Southampton Statistical Sciences Research Institute
University of Southampton

Transformed Variables in Survey Sampling

Raymond L. Chambers, Alan H. Dorfman

Abstract

It can happen, especially in economic surveys, that we are interested in estimating the population mean or total of a variable Y, based on a sample, when a linear model seems appropriate, not for Y itself, but for a strictly monotone transformation of Y. In the present paper, we mainly focus on the important case where the transformation is logarithmic, but some new ideas introduced are not limited to that case. Currently available methods, based on the lognormal distribution, are reviewed, and two new methods introduced, one based on the idea of "smearing" (Duan, 1983), which do not require the lognormal assumption. Theoretical biases and variances are given, with suggestions for sample design and variance estimation, and a practical measure for reducing sensitivity to deviant points is suggested. We evaluate and compare the different estimators we describe in an extensive empirical study based on four economic populations taken from the UK Monthly Wages and Salaries Survey.

S^{3}RI Methodology Working Paper M03/21

Transformed Variables in Survey Sampling

Raymond L. Chambers[1] & Alan H. Dorfman[2]

(1) Southampton Statistical Sciences Research Institute, University of Southampton

(2) U.S. Bureau of Labor Statistics, Washington DC

ABSTRACT

It can happen, especially in economic surveys, that we are interested in estimating the population mean or total of a variable *Y*, based on a sample, when a linear model seems appropriate, not for *Y* itself, but for a strictly monotone transformation of *Y*. In the present paper, we mainly focus on the important case where the transformation is logarithmic, but some new ideas introduced are not limited to that case. Currently available methods, based on the lognormal distribution, are reviewed, and two new methods introduced, one based on the idea of "smearing" (Duan, 1983), which do not require the lognormal assumption. Theoretical biases and variances are given, with suggestions for sample design and variance estimation, and a practical measure for reducing sensitivity to deviant points is suggested. We evaluate and compare the different estimators we describe in an extensive empirical study based on four economic populations taken from the UK Monthly Wages and Salaries Survey.

KEY WORDS: balanced sampling, jackknife variance estimator, lognormal linear model, prediction, probability proportional to size sampling, residuals, simple random sampling, stratified random sampling, smearing estimator, outlier robust estimator

1. Introduction

Given a population of N units, we wish to predict the finite population total $T = \sum_{i=1}^{N} y_i$ of a variable of interest Y, based on a sample s of size n from that population. In addition to the sampled values of Y, we have auxiliary information in the form of population values x_i, $i = 1, .., N$ of a covariate X. The standard approach to this task (see Royall, 1982) assumes a linear relationship between Y and X. Often, however, there is good reason to think that the relationship between Y and X themselves is not linear, but linear in another scale of measurement, so that we have

$$h(Y) = \beta_0 + \beta_1 g(X) + \varepsilon \,, \tag{1}$$

where β_0, β_1 are unknown parameters, we allow for a transform of X (possibly X itself), and the errors ε have mean 0 and variance σ^2. The question then becomes: how do we make an inference concerning T, based on the available data, using this model? Allowing for transformation of X does not of course by itself carry us beyond the standard linear model; the essential difficulty posed by (1) is in handling the transform of the dependent variable Y. In the present paper we focus mainly on the case where h is the (natural) logarithm *log*, and we also assume that $g(x) = log(x)$, so that the special case of interest is the log-log model

$$\log(Y) = \mathbf{Z}'\beta \;\; + \varepsilon \,, \tag{2}$$

where $\mathbf{Z}' = (1 \quad \log(X))$ and $\beta = (\beta_0 \beta_1)'$.

The use of transformations in inference has a long history, and has been much studied (e.g. Deming 1984 [original publication 1943], Carroll and Ruppert 1988), but not a great deal has been done in the sampling context. Chen and Chen (1996) considered an approach based on empirical likelihood, restricting its use to attainment of confidence intervals. Their results improved on earlier coverage attained using robust variance estimators based on a linear model

(Royall and Cumberland, 1985). Karlberg (2000a, 2000b) assumed the errors ε were normal (so that Y has a lognormal distribution) and developed predictors with negligible biases. We review predictors that assume lognormality in Section 2. Section 3 introduces two new predictors of total: a SMEARING predictor, based on ideas in Duan (1983), and a ratio-adjusted-for-sample-total (RAST) predictor. Approximations to their biases and variances are given; the respective jackknife variance estimators are approximately unbiased for these variances. Vulnerability to data values that deviate from the model is noted, and modifications that improve the robustness of the proposed methods are described. Section 4 describes an extensive empirical study, evaluating several of the approaches proposed in this paper. Section 5 states conclusions.

2. Predictors based on the lognormal model

A too simple response to model (2) is to use optimal linear methods to get an (ordinary least squares) estimate $\mathbf{b}_{ols}$ of β, back-transform to get predicted values of Y at non-sample values, and use these to predict $T_r = \sum_r y_i$, the non-sample component of T. Here r denotes the set of non-sampled population units. This gives

$$\hat{T}_A = \sum_s y_i + \hat{T}_{r,A} = \sum_s y_i + \sum_r h^{-1}(\mathbf{z}_i'\mathbf{b}_{ols}) = \sum_s y_i + \sum_r e^{\mathbf{z}_i'\mathbf{b}_{ols}},$$

the naïve back-transformation predictor of T.

That this is not very satisfactory is readily seen. Suppose the errors are normally distributed, $\varepsilon \sim N(0,\sigma^2)$. Then Y has a lognormal distribution, and we have $E(Y \mid X) = e^{\mathbf{z}'\beta+\sigma^2/2}$, so that $E(T_r) = \sum_r e^{\mathbf{z}_i'\beta+\sigma^2/2}$. $\hat{T}_A$ will be biased low, since $E\left(\hat{T}_{r,A}\right) = \sum_r e^{\mathbf{z}_i'\beta+\mathbf{z}_i'\,\mathrm{var}(\mathbf{b}_{ols})\mathbf{z}_i/2}$, and $\mathbf{z}_i'\,\mathrm{var}(\mathbf{b}_{ols})\mathbf{z}_i$ is of lower order than σ^2.

This suggests as remedy, what might be called the naïve lognormal predictor,

$$\hat{T}_B = \sum_s y_i + \sum_r e^{\mathbf{z}_i'\mathbf{b}_{ols} + s^2/2},$$

where $s^2 = (n-2)^{-1}\sum_s (\log(y_i) - \mathbf{z}_i'\mathbf{b}_{ols})^2$. However, this is also biased for *T*:

$$\begin{aligned} E(e^{\mathbf{z}_i'\mathbf{b}_{ols} + s^2/2} - y_i) &\approx \frac{e^{\mathbf{z}_i'\beta + \sigma^2/2}}{2}\left(\mathbf{z}_i' var(\mathbf{b}_{ols})\mathbf{z}_i' + \frac{1}{4}var(s^2)\right) \\ &= \frac{e^{\mathbf{z}_i'\beta + \sigma^2/2}}{2}\left(\sigma^2\mathbf{z}_i'(\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{z}_i + \frac{\sigma^4}{2(n-2)}\right). \end{aligned}$$

Here $\mathbf{Z}_s$ denotes the matrix of sample values of *Z*. If $k_i = 1 + \frac{1}{2}\left(\sigma^2\mathbf{z}_i'(\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{z}_i + \frac{\sigma^4}{2(n-2)}\right)$, then

$E\left(\frac{e^{\mathbf{z}_i'\mathbf{b}_{ols} + s^2/2}}{k_i} - y_i\right) \approx 0$, and we can use the estimate of this factor

$$\hat{k}_i = 1 + \frac{1}{2}\left(s^2\mathbf{z}_i'(\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{z}_i + \frac{s^4}{2n}\right) = 1 + \frac{s^2 a_{ii}}{2} + \frac{s^4}{4n},$$

where $a_{ii} = \mathbf{z}_i'(\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{z}_i$, to get the first order bias corrected lognormal predictor

$$\hat{T}_C = \sum_s y_i + \sum_r \hat{k}_i^{-1} e^{\mathbf{z}_i'\mathbf{b}_{ols} + s^2/2}.$$

Karlberg (2000a, 2000b) employs something very close to this:

$$\hat{T}_K = \sum_s y_i + \sum_r e^{\mathbf{z}_i'\mathbf{b}_{ols} + \frac{s^2}{2}(1-a_{ii}) - \frac{s^4}{4n}} = \sum_s y_i + \sum_r \hat{l}_i^{-1} e^{\mathbf{z}_i'\mathbf{b}_{ols} + s^2/2},$$

where $\hat{l}_i = e^{\frac{s^2 a_{ii}}{2} + \frac{s^4}{4n}} \approx \hat{k}_i$. Under the lognormal assumption, this predictor has O(n^{-2}) bias, and can be expected to perform well, provided the lognormal model holds, or nearly holds.

3. The RAST and SMEARING Predictors

The preceding transformation-based predictors use bias adjustments that assume a normal distribution for the transformed variable. We introduce two new predictors that escape this restriction and have other desirable properties.

3.1 Ratio Adjustment by Sample Totals (RAST)

A method of predicting the non-sample total T_r of Y should be able to exactly recover the (known) sample total of this variable. If it does, then the method yields an unbiased predictor of this sample total, and we can anticipate that it will also give a close to unbiased predictor of T_r, and hence of T. Let $\hat{y}_i$ denote the predicted value of y_i under the method of interest. Then this requirement translates into the condition $\sum_s y_i = \sum_s \hat{y}_i$.

