Communications of the Korean Statistical Society Vol. 15, No. 5, 2008, pp. 783–791

# Model-Based Prediction of the Population Proportion and Distribution Function Using a Logistic Regression[†]

Mingue Park[1)]

**Abstract**

Estimation procedure of the finite population proportion and distribution function is considered. Based on a logistic regression model, an approximately model-optimal estimator is defined and conditions for the estimator to be design-consistent are given. Simulation study shows that the model-optimal design-consistent estimator defined under a logistic regression model performs well in estimating the finite population distribution function.

*Keywords:* MLE; design consistency; distribution function estimation; model-based approach.

## 1. Introduction

One of the common finite population parameters of interest in practice is the population proportion. In designing a survey, many questions are designed to use a nominal scale to estimate the proportion of the subpopulation that has a certain characteristic of interest. Because the finite population proportion can be expressed as a population mean of the dichotomous variable that takes a value 0 or 1, the study on the population proportion estimation has been done as a special case of the population mean estimation. One of a few studies focusing on finite population proportion estimation is from Valliant *et al.* (2000). They considered the logistic regression model to obtain a maximum likelihood type predictor and investigated model properties of the estimator. But the design properties of the estimator were not investigated.

Estimation of the population mean of a dichotomous variable has been intensively studied to obtain an efficient estimator of the population distribution function. The distribution function at $|t| < \infty$, $F_N(t) = N^{-1} \sum_{i \in U} I[z_i \leq t]$ is a mean of dichotomous variables where the indicator function $I[z_i \leq t]$ is one if $z_i \leq t$ and zero elsewhere and $U$ is a set of indices in the population. Chambers and Dunstan (1986) suggested a model-based predictor of the population distribution function under a linear regression model for a simple random sample. Rao *et al.* (1990) suggested a design-consistent model-assisted estimator of the population distribution function under a linear regression

† This research is supported by a Korea University Grant

1) Assistant Professor, Department of Statistics, Korea University, Seoul 136-701, South Korea. Corresponding author: mpark2@korea.ac.kr

model. Wu and Sitter (2001) considered a nonlinear model as well as a linear model to describe the relationship between the variable of interest and auxiliary variables. They derived a regression type estimator using the distribution of the predicted values of unobserved observations. Recently Harms and Duchesne (2006) suggested a regression type estimator where the population distribution function of the auxiliary variables is used as an auxiliary information. Many theoretical and empirical studies show that there is no single sharp winner among the estimators developed in estimating the population distribution function. See, Chambers *et al.* (1992).

In this paper, we consider a model-optimal estimation procedure in estimating the population proportion under a logistic regression model. We give the condition under which the model-optimal estimator is also design consistent. An approximation of the estimator is derived to obtain a reasonable variance estimator. Through a simulation study, we compare the small sample performance of the estimators in estimating the population proportion and the population distribution function.

## 2. Model-Optimal Prediction

Assume that the finite population of binary variables, $y_i,\ i = 1, \ldots, N$, is a random sample from a superpopulation model

$$y_i \sim \text{Bernoulli}(p_i) \tag{2.1}$$

and

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i\boldsymbol{\beta}, \tag{2.2}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. Note that $p_i = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})]^{-1}$. The superpopulation model described by (2.1) and (2.2) is known as a logistic regression model.

Under the model described in (2.1) and (2.2), consider the estimation of the population mean of $y_i$, $\bar{y}_N = N^{-1}\sum_{i\in U} y_i$. If $(y_i, \mathbf{x}_i)$, $i \in U$ are available, the maximum likelihood estimator(MLE) of $\boldsymbol{\beta}$ is obtained by maximizing the log-likelihood function

$$l(\boldsymbol{\beta}) = \sum_{i\in U} \ln\left(\frac{p_i}{1-p_i}\right) y_i + \ln(1-p_i). \tag{2.3}$$

Let $\boldsymbol{\beta}_N$ denote the MLE obtained by solving the likelihood equation

$$\sum_{i\in U}(y_i - p_{iN})\mathbf{x}_i = \sum_{i\in U} g\left(\mathbf{z}_i : \boldsymbol{\beta}_N\right) = 0, \tag{2.4}$$

where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and $p_{iN} = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta}_N)]^{-1}$. Note that $\boldsymbol{\beta}_N$ is a model-consistent and asymptotically model-efficient estimator of the superpopulation parameter $\boldsymbol{\beta}$.

