Discussion Papers No. 383, August 2004
Statistics Norway

Jan F. Bjørnstad and Elinor Ytterstad

Two-Stage Sampling from a Prediction Point of View

Abstract:
This paper considers the problem of estimating the population total in two-stage cluster sampling when cluster sizes are unknown, making use of a population model arising basically from a variance component model. The problem can be considered as one of predicting the unobserved part Z of the total, and the concept of predictive likelihood is studied. Prediction intervals and a predictor for the population total are derived for the normal case, based on predictive likelihood. The predictor obtained from the predictive likelihood is shown to be approximately uniformly optimal for large sample size and large number of clusters, in the sense of uniformly minimizing the mean square error in a partially linear class of model-unbiased predictors. Three prediction intervals for Z based on three similar predictive likelihoods are studied. For a small number n_0 of sampled clusters they differ significantly, however, for large n_0 the three intervals are practically identical. Model-based and design-based coverage properties of the prediction intervals are studied based on a comprehensive simulation study. Roughly, the simulation study indicates that for large sample sizes the coverage measures achieve approximately the nominal level $1 - \alpha$ and are slightly less than $1 - \alpha$ for moderately large sample sizes. For small sample sizes the coverage measures are about 95% of the nominal level.

Keywords: Survey sampling, population model, predictive likelihood, optimal predictor, prediction intervals, simulation

JEL classification: C42, C13, C15

Address: Jan F. Bjørnstad, Statistics Norway, Division for Statistical Methods and Standards. E-mail: jab@ssb.no

Elinor Ytterstad, University of Tromsø, Department of Mathematics and Statistics, N-9037 Tromsø, E-mail: Elinor.Ytterstad@matnat.uit.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1. Introduction

Two-stage surveys are used in sampling from finite populations of, say, N primary units or clusters, where each cluster consists of m_i units. N is assumed known. As mentioned by Kelly and Cumberland (1990) and Valliant, Dorfman and Royall (2000, ch. 8.9), it often happens that the m_i's are unknown before sampling, and this is the case we consider in this paper. Let y_{ij} be the value of the variable of interest for unit j of i'th cluster. The problem is to estimate the total

$$t = \sum_{i=1}^{N} \sum_{j=1}^{m_i} y_{ij} \ .$$

An example is considered in Thomsen, Tesfu and Binder (1986) and Thomsen and Tesfu (1988), with t being the size of a particular population. The clusters are certain administrative units, the units are households and y_{ij} is the number of persons in household j of the i'th administrative unit.

We assume that, before sampling, other measures of the sizes of the clusters are available to us. Let $x_1,...,x_N$ be these measures with $X = \sum_{i=1}^{N} x_i$. Kelly and Cumberland (1990) consider a case where the clusters are blocks of dwelling units and x_i is the number of units in block i from a previous census.

The sampling plan is as follows: At stage 1 a sample s of size n_0 of the clusters (1,..., N) is selected according to some sampling design. At stage 2 we select for each $i \in s$, a sample s_i of size n_i of units using possibly a different sampling design than at stage 1. The designs are assumed to be non-informative, i.e., they do not depend on the y_{ij}'s and the m_i's. E.g., in Thomsen and Tesfu (1988) the two-stage sampling plan is to use pps-sampling at stage 1 (letting selection probabilities of clusters be proportional to the x_i's) and simple random sampling (srs) at stage 2. This is a common two-stage sampling plan, as also mentioned by Kelly and Cumberland (1990). Usually, the second stage sample sizes are the same, leading to approximately equal selection probabilities for all units provided the ratios x_i/m_i are not too different. When the m_i's are known, one often used sampling plan is to let the first stage selection probabilities be proportional to m_i, and then srs with same sample sizes at stage 2 yielding equal selection probabilities for all units. As mentioned by Valliant et al. (2000, ch.8.1), equal sample sizes at stage 2 has many advantages and is probably the most common allocation of sample units in practice.

The total sample size is $n = \sum_{i \in s} n_i$ and our data now consists of $y(\mathbf{s}) = \left(y_{ij}\right)_{i \in s, j \in s_i}$ and the vector $m(s) = (m_i)_{i \in s}$, where $\mathbf{s} = \{s, s_i : i \in s\}$. Let $y = \left(y(\mathbf{s}), m(s)\right)$. For the pps-srs sampling plan mentioned above, a commonly used design-unbiased estimator of *t* is the Horvitz-Thompson estimator (see, e.g., Thomsen et al. 1986, Kelly and Cumberland,1990, and Särndal, Swensson and Wretman, 1992)

$$\hat{t}_{HT} = \frac{X}{n_0} \sum_{i \in s} \frac{m_i \bar{y}_i}{x_i} \tag{1}$$

where $\bar{y}_i = \sum_{j \in s_i} y_{ij} / n_i$.

In this paper a population model is adopted, regarding m_i, y_{ij} as realized values of random variables M_i, Y_{ij} for $j = 1, ..., M_i$ and $i = 1, ..., N$. The M_i's are assumed independent of all Y_{ij}, and furthermore:

$$\begin{aligned} &E(M_i) = \beta x_i,\ V(M_i) = \sigma^2 v(x_i) \\ &Cov(M_i, M_j) = 0 \\ &E(Y_{ij}) = \mu,\ \ V(Y_{ij}) = \tau^2 \\ &Cov(Y_{ij}, Y_{ik}) = \rho\tau^2 \quad \text{if } k \neq j, \rho \geq 0 \\ &Cov(Y_{ij}, Y_{lk}) = 0 \quad \text{if } l \neq i. \end{aligned} \tag{2}$$

Since the variance of a cluster total is nonnegative, we must have $\rho \geq -1/(\max m_i - 1)$ as also noted by Kelly and Cumberland (1990). It is therefore a minor restriction to assume a nonnegative ρ. Also, usually $v(x) = x^g$ with $0 \leq g \leq 2$. In fact, it is typically assumed that $v(x) = x$ (see e.g., Royall,1986, Kelly and Cumberland,1990, and Valliant et al.,2000, ch. 8.9).

A more general model is to let ρ and τ vary with the clusters, having cluster parameters ρ_i, τ_i. However, we then have the problem of estimating these parameters. Without further assumptions we are only able to estimate $(1 - \rho_i)\tau_i^2$. As noted by Valliant el al. (2000, ch. 8.1), it is often sensible to adopt model (2), especially after suitable stratification that also may allow μ to be different for different parts of the population.

The model (2) for the Y_{ij}'s arises naturally from expressing Y_{ij} in the following way:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where all μ_i, ε_{ij} are independent with

$$E(\mu_i) = \mu, V(\mu_i) = \tau_b^2 \text{ and } E(\varepsilon_{ij}) = 0, V(\varepsilon_{ij}) = \tau_w^2. \quad (3)$$

Here, $V(\mu_i) = \tau_b^2$ expresses the variability between the clusters, and $V(\varepsilon_{ij}) = \tau_w^2$ expresses the variability within the clusters. Then $\tau^2 = \tau_b^2 + \tau_w^2$ and the intraclass correlation is given by:

$$\rho = \tau_b^2 / \tau^2,$$

the proportion of the total variability due to the variability between the clusters.

The total *t* is now a realized value of a random variable *T*, where *T* can be expressed as

$$T = \sum_{i \in s} \sum_{j \in s_i} Y_{ij} + Z \text{ with}$$

$$Z = \sum_{i \in s} \sum_{j \notin s_i} Y_{ij} + \sum_{i \notin s} \sum_{j=1}^{M_i} Y_{ij}. \quad (4)$$

Expressing the *T* on this form, we see that the problem can be expressed as one of *predicting* the unobserved value *z* of the random variable *Z*. It is often clarifying to write a predictor $\hat{T}$ of *T* on the form

$$\hat{T} = \sum_{i \in s} \sum_{j \in s_i} Y_{ij} + \hat{Z} \quad (5)$$

where $\hat{Z}$ then implicitly is a predictor of Z. Considering the modified Horvitz-Thompson estimate $\hat{T}_{HT}$ given by (1) on the form (5), we can use the following expression, with $X_s = \sum_{i \in s} x_i$,

$$\hat{T}_{HT} = \sum_{i \in s} \sum_{j \in s_i} Y_{ij} + \sum_{i \in s} (m_i \frac{X_s}{n_0 x_i} - n_i) \overline{Y}_i + \sum_{i \notin s} x_i \left(\frac{1}{n_0} \sum_{j \in s} \frac{m_j}{x_j} \overline{Y}_j \right).$$

The last term predicts $\sum_{i \notin s} \sum_{j=1}^{M_i} Y_{ij}$ while the second term predicts $\sum_{i \in s} \sum_{j \notin s_i} Y_{ij}$. From this point of view $\hat{T}_{HT}$ does not look like a reasonable predictor.

Modeling the population in survey sampling has been and still is somewhat controversial, although most statisticians seem to agree on using modeling in developing statistical methods while evaluation is done with respect to the sampling design. An important aspect of this issue is that the likelihood principle in a sense makes it necessary to model the population. Without a model the only stochastic elements are the samples $\mathbf{s} = \{s, s_i : i \in s\}$, and the likelihood function is then flat (see, e.g., Cassel, Särndal and Wretman, 1977). This means that from the likelihood principle point of view the data contains no information about the unobserved y_{ij}'s and m_i's. To make inference we therefore need to

relate the data to the unobserved values somehow, and the most natural way of doing so is to formulate a model (see also remarks by Berger and Wolpert, 1988, p. 114 and Bjørnstad, 1996).

The random variables observed are $Y(\mathbf{s})$, $M(s)$ and $\mathbf{s}$, where $\mathbf{s}$ now is ancillary. The likelihood principle implies that inference should depend only on the actual $\mathbf{s}$ observed and not on the sampling design. This is usually called the prediction approach to survey sampling and will be adopted in this paper. Hence, theoretical considerations are conditional on given $\mathbf{s}$. The prediction approach aims at choosing a predictor that is good for the actual $\mathbf{s}$ obtained and has given significant contributions to a better understanding of several problems in survey sampling, some of which are mentioned in Thomsen and Tesfu (1988) and Valliant et al. (2000). It also enables one to use more conventional statistical methods, although the problem is not to make inferences about θ but rather predict Z. Hence, θ basically plays the role of a nuisance parameter.

To predict Z we shall use predictive likelihood based methods, a non-Bayesian likelihood approach to prediction problems in general. One can argue that in the context of a population model, survey sampling provides one of the more natural "prediction" problems in statistics. Predictive likelihood can therefore serve as a basis for essentially all problems of this kind in survey sampling. Some major references to the general theory of predictive likelihood are Hinkley (1979), Mathiasen (1979) and Butler (1986). A review of some of the suggested likelihoods is given in Bjørnstad (1990, 1998). Predictive likelihood is discussed from the perspective of the likelihood principle for prediction in Bjørnstad (1996). Bolfarine and Zacks (1992) consider methods based on predictive likelihood in survey sampling.

Section 2 introduces the concept of predictive likelihood and shows how predictors and prediction intervals can be constructed from a predictive likelihood, and in Section 3 a predictive likelihood is derived for the normal model. Considering a predictive likelihood for Z directly does not work, mainly because Z is a sum of a *stochastic* number of random variables. Therefore, predictor and prediction interval will be obtained from a joint predictive likelihood for Z and the vector $M(\bar{s}) = (M_i)_{i \notin s}$.

In Section 3.3 optimality theory for a class ℓ of predictors linear in the Y_{ij}'s, but not simultaneously in both Y_{ij}'s and M_i's, under the distribution-free model (2) is developed.

In Section 4 three prediction intervals for Z based on similar predictive likelihoods are studied and a comprehensive simulation study for estimating confidence levels, both model-based and design-based is undertaken. The prediction intervals are evaluated by four different measures; the model-based

coverage C_m, the design-based coverage C_d, the unconditional coverage C (expected design-based coverage), and the conditional coverage given the data.

2. Predictive likelihood

We shall here give a brief general introduction to the concept of predictive likelihood. For a more complete exposition we refer to Bjørnstad (1990, 1998). Let $Y = y$ be the data. The problem is to predict the unobserved or future value z of a random variable Z usually by a predictor and confidence interval for Z. It is assumed that (Y,Z) has a probability density or mass function (pdf) $f_\theta(y,z)$. In general we let $f_\theta(\cdot)$ and $f_\theta(\cdot\,|\,\cdot)$ denote the pdf and conditional pdf of the enclosed variables. The likelihood basis in prediction is the generalized joint likelihood for the two unknown quantities, z and θ. In Bjørnstad (1996) it is shown that the joint likelihood function is given by $l_y(z,\theta) = f_\theta(y,z)$. With this likelihood, the corresponding likelihood principle is implied by the sufficiency principle for prediction and the conditionality principle, generalizing the fundamental result by Birnbaum (1962) for parametric likelihood. The aim is to develop a partial likelihood for z, $L(z|y)$, by eliminating θ from l_y. Any such likelihood is called a predictive likelihood and gives rise to one particular prediction method.

