Ridge Regression Estimation for Survey Samples

Mingue Park* and Min Yang

Korea University and University of Missouri

Abstract

A procedure for constructing a vector of regression weights is considered. Under the regression superpopulation model, the ridge regression estimator that has minimum model mean squared error is derived. Through a simulation study, the ridge regression weights, regression weights, quadratic programming weights and raking ratio weights are compared. The ridge regression procedure with weights bounded by zero performed very well.

Keywords: Regression superpopulation model, Ridge regression, Model MSE, Design consistency

1 Introduction

In survey sampling, information about the population, often called auxiliary information, is commonly incorporated by means of regression estimation or raking. A review of the use of auxiliary information in regression estimation for sample surveys is given by Fuller (2002). The raking method is credited to Deming and Stephan (1940). One regression estimator of the population mean can be defined as a linear estimator, $\bar{y}_{lin} = \sum_{i=1}^{n} w_i \, y_i,$ where the w_is' minimize

$$\sum_{i=1}^{n}(w_i - \alpha_i)^2 \alpha_i^{-1} \quad , \tag{1.1}$$

*Address correspondence to Mingue Park, Department of Statistics, Korea University, Seoul, 136-701, South Korea; E-Mail:mpark2@korea.ac.kr

subject to the vector of constraints

$$\sum_{i=1}^{n} w_i \mathbf{x}_i = \bar{\mathbf{x}}_N \ , \tag{1.2}$$

$\alpha_i = \left(\sum_{j=1}^{n} \pi_j^{-1}\right)^{-1} \pi_i^{-1}$, $\mathbf{x}$ is a vector of auxiliary variables, $\bar{\mathbf{x}}_N$ is the population mean of $\mathbf{x}$ and the π_i's are the selection probabilities.

In defining the regression estimator, regression superpopulation model is often introduced to describe a population and (or) as a basis of the estimation of the population characteristics. If the auxiliary variables have strong linear relationship with study variables then regression estimator permits a significant gain in efficiency by the constraints (1.2). In a large scale survey, many variables of interest are considered and it is not possible to construct a single model that is appropriate for all variables of interest. In such a case, if relevant auxiliary variables to a certain set study variables are omitted from the model then estimators may have substantial model bias. On the other hand, if one attempts to use a large vector of the auxiliary variables, some of the weights for the estimator could be extremely large or negative or it may be impossible to satisfies the restrictions on weights (1.2). If the regression weights are to be used to estimate a finite population total in a general purpose survey, it seems reasonable that no individual weight should be less than one. Also, it seems reasonable, on robustness ground, to avoid very extreme weights.

There are several ways to reduce the range of regression weights directly. One is to modify the w_i defining the estimator so that there are no negative weights and no large weights. Huang and Fuller (1978) is an early paper defining such a procedure. A number of procedures build on the fact that the weights can be defined as values that optimize some function. Deville and Särndal (1992) considered several objective functions that can be used to construct weights. Singh and Mohl (1996) compared several nonnegative regression type estimators through numerical examples. Using conditional inclusion probabilities, Park and Fuller (2005) introduced a set of regression weights that are positive in most samples.

Another modification of regression weights is to reduce the number of constraints or to relax some of the restrictions used in constructing the estimator. Bardsley and Chamber (1984) introduced a ridge regression estimator in which the restrictions (1.2) were added to the objective function (1.1) with a certain coefficient matrix. However no suggestion for choosing the coefficient matrix was made.

In this paper, we consider a procedure that replaces some of the linear restrictions in (1.2) with added components in the objective function (1.1). We derive the coefficient matrix for the added components such that the defined ridge regression estimator has approximately the minimum model mean squared error (MSE). Using quadratic programming, we generate nonnegative ridge regression weights. To investigate the performance of alternative regression type estimators when the model is misspecified, we compare regression weights, ridge regression weights, quadratic programming weights and raking ratio weights as weights for estimating the population percentiles.

2 Ridge Regression Estimation

Consider the regression superpopulation model

$$y_i = \beta_0 + \mathbf{x}_i \boldsymbol{\beta} + e_i \ , \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$ and $(\beta_0, \boldsymbol{\beta}')'$ is the vector of regression coefficients. Assume we have a working diagonal covariance matrix for $(e_1, \cdots, e_n)$, denoted by $\boldsymbol{\Phi} = \text{Diag}(\phi_{11}, \cdots, \phi_{nn})$, for the model (2.1). It is often assumed, in practice, that $\boldsymbol{\Phi}$ is known up to a constant. The possible choices of $\boldsymbol{\Phi}$ are well discussed in Särndal, Swensson and Wretman (1991, Ch. 5, 6). A ridge regression estimator of the vector of regression coefficients was originally proposed by Hoerl and Kennard (1970) to construct a nonsensitive estimator for the regression coefficients when there is multicollinearity among predictors or when sample size is small relative to the number of predictors. Bardsley and Chamber (1984) introduced a procedure that relaxes the constraints on weights (1.2) and showed that the procedure is a type of ridge estimator.

