceptionally egregious in their tax determinations. For example, if the IRS

CV does not rely upon distributional assumptions for the estimate. The CV and its related RP are descriptions of how much estimates may vary in relationship to their size.

3) Mean Squared of Error of the Estimators (MSE)

The mean squared error is given by:

$$MSE = \frac{1}{5{,}000} \sum_{j=1}^{5{,}000} \left(\hat{Y}_j - Y\right)^2,$$

where Y is the true know value from the simulated population. While Y, is not typically known in practice, it is useful in simulations to compare the average square of the difference between the estimates and the actual true value being estimated. For an SE, CV or RP, the amount of variability of an estimate is calculated from variability of the estimate in comparison to amount calculated from sample results — such as a sample mean. However, MSE is calculated from the estimate's variability compared to the actual known value in the population.

4) Confidence Interval Coverage:

$$Confidence\ Interval\ Coverage\ \% \ = \ C/5{,}000$$

where C is the number of the 5,000 iterations containing the actual total qualifying expenses in the population inside the 90% two-sided confidence interval around the estimated qualifying expenses.

5) Average Bias%:

The average percent bias is given by:

$$Average\ Bias\% = \tfrac{1}{5{,}000} \textstyle\sum_{j=1}^{5{,}000} \left(\hat{Y}_j - Y\right).$$

The average bias is the mean error. In theory, for unbiased estimates, the average bias approaches zero as the number of iterations approaches infinity. The average bias was divided by the true population Y in these analyses in order to compare results more easily across different levels of p.

5. Results

5.1 Average SE

The table below summarizes the average SE results across all scenarios. Note in general the SE for the DIFF tends to decrease as p increases. For each method, the table shading is lightest for the lowest SE in the scenario and darkens as the SE increases. Opposite of DIFF, for MPU the SEs increases as p increases. Meanwhile, the model-based estimations have roughly the same SEs for p=10% and p=90% as well as similar SEs for p=25% and p=75%. This is not surprising after considering the lengths of the residuals in Figure 3.3.2. The largest SEs for the model-based estimates are at p=50% in each scenario.

Also observe that the model-based estimator, with very few exceptions, has the lowest SE for each scenario. Therefore, it would typically be chosen as the estimator to use for federal tax purposes according to Revenue Procedure 2011 guidance of choosing the estimator with the lowest SE when multiple estimates are appropriate to use.

Table 5.1.1: Average SE for Each Scenario

Method	p	N=100		N=1,000			N=10,000		
		n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
DIFF	10%	199,729	131,255	2,345,753	1,676,749	901,070	24,125,538	17,566,633	9,734,750
	25%	183,865	121,072	2,301,020	1,653,817	887,382	23,311,803	17,044,904	9,391,062
	50%	188,114	122,821	1,994,217	1,446,558	780,598	20,388,774	15,025,488	8,472,016

draws a sample in audit and extrapolates the taxpayer owes an amount with 20% relative precision, the IRS will then discount that amount by 20% due to the imprecision of their estimate.

Method	p	N=100		N=1,000			N=10,000		
		n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
	75%	131,607	88,929	1,406,031	1,072,806	600,266	14,949,419	11,390,461	6,596,780
	90%	64,487	48,378	756,079	617,477	370,891	8,000,103	6,657,076	4,160,306
MPU	10%	79,995	59,974	762,972	628,556	373,254	7,747,104	6,342,038	4,014,787
	25%	148,473	102,139	1,396,590	1,042,263	578,272	14,705,938	11,085,304	6,522,921
	50%	156,223	105,416	2,004,914	1,469,950	794,651	20,734,369	15,174,504	8,514,295
	75%	199,865	130,656	2,271,334	1,633,882	874,783	23,274,404	17,001,190	9,442,617
	90%	204,395	133,183	2,345,093	1,686,193	900,178	24,154,767	17,455,480	9,738,846
MOD	10%	69,559	50,138	679,757	523,515	292,867	6,694,385	5,157,277	2,944,732
	25%	99,422	68,076	1,088,167	777,910	425,871	11,073,711	7,939,751	4,378,046
	50%	116,226	78,975	1,278,979	904,530	492,113	12,972,607	9,185,349	5,027,578
	75%	104,622	71,532	1,092,837	777,349	425,634	10,993,544	7,895,515	4,350,289
	90%	65,627	48,072	680,720	517,520	292,614	6,857,889	5,319,097	3,022,735

5.2 Average CV

See the figure below for typical average CV findings. This one is for a population of 100 expenses and a sample size of 15. The figure shows the average CVs for the five levels of p and the three estimation methods.

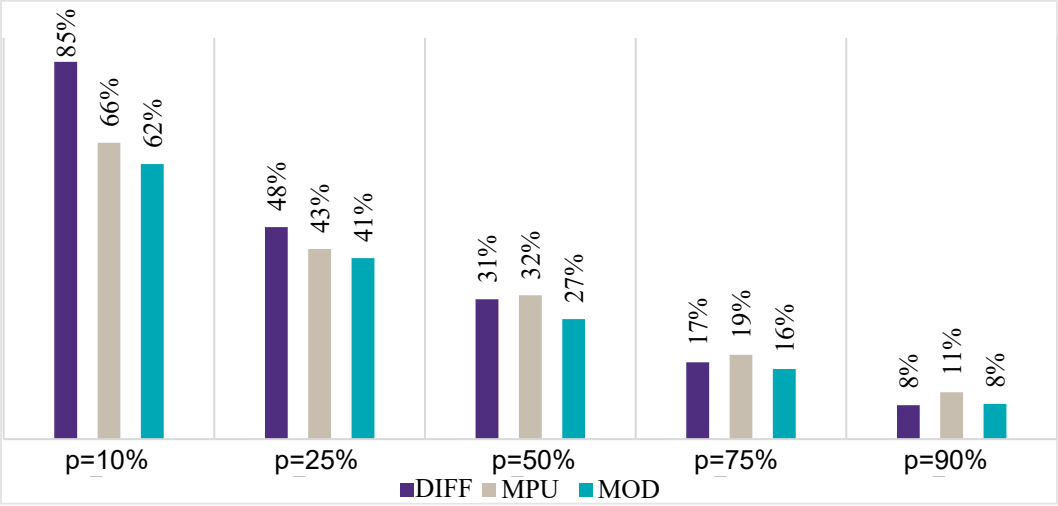

Figure 5.2.1: Average CV Results for N=100, n=15

It is true that the model-based estimator has a larger CV when $\boldsymbol{p}$ is small than it does when $\boldsymbol{p}$ is large. This may be the source of confusion of why some believed the model-based estimator performs poorly in cases of small $\boldsymbol{p}$.