None of the lognormal predictors discussed in the previous section possess this property. However, for an arbitrary estimator $\mathbf{b} = (b_0 b_1)'$ of $\boldsymbol{\beta}$, it is not difficult to modify the naïve back-transformation predictor so that it does. Put $\gamma(\mathbf{b}) = \log \sum_s y_i - \ln \sum_s e^{\mathbf{z}_i'\mathbf{b}}$ and define $\mathbf{b}^* = (b_0 + \gamma(\mathbf{b}) b_1)'$. It is easy to see that $\sum_s e^{\mathbf{z}_i'\mathbf{b}^*} = \sum_s y_i$. The resulting predictor of T is

$$\hat{T}_{RAST} = \sum_s y_i + \sum_r e^{b_0^* + b_1^* \ln x_i} = \sum_s y_i + \frac{\sum_s y_i}{\sum_s x_i^{b_1}} \sum_r x_i^{b_1}, \qquad (3)$$

which we term the Ratio Adjustment by Sample Total (RAST) predictor. More generally, we can consider using weighted sample sums in the numerator and denominator of the second term. Even more general, for the model (1), is

$$\hat{T}_{RAST} = \sum_s y_i + \frac{\sum_s w_i y_i}{\sum_s w_i h^{-1}(\mathbf{z}_i'\mathbf{b})} \sum_r h^{-1}(\mathbf{z}_i'\mathbf{b}). \qquad (4)$$

We assume the weights w_i when normalized to be of order n^{-1}.

3.2 The SMEARING Predictor

For predicting a Y-value at $\mathbf{Z}$, where Y obeys the model (2), Duan (1983) suggested estimating $E(Y \mid \mathbf{Z}) = \int e^{\mathbf{Z}'\beta+\varepsilon} dF(\varepsilon)$ by $\hat{E}(Y \mid \mathbf{Z}) = n^{-1}\sum_s e^{\mathbf{Z}'\mathbf{b}_{ols}+R_i}$, where the R_i are the sample residuals from the ordinary least squares (*ols*) fit of ln (y_i) on $\mathbf{z}_i$. For an arbitrary estimator $\mathbf{b} = (b_0 b_1)'$ of $\boldsymbol{\beta}$ this leads naturally to the corresponding SMEARING predictor of the population total:

$$\begin{aligned} \hat{T}_{SMEAR} &= \sum_s y_i + \sum_r \hat{E}(y_i \mid \mathbf{z}_i) \\ &= \sum_s y_i + \sum_r n^{-1}\sum_s e^{\mathbf{z}_i'\mathbf{b}+R_i} \\ &= \sum_s y_i + \left(\sum_r x_i^{b_1}\right)\left(n^{-1}\sum_s \frac{y_i}{x_i^{b_1}}\right). \end{aligned} \tag{5}$$

Observe that for the log-log model, the RAST predictor in (3) is a ratio of means estimator, and the SMEARING predictor in (5), a mean of ratios estimator, in the auxiliary x^{b_1}. Again we can easily extend this to a weighted version. The generalization for the model (1) is

$$\hat{T}_{SMEAR} = \sum_s y_i + \sum_{j\in r}\sum_{i\in s} w_i h^{-1}\left(\mathbf{z}_j'\mathbf{b} + R_i\right) \tag{6}$$

where the weights w_i add to 1 and are of order n^{-1}.

3.3 Biases

3.3.1 Bias of the RAST Predictor

We consider first the log-log model (2). One readily sees that

$$E(T_r) = e^{\beta_0} E\left(e^{\varepsilon}\right)\sum_r x_i^{\beta_1}. \tag{7}$$

Assume $\mathbf{b} = \beta + O_p\left(n^{-1/2}\right)$. The non-sample part the RAST predictor (3) can then be written

$$\hat{T}_{r,RAST} = \frac{\sum_s w_i e^{\beta_0} x_i^{\beta_1} e^{\varepsilon_i} \sum_r x_i^{\beta_1} x_i^{O_p\left(n^{-1/2}\right)}}{\sum_s w_i x_i^{\beta_1} x_i^{O_p\left(n^{-1/2}\right)}} = \frac{\sum_s w_i e^{\beta_0} x_i^{\beta_1} e^{\varepsilon_i} \sum_r x_i^{\beta_1}}{\sum_s w_i x_i^{\beta_1}}\left(1 + O_p\left(n^{-1/2}\right)\right)$$

with expectation $E\left(\hat{T}_{r,RAST}\right) = e^{\beta_0} E\left(e^{\varepsilon}\right) \sum_r x_i^{\beta_1} \left(1 + O\left(n^{-1/2}\right)\right)$, so that the RAST predictor (3) for the log transform is almost unbiased under (1). Using second order Taylor expansions of $x_i^{b_1}$, we find that the multiplier $\left(1 + O_p\left(n^{-1/2}\right)\right) \approx 1 + A\left(b_1 - \beta_1\right) + B\left(b_1 - \beta_1\right)^2$, where A is a constant and

$$B = -\frac{\sum_s w_i x_i^{\beta_1} \log\left(x_i\right)}{\sum_s w_i x_i^{\beta_1}} \frac{\sum_r x_i^{\beta_1} \log\left(x_i\right)}{\sum_r x_i^{\beta_1}} + \frac{1}{2} \frac{\sum_r x_i^{\beta_1} \left(\log\left(x_i\right)\right)^2}{\sum_r x_i^{\beta_1}} - \frac{1}{2} \frac{\sum_s w_i x_i^{\beta_1} \left(\log\left(x_i\right)\right)^2}{\sum_s w_i x_i^{\beta_1}}.$$

We note that, under "weighted balance" (see below) the last two terms of B will tend to cancel, and the RAST predictor will have a negative $O(Nn^{-1})$ bias, provided enough of the x_i exceed 1.

The bias of the generalized RAST predictor (4) is not in general necessarily of low order, even under (1). However, under *weighted balance*, it is almost unbiased, even if the model (1) does not hold, as the following development shows.

We suppose the values of X in the population can be characterized by a density $d_P(x)$, and the sample values by a density $d_s(x)$. If the sample fraction is small, then the non-sample density $d_r(x)$ approximates $d_P(x)$. Suppose also the weights w_i derive from a function $w(x)$. Then we say the sample has *approximate weighted balance* if $w(x)d_s(x) \propto d_P(x)$. We add the qualifier "approximate" since these functions are smooth approximations to the granular reality. For further discussion of weighted balance, and methods for achieving it, see Valliant *et al*. (2000, Chapter 3).

Now suppose the working model is $h\left(y_i\right) = \mathbf{z}_i'\beta + \varepsilon_i$, with $\varepsilon_i \sim \left(0, \sigma^2\right)$, with y_i's independent, but the truth is

$$g\left(y_i\right) = m\left(x_i\right) + v\left(x_i\right)^{1/2} \eta_i, \tag{8}$$

with $\eta_i \sim \left(0, \tau^2\right)$, again with y_i's independent. Then

$$E\left(T_r\right) = \sum_r E\left\{g^{-1}\left(m\left(x_i\right) + v\left(x_i\right)^{1/2} \eta_i\right)\right\} = \sum_r \Omega\left(x_i\right),$$

with $\Omega(x_i) \equiv E\left\{g^{-1}\left(m(x_i) + v(x_i)^{1/2}\eta_i\right)\right\}$. We can therefore write $E(T_r) \approx (N-n)\int \Omega(x)\, d_r(x)dx$.

On the other hand, for the weighted RAST predictor given by $\hat{T}_{r,RAST} = \sum_r h^{-1}(\mathbf{z}_i'b)\dfrac{\sum_s w_i y_i}{\sum_s w_i h^{-1}(\mathbf{z}_i'b)}$,

we have, for *n*, *N* large,

$$E\left(\hat{T}_{r,RAST}\right) \approx \sum_r E\left\{h^{-1}(\mathbf{z}_i'b)\right\}\frac{\sum_s w_i E(y_i)}{\sum_s w_i E\left\{h^{-1}(\mathbf{z}_i'b)\right\}} = \sum_r \Psi(x_i)\frac{\sum_s w_i E(y_i)}{\sum_s w_i \Psi(x_i)},$$

where we have set $\Psi(x_i) = E\left(h^{-1}(\mathbf{z}_i'b)\right)$. We can write

$$E\left(\hat{T}_{r,RAST}\right) \approx (N-n)\int \Psi(x)\, d_r(x)dx \frac{\int w(x)\Omega(x)\, d_s(x)dx}{\int w(x)\Psi(x)\, d_s(x)dx},$$

and it is readily seen that this is $(N-n)\int \Omega(x)\, d_r(x)dx \approx E(T_r)$, under weighted balance, for $d_r(x) \approx d_P(x)$.

3.3.2 Bias of the SMEARING Predictor

For the log-log version (5) of the SMEARING predictor, we find that, to second order,

$$\hat{T}_{r,SMEAR} \approx e^{\beta_0}\left\{\sum_s w_i e^{\varepsilon_i}\sum_r x_i^{\beta_1} + C(b_1-\beta_1) + D(b_1-\beta_1)^2\right\},$$

where *C* is a constant and

$$D = -\sum_r x_i^{\beta_1}\log(x_i)\sum_s w_i \log(x_i)e^{\varepsilon_i} + \frac{1}{2}\sum_r x_i^{\beta_1}\left(\log(x_i)\right)^2\sum_s w_i e^{\varepsilon_i} + \frac{1}{2}\sum_r x_i^{\beta_1}\sum_s w_i\left(\log(x_i)\right)^2 e^{\varepsilon_i}.$$

The following result is helpful in assessing *D*.