To define a ridge regression estimator, consider the procedure that replaces the restrictions (1.2) with an added component in the objective function (1.1) with a coefficient diagonal matrix $\boldsymbol{\Psi}$. That is, the weights for the ridge regression estimator is obtained by minimizing

$$Q = (\mathbf{w} - \boldsymbol{\alpha})' \boldsymbol{\Phi} (\mathbf{w} - \boldsymbol{\alpha}) + (\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N) \boldsymbol{\Psi} (\mathbf{w}'\mathbf{X} - \bar{\mathbf{x}}_N)' \ , \tag{2.2}$$

with a constraint $\mathbf{w}'\mathbf{J}_n - 1 = 0$, where $\mathbf{w} = (w_1, \cdots, w_n)'$, $\mathbf{X} = (\mathbf{x}_1', \cdots, \mathbf{x}_n')'$, $\bar{\mathbf{x}}_N = (\bar{x}_{1,N}, \cdots, \bar{x}_{p,N})$, $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$, $\mathbf{\Psi}$ is a positive definite diagonal matrix, and $\mathbf{J}_n$ is the column vector of ones. The solution for the vector of weights is

$$\mathbf{w} = \boldsymbol{\alpha} + \mathbf{\Phi}^{-1/2}(\mathbf{I} - \mathbf{P}_{J_n^*})\mathbf{X}^* \left[\mathbf{X}^{*\prime}(\mathbf{I} - \mathbf{P}_{J_n^*})\mathbf{X}^* + \mathbf{\Psi}^{-1}\right]^{-1} (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)' \tag{2.3}$$

where $\mathbf{P}_{J_n^*} = \mathbf{J}_n^*(\mathbf{J}_n^{*\prime}\mathbf{J}_n^*)^{-1}\mathbf{J}_n^{*\prime}$, $(\mathbf{J}_n^*, \mathbf{X}^*) = \mathbf{\Phi}^{-1/2}(\mathbf{J}_n, \mathbf{X})$, $\bar{\mathbf{x}}_\pi = \sum_{i=1}^n \alpha_i \mathbf{x}_i$ and α_i is of (1.1). The ridge regression estimator defined with the vector of weights (2.3) is

$$\bar{y}_{rreg} = \mathbf{w}'\mathbf{y} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}_{rid}\ , \tag{2.4}$$

where

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{rid} &= \left[\mathbf{X}^{*\prime}\left(\mathbf{I} - \mathbf{P}_{J_n^*}\right)\mathbf{X}^* + \mathbf{\Psi}^{-1}\right]^{-1}\left[\mathbf{X}^{*\prime}\left(\mathbf{I} - \mathbf{P}_{J_n^*}\right)\mathbf{y}^*\right] \\ &= \left[\sum_{i=1}^n \phi_{ii}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_\phi)'(\mathbf{x}_i - \bar{\mathbf{x}}_\phi) + \mathbf{\Psi}^{-1}\right]^{-1}\left[\sum_{i=1}^n \phi_{ii}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}_\phi)' y_i\right]\end{aligned} \tag{2.5}$$

$\bar{\mathbf{x}}_\phi = \left(\sum_{i=1}^n \phi_{ii}^{-1}\right)^{-1} \sum_{i=1}^n \phi_{ii}^{-1}\mathbf{x}_i$, $\mathbf{y}^* = \mathbf{\Phi}^{-1/2}\mathbf{y}$ and $\mathbf{y} = (y_1, \cdots, y_n)'$. Note that $\hat{\boldsymbol{\beta}}_{rid}$ has the form of a generalized ridge estimator of $\boldsymbol{\beta}$ in (2.1).

3 Optimal Coefficient Matrix $\mathbf{\Psi}$

In this scetion, we show that deriving a linear estimator that has the minimum model MSE is equivalent to deriving the optimal value of $\mathbf{\Psi}$ for the ridge regression estimator of (2.4) and define the optimal ridge regression estimator under the multiple regression superpopulation model. To motivate the procedure of deriving the optimal $\mathbf{\Psi}$ that minimizes the MSE of the ridge regression estimator, consider a single x-variable. Assume the linear model (2.1) with a single explanatory variable x_i where e_i's are independently distributed with mean zero and variance σ_e^2, and β_0 and β are the regression coefficients. Consider a linear estimator of the population mean of y, $\bar{y}_{lin} = \sum_{i=1}^n w_i\, y_i$, where $\sum_{i=1}^n w_i = 1$. Then the error of the linear estimator in estimating the population mean of y is $\bar{y}_{lin} - \bar{y}_N = \sum_{i=1}^n\ w_i e_i - \bar{e}_N + \left(\sum_{i=1}^n\ w_i x_i - \bar{x}_N\right)\beta$, where $\bar{y}_N$, $\bar{x}_N$ and $\bar{e}_N$ are the population means of y, x and e, respectively. The model MSE of the linear estimator for the known $\bar{x}_N$ and conditional on the sample x's, if we ignore the finite population correction factor, is

$$\begin{aligned} \mathrm{E}\left\{\left(\sum_{i=1}^{n} w_i y_i - \bar{y}_N\right)^2 \middle| \mathbf{x}\right\} &= \sum_{i=1}^{n} w_i^2\sigma_e^2 + \left(\sum_{i=1}^{n} w_i x_i - \bar{x}_N\right)^2 \beta^2 \qquad (3.1) \\ &= \left(1-R^2\right)\,\sigma_y^2 \left[\sum_{i=1}^{n} w_i^2 + \frac{R^2}{\left(1-R^2\right)\,\sigma_x^2}\left(\sum_{i=1}^{n} w_i x_i - \bar{x}_N\right)^2\right], \end{aligned}$$

where σ_y^2 and σ_x^2 are the population variances of y and x, σ_{xy} is the population covariance between x and y and $R^2 = (\sigma_x^2\sigma_y^2)^{-1}\sigma_{xy}^2$. Let R^2 and σ_x^2 be known and the quadratic function of weights can be approximated by $\sum_{i=1}^{n}(w_i-\alpha_i)^2\alpha_i^{-1} \approx \sum_{i=1}^{n}(w_i-\alpha_i)^2\bar{\alpha}^{-1} \approx \sum_{i=1}^{n} w_i^2\bar{\alpha}^{-1} - n\bar{\alpha}$, where α_i is the initial weight and $\bar{\alpha}$ is the sample mean of the α_i. Then the set of weights which minimizes the objective function (2.2), with $\mathbf{X} = (x_1,\cdots,x_n)'$, $\mathbf{\Phi} = \mathrm{diag}(\alpha_1^{-1},\cdots,\alpha_n^{-1})$, $\mathbf{\Psi} = \psi\mathbf{I}$ and

$$\psi = [(1-R^2)\bar{\alpha}\sigma_x^2]^{-1}R^2, \qquad (3.2)$$

would minimizes the model MSE of the linear estimator $\bar{y}_{lin}$, where $\mathbf{I}$ is an identity matrix. Thus, the procedure that replaces the linear restriction of (1.2) with an added component in the objective function (1.1) with the coefficient ψ of (3.2) is approximately equivalent to finding weights that minimize the MSE of the linear estimator.