With a sample of size $n$, we obtain a design consistent estimator of $\boldsymbol{\beta}_N$, $\hat{\boldsymbol{\beta}}$, by solving the weighted likelihood equation

$$\sum_{i\in A} \alpha_i(y_i - \hat{p}_i)\mathbf{x}_i = \sum_{i\in A} \alpha_i g(\mathbf{z}_i : \hat{\boldsymbol{\beta}}) = 0, \tag{2.5}$$

where $\hat{p}_i = \left[1 + \exp(-\mathbf{x}_i\hat{\boldsymbol{\beta}})\right]^{-1}$, $\alpha_i = \pi_i^{-1}$, $\pi_i$ is an inclusion probability and $A$ is a set of indices in the sample. Note that $\sum_{i\in A}\alpha_i g(\mathbf{z}_i : \boldsymbol{\beta})$ is design unbiased to $\sum_{i\in U} g(\mathbf{z}_i : \boldsymbol{\beta})$.

Under the assumption that auxiliary variable $\mathbf{x}_i$ is known for all $i \in U$, consider the predictor of the population proportion based on the MLE of $p_i$,

$$\bar{y}_{mle} = N^{-1}\left(\sum_{i\in A} y_i + \sum_{j\in U\setminus A}\hat{p}_j\right), \tag{2.6}$$

where $U\setminus A$ is a set of indices that are not selected in the sample and $\hat{p}_j$ is defined as (2.5). Valliant *et al.* (2000) suggested an estimator of the form (2.6) and investigated the properties of the estimator using a prediction approach.

In a large scale survey, design consistency is often required to an estimator for the protection from model failure. Under the assumptions on the sequence of populations, samples and sampling designs described in Isaki and Fuller (1982) or Park and Yang (2008), we define a sequence of estimators $\hat{\theta}_N$ of the population parameter $\theta_N$ to be design consistent, if for all $\epsilon > 0$, $\lim_{N\to\infty} P\{|\hat{\theta}_N - \theta_N| > \epsilon \,|\, \mathcal{F}_N\} = 0$, where the notation indicates that the $N^{th}$ finite population, $\mathcal{F}_N = (\mathbf{z}_1', \mathbf{z}_2', \ldots, \mathbf{z}_N)'$, is held fixed and the probability depends only on the sampling design. If the sampling design and the super-population model satisfy certain conditions, the estimator (2.6) is design-consistent.

**Result**: Assume $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ is a sequence of independent and identically distributed random vectors with finite fourth moment. Assume that the sequence of the finite populations and samples satisfies

$$\mathrm{Var}\left[n_N^{\frac{1}{2}} N^{-1}\left(\sum_{i\in A}\alpha_i\mathbf{z}_i - \sum_{i\in U}\mathbf{z}_i\right)\,\middle|\,\mathcal{F}_N\right] = \mathbf{V}_{zz,N}, \tag{2.7}$$

where $\mathbf{V}_{zz,N}$ is positive semi-definite *a.s.*. Assume for all $\boldsymbol{\beta}$ in a closed set $\mathcal{B}$ containing $\boldsymbol{\beta}_N$ as an interior point,

$$N^{-1}\sum_{i\in A}\alpha_i\mathbf{H}(\mathbf{z}_i, \boldsymbol{\beta}) = N^{-1}\sum_{i\in U}\mathbf{H}(\mathbf{z}_i, \boldsymbol{\beta}) + O_p\left(n_N^{-\frac{1}{2}}\right) \tag{2.8}$$

and

$$\lim_{N\to\infty} N^{-1}\sum_{i\in U}\mathbf{H}(\mathbf{z}_i, \boldsymbol{\beta}) = \mathbf{H}(\boldsymbol{\beta}) \quad a.s., \tag{2.9}$$

for some positive definite matrix $\mathbf{H}(\boldsymbol{\beta})$, where $\mathbf{H}(\mathbf{z}_i, \boldsymbol{\beta}) = \partial\mathbf{g}(\mathbf{z}_i : \boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and the function $\mathbf{g}(\cdot)$ is in (2.4) and (2.5). Assume

$$p\lim_{N\to\infty} N\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right)\,\middle|\,\mathcal{F}_N = 0 \quad a.s., \tag{2.10}$$

where $\boldsymbol{\beta}_N$ and $\hat{\boldsymbol{\beta}}$ are solutions of (2.4) and (2.5) respectively. If there exist a column vector $\boldsymbol{\gamma}_N$ such that

