Variance Estimation Under Two–Phase Sampling

by **Q. T. Thach** and **N. G. N. Prasad**

Q. T. Thach
Department of Community Medicine
The University of Hong Kong
Patrick Manson Building South Wing
7 Sassoon Road
Hong Kong

N. G. N. Prasad
Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta, T6G–2G1
Canada

This paper considers a new variance estimator for ratio estimation based on the empirical likelihood approach under simple random sampling without replacement both in single and two–phase sampling. We will use this approach to choose the probability weights under constraints formulated from the information on the auxiliary variable.

1 Introduction

In survey sampling, we are often interested in estimating the mean value of a characteristic Y of a particular population when the information on the auxiliary variable X, correlated with Y, is already available or can be easily observed. For such situations, the estimation for the mean value of Y through ratio and regression techniques has been discussed in the literature for two different cases (see Cochran, 1977) (1) Single–phase case: When the population mean of the characteristics X is already known and the information on Y is observed for units in a sample of size n; (2) Two–phase case: When the population mean of the characteristics X is not known and it is estimated by taking a large random sample of size n' and observing X, then drawing a subsample of size n observing Y.

Two–phase sampling is generally employed when it is economically feasible to take a large preliminary sample in which an auxiliary variable X, correlated with a characteristic of interest Y, is measured alone. The initial sample gives an estimate $\bar{x}_{n'}$ of the population mean $\bar{X}$, while the subsample in which Y is measured is employed to estimate the population mean $\bar{Y}$ through ratio or regression estimation using $\bar{x}_{n'}$. For example, in a survey that estimates the total wheat yield in a given locality in Canada, one might use a large sample of n' farms to estimate the total area under wheat cultivation and a subsample of n farms to determine the actual yield.

Chen and Qin (1993) employed the empirical likelihood method to use summary information on the auxiliary variable at the estimation stage. Benhin and Prasad (1997) extended the empirical likelihood to double sampling when two auxiliary variables were available.

Turning to variance estimation under the ratio method, Rao and Sitter (1995) proposed a new linearization variance estimator for a ratio estimator that made better use of the sample data than the standard textbook formula. They also obtained a jackknife variance estimator and concluded through a simulation study that their conditional and unconditional variances had better properties than the standard formula (see Sukhatme and Sukhatme,1970). Subsequently, Sitter (1997) extended this method to regression estimation along the same lines as ratio estimation. He showed under a model proposed by Dorfman (1994) that the resulting variance estimators were design–unbiased and approximately model–unbiased. For more information on variance estimation under two–phase sampling under model-based approach, see Dorfman (1994).

This paper considers a new variance estimator for ratio estimation based on the empirical likelihood approach under simple random sampling without replacement both in single and two–phase sampling. We will use this approach to choose the probability weights under constraints formulated from the information on the auxil-

iary variable.

In Section 2 and Section 3, we review variance estimators available for ratio estimation in single and double sampling. The proposed variance estimator is derived under each case using empirical likelihood. We extend the empirical likelihood method to regression estimation in Section 4. A simulation study to examine the unconditional and conditional repeated sampling properties of the proposed variance estimator in two phase sampling is presented in Section 5.

2 Variance estimator of the ratio estimator under single–phase sampling

Suppose that a population consists of N distinct units with values (y_i, x_i), where $x_i > 0$ $(i = 1, \ldots, N)$. Denote the population means of Y and X, respectively, by $\bar{Y}$ and $\bar{X}$. To estimate $\bar{Y}$ under simple random sampling of size n, it is customary to use the ratio estimator $\bar{y}_{rs} = (\bar{y}/\bar{x})\bar{X}$, where $\bar{y}$ and $\bar{x}$ are the sample means of y and x. The variance of $\bar{y}_{rs}$ is approximated by (Cochran, 1977, p. 155) and given by

$$V(\bar{y}_{rs}) \doteq \left(\frac{1}{n} - \frac{1}{N}\right) S_D^2, \tag{2.1}$$

where $S_D^2 = (N-1)^{-1} \sum_{i=1}^{N} D_i^2$, $D_i = Y_i - RX_i$ with $R = \bar{Y}/\bar{X}$. Two commonly used estimators of $V(\bar{y}_{rs})$ are

$$v_0(\bar{y}_{rs}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2 \quad \text{and} \quad v_1(\bar{y}_{rs}) = \left(\frac{1}{n} - \frac{1}{N}\right) \left(\frac{\bar{X}}{\bar{x}}\right)^2 s_d^2, \tag{2.2}$$

where

$$s_d^2 = \frac{1}{n-1} \sum_{i \in s} d_i^2 = s_y^2 - 2rs_{xy} + r^2 s_x^2, \tag{2.3}$$

with $r = \bar{y}/\bar{x}$, $d_i = y_i - rx_i$. Although the original motivation for $v_1(r) = v_1(\bar{y}_{rs})/\bar{X}^2$ as a variance estimator of the ratio R is the unavailability of $\bar{X}$, it is not clear that $v_1(\bar{y}_{rs})$ is indeed worse than $v_0(\bar{y}_{rs})$ (see Cochran, 1977 and Rao and Rao, 1971). Chen and Qin (1993) applied the empirical likelihood approach in conjunction with summary information on the auxiliary variable in improving the customary estimator under simple random sampling. They showed that the empirical likelihood estimator was asymptotically equivalent to the linear regression estimator when the population mean of the auxiliary variable was known (see Hartley and Rao, 1968). To use the

empirical likelihood method as described in the previous chapter, we maximize the empirical likelihood

$$L(F) = \prod_{i=1}^{n} p_i, \tag{2.4}$$

where $p_i = P(Y = y_i)$. The p_i's are subject to $0 \leq p_i \leq 1$ and $\sum_{i=1}^{n} p_i = 1$. With these weights an empirical likelihood estimator for S_D^2 is given by

$$s_d^2(el) = \sum_{i=1}^{n} p_i d_i^2. \tag{2.5}$$

The resulting empirical likelihood variance estimator for $\bar{y}_{rs}$ is

$$v_3(\bar{y}_{rs}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2(el). \tag{2.6}$$

3 The two–phase sampling procedure

Assume that a simple random sample s' of size n' is selected without replacement from a population of N units and x_i is observed for $i \in s'$. A simple random subsample s of size n is then selected without replacement from s' and y_i is observed for $i \in s$. Several estimates of $\bar{Y} = \sum_{i=1}^{N} Y_i/N$ can be formed. The simplest is the usual biased ratio estimate with the population mean $\bar{X}$ replaced by its estimates $\bar{x}_{n'}$, given by $\bar{y}_{rt} = (\bar{y}_n/\bar{x}_n)\bar{x}_{n'} = r\bar{x}_{n'}$, where $\bar{y}_n$ and $\bar{x}_n$ are the means for s and $\bar{x}_{n'}$ for s'.

3.1 The ratio estimator and some preliminary results

The estimator $\bar{y}_{rt}$ is design-consistent for $\bar{Y}$, i.e., $p\lim_{n\to\infty}{}_{\pi}(\bar{y}_{rt} - \bar{Y}) = 0$, where π denotes the probability space generated by the sampling scheme. The variance of $\bar{y}_{rt}$ is approximated by a standard formula and is given by

$$V(\bar{y}_{rt}) \doteq \left(\frac{1}{n} - \frac{1}{n'}\right) S_D^2 + \left(\frac{1}{n'} - \frac{1}{N}\right) S_y^2, \tag{3.1}$$

where

$$S_D^2 = \frac{1}{N-1}\sum_{i=1}^{N} D_i^2 = S_y^2 - 2RS_{xy} + R^2 S_x^2, \quad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2, \quad S_{xy}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y}),$$

with $D_i = y_i - Rx_i$ and $R = \bar{Y}/\bar{X}$. Note that if $n' = n$, $(\bar{y}_n/\bar{x}_n)\bar{x}_{n'} = \bar{y}_n$ and so (3.1) reduces to $(1/n - 1/N)S_y^2$, which is the variance of $\bar{y}_n$ under simple random sampling in single–phase sampling. It is observed that if $n' = N$, the estimator is the ratio estimator under single–phase sampling, and the variance reduces to the approximate formula for its variance. It follows that the estimate $\bar{y}_{rt}$ based on two–phase sampling is more efficient than the estimate $\bar{y}_n$ based on simple random sampling when no auxiliary variable is used, if

$$R^2 S_x^2 - 2RS_{xy} < 0,$$

i.e., if

$$\rho_{xy}\frac{C_y}{C_x} > \frac{1}{2},$$

where ρ_{xy} is the population correlation coefficient between x and y, and C_x and C_y are population coefficients of variation of x and y, respectively. A design-consistent estimator of the variance estimator of $\bar{y}_{rt}$ is given by

$$v_0(\bar{y}_{rt}) = \left(\frac{1}{n} - \frac{1}{n'}\right) s_d^2 + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2, \tag{3.2}$$

where

$$s_d^2 = \frac{1}{n-1}\sum_{i\in s} d_i^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_n)^2, \tag{3.3}$$

with $d_i = y_i - rx_i$. The second term in $v_0(\bar{y}_{rt})$ is obtained by using the sample variance s_y^2 to estimate the population variance S_y^2.