The weights and corresponding ridge regression estimator with $\alpha_i = n^{-1}$, by (2.3) and (2.4), are $w_i = n^{-1} + \psi\,[n+\psi s_{xx}]^{-1}\,(\bar{x}_N - \bar{x}_n)(x_i - \bar{x}_n)$, and

$$\bar{y}_{rreg} = (1-\gamma)\bar{y}_n + \gamma\bar{y}_{reg} \ , \qquad (3.3)$$

where $\bar{y}_{reg} = \bar{y}_n + (\bar{x}_N - \bar{x}_n)\,\hat{\beta}$, $\hat{\beta} = s_{xx}^{-1}s_{xy}$, $\gamma = [n+\psi s_{xx}]^{-1}\,\psi s_{xx}$, $s_{xx} = \sum_{i=1}^{n}(x_i-\bar{x}_n)^2$, $s_{xy} = \sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)$ and ψ of (3.2). The ridge regression estimator $\bar{y}_{rreg}$ of (3.3) is a linear combination of the sample mean and the ordinary least square (OLS) regression estimator. If $\psi \to \infty$, which implies that the correlation between x and y becomes one, then $\bar{y}_{rreg}$ converges to the regression estimator. If $\psi \to 0$, implying that the correlation becomes zero, then $\bar{y}_{rreg}$ converges to the sample mean.

With the relationship between the model MSE of a linear estimator $\sum_{i=1}^n w_i y_i$ and the model MSE of ridge regression estimator, optimal γ and the corresponding ψ can be also obtained directly by minimizing the model MSE of the ridge regression estimator. Under the linear model, the conditional MSE of $\bar{y}_{rreg}$ is

$$\begin{aligned} \mathrm{E}\{(\bar{y}_{rreg} - \bar{y}_N)^2|\mathbf{x}\} =& \mathrm{E}\{[\bar{e}_n - \bar{e}_N + (1-\gamma)(\bar{x}_n - \bar{x}_N)\beta + \gamma(\bar{x}_N - \bar{x}_n)(\hat{\beta} - \beta)]^2|\mathbf{x}\} \\ \approx& \left(1 - \frac{n}{N}\right)\frac{\sigma_e^2}{n} + (1-\gamma)^2(\bar{x}_n - \bar{x}_N)^2 R^2 \frac{\sigma_y^2}{\sigma_x^2} + \gamma^2(\bar{x}_n - \bar{x}_N)^2 \frac{(1-R^2)\sigma_y^2}{(n-2)\sigma_x^2} \; . \end{aligned} \tag{3.4}$$

The approximation is due to the approximation of $(n-2)^{-1}s_{xx}$ by σ_x^2. If we differentiate the approximate conditional MSE of (3.4) with respect to γ and set the result equal to zero, then the MSE is minimized with

$$\gamma_{opt} = [1 + (n-3)R^2]^{-1}(n-2)R^2 \; . \tag{3.5}$$

The corresponding ψ is $\psi_{opt} = \{(1-R^2)\;[s_{xx}/(n-2)]\}^{-1}(nR^2)$. Given the ψ of (3.2), we confirm that the γ of (3.3) is equal to γ_{opt} of (3.5). In practice, the population correlation coefficient R^2 is often unknown. A design consistent estimator of R^2 such as the sample correlation coefficient can be used to define the $\bar{y}_{rreg}$.

For the case that multiple auxiliary variables are available, we consider a procedure that replaces some of the linear restrictions in (1.2), with a component in (1.1) that is a positive definite quadratic form in the replaced restrictions. Assume the linear model (2.1). Let $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)'$ be the column vector of initial weights given in (1.1) and assume the population mean of $\mathbf{x}$, $\bar{\mathbf{x}}_N$, is known. Let the matrix of observations on the auxiliary variables, $\mathbf{X}$, be partitioned as $(\mathbf{X}_1, \mathbf{X}_2)$, where $\mathbf{X}_1$ is the set of $p_1 < p$ variables for which exact constraints are required and $\mathbf{X}_2$ is the set of $p_2 = p - p_1$ variables for which the constraints can be relaxed. The partition of auxiliary variables $\mathbf{X}$ could be based on the strength of the relationship of auxiliary variables with the set of variables of interest or based on the relative number of auxiliary variables to sample size. For example, if the researcher believes that subset of auxiliary variables, denoted by $\mathbf{X}_1$, are significantly correlated with the study variables, he or she will keep the exact constraints for these variables and relax the constraints for the remaining auxiliary variables. If the number of auxiliary variables is large then we can obtain a stable solution for the regression coefficient estimator and an stable estimator of the population means of the study variables by relaxing the subset or all linear constraints.