Different ways of eliminating θ give rise to different L. The two main type of suggestions are the conditional predictive likelihood L_c , essentially suggested by Hinkley (1979), and the profile predictive likelihood L_p, first considered by Mathiasen (1979). Let $R = r(Y,Z)$ denote a minimal sufficient statistics for (Y,Z). Then

$$L_c(z\,|\,y) = f_\theta(y,z) / f_\theta(r(y,z)) \tag{6}$$

$$L_p(z\,|\,y) = \max_\theta f_\theta(y,z) = f_{\hat\theta_z}(y,z)\,. \tag{7}$$

Typically, L_c and L_p are quite similar when sufficiency provides a genuine reduction and the dimension of θ is small.

In linear models, L_p will ignore the number of parameters and can be misleadingly precise. A modification of L_p, L_{mp}, that adjusts for this was suggested by Butler (1986, rejoinder), see also Bjørnstad (1990). If Y,Z are independent, Y consisting of n independent observations and Z being an m-dimensional vector of independent variables, then L_{mp} is given by

$$L_{mp}(z\,|\,y) = L_p(z\,|\,y)\cdot |\,I^z(\hat\theta_z)\,|^{1/2} \,/\, |\,H_z H_z'\,|^{1/2}\,. \tag{8}$$

Here, $I^z(\theta) = \{I^z_{ij}(\theta)\}$ is the "observed" information-matrix based on (y,z), i.e. $I^z_{ij}(\theta) = -\partial^2 \log f_\theta(y,z)/\partial\theta_i\partial\theta_j$. $H_z = H_z(\hat{\theta}_z)$, and $H_z(\theta)$ is the k x (n+m) matrix of second-order partial derivatives of $\log f_\theta(y,z)$ with respect to k-dimensional θ and (y,z).

We shall assume that any L considered is normalized as a probability distribution in Z. The mean and variance of L are then called the predictive expectation and the predictive variance of Z, denoted by $E_p(Z)$ and $V_p(Z)$. $E_p(Z)$ is then a natural predictor for z, called the mean predictor. $L(z|y)$ also gives us an idea on how likely different z-values are in light of the data, and can be used to construct prediction intervals for z. An interval (a_y, b_y) is a (1-α) predictive interval based on $L(z|y)$ if

$$\int_{a_y}^{b_y} L(z \mid y)dz = 1-\alpha.$$

A simplified (1-α) predictive interval is of the form

$$E_p(Z) \pm u_{\alpha/2}\sqrt{V_p(Z)} \tag{9}$$

where $u_{\alpha/2}$ is the upper α/2-point in the actual (exact or approximate) conditional distribution, given y, of $(Z - E_\theta(Z \mid y))/\sqrt{V_\theta(Z \mid y)}$.

3. Predictive likelihood and predictor in two-stage sampling

3.1 Predictive likelihood for mixtures

In two-stage sampling, Z is given by (4), and is the sum of two mixtures. Therefore, instead of considering a predictive likelihood for Z directly, we look at a joint predictive likelihood for Z and $M(\bar{s})$. It has the following form

$$L(z, m(\bar{s}) \mid y) = L_{m(\bar{s})}(z \mid y)L(m(\bar{s}) \mid y). \tag{10}$$

$L_{m(\bar{s})}(z \mid y)$ is a predictive likelihood for z conditional on $M(\bar{s}) = m(\bar{s})$, i.e., based on $f_\theta(y, z \mid m(\bar{s}))$. Since $f_\theta(y, z \mid m(\bar{s})) = f_{\mu,\tau,\rho}(y(\mathbf{s}), z \mid m(s), m(\bar{s}))f_{\beta,\sigma}(m(s))$, $L_{m(\bar{s})}(z \mid y)$ is, in fact, based on $f_{\mu,\tau,\rho}(y(\mathbf{s}), z \mid m(s), m(\bar{s}))$. $L(m(\bar{s}) \mid y)$ is a predictive likelihood for $m(\bar{s})$ based on $f_\theta(y, m(\bar{s}))$. The predictive likelihood for Z is given by the marginal in (10), e.g., in case of a continuous model for M_i,

$$L(z \mid y) = \int L(z, m(\bar{s}) \mid y) dm(\bar{s}) \tag{11}$$

Then E_p and V_p follow the usual rules for double expectation, i.e.,

$$E_p(Z) = E_p\{E_p(Z \mid M(\bar{s}))\} \tag{12}$$

$$V_p(Z) = E_p\{V_p(Z \mid M(\bar{s}))\} + V_p\{E_p(Z \mid M(\bar{s}))\}.$$

In (12), $E_p(Z \mid m(\bar{s}))$ and $V_p(Z \mid m(\bar{s}))$ are the predictive mean and variance for Z from $L_{m(\bar{s})}(z \mid y)$. In principle we can derive $L(z|y)$ as the marginal likelihood in (11). The advantage of (12) is that we are able to obtain $E_p(Z)$ and $V_p(Z)$ without actually deriving $L(z|y)$.

Under the model (2) we can factorize $f_\theta(y, z, m(\bar{s})) = f_{\mu,\tau,\rho}(y(\mathbf{s}), z \mid m(s), m(\bar{s})) f_{\beta,\sigma}(m(s), m(\bar{s}))$ and it is readily seen that applying L_p, given by (7), to the terms on the right hand side in (10) in fact gives us $L_p(z, m(\bar{s}) \mid y) = \max_\theta f_\theta(y, z, m(\bar{s}))$, i.e.,

$$L_p(z, m(\bar{s}) \mid y) = L_{m(\bar{s}),p}(z \mid y) L_p(m(\bar{s}) \mid y). \tag{13}$$

It follows that $E_p(Z)$ and $V_p(Z)$ based on $L_p(z, m(\bar{s}) \mid y)$ can be derived by (12). We note that L_c, given by (6), has the same property, i.e., $L_c(z, m(\bar{s}) \mid y) = L_{m(\bar{s}),c}(z \mid y) L_c(m(\bar{s}) \mid y)$.

3.2 Normal model

It is now assumed that model (2) holds with Y_{ij} and M_i normally distributed. We shall first consider the second likelihood in (10), $L(m(\bar{s}) \mid y)$, using the profile predictive likelihood L_p. Let $t_\nu^{(k)}(\Sigma)$ denote the k-dimensional multivariate t-distribution with ν degrees of freedom and variance-covariance matrix Σ, i.e., $t_\nu^{(k)}(\Sigma)$ is the distribution of $(\mathbf{U}/W)\sqrt{\nu}$ where $\mathbf{U} \sim N_k(0, \Sigma)$ and W^2 has a chi-square distribution with ν degrees of freedom. Let $X(\bar{s})$ be the vector of $(x_i : i \notin s)$. Then $L_p(m(\bar{s}) \mid y)$ leads to a multivariate t-distribution, such that $[M(\bar{s}) - \hat{\beta} X(\bar{s})]/\hat{\sigma} \sim t_{n_0}^{(N-n_0)}(V)$, where the maximum likelihood estimators (MLE) are, with $W_s = \sum_{i \in s} x_i^2 / v(x_i)$, $\hat{\beta} = \frac{1}{W_s} \sum_{i \in s} m_i x_i / v(x_i)$, the best unbiased estimator uniformly minimizing the variance, and $\hat{\sigma}^2 = \frac{1}{n_0} \Sigma_{i \in s} (m_i - \hat{\beta} x_i)^2 / v(x_i)$. $V = (v_{ij})$ with $v_{ii} = v(x_i) + x_i^2 / W_s$ and $v_{ij} = x_i x_j / W_s$ for $i \neq j$. It follows that $E_p(M_i) = \hat{\beta} x_i$,

$V_p(M_i) = \frac{n_0}{n_0-2}\hat{\sigma}^2(v(x_i) + x_i^2 / W_s)$ and the predictive covariances are given by

$Cov_p(M_i, M_j) = \frac{n_0}{n_0-2}\hat{\sigma}^2 \cdot x_i x_j / W_s$ for $i \neq j$. This implies that

$$E_p(\sum_{i \notin s} M_i) = \hat{\beta} X_{\bar{s}} \tag{14}$$

$$V_p(\sum_{i \notin s} M_i) = \frac{n_0}{n_0 - 2}\hat{\sigma}^2\left(v_{\bar{s}} + \frac{X_{\bar{s}}^2}{W_s}\right)$$

where $X_{\bar{s}} = \sum_{i \notin s} x_i$ and $v_{\bar{s}} = \sum_{i \notin s} v(x_i)$.

L_c and L_{mp} (for $M_i / \sqrt{v(x_i)}, i \notin s$), given by (8), lead to moments similar to (14) with n_0 - 2 replaced by n_0 - 5 and n_0 - 4 respectively.

Let us now consider the first term in (10), $L_{m(\bar{s})}(z \mid y)$ based on $f_{\mu,\tau,\rho}(y(\mathbf{s}), z \mid m(s), m(\bar{s}))$. For this likelihood we will restrict attention to L_p, i.e., deriving $L_{m(\bar{s}),p}(z \mid y)$. The MLE $\hat{\mu}, \hat{\tau}^2, \hat{\rho}$ can be expressed in the following way, with $SSE = \sum_{i \in s} \sum_{j \in s_i} (y_{ij} - \bar{y}_i)^2$:

$$\hat{\mu} = \sum_{i \in s} \frac{n_i}{1 - \hat{\rho} + n_i \hat{\rho}} \bar{y}_i \Big/ \sum_{i \in s} \frac{n_i}{1 - \hat{\rho} + n_i \hat{\rho}} \tag{15}$$

$$\hat{\tau}^2 = \frac{1}{n}\left(\frac{SSE}{1 - \hat{\rho}} + \sum_{i \in s} \frac{n_i (\bar{y}_i - \hat{\mu})^2}{1 - \hat{\rho} + n_i \hat{\rho}}\right)$$

and $\hat{\rho}$ is found numerically, maximizing

$$-(n/2)\log\hat{\tau}^2 - (1/2)\sum_{i \in s} \log(1 + (n_i - 1)\hat{\rho}) - ((n - n_0)/2)\log(1 - \hat{\rho}).$$

When $n_i = c$, for all $i \in s$, then $\hat{\mu} = \bar{y} = \sum_{i \in s} \bar{y}_i / n_0$, $\hat{\tau}^2 = SS / n$, $\hat{\rho} = \max(0, 1 - \frac{c}{c-1} \cdot \frac{SSE}{SS})$, where $SS = \sum_{i \in s} \sum_{j \in s_i} (y_{ij} - \bar{y})^2$.

Consider first the case when ρ and τ are known. Then $\hat{\mu}$ is given by (15) with ρ replacing $\hat{\rho}$. In this case $L_{m(\bar{s}),p}(z \mid y)$ is such that Z is normally distributed with predictive mean and predictive variance

$$E_p(Z \mid m(\bar{s})) = \sum_{i \in s} (m_i - n_i) \cdot \left(\frac{1 - \rho}{1 - \rho + n_i \rho} \hat{\mu} + \frac{n_i \rho}{1 - \rho + n_i \rho} \bar{y}_i \right) + \hat{\mu} \sum_{i \notin s} m_i \tag{16}$$

$$V_p(Z \mid m(\bar{s})) = V(Z \mid y, m(\bar{s})) + \frac{\tau^2}{\sum_{i \in s} \frac{n_i}{1 - \rho + n_i \rho}} \left(\sum_{i \notin s} m_i + \sum_{i \in s} (m_i - n_i) \cdot \frac{1 - \rho}{1 - \rho + n_i \rho} \right)^2 . \tag{17}$$

Here, $V(Z \mid \cdot)$ denotes the usual variance in the conditional distribution of *Z*. When ρ, τ are unknown, $L_{m(\bar{s}),p}(z \mid y)$ will for large n_0 be approximately such that *Z* is normally distributed with $E_p(Z \mid m(\bar{s}))$ and $V_p(Z \mid m(\bar{s}))$ given by (16) and (17) with $\hat{\rho}, \hat{\tau}^2$ replacing ρ, τ^2. Recall that $\bar{y}_i = \sum_{j \in s_i} y_{ij} / n_i$ and $\hat{\theta} = (\hat{\mu}, \hat{\tau}, \hat{\rho}, \hat{\beta}, \hat{\sigma})$ the MLE of $\theta = (\beta, \sigma, \mu, \tau, \rho)$. Then the conditional expected value of *Z* given the data, estimated at $\hat{\theta}$, is equal to

$$E_{\hat{\theta}}(Z \mid y) = \sum_{i \in s} (m_i - n_i) \left(\frac{1 - \hat{\rho}}{1 - \hat{\rho} + n_i \hat{\rho}} \hat{\mu} + \frac{n_i \hat{\rho}}{1 - \hat{\rho} + n_i \hat{\rho}} \bar{y}_i \right) + \hat{\mu} \sum_{i \notin s} (\hat{\beta} x_i) \tag{18}$$