3.2 Linearization variance estimator

Rao and Sitter (1995) proposed a linearization variance estimator of $\bar{y}_{rt}$ that made better use of the sample data than the standard one, v_0. They first expressed S_y^2 as

$$\begin{aligned}
S_y^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - R\bar{X})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - Rx_i + Rx_i - R\bar{X})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left\{(y_i - Rx_i)^2 + 2R(y_i - Rx_i)(x_i - \bar{X}) + R^2 S_x^2\right\} \\
&= S_D^2 + 2RS_{Dx} + R^2 S_x^2,
\end{aligned} \tag{3.4}$$

where S_D^2 and S_x^2 were the corresponding population variances of D_i and x_i, S_{Dx} was the population covariance between D_i and x_i. Then the sample variance of s_y^2 can be written as

$$s_y^2 = s_d^2 + 2rs_{dx} + r^2 s_x^2, \tag{3.5}$$

where s_d^2, s_{dx} and s_x^2 are the sample analogues of S_D^2, S_{Dx} and S_x^2 based on subsample s. It follows from (3.4) and (3.5) that an alternative estimator of S_y^2 that makes more use of the sample data is obtained by using

$$s_x'^2 = \frac{1}{n'-1}\sum_{i\in s'}(x_i - \bar{x}_{n'})^2$$

in place of s_x^2. The linearization variance estimator of $\bar{y}_{rt}$ is

$$v_1(\bar{y}_{rt}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{1}{n'} - \frac{1}{N}\right) r s_{dx} + \left(\frac{1}{n'} - \frac{1}{N}\right) r^2 s_x'^2. \tag{3.6}$$

This variance estimator is also design-consistent. We can rewrite (3.1) using (3.5) as

$$v_0(\bar{y}_{rt}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{1}{n'} - \frac{1}{N}\right) r s_{dx} + \left(\frac{1}{n'} - \frac{1}{N}\right) r^2 s_x^2. \tag{3.7}$$

3.3 Jackknife variance estimator

Another approach to variance estimation is to use a jackknife technique. Rao and Sitter (1995) proposed a jackknife method which entailed recalculating $\bar{y}_{rt}$ with the jth element removed for each $j \in s'$ and then using the variance of these n' jackknife values, $\bar{y}_{rt}(j)$. Clearly, deleting unit j will affect $\bar{x}_n$ and $\bar{y}_n$ only for $j \in s$ but not for $j \in s' - s$, while it will affect $\bar{x}_{n'}$ for all $j \in s'$. Thus, they defined

$$\bar{y}_{rt}(j) = \{\bar{y}_n(j)/\bar{x}_n(j)\}\bar{x}_{n'}(j)$$

for all $j \in s'$, where

$$\bar{x}_n(j) = \begin{cases} \frac{n\bar{x}_n - x_j}{n-1} & \text{if} \quad j \in s \\ \bar{x}_n & \text{if} \quad j \in s' - s, \end{cases} \qquad \bar{y}_n(j) = \begin{cases} \frac{n\bar{y}_n - y_j}{n-1} & \text{if} \quad j \in s \\ \bar{y}_n & \text{if} \quad j \in s' - s, \end{cases}$$

and $\bar{x}_{n'}(j) = (n'\bar{x}_{n'} - x_j)/(n'-1)$ for all $j \in s'$. Now the usual jackknife method to $\bar{y}_{rt}(j)$ will yield the following variance estimator:

$$v_J(\bar{y}_{rt}) = \frac{n'-1}{n'}\sum_{j\in s'}\{\bar{y}_{rt}(j) - \bar{y}_{rt}\}^2. \tag{3.8}$$

This jackknife estimator ignores the finite population corrections $1 - n/N$ and $1 - n'/N$.

For a nonlinear parameter $\theta = g(\bar{Y})$, a jackknife variance estimator is obtained by replacing $\bar{y}_{rt}(j)$ and $\bar{y}_{rt}$ in (3.8) by $\hat{\theta}_{rt}(j) = g(\bar{Y}_{rt}(j))$ and $\hat{\theta}_{rt} = g(\bar{y}_{rt})$. A linearized version of v_J, for large n, is obtained by noting that

$$\bar{y}_{rt}(j) - \bar{y}_{rt} = \begin{cases} -r\left(\frac{x_j - \bar{x}_{n'}}{n'-1}\right) - \frac{\bar{x}_{n'}(j)}{\bar{x}_n(j)}\left(\frac{y_j - r x_j}{n-1}\right) & \text{if} \quad j \in s \\ -r\left(\frac{x_j - \bar{x}_{n'}}{n'-1}\right) & \text{if} \quad j \in s' - s, \end{cases} \tag{3.9}$$

and assuming $\bar{x}_{n'}(j)/\bar{x}_n(j) \doteq \bar{x}_{n'}/\bar{x}_n$ in (3.9). From (3.8) and (3.9), we get

$$v_J(\bar{y}_{rt}) \doteq \left(\frac{\bar{x}_{n'}}{\bar{x}_n}\right)^2 \frac{s_d^2}{n} + 2\left(\frac{\bar{x}_{n'}}{\bar{x}_n}\right)\frac{r s_{dx}}{n'} + \frac{r^2 s_x'^2}{n'}. \tag{3.10}$$

Ignoring the finite population corrections and comparing (3.9) and (3.10), it now follows that v_J is also design-consistent for $V(\bar{y}_{rt})$ since $\bar{x}_{n'}/\bar{x}_n \doteq 1$ for large n. It also follows from (3.9) and (3.10) that another design-consistent linearization variance estimator, when the finite population corrections are not ignorable, is given by

$$v_2(\bar{y}_{rt}) \doteq \left(\frac{\bar{x}_{n'}}{\bar{x}_n}\right)^2 \left(\frac{1}{n} - \frac{1}{N}\right) s_d^2 + 2\left(\frac{1}{n'} - \frac{1}{N}\right)\left(\frac{\bar{x}_{n'}}{\bar{x}_n}\right) r s_{dx} + \left(\frac{1}{n'} - \frac{1}{N}\right) r^2 s_x'^2. \tag{3.11}$$

Rao and Sitter (1995) noted that if the finite population corrections could be ignored, v_J should perform well conditionally given, $\bar{x}_{n'}/\bar{x}_n$, since it was asymptotically equivalent to v_2.

In the next section, we propose two alternative variance estimators. One of them is a modification of v_0 while the other one is suggested by the empirical likelihood principle. Both utilize the information collected in the first phase as supplementary information in order to improve the precision of variance estimator of population characteristics.

3.4 The empirical likelihood for the double sampling

Since no auxiliary information is available beyond the initial sample s', we maximize the empirical conditional likelihood given by

$$L(s|s') = \prod_{i \in s} p_i \tag{3.12}$$

subject to

$$p_i \geq 0, \quad \sum_{i \in s} p_i = 1 \quad \text{and} \quad \sum_{i \in s} p_i w_i = 0, \tag{3.13}$$

with $w_i = x_i - \bar{x}_{n'}$. Then the empirical likelihood-based estimator for the sample variance s_y^2 and sample covariance s_{xy} are obtained by replacing n^{-1} with the p_i's in the plug-in estimator, i.e.

$$s_y^2(el) = \sum_{i \in s} p_i (y_i - \bar{y})^2, \tag{3.14}$$

$$s_{xy}(el) = \sum_{i \in s} p_i (x_i - \bar{x})(y_i - \bar{y}). \tag{3.15}$$

We use arguments of Rao and Sitter (1995) to obtain the variance estimator of $\bar{y}_{rt}$. First we observe that $\sum_{i=1}^N D_i = 0$ and that S_D^2 is expressed as

$$\begin{aligned} S_D^2 &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - RX_i)^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y} - R(X_i - \bar{X}))^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \{(Y_i - \bar{Y})^2 - 2R(X_i - \bar{X})(Y_i - \bar{Y}) + R^2 (X_i - \bar{X})^2\} \\ &= S_y^2 - 2RS_{xy} + R^2 S_x^2, \end{aligned} \tag{3.16}$$

where S_{xy} is the population covariance between x_i and y_i. Thus $v_0(\bar{y}_{rt})$ in (3.2) can be re-expressed as

$$v_0(\bar{y}_{rt}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 - 2r \left(\frac{1}{n} - \frac{1}{n'}\right) s_{xy} + r^2 \left(\frac{1}{n} - \frac{1}{n'}\right) s_x^2. \tag{3.17}$$

The resulting variance estimator based on empirical likelihood is given as

$$v_4(\bar{y}_{rt}) \doteq \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2(el) - 2r \left(\frac{1}{n} - \frac{1}{n'}\right) s_{xy}(el) + r^2 \left(\frac{1}{n} - \frac{1}{n'}\right) s_x'^2. \tag{3.18}$$

Intuitively, one would expect the variance estimator based on the empirical likelihood to be more efficient than the Rao–Sitter estimator, since it makes use of extra information, i.e., the knowledge of the mean of a subsample of x.

An alternative variance estimator of $V(\bar{y}_{rt})$ can also be obtained. We note that when the y_i's are exactly proportional to x_i's for $i = 1, \ldots, N$, i.e., $y_i = kx_i$, with k as a constant, then the variance $V(\bar{y}_{rt})$ reduces to $(1/n' - 1/N)k^2 S_x^2$, which could be estimated by $k^2(1/n' - 1/N)s_x'^2$. Putting $y_i = kx_i$ in (3.2), we get $v_0(\bar{y}_{rt}) = k^2(1/n' - 1/N)s_x^2$, which is less efficient than $k^2(1/n' - 1/N)s_x'^2$. In view of this, we propose a modified estimator of v_0 given by

$$v_3(\bar{y}_{rt}) = \left(\frac{1}{n} - \frac{1}{n'}\right) s_d^2 + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \frac{s_x'^2}{s_x^2}. \tag{3.19}$$

Note that $v_3(\bar{y}_{rt})$ reduces to $k^2(1/n' - 1/N)s_x'^2$ when $y_i = kx_i$.