Lemma. For a_i and u_i non-negative,

$$\sum_{j=1}^J a_j b_j^2 \sum_{i=1}^I u_i + \sum_{j=1}^J a_j \sum_{i=1}^I u_i c_i^2 \geq 2\sum_{j=1}^J a_j b_j \sum_{i=1}^I u_i c_i.$$

This inequality holds trivially for $I = J = 1$, and can be proved by induction: assuming it holds for fixed *I, J*, one shows it holds for $I + 1$ and *J.* The expression is symmetric in the *i* and *j* terms, so that, likewise, its holding for any *I, J* implies it holds for *I* and $J + 1$.

Letting $a_j = x_j^{\beta_1}$, $u_i = w_i e^{\varepsilon_i}$, and $b_k = c_k = \log(x_k)$ we see that in general *D* is positive, so that $\hat{T}_{r,SMEAR}$ has a positive bias of order Nn^{-1}.

The general SMEARING predictor (6) will be first order unbiased under model (1). The j^{th} non-sample term of $\hat{T}_{r,SMEAR}$ can be written

$$\sum_s w_i \left\{ h\left(\beta_0 + \beta_1 x_j + \varepsilon_i\right) + \delta_{ij} h'\left(\beta_0 + \beta_1 x_j + \varepsilon_i\right) + O_p\left(\delta_{ij}^{\ 2}\right) \right\},$$

where $\delta_{ij} = \left(b_1 - \beta_1\right)\left(x_j - x_i\right)$. The expectation of the first term coincides with the expectation of the j^{th} term of T_r. Duan (1983) has shown that, under mild conditions, the SMEARING predictor at a point is weakly consistent, and this carries over to the predictor of total.

What if the working model is wrong? As above, let the working model be $h\left(y_i\right) = \mathbf{z}_i'\beta + \varepsilon_i$, with $\varepsilon_i \sim \left(0,\sigma^2\right)$, and y_i's independent, and suppose the truth is (8), so that, again, $E\left(T_r\right) \approx \left(N - n\right)\int \Omega(x)\, d_r(x)dx$ as in the development after equation (8). As an alternative to the SMEARING predictor, we consider its "twiced" version with prediction component of the form

$$\tilde{T}_r = \sum_r \sum_s \varphi_{ij} h^{-1}\left(z_j'\hat{\beta} + R_i\right) + \sum_s w_{i'} \left\{ Y_{i'} - \sum_s \varphi_{ii'} h^{-1}\left(z_{i'}'\hat{\beta} + R_i\right) \right\}.,$$

where the φ-weights are positive, add to 1 for each *j*, and are of order 1/*n.*

Letting $g\left(x_i, x_j\right) = E\left\{ h^{-1}\left(z_j'\hat{\beta} + R_i\right) \right\}$, we write its expectation as

$$\begin{aligned} E\left(\tilde{T}_r\right) &= \sum_r \sum_s \varphi_{ij} g\left(x_i, x_j\right) + \sum_s w_{i'} \left\{ \Omega\left(x_{i'}\right) - \sum_s \varphi_{ii'} g\left(x_i, x_{i'}\right) \right\} \\ &\approx \left(N - n\right) n \int\int \varphi(u,v) g(u,v)\, d_s(u) d_r(v) du dv \\ &\quad + n \int w(v) \Omega(v)\, d_s(v) dv - \left(N - n\right) \int\int w(v) \varphi(u,v) g(u,v)\, d_s(u) d_s(v) du dv . \end{aligned}$$

This reduces to $(N-n)\int \Omega(x)\, d_r(x)dx \approx E(T_r)$, if the w-weights are such that

$$nw(v)d_s(v) = (N-n)d_r(v). \qquad (9)$$

That is, weighted balance, with the appropriate size adjustment, yields approximate unbiasedness of the twiced SMEARING estimator, despite mis-specification of the model.

In the simulations described below, we explored "histogram weights", defined by letting w_i be the number of non-sample units j having $|\hat{y}_j - \hat{y}_i| \le R/n$, for $i \in s$. This should yield an approximate version of the weights (9); these weights are like kernel weights in non-parametric regression estimation; furthermore, by basing them on (tentative) fitted values, the "curse of dimensionality" would be avoided, were Y dependent on more than one X variable. For the φ-weights, we tried both $\varphi_i = 1/n$, referred to as the "plain vanilla" version below, and also $\varphi_i = w_i$.

3.4 Variances and Variance Estimation

In the simulation studies described in Section 4 below, we use a jackknife variance estimator to form confidence intervals. Ignoring lower order terms, we here show the unbiasedness of this estimator for the variances of the RAST and SMEARING predictors of total. We do this for the general case, assuming in each case unbiasedness (to low order) of the corresponding predictors, which holds for the SMEARING predictor under (1), and for the RAST predictor in the log-log case and in general under favourable weighting structure (see above). We also assume the sampling fraction is sufficiently small so that $\mathrm{var}(\hat{T}_r - T_r) \approx \mathrm{var}(\hat{T}_r)$.

Then in the case of the RAST predictor, we have

$$\hat{T}_{r,RAST} \approx \sum_s w_i y_i \frac{\sum_r h^{-1}(\beta_0 + \beta_1 x_i)}{\sum_s w_i h^{-1}(\beta_0 + \beta_1 x_i)},$$

so that

$$\mathrm{var}\left(\hat{T}_{r,RAST}\right) \approx \sum_s w_i^2 \,\mathrm{var}(y_i)\left\{\frac{\sum_r h^{-1}(\beta_0+\beta_1 x_i)}{\sum_s w_i h^{-1}(\beta_0+\beta_1 x_i)}\right\}^2 .$$

The jackknife variance estimator is $v_{J,RAST} = \frac{n-1}{n}\sum_{j\in s}\left(\hat{T}_{r,RAST[\,j]} - \hat{T}_{r,RAST}\right)^2$, where we take

$$\hat{T}_{r,RAST[\,j]} \approx \sum_{s-\{j\}} w_i y_i \frac{\sum_{r+\{j\}} h^{-1}(\beta_0+\beta_1 x_i)}{\sum_{s-\{j\}} w_i h^{-1}(\beta_0+\beta_1 x_i)}.$$

It follows that

$$\hat{T}_{r,RAST[\,j]} - \hat{T}_{r,RAST} \approx \frac{\sum_r h^{-1}(\beta_0+\beta_1 x_i)}{\sum_s w_i h^{-1}(\beta_0+\beta_1 x_i)}\left\{\sum_{s-\{j\}} w_i y_i \frac{w_j h^{-1}(\beta_0+\beta_1 x_j)}{\sum_{i\in s-\{j\}} w_i h^{-1}(\beta_0+\beta_1 x_i)} - w_j y_j\right\}.$$

Under the assumption that $\hat{T}_{RAST}$ is (nearly) unbiased, we have

$$\begin{aligned} E\left(\hat{T}_{r,RAST[\,j]} - \hat{T}_{r,RAST}\right)^2 &\approx \mathrm{var}\left(\hat{T}_{r,RAST[\,j]} - \hat{T}_{r,RAST}\right) \\ &\approx \left\{\frac{\sum_r h^{-1}(\beta_0+\beta_1 x_i)}{\sum_s w_i h^{-1}(\beta_0+\beta_1 x_i)}\right\}^2 \left\{\sum_{s-\{j\}} w_i \,\mathrm{var}(y_i)\left(\frac{w_j h^{-1}(\beta_0+\beta_1 x_j)}{\sum_{i\in s-\{j\}} w_i h^{-1}(\beta_0+\beta_1 x_i)}\right)^2 + w_j^2 \,\mathrm{var}(y_j)\right\} \\ &\approx \left\{\frac{\sum_r h^{-1}(\beta_0+\beta_1 x_i)}{\sum_s w_i h^{-1}(\beta_0+\beta_1 x_i)}\right\}^2 \left\{w_j^2 \,\mathrm{var}(y_j)\right\} \end{aligned}$$

since the omitted sum is O(n^{-1}) times the order of $w_j^2 \,\mathrm{var}\left(y_j\right)$. The approximate unbiasedness of $v_{J,RAST}$ follows directly from this. For the SMEARING predictor, we take

$$\hat{T}_{r,SMEAR[\,j]} \approx \frac{\sum_{k\in r+\{j\}}\sum_{i\in s-\{j\}} w_i h^{-1}(\beta_0+\beta_1 x_k + R_i)}{\sum_{i\in s-\{j\}} w_i}.$$

The argument is then similar to that for the RAST predictor and is omitted.

3.5 Dealing with outliers

All the predictors developed thus far assume that the linear model (2) for log(Y) in terms of log(X) fits well, or at least that Y is well behaved with respect to some underlying true model. However, the reality is that the sample data typically include a substantial number of "special" values (e.g. zero) and outliers. The logarithmic transformation effectively controls the influence of raw-scale outliers, but is then susceptible to log-scale outliers (e.g. values near zero). These values can have a large effect on back-transformed predictions.

In order to control the influence of such outliers, we use robust methods of parameter estimation. In particular, the simulation study reported on in the next section was carried out using R (Ihaka and Gentleman, 1996), and we estimated β in (2) using the *rlm* function, which is part of the *MASS* robust statistics library (Venables and Ripley, 1994). We used a biweight influence function with tuning constant $c = 4.685$ and calculated the standard deviation s of the residuals using the MAD estimate output by *rlm*.