However, lower levels of $\boldsymbol{p}$ have larger CVs for all three estimators. The size of the estimate is in the denominator of the CV calculation. A lower value of $\boldsymbol{p}$ will have a smaller estimate. Dividing by a smaller estimate yields a larger CV.

In fact, the model-based estimator consistently had the smaller CV of three estimation methods tested. The table below lists the average CVs for each of the 40 scenarios. It is shaded to create a heat map where lighter shades of blue are closer to zero and the darker colors are assigned to larger average CVs.

Note the largest and darkest shaded average CVs are for the difference estimator with smaller values of p. Note that the average SEs were higher for these scenarios with the difference estimator. That, combined with the estimated amount being smaller for smaller values of p, the CV — which is the ratio of the SE to the estimate — becomes quite large in these settings.

However, the average CVs for the difference estimator tend to be much smaller for larger values of p. Again, from Table 5.1.1, note that in these scenarios, the SE tends to be much smaller. That in combination with larger estimates for larger levels of p, the average CVs for the difference estimator are quite good (low) with higher levels of p. However, even when it is at its best, the average CV of the difference estimator never falls below the model-based estimator's average.

Unsurprisingly, it is also noted that average CVs were smaller for scenarios with a sample size of 100, which was the largest size in the study.

Table 5.2.2: Average CVs for Each Scenario
(Lightest Shading is Near Zero, Darkest is Near 1000%)

Method	p	N=100		N=1,000			N=10,000		
		n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
DIFF	10%	1183%	623%	951%	525%	298%	973%	489%	474%
	25%	256%	83%	629%	310%	56%	530%	289%	78%
	50%	127%	39%	102%	144%	17%	144%	50%	18%
	75%	24%	14%	37%	17%	9%	28%	17%	9%
	90%	9%	6%	9%	7%	4%	11%	8%	5%
MPU	10%	67%	56%	68%	67%	39%	67%	69%	43%
	25%	59%	39%	63%	46%	25%	63%	47%	26%
	50%	42%	27%	41%	29%	16%	43%	31%	17%
	75%	31%	20%	31%	22%	12%	32%	23%	13%
	90%	25%	16%	26%	19%	10%	27%	19%	11%
MOD	10%	63%	51%	62%	56%	29%	62%	59%	32%
	25%	48%	30%	51%	34%	18%	50%	34%	17%
	50%	29%	19%	26%	18%	10%	28%	19%	10%
	75%	17%	11%	15%	11%	6%	15%	11%	6%
	90%	8%	6%	8%	6%	3%	8%	6%	3%

5.3 MSEs

On the next page are the MSE results for the five levels of ***p*** and a population size of 100 with a sample of 15. Note that the model-based estimator consistently yielded the smallest MSE at each level of ***p***. This was found through-out all of the scenarios as shown in the table just below the figure.

To compare the methods more efficiently across the scenarios, the MSE from each scenario for the difference and MPU estimators were divided by the corresponding MSE for the model-based estimate in the scenarios. The results of this ratio are presented in the table under the figure. Each one is larger than 100%. That is, every MSE from a difference or MPU estimate was larger than the model-based estimate in these simulations.

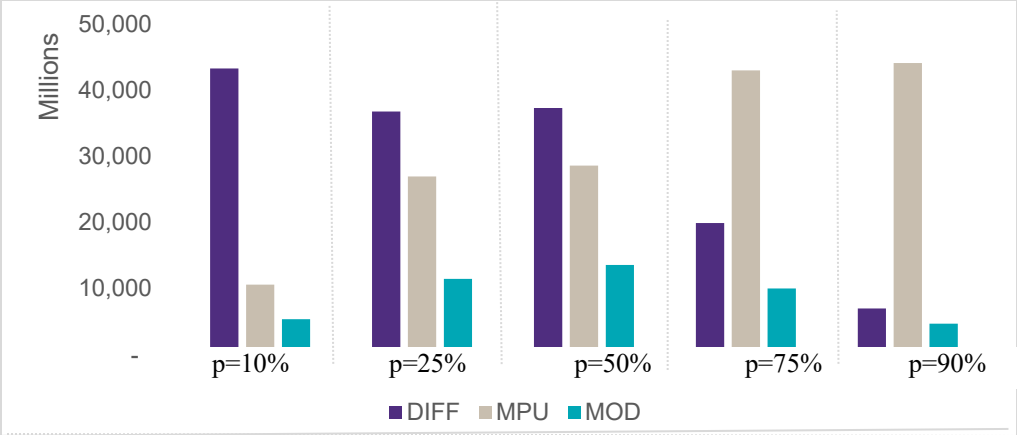

Figure 5.2.1: MSE for N=100, n=15

Table 5.3.1: Ratio of MSE Compared to the MSE of the Model-Based Estimator

p	Method	N=100		N=1,000			N=10,000		
		n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
10%	DIFF	801%	742%	877%	924%	890%	1,124%	1,141%	1,082%
	MPU	196%	193%	229%	235%	224%	215%	208%	208%
20%	DIFF	320%	262%	449%	442%	394%	463%	490%	474%
	MPU	235%	199%	229%	236%	208%	233%	237%	230%
50%	DIFF	274%	253%	287%	285%	266%	299%	298%	306%
	MPU	210%	203%	274%	270%	245%	303%	304%	302%
75%	DIFF	199%	190%	250%	245%	239%	235%	232%	225%
	MPU	429%	406%	435%	426%	437%	453%	483%	472%
90%	DIFF	149%	146%	177%	177%	170%	218%	215%	223%
	MPU	936%	935%	1,073%	1,011%	957%	1,012%	1,031%	1,061%

Therefore, on average, the model-based estimation method produced estimates closer to the true values than the difference and the MPU methods.