4 The regression estimator

In this section, we will consider the extension of the ratio method of estimation in two–phase sampling to the case of linear regression estimation under the empirical likelihood framework. To this end, consider the two–phase sampling scheme described in Section 2. The simple linear regression estimator for two–phase sampling defined by

$$\bar{y}_{lr} = \bar{y}_n + b(\bar{x}_{n'} - \bar{x}_n), \tag{4.1}$$

where $b = s_{xy}/s_x^2$ is the least square regression coefficient of y_i on x_i based on s. This estimator is design consistent for $\bar{Y}$. A design consistent linearization variance estimator of $\bar{y}_{lr}$ is given by the standard formula

$$v_0(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{n'}\right) s_{d'}^2, + \left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 \tag{4.2}$$

where $s_{d'}^2 = (n-1)^{-1} \sum_{i \in s} d_i'^2$ and s_y^2 are the sample variances of $d_i' = y_i - \bar{y} - b(x_i - \bar{x}_n)$ and y_i. Alternatively, $v_0(\bar{y}_{lr})$ can be expressed as

$$v_0(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^2 - 2b \left(\frac{1}{n} - \frac{1}{n'}\right) s_{xy} + b^2 \left(\frac{1}{n} - \frac{1}{n'}\right) s_x^2. \tag{4.3}$$

Sitter (1997) proposed three variance estimators for regression estimation along the same lines as the ratio estimation. The linearization and jackknife variance estimators in this case are given, respectively, by

$$v_1(\bar{y}_{lr}) \quad = \left(\tfrac{1}{n} - \tfrac{1}{N}\right) s_{d'}^2 + \left(\tfrac{1}{n'} - \tfrac{1}{N}\right) b^2 s_x'^2, \tag{4.4}$$

and

$$v_J(\bar{y}_{lr}) \doteq \frac{s^2_{d'}}{n} + \frac{b^2 s'^2_x}{n'} + \left\{\frac{\bar{x}_{n'} - \bar{x}_n}{(n-1)s^2_x}\right\}^2 \sum_{j\in s} \frac{d'^2_j (x_j - \bar{x}_n)^2}{(1-k_j)^2} + R, \tag{4.5}$$

where $k_j = 1/n + (x_j - \bar{x}_n)^2/\{(n-1)s^2_x\}$, and

$$R = \frac{2}{n}\left\{\frac{1}{n}\sum_{i\in s}\frac{d'^2_j a_j}{(1-k_j)} + \frac{b}{n'-1}\sum_{j\in s}\frac{d'_j a_j (x_j - \bar{x}_{n'})}{(1-k_j)}\right\}, \tag{4.6}$$

with $a_j = \{n(x_j - \bar{x}_n)(\bar{x}_{n'} - \bar{x}_n)\}/\{(n-1)s^2_x\}$.

Also, noting that the first two terms on the right hand side of (4.5), and comparing these to (4.4), a linearized version of $v_J(\bar{y}_{lr})$ when the finite population corrections are not ignorable is given by

$$\begin{aligned} v_2(\bar{y}_{lr}) = &\left(\frac{1}{n} - \frac{1}{N}\right) s^2_{d'} + \left(\frac{1}{n} - \frac{1}{N'}\right)\frac{b^2 s'^2_x}{n'} \\ &+ \left\{\frac{\bar{x}_{n'} - \bar{x}_n}{(n-1)s^2_x}\right\}^2 \sum_{j\in s}\frac{d'^2_j (x_j - \bar{x}_n)^2}{(1-k_j)^2} + R. \end{aligned} \tag{4.7}$$

In a similar motivation as in Section 3.4, a variance estimator for $\bar{y}_{lr}$ based on the empirical likelihood approach is

$$v_3(\bar{y}_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2_y(el) - 2b\left(\frac{1}{n} - \frac{1}{n'}\right) s_{xy}(el) + b^2\left(\frac{1}{n} - \frac{1}{n'}\right) s'^2_x, \tag{4.8}$$

where $s^2_y(el)$ and $s_{xy}(el)$ are defined, respectively, by (3.14) and (3.15).

5 A simulation study

We study the finite sample properties of various variance estimators through a simulation study. We adopt the model and parameter settings used by Rao and Sitter (1995). The model we consider is

$$y_i = \beta x_i + x_i^{1/2}\varepsilon_i,$$

where $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$, $x_i \overset{ind}{\sim}$ gamma(a, b) and ε_i and x_i are independent of each other. Thus the mean, the variance and coefficient of variation of x are given by $\mu_x = ab$, $\sigma^2 = ab^2$ and $C_x = \sigma_x/\mu_x = a^{-1/2}$, respectively. Furthermore, the mean and variance of y are $\mu_y = \beta\mu_x$ and $\sigma^2_y = \beta^2\sigma^2_x + \mu_x\sigma^2$, and the correlation between x and y is $\rho = \beta\sigma_x/\sigma_y$.

We confine our simulation study for $n = 20$, $n' = 100$, and $n = 80$, $n' = 400$. We generate $R = 10,000$ independent two–phase random samples according to the above model with $\beta = 1.0$ and $\mu_x = 100$ and σ and σ_x chosen to match specified values of ρ and C_x. Here, we ignore the finite population corrections since the two–phase samples are generated from an infinite population. The Monte Carlo estimator of true mean squared error of $\bar{y}_{rt}$, is computed using

$$MSE(\bar{y}_{rt}) = \frac{1}{R}\sum_{t=1}^{R}(\bar{y}_{rt}^{(t)} - \mu_y)^2, \tag{5.1}$$

where $\bar{y}_{rt}^{(t)}$ denotes the value of $\bar{y}_{rt}$ for the t-th Monte Carlo run. The Monte Carlo estimate of mean squared error of a specified estimator say v is computed using

$$MSE(v) = \frac{1}{R}\sum_{t=1}^{R}(v^{(t)} - MSE(\bar{y}_{rt}))^2. \tag{5.2}$$

Table 1 gives the values of $MSE(v)/MSE(v_0)$ for $v = v_1, \ldots, v_4$ and v_j for different values of ρ and C_x where for convenience, $v_t = v_t(\bar{y}_{rt})$, $t = 0, \ldots, 4$ and $v_J = v_J(\bar{y}_{rt})$. It is clear from Table 1 that v_4 is substantially more efficient than other variance estimators. On the other hand, v_3 is more efficient than v_1, v_2 and v_J only for $C_x = 1.4, 1.0, 0.5, 0.33$ and $\rho = 0.8$ and substantially more efficient than v_0 for all values of ρ and C_x. Note that v_J is more efficient than v_0 only for large $n = 80$ as the factor $\bar{x}_{n'}/\bar{x}_n \doteq \bar{x}_{n'}(j)/\bar{x}_n(j) \doteq 1$ becomes more stable.

We also investigate the conditional properties of each variance estimator along the lines of Rao and Sitter (1995). The 10,000 simulated samples are first ordered on the values of $\bar{x}_{n'}/\bar{x}_n$ and then grouped into 20 successive groups each containing $G = 1,000$ samples. For each group, the simulated conditional mean squared error of $\bar{y}_{rt}$ and conditional mean of $v_t, t = 0, \ldots, 4$ and v_J are calculated, respectively,

$$MSE_c = \frac{1}{G}\sum_{g=1}^{G}\{\bar{y}_{rt}^{(g)} - \mu_y\}^2 \quad \text{and} \quad E_c v_t = \frac{1}{G}\sum_{g=1}^{G} v_t^{(g)}. \tag{5.3}$$

For each of the 20 groups, the values of $E_c v_t$ for $t = 0, \ldots, 4$, Ev_J and MSE_c are plotted against the group averages of $\bar{x}_{n'}/\bar{x}_n$ for 12 selected values of ρ and C_x. Figures 1–12 with $n' = 100$ and $n = 20$ show these results. The case $n' = 400$ and $n = 20$ produce similar plots and therefore were omitted. It is clear from these plots that v_1, v_2, v_3, v_4 and v_J perform well in tracking the conditional MSE when $\bar{x}_{n'}/\bar{x}_n$ is between 0.9 and 1.4 with v_J and v_4 slightly better, i.e. they exhibit a similar pattern to the conditional MSE. However, v_0 is able to track the conditional MSE only when

$\bar{x}_{n'}/\bar{x}_n$ is near 1. This means that with a balanced design, v_0 does not deviate much from the conditional MSE.

It is noticed that v_1, v_2, v_3, v_4 and v_J perform poorly in tracking the conditional MSE when $\bar{x}_{n'}/\bar{x}_n \leq 0.9$ and also when $\bar{x}_{n'}/\bar{x}_n \geq 1.4$. Whereas, v_0 leads to significant overestimation of conditional MSE when $\bar{x}_{n'}/\bar{x}_n \leq 0.9$ and lead to significant underestimation when $\bar{x}_{n'}/\bar{x}_n \geq 1.2$. Thus, all things considered, v_1, v_2, v_3, v_4 and v_J behave more closely to the conditional MSE than do v_0.

The simulation study suggests that the proposed variance estimator v_4 provides more stable standard errors for ratio estimation. It has a competitive conditional performance, having smallest unconditional MSE. The commonly used estimator v_0 fails on both grounds.

Table 1: Mean square error of v_1, v_2, v_3, v_4 relative to v_0.