For the RAST and SMEARING predictors, we can go one step further, discounting outlying terms that enter into the RAST or SMEARING adjustment terms by using the outlier robust weights $\{w_i\}$, output by *rlm*. This leads to robust versions of these predictors such as:

$$\hat{T}_{RAST}^{rob} = \sum_s y_i + \sum_s w_i y_i \frac{\sum_r x_i^{b_1^{rob}}}{\sum_s w_i x_i^{b_1^{rob}}},$$

$$\hat{T}_{SMEAR}^{rob} = \sum_s y_i + \left(\sum_r x_i^{b_1^{rob}}\right)\left(\sum_s w_i \frac{y_i}{x_i^{b_1^{rob}}} \Big/ \sum_s w_i\right)$$

where b_1^{rob} is the robust estimate of β_1 output by *rlm*.

Thus those sample units that are effectively down-weighted as outliers in the log-scale in the course of robust estimation of the regression parameters are also down-weighted in the RAST

and SMEARING adjustments. These weights are not of course the weights described in section 3.3 above to achieve weighted balance. Estimators that incorporate histogram weights will be codified with an "H", those that incorporate robust weights, with an "R". The former (and twicing, in the case of SMEAR) is meant to deal with *global* deviations from the working model; the latter is intended to handle *local* deviations from the model. It is possible to incorporate both weights, for example the Twiced Robustified SMEARING estimator:

$$\text{SM/RH(2):} \qquad \begin{aligned} &\sum_s y_i + R_w \sum_r \tilde{y}_{Oi} + \sum_s w_i (y_i - R_w \tilde{y}_{Oi}) \\ &\tilde{y}_{Oi} = \exp\left(\tilde{\alpha}_O + \tilde{\beta}_O \log(x_i)\right) \\ &R_w = \frac{\sum_s w_i y_i / \tilde{y}_{Oi}}{\sum_s w_i} \end{aligned}$$

where $\hat{\alpha}_O$ and $\hat{\beta}_O$ are outlier robust estimates and w_i is a histogram weight based on the sample $\tilde{y}_{Oi}$ values.

4. Simulation Study

We carried out an extensive simulation study on four populations of businesses drawn from the UK's Monthly Wages and Salaries Survey (MWSS). These were the businesses making up two sectors of the MWSS sample, labelled A with population size $N = 768$, and B with $N = 1005$. For each sector, we considered two dependent variables Y, wages (*WAGES*) and number employed (*EMP*) at the time of the survey. For each, the dependent variable X was employment as measured on the UK Inter Departmental Business Register, the sampling frame for the MWSS, at the time of selection of the MWSS sample. This is denoted *Register EMP* below. The populations are represented graphically in Figures 1 and 2. For confidentiality reasons, the actual values have been re-scaled and the plots do not show a scale. It is readily apparent that the log-log

transformation yields something close to a homoscedastic linear fit, but with various anomalies peculiar to each population.

Each population was independently sampled 1000 times using (a) simple random sampling without replacement (SRSWOR), (b) size stratified random sampling (SizeSTRS), with size defined by $X = Register\ EMP$, (c) systematic probability proportional to size sampling (SYSPPS), with X as size variable, and finally (d) restricted "overbalanced" PPSSYS sampling that give samples that are nearly balanced with respect to inverse X weights (details in the Appendix.) In all cases sample sizes were $n = 50$. In the stratified case, we employed 4 strata, with strata boundaries cutting off approximately equal stratum X-totals. The "top" stratum was completely enumerated, with SRSWOR for the remaining strata. The sector B allocation was 15, 15, 15, 5, and the sector A allocation, 13,13,12,12.

For all designs, we considered 10 predictors of T. These were the expansion estimator (EE), the ratio estimator (RE), the naïve back-transform predictor (TA), the Karlberg lognormal model-based predictor (TK), the RAST predictor (RA), the SMEARING predictor (SM), and robust versions of the last four, signified by TA/R, TK/R, RA/R and SM/R respectively. In the case of stratified sampling, we used both stratified versions of these predictors as well as versions that ignored the strata (i.e. stratification was treated purely as a sampling device). In the latter case we calculated the across-stratum ratio estimator (RE/stratum weighted) since this is a more suitable comparator commensurate with "survey practice" for this case. Similarly, for the unequal probability design SYSPPS, the baseline comparator was the Horvitz-Thompson ratio estimator (RE/pi-weighted).

Additionally, for STRS we added unstratified versions of the histogram weighted predictors, namely RA/H and RA/RH, SM/H, SM/RH, SM/H(2), SM/RH(2), SM/H(2v), SM/RH(2v), where (2) refers to twicing, and "v" to the "plain vanilla" version of the choice of

φ - weights (see above.) These same additional 8 predictors were also calculated for PPSSYS and overbalanced samples.

For variance estimation, we used the Jackknife for all transformation-based predictors. The conventional design-based variance estimator was used for EE, while for RE and RE/Across we used the robust variance estimator suggested by Royall and Cumberland (1981). Variances were summed by stratum for the stratified versions of the estimators.

Our measures of simulation performance are given in Table 1. Tables 2 - 5 give the values of these performance measures for the various predictors under the several sampling schemes for the four populations. The best values of RMSE and Ratio Dominance (i.e. stochastic dominance relative to the "usual" ratio estimator for the design, see Table 1) are boldfaced in these tables. Here "best" means respectively "having RMSE within 10% of the smallest RMSE achieved for a particular design", and "Ratio Dominance greater than 50%", again for the particular design. Some observations on these results are:

1. Transformation-based predictors should be treated with care. The naïve back-transformed predictor (TA) was very biased on a number of occasions, while the "unbiased" lognormal model-based predictor proposed by Karlberg (TK) appeared to be very outlier sensitive (as indicated by improved performance when robustified.) This sensitivity was shared by the SMEARING predictor (SM). As anticipated, the RAST predictor (RA) controlled transformation bias, but was rather inefficient.
2. In contrast, outlier robust versions of transformation-based predictors generally worked well. In particular the variability of RAST predictor was reduced without an increase in bias. There were also substantial reductions in both bias and variability for the Karlberg and SMEARING predictors. Overall, the robust SMEARING predictor worked very well.

This effect of robustifying seems to hold as well for the "H"-weighted and twiced predictors.

3. The heteroskedasticity robust standard error estimates used for the various ratio estimators tended to be biased low, with corresponding undercoverage of associated confidence intervals. This also applied to the standard errors for the expansion estimators (with or without stratification). In contrast, jackknife standard errors seemed to be much better. Their associated coverage was often very good. The exception was for $Y = WAGES$ in the A population (Table 3), where the robust unstratified transformation-based predictors based on the SizeSTRS design appeared to have higher biases, leading to a reduction in coverage. This was particularly the case for R/non-H predictors. Use of the histogram weights and twicing (for the SMEARING predictor) improves coverage. In Table 3 we included the effect on the robustified (R) predictors of using the corresponding non-robustified variance estimates (these are the figures in parentheses in the 2-sigma coverage column). This improves coverage, sometimes appreciably, at the cost of a widening of intervals. This may be a good device in practice. It should be noted that there is no undercoverage (in fact we have the contrary) when the SYSPPS design is restricted to only choosing overbalanced samples.

4. Coverage of confidence intervals was uneven. Intervals were often too large, especially in the case of the unstratified predictors under STRS, and for the overbalanced samples. In the overbalanced samples particularly, it may be noted that the average size of intervals, as measured by %Relative Av SE, is about the same as for the PPS samples, despite the fact that by and large the point predictions have sharply lower RMSE.

5. Good RMSE performance often did not translate into stochastic dominance. However, outlier robust versions of RAST, SMEARING and Karlberg predictors generally

dominated the associated ratio estimator. This is brought out in Table 6, which summarizes the number of times each estimator was best with respect to RMSE and Ratio Dominance across the four populations.

We now focus on RMSE. Table 6 gives the "winners" for each design/population combination. It is clear that no single predictor dominates.

6. What is most notable is the impact of sample design. In all cases the RMSE for SRS-based predictors exceeds that of the STRS/stratified predictors, which in turn exceeds that of the STRS/unstratified predictors, which is about the same as the PPS-based predictors, which, finally, exceeds the RMSE under overbalanced sampling.

7. The table also suggests that the use of the histogram weights and twicing is more effective in the $Y = WAGES$ populations than in the $Y = EMP$ populations. This makes sense since the latter provides a cleaner linear fit on the log scale. (In fact, the hypothesis of a zero quadratic term in the log-log model is convincingly rejected for the A/Y=$WAGES$ population.)

Restricting attention to just the three better designs, we give all the "near winners" in Table 7, that is, those predictors whose RMSE was within 1.1 of the smallest for the given design/population. Again, robustified predictors dominate and twicing and use of histogram weights lead to better results for the messier *WAGES* populations, and we also note:

8. Robust SMEARING (SM/R) is best for the conventional sample designs, in the sense of appearing most often (3 times for STRS/Unstratified and 3 times for SYPPS) but TK/R and RA/RH are close behind (both 2 and 3 times, respectively.)

9. The conventional expansion and ratio estimators and the naïve predictor TA are not contenders, although RE/pi-weighted and TA/R each appear once.

Some further insight is given by Table 8, which lists all predictors based on the overbalanced sample design which have RMSE within 1.1 of the smallest RMSE achieved under the conventional designs (SYSPPS or STRS/Unstratified) for each population. The percentage of the minimal RMSE is also given in the table. We note:

10. There are serious gains available from narrowing the scope of the sampling design to selection of overbalanced samples. Except for the "too nice" B/*Y=EMP* population, many predictors are more efficient than the best that the conventional designs can offer. For A/*Y=WAGES* only SM/H and SM/H(2v) are out of the running.
11. Except for the B/*Y=EMP* population, the estimators SM/RH(2), RA/RH and TK/R are most consistently near best, in that order, in the overbalanced samples. Only the last is good for all 4 populations, though.