This is more easily discerned from the box plots.[10] In the figure below, the true value in the population is depicted by the long horizontal line for each level p. The range of estimates clearly shows those resulting from model-based estimation tend to be much closer to the actual value in the population compared to the alternatives of the difference and MPU methods. The inter-quartile ranges are narrower for the model approach compared to the other two estimators. Even the model-based "outliers" — those estimates beyond 1.5 times their interquartile range — tend to be closer to the actual value in most settings than the non-outliers of the other two methods.

Another familiar pattern is illustrated in the figure below. The difference estimator performs better than MPU when p exceeds 50% and MPU performs better than the difference method when p is below 50%. This was noted in other measures above as well. It is just more visually evident in the

[10] The estimated amounts calculated in the 5,000 iterations are depicted by the vertical spread for each of the three estimators at each level of p. Each box demarks the interquartile range - from the 25th to the 75th percentiles of the estimated values. For each method and level of p, the two short lines (one below the box and one above the box) are drawn at a distance from each box that is 1.5 times the difference between the 75th and 25th percentile. Per Tukey's definition, the estimated amounts beyond 1.5 times the interquartile range are considered outliers.

box plots. However, in each set of box plots, note that the spread of the estimated values from the model-based estimator is typically much lower than either of the design-based estimators.

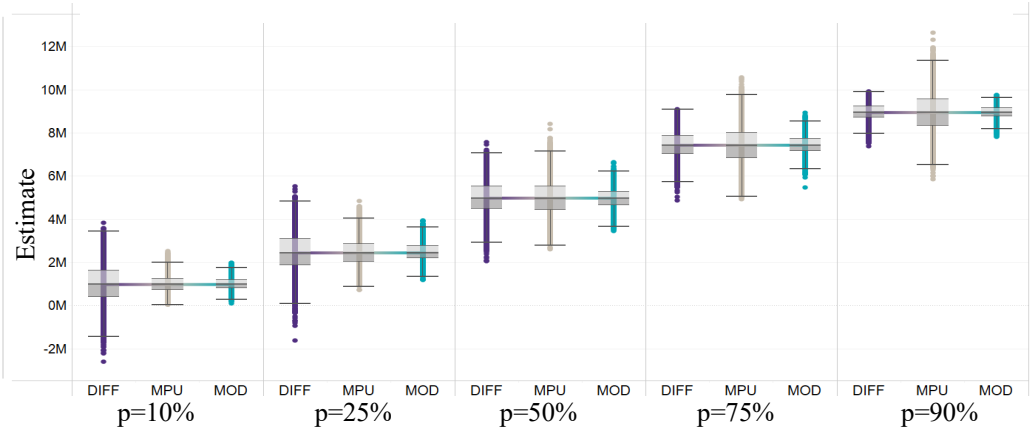

Figure 5.2.2: Box Plots of Estimates, N=1,000, n=100

5.4 Confidence Interval Coverage

In the figure below, the percent of confidence intervals containing the true population value is compared for the five levels of p when $N = 100$ and $n = 15$. Note, that there is indeed poor coverage with three of the five scenarios falling short of 90% with the model-based approach. This could be why there was concern in the other analysis with the use of model-based estimation. However, *all* of the scenarios are less than 90% with the other two alternatives. This was typical of the findings in the other scenarios.

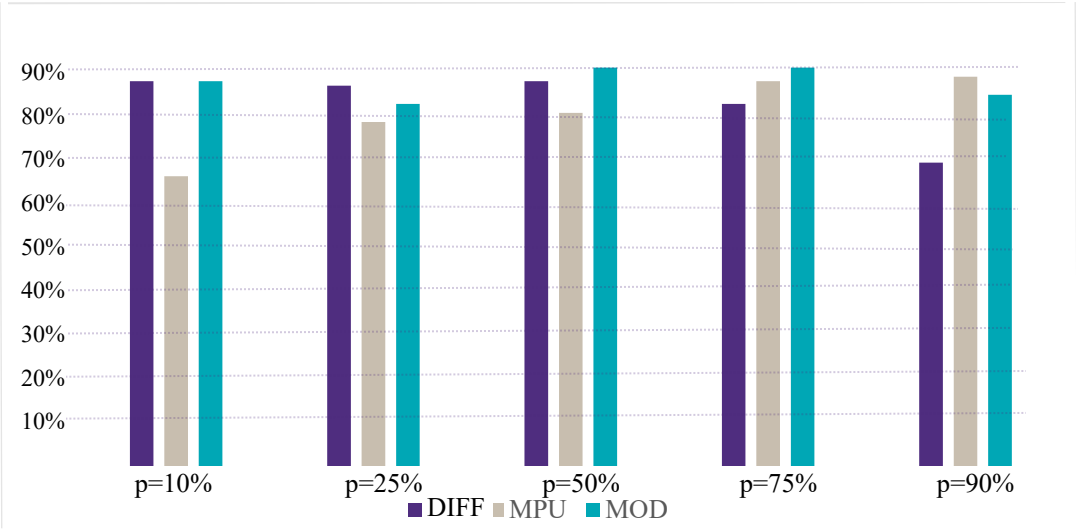

Figure 5.3.1: 90% Two-Sided Confidence Interval Coverage, N=100, n=15

A summary of results from the 40 scenarios is provided in the table below. Again, the table was made into a heat map with confidence interval coverage above 90% shown in blue, equal to 90% (within rounding) in white, and below 90% in red. Lighter colors are closer to 90% while darker colors are further away.