	$n = 20,\ n' = 100$				$n = 80,\ n' = 400$			
	C_x				C_x			
ρ	1.4	1.0	0.5	0.33	1.4	1.0	0.5	0.33
	$MSE(v_1)/MSE(v_0)$							
0.9	0.51	0.54	0.63	0.67	0.52	0.55	0.63	0.66
0.8	0.73	0.77	0.84	0.88	0.72	0.76	0.84	0.88
0.7	0.82	0.86	0.93	0.94	0.85	0.86	0.92	0.94
	$MSE(v_2)/MSE(v_0)$							
0.9	0.55	0.57	0.64	0.68	0.53	0.55	0.63	0.65
0.8	0.87	0.86	0.86	0.87	0.74	0.78	0.83	0.87
0.7	0.98	0.94	0.98	0.94	0.89	0.88	0.93	0.94
	$MSE(v_J)/MSE(v_0)$							
0.9	0.89	0.77	0.73	0.74	0.61	0.58	0.65	0.67
0.8	1.62	1.24	1.01	1.00	0.85	0.86	0.87	0.89
0.7	1.86	1.35	1.18	1.07	1.03	0.98	0.95	0.96
	$MSE(v_3)/MSE(v_0)$							
0.9	0.52	0.63	0.79	0.90	0.57	0.63	0.81	0.88
0.8	0.65	0.71	0.82	0.85	0.66	0.71	0.84	0.93
0.7	0.67	0.78	0.88	0.92	0.73	0.75	0.90	0.95
	$MSE(v_4)/MSE(v_0)$							
0.9	0.42	0.48	0.59	0.64	0.46	0.49	0.60	0.63
0.8	0.60	0.66	0.75	0.78	0.61	0.66	0.78	0.85
0.7	0.64	0.74	0.85	0.88	0.71	0.74	0.87	0.91