5. Summary

Using models for transformed data to handle non-linearity can bring gains in the prediction of finite population totals. However, outliers in the transformed scale can have a much more dramatic effect on transformation-based predictors than raw-scale outliers have on linear predictors. Our empirical results suggest that the robustified SMEARING, RAST, and Karlberg predictors are the preferred predictors for the log-log model (2), with further modification using twicing and histogram weights, where the log-log model possibly holds less strictly. In particular, it seems that SM/R, and SM/RH(2) in the messier WAGES populations, are the most consistently reliable, with TK/R and RA/RH not far behind. Efficiencies depend strongly on the sample design. A jackknife variance estimator does a reasonable job of estimating the precision of estimators, although further work on variance estimation is desirable to reduce instances of below nominal coverage as well as too long interval length. The RAST and SMEARING estimators can

also be applied to transform models other than the logarithmic transform, and the theoretical analysis reported in this paper leads us to anticipate good results. Empirical testing of their behaviour in this case, however, remains for further investigation.

References

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, New York: Chapman and Hall.

Chen, G. and Chen, J. (1996). A transformation method for fintie population sampling calibrated with empirical likelihood, *Survey Methodology*, 22, 139-146.

Deming, W. E. (1984). *Statistical Adjustment of Data*, New York: Dover [original publication 1943]

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association*, **78**, 605-610

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299 – 314.

Karlberg, F. (2000a). Population total prediction under a lognormal superpopulation model, *Metron*, 53-80.

Karlberg, F. (2000b). Survey estimation for highly skewed population in the presence of zeroes, *Journal of Official Statistics*, **16**, 229-241.

Royall, R. M. (1982). Finite populations, Sampling from. Entry in the *Encyclopedia of Statistical Sciences* (eds. Johnson and Kotz), New York: Wiley.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance, *Journal of the American Statistical Association* **76**, 66 - 88.

Royall, R. M. and Cumberland, W. G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association* **80**, 355 - 359.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*, New York: John Wiley

Venables, W.N. and Ripley, B.D. (1994). *Modern Applied Statistics with S-PLUS*. New York: Springer.

Appendix – Weighted Balance

On the original scale of X and Y it is clear that variances of Y given X increase sharply with X. If we suppose that they increase proportionally to X^2, then, in standard non-transformed modelling, weighted balance, with weights inversely proportional to X, gives protection against misspecification of the model, and also lowest variances for estimates of totals (Valliant, *et al.* 2000). This particular form of balance has been referred to in the literature as *overbalance*, and is defined by

$$\frac{\sum_s x_i^{-1} x_i^K}{n} = \frac{\sum_P x_i^K}{\sum_P x_i},$$

or, equivalently,

$$\frac{\sum_P x_i^K}{\sum_P x_i}\left[\frac{\sum_P x_i}{\sum_P x_i^K}\left(n^{-1}\sum_s x_i^{K-1}\right)-1\right]=0,$$

for K = 0, 2, 3,… etc. SYSPPS sampling with X size variable *aims* at overbalance, but does not actually achieve it for most samples. Figure 3 indicates the extent to which samples selected via this design deviated from overbalance for the A and B populations. In these plots, (deviation from) K-order overbalance is measured by the term in square brackets above (that is, the relative difference of the sample moment and the corresponding population ratio), with a value of zero indicating that a sample is exactly overbalanced at that order. The 100 samples (out of the 1000 actually drawn) with smallest values of $\sqrt{0\text{th order overbalance}^2 + 2\text{nd order overbalance}^2}$ were taken as defining an overbalanced sampling strategy. Based on the results in Table 8, such a sampling strategy seems to be a promising way to proceed.

Table 1. Performance measures used in the simulation study.

Measure	Description
% Relative Bias	Average simulation error, expressed as a percentage of the target population total.
% Relative RMSE	Square root of average squared simulation error, expressed as a percentage of the target population total.
% Relative Av SE	Average simulation estimated standard error, expressed as a percentage of the target population total.
2-sigma Coverage	Proportion of simulation "2-sigma" confidence intervals that include the target population total.
Ratio Dominance	Proportion of times a predictor stochastically dominates (i.e. has smaller absolute simulation error than) the corresponding ratio estimator. We compared with the simple ratio estimator (RE) for the SRSWOR design, the stratified ratio estimator (RE) for stratified predictors based on the SizeSTRS design, the across-stratum ratio estimator (RE/stratum weighted) for the unstratified predictors based on the SizeSTRS design and the inverse pi-weighted ratio estimator (RE/pi-weighted) for the PPSSYS design.

Table 2. Simulation results for Sector A, *Y = EMP*.

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SRSWOR design					
EE	3.82	65.15	48.54	0.72	0.13
RE	0.72	**12.38**	9.21	0.86	.
TA	0.36	17.44	17.18	0.96	0.33
TA/R	5.58	**12.30**	11.95	0.97	0.48
TK	13.13	22.61	18.27	0.95	0.28
TK/R	8.95	14.54	12.52	0.95	0.41
RA	1.34	15.53	15.98	0.94	0.37
RA/R	3.96	**12.90**	13.21	0.93	**0.52**
RA/H	11.29	20.46	17.31	0.94	0.31
RA/RH	10.56	15.74	13.14	0.94	0.38
SM	14.04	26.30	21.00	0.95	0.29
SM/R	9.12	14.66	12.51	0.95	0.40
SM/H	15.25	27.75	21.67	0.94	0.27
SM/RH	9.62	14.95	12.58	0.94	0.40
SizeSTRS design / Stratified predictors					
EE	0.15	9.23	8.66	0.89	0.25
RE	0.11	4.85	3.76	0.89	.
TA	-2.52	5.58	5.29	0.94	0.38
TA/R	-1.35	**4.19**	5.06	0.94	0.47
TK	0.73	5.19	5.47	0.98	**0.52**
TK/R	-0.13	**4.24**	5.26	0.96	**0.55**
RA	0.24	5.21	5.49	0.97	0.46
RA/R	0.06	4.76	5.58	0.95	**0.53**
SM	0.82	6.84	6.01	0.97	0.48
SM/R	-0.12	**4.27**	5.28	0.96	**0.56**

Table 2 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SizeSTRS design / Unstratified predictors					
RE/stratum weighted	0.03	4.79	4.17	0.93	
RE/unweighted	-3.64	3.82	2.85	0.97	0.39
TA	-4.54	5.17	4.99	0.98	0.27
TA/R	-2.15	2.79	3.82	1.00	**0.59**
TK	-0.53	2.59	4.47	1.00	**0.68**
TK/R	-1.29	**2.25**	3.81	1.00	**0.73**
RA	-1.89	3.47	4.93	1.00	**0.59**
RA/R	-1.80	2.62	4.10	1.00	**0.64**
RA/H	-0.07	3.40	4.81	1.00	**0.69**
RA/RH	-0.59	**2.48**	4.17	1.00	**0.70**
SM	-1.42	3.53	4.65	1.00	**0.60**
SM/R	-1.29	**2.25**	3.82	1.00	**0.73**
SM/H	0.19	6.72	5.62	1.00	**0.63**
SM/RH	-0.93	**2.18**	3.94	1.00	**0.75**
SM/H(2v)	0.76	4.43	5.57	1.00	**0.54**
SM/RH(2v)	-1.36	3.22	4.95	1.00	**0.59**
SM/H(2)	-0.03	4.46	5.51	1.00	**0.56**
SM/RH(2)	-0.48	2.65	4.27	1.00	**0.67**
SYSPPS design					
RE/pi-weighted	-0.22	**3.07**	3.15	0.92	
TA	-4.39	5.45	5.44	0.97	0.21
TA/R	-1.70	3.25	4.11	1.00	0.41
TK	0.03	**2.87**	4.58	1.00	**0.59**
TK/R	-0.79	**3.09**	4.05	1.00	0.46
RA	-1.14	4.03	5.49	1.00	0.47
RA/R	-0.87	**3.00**	4.42	1.00	0.51
RA/H	0.48	3.52	5.18	1.00	0.48
RA/RH	-0.06	**3.05**	4.46	1.00	0.49
SM	-1.39	3.32	4.69	1.00	0.47
SM/R	-0.81	**3.09**	4.07	1.00	0.45
SM/H	0.56	3.62	5.22	1.00	**0.56**
SM/RH	-0.10	**2.92**	4.53	0.99	**0.51**
SM/H(2v)	2.94	5.87	6.96	0.98	0.35
SM/RH(2v)	-2.82	4.58	6.38	1.00	0.28
SM/H(2)	0.39	4.26	5.97	1.00	0.29
SM/RH(2)	-0.04	3.31	4.58	1.00	0.44

Table 2 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
		100 best "overbalanced PPSSYS samples			
RE/pi-weighted	0.98	1.29	3.57	1.00	.
TA	-2.44	2.69	4.53	1.00	20
TA/R	-2.23	2.53	3.52	1.00	0.25
TK	0.15	**1.09**	4.24	1.00	**0.65**
TK/R	-1.56	1.97	3.49	1.00	0.35
RA	-0.03	1.66	4.42	1.00	**0.51**
RA/R	-1.69	2.34	3.68	1.00	0.33
RA/H	2.86	3.02	6.43	1.00	0.03
RA/RH	-1.37	1.54	3.53	1.00	0.37
SM	-0.03	**1.05**	4.39	1.00	**0.74**
SM/R	-1.57	1.98	3.50	1.00	0.35
SM/H	1.73	2.07	6.07	1.00	0.18
SM/RH	-1.42	1.78	3.53	1.00	0.34
SM/H(2v)	6.45	6.78	10.14	1.00	0.00
SM/RH(2v)	-2.33	2.41	4.46	1.00	0.17
SM/H(2)	4.27	4.40	7.17	1.00	0.00
SM/RH(2)	-1.30	1.43	3.63	1.00	0.41

Table 3. Simulation results for Sector A, *Y = WAGES*. Values in brackets for 2-sigma coverage are obtained by combining robustified predictors with variance estimates associated with corresponding non-robustified predictors.