Note that only the model-based estimates exceeded 90%. There are also seven scenarios that had 90% confidence (within rounding) while the other estimators had only two each. The worst coverage

for the model-based estimation was 78%, 61% for MPU and 58% for the difference estimator. While 78% is far from 90%, it is closer than the other two values.

Table 5.4.1: Confidence Interval Coverage
(Lighter Colors are Closer to 90%. Darker Colors are Further Away.
Shades of Red are Below 90%. Shades of Blue are Above 90%.)

Method	p	N=100		N=1,000			N=10,000		
		n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
DIFF	10%	87%	89%	88%	88%	90%	86%	88%	89%
	25%	86%	89%	86%	87%	89%	86%	88%	90%
	50%	87%	89%	84%	87%	88%	83%	87%	89%
	75%	82%	87%	77%	82%	88%	77%	83%	88%
	90%	69%	74%	58%	74%	84%	62%	73%	83%
MPU	10%	66%	78%	62%	72%	85%	61%	72%	83%
	25%	78%	87%	77%	82%	88%	76%	81%	88%
	50%	80%	86%	85%	86%	89%	84%	87%	89%
	75%	87%	89%	87%	89%	90%	87%	88%	89%
	90%	88%	89%	87%	89%	90%	87%	88%	89%
MOD	10%	87%	90%	82%	86%	89%	78%	78%	91%
	25%	82%	85%	89%	87%	87%	90%	89%	91%
	50%	90%	91%	88%	90%	88%	89%	91%	89%
	75%	90%	92%	91%	89%	90%	90%	88%	88%
	90%	84%	89%	79%	79%	91%	79%	81%	88%

There are scenarios where the other two methods do have better coverage than model-based; however, those estimates are not the ones that would be used in the setting. For example, for all sample sizes, when $N = 1{,}000$ and $p = 10\%$, the difference estimator has better confidence coverage than the model-based estimator. The width of the confidence interval is dependent on the standard error. When p is below 50%, the difference interval has larger standard errors than the MPU estimator; this creates wider confidence intervals for the difference estimator and therefore better coverage.

However, that does not make it a better estimator. See the box plots above. Guidance in the IRS Revenue Procedure 2011-42 dictates using the estimator with the smaller standard error when choosing among those that are appropriate to use (meet the assumptions for their use). Of the design-based choices, that would be the MPU estimator in these cases – but note the MPU estimator has worse coverage than model-based for the same scenario $N = 1{,}000$ and $p = 10\%$.

It should be noted that in general poor confidence interval coverage was noted in *all* of the estimation methods. The problem lies with construction of the confidence intervals using the student's t-distribution which assumes an underlying normal distribution of the estimates.

The figure below shows the distribution of estimates for $\boldsymbol{N = 10{,}000}$, $\boldsymbol{n = 15}$, and $\boldsymbol{p = 90\%}$. It is clear the distributions of the estimates are not normally distributed bell-shaped curves. Even the MPU shows an asymmetric longer tail on the right than the left.

It is interesting to note though, as expected from the MSE analysis and that p exceeds 50% in this scenario, that the MPU has a wider range of estimated values than the difference estimator. In addition, as illustrated below and by the MSEs and the box plots, the model-based estimator has a narrower range of values than the other two. Its estimates are close to the actual value more frequently and do not vary as far from it. The difference estimator, which would be the one used

according to the revenue procedure, is more frequently far from the true value than it is close; although, its average over all 5,000 estimates does fall near actual population value.

The solid black lines marks where the true population value is, the dotted red lines note the average estimated amount.) In the figure below these lines fall on top of each other, which is expected for unbiased estimates and a large number of iterations.

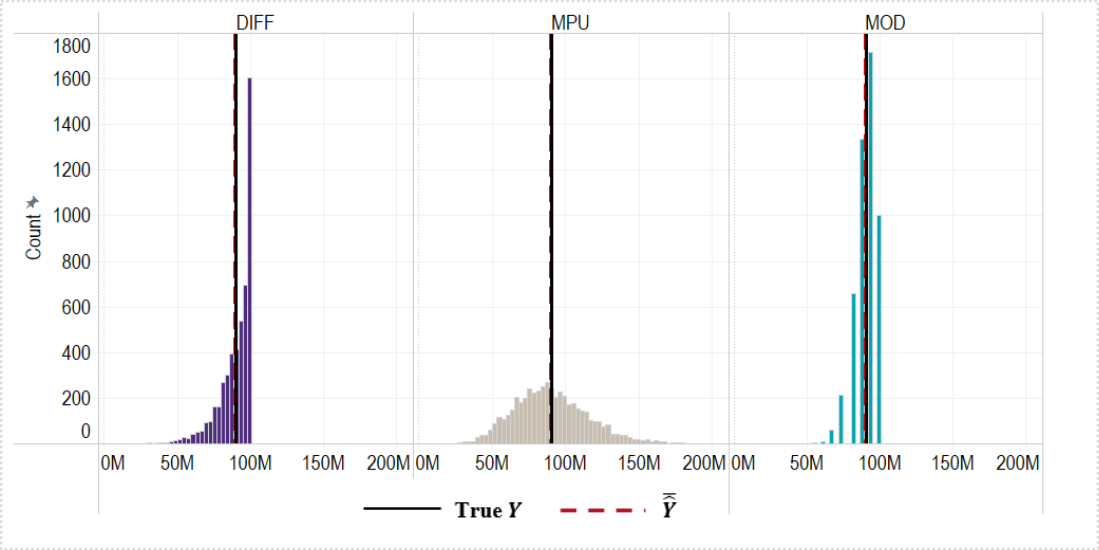

Figure 5.3.2: Distribution of Estimates for N=10,000, n=15, p=90%

5.5 Bias

While reviewing the distribution plots, a discrepancy in the model-based estimators was noted sporadically in a few of the scenarios. For example, the figure below shows the distribution plots for the scenario where N=100, n=15, and p=25%. The average estimate over the 5,000 iterations does not quite align with the actual value for the model-based estimator.