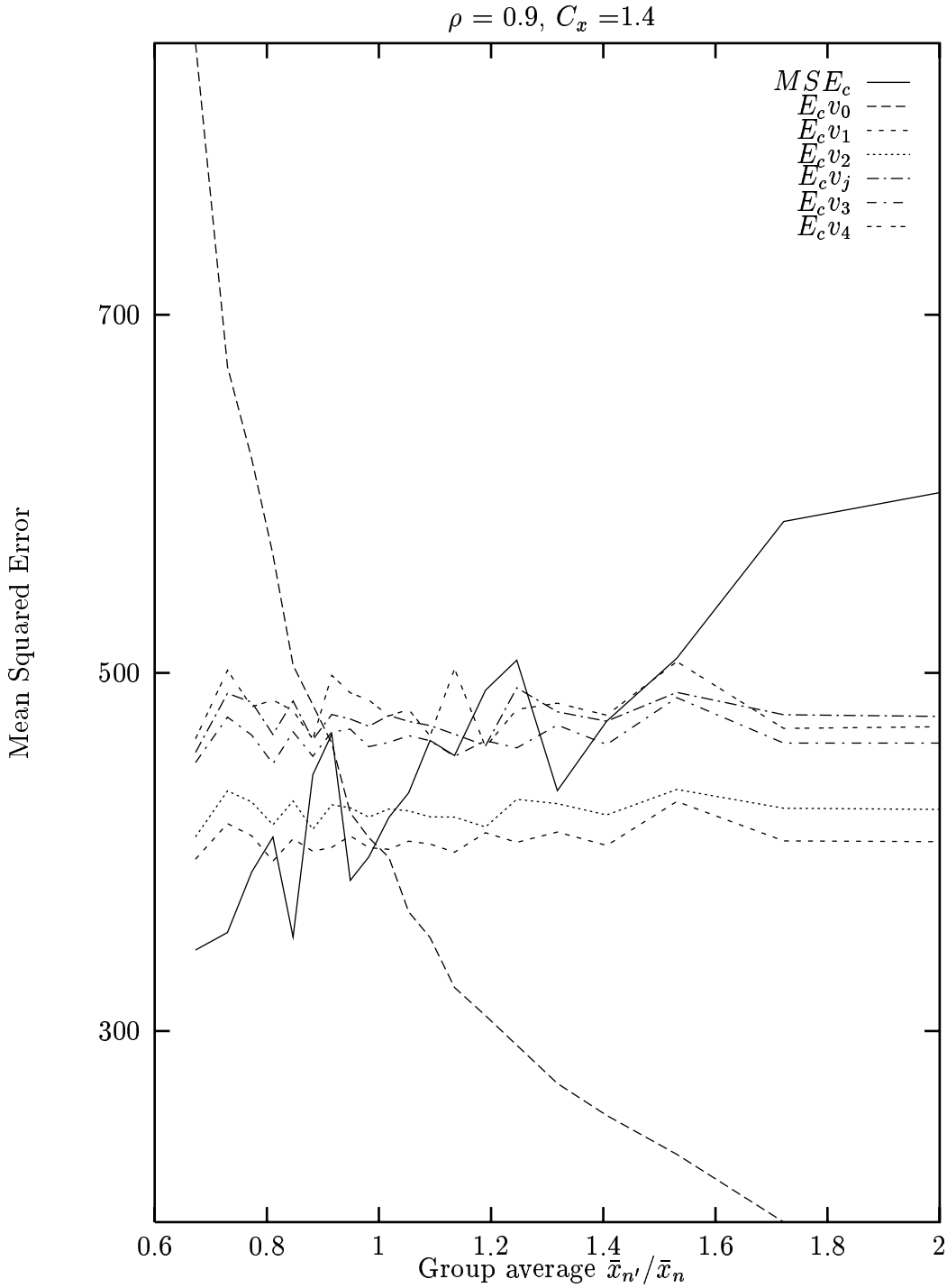

Figure 1: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

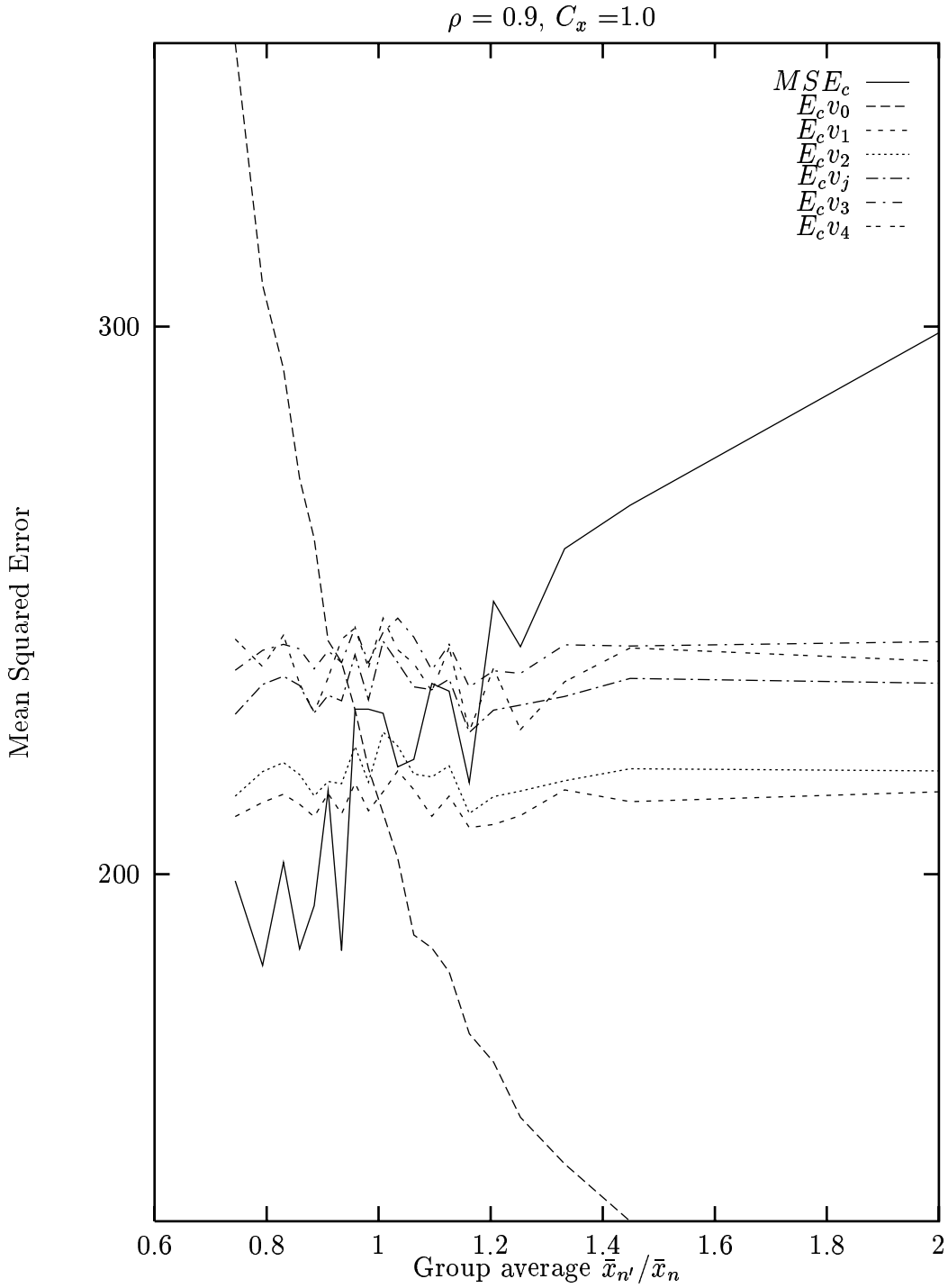

Figure 2: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

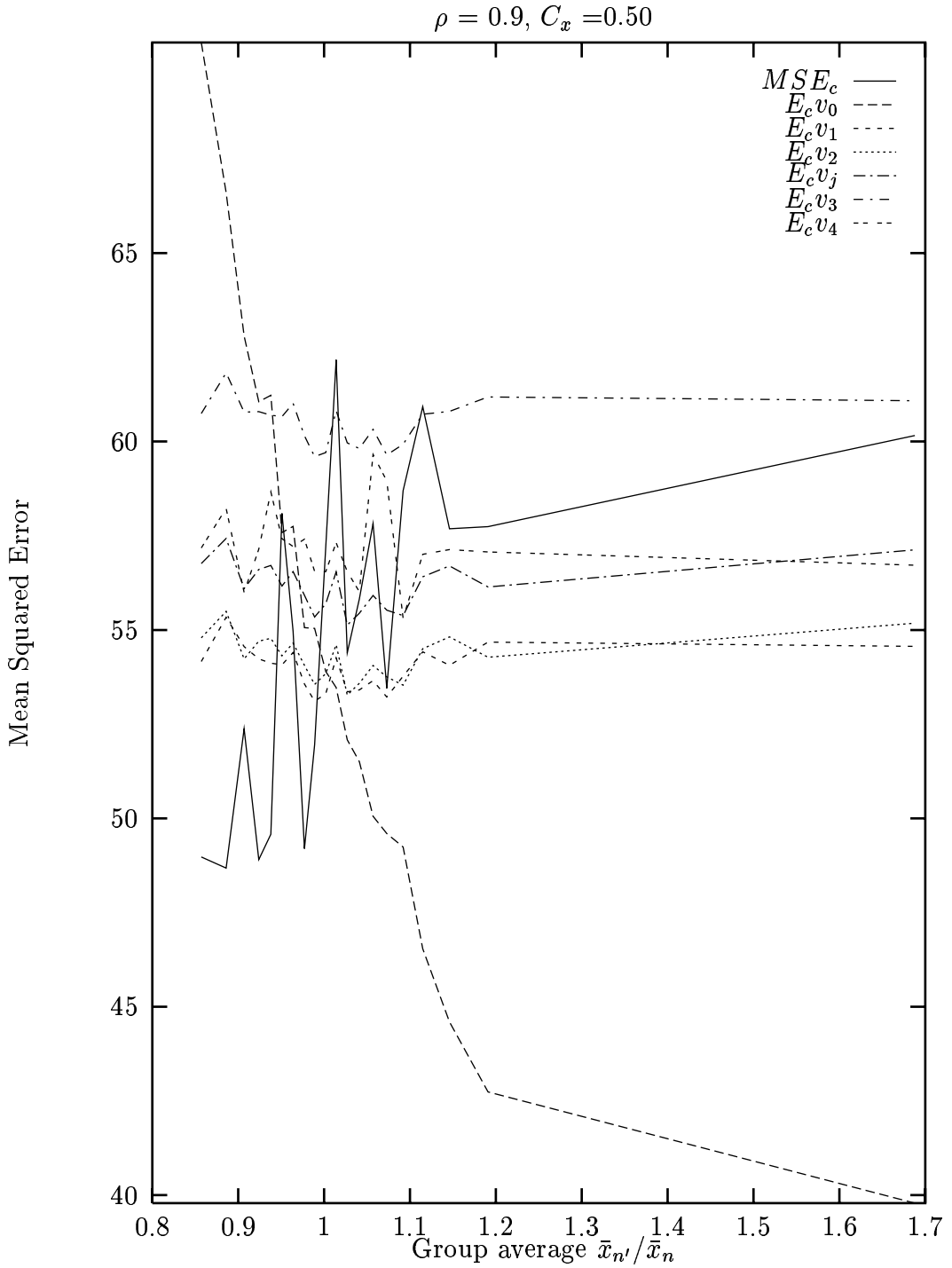

Figure 3: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$, versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