A generalization of (1.1) and (1.2) is the function

$$(\mathbf{w}-\boldsymbol{\alpha})'\boldsymbol{\Phi}(\mathbf{w}-\boldsymbol{\alpha})+(\mathbf{w}'\mathbf{X}_2-\bar{\mathbf{x}}_{2,N})\boldsymbol{\Psi}(\mathbf{w}'\mathbf{X}_2-\bar{\mathbf{x}}_{2,N})' \ , \tag{3.6}$$

and the constraint

$$\mathbf{w}'(\mathbf{J}_n,\mathbf{X}_1)-(1,\bar{\mathbf{x}}_{1,N})=(0,\mathbf{0}) \ , \tag{3.7}$$

where $\mathbf{w}=(w_1,\cdots,w_n)'$, $\mathbf{X}_1=(\mathbf{x}'_{1,1},\cdots,\mathbf{x}'_{1,n})'$, $\mathbf{X}_2=(\mathbf{x}'_{2,1},\cdots,\mathbf{x}'_{2,n})'$, $\bar{\mathbf{x}}_{1,N}=(\bar{x}_{1,N},\cdots,\bar{x}_{p_1,N})$, $\bar{\mathbf{x}}_{2,N}=(\bar{x}_{p_1+1,N},\cdots,\bar{x}_{p,N})$, $\mathbf{x}_{1,i}=(x_{i1},\cdots,x_{ip_1})$, $\mathbf{x}_{2,i}=(x_{i(p_1+1)},\cdots,x_{ip})$, $\boldsymbol{\Phi}$ is a known positive definite diagonal matrix and $\boldsymbol{\Psi}$ is a positive definite diagonal matrix to be determined, and $\mathbf{J}_n$ is the column vector of ones. We add the column of ones in the constraints (3.7) so that the resulting estimator is location invariant. The solution for $\mathbf{w}$ is

$$\mathbf{w}=\boldsymbol{\alpha}+\boldsymbol{\Phi}^{-1/2}(\mathbf{I}-\mathbf{P}_{J_n^*})\mathbf{X}^*\left[\mathbf{X}^{*\prime}(\mathbf{I}-\mathbf{P}_{J_n^*})\mathbf{X}^*+\boldsymbol{\Psi}^{\dagger}\right]^{-1}(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_\pi)' \ , \tag{3.8}$$

where $\boldsymbol{\Psi}^{\dagger}=\text{Diag}\left(\mathbf{0},\boldsymbol{\Psi}^{-1}\right)$, $\mathbf{X}=(\mathbf{X}_1,\mathbf{X}_2)$, and $\mathbf{P}_{J_n^*}$, $\mathbf{J}_n^*$ and $\mathbf{X}^*$ are defined in (2.3). The ridge regression estimator defined with the vector of weights (3.8) is

$$\bar{y}_{rreg}=\mathbf{w}'\mathbf{y}=\bar{y}_\pi+(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_\pi)\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}} \, , \tag{3.9}$$

where

$$\boldsymbol{\Gamma}=\left[\sum_{i=1}^{n}\phi_{ii}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)'(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)+\boldsymbol{\Psi}^{\dagger}\right]^{-1}\left[\sum_{i=1}^{n}\phi_{ii}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)'(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)\right] \ , \tag{3.10}$$

and $\hat{\boldsymbol{\beta}}=\left[\sum_{i=1}^{n}\phi_{ii}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)'(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)\right]^{-1}\left[\sum_{i=1}^{n}\phi_{ii}^{-1}(\mathbf{x}_i-\bar{\mathbf{x}}_\phi)'y_i\right]$. Note that the ridge regression estimator (3.9) with $y=\mathbf{x}$ is $\bar{\mathbf{x}}_{rreg}=\bar{\mathbf{x}}_\pi+(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_\pi)\boldsymbol{\Gamma}$, and is not equal to $\bar{\mathbf{x}}_N$ unless $\boldsymbol{\Gamma}=\mathbf{I}$. That is, by relaxing the linear constraints on weights, the estimator does not generate the exact population mean of $\mathbf{x}$ when the estimator is applied to auxiliary variables. The error of $\bar{y}_{rreg}$ in estimating $\bar{y}_N$ is

$$\begin{aligned}\bar{y}_{rreg}-\bar{y}_N&=\bar{\mathbf{x}}_\pi\boldsymbol{\beta}+\bar{e}_\pi+(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_\pi)\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}-\bar{\mathbf{x}}_N\boldsymbol{\beta}-\bar{e}_N\\&=\bar{e}_\pi-\bar{e}_N+(\bar{\mathbf{x}}_\pi-\bar{\mathbf{x}}_N)(\mathbf{I}-\boldsymbol{\Gamma})\boldsymbol{\beta}+(\bar{\mathbf{x}}_N-\bar{\mathbf{x}}_\pi)\boldsymbol{\Gamma}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \, .\end{aligned} \tag{3.11}$$

Without loss of generality, assume $\mathbf{X}$ is transformed such that $\mathbf{X}_1^{*\prime}\left(\mathbf{I}-\mathbf{P}_{J_n^*}\right)\mathbf{X}_2^* = 0$, and a variance estimator $\hat{\mathrm{V}}\{\bar{\mathbf{x}}_{2\pi}\}$ and $\mathbf{X}_2^{*\prime}\left(\mathbf{I}-\mathbf{P}_{J_n^*}\right)\mathbf{X}_2^*$ are diagonal matrices. For simultaneous diagonalization of two symmetric matrices, see Harville (1997, Ch. 21). Then, the matrix $\mathbf{\Gamma}$ defined in (3.10) is $\mathbf{\Gamma} = \mathrm{Diag}\left(\mathbf{I},\mathbf{\Gamma}_{22}\right)$, where $\mathbf{\Gamma}_{22} = \left[\mathbf{\Psi}^{-1} + \mathbf{X}_2^{*\prime}\left(\mathbf{I}-\mathbf{P}_{J_n^*}\right)\mathbf{X}_2^*\right]^{-1}\mathbf{X}_2^{*\prime}\left(\mathbf{I}-\mathbf{P}_{J_n^*}\right)\mathbf{X}_2^*$. Thus the conditional MSE of the $\bar{y}_{rreg}$ under the model (2.1) can be approximated by, if we ignore the finite population correction factor,