From this figure, it is clear that the model-based approach is still closer to the actual value far more frequently than the other two estimators and it has a narrower range of estimated values than the wider spread of the other two.

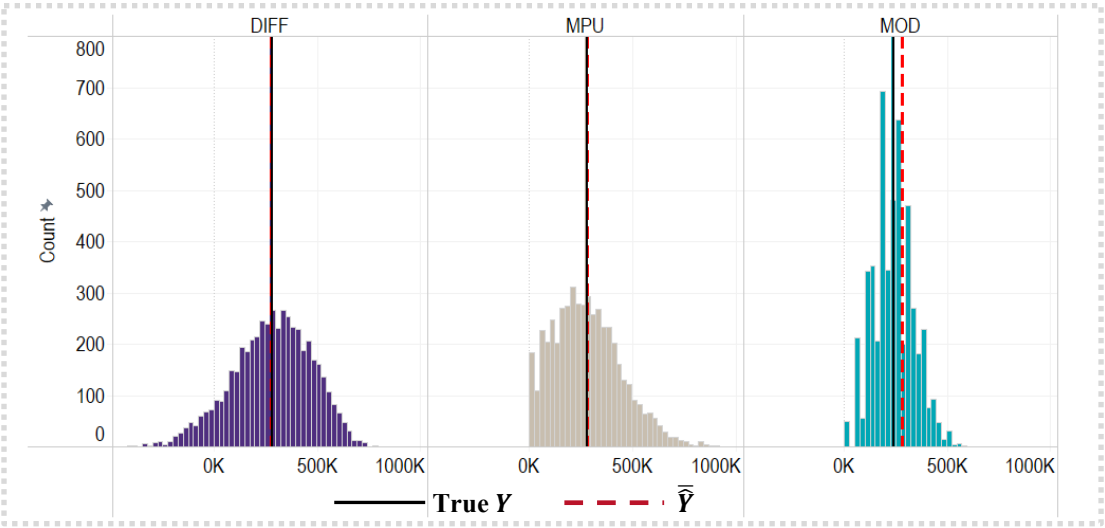

Figure 5.4.1: Distribution of Estimates for N=100, n=15, p=25%

As a side note, the apparent missing left tail of the MPU estimator was due to no estimates falling below zero. It is impossible for any actual values to be less than zero the way the data were constructed for these simulations. The scales of the horizontal axes are the same for all three plots. The difference estimator falls below zero in many of the iterations due to nonsense resulting from the calculation when the estimate of the non-qualifying amount exceeds the total sum of the expenses (auxiliary variable, x) in the population.

However, that is a distraction from the more pressing question: why is there bias at all? In Section 3.7, it was shown that the estimate model is unbiased. This led to many more questions to consider:

1 How bad/frequent was the apparent bias?
2 When does it occur and why?
3 What does "model-unbiased" truly mean?
4 Was the model mis-specified?
5 Were the calculations in Section 3.7 demonstrating model-unbiased, correct?
6 Is this a result of the inconsistency identified in Section 3.6?
7 Were the data simulated correctly — as described in Section 6.1 — according to the assumptions laid out in Section 3.2?

The table below summarizes the bias percentage (difference between the average estimate and actual value expressed as a percent) of the model-based estimator after the 5,000 iterations in each of the 40 scenarios. There was little/negligible bias found in the higher levels of $\boldsymbol{p}$ and in the largest population size. The problematic scenarios were in the smallest population for lower levels of $\boldsymbol{p}$. Interestingly, sample size, was less of a factor.

Table 5.5.1: Bias% of Model-Based Estimator Across Scenarios
(Lighter Colors are Closer to Zero. Darker Colors are Further Away.
Shades of Red Under-Estimated on Average. Shades of Blue Over-Estimated on Average.)

	N=100		N=1,000			N=10,000		
p	n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
10%	3.6%	3.8%	2.0%	3.5%	2.1%	1.3%	2.6%	1.8%
25%	15.0%	12.2%	7.2%	6.4%	6.1%	0.6%	0.5%	-0.2%
50%	-9.2%	-7.7%	-3.4%	-3.4%	-3.1%	1.5%	1.2%	1.3%
75%	-0.2%	0.0%	-0.7%	-0.9%	-0.5%	-1.5%	-1.3%	-1.2%
90%	1.0%	0.8%	0.7%	0.8%	0.9%	-0.2%	-0.1%	-0.1%

The smallest sample size of 15 was problematic for the smallest population of 100 as it had less bias for the mid-size population of 1,000 and it had small to negligible bias for the largest population of 10,000. While sample size did play a role, the population size was more influential. Figure 4.1.1 was reconsidered. The largest population was a smoother distribution compared to the choppier smaller populations. Was that somehow reducing bias?

The table below lists the actual $\boldsymbol{p}$ and actual $\boldsymbol{R}'$ attained in the population construction, as well as the average estimated beta, $\bar{\hat{\boldsymbol{\beta}}} = \sum_{j=1}^{5{,}000} \hat{\boldsymbol{\beta}} \,/5000$, from each of the 40 scenarios. The lighter shadings are values close to the intended p, while darker shading is further away. Blue shading indicates values above the intended p, while red indicates values below.

Table 5.4.2: Actual p, Actual R', and Average $\hat{\beta}$ by Scenario
(Lighter Colors are Closer to the intended p. Darker Colors are Further Away.
Shades of Red are Below the Intended p, Shades of Blue are Above.)