Let $V_{\hat{\theta}}(Z \mid y)$ denote the estimated conditional variance of *Z* given the data. It now follows, from (12) - (14), (16) - (18) that, approximately, $L_p(z, m(\bar{s}) \mid y)$ has

$$E_p(Z) = E_{\hat{\theta}}(Z \mid y)$$

and

$$V_p(Z) = V_{\hat{\theta}}(Z \mid y)$$
$$+ \frac{\hat{\tau}^2}{\sum_{i \in s} \frac{n_i}{1 - \hat{\rho} + n_i \hat{\rho}}} \left[\hat{\beta} X_{\bar{s}} + (1 - \hat{\rho}) \sum_{i \in s} \frac{m_i - n_i}{1 - \hat{\rho} + n_i \hat{\rho}} \right]^2 + \hat{\sigma}^2 \left(\hat{\mu}^2 \cdot \frac{X_{\bar{s}}^2}{W_s} + \hat{\rho} \hat{\tau}^2 \cdot \frac{\sum_{i \notin s} x_i^2}{W_s} \right) + h(2). \tag{19}$$

Here,

$$V_\theta(Z \mid y) = \tau^2 (1 - \rho) \sum_{i \in s} (m_i - n_i) \left(1 + (m_i - n_i) \frac{\rho}{1 + (n_i - 1)\rho} \right)$$
$$+ \tau^2 (\beta X_{\bar{s}} + \rho \sigma^2 v_{\bar{s}} + \rho \sum_{i \notin s} \beta x_i (\beta x_i - 1)) + \mu^2 \sigma^2 v_{\bar{s}}$$

and

$$h(k) = \frac{n_0}{n_0 - k} \hat{\sigma}^2 \cdot \left(\frac{\hat{\tau}^2}{\sum_{i \in s} \frac{n_i}{1 - \hat{\rho} + n_i \hat{\rho}}} + \frac{k}{n_0} \hat{\mu}^2 \right) \left(v_{\bar{s}} + \frac{X_{\bar{s}}^2}{W_s} \right) + \frac{k}{n_0 - k} \hat{\rho} \hat{\tau}^2 \hat{\sigma}^2 \cdot \left(v_{\bar{s}} + \frac{1}{W_s} \sum_{i \notin s} x_i^2 \right).$$

The predictive likelihood

$$L_{p,c}(z, m(\bar{s}) \mid y) = L_{m(\bar{s}),p}(z \mid y) L_c(m(\bar{s}) \mid y)$$

leads to the same $E_p(Z)$ while $V_p(Z)$ equals (19) with $h(5)$ instead of $h(2)$. With

$$L_{p,mp}(z, m(\bar{s}) \mid y) = L_{m(\bar{s}),p}(z \mid y) L_{mp}(m(\bar{s}) \mid y)$$

we get the same $E_p(Z)$ and $V_p(Z)$ equal to (19) with $h(4)$.

Let $\hat{w}_i = (n_i\hat{\rho})/(1-\hat{\rho}+n_i\hat{\rho})$. Writing the predictor $\hat{Z}_0 = E_{\hat{\theta}}(Z \mid y)$, given by (18), as

$$\hat{Z}_0 = \sum_{i \in s} \sum_{j \notin s_i} \left((1-\hat{w}_i)\hat{\mu} + \hat{w}_i \bar{y}_i\right) + \sum_{i \notin s} (\hat{\beta} x_i)\hat{\mu} \qquad (20)$$

we see from (4) that predicting Z by $\hat{Z}_0$ means that for $i \notin s$ each unobserved Y_{ij} is predicted by $\hat{\mu}$ and M_i is predicted by $\hat{\beta} x_i$. For $i \in s$, $j \notin s_i$, Y_{ij} is predicted by $\hat{w}_i \bar{y}_i + (1-\hat{w}_i)\hat{\mu}$. This predictor shrinks the natural estimate $\bar{y}_i$ towards $\hat{\mu}$. Using the representation (3) of the model, we note that $(1-\rho)/(1-\rho+n_i\rho) = Var(\bar{Y}_i \mid \mu_i)/(Var(\bar{Y}_i \mid \mu_i) + Var(\mu_i))$. Hence, for $i \in s$, the smaller $Var(\mu_i)$ is compared to $Var(\bar{Y}_i \mid \mu_i)$, the more weight we put on $\hat{\mu}$ to predict Y_{ij} for $j \notin s_i$. Or, in other words, the smaller the variability is *between* the clusters compared to the variability *within* the clusters, the more $\bar{y}_i$ shrinks towards $\hat{\mu}$.

3.3 Some optimality considerations

All three predictive likelihoods for the model (2), with normally distributed Y_{ij} and M_i, give the same predictor for the population total T,

$$\hat{T}_0 = \sum_{i \in s} \sum_{j \in s_i} y_{ij} + \hat{Z}_0$$

with $\hat{Z}_0$ given by (20).

The optimality considerations are conditional on $\mathbf{s} = \{s, s_i : i \in s\}$, and $E_\theta(\cdot)$ is used to denote $E_\theta(\cdot|\mathbf{s})$. Let $\ell = \{\hat{T} : \hat{T} = \sum_{i \in s} \sum_{j \in s_i} a_{ij} Y_{ij}\}$ be a class of "partially" linear predictors, where each a_{ij} is a function of $M(s)$. We shall restrict attention to the class of model-unbiased predictors in ℓ, i.e.,

$$\ell_u = \{\hat{T} \in \ell : E_\theta(\hat{T} - T) = 0, \forall \theta\}.$$

We shall now consider the distribution-free model (2). The parameter estimates of (β,σ^2) are still valid, $\hat{\beta}$ now the best linear unbiased (BLU) estimator and $\frac{n_0}{n_0-1}\hat{\sigma}^2$ still unbiased. Regarding the MLE $\hat{\mu}$, given by (15), it is readily seen that with ρ replacing $\hat{\rho}$, $\hat{\mu}$ is the BLU estimator as also noted by Kelly and Cumberland (1990). What remains is to derive alternative estimators for ρ and τ^2. Here one can use an ANOVA approach, as in Valliant et al. (2000, ch. 8.3) or Kelly and Cumberland (1990). When $n_i = c$ for all $i \in s$, these two ANOVA approaches yield the same estimators $\hat{\rho}_{av}, \hat{\tau}^2_{av}$ satisfying

$$\hat{\rho}_{av}\hat{\tau}^2_{av} = \frac{1}{c}\left(\frac{SS - SSE}{n_0 - 1} - \frac{SSE}{n - n_0}\right)$$

$$(1-\hat{\rho}_{av})\hat{\tau}^2_{av} = \frac{SSE}{n - n_0}.$$

It follows that, approximately, (for large n_0 with $(n_0 - 1)/n_0 \approx 1$), $\hat{\tau}^2_{av} = SS/n$ and $\hat{\rho}_{av} = 1 - \frac{c}{c-1}\cdot\frac{SSE}{SS}$; the same as the MLE in the normal model.

With these new parameter estimates $\hat{T}_0$ is clearly a reasonable predictor also for this distribution-free model, e.g., Kelly and Cumberland (1990) suggests using this predictor (see also Valliant et al., 2000, ch.8.9). The optimal procedure at θ, $\hat{T}_\theta$, in ℓ_u is defined to be the predictor in ℓ_u that minimizes $E_\theta(\hat{T} - T)^2$ for $\hat{T} \in \ell_u$. If $\hat{T}_\theta$ does not depend on θ it is uniformly optimal.

We see that, by using that $E(\hat{T} - T) = EE(\hat{T} - T \mid \mathbf{M})$, with $\mathbf{M} = (M_1, ..., M_N)$

$$\hat{T} \in \ell_u \Leftrightarrow E_\beta\left(\sum_{i\in s}\sum_{j\in s_i} a_{ij}\right) = \beta X, \forall\beta$$

We note the following result.

Lemma 1. The optimal predictor $\hat{T}_\theta$ must be a member of the class

$$\ell_{0u} = \{\hat{T} \in \ell_u : \hat{T} = \sum_{i\in s} b_i\bar{Y}_i; b_i \text{ is function of } M(s), i \in s\}$$

and $$E_\beta\left(\sum_{i\in s} b_i\right) = \beta X, \forall\beta \qquad (21)$$

Proof. Using the rule $V(\hat{T}-T) = EV(\hat{T}-T\,|\,\mathbf{M}) + VE(\hat{T}-T\,|\,\mathbf{M})$, we see that, with $\bar{a}_i = \sum_{j\in s_i} a_{ij}/n_i$,

$$\hat{T}\in \ell_u \;\Rightarrow E_\theta(\hat{T}-T)^2 = V_\theta(\hat{T}-T)$$
$$= \tau^2(1-\rho)E_\theta[\sum_{i\in s}\sum_{j\in s_i} a_{ij}^2] + \rho\tau^2\sum_{i\in s} E_\theta(n_i\bar{a}_i)^2$$
$$- 2\rho\tau^2 E_\theta(\sum_{i\in s} M_i n_i\bar{a}_i) + \mu^2 V(\sum_{i\in s} n_i\bar{a}_i - \sum_{i\in s} M_i) + \psi \qquad (22)$$

Here, ψ is a function of the parameters only. Since $\sum_{j\in s_i} a_{ij}^2 \ge n_i\bar{a}_i^2$, it follows that $\hat{T}_\theta$ must have $a_{ij} = a_i$, for all $j\in s_i$, and $b_i = n_i a_i$. ♦

We restrict attention to the class $\mathcal{L}_u$ of model-unbiased predictors in ℓ_u where each a_{ij} is a linear function of $M(s)$. We note that $\hat{T}_{HT}$, given by (1), is a member of $\mathcal{L}_u$. Then, from Lemma 1, it is sufficient to consider the class

$$L_{0u} = \{\hat{T}\in \ell_u : \hat{T} = \sum_{i\in s} b_i\bar{Y}_i \text{ and } b_i = c_i + \sum_{j\in s} c_{ij}M_j\}.$$

From (21),

$$\hat{T}\in L_{0u} \Leftrightarrow E_\beta(\sum_{i\in s} b_i) = \beta X, \forall\beta \Leftrightarrow \sum_{i\in s} c_i + \beta\sum_{i\in s}\sum_{j\in s} c_{ij}x_j = \beta X, \forall\beta.$$

Hence,

$$\hat{T}\in L_{0u} \Leftrightarrow \sum_{i\in s} c_i = 0 \text{ and } \sum_{i\in s}\sum_{j\in s} c_{ij}x_j = X. \qquad (23)$$

We note that $\hat{T}_0$ can be expressed as $\sum_s b_i^0\bar{Y}_i$ and b_i^0 is linear in $M(s)$. $\hat{T}_0$ satisfies (23) with $c_i = 0$ and hence is model-unbiased when ρ is known (e.g., when $\rho = 0$) and approximately model-unbiased otherwise.

Lemma 2. The optimal predictor $\hat{T}_\theta$ in L_{0u} minimizes with respect to $\mathbf{c} = (c_i, i\in s; c_{ij}, i\in s, j\in s)$, subject to condition (23),

$$Q(\mathbf{c}) = \tau^2\sum_{i\in s}\tfrac{1}{\phi_i}(V(b_i) + (Eb_i)^2) - 2\rho\tau^2\sum_{i\in s} E(M_i b_i) + \mu^2 V[\sum_{i\in s}(b_i - M_i)]$$

where $\phi_i = \dfrac{n_i}{1-\rho+n_i\rho}$.

Proof. For $\hat{T} \in L_{ou}$, we see from (22), using (21),

$$E_\theta(\hat{T}-T)^2 = \tau^2(1-\rho)E_\theta\sum_{i\in s}(b_i^2/n_i) + \rho\tau^2\sum_{i\in s}[V_\theta(b_i)+(E_\theta b_i)^2]$$
$$- 2\rho\tau^2 E_\theta\sum_{i\in s}M_i b_i + \mu^2 V(\sum_{i\in s}b_i - \sum_{i\in s}M_i) + \psi$$

$$= \sum_{i\in s}(\frac{\tau^2(1-\rho)}{n_i} + \rho\tau^2)[V_\theta(b_i)+(E_\theta b_i)^2]$$
$$- 2\rho\tau^2 E_\theta\sum_{i\in s}M_i b_i + \mu^2 V(\sum_{i\in s}b_i - \sum_{i\in s}M_i) + \psi$$

Result follows since $(1-\rho)/n_i + \rho = 1/\phi_i$. ♦

Let now $\phi_s = \sum_s \phi_i$, $\alpha = \tau^2/(\tau^2 + \phi_s\mu^2)$ and $\hat{m}_i = (1-\alpha)m_i + \alpha\hat{\beta}x_i$. Then the following result holds.