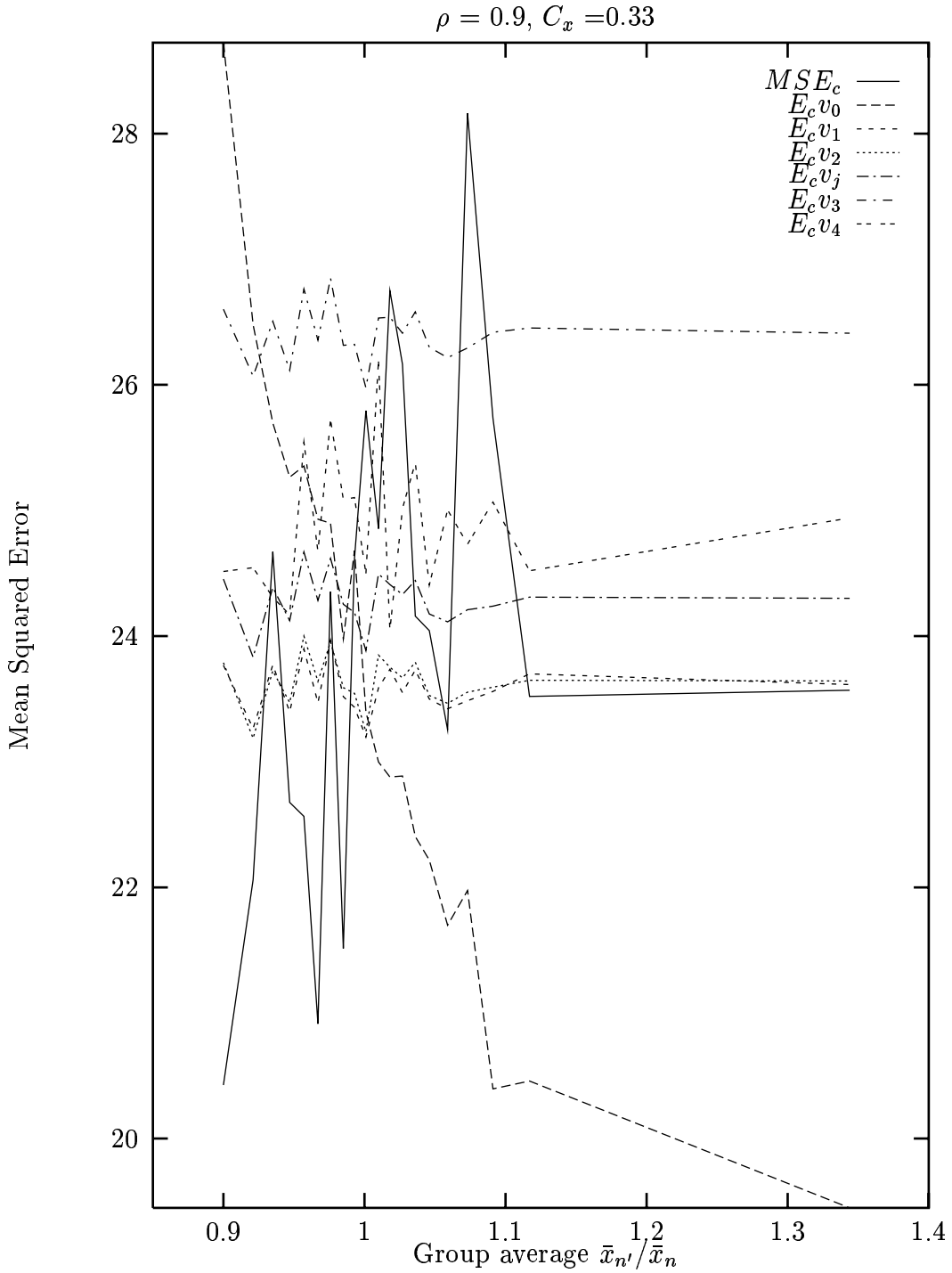

Figure 4: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

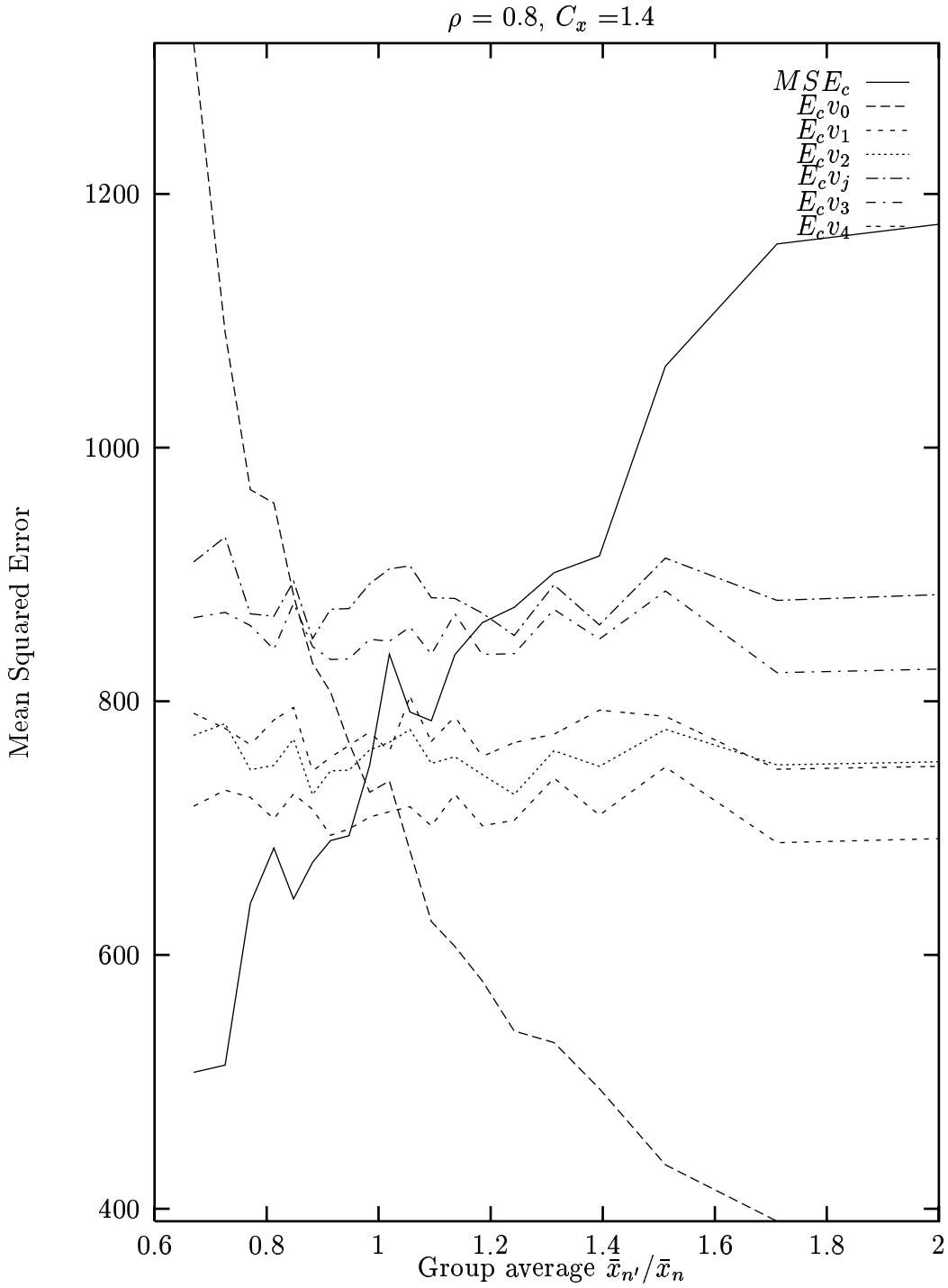

Figure 5: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

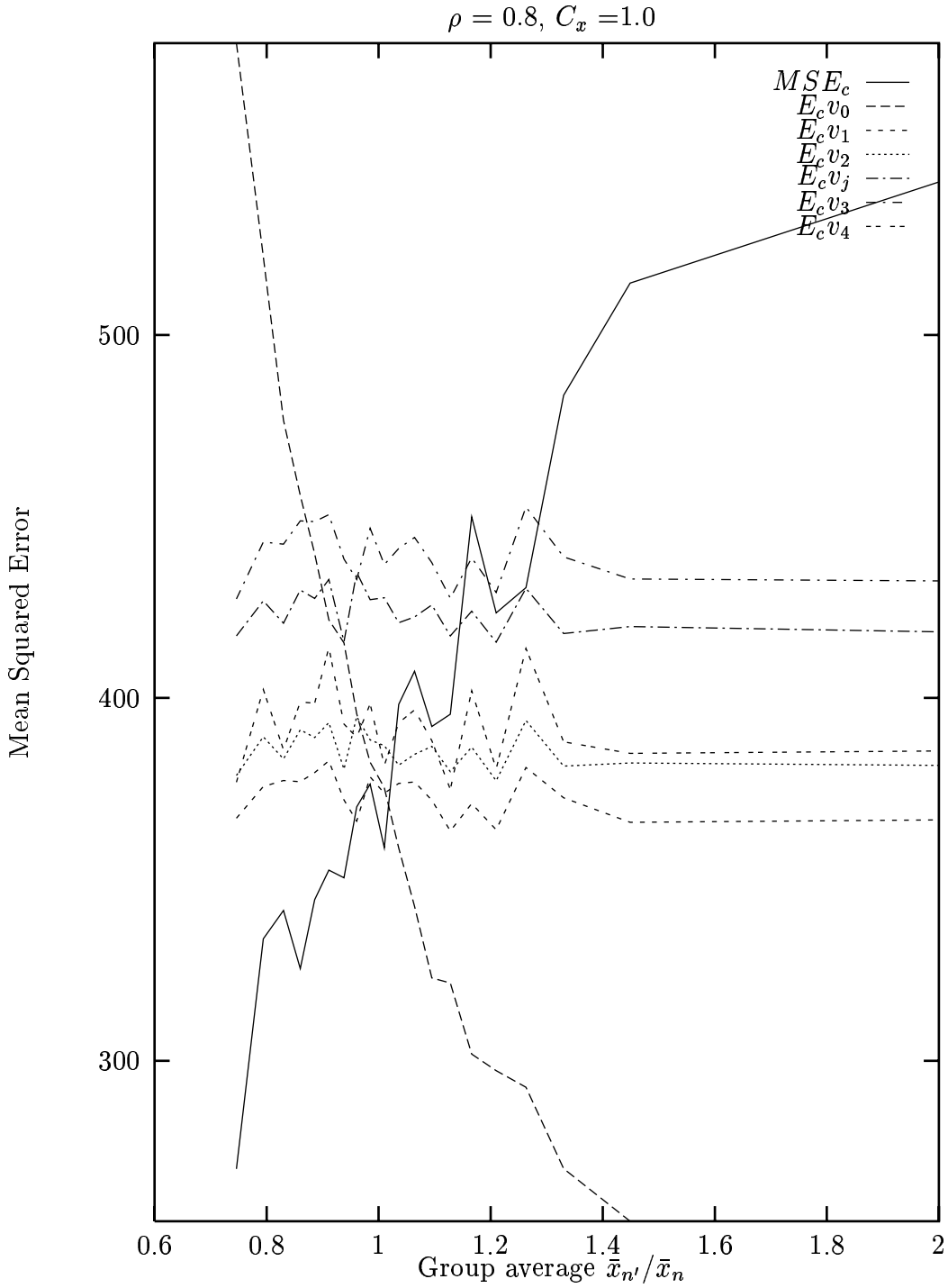

Figure 6: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

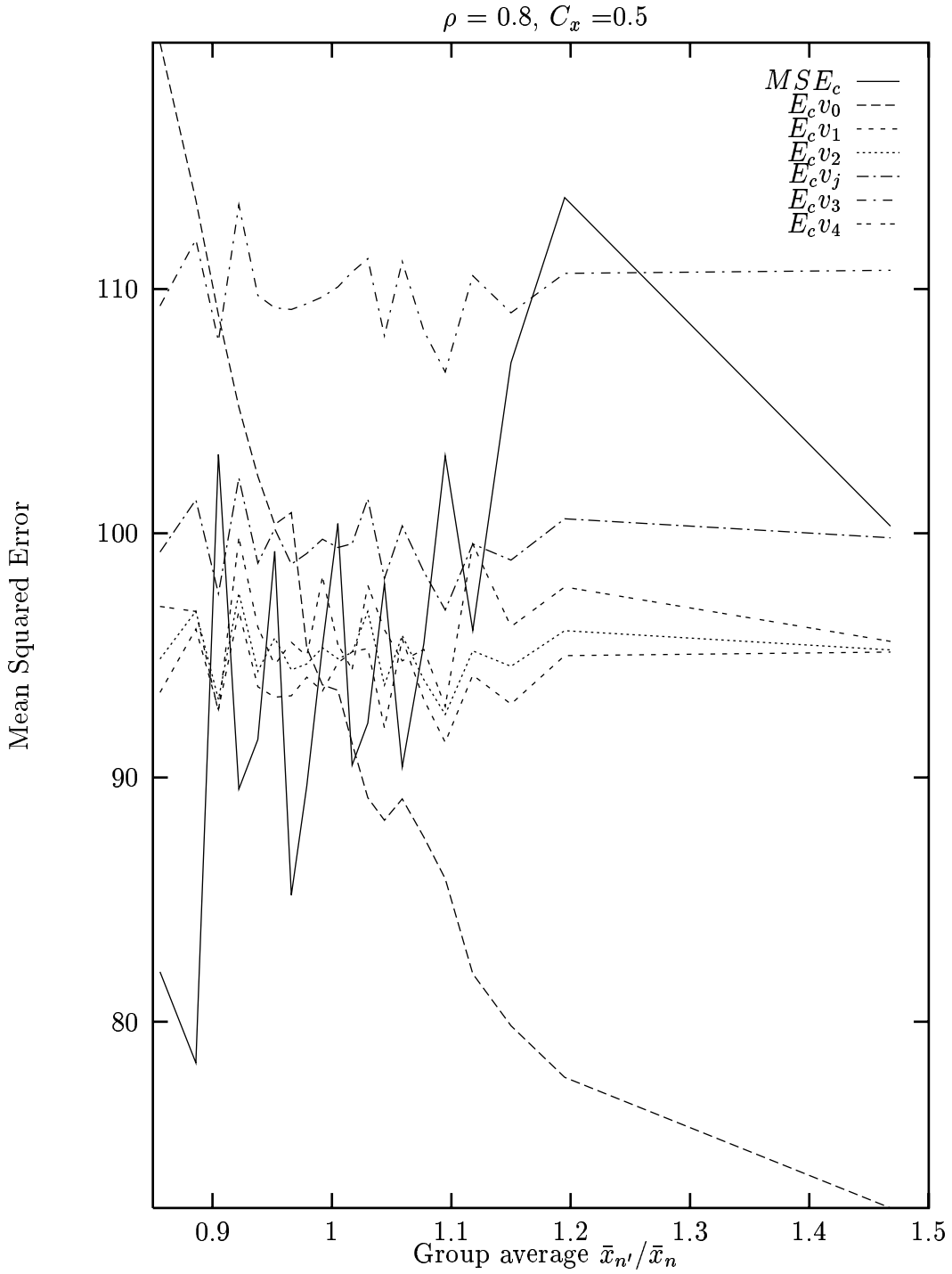