$$\mathbf{L}_\pi = \mathbf{X}_N\boldsymbol{\gamma}_N, \tag{2.11}$$

where $\mathbf{L}_\pi = (1-\pi_1, \ldots, 1-\pi_N)'$ and $\mathbf{X}_N = (\mathbf{x}_1', \ldots, \mathbf{x}_N')'$. Then

$$\bar{y}_{mle} = \bar{y}_N + O_p\left(n_N^{-\frac{1}{2}}\right). \tag{2.12}$$

**Proof**: By assumption (2.11) and the fact that $\hat{\boldsymbol{\beta}}$ is the solution of (2.5),

$$\begin{aligned}\sum_{i\in U}\hat{p}_i + \sum_{i\in A}\alpha_i(y_i-\hat{p}_i) &= \sum_{i\in U}\hat{p}_i + \sum_{i\in A}(y_i-\hat{p}_i) + \sum_{i\in A}\alpha_i(y_i-\hat{p}_i)(1-\pi_i)\\ &= \sum_{i\in U}\hat{p}_i + \sum_{i\in A}(y_i-\hat{p}_i)\\ &= \sum_{i\in A} y_i + \sum_{i\in U\setminus A}\hat{p}_i.\end{aligned} \tag{2.13}$$

Note that $|\mathbf{g}(\cdot)| < |\mathbf{x}_i|$ and functions $\mathbf{g}(\cdot)$ and $\mathbf{H}(\cdot)$ are continuous in $\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$ in a closed interval $\mathcal{B}$ containing $\boldsymbol{\beta}_N$ as an interior point, where $\boldsymbol{\beta}_N$ is the solution of (2.4) and $|\cdot|$ is a norm of a vector. By assumption (2.8) and Theorem 1.3.8 in Fuller (2008),

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = \left[\sum_{i\in A}\alpha_i\mathbf{H}(\mathbf{z}_i,\boldsymbol{\beta}_N)\right]^{-1}\sum_{i\in A}\alpha_i g(\mathbf{z}_i,\boldsymbol{\beta}_N) + o_p\left(n_N^{-\frac{1}{2}}\right). \tag{2.14}$$

For $\hat{\boldsymbol{\beta}} \in \mathcal{B}$, by a Taylor expansion,

$$p(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) = p_{iN} - p_{iN}(1-p_{iN})\mathbf{x}_i(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + o_p\left(n_N^{-\frac{1}{2}}\right). \tag{2.15}$$

Thus,

$$\begin{aligned}&N^{-1}\left[\sum_{i\in U}\hat{p}_i + \sum_{i\in A}\alpha_i(y_i-\hat{p}_i)\right]\\ &= N^{-1}\left[\sum_{i\in U}p_{iN} + \sum_{i\in A}\alpha_i(y_i-p_{iN}) + \left(\sum_{i\in A}\alpha_i\mathbf{q}_i - \sum_{i\in U}\mathbf{q}_i\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_N\right)\right] + o_p\left(n_N^{-\frac{1}{2}}\right)\\ &= N^{-1}\left[\sum_{i\in U}p_{iN} + \sum_{i\in A}\alpha_i(y_i-p_{iN})\right] + o_p\left(n_N^{-\frac{1}{2}}\right)\\ &= N^{-1}\left[\left(\sum_{i\in U}p_{iN} - \sum_{i\in A}\alpha_i p_{iN}\right) + \sum_{i\in A}\alpha_i y_i\right] + o_p\left(n_N^{-\frac{1}{2}}\right)\\ &= \bar{y}_N + O_p\left(n_N^{-\frac{1}{2}}\right),\end{aligned} \tag{2.16}$$

where $\mathbf{q}_i = p_{iN}(1-p_{iN})\mathbf{x}_i$. □

Note that function $\mathbf{g}(\cdot)$ in (2.4) and (2.5) is a vector of the derivative of the objective function (2.3) and its design unbiased quantity and thus, assumption (2.10) is obtained by using a property of the objective function. See also Gallant (1987). The conditions (2.7)

and (2.8) are the conditions for the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) of the population mean of a finite variable to be design consistent. For details on condition (2.7), see Isaki and Fuller (1982). If the column of ones is in the column space of $\mathbf{X}_N$, the condition (2.11) is equivalent such that the column of inclusion probabilities is in the column space of $\mathbf{X}_N$. If we select a stratified random sample and stratum indicators are used as a set of auxiliary variables, a common practice, then condition (2.11) is satisfied. Also condition (2.11) is satisfied for equal probability sampling designs. Thus, if the design and model satisfy condition (2.11), then a model-optimal estimator (2.6) is also robust to model failure in a large sample framework. That is, the estimator $\bar{y}_{mle}$ converges in probability to $\bar{y}_N$ even when the logistic regression model fails to describe the finite population. A design consistent estimator of the variance of $\bar{y}_{mle}$ can be obtained using the expression (2.16). One variance estimator is the Horvitz-Thompson variance estimator of $N^{-1}\sum_{i\in A}\alpha_i\hat{e}_i$, where $\hat{e}_i = y_i - \hat{p}_i$. For details, see Särndal *et al.* (1992).