$$\begin{aligned}\mathrm{E}\{(\bar{y}_{rreg}-\bar{y}_N)^2|\mathbf{X}\} \approx{}& \boldsymbol{\alpha}'\mathbf{\Phi}\boldsymbol{\alpha} + \left(\bar{\mathbf{x}}_{1,\pi}-\bar{\mathbf{x}}_{1,N}\right)\left[\mathrm{V}\{\hat{\boldsymbol{\beta}}_1|\mathbf{X}\}\right]\left(\bar{\mathbf{x}}_{1,\pi}-\bar{\mathbf{x}}_{1,N}\right)' \\ &+ \boldsymbol{\beta}_2'\left(\mathbf{I}-\mathbf{\Gamma}_{22}\right)\left(\bar{\mathbf{x}}_{2,\pi}-\bar{\mathbf{x}}_{2,N}\right)'\left(\bar{\mathbf{x}}_{2,\pi}-\bar{\mathbf{x}}_{2,N}\right)\left(\mathbf{I}-\mathbf{\Gamma}_{22}\right)\boldsymbol{\beta}_2 \\ &+ \left(\bar{\mathbf{x}}_{2,\pi}-\bar{\mathbf{x}}_{2,N}\right)\mathbf{\Gamma}_{22}\left[\mathrm{V}\{\hat{\boldsymbol{\beta}}_2|\mathbf{X}\}\right]\mathbf{\Gamma}_{22}\left(\bar{\mathbf{x}}_{2,\pi}-\bar{\mathbf{x}}_{2,N}\right)' \quad .\end{aligned} \tag{3.12}$$

The approximation is due to the approximation of the covariance between $\bar{e}_\pi$ and $\hat{\boldsymbol{\beta}}$ by zero. If the initial weights and variance matrix satisfy the condition, $\boldsymbol{\alpha}'\mathbf{\Phi} = c\mathbf{J}_n$ for some c, the covariance is exactly zero. By replacing $(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N)'(\bar{\mathbf{x}}_\pi - \bar{\mathbf{x}}_N)$ with $\mathrm{V}\{\bar{\mathbf{x}}_\pi\}$, and replacing conditional variances of regression coefficient estimators with unconditional variances, we obtain an approximate unconditional MSE of $\bar{y}_{rreg}$ as

$$\begin{aligned}\mathrm{E}\left\{(\bar{y}_{rreg}-\bar{y}_N)^2\right\} \approx{}& \boldsymbol{\alpha}'\mathbf{\Phi}\boldsymbol{\alpha} + \mathrm{Tr}\left\{\left[\mathrm{V}\left(\hat{\boldsymbol{\beta}}_1\right)\right]\left[\mathrm{V}\left(\bar{\mathbf{x}}_{1,\pi}\right)\right]\right\} \\ &+ \sum_{i=p_1+1}^{p}\beta_i^2(1-\gamma_{ii})^2\left[\mathrm{V}\{\bar{x}_{i,\pi}\}\right] + \sum_{i=p_1+1}^{p}\gamma_{ii}^2\left[\mathrm{V}\{\hat{\beta}_i\}\right]\left[\mathrm{V}\{\bar{x}_{i,\pi}\}\right] \quad ,\end{aligned} \tag{3.13}$$

where γ_{ii} is a diagonal element of $\mathbf{\Gamma}$ and $\mathrm{V}\left(\bar{\mathbf{x}}_\pi\right)$ is a model variance of $\bar{\mathbf{x}}_\pi$ under a certain model assumption on $\mathbf{x}_i$ such as $\mathbf{x}_i$ are independently distributed with common mean and finite variance. If $\mathbf{\Phi} = \phi\mathbf{I}$ and $\mathbf{x}_i$ are independent multivariate normal random variables with a common mean and covariance, then the conditional model MSE can be calculated using the moments of the Wishart and Hotelling's T^2 distributions. Note that $\left[\mathrm{V}\{\hat{\boldsymbol{\beta}}|\mathbf{X}\}\right]^{-1}$ has a Wishart distribution and $c\ (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\left[\mathrm{V}\{\hat{\boldsymbol{\beta}}|\mathbf{X}\}\right](\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)'$ has a Hotelling's T^2 distribution for some constant c under the normal assumption on $\mathbf{x}$. See Ch.5 and Ch.6 of Anderson (1984).

The diagonal $\mathbf{\Gamma}$ that minimizes the approximate model MSE defined in (3.13) is $\gamma_{jj,opt} = \left[\beta_j^2 + \mathrm{V}(\hat{\beta}_j)\right]^{-1}\beta_j^2$, for $j = p_1+1, \cdots, p$. By the relationship between $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ in (3.10), the optimal ψ_{jj} is $\psi_{jj,opt} = \left[s_{j\phi j}\mathrm{V}(\hat{\beta}_j)\right]^{-1}\beta_j^2$, where $s_{j\phi j}$ is the j-th diagonal element of $\mathbf{X}_2^{*\prime}\left(\mathbf{I}-\mathbf{P}_{J_n^*}\right)\mathbf{X}_2^*$.

Husain (1969) considered the objective function (3.6) and the constraints (3.7) for a simple random sample from a normal distribution with $\mathbf{\Phi} = \phi\mathbf{I}$ and $\mathbf{\Psi}^{-1} = \psi^{-1}\mathbf{I}$ and derived the ψ that minimizes the MSE of the estimator of the form (3.9). For a diagonal matrix $\mathbf{\Psi}$, Rao and Singh (1997) introduced the tolerance matrix $\mathbf{\Delta} = \text{diag}(\delta_1, \cdots, \delta_p)$ where each diagonal element δ_i is the tolerance for the i-th linear constraint on x. In practice, the regression coefficients and the variances of the regression coefficient estimators are unknown. Design consistent estimators of these unknown parameters can be used to define the ridge regression estimator.