Theorem. The optimal predictor at θ in $\mathcal{L}_u$ is given by

$$\hat{T}_\theta = \sum_{i\in s}(\hat{m}_i(1-w_i)\hat{\mu}_\rho + m_i w_i \bar{y}_i) + \hat{\mu}_\rho\sum_{i\notin s}\hat{\beta}x_i \qquad (24)$$

i.e., $\hat{T}_\theta = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \hat{Z}_\theta$ where

$$\hat{Z}_\theta = \sum_{i\in s}[(\hat{m}_i(1-w_i)\hat{\mu}_\rho + m_i w_i \bar{y}_i) - n_i\bar{y}_i] + \hat{\mu}_\rho\sum_{i\notin s}\hat{\beta}x_i .$$

Here $w_i = \rho\phi_i$ and $\hat{\mu}_\rho = \sum_s \phi_i\bar{y}_i/\phi_s$.

Remarks. (I) The optimal predictor at θ depends only on ρ and the coefficient of variation τ/μ, and is hence uniformly optimal in (μ,β,σ) if ρ and τ/μ are assumed known.

(II) The expression for $\hat{T}_\theta$ means that for $i\in s$, $\sum_{j=1}^{m_i} y_{ij}$ is estimated by $(\hat{m}_i(1-w_i)\hat{\mu}_\rho + m_i w_i\bar{y}_i)$, and for $i\notin s$, $\sum_{j=1}^{m_i} y_{ij}$ is estimated by $\hat{\mu}_\rho\hat{\beta}x_i$, i.e. m_i is estimated by $\hat{\beta}x_i$ and each y_{ij} by $\hat{\mu}_\rho$.

(III) Let $\hat{\mu}_i = (1-w_i)\hat{\mu}_\rho + w_i\bar{y}_i$. Then an alternative expression to (24) is:

$$\hat{T}_\theta = \sum_{i\in s} m_i\hat{\mu}_i + \hat{\mu}_\rho\sum_{i\notin s}\hat{\beta}x_i + R$$

where $R = \alpha\hat{\mu}_\rho \sum_s (\hat{\beta}x_i - m_i)(1 - w_i)$.

Now, $\sum_{i\in s} n_i\hat{\mu}_i = \sum_{i\in s} n_i\bar{y}_i$ and therefore $\hat{T}_\theta = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \hat{Z}_\theta$ where

$$\hat{Z}_\theta = \sum_{i\in s}(m_i - n_i)\hat{\mu}_i + \hat{\mu}_\rho \sum_{i\notin s}\hat{\beta}x_i + R.$$

It can be shown that R is typically of order $1/(N\sqrt{n_0})$ or less compared to $\hat{T}_0$. Hence $\hat{T}_\theta \approx \hat{T}_0$ if $\hat{\rho} \approx \rho$ and $\hat{T}_0$ is approximately uniformly optimal for large $N\sqrt{n_0}$ and large sample size n.

(IV) Valliant et al. (2000, ch. 8.9) and Kelly and Cumberland (1990) considers optimality for a completely linear class in $\bar{Y}_i$ and M_i. This is a rather restrictive class, excluding interesting estimators that are linear in $M_i\bar{Y}_i$, e.g., $\hat{T}_{HT}$, and also $\hat{T}_0$ even when ρ is known. Neither does the class include $\hat{T}_\theta$. They show that the optimal predictor in this class at θ is given by $\hat{T}^*_\theta = \sum_{i\in s}\sum_{j\in s_i} y_{ij} + \hat{Z}^*_\theta$, where $\hat{Z}^*_\theta = \sum_{i\in s}[(m_i - \beta x_i w_i)\mu - n_i(1 - w_i)\mu + (\beta x_i - n_i)w_i\bar{y}_i) + \mu\sum_{i\notin s}\beta x_i$, depending on the parameters β,μ,ρ. From remark (III), the theorem shows that if βx_i is replaced by m_i for $i \in s$ and μ,β is estimated by $\hat{\mu}_\rho, \hat{\beta}$ the predictor is approximately uniformly optimal in the class $\mathcal{L}_u$ for known ρ.

(V) As mentioned in Section 1, usually the model for the M_i's is to assume $v(x) = x$, leading to the ratio estimator for β, $\hat{\beta}_r = \sum_{i\in s} m_i / \sum_{i\in s} x_i$. As mentioned earlier, a commonly used sampling design is pps-sampling at stage 1 and srs at stage 2. Then the individual selection probabilities for unit j of the i'th cluster are given by $\pi_{ij} = n_0(x_i / X)(n_i / m_i)$. In surveys consisting of persons, it is customary to let all persons have the same selection probabilities. This is not possible exactly, since m_i is unknown when determining the size n_i of s_i. However, choosing $n_i = n/n_0 = c$ for all $i \in s$ will lead to $\pi_{ij} \approx n/X$, since x_i / m_i is approximately constant in i. Hence, a typical case is letting $v(x) = x$ and $n_i = c$ for all $i \in s$. Then $\hat{\mu}_\rho = \bar{y}$ and $R = 0$, and with ρ known, we have that $\hat{T}_0 = \hat{T}_\theta, \forall\theta$, and hence $\hat{T}_0$ is uniformly optimal for known ρ.

(VI) Consider the simplified model of negligible intraclass correlation, assuming $\rho = 0$. The simulation study in section 4, assuming normal model, for predictive intervals of the form (9) indicates that this is a valid assumption when $\rho \le 0.01$ (but not if ρ is assumed to be larger than 0.05). Then $\hat{Z}_0 = \sum_{i\in s}(m_i - n_i)\bar{y} + \bar{y}\sum_{i\notin s}\hat{\beta}x_i$ and $\hat{Z}_\theta = \sum(\hat{m}_i - n_i)\bar{y} + \bar{y}\sum_{i\notin s}\hat{\beta}x_i$. In this case, $R = \alpha\bar{y}\sum_{i\in s}(\hat{\beta}x_i - m_i)$ is of order $1/Nn$ and $\hat{T}_0$ is approximately uniformly optimal for large Nn, which is practically always the case. In fact, it can be shown that $V(\hat{T}_0 - T) - V(\hat{T}_\theta - T) = \frac{\alpha}{n}\sigma^2\tau^2\left(\sum_{i\in s} v(x_i) - X_s^2/w_s\right)$ which is of order $1/n^2$ and $\{V(\hat{T}_0 - T) - V(\hat{T}_\theta - T)\}/V(\hat{T}_\theta - T)$ is typically neglible even for moderate sized n. When $v(x) = x$, $\hat{Z}_\theta = \hat{Z}_0$ and $\hat{T}_0$ is exactly uniformly optimal.

Proof of Theorem. By Lagrange's method we shall minimize

$$F = Q - 2\lambda_1\sum_{i\in s} c_i - 2\lambda_2\left(\sum_{i\in s}\sum_{j\in s} c_{ij}x_j - X\right) .$$

The equations for partial derivatives are as follows:

$$\partial F/\partial c_i = 0 \Leftrightarrow c_i + \beta\sum_{k\in s} c_{ik}x_k = \phi_i(\rho\beta x_i + \lambda_1^*), \text{ with } \lambda_1^* = \lambda_1/\tau^2, \text{ for } i \in s.$$

Summing over $i \in s$ gives

$$\lambda_1^* = \frac{\beta(X - \rho\sum_s x_i\phi_i)}{\phi_s} .$$

Hence:

$$c_i = \rho\beta x_i\phi_i + \phi_i\frac{\beta(X - \rho\sum_s x_i\phi_i)}{\phi_s} - \beta\sum_{k\in s} c_{ik}x_k \tag{25}$$

For $j \ne i$, with $c_{\circ j} = \sum_{k\in s} c_{kj}$:

$$\partial F/\partial c_{ij} = 0 \Leftrightarrow c_{ij} = \frac{\phi_i\mu^2}{\tau^2}(1 - c_{\circ j}) - \frac{\phi_i\beta^2 x_j}{\phi_s\sigma^2 v(x_j)}(X - \rho\sum_s x_k\phi_k) + \phi_i\frac{x_j}{\tau^2\sigma^2 v(x_j)}\lambda_2 \tag{26}$$

$$\partial F/\partial c_{ii} = 0 \Leftrightarrow c_{ii} = \rho\phi_i + \frac{\phi_i\mu^2}{\tau^2}(1 - c_{\circ i}) - \frac{\phi_i\beta^2 x_i}{\phi_s\sigma^2 v(x_i)}(X - \rho\sum_s x_k\phi_k) + \phi_i\frac{x_i}{\tau^2\sigma^2 v(x_i)}\lambda_2 \tag{27}$$

From (26) and (27) we can determine $c_{\circ j}$ as a function of λ_2 :

$$c_{\circ j} = 1-\alpha+\alpha[\rho\phi_j - \frac{\beta^2 x_j}{\sigma^2 v(x_j)}(X-\rho\sum_s x_k\phi_k) + \frac{x_j}{\tau^2\sigma^2 v(x_j)}\phi_s\lambda_2].$$

Since $\sum_s c_{\circ j}x_j = X$, we find $\lambda_2^* = \lambda_2/(\tau^2\sigma^2)$:

$$\lambda_2^* = \frac{1}{\phi_s W_s}[\frac{X-(1-\alpha)X_s}{\alpha} + \frac{\beta^2}{\sigma^2}W_s(X-\rho\sum_s x_k\phi_k) - \rho\sum_s x_k\phi_k] \tag{28}$$

implying that

$$c_{\circ j} = 1-\alpha+\frac{x_j}{v(x_j)W_s}(X-(1-\alpha)X_s) + \alpha\rho\phi_j - \frac{x_j}{v(x_j)W_s}\alpha\rho\sum_s x_k\phi_k \tag{29}$$

From (26), (28) and (29) it follows that for $j \neq i$,

$$c_{ij} = \frac{\phi_i}{\phi_s}[(1-\alpha)(1-\rho\phi_j) + \frac{x_j}{v(x_j)W_s}(X-(1-\alpha)X_s) - \alpha\rho\frac{x_j}{v(x_j)W_s}\sum_s x_k\phi_k] \tag{30}$$

and, from (27),

$$c_{ii} = \rho\phi_i + \frac{\phi_i}{\phi_s}[(1-\alpha)(1-\rho\phi_i) + \frac{x_i}{v(x_i)W_s}(X-(1-\alpha)X_s) - \alpha\rho\frac{x_i}{v(x_i)W_s}\sum_s x_k\phi_k] \tag{31}$$

From (30) and (31) we find that

$$\sum_{k\in s} c_{ik}x_k = \frac{\phi_i}{\phi_s}[X-\rho\sum_s x_k\phi_k] + \rho\phi_i x_i$$

and using (25) we see that $c_i = 0$. Then from (30) & (31), the optimal predictor at θ equals

$$\hat{T}_\theta =$$

$$\sum_{i\in s}\bar{y}_i(\rho\phi_i m_i + \frac{\phi_i}{\phi_s}\sum_{j\in s}[(1-\alpha)(1-\rho\phi_j) + \frac{x_j}{v(x_j)W_s}(X-(1-\alpha)X_s) - \alpha\rho\frac{x_j}{v(x_j)W_s}\sum_s x_k\phi_k]m_j)$$

$$= \sum_{i\in s}\bar{y}_i w_i m_i + \hat{\mu}_\rho\sum_{j\in s}[(1-\alpha)(1-w_j) + \frac{x_j}{v(x_j)W_s}(X-(1-\alpha)X_s) - \alpha\frac{x_j}{v(x_j)W_s}\sum_s x_k w_k]m_j$$

$$= \sum_{i\in s}\bar{y}_i w_i m_i + \sum_{j\in s}(1-\alpha)m_j(1-w_j)\hat{\mu}_\rho + \hat{\beta}(X_{\bar{s}} + \alpha X_s)\hat{\mu}_\rho - \alpha\hat{\beta}\hat{\mu}_\rho\sum_s x_k w_k$$

$$= \sum_{i\in s}\bar{y}_i w_i m_i + \sum_{i\in s}[(1-\alpha)m_i + \alpha\hat{\beta}x_i](1-w_i)\hat{\mu}_\rho + \hat{\mu}_\rho\sum_{i\notin s}\hat{\beta}x_i$$

$\Rightarrow$

$$\hat{T}_\theta = \sum_{i\in s}(\hat{m}_i(1-w_i)\hat{\mu}_\rho + m_i w_i\bar{y}_i) + \hat{\mu}_\rho\sum_{i\notin s}\hat{\beta}x_i \ . \ \blacklozenge$$

4. Prediction intervals based on predictive likelihood

4.1 Coverage measures and simulation set-up

We consider model (2) under the normality assumptions in Section 3.2. It can be shown that, conditional on y, $(Z - E_\theta(Z \mid y))/\sqrt{V_\theta(Z \mid y)}$ is asymptotically $N(0,1)$ as $N - n_0 \to \infty$ provided the x_i's are bounded as $N - n_0 \to \infty$. Hence $Z|y$ is approximately normal for large $N - n_0$, and the (1-α) predictive interval given by (9) becomes

$$E_{\hat{\theta}}(Z \mid y) \pm u_{\alpha/2}\sqrt{V_p(Z)}$$

where $u_{\alpha/2}$ is the upper $\alpha/2$ -point in $N(0,1)$. This amounts to regarding $N(E_p(Z), V_p(Z))$ as a predictive distribution for Z. $V_p(Z)$ equals (19) if the interval is based on $L_p(z, m(\bar{s}) \mid y)$, while $L_{p,c}$ has (19) with $h(5)$ and $L_{p,mp}$ has (19) with $h(4)$. Let us denote these prediction intervals by I_p, I_{pc} and I_{mp}. Clearly $I_p \subset I_{mp} \subset I_{pc}$.