## 3. Simulation Study

To investigate the performance of the estimation procedures, we first considered the finite population generated from a logistic regression model. The two binary variables, $Y_1$ and $Y_2$, were generated from the Bernoulli distribution, $y_i \sim$ Bernoulli$(p_i)$, where

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i\boldsymbol{\beta} = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2, \tag{3.1}$$

$x_{1i} \sim N(2, 1)$ and $x_{2i} \sim N(1, 1)$. For the variable $Y_1$, we used $\boldsymbol{\beta} = (-2, 2, 0)'$ and $\boldsymbol{\beta} = (-0.5, 1, -0.5)'$ was used for the variable $Y_2$. A finite population of size $N = 1,000$ was generated for the variables $Y_1$ and $Y_2$. For the simulation study, 10,000 simple random samples of size $n = 100$ were selected from the finite population. We considered four estimators for the population mean of $Y$.

1. The Horvitz-Thompson estimator

$$\bar{y}_{HT} = n^{-1}\sum_{i\in A} y_i. \tag{3.2}$$

2. The regression estimator

$$\bar{y}_{reg} = \sum_{i\in A} w_i y_i = \sum_{i\in A}\left[n^{-1} + (\bar{\mathbf{z}}_N - \bar{\mathbf{z}}_{HT})\left(\sum_{i=1}^{n}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\mathbf{z}_i\right] y_i, \tag{3.3}$$

   where $\mathbf{z}_i = (1, x_{1i})$, $\bar{\mathbf{z}}_N$ is the population mean of $\mathbf{z}_i$ and $\bar{\mathbf{z}}_{HT} = n^{-1}\sum_{i\in A}\mathbf{z}_i$.

3. The model calibration estimator (Wu and Sitter, 2001)

$$\bar{y}_{mc} = \bar{y}_{HT} + \left(\bar{\hat{p}}_N - \bar{\hat{p}}_{HT}\right)\hat{\gamma}, \tag{3.4}$$

   where $\hat{p}_i = \{1 + \exp[-(\mathbf{z}_i\hat{\boldsymbol{\beta}}_z)]\}^{-1}$, $\hat{\gamma} = [\sum_{i\in A}(\hat{p}_i - \bar{\hat{p}}_{HT})^2]^{-1}\sum_{i\in A}(\hat{p}_i - \bar{\hat{p}}_{HT})y_i$.

Table 3.1: Monte Carlo properties of the estimators

| | Variable | Estimators | | | |
|---|---|---|---|---|---|
| | | $\bar{y}_{HT}$ | $\bar{y}_{reg}$ | $\bar{y}_{mc}$ | $\bar{y}_{mle}$ |
| Relative Bias in percent(%) | $Y_1$ | 0.028 | 0.172 | 0.017 | 0.014 |
| | $Y_2$ | 0.018 | 0.050 | −0.012 | −0.017 |
| Relative MSE | $Y_1$ | 1.000 | 0.762 | 0.708 | 0.707 |
| | $Y_2$ | 1.000 | 0.856 | 0.849 | 0.847 |

4. The model-based design-consistent predictor, $\bar{y}_{mle}$.
   Note that we used information on $x_{1i}$ only in defining estimators for both variables $Y_1$ and $Y_2$.

Table 3.1 shows the Monte Carlo relative bias and the relative MSE of the estimators. The relative bias in percent is $[\bar{y}_N^{-1}(\hat{E}(\bar{y}) - \bar{y}_N)] \times 100$ and the relative MSE is the Monte Carlo MSE of the estimators relative to that of $\bar{y}_{HT}$. Generally, the regression estimator shows the best performance when the variable of interest has a linear relationship with auxiliary variables. However, the variable of interest is an indicator variable in our simulation study so the regression estimator has the large bias, although the largest one is less than 0.2%. Because the estimators are constructed based on $x_1$ only, the MSEs of the estimators for the $Y_2$ are larger than the MSEs of the estimator for $Y_1$. For both variables $Y_1$ and $Y_2$, $\bar{y}_{mc}$ and $\bar{y}_{mle}$ that are based on a logistic regression model show the better performances than $\bar{y}_{reg}$ and $\bar{y}_{HT}$.