Figure 7: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

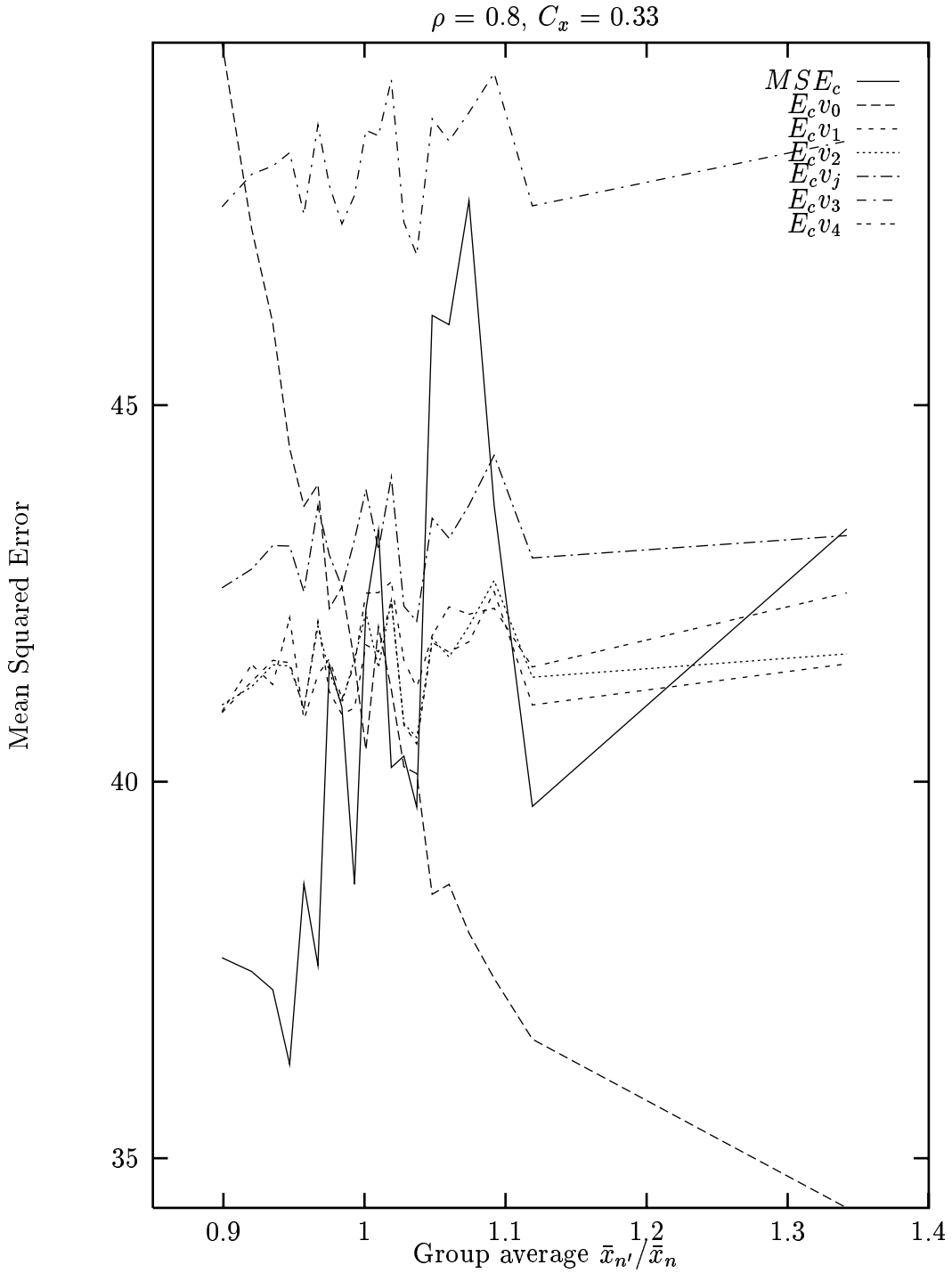

Figure 8: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

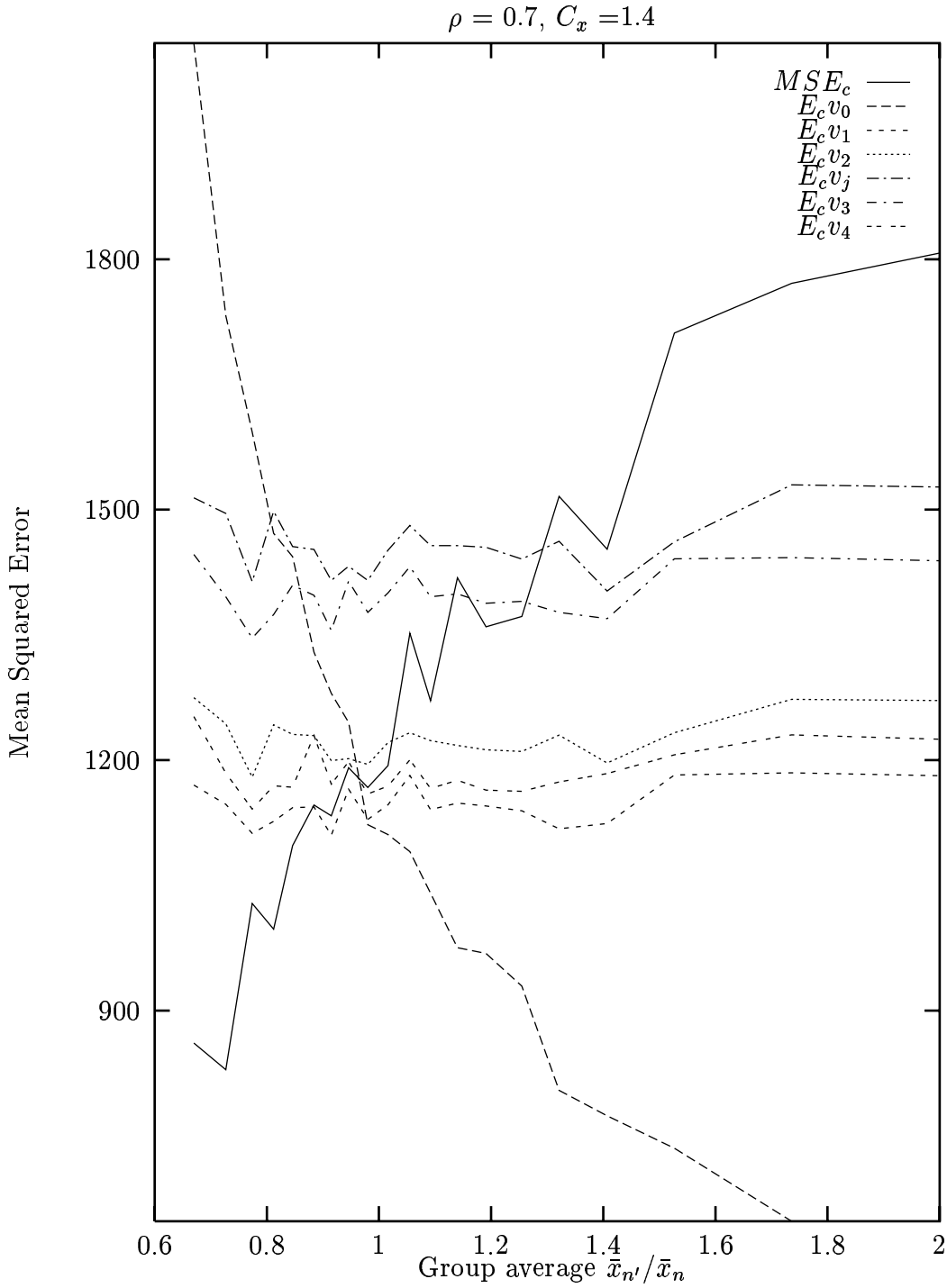

Figure 9: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and $n = 20$.

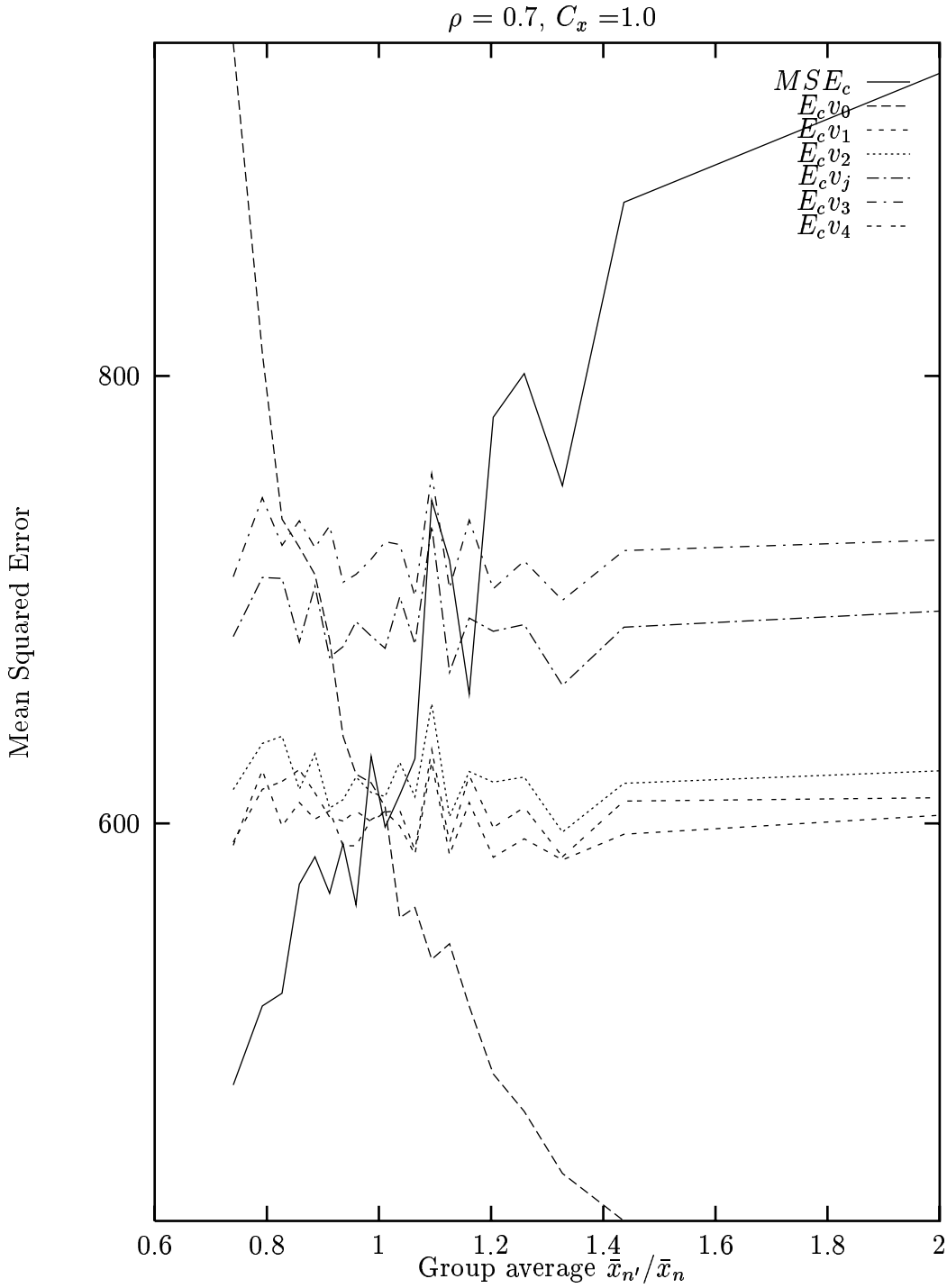

Figure 10: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and n = 20.