For large n_0 there is practically no difference between these intervals. However, for small n_0 they do differ. To find out how the intervals perform a comprehensive simulation study is undertaken. The prediction intervals are evaluated by four different measures, (i) the model-based coverage, (ii) the design-based coverage, (iii) the unconditional coverage, and (iv) the conditional coverage given the data and the guarantee of conditional coverage 1- α. The unconditional coverage is the expected design-based coverage and is a measure of how the prediction interval does as a *method* in long run behaviour in repeated surveys, when regarding the population distribution as a model for how the y-variable varies over time. This, of course, has been the standard justification for the design-based coverage, but it is not a correct interpretation. Rather, since the design-based coverage is for fixed y-values, it only measures the coverage of the interval in *hypothetical* repeated surveys with fixed y-values. It can not be used as a measure of coverage in the long run. The precise definitions of the various coverage measures for a prediction interval $I(y,\mathbf{s})$ for the values z of Z are as follows:

I. The model-based coverage $C_m = P_m(Z \in I(Y,\mathbf{s}) \mid \mathbf{s})$, over the joint distribution of (Y,Z)
II. The design-based coverage $C_d = P_d(z \in I(y,\mathbf{S}) \mid y,z)$, over the sampling design, regarding the total sample $\mathbf{S}$ as the stochastic variable.
III. The unconditional coverage $C = P(Z \in I(Y,\mathbf{S}))$ $(= E(C_d) = E(C_m))$

IV. The conditional coverage $C(\theta \mid y,\mathbf{s}) = P(Z \in I(y,\mathbf{s}) \mid y,\mathbf{s})$
and the guarantee of conditional coverage 1 - $\alpha = P(C(\theta \mid Y,\mathbf{s}) \geq 1-\alpha \mid \mathbf{s})$.

The two-stage sampling plan used in the simulation study is as follows:

1. At stage 1, n_0 clusters are drawn proportional to the x_i's, using the S-Plus function sample().
2. At the second stage, simple random sampling is used with equal sample sizes for each selected cluster.

As mentioned in Section 1, this is a commonly used sampling plan when the cluster sizes are unknown before sampling leading to approximately equal selection probabilities for the units. For the simulation study we assume that $v(x) = x$ in the model for the M_i's.

Note. An alternative confidence interval for Z is obtained by using an estimate $\hat{V}(\hat{Z}_0 - Z)$ of $Var(\hat{Z}_0 - Z)$ instead of $V_p(Z)$. Let $\hat{V}(\hat{Z}_0 - Z)$ be obtained by essentially replacing the unknown parameters in $Var(\hat{Z}_0 - Z)$ with their estimates. For the sampling plan and model used in the simulation study it can be shown that, for known ρ, $V_p(Z) > \hat{V}(\hat{Z}_0 - Z)$ and the relative difference can be substantial especially when the proportion $f_s = X_s/X$ is small. Hence such an interval will always give lower coverage than the predictive interval for all four coverage measures. As an example, consider the case where $n_0 = 10$, $n_i = 20$, $N = 110$, $X = 10^4$ and $f_s = 0.1$.Then the square root of the ratio $V_p(Z)/\hat{V}(\hat{Z}_0 - Z)$ is at least 1.11 when $\rho = .1$, $\hat{\tau}^2 / \hat{\mu}^2 = .2$ and $\hat{\sigma} / \hat{\beta} = 1$.

The approximations to $L_{m(\bar{s}),p}$ and to the distribution of Z given y are not valid for small n_0 and small N- n_0. The simulation study considers therefore mainly moderate and large n_0 and N- n_0. Table 1 describes the set-up. The N= 400- case corresponds roughly to a population of the size of Norway with about 400 municipalities. The four coverage measures are considered for a range of parameter values.

Table 1. Moderate and large size cases for the simulation study. N is the number of clusters in the population, n_0 is the number of clusters in the sample, and x_i is the size measure for cluster i . The cluster sample size n_i is equal to 20, and in the variance model for the cluster sizes $v(x) = x$.

$n_i = 20,\ v(x_i) = x_i$

$N = 50$ $n_0 = 10, 40$		$N = 400$ $n_0 = 10, 40, 100$	
i	x_i	i	x_i
1-7	100	1-50	1000
8-19	500	51-180	2000
20-26	1000	181-260	4000
27-40	1500	261-330	6000
41-45	2000	331-365	10000
46-50	5000	366-385	50000
		386-395	100000
		396-400	250000

The chosen parameter values for the simulation cases in Table 1 are presented in Table 2. Since x_i can be regarded as a preliminary estimate of the actual size m_i of cluster i, the regression coefficient β equals 1 in most of the simulations. Also, we consider y to be essentially positive valued. Then, at least, $\mu - 3\tau \geq 0$. Hence, we let the maximum value of the coefficient of variation τ/μ to be 1/3. To avoid negative m_i in the simulations, we shall assume that $\beta x_i - 4.5\sigma\sqrt{x_i} \geq 0$, i.e., $\sigma / \beta \leq \sqrt{x_i} / 4.5$. With $\beta = 1,\ \sigma \leq \sqrt{x_i} / 4.5$. Hence, for the case of $N = 50,\ \sigma \leq 2$, and for the case of $N = 400,\ \sigma \leq 7$.

Table 2. Parameter values in the simulation study

$1-\alpha$	.95	
β	.8 ,1, 1.2	
σ/β	1, 1.5, 2	if $N = 50$
	1, 4, 7	if $N = 400$
μ	3, 6	
τ/μ	1/50, 1/30, 1/10, 1/6, ¼, 1/3	
ρ	0, .001, .005, .01, .02, .03, .04, .05, .1, .5, .9	

Comparisons of the three prediction intervals are presented only for unconditional coverage C. All other presentations for the cases in table 1 (as well as Table 5) concern I_{mp}. There will also be a

broader range of parameter values for C than for the other coverage measures. Regarding the purely model-based C_m and design-based C_d, one may be unfortunate with **s**, for C_m, or the simulated population for C_d for some of the chosen values of the parameters. Therefore, these measures are not used as much as C when coverage dependency on the parameter values is studied.

4.2 Simulation results for the unconditional coverage

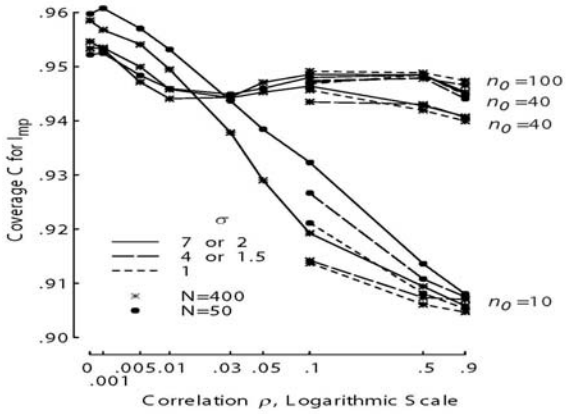
The three intervals I_p, I_{mp} and I_{pc} differ only in the component $h(k)$ such that $I_p \subset I_{mp} \subset I_{pc}$. The term $h(k)$ contributes less than 1/1000 of the interval width, and decreases much more rapidly than $V_p(Z)$ as a function of n_0. This causes a decreasing difference between the intervals as n_0 increases, and for large n_0 there are practically no differences between the three prediction intervals, as seen in Table 3. The estimated values of C are based on 100,000 repetitions of simulating population values and drawing the two-stage sample.

Table 3. The unconditional coverage C for I_p, I_{mp} and I_{pc}. Parameter values: μ=3, τ=1, β=1, ρ = .5 .

	N=50, σ=2		N=400, σ=7		
Prediction interval	n_0 =10	n_0=40	n_0=10	n_0 =40	n_0=100
I_p	.9110	.9483	.9084	.9426	.9484
I_{mp}	.9136	.9485	.9094	.9428	.9485
I_{pc}	.9155	.9486	.9101	.9428	.9485

From Figure 1, we see that for moderate and large n_0 (≥ 40) over the range $.05 \leq \rho \leq .9$ and different values of σ, C is constant slightly below the nominal confidence level 1-α = .95, while for small n_0 (= 10) C is decreasing with ρ and increasing with σ. We note that, in practice, a small correlation is not unusual. When ρ is very close to 0, C is slightly larger than .95. This is due to the fact that ρ will be over-estimated when close to 0, causing wider intervals since $V_p(Z)$ is increasing as a function of $\hat{\rho}$.

Figure 1. The unconditional coverage *C* for I_{mp} as a function of ρ. Parameter values: $\beta = 1$, $\mu=3$ and $\tau/\mu=1/3$.



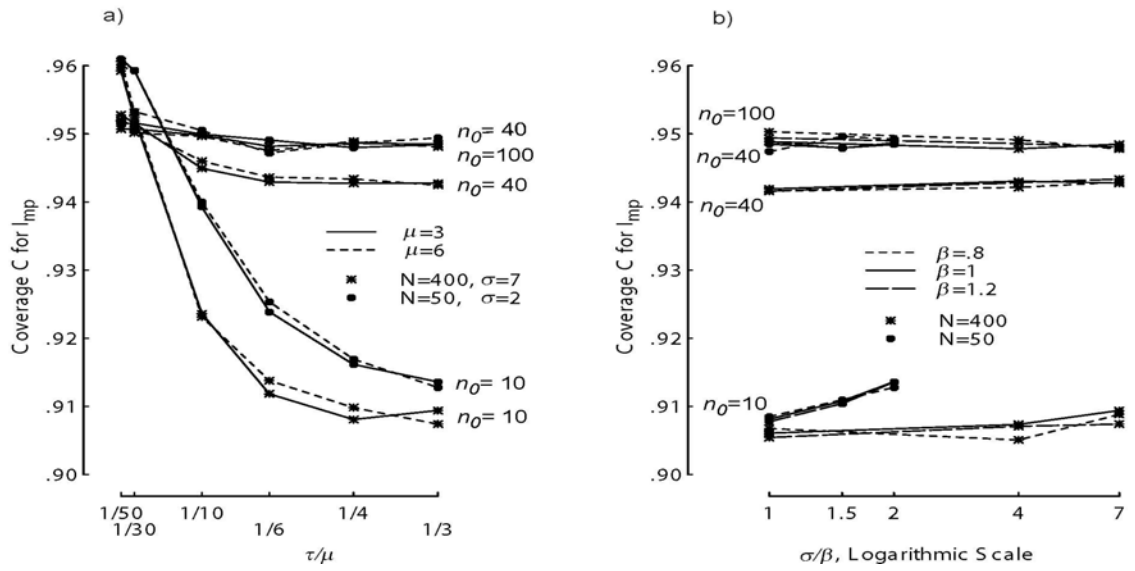
Let us now consider more closely *C* as a function of n_0 and *N*. Except for extremely small values of ρ it increases as a function of n_0, as seen in Figure 1 (as well as Figures 2,3). It seems also to be slightly connected with *N* as described in the function $.95(1-\frac{N-n_0}{N}\cdot\frac{1}{n_0^{(1+1/3)}})$. The similarity is demonstrated in Table 4. The relationship between *C* and (n_0, *N*) was found by fitting a linear regression to a transformation of *C*.