We also considered an estimation of the population distribution function to investigate the small sample performance of the estimators. The variable of interest for the estimation of the population distribution function is a dichotomous variable and thus the estimator based on the logistic regression model could be a reasonable choice. In many previous studies, the performances of the population distribution function estimators were compared when the variable of interest has a linear relationship with auxiliary variables. We also considered a regression superpopulation model for $y$

$$y_i = 2 + x_i + e_i, \tag{3.5}$$

where $e_i \sim N(0, 4)$. To define a symmetric population distribution, we generated $x_i$ from the normal distribution with a mean of zero and a variance of nine. Let denote the symmetric population based on normal $x$ as $Y_3$. For the variable $Y_4$ having a skewed distribution, we generated $x_i$ from the $\chi^2$ distribution with two degrees of freedom. For each variable $Y_3$ and $Y_4$, we define seven indicator variables

$$I_{iP} = \begin{cases} 1, & \text{if } \ y_i \le q_{iP}, \\ 0, & \text{elsewhere}, \end{cases} \tag{3.6}$$

for $i = 3, 4$, $P = 0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95$, where $q_{iP}$ is the minimum value that satisfies $N^{-1}\sum_{i\in U} I(y_i < q_{iP}) = P$. Note that the population mean of the indicator

Table 3.2: Monte Carlo properties of the estimators: Normal population

| | Relative Bias in percent | | | | Relative MSE | | |
|---|---|---|---|---|---|---|---|
| $P$ | $\bar{y}_{HT}$ | $\bar{y}_{reg}$ | $\bar{y}_{mc}$ | $\bar{y}_{mle}$ | $\bar{y}_{reg}$ | $\bar{y}_{mc}$ | $\bar{y}_{mle}$ |
| 0.05 | 0.052 | −2.749 | 1.614 | 0.457 | 0.805 | 0.721 | 0.632 |
| 0.10 | −0.388 | −2.056 | −0.429 | −0.252 | 0.746 | 0.729 | 0.579 |
| 0.25 | −0.015 | −0.560 | −0.070 | −0.053 | 0.644 | 0.678 | 0.570 |
| 0.50 | 0.138 | 0.132 | 0.045 | 0.087 | 0.605 | 0.649 | 0.562 |
| 0.75 | 0.085 | 0.262 | 0.024 | 0.038 | 0.668 | 0.682 | 0.590 |
| 0.90 | 0.049 | 0.192 | 0.016 | 0.008 | 0.809 | 0.842 | 0.694 |
| 0.95 | 0.033 | 0.154 | −0.055 | 0.005 | 0.854 | 0.802 | 0.715 |

Table 3.3: Monte Carlo properties of the estimators: $\chi^2(2)$ population

| | Relative Bias in percent | | | | Relative MSE | | |
|---|---|---|---|---|---|---|---|
| $P$ | $\bar{y}_{HT}$ | $\bar{y}_{reg}$ | $\bar{y}_{mc}$ | $\bar{y}_{mle}$ | $\bar{y}_{reg}$ | $\bar{y}_{mc}$ | $\bar{y}_{mle}$ |
| 0.05 | −0.350 | −1.504 | NA | −0.216 | 0.969 | NA | 0.975 |
| 0.10 | −0.201 | −1.257 | NA | −0.105 | 0.950 | NA | 0.947 |
| 0.25 | −0.040 | −0.900 | 0.040 | −0.014 | 0.904 | 1.506 | 0.869 |
| 0.50 | −0.065 | −0.644 | −0.061 | −0.065 | 0.811 | 0.818 | 0.758 |
| 0.75 | −0.056 | −0.313 | −0.066 | −0.074 | 0.717 | 0.750 | 0.678 |
| 0.90 | −0.017 | −0.024 | −0.023 | −0.042 | 0.715 | 1.029 | 0.654 |
| 0.95 | −0.025 | 0.046 | 0.140 | −0.052 | 0.762 | 14.816 | 0.689 |

variable (3.6) is the population distribution function. To define a model calibration estimator of the form (3.4), we modified the estimator as Wu and Sitter (2001) did in their simulation such that the estimator satisfies the property of calibration. That is, if the estimator is applied to estimate the population distribution function of an auxiliary variable, it gives the true population distribution function of the auxiliary variable. The modified model calibration estimator is defined as (3.4) by replacing $\hat{p}_k$ with $I(\hat{\delta}_0 + \hat{\delta}_1 x_k < q_{iP})$, where $(\hat{\delta}_0, \hat{\delta}_1)'$ is the ordinary least squares estimator of the linear regression coefficient. For a simulation study, 10,000 simple random samples of size 100 were selected.