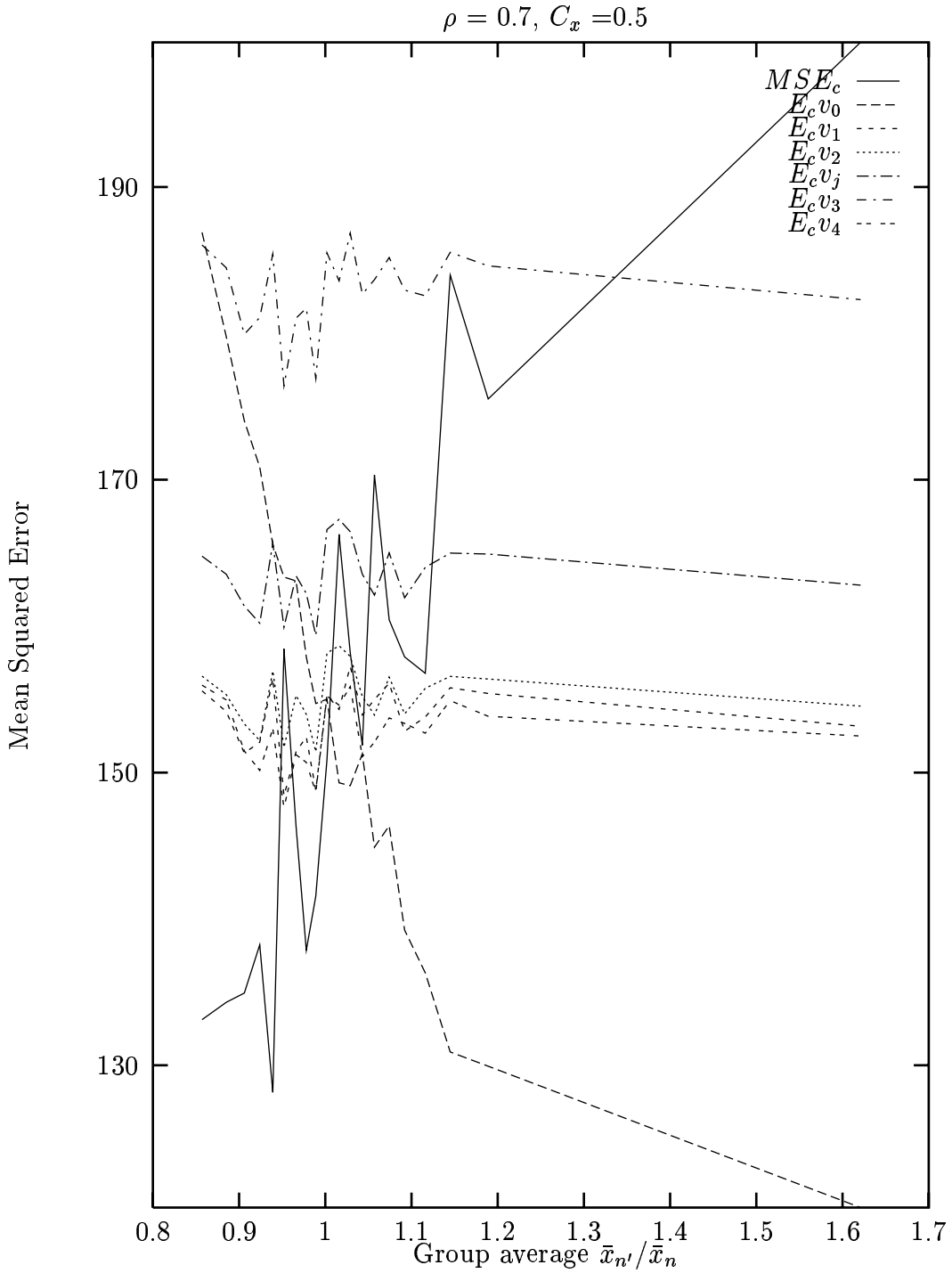

Figure 11: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and n = 20.

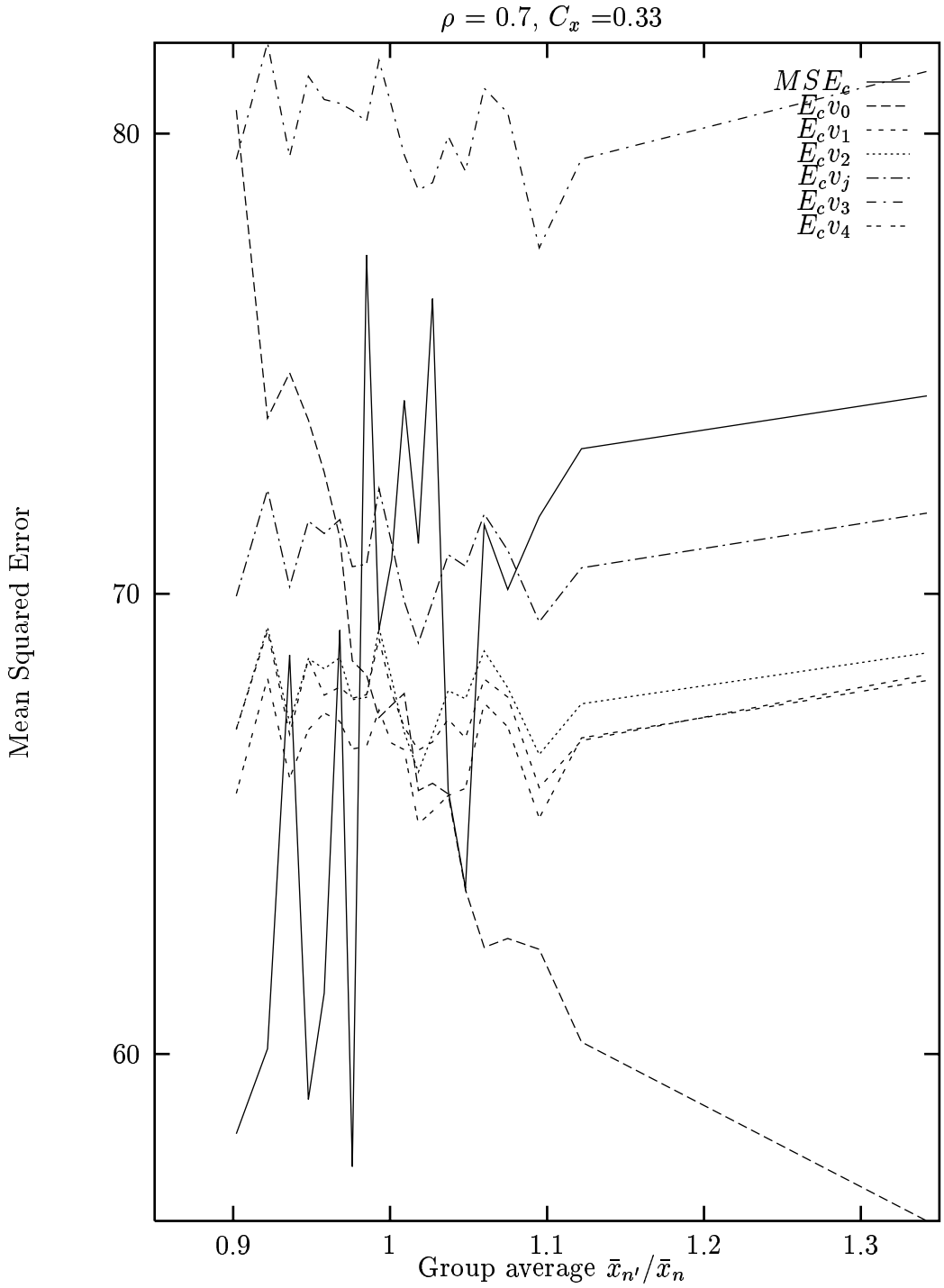

Figure 12: Conditional means $E_c v_0$, $E_c v_1$, $E_c v_2$, $E_c v_3$, $E_c v_4$, $E_c v_j$ and conditional mean squared error (MSE_c) of $\bar{y}_{rt}$ versus group average $\bar{x}_{n'}/\bar{x}_n$ with n'=100 and n = 20.

References

Benhin, E., and Prasad, N. G. N. (1997). Empirical likelihood estimation in two-phase sampling using two auxiliary variables. Unpublished manuscript.

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**, 107–116.

Cochran, W. G. (1977). *Sampling Techniques*, third edition, John Wiley and Sons, New York.

Dorfman, A. H. (1994). A note on variance estimation for the regression estimator in double sampling. *J. Am. Statist. Assoc.*, **89**, 137–140.

Hartley, H. O., and Rao, J. N. K. (1968). Covariate measurement error in generalized linear models. *Biometrika*, **55**, 547–557.

Rao, J. N. K., and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453–460.

Rao, P. S. R. S., and Rao, J. N. K. (1971). Small sample results for ratio estimators. *Biometrika*, **58**, 625–630.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *J. Am. Statist. Assoc.*, **92**, 780–787.

Sukhatme, P. V., and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications*, second edition, Iowa State University Press, Ames.