Table 4. The unconditional coverage *C* for I_{mp}. Parameter values: $\mu=3$, $\tau=1$, $\rho=0.5$, $\beta=1$, $\sigma=2$ or 7

C	.94845	.94847	.94551	.94276	.94084	.93089	.92822	.91358	.90937
n_0	40	100	60	40	25	20	15	10	10
N	50	400	400	400	50	400	50	50	400
$.95(1-\frac{N-n_0}{N}\cdot\frac{1}{n_0^{(1+1/3)}})$	.94861	.94847	.94656	.94375	.94350	.93338	.93202	.91472	.90701

Figure 2 shows that *C* depends on μ and τ only for small samples and then through τ/μ with higher values when τ/μ is extremely small (≅.02). This is also found in Figure 4 below for very small n_0 and *N*. Also from figure 2, we see that *C* seems to be constant over σ/β for large populations, when the sample size is moderate or large n_0 (≥ 40), and there is a slight increase in *C* with σ/β when $N = 50$ and $n_0 = 10$.

Figure 2. The unconditional coverage *C* for I_{mp} as a function of τ/μ, when $\beta = 1$, in 2a), and σ/β, when $\mu = 3$, $\tau = 1$, in 2b). The value of ρ equals .5.

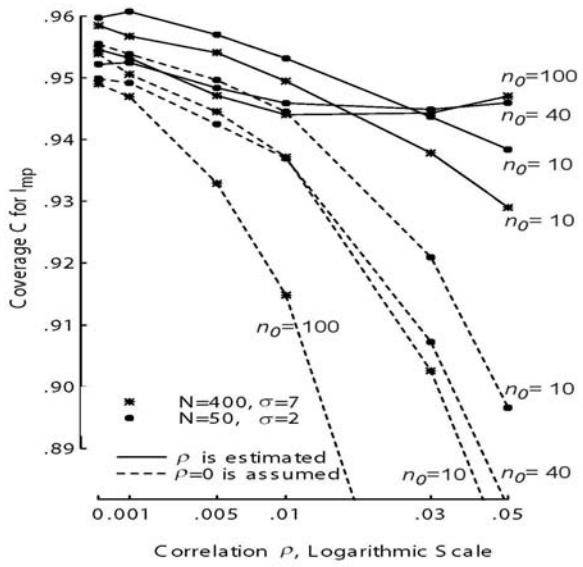


To summarize the main features of the simulation results presented in Table 3 and Figures 1,2: The coverage *C* which is the expected design-based coverage depends on the parameters basically through the coefficients of variation σ/β and τ/μ and the intraclass correlation coefficient ρ. For $n_0 \geq 40$ and $N \geq 50$ the prediction intervals has *C* essentially constant in the parameters and approximately equal to the nominal confidence level 1- α. Hence, one may say that the prediction intervals are calibrated with respect to long-run behaviour.

For extremely small ρ the parameter will be over-estimated as mentioned above. The effect of over-estimation and the necessity of a correlation parameter in the model is closely studied and presented in Figure 3. It seems that for $\rho \leq .01$, one may simplify the model by assuming $\rho = 0$. We note that $C > .9$ for $\rho \leq .04$ for the case of $n_0 = 10$ and $N = 50$. In general, it seems that the loss of information provided by the prediction interval with this simplifying assumption is the least for small samples/small populations. To interpret these values of ρ, it may be helpful to express the model (2) as in (3). We can express the ratio of within to between variability as $\tau_w^2 / \tau_b^2 = (1-\rho)/\rho$. For example, when $\rho = .01$ this ratio is equal to 99 and $\tau_w / \tau_b = \sqrt{99} = 9.95$. That is, the value $\rho = .01$ means that the variability within the clusters is about 10 times larger than between the clusters. If this is a reasonable assumption, we may simplify the model by letting $\rho = 0$. We also see that $\rho \leq .04$

corresponds to $\tau_w / \tau_b \geq \sqrt{24} = 4.90$. In general, if $\tau_w >> \tau_b$ then $\rho \approx 0$ and if $\tau_w << \tau_b$, then $\rho \approx 1$. For these two extreme values of ρ, we have of course also $\rho = 0 \Leftrightarrow \tau_b = 0$ and $\rho = 1 \Leftrightarrow \tau_w = 0$.

Figure 3. The unconditional coverage *C* for I_{mp} as a function of ρ, $0 \leq \rho \leq .05$, for the two cases i) $\rho = 0$ is assumed and ii) ρ is estimated. Parameter values: $\mu = 3$, $\tau = 1$, $\beta = 1$.



Since the prediction intervals are based on asymptotic considerations, we do not expect them to be calibrated with respect to the coverage measures for small n_0 and small N. Still, it is of interest to get an idea of how the coverage properties of the different intervals are in this case. To do this we consider C for the set-up in Table 5.

Table 5. Small size cases for the simulation study. N is the number of clusters in the population, n_0 is the number of clusters in the sample, and x_i is the size measure for cluster i. In case a, the second stage sample sizes are all equal to 3 or all equal to 10. In case b, the second stage sample sizes are all equal to 10 or all equal to 400.

$N = 10$, $n_0 = 6$, $v(x_i) = x_i$			
Case a $n_i = 3, 10$		*Case b* $n_i = 10, 400$	
i	x_i	i	x_i
1 -3	50	1 - 3	5000
4, 5	30	4, 5	3000
6 – 8	100	6 - 8	10000
9, 10	50	9, 10	5000

Estimated values of C are presented in Figures 4 and 5, and are based on 10,000 repetitions of simulating population values and drawing the two-stage sample. It seems clear that I_p is too short generally while for I_{mp} it is only for rather extremely small values of τ/μ or ρ or large values of σ/β that the coverage is close to the nominal confidence level, and in those cases I_{pc} is typically too wide. As a general impression for all three intervals, it seems that the coverage C is about 95% of the nominal level for these small size cases.

Figure 4. The unconditional coverage C for I_p, I_{mp} and I_{pc}. N = 10, n_0 = 6, ρ = .5. In a) as a function of τ/μ when β=1, σ=1. In b) as a function of σ/β when μ =3, τ=1.

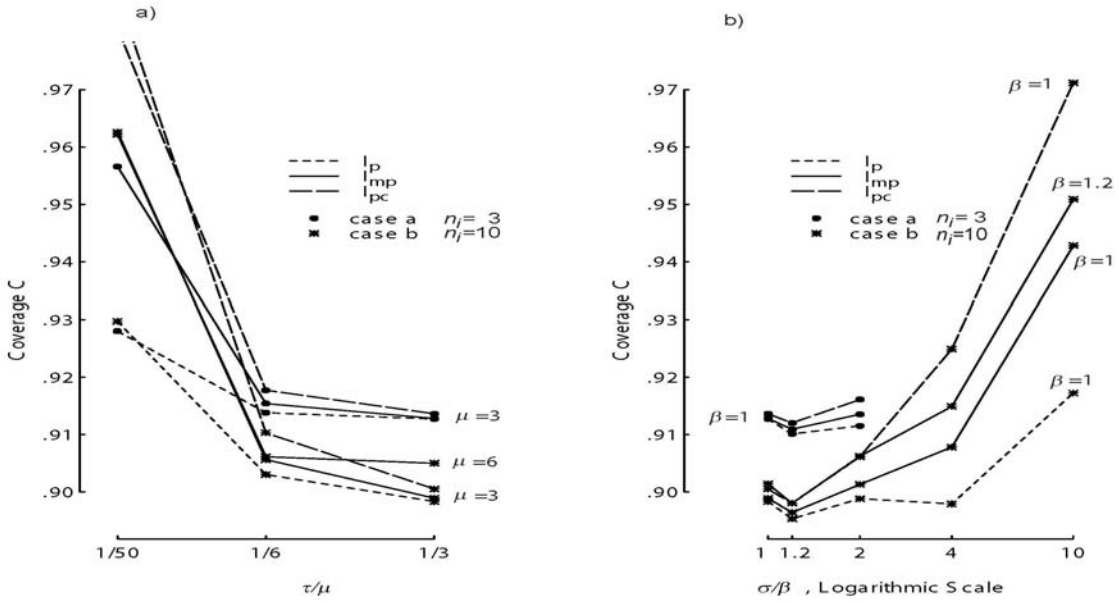
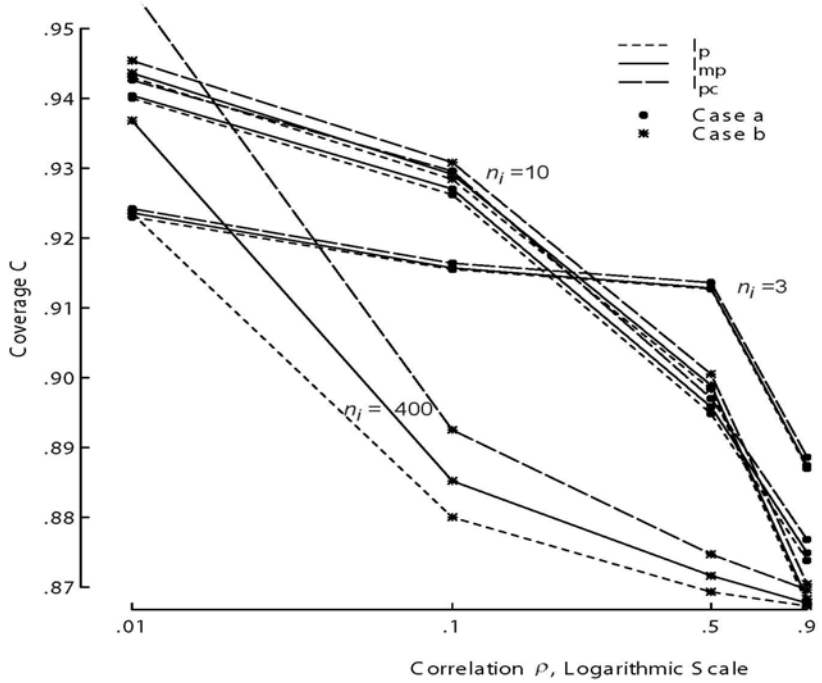


Figure 5. The unconditional coverage C for I_p, I_{mp} and I_{pc}. N = 10, n_0 = 6. As a function of ρ when β=1, σ=1, μ=3, τ=1.

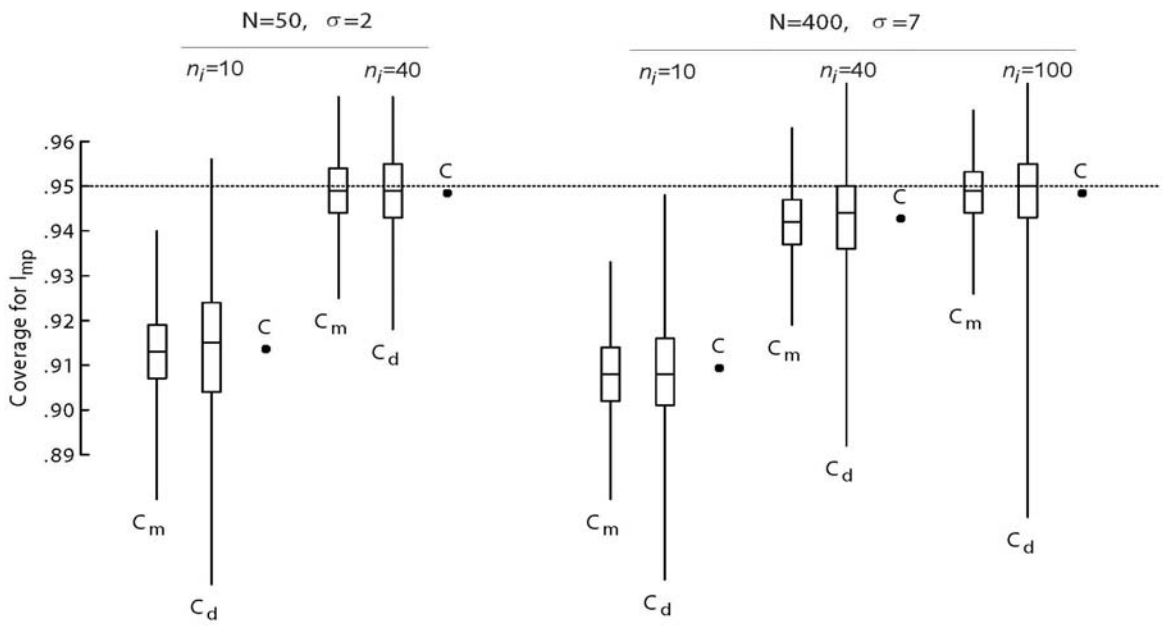


The statistical computing package S-PLUS 6 for UNIX is used in the simulation study and for creating the graphical displays.