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SRSWOR design					
EE	3.09	55.18	42.57	0.72	0.29
RE	6.87	31.14	22.57	0.86	
TA	1.96	24.89	23.99	0.96	**0.55**
TA/R	-0.17	**20.09**	20.71	0.93	**0.60**
TK	28.76	45.46	32.01	0.98	0.40
TK/R	12.20	26.98	25.30	0.98	**0.59**
RA	5.26	33.55	31.64	0.93	**0.55**
RA/R	-0.20	25.38	25.89	0.90	**0.64**
SM	50.48	104.23	54.15	0.98	0.36
SM/R	12.24	27.03	24.73	0.98	**0.58**
SizeSTRS design / Stratified predictors					
EE	0.51	14.05	12.74	0.87	0.36
RE	1.03	12.23	9.46	0.89	
TA	-6.32	11.38	11.03	0.85	0.32
TA/R	-6.75	**11.08**	11.08	0.82	0.33
TK	2.47	12.71	12.87	0.96	**0.50**
TK/R	-2.00	**10.11**	12.47	0.93	0.47
RA	1.68	12.25	13.98	0.94	0.47
RA/R	-2.52	**10.54**	12.84	0.91	0.42
SM	7.75	32.90	19.21	0.96	0.48
SM/R	-1.74	**10.37**	12.98	0.93	0.47

Table 3 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage		Ratio Dominance
SizeSTRS design / Unstratified predictors						
RE/stratum weighted	0.31	10.80	9.83	0.91		.
RE/unweighted	-18.60	18.97	4.45	0.01		0.08
TA	-12.75	13.65	7.37	0.63		0.18
TA/R	-12.15	12.83	6.47	0.58	(0.68)	0.20
TK	-3.74	**8.17**	9.18	0.88		**0.58**
TK/R	-7.97	9.05	6.49	0.79	(0.85)	0.40
RA	-7.94	10.15	8.28	0.81		0.41
RA/R	-9.38	10.41	6.66	0.75	(0.81)	0.34
RA/H	-0.84	9.37	10.65	0.95		**0.58**
RA/RH	-5.26	**7.81**	8.04	0.91	(0.95)	**0.52**
SM	-1.66	17.19	11.88	0.87		**0.54**
SM/R	-8.47	9.54	6.54	0.77	(0.85)	0.36
SM/H	4.54	51.58	17.52	0.91		0.46
SM/RH	-7.02	**8.41**	7.01	0.85	(0.91)	0.44
SM/H(2v)	-0.28	10.13	11.43	0.96		**0.57**
SM/RH(2v)	-5.32	**7.87**	8.12	0.91	(0.95)	**0.52**
SM/H(2)	-0.15	12.82	13.05	0.96		**0.57**
SM/RH(2)	-5.25	**7.81**	8.05	0.92	(0.96)	**0.53**
SYSPPS design						
RE/pi-weighted	-0.38	13.67	7.91	0.86		.
TA	-10.90	11.97	7.64	0.73		0.16
TA/R	-10.56	11.30	6.64	0.71	(0.79)	0.13
TK	-2.75	**7.26**	8.67	0.86		0.48
TK/R	-6.75	8.05	6.85	0.84	(0.86)	0.27
RA	-4.11	8.07	8.74	0.90		0.44
RA/R	-6.32	8.10	7.09	0.85	(0.88)	0.35
RA/H	-2.59	8.20	9.39	0.89		0.38
RA/RH	-5.08	**7.70**	7.98	0.89	(0.90)	0.40
SM	-2.56	**7.86**	9.30	0.86		0.39
SM/R	-6.69	8.06	6.99	0.84	(0.86)	0.27
SM/H	-3.44	13.35	9.68	0.88		0.36
SM/RH	-5.61	**7.95**	8.00	0.88	(0.89)	0.37
SM/H(2v)	-2.61	10.21	12.31	0.90		0.33
SM/RH(2v)	-4.68	8.79	9.92	0.90	(0.92)	0.36
SM/H(2)	-1.65	10.09	10.60	0.91		0.38
SM/RH(2)	-4.68	**7.81**	8.37	0.90	(0.91)	0.42

Table 3 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
		100 best "overbalanced PPSSYS samples			
RE/pi-weighted	2.25	5.28	7.74	1.00	.
TA	-7.14	7.51	6.90	0.98	0.25
TA/R	-7.78	7.96	5.88	0.94	0.20
TK	-0.43	**3.17**	7.59	1.00	**0.55**
TK/R	-4.70	5.09	5.93	0.98	0.33
RA	-1.62	3.82	7.83	1.00	0.47
RA/R	-5.23	5.78	6.38	0.98	0.30
RA/H	1.54	5.56	9.51	1.00	0.40
RA/RH	-2.76	3.71	7.17	1.00	0.46
SM	-0.40	3.52	7.95	1.00	**0.53**
SM/R	-4.74	5.13	5.96	0.98	0.33
SM/H	0.42	8.72	11.18	1.00	0.26
SM/RH	-4.55	5.32	7.33	0.98	0.32
SM/H(2v)	3.15	8.27	12.77	1.00	0.10
SM/RH(2v)	-2.34	3.85	8.86	1.00	**0.56**
SM/H(2)	2.57	4.24	9.31	1.00	0.40
SM/RH(2)	-1.66	**3.01**	7.32	1.00	**0.60**

Table 4. Simulation results for Sector B, *Y = EMP*.

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SRSWOR design					
EE	-0.07	38.01	31.73	0.81	0.23
RE	2.28	14.93	11.78	0.90	
TA	-11.33	17.63	13.46	0.85	0.43
TA/R	-1.79	**11.33**	11.85	0.93	**0.59**
TK	2.74	15.14	14.84	0.95	0.47
TK/R	3.55	**11.97**	12.13	0.95	**0.59**
RA	1.30	15.38	15.37	0.93	0.45
RA/R	3.49	13.88	14.14	0.93	**0.56**
SM	2.60	15.85	15.56	0.95	0.45
SM/R	3.72	**12.12**	12.12	0.95	**0.58**
SizeSTRS design / Stratified predictors					
EE	-0.31	11.86	11.18	0.90	0.29
RE	-0.16	7.41	6.43	0.93	
TA	-7.92	10.55	7.89	0.84	0.26
TA/R	-3.46	7.13	8.27	0.95	0.46
TK	1.86	7.86	8.59	0.98	0.44
TK/R	0.29	**6.33**	8.44	0.98	**0.56**
RA	-0.44	7.76	8.35	0.97	0.45
RA/R	0.24	**6.62**	8.54	0.98	**0.58**
SM	0.87	8.16	8.81	0.98	0.41
SM/R	0.24	**6.43**	8.41	0.98	**0.57**

Table 4 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SizeSTRS design / Unstratified predictors					
RE/stratum weighted	-0.26	7.30	6.57	0.94	
RE/unweighted	-9.34	9.99	7.38	0.89	0.17
TA	-9.34	10.66	7.10	0.80	0.20
TA/R	-3.12	5.28	6.83	0.99	**0.56**
TK	2.42	5.97	7.21	1.00	**0.56**
TK/R	0.88	**4.25**	6.58	1.00	**0.64**
RA	-3.69	6.64	8.62	0.99	0.46
RA/R	-2.50	4.83	8.23	1.00	**0.60**
RA/H	0.01	7.54	8.39	0.97	0.42
RA/RH	0.90	5.48	7.17	0.99	**0.58**
SM	-0.32	5.63	7.17	0.99	**0.57**
SM/R	0.62	**4.21**	6.50	1.00	**0.66**
SM/H	0.17	8.89	9.71	0.99	0.44
SM/RH	-0.08	**4.54**	7.50	1.00	**0.66**
SM/H(2v)	-0.06	6.89	7.89	0.97	0.47
SM/RH(2v)	3.46	6.33	7.84	0.99	**0.51**
SM/H(2)	0.05	7.34	8.34	0.97	0.45
SM/RH(2)	0.56	4.95	7.05	0.99	**0.62**
SYSPPS design					
RE/pi-weighted	-0.03	6.06	5.98	0.95	
TA	-8.13	9.86	7.01	0.83	0.23
TA/R	-2.08	**4.23**	6.34	1.00	**0.62**
TK	2.31	6.30	7.07	1.00	**0.51**
TK/R	1.27	**4.00**	6.11	0.99	**0.61**
RA	-2.92	7.40	8.60	0.99	0.37
RA/R	-0.97	**4.36**	7.51	1.00	**0.59**
RA/H	-0.20	7.35	7.76	0.98	0.25
RA/RH	0.83	4.90	6.75	1.00	**0.51**
SM	-0.12	5.82	6.98	1.00	**0.52**
SM/R	1.14	**3.98**	6.19	0.99	**0.62**
SM/H	-0.33	8.57	9.11	0.99	0.46
SM/RH	0.29	4.59	7.42	1.00	**0.55**
SM/H(2v)	-0.16	8.14	8.50	0.98	0.22
SM/RH(2v)	0.08	5.21	7.46	1.00	0.48
SM/H(2)	-0.10	8.15	8.17	0.98	0.21
SM/RH(2)	0.91	5.00	6.77	1.00	**0.50**