N	n	Intended p	10%	25%	50%	75%	90%
100	15	Actual p	12.0	25.0	47.0	71.0	89.0
		Actual R'	12.6	30.4	42.4	70.9	90.1
		$\bar{\hat{\beta}}$	12.1	25.1	47.0	71.1	89.1

N	n	Intended p	10%	25%	50%	75%	90%
		Actual p	12.0	25.0	47.0	71.0	89.0
	30	Actual R'	12.6	30.4	42.4	70.9	90.1
		$\bar{\hat{\beta}}$	11.9	25.1	47.1	71.0	89.1
		Actual p	11.3	24.4	52.2	74.9	89.9
	15	Actual R'	11.6	26.2	50.5	74.3	90.8
		$\bar{\hat{\beta}}$	11.3	24.3	52.2	74.9	90.1
		Actual p	11.3	24.4	52.2	74.9	89.9
1,000	30	Actual R'	11.6	26.2	50.5	74.3	90.8
		$\bar{\hat{\beta}}$	11.2	24.5	52.2	75.0	90.0
		Actual p	11.3	24.4	52.2	74.9	89.9
	100	Actual R'	11.6	26.2	50.5	74.3	90.8
		$\bar{\hat{\beta}}$	11.3	24.4	52.2	74.7	89.9
		Actual p	9.6	25.5	49.9	74.9	89.7
	15	Actual R'	9.9	25.5	50.5	74	89.7
		$\bar{\hat{\beta}}$	9.7	25.3	49.7	75.2	89.9
		Actual p	9.6	25.5	49.9	74.9	89.7
10,000	30	Actual R'	9.9	25.5	50.5	74	89.7
		$\bar{\hat{\beta}}$	9.6	25.4	49.9	75	89.8
		Actual p	9.6	25.5	49.9	74.9	89.7
	100	Actual R'	9.9	25.5	50.5	74	89.7
		$\bar{\hat{\beta}}$	9.7	25.5	49.9	74.9	89.7

The lightest shadings of blue and pink show that percentages for the actual p, actual R', and average estimated beta, $\bar{\hat{\boldsymbol{\beta}}}$, are all within a percentage point of the intended p when p=90% and/or when N=10,000. These are also within a percentage point for most sample sizes when N=1,000 and p =25% or p=75%. The smallest population, when N=100, had the farthest values from the intended p.

While exploring whether consistency was the issue for the observed bias — and whether that was related to population size, it became apparent that the mechanism of creating the simulated levels of $\boldsymbol{p}$ was part of the issue. In each scenario, every record in the population was given a chance of qualifying; the chance was $\boldsymbol{p}$. As expected, $\bar{\hat{\boldsymbol{\beta}}}$, the estimated beta, averaged over the 5,000 iterations and was indeed approximately equal to the actual $\boldsymbol{p}$ simulated in the population but, not necessarily the intended p. The population construction did not necessarily result in precisely $\boldsymbol{p}$ percent of the records qualifying nor did the resulting actual $\boldsymbol{p}$ equal $\boldsymbol{R'} = \sum_{i=1}^{N} \boldsymbol{Y_i} / \sum_{i=1}^{N} \boldsymbol{X_i}$.

The presence or absence of a few large records in the tails of the distribution could swing R' to differ from the actual p and intended p — as noted in the smaller population with several cells shaded either a dark blue or dark red.

For example, when intended p=25%, N=100, and n=30, the actual p is within rounding equal to the intended p, but the actual R'=30.4%, more than 5% above the intended p in for the population. This explains the root of the inconsistency when estimating R' from $\hat{\beta}$.

The presence or absence of a few large records in the tails of the distribution could swing R' to differ from the actual p.

The simulated populations were updated to force the attained p to be closer to the intended p in each scenario. To achieve this, the population of size N was divided into Np groups of equal counts (or nearly equal within rounding). Then a random record within each group was assigned $Y_i = x_i$ while

the remaining Y_i in the group were assigned to be equal to zero. This forced the p attained in the population to equal the intended p.

In addition, for diagnostic purposes, to also force the actual R' to be closer to the actual p - at least for the next set of analyses, the population listing was sorted from largest to smallest value of x_i prior to dividing the population into the Np groups.

This forced a more even distribution of large and small qualifying records in the population for the next set of analyses. It constructed a population where high valued expenses *actually were* qualifying *just as frequently* as low valued expenses — rather than *merely given an equal chance* to qualify during the population construction.

The 5000 iterations of sample draws and estimations were then repeated for each of the 40 scenarios. The table below summarizes the resulting average estimated β together with the actual p and R' in the population.

Note in the updated populations, every actual p is now equal to the intended p. Every value of R' is now within a percentage point of the intended p, as are all values of $\bar{\hat{\beta}}$.

Table 5.4.3: Actual p, Actual R', and Average $\hat{\beta}$ by Scenario on Updated Population
Lighter Colors are Closer to the intended p. Darker Colors are Further Away. Shades of Red are Below the Intended p, Shades of Blue are Above.

	n	Intended p	10%	25%	50%	75%	90%
100	15	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	10.3	24.2	50.4	75.0	90.1
		$\bar{\hat{\beta}}$	9.9	24.9	49.8	75.2	90.0
	30	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	10.3	24.2	50.4	75.0	90.1
		$\bar{\hat{\beta}}$	9.9	25.0	50.1	75.1	90.1
1,000	15	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	9.9	24.6	50.2	74.9	90.2
		$\bar{\hat{\beta}}$	10.1	25.1	49.7	74.9	89.9
	30	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	9.9	24.6	50.2	74.9	90.2
		$\bar{\hat{\beta}}$	10.2	24.9	49.9	75.1	90.0
	100	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	9.9	24.6	50.2	74.9	90.2
		$\bar{\hat{\beta}}$	10.0	25.1	49.9	75.0	90.0
10,000	15	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	10.1	24.9	49.9	75.1	90.0
		$\bar{\hat{\beta}}$	10.1	25.1	50.2	75.1	90.1
	30	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	10.1	24.9	49.9	75.1	90.0
		$\bar{\hat{\beta}}$	10.1	25.1	50.0	75.1	89.9
	100	Actual p	10.0	25.0	50.0	75.0	90.0
		Actual R'	10.1	24.9	49.9	75.1	90.0
		$\bar{\hat{\beta}}$	10.0	25.0	50.0	74.9	90.0

Furthermore, as demonstrated in the table below, the bias in the updated population was significantly reduced when the actual $\boldsymbol{R'}$ was closer to the actual $\boldsymbol{p}$; it is minimal/negligible in this updated data.