4 Design Consistency of a Ridge Regression Estimator

To investigate the large sample properties of the ridge regression estimator, we consider a sequence of populations, samples, and sampling designs. The set of indices for the N-th finite population is $U_N = \{1, \cdots, N\}$, where $N = 1, 2, \cdots$. Associated with j-th element of the N-th finite population is a vector of characteristics, denoted by $\mathbf{y}_{jN}$. Let $\mathcal{F}_N = \{\mathbf{y}_{1N}, \cdots, \mathbf{y}_{NN}\}$ be the set of vectors for the N-th finite population. The population mean of y for the N-th finite population is $\bar{y}_N = N^{-1}\sum_{i=1}^{N} y_{iN}$. Let A_N denote the set of indices appearing in the sample selected from the N-th finite population. The sample size is denoted by n_N. We assume that samples are selected according to the probability rule $P_N(\cdot)$. Under the specified sequence of populations, samples, and sampling designs, we define a sequence of estimators $\hat{\theta}_N$ of the population mean $\bar{y}_N$ to be design consistent, if for all $\epsilon > 0$, $\lim_{N\to\infty} P\left\{|\hat{\theta}_N - \bar{y}_N| > \epsilon \;\middle|\; \mathcal{F}_N\right\} = 0$, where the notation indicates that N-th finite population is held fixed and the probability depends only on the sampling design.

The optimal diagonal matrix $\mathbf{\Gamma}$ can be estimated using the design consistent estimators of β_j and $V\left(\hat{\beta}_j\right)$. The estimated optimal matrix is denoted by $\widehat{\mathbf{\Gamma}} = \text{diag}\left(\hat{\gamma}_{jj,opt}\right)$, where

$$\hat{\gamma}_{jj,opt} = \left[\hat{\beta}_j^2 + \widehat{V}\left(\hat{\beta}_j\right)\right]^{-1} \hat{\beta}_j^2, \tag{4.1}$$

and $\widehat{V}\left(\hat{\beta}_j\right)$ is a design consistent estimator of the variance of $\hat{\beta}_j$. Design consistent variance estimator of the regression coefficient estimator is given in Fuller (1975). See also Särndal, Swensson and Wretman (1991). The estimator defined in (3.9) with $\hat{\gamma}_{jj,opt}$, $j = p_1 + 1, \cdots, p$, is asymptotically equivalent to the regression estimator and is design consistent.

Theorem. Let $\{\mathcal{F}_N, A_N\}$ be a sequence of populations and samples such that

$$(\bar{y}_\pi \ , \ \bar{\mathbf{x}}_\pi) - (\bar{y}_N \ , \ \bar{\mathbf{x}}_N) = O_p\left(n_N^{-\frac{1}{2}}\right) \ , \tag{4.2}$$

where $(\bar{y}_\pi \ , \ \bar{\mathbf{x}}_\pi) = \sum_{i \in A_N} \alpha_i (y_i \ , \ \mathbf{x}_i)$, n_N is the sample size for the N-th sample, $\boldsymbol{\alpha}' = (\alpha_1, \cdots, \alpha_{n_N})$, $\alpha_i = \left(\sum_{j \in A_N} \pi_j^{-1}\right)^{-1} \pi_i^{-1}$, π_i is the probability that element i is selected for the sample, and $(\bar{y}_N, \bar{\mathbf{x}}_N)$ is the population mean of $(y, \mathbf{x})$. Assume there exists a sequence $\mathbf{M}_{z\phi z,N}$ and a sequence $q_{j,N}$ such that

$$\mathbf{M}_{z\phi z,n} - \mathbf{M}_{z\phi z,N} = O_p\left(n_N^{-\frac{1}{2}}\right) \ , \ \lim_{N\to\infty} \mathbf{M}_{z\phi z,N} = \mathbf{M}_{z\phi z} \ , \tag{4.3}$$

$$n_N \widehat{V}\left(\hat{\beta}_j\right) - q_{j,N} = O_p\left(n_N^{-\frac{1}{2}}\right) \ , \ \lim_{N\to\infty} q_{j,N} \ = q_j \tag{4.4}$$

for $j = 1, \cdots, p$, where $\mathbf{M}_{z\phi z,n} = n_N^{-1} \mathbf{Z}^{*\prime} \left(\mathbf{I} - \mathbf{P}_{J_n^*}\right) \mathbf{Z}^*$, $\mathbf{Z}^* = (\mathbf{y}^*, \mathbf{X}^*)$, $\mathbf{M}_{z\phi z}$ is a positive definite matrix. Then, the ridge regression estimator (3.9) with $\hat{\gamma}_{jj}$ of (4.1) satisfies

$$\bar{y}_{rreg} = \bar{y}_{reg} + O_p\left(n_N^{-1}\right) \tag{4.5}$$

$$= \bar{y}_N + O_p\left(n_N^{-1/2}\right) \tag{4.6}$$

where $\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}$.

Proof. By adding and subtracting $(\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}}$, the estimator defined in (3.9) with $\hat{\boldsymbol{\Gamma}}$ can be written as

$$\bar{y}_{rreg} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\hat{\boldsymbol{\beta}} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\left(\hat{\boldsymbol{\Gamma}} - \mathbf{I}\right)\hat{\boldsymbol{\beta}} \ . \tag{4.7}$$

For $j = p_1 + 1, \cdots, p$,

$$[\hat{\beta}_j^2 + \widehat{V}(\hat{\beta}_j)]^{-1}\hat{\beta}_j^2 = 1 + O_p\left(n_N^{-\frac{1}{2}}\right) \tag{4.8}$$

because $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_p\left(n_N^{-\frac{1}{2}}\right)$ by (4.3), and $n_N^{-1}[n_N \widehat{V}(\hat{\beta}_j)] \to 0$ by (4.4), where $\boldsymbol{\beta}_N = \mathbf{M}_{x\phi x,N}^{-1}\mathbf{M}_{x\phi y,N}$. Result (4.5), then, follows by (4.7), (4.8) and (4.2). By (4.3), $\bar{y}_{reg} = \bar{y}_\pi + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_\pi)\boldsymbol{\beta}_N + O_p\left(n_N^{-1}\right)$ and thus result (4.6) is immediate by the assumption (4.2). ■