Table 4 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
		100 best "overbalanced PPSSYS samples			
RE/pi-weighted	-2.10	6.75	5.96	0.96	.
TA	-10.46	11.66	7.07	0.65	0.17
TA/R	-3.98	4.94	6.34	1.00	**0.73**
TK	-0.45	5.74	7.10	1.00	**0.64**
TK/R	0.50	**3.67**	6.24	1.00	**0.64**
RA	-5.82	9.49	8.80	1.00	0.36
RA/R	-2.09	4.07	7.30	1.00	0.81
RA/H	-3.50	7.28	7.39	0.96	0.30
RA/RH	-1.72	5.37	6.87	1.00	**0.59**
SM	-2.36	6.67	7.18	1.00	**0.58**
SM/R	0.24	**3.57**	6.27	1.00	**0.64**
SM/H	-1.26	8.55	9.07	1.00	0.46
SM/RH	-2.01	4.73	7.13	1.00	**0.75**
SM/H(2v)	-3.83	7.72	8.01	0.91	0.29
SM/RH(2v)	-3.56	7.14	7.71	1.00	0.34
SM/H(2)	-4.20	7.80	7.76	0.91	0.29
SM/RH(2)	-1.69	5.52	6.93	1.00	**0.56**

Table 5. Simulation results for Sector B, *Y = WAGES*.

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SRSWOR design					
EE	0.40	38.73	33.32	0.79	0.40
RE	5.63	**30.91**	23.48	0.87	
TA	-23.67	**30.84**	19.96	0.66	0.41
TA/R	-16.40	**28.27**	23.42	0.77	0.47
TK	12.42	**30.62**	28.48	0.96	**0.54**
TK/R	11.56	**30.30**	30.66	0.95	**0.55**
RA	5.23	31.79	29.92	0.89	0.47
RA/R	5.34	32.06	30.91	0.89	**0.50**
SM	9.94	**29.93**	28.98	0.95	**0.55**
SM/R	7.66	**28.53**	27.90	0.94	**0.58**
SizeSTRS design / Stratified predictors					
EE	0.14	16.26	16.35	0.91	0.38
RE	0.27	**13.05**	12.62	0.93	
TA	-21.77	24.37	12.90	0.57	0.18
TA/R	-18.16	21.68	15.68	0.71	0.22
TK	2.15	15.02	16.97	0.96	0.45
TK/R	2.23	15.79	21.93	0.97	0.45
RA	-0.53	**13.67**	16.13	0.95	0.46
RA/R	-1.78	**13.60**	17.61	0.95	0.45
SM	0.97	15.39	17.51	0.96	0.41
SM/R	-1.18	**14.16**	18.39	0.95	0.44

Table 5 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
SizeSTRS design / Unstratified predictors					
RE/stratum weighted	0.14	12.96	13.01	0.94	.
RE/unweighted	-25.63	26.78	13.02	0.51	0.09
TA	-27.71	28.64	10.11	0.23	0.08
TA/R	-24.87	26.08	11.00	0.39	0.11
TK	0.82	**11.04**	14.48	0.98	**0.58**
TK/R	0.69	13.55	20.69	0.96	0.48
RA	-13.97	16.85	14.60	0.89	0.31
RA/R	-13.69	16.46	14.13	0.88	0.32
RA/H	-2.12	14.05	15.76	0.93	0.45
RA/RH	-4.04	12.90	14.80	0.94	**0.50**
SM	-3.98	**10.95**	13.41	0.96	**0.56**
SM/R	-5.64	**11.29**	13.21	0.95	**0.53**
SM/H	-8.27	16.50	16.89	0.89	0.34
SM/RH	-10.41	14.94	15.23	0.90	0.35
SM/H(2v)	-2.53	12.84	14.80	0.94	**0.50**
SM/RH(2v)	-3.24	**11.83**	14.44	0.96	**0.53**
SM/H(2)	-3.56	13.64	15.47	0.93	0.45
SM/RH(2)	-6.00	12.83	14.44	0.92	0.46
SYSPPS design					
RE/pi-weighted	0.15	11.32	11.50	0.97	.
TA	-22.47	24.25	10.66	0.43	0.12
TA/R	-19.13	21.31	11.45	0.64	0.16
TK	6.12	12.47	14.76	0.99	0.46
TK/R	2.91	**10.33**	14.35	0.99	**0.56**
RA	-4.97	12.38	15.84	0.99	0.44
RA/R	-4.41	11.14	15.46	0.99	0.48
RA/H	-0.72	**10.19**	14.73	0.97	**0.52**
RA/RH	-1.40	**9.53**	14.24	0.98	**0.59**
SM	1.26	10.78	13.74	0.99	**0.56**
SM/R	0.41	**9.97**	13.76	0.99	**0.56**
SM/H	-4.48	15.53	17.72	0.97	0.29
SM/RH	-5.31	12.45	16.62	0.97	0.35
SM/H(2v)	-1.23	10.79	16.03	0.96	0.49
SM/RH(2v)	-2.42	**9.90**	14.81	0.97	**0.54**
SM/H(2)	0.47	**10.17**	15.03	0.97	**0.56**
SM/RH(2)	-0.60	**9.47**	14.32	0.98	**0.59**

Table 5 (continued)

Predictor	% Relative Bias	% Relative RMSE	% Relative Av SE	2-sigma Coverage	Ratio Dominance
		100 best "overbalanced PPSSYS samples			
RE/pi-weighted	1.52	11.66	11.93	1.00	.
TA	-23.36	24.02	11.10	0.42	0.11
TA/R	-16.98	17.86	11.66	0.85	0.16
TK	10.62	14.03	16.11	1.00	0.28
TK/R	4.01	7.83	14.39	1.00	**0.63**
RA	-10.59	14.99	14.74	1.00	0.18
RA/R	-7.72	11.91	14.68	1.00	0.21
RA/H	-4.00	8.04	13.16	1.00	0.43
RA/RH	-2.39	**5.51**	13.86	1.00	**0.63**
SM	-0.02	9.57	13.69	1.00	0.32
SM/R	0.73	6.83	13.94	1.00	**0.52**
SM/H	-0.59	23.39	20.95	1.00	0.13
SM/RH	-4.05	11.14	20.41	1.00	0.31
SM/H(2v)	-5.06	8.65	14.09	1.00	0.36
SM/RH(2v)	-4.61	6.74	14.69	1.00	0.40
SM/H(2)	-4.72	6.60	13.39	1.00	0.48
SM/RH(2)	-1.81	**5.15**	13.60	1.00	**0.67**

Table 6. Predictors achieving minimum % Relative RMSE for each Design/Population

Population	A/*Y=EMP*		A/*Y=WAGES*		B/*Y=EMP*		B/*Y=WAGES*	
Design	**Predictor**	**minimum % Rel RMSE**	**Predictor**	**minimum % Rel RMSE**	**Predictor**	**minimum % Rel RMSE**	**Predictor**	**minimum % Rel RMSE**
SRSWOR	TA/R	12.30	TA/R	20.09	TA/R	11.33	TA/R	28.27
SizeSTRS Stratified	TA/R	4.19	TK/R	10.11	TK/R	6.33	RE	13.05
SizeSTRS Unstratified	SM/RH	2.18	RA/RH SM/RH(2)	7.81	SM/R	4.21	SM	10.95
SYSPPS	TK	2.87	TK	7.26	SM/R	3.98	SM/RH(2)	9.47
Over-balanced	SM	1.05	SM/RH(2)	3.01	SM/R	3.57	SM/RH(2)	5.15

Table 7. Predictors achieving Near Best values (< 1.1 minimum value) of % Rel RMSE for 3 Designs

Population	A/*Y=EMP*		A/*Y=WAGES*		B/*Y=EMP*		B/*Y=WAGES*	
Design	**Predictor**	**% Rel RMSE**	**Predictor**	**% Rel RMSE**	**Predictor**	**% Rel RMSE**	**Predictor**	**% Rel RMSE**
STRS	SM/RH	2.18	RA/RH	7.81	SM/R	4.21	SM	10.95
Unstratified	TK/R	2.25	SM/RH(2)	7.81	TK/R	4.25	TK	11.04
	SM/R	2.25	SM/RH(2v)	7.87	SM/RH	4.54	SM/R	11.29
	RA/RH	2.48	SM/RH	8.41			SM/RH(2v)	11.83
SYSPPS	TK	2.87	TK	7.26	SM/R	3.98	SM/RH(2)	9.47
	SM/RH	2.92	RA/RH	7.70	TK/R	4.00	RA/RH	9.53
	RA/R	3.00	SM/RH(2)	7.81	TA/R	4.23	SM/RH(2v)	9.90
	RA/RH	3.05	SM	7.86	RA/R	4.36	SM/R	9.97
	RE/pi-weighted	3.07	SM/RH	7.95			SM/H	10.17
							RA/H	10.19
	TK/R	3.09					TK/R	10.33
	SM/R	3.09						
Overbalanced	SM	1.05	SM/RH(2)	3.01	SM/R	3.57	SM/RH(2)	5.15
	TK	1.09	TK	3.17	TK/R	3.67	RA/RH	5.51

Table 8. Overbalanced Samples. Predictors with % Rel RMSE < 1.1 of minimum % Rel RMSE of predictors based on both SYSPPS and STRS/unstratified designs (percentages of minimum % Rel RMSE shown). Entries in square brackets lie outside the tolerance range, but represent the next best performers.