Table 3.2 shows the Monte Carlo properties of the estimators for the normal population. The relative bias is $[P^{-1}(\hat{E}(\bar{y}) - P)] \times 100$. For the middle or large value of $P$, all estimators have the negligible bias. For $P = 0.05$ and $P = 0.10$, regression estimator has the significant bias. The bias of $\bar{y}_{mle}$ is smaller than $\bar{y}_{mc}$ for all small $P$s. The relative MSE is the Monte Carlo MSE of the estimator relative to that of $\bar{y}_{HT}$. For this symmetric population, $\bar{y}_{mle}$ is the most efficient estimator for all values of $P$.

Table 3.3 shows the Monte Carlo properties of the estimators in estimating the distribution function of a skewed distribution. In this skewed population, $\bar{y}_{mc}$ is not defined for $P = 0.05$ and $P = 0.10$ because no observation in the sample satisfies $I(\hat{\delta}_0 + \hat{\delta}_1 x_k < q_{iP})$. Also $\bar{y}_{mc}$ did not have solution in 3776 samples for $P = 0.25$, 95 samples for $P = 0.90$ and 702 samples for $P = 0.95$. For $P$ =0.25, 0.90 and 0.95, we used $\bar{y}_{HT}$ when $\bar{y}_{mc}$ did not have a solution. Because $\bar{y}_{mc} = \bar{y}_{HT}$ when no solution for $\bar{y}_{mc}$ is available, the

estimator $\bar{y}_{mc}$ for $P$ =0.25, 0.90 and 0.95 has a large MSE. For all $P$, $\bar{y}_{HT}$ and $\bar{y}_{mle}$ have a bias of less than 0.4%. The regression estimator has a significant bias for $P = 0.05$ and $P = 0.10$. Because the population is skewed to the right, simple random sample tends to have less units that are quite far from the mean and thus we have many negative biases. Regression estimator has the relatively large absolute bias for all $P$ except $P = 0.90$ and $P = 0.95$. For a skewed distribution, $\bar{y}_{mle}$ has the smallest MSE in all $P$ except for $P = 0.05$. For this skewed population, the regression estimator shows a better performance than $\bar{y}_{mc}$ with respect to MSE.

## 4. Discussion

As an estimator of the population proportion, we consider the approximately model-optimal estimator based on the MLE of the logistic regression coefficient. Under the reasonable conditions on the sampling design and model, the efficient estimator obtained under the logistic regression model is also design-consistent. Simulation study shows that the defined estimator performs well in estimating the population proportion even when the significant auxiliary variable is missed in defining the estimator. The model-efficient design-consistent estimator also shows a better performance in estimating the population distribution than other estimators. However one should be cautious in generalizing the simulation results because the performances of the estimators were compared in a limited simulation set up. More general sampling designs and different parameters can be considered in the future study. Due to the poor performance in a simulation study, it is not recommended to use the model calibration estimator in predicting the skewed population distribution function.

## Acknowledgment

The author thank the associate editor and referees for the comments that significantly improved this paper.

## References

Chambers, R. L., Dorfman, A. H. and Hall, P. (1992). Properties of estimators of finite population distribution function, *Biometrika,* **79**, 577–582.

Chambers, R. L. and Dunstan, R. (1986). Estimating distribution functions from survey data, *Biometrika,* **73**, 597–604.

Fuller, W. A. (2008). *Sampling Statistics,* Iowa State University, Ames, IA.

Gallant, A. R. (1987). *Nonlinear Statistical Models,* John Wiley & Sons, New York.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles, *Survey methodology,* **32**, 37–52.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association,* **47**, 663–685.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model, *Journal of the American Statistical Association,* **77**, 89–96.

Park, M. and Yang, M. (2008). Ridge regression estimation for survey samples, *Communications in Statistics - Theory and Method,* **37**, 532–543.

Rao, J. N. K., Kovar, J. G. and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information, *Biometrika,* **77**, 365–375.

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

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

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *Journal of the American Statistical Association,* **96**, 185–193.

[Received July 2008, Accepted August 2008]