Table 5.4.4: Reduced Bias% of Model-Based Estimator Across Scenarios for Updated Populations

	N=100		N=1,000			N=10,000		
p	n=15	n=30	n=15	n=30	n=100	n=15	n=30	n=100
10	-3.6	-2.8	2.6	2.8	1.1	0.1	-0.3	-1.2
25	2.3	2.0	1.8	0.9	1.6	0.8	0.7	0.0
50	-1.2	-0.4	-0.9	-0.5	-0.5	0.5	0.1	0.2
75	0.3	0.1	0.0	0.2	0.1	0.1	-0.1	-0.3
90	0.0	0.1	-0.3	-0.2	-0.2	0.1	-0.1	0.0

The figure below shows the updated distributions for the same scenario as Figure 7.5, where $\boldsymbol{N = 100, n = 15}$, and $\boldsymbol{p = 25\%}$. The average estimate, $\bar{\bar{Y}}$ now closely aligns with the true value for the model-based estimator.

From these results, it was concluded that consistency was an issue influencing the bias in the original simulated populations.

The important question is whether this inconsistency could be a problem in practice. The answer is probably not.

First, even though it created some bias in the original populations, the model-based estimates were less variable and tended to be closer to the true values than the two alternatives of the difference and MPU estimators that would be most likely used if model-based estimation were not applied. This was demonstrated by the distributional figures, the box plots, average SEs, average CVs and the MSEs.

Second, in practice, the sample design is typically stratified by the size of the expenses, $\boldsymbol{x_i}$. When $\boldsymbol{R'}$ is significantly different across size strata, separate models are constructed by stratum, or by groups of strata with similar $\boldsymbol{R'}$. Therefore, in general, the stratification provides natural embedded protection from extreme differences resulting from the presence or absence of a few extreme large values selected in the sample.

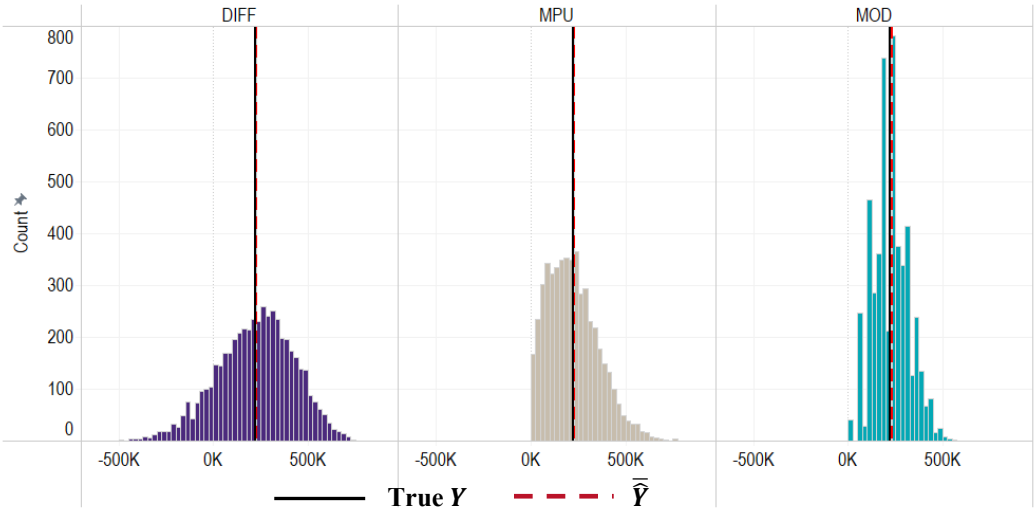

Figure 5.4.2: Distribution of Estimates in Revised Population, N=100, n=15, p=25%

6. Conclusions

None of the estimators consistently have 90% coverage or better. Confidence interval coverage is relatively equal across the three estimators for p values in the middle range, 25%, 50%, and 75%. It is more uneven with extreme values of p (10%, and 90%). The problem is in using the t-distribution as p gets further from 50%.

When deciding which estimate to use for federal tax purposes out of two or more statistically appropriate choices, the guidelines in the IRS Revenue Procedure 2011-42 has taxpayers using the estimator with the smallest standard error. When $p < 50\%$, the MPU would be used — but the model-based estimator has better confidence interval coverage in this case. When $p > 50\%$, the difference estimator would be used, but again, the model-based approach has better confidence interval coverage in these cases. Hence, in federal tax settings, the model-based estimate has better confidence coverage than the alternative that would be used from design-based approaches.

Some bias was found in the model-based estimator, despite being regarded as "model unbiased". However, across the board, the model-based estimator outperforms the design-based estimators in CV, SE, and MSE. The distributions, especially illustrated by the box plots, also demonstrate that the model-based estimates are less variable and are closer to the actual values more frequently in the scenarios tested. The model-based estimate was still overall the better choice in all scenarios tested.

That said, use caution in practice when the sampled $\hat{\boldsymbol{p}} = \hat{\boldsymbol{\beta}}$ differs too far from the $\boldsymbol{r'} = \sum_{i=1}^{n} \boldsymbol{y_i} / \sum_{i=1}^{n} \boldsymbol{x_i}$, found in the sample. This can result in an inconsistent and biased estimate. In the scenarios tested in this analysis, despite the bias from the inconsistency, the model-based estimator still performed better than the alternative that would be used according to IRS Revenue Procedure 2011-42. Alternatively, a stratified approach could resolve the issue. When there are disparate $\boldsymbol{r'_h}$ observed or anticipated, where the strata are enumerated by h=1, 2, …, separate stratum estimates can be calculated.

In terms of the original research questions regarding the efficacy of model-based estimation when applied to all or nothing data, the table summarizes the findings.