It is well known that regression estimator is superior to $\bar{y}_\pi$ if the multiple regression coefficient between y and $\mathbf{x}$ is greater than p/n_N, where p is the dimension of $\mathbf{x}$ and n_N is a sample size. See Park (2002). Thus, by Theorem 1, we conclude that ridge regression estimator is always almost superior to $\bar{y}_\pi$ because it is asymptotically equivalent to regression estimator. Theorem 1 also provides robust property of the ridge regression estimator to model failure in large sample framework. That is, ridge regression estimator approaches to the true value of the population mean of variable of interest as sample size increase even when the assumed superpopulation model is not true.

5 Simulation Study

We conduct a simulation study to compare the alternative methods of constructing regression weights. A population with six post strata was considered. 30,000 simple random samples of size 60 were selected. The regression superpopulation model (2.1) was used to generate the study variable Y in which $\mathbf{x}_i = (I_{i1}, I_{i2}, I_{i3}, I_{i4}, I_{i5}, x_i)$, $I_{ih} = 1$ if i-th element is in stratum h, and $I_{ih} = 0$ elsewhere, for $h = 1, \cdots, 5$, x_i is the value generated from the χ^2 distribution with two degrees of freedom, $\boldsymbol{\beta} = (0,0,0,0,0,1)'$ and e_i has a normal distribution with mean zero and variance one. All six auxiliary variables are considered as ones for which the constraints would be relaxed. The parameters being estimated are those of the infinite generating mechanism. Along with the variable Y generated from the assumed model, the estimated percentiles of the distribution function of x were also considered to investigate the small sample properties of the procedures when linear regression superpopulation model is not appropriate. Five estimation procedures were considered: 1. Ordinary least square regression (OLS-Reg); 2. Ridge regression (Ridge-Reg); 3. Quadratic programming for regression weights (QP-Reg); 4. Quadratic programming for ridge regression weights (QP-Ridge); 5. Raking regression (Raking-Reg).

The weights for the ordinary least squares regression estimator are

$$w_i = n^{-1} + (\bar{\mathbf{x}}_N - \bar{\mathbf{x}}_n)\left[\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_n)'(\mathbf{x}_i - \bar{\mathbf{x}}_n)\right]^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_n)'.$$

Ridge regression weights were calculated by (3.9) with the estimated $\hat{\gamma}_{jj,opt}$ of (4.1). In defining the optimal γ, we used the regression coefficient estimators and estimated model variances of regression coefficient estimators for the study variable Y. That is, a single vector of ridge regression weights was calculated and used to estimate all parameters of interest. In large scale survey, where we have large number of variables of interest, a single set of weights is usually used to construct estimators for all variables of interest. Quadratic programming was used to derive the nonnegative regression and ridge regression weights. In quadratic programming for regression estimator, weights that minimize (1.1) subject to the linear constraint (1.2) and range restriction $w_i \geq 0$ were derived. In quadratic programming for ridge regression estimator, weights that minimize (3.6) subject to the constraint $\sum_i^n w_i = 1$ and range restriction $w_i \geq 0$ were derived. The weights for raking regression were derived by minimizing $\sum_{i=1}^n w_i \log(nw_i) - w_i + n^{-1}$, subject to the constraint (1.2).

Table 1: Monte Carlo MSE of the estimators of the mean of Y.

	OLS-Reg	Ridge-Reg	QP-Reg	QP-Ridge	Raking-Reg
MSE $\times 10^2$	1.864	1.775	1.866	1.775	1.870

Table 1 shows the MSE of the estimators of the population mean of the variable Y. Ridge regression procedures have the smallest MSE because the estimated MSE of a linear estimator is the objective function minimized by these procedures. The estimated biases of OLS regression estimator and ridge regression estimator are -0.72×10^{-4} and -2.9×10^{-4} respectively. Although ridge regression gives larger bias but both estimated biases are not significantly different from zero.

Table 2: Monte Carlo Mean and Variance of the minimum and maximum weight.

	Minimum Weight		Maximum Weight	
Procedure	Mean ($\times 10^2$)	Variance ($\times 10^5$)	Mean($\times 10^2$)	Variance($\times 10^5$)
OLS-Reg	0.84	0.14	3.19	1.09
Ridge-Reg	1.08	0.13	2.53	0.42
QP-Reg	0.85	0.11	3.19	1.09
QP-Ridge	1.09	0.12	2.53	0.42
Raking-Reg	0.96	0.05	3.40	1.44

Table 2 shows the properties for the minimum and maximum of the weights. Among 30,000 samples, 1,066 samples have at least one negative OLS regression weight. At least one ridge regression weight is negative in 382 samples. Ridge regression has the largest average minimum weight and OLS regression has the smallest average minimum weight. Ridge regression has the smallest maximum average weight with the smallest variance. Raking regression has the largest maximum average weight with the largest variance. Raking regression has the largest range of weights.

Table 3: Monte Carlo Standardized Bias in Percentile Estimators.