4.3 The coverage measures C_m, C_d, $C(\theta \mid y,\mathbf{s})$, $P(C(\theta \mid Y,\mathbf{s}) \geq 1-\alpha \mid \mathbf{s})$

We shall consider the moderate and large sample cases in Table 1, with parameter values as presented in Table 2. Box-plot of C_m -estimates from 1000 two-stage samples of size n_0 according to the sampling plan described in Section 4.1 are shown in Figure 6. C_d - estimates from 1000 populations are also presented as a Box-plot in Figure 6, showing more variability than C_m. Finally, the "joint" coverages C over sample distribution and population distribution of Y_{ij}, M_i in Table 3 are added to Figure 6. Since $C = E(C_m) = E(C_d)$, the median values in the box plots and C all estimate approximately the same coverage probability. As expected then, there is very little difference between these values in figure 6. We note that the variability in the coverages decreases as n_0 increases, but is still very large, especially for C_d when n_0 = 100 and N = 400 compared to n_0 = 40 and N = 50. Also, all mean values are less than .95, although just very slightly for large sample/large population case.

Figure 6. Box-plots of 1000 C_d - and C_m estimates for I_{mp} with $\beta = 1$, $\mu = 3$, $\tau = 1$, $\rho = .5$.



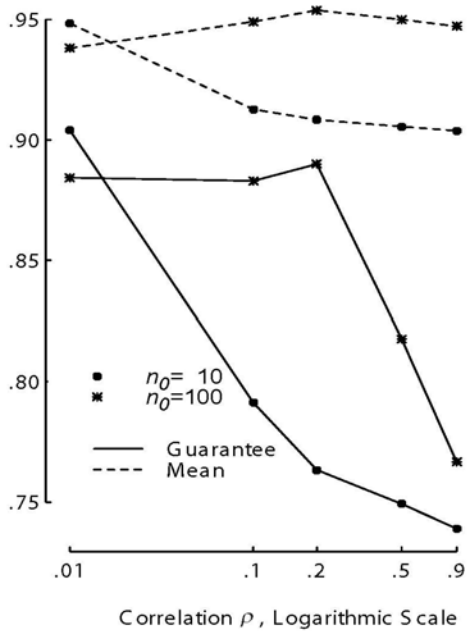
Next we consider the conditional coverage $C(\theta \mid y,\mathbf{s}) = P(Z \in I(y,\mathbf{s}) \mid y,\mathbf{s})$ and the guarantee of conditional coverage .95, $P(C(\theta \mid Y,\mathbf{s}) \geq .95 \mid \mathbf{s})$. For each case in Table 6 a new sample **s** is drawn according to the two-stage sampling plan in Section 4.1 and Z is simulated 1000 times, from its conditional distribution given the data with the same parameters as for the sampled data, to estimate $C(\theta \mid y,\mathbf{s}) = P_\theta(Z \in (y,\mathbf{s}) \mid y,\mathbf{s})$. This is repeated 1000 times, with respect to the distribution of Y to estimate $P_\theta(C(\theta \mid Y,\mathbf{s}) \geq 1-\alpha \mid \mathbf{s})$. For each case, the simulation is done three times with different **s** to see how much the results may vary for different **s**.

Table 6. The conditional coverage and guarantee of conditional coverage .95 for I_{mp}. (1000 Z are drawn using true parameter values, for each of 1000 samples), $\mu = 3$, $\beta = 1$, $\rho = .5$.

Mean $C(\theta \mid y, \mathbf{s})$	$P(C(\theta \mid Y, \mathbf{s}) \geq .95 \mid \mathbf{s})$	τ / μ	σ / β	N	n_0
.949, .942, .948	.863, .851, .775	1/3	2	400	100
.956, .954, .950	.865, .783, .828	1/3	4	400	100
.948, .943, .948	.814, .756, .787	1/3	7	400	60
.944, .944, .944	.734, .749, .776	1/3	7	400	40
.926, .928, .937	.726, .712, .784	1/3	7	400	20
.948, .948, .948	.844, .813, .782	1/3	2	50	40
.947, .942, .939	.731, .727, .709	1/3	2	50	25
.927, .931, .927	.684, .703, .688	1/3	2	50	15
.916, .903, .906	.687, .659, .663	1/3	2	50	10
.953, .950, .949	.729, .700, .702	1/50	7	400	100
.953, .952, .952	.740, .727, .728	1/30	7	400	100
.953, .951, .948	.809, .804, .787	1/10	7	400	100
.952, .956, .946	.813, .853, .800	1/6	7	400	100
.955, .950, .946	.827, .834, .803	1/4	7	400	100

Figure 7 shows graphs of the mean conditional coverage and the guarantee of conditional coverage .95 as a function of ρ. The simulations are done as for Table 6, and the graphs are based on the average of ten simulations with different **s**.

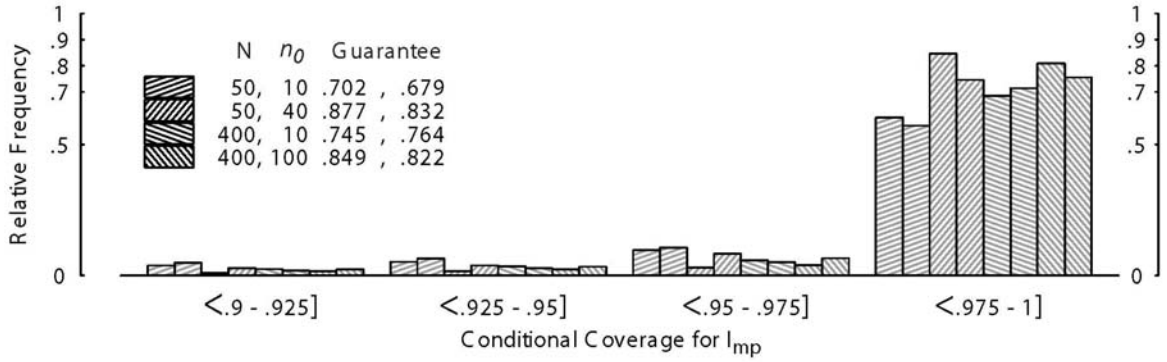
Figure 7. The guarantee of conditional coverage .95 and the mean conditional coverage for I_{mp} as functions of ρ. N = 400, μ =3, τ = 1, β = 1, σ = 7.



For constant parameter values, the guarantee of conditional coverage .95 is increasing in n_0 for a fixed N, except when ρ is close to 0. We also see that the guarantee of 95% conditional coverage is much less than 95%. This is to be expected of prediction intervals with unconditional coverage C equal to .95, as mentioned in Bjørnstad (1990). If we want 95% conditional coverage 95% of the times, C needs to be much larger than .95. Table 6 and Figure 7 shows that the guarantee essentially decreases with ρ. It also seems that the guarantee is increasing in τ/μ for small values until it stabilizes. We note that the mean $C(\theta \mid y, \mathbf{s})$ is an estimate of C_m and shows a similar pattern as C (in Figures 1,2) when N = 400 as function of ρ and τ/μ.

The conditional coverage is stochastic in Y, and to get an idea of how the *distribution* of the conditional coverage is compared to the *probability* of being larger than .95, Figure 8 shows simulated histograms of 4 different cases of (N,n_0) together with the guarantee. The simulation is done in the same way as for Table 6. For each case, the simulation is performed twice with **s** selected according to the two-stage design in Section 4.1. We see that, except when N = 50, n_0 = 10, the conditional coverage is larger than .975 at least 70% of the time. It is also possible to show with repeated simulations of 1000 Z-values that the distribution of the conditional coverage fits well a beta distribution.

Figure 8. Histogram of conditional coverage for I_{mp}. Parameter values: $\mu = 3, \tau = 1, \beta = 1, \sigma = 2$ when N = 50, σ = 7 when N = 400, ρ = .5.



In practice we do not know, of course, the true values of the parameters. Instead we may derive the conditional level and the guarantee at estimated parameter values. Then it is useful to have an idea on how these measures perform within one standard error (*s.e.*) of the true values, and also at the estimated values. Table 7 presents *s.e.* for the different parameters, from the simulation used to derive table 6, based on 1000 draws of the data $Y=y$.

Table 7. Estimated standard errors for the estimates of parameter values based on 1000 repetitions

N	n_0	$\mu = 3$	$\tau = 1$	$\beta = 1$	$\sigma = 7$ *or* 2	$\rho = .5$
400	100	.073	.039	.0038	.507 (σ=7)	.040
400	40	.114	.059	.0046	.768 (σ=7)	.061
400	10	.230	.115	.0076	1.517 (σ=7)	.127
50	40	.114	.061	.0076	.223 (σ=2)	.062
50	10	.237	.112	.0136	.443 (σ=2)	.124

To get an indication of how $C(\theta \mid y, \mathbf{s})$ will behave as function of θ within 1 *s.e.* of the true values, table 8 presents results where the parameters are changed one at a time. Also in Table 8 are included the results for the guarantee of estimated conditional coverage. Here, Z is drawn 1000 times from the conditional distribution given the data with parameter θ^* different from the true θ used to sample the data, to estimate $C(\theta^* \mid y, \mathbf{s}) = P_{\theta^*}(Z \in (y, \mathbf{s}) \mid y, \mathbf{s})$. This is repeated 1000 times, with respect to the true

distribution of Y to estimate $P_\theta(C(\theta^*|Y,\mathbf{s}) \geq 1-\alpha \,|\, \mathbf{s})$. For each case in Table 8 a new sample **s** is drawn three times according to the two-stage sampling plan in Section 4.1.

Table 8. The conditional coverage and guarantee of conditional coverage .95 for I_{mp}, when Z are drawn from a population with non-true parameter values. Standard error (s.e.) from table 7.
True values: μ=3, τ=1, β=1, σ=7 when N = 400, σ =2 when N = 50, ρ = .5

Mean $C(\theta^*\|y,\mathbf{s})$	$P_\theta(C(\theta^*\|Y,\mathbf{s}) \geq 1-\alpha \,\|\, \mathbf{s})$	N, n_0	*Parameter values* θ^* *for* Z
.949, .955, .958	.805, .827, .816	*400,100*	*true*
.919, .911, .917	.755, .792, .753	*400,100*	μ *=3+s.e., else true*
.948, .951, .949	.833, .845, .817	*400,100*	τ *=1+s.e., else true*
.947, .953, .944	.804, .831, .802	*400,100*	β *=1+s.e., else true*
.953, .954, .942	.841, .833, .800	*400,100*	σ *=7+s.e., else true*
.945, .944, .938	.829, .761, .781	*400,100*	ρ *= .5+s.e., else true*
.986, .981, .985	.921, .931, .928	*400,100*	$\hat{\theta}$
.998, .998, .995	.997, .998, .998	*400,40*	$\hat{\theta}$
.999998, .999994, .999996	1, 1, 1	*400,10*	$\hat{\theta}$
.953, .953, .959	.853, .843, .823	*50,40*	$\hat{\theta}$
.9997, .9997, .9997	1, 1, 1	*50,10*	$\hat{\theta}$

It seems that the guarantee is most sensitive to change in μ. From the simulated results of $E_\theta C(\hat{\theta}\,|\,Y,\mathbf{s})$ and $P_\theta(C(\hat{\theta}\,|\,Y,\mathbf{s}) \geq 1-\alpha \,|\, \mathbf{s})$ we also see that, except for the case N = 50, n_0 = 40, the estimated conditional coverages overestimate the true conditional coverages (see Table 6) by a large degree. It is therefore questionable how interesting this measure is in practice as regarding the coverage property of a given prediction interval.

It is of interest to see how the parameter estimates affect the true conditional coverage. In order to do this we performed a simulated logistic regression analysis for N = 400, n_0 = 100, based on 1000 simulated observations of Z for each of 1000 simulated Y-values. This gives us 1000 observations in the regression analysis and for each observation of Y, the number of Z-values in the given prediction interval as the dependent variable with explanatory variables being functions of parameter estimates $\hat{\mu}, \hat{\tau}$, etc., and. It turns out that $(\hat{\mu}-\mu)^2$ has a large negative effect, decreasing $C(\theta\,|\,y,\mathbf{s})$ to a high degree, while $C(\theta\,|\,y,\mathbf{s})$ seems to increase with increasing estimates of ρ and τ. The estimates

of β, σ do not seem to influence the conditional coverage significantly. When ρ is underestimated or $\hat{\mu}$ is very different from μ, $C(\theta \mid y, \mathbf{s})$ can be very low compared to the nominal level. Still, typically, the estimated conditional coverage $C(\hat{\theta} \mid y, \mathbf{s})$ will still be close to the nominal confidence level. Hence, as also noted from Table 8, $C(\hat{\theta} \mid y, \mathbf{s})$ may not be an informative measure on the conditional property of the prediction interval for a given data set. Therefore, a data-based measure of coverage should be a plot of $C(\theta \mid y, \mathbf{s})$ as a function of θ. This measure also satisfies the likelihood principle for prediction, as noted by Bjørnstad (1996).