Table 8.1: Summary of Conclusions on Impetus Questions

	Question	**Findings**
1	Accuracy and Appropriateness	**The theory is fine** – with the correct variance structure. Additionally, **model-based estimates were consistently closer to the true values** when compared to the design-based estimator alternatives that would most likely be applied in the same settings.
2	Bias	Despite the theory demonstrating model-unbiased, **some bias was noted** stemming from $\hat{\beta}$ that is an inconsistent estimate of the ratio of total qualifying to total non-qualifying monetary amounts and simulated data that did not have populations with an equal outcome of large and small values with equal qualifying rates. However, even in the presence of this issue, **the model-based still proved to be a more accurate estimator than the alternatives** in the scenarios tested in this analysis.
3	Confidence Interval Coverage	**All of the estimators had poor coverage.**

		Of the estimators analyzed, **the model-based approach had better coverage** than the design-based estimator that would be applied according to IRS Revenue Procedure 2011-42.
4	Infrequent Events When p is Small	**All estimators had poor CVs** (and would therefore, also have poor RPs) for the lowest value of p tested in this analysis. **The model-based approach consistently had better CVs than the alternatives.** In actuality, the **residuals** for say, p=10% **are mirror images** of the residuals for p=90%. The **model based estimators had better SEs and MSEs** than the other estimators.

In short, there was nothing found in the analysis that would lead these researchers to believe the design-based alternatives of a difference or MPU estimator would be a better choice than a model-based estimator in the presence of all or nothing data.

The model-based estimation approach provided the more accurate estimates compared to these alternatives, had better confidence interval coverage and on average were closer to the true value than the design-based approaches.

7. Next Steps

Every answer yields further questions, and more analyses can always be performed. Below is a list for future research.

1. Simulate more populations with the auxiliary variable varying the distribution parameters such as skewness and kurtosis and distance between p and R'.

2. Expand analyses to include "all or nothing, or something in between." In such analysis,

$$= \begin{cases} x_i & ,with\ probability\ p \\ 0 < z_i < x_i & ,\ with\ probability\ q \\ 0 & ,with\ probability\ 1-p-q \end{cases}$$

   Many taxpayer's facts and circumstances fall into this more complex structure. It tends to have smaller variances though because the residuals from the second outcome are shorter than the residuals from an all or nothing setting.

3. Improve confidence interval coverage.
   a) Determine more appropriate distribution statistics than the t-distribution when constructing confidence intervals.
   b) Explore and compare non-distributional techniques for determining confidence intervals, such as bootstrapping or jackknifing.

4. Further analyze bias.
   a) Investigate the impact of consistency issues on bias and hindrance to convergence more thoroughly.
   b) Determine a rule of thumb for settings when there may be too much bias.

5. Add stratified sample designs to the analyses, including a certainty stratum for extreme values.
    a) Extrapolating by stratum, will improve the variances for the difference and MPU estimators. It will be interesting to determine whether under stratified designs, the model based so clearly outperforms the other design based approaches.
    b) Stratification should reduce the consistency issue and therefore the bias found in simulations.
    c) Potentially, this could also make the estimates more normally distributed on a stratum-by-stratum basis and thus possibly improve confidence interval coverage of the total estimates taken from the sum over the strata.

8. Acknowledgements

Many thanks to Richard Valliant, Zhenyu (Rick) Liu, and Phillip Kott. Rick Liu gave the researchers many issues to consider. Richard Valliant encouraged the authors to perform the theoretical statistical derivations. Phil Kott encouraged the exploration of using model-based estimation.

9. References

Batcher M. & Liu Y. (2003), "Ratio Estimation of Small Samples Using Deep Stratification" *Proceedings of the 2003 Joint Statistical Meetings, Survey Methodology Section,* ASA: Alexandria, VA

Brewer K. (1999, "Design-based or Prediction-based Inference? Stratified Random vs. Stratified Balanced Sampling" *International Statistical Review*, 67, 1, 35-47.

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

Dorfman A., Valliant R., and Royall R. (2000), *Finite Population Sampling and Inference: A Prediction Approach*, John Wiley & Sons: New York, NY

IRS Revenue Procedure 2011-42 https://www.irs.gov/pub/irs-drop/rp-11-42.pdf

IRS Internal Revenue Manual Part 4. Examining Process Chapter 47. Computer Audit Specialist Section 3. Statistical Sampling Auditing Techniques (IRM 4.47.3) https://www.irs.gov/irm/part4/irm_04-047-003

Liu Y., Batcher M., and Rotz W. (2001), "Application of the Hypergeometric Distribution in a Special Case of Rare Events" *Proceedings of the 2001 Joint Statistical Meetings,* ASA: Alexandria, VA

Mosteller F., Tukey J. (1982) Understanding Robust and Exploratory Data Analysis 1st Edition John Wiley & Sons: New York, NY

Lohr S. (1999), *Sampling: Design and Analysis*, Duxbury Press: Pacific Grove, CA

Rotz W., Yang J., and Joshee A., (2006), "Degrees of Freedom and Confidence Interval Coverage in Complex Model Based Sampling and Estimation" *Proceedings of the 2006 Joint Statistical Meetings - Section on Survey Research Methods.* ASA: Alexandria, VA

Rotz W., Joshee A., and Yang J., (2006) "Confidence Interval Coverage in Complex Model Based Estimation" *Proceedings of the 2006 Joint Statistical Meetings, Survey Methodology Section* ASA Alexandria, VA

Royall R. and Cumberland W. (1981), "An Empirical Study of Ratio Estimator and Estimators of Its Variance" *Journal of the American Statistical Association* 76, 373, 66-77.

Royall R. and Cumberland W. (1981), "The Finite-Population Linear Regression Estimator and Estimators of Its Variance – An Empirical Study" Journal of the American Statistical Association 76, 376, 924-930.

Royall R. and Herson J. (1973), "Robust Estimation in Finite Populations" *Journal of the American Statistical Association* 68, 344, 880-889.

Sarndal, Swenson, and Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag: New York, NY

Tam S. & Chan N. (1984), "Screening of Probability Samples" *International Statistical Review*, 52, 3, 301-308.