	Percentile								
Procedure	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
OLS-Reg	-4.15	-4.01	-3.56	-2.69	-1.58	-1.16	1.45	5.54	19.25
Ridge-Reg	-3.90	-4.02	-3.60	-2.64	-1.55	-1.13	1.45	5.55	19.25
QP-Reg	-4.05	-3.94	-3.50	-2.64	-1.55	-1.11	1.48	5.48	18.54
QP-Ridge	-3.87	-4.00	-3.58	-2.62	-1.54	-1.10	1.48	5.55	18.96
Raking-Reg	-2.77	-2.79	-2.41	-1.78	-1.10	-0.93	0.63	3.76	15.21

Table 3 contains the Monte Carlo bias of the estimators for the percentiles of the distribution of x, where the table entries are $\{\min\left[p\ ,\ (1-p)\right]\}^{-1}\left[\hat{E}\left(\hat{p}\right)-p\right]\times 100$, and p is the true percentiles. Thus, the Monte Carlo estimated relative bias of the regression estimator of the 0.01 percentile is -4.15$\%$. The OLS regression estimator has the largest biases for $p = 0.01,\ 0.25,\ 0.50,\ 0.75$ and 0.99, and ridge regression has the largest biases for $p = 0.05,\ 0.10$ and 0.95. Except $p = 0.01$, the biases of OLS regression and ridge regression procedures do not differ significantly. Raking ratio has the smallest bias for all percentiles.

Table 4 gives the Monte Carlo relative MSE of the estimators where the table entries are $\{\min\left[p\ ,\ (1-p)\right]\}^{-2}\left[\hat{E}\left(\hat{p}\right)-p\right]^{2}\times 100$. Ridge regression estimator has the smallest MSE in all percentiles. For extreme percentiles, raking regression has the largest MSE. In the middle percentiles, OLS regression has the largest MSE. Note that the raking regression has a significantly smaller bias for the extreme percentiles.

Table 5 gives the Monte Carlo MSE for the 1,066 samples with negative OLS-regression weights. Quadratic programming for the ridge regression is superior to other nonnegative weight procedures for all percentiles except $p = 0.50,\ 0.99$. Especially, the efficiency of quadratic programming with ridge regression relative to other procedures is outstanding for lower percentiles. For the middle and large percentiles, the performances of raking regression and quadratic programming with ridge regression are comparable.

Table 4: Monte Carlo Relative MSE of Percentile Estimators.

	Percentile								
Procedure	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
OLS-Reg	171.78	32.34	14.75	4.34	1.06	2.06	7.39	21.07	138.44
Ridge-Reg	159.42	29.86	13.60	4.03	0.99	1.91	6.90	19.84	127.54
QP-Reg	172.86	32.67	14.87	4.35	1.06	2.05	7.39	21.04	137.54
QP-Ridge	159.48	29.87	13.61	4.03	0.99	1.90	6.90	19.82	129.05
Raking-Reg	175.45	32.88	14.81	4.24	1.02	2.05	7.27	20.92	138.95

Table 5: Monte Carlo Relative MSE of Percentile Estimators for Samples with at least One Negative OLS-Reg weight.

	Percentile								
Procedure	0.01	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.99
OLS-Reg	178.62	38.06	17.71	5.74	1.67	2.40	9.83	28.22	160.08
Ridge-Reg	172.12	35.90	16.52	5.29	1.50	1.94	8.62	23.06	104.36
QP-Reg	209.02	47.42	21.11	5.91	1.60	2.14	9.88	27.34	134.92
QP-Ridge	173.78	36.22	16.61	5.25	1.46	1.83	8.62	22.65	90.98
Raking-Reg	222.72	48.59	20.60	5.32	1.37	2.00	8.80	22.81	87.86

6 Summary and discussion

We considered a situation in which many variables are of interest and thus a single regression superpopulation model may fail to explain the relationships between study variables and auxiliary variables. For such a situation, construction of regression type estimator, that is approximately efficient for the main variables of interest and is also robust to model misspecification, is considered.

Under the model in which many auxiliary variables are used, ridge regression estimator is derived by relaxing the linear constraints.

Through a simulation study, ridge regression weight is compared to alternative regression type weights. The ridge regression weights have the largest minimum weight and the smallest maximum weight. The OLS-regression has the smallest minimum weights and raking regression has the largest maximum weight. In estimating the population percentiles of a skewed distribution, raking ratio has smaller bias than other procedures. The ridge regression estimator with the optimal coefficient matrix derived under the assumed regression superpopulation model has the smallest MSE for all percentiles. By relaxing the linear constraints on weights with an appropriate coefficient matrix, the ridge regression estimator not only generates the optimal estimator for the population parameter for the assumed model but it also shows better performance in estimating the population percentiles in which the assumed model is not appropriate. Based on the results of the simulation study, the use of the ridge regression estimator with the optimal coefficient matrix for the important variables could be recommended for large scale surveys in which a large number of variables are considered.

7 References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis,* 2nd ed. New York: Wiley.

Bardsley, P. and Chambers, R. L. (1984). Multipurpose estimation from unbalanced samples. Applied Statistics, 33: 290–299.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Annals of Mathematical Statistics, 11: 427–444.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87: 376–382.

Fuller, W. A. (1975). Regression analysis for sample survey. Sankhya Series C, 37: 117–132.

Fuller, W. A. (2002). Regression estimation for survey samples. Survey Methodology, 28: 5–23.

Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12: 55–67.

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation for sample survey data. Proceedings of the Social Statistical Section, American Statistical Association. 300–305.

Husain, M. (1969). Construction of regression weights for estimation in sample surveys. Unpublished M.S. thesis, Iowa State University, Ames, Iowa.

Park, M. (2002). Regression estimation of the mean in survey sampling. Unpublished Ph.D. dissertation, Iowa State University, Ames, Iowa.

Park, M., and Fuller, W.A. (2005). Towards nonnegative regression weights for survey samples. Survey Methodology, 31: 85–93.

Rao, J. N. K. and Singh, A. C. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. Proceedings of the section on survey research methods, American Statistical Association. 57–64.

Särndal, C.E., Swensson, B. and Wretman, J. (1991). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Singh, A.C., and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. Survey Methodology, 22: 107–115.