A/*Y=EMP*		A/*Y=WAGES*		B/*Y=EMP*		B/*Y=WAGES*	
Predictor	**% of min RMSE**	**Predictor**	**% of min RMSE**	**Predictor**	**% of min RMSE**	**Predictor**	**% of min RMSE**
SM	48	SM/RH(2)	41	SM/R	90	SM/RH(2)	54
TK	50	TK	44	TK/R	92	RA/RH	58
RE/pi-weighted	59	SM	48	RA/R	102	SM/H(2)	70
SM/RH(2)	66	RA/RH	51	[RA/RH]	[135]	SM/RH(2)	71
RA/RH	71	RA	53	[SM/RH(2)]	[139]	SM/R	72
RA	76	SM/RH(2v)	53			TK/R	83
SM/RH	82	SM/H(2)	58			RA/H	85
TK/R	90	TK/R	70			SM/H(2v)	91
SM/R	91	SM/R	71			SM	101
SM/H	95	RE/pi-weighted	73				
RA/R	107	SM/RH	73				
		RA/H	77				
		RA/R	80				
		TA	103				
		TA/R	110				

Figure 1: The sector A populations from the Monthly Wages and Salaries Survey

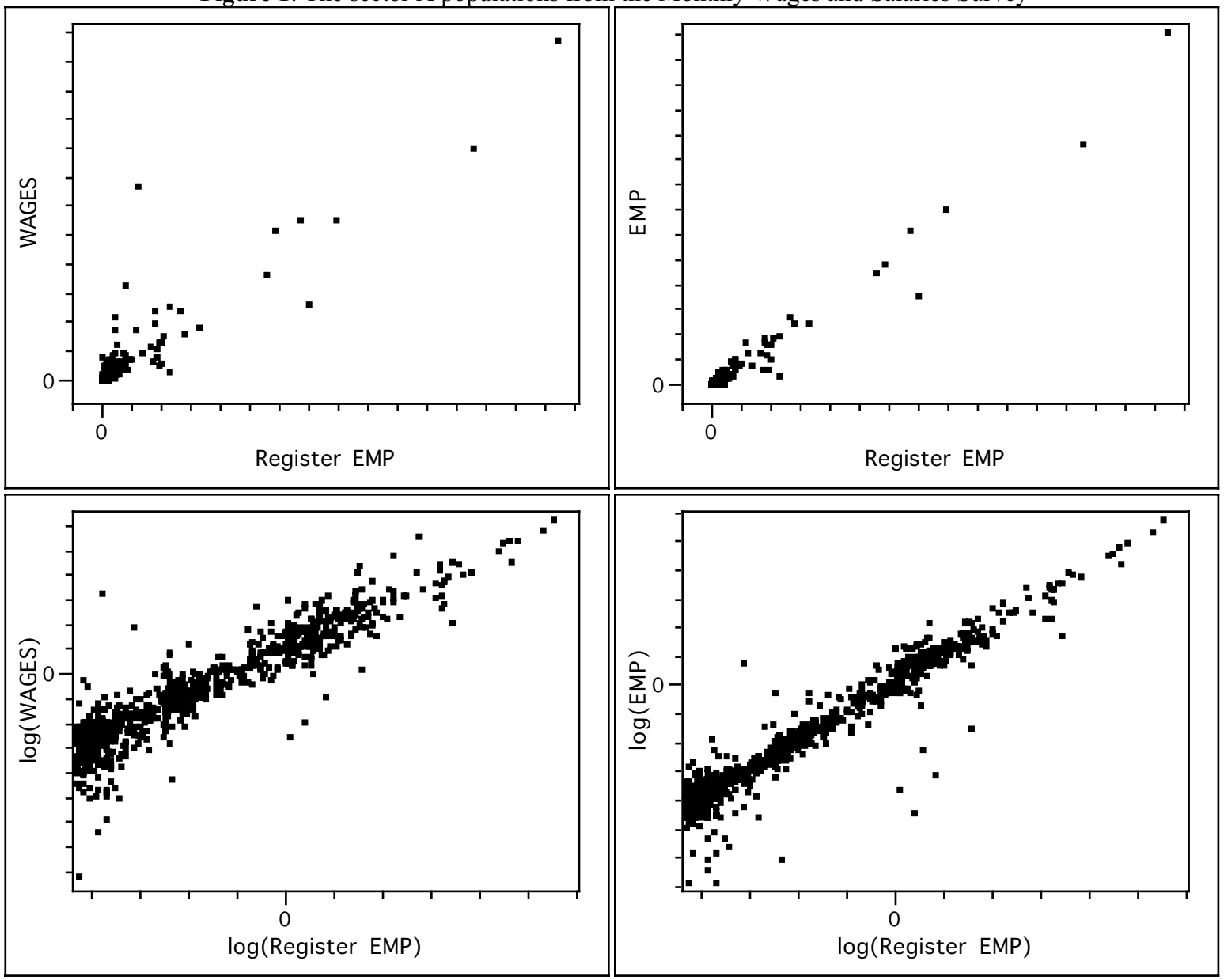

Figure 2: The sector B populations from the Monthly Wages and Salaries Survey

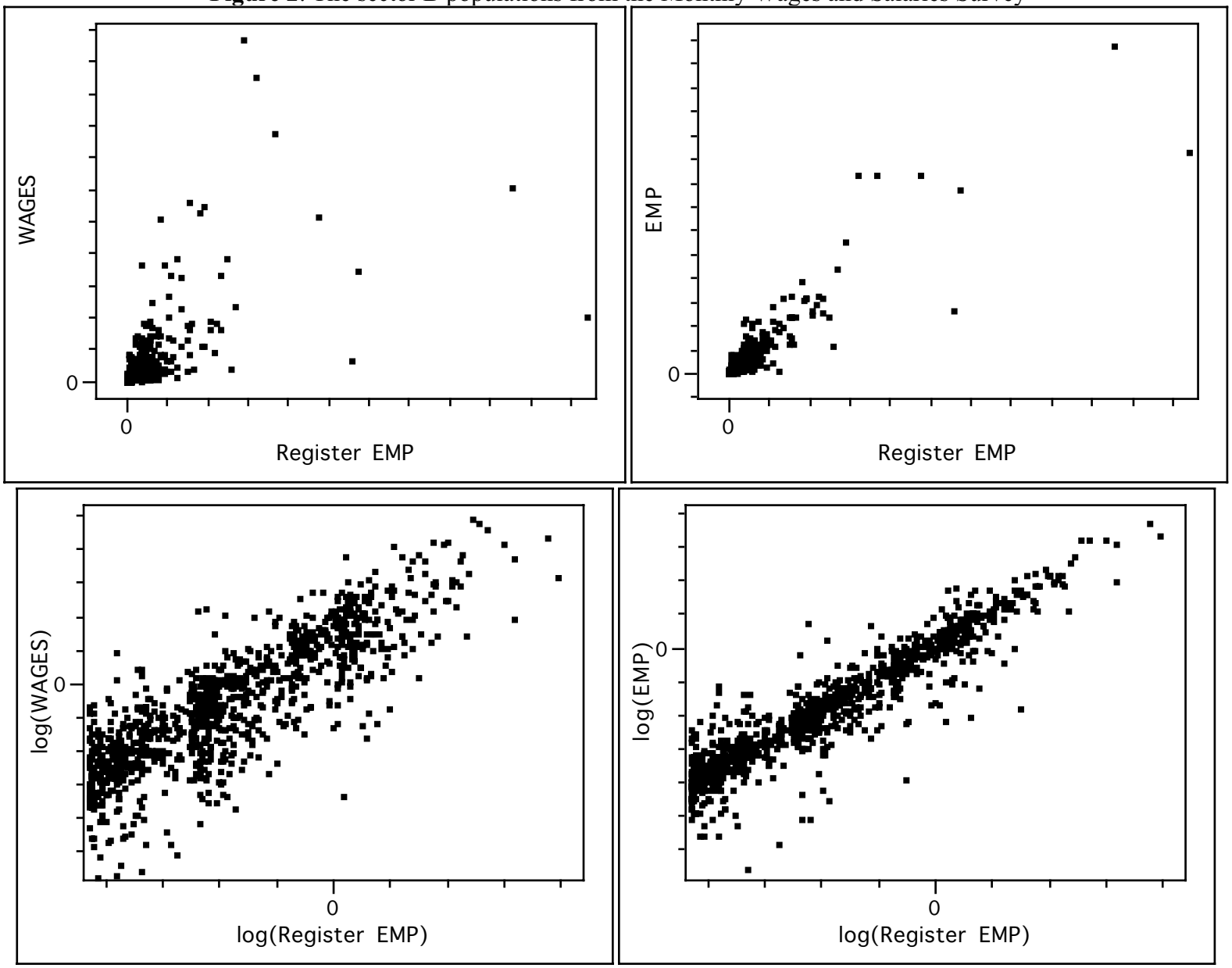

Figure 3: Zero order overbalance (x axis) plotted against second order overbalance (y axis) for the 1000 samples selected via PPSSYS. Overbalance calculations exclude certainty units. Note that in the A population, no sample achieved both zero and second order overbalance. Points indicated in red correspond to the 100 best overbalanced PPSSYS samples.