4.4. A summary of the simulation study

We concentrate on the moderate and large size cases in Table 1. The three prediction intervals are compared using the unconditional coverage C and they achieve practically the same C-level, even for the moderate sample size of 200, i.e., when $n_0 = 10$. In this case C is typically around .91 when $\rho \geq .1$, 96% of the nominal level of .95. The coverage C is at least .94, when the sample size is at least 800 ($n_0 = 40$) and approximately .95 for the large sample case of 2000 ($n_0 = 100$) for either case of population size and through out the range of the intraclass correlation coefficient ρ. The only exception occurs when ρ is very close to 0. Then ρ is typically overestimated, leading to slightly larger prediction intervals than "necessary" with C ranging from .955 to .96. In general when ρ is close to 0, say less than .01 (i.e., the variability within the clusters is expected to be about 10 times larger than between the clusters) one may simplify the model buy assuming that $\rho = 0$ when deriving the prediction intervals.

The simulation results indicate that the unconditional coverage is essentially independent of all parameters except for ρ when n_0 is at least 40. For the case of $n_0 = 10$, C seems to depend on μ, τ and β, σ through the coefficients of variation τ/μ and σ/β, decreasing in τ/μ and slightly increasing in σ/β. Even though C is approximately .91 for most parameter values, when τ/μ is very small C achieves the nominal level of .95 also in this case.

Regarding the model-based coverage C_m, the box-plot shows that for N=50, $n_0 = 10$ it varies in the range .913 ± .006 for 50% of the selected samples, while for N=400, $n_0 = 10$ the similar range is .908 ± .006, for N=400, $n_0 = 40$ the range is .942 ± .005 and for N=400, $n_0 = 100$ & N=50, $n_0 = 40$ it is .949 ± .005. Hence, C_m shows approximately the same pattern as C. The interquartile ranges for the design-based coverage C_d are .915 ± .010, .908 ± .008, .944 ± .007 and .949 ± .006 for the same four cases respectively, showing a larger variability than C_m.

For large sample sizes, the coverage measures C, C_m, and C_d achieve approximately the nominal level 1-α in most cases. The coverage measures are slightly less than 1-α for moderately sized n_0, and about 95% of the nominal level for small n_0, except for $\rho \approx 0$. It therefore seems that the critical value $u_{\alpha/2}$ from $N(0,1)$ in the prediction intervals is slightly too small for small and moderate n_0. Regarding the interval I_{mp}, based on the predictive likelihood L_{mp}, we note that L_{mp} is a mixture of a normal distribution and a multivariate t-distribution with n_0 - 2 degrees of freedom. One alternative is therefore to use the upper $\alpha/2$-point in the $t(n_0\text{-}2)$-distribution instead of $u_{\alpha/2}$ when n_0 is not large, at least when n_0 is less than 40. Some preliminary simulations regarding C_m indicate that this choice may work well for small values of n_0.

The guarantee of conditional coverage 1 - α is essentially increasing in n_0 and decreasing in ρ. The conditional coverage is typically larger than .975 when the nominal level is .95, and guarantee of conditional coverage .95 varies from about .70 to .90 with largest values for ρ close to 0.

5. Concluding remarks

We have considered two-stage cluster sampling with unknown cluster sizes before sampling, deriving predictor and prediction intervals from a predictive likelihood. Optimality properties of the predictor and coverage properties of the prediction intervals have been studied.

Predictive likelihood is derived for the model (2) assuming that the random variables are normally distributed. Considering a predictive likelihood for the unobserved part Z of the population directly does not work, mainly because Z is a sum of a *stochastic* number of random variables. Therefore, predictor and prediction interval is obtained from a joint predictive likelihood for Z and the vector $M(\bar{s}) = (M_i)_{i \notin s}$. The predictor obtained from the predictive likelihood is $\hat{Z}_0 = E_{\hat{\theta}}(Z \mid y)$, where $\hat{\theta}$ is the vector of MLE for the parameters in the model.

Optimality theory for a class of model-unbiased predictors linear in the Y_{ij}'s, but not simultaneously in both Y_{ij}'s and the clusters sizes M_i's, under the distribution-free model (2) is developed. It is shown that the predictive likelihood-based $\hat{Z}_0$ (with the intraclass correlation ρ estimated by an ANOVA approach instead of MLE) is approximately uniformly optimal for large sample size and large number of clusters, in the sense of uniformly minimizing the mean square error in the class considered for the distribution-free model (2), generalizing results in Kelly and Cumberland (1990) and Valliant et al. (2000). A typical model for the M_i's is to let $v(x) = x$. If also the sample sizes at the second stage are

equal (the common case for the pps-srs sampling plan), then $\hat{Z}_0$ is uniformly optimal for known ρ. In particular this holds in the simplified model where it can be assumed that the intraclass correlation is negligible, i.e. we let $\rho = 0$.

Three prediction intervals for Z based on three similar predictive likelihoods are studied. They are all of the form $\hat{Z}_0 \pm u_{\alpha/2}\sqrt{V_p(Z)}$, where $u_{\alpha/2}$ is the upper $\alpha/2$-point of $N(0,1)$, and $V_p(Z)$ is the variance in the normalized predictive likelihood. For small number n_0 of sampled clusters they differ significantly, however, for large n_0 the three intervals are practically identical. A comprehensive simulation study for estimating confidence levels, both model-based and design-based is undertaken. The prediction intervals are evaluated by four different measures; the model-based coverage C_m, the design-based coverage C_d, the unconditional coverage C (expected design-based coverage), and the conditional coverage given the data and the guarantee of conditional coverage 1- α. Roughly, the simulation study indicates that for large sample sizes (about 2000), C,C_m and C_d achieve approximately the nominal level 1 - α and are slightly less than 1 - α for moderately large sample sizes (about 800) For small sample sizes (about 200) these coverage measures are about 95% of the nominal level.

References

Berger, J.O. and Wolpert, R.L. (1988): *The likelihood Principle (2nd Edition)*, IMS Lecture Notes-Monograph Series, Vol. 6.

Birnbaum, A. (1962): On the Foundations of Statistical Inference (with discussion), *Journal of the American Statistical Association*, **57**, 269-306.

Bjørnstad, J.F. (1990): Predictive Likelihood: A Review (with discussion), *Statistical Science*, **5**, 242-265.

Bjørnstad, J.F. (1996): On the Generalization of the Likelihood Function and the Likelihood Principle, *Journal of the American Statistical Association*, **91**, 791-806.

Bjørnstad, J.F. (1998): "Predictive Likelihood", in *Encyclopedia of Statistical Sciences* (ed. Kotz, S., Read, C.R., and Banks, D.L.), Update volume 2, 539-545.

Bolfarine, H. and Zacks, S. (1992): *Prediction Theory for Finite Populations*, Springer.

Butler, R.W. (1986): Predictive Likelihood Inference with Applications (with discussion), *Journal of the Royal Statistical Society,* Ser. B, **48**, 1-38.

Cassel, C.-M., Särndal, C.-E., and Wretman, J. (1977): *Foundations of Inference in Survey Sampling,* Wiley.

Hinkley, D.V. (1979): Predictive Likelihood, *The Annals of Statistics,***7**, 718-728 (corrig. **8**, 694).

Kelly, E.J. and Cumberland, W.G. (1990): Prediction Theory Approach to Multistage Sampling when Cluster Sizes are Unknown, *Journal of Official Statistics,* **6**, 437-449.

Mathiasen, P.E. (1979): Prediction Functions, *Scandinavian Journal of Statistics*, **6**, 1-21.

Royall, R.M. (1986): The Prediction Approach to Robust Variance Estimation in Two-Stage Cluster Sampling, *Journal of the American Statistical Association*, **81**, 119-123.

S-PLUS 6 for UNIX (1998-2000), Mathsoft, Inc., Seattle, WA.

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

Thomsen, I., Tesfu, D. and Binder, D.A. (1986): Estimation of design effects and intraclass correlations when using outdated measures of size, *International Statistical Review,* **54**, 343-349.

Thomsen, I. and Tesfu, D. (1988): On the use of models in sampling from finite populations, *Handbook of Statistics,* **6**, 369-397.

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

Recent publications in the series Discussion Papers

292 E. Biørn, K-G. Lindquist and T. Skjerpen (2000): Heterogeneity in Returns to Scale: A Random Coefficient Analysis with Unbalanced Panel Data

293 K-G. Lindquist and T. Skjerpen (2000): Explaining the change in skill structure of labour demand in Norwegian manufacturing

294 K. R. Wangen and E. Biørn (2001): Individual Heterogeneity and Price Responses in Tobacco Consumption: A Two-Commodity Analysis of Unbalanced Panel Data

295 A. Raknerud (2001): A State Space Approach for Estimating VAR Models for Panel Data with Latent Dynamic Components

296 J.T. Lind (2001): Tout est au mieux dans ce meilleur des ménages possibles. The Pangloss critique of equivalence scales

297 J.F. Bjørnstad and D.E. Sommervoll (2001): Modeling Binary Panel Data with Nonresponse

298 Taran Fæhn and Erling Holmøy (2001): Trade Liberalisation and Effects on Pollutive Emissions and Waste. A General Equilibrium Assessment for Norway

299 J.K. Dagsvik (2001): Compensated Variation in Random Utility Models

300 K. Nyborg and M. Rege (2001): Does Public Policy Crowd Out Private Contributions to Public Goods?

301 T. Hægeland (2001): Experience and Schooling: Substitutes or Complements

302 T. Hægeland (2001): Changing Returns to Education Across Cohorts. Selection, School System or Skills Obsolescence?

303 R. Bjørnstad: (2001): Learned Helplessness, Discouraged Workers, and Multiple Unemployment Equilibria in a Search Model

304 K. G. Salvanes and S. E. Førre (2001): Job Creation, Heterogeneous Workers and Technical Change: Matched Worker/Plant Data Evidence from Norway

305 E. R. Larsen (2001): Revealing Demand for Nature Experience Using Purchase Data of Equipment and Lodging

306 B. Bye and T. Åvitsland (2001): The welfare effects of housing taxation in a distorted economy: A general equilibrium analysis

307 R. Aaberge, U. Colombino and J.E. Roemer (2001): Equality of Opportunity versus Equality of Outcome in Analysing Optimal Income Taxation: Empirical Evidence based on Italian Data

308 T. Kornstad (2001): Are Predicted Lifetime Consumption Profiles Robust with respect to Model Specifications?

309 H. Hungnes (2001): Estimating and Restricting Growth Rates and Cointegration Means. With Applications to Consumption and Money Demand

310 M. Rege and K. Telle (2001): An Experimental Investigation of Social Norms

311 L.C. Zhang (2001): A method of weighting adjustment for survey data subject to nonignorable nonresponse

312 K. R. Wangen and E. Biørn (2001): Prevalence and substitution effects in tobacco consumption. A discrete choice analysis of panel data

313 G.H. Bjertnær (2001): Optimal Combinations of Income Tax and Subsidies for Education

314 K. E. Rosendahl (2002): Cost-effective environmental policy: Implications of induced technological change

315 T. Kornstad and T.O. Thoresen (2002): A Discrete Choice Model for Labor Supply and Child Care

316 A. Bruvoll and K. Nyborg (2002): On the value of households' recycling efforts

317 E. Biørn and T. Skjerpen (2002): Aggregation and Aggregation Biases in Production Functions: A Panel Data Analysis of Translog Models

318 Ø. Døhl (2002): Energy Flexibility and Technological Progress with Multioutput Production. Application on Norwegian Pulp and Paper Industries

319 R. Aaberge (2002): Characterization and Measurement of Duration Dependence in Hazard Rate Models

320 T. J. Klette and A. Raknerud (2002): How and why do Firms differ?

321 J. Aasness and E. Røed Larsen (2002): Distributional and Environmental Effects of Taxes on Transportation

322 E. Røed Larsen (2002): The Political Economy of Global Warming: From Data to Decisions

323 E. Røed Larsen (2002): Searching for Basic Consumption Patterns: Is the Engel Elasticity of Housing Unity?

324 E. Røed Larsen (2002): Estimating Latent Total Consumption in a Household.

325 E. Røed Larsen (2002): Consumption Inequality in Norway in the 80s and 90s.

326 H.C. Bjørnland and H. Hungnes (2002): Fundamental determinants of the long run real exchange rate:The case of Norway.

327 M. Søberg (2002): A laboratory stress-test of bid, double and offer auctions.

328 M. Søberg (2002): Voting rules and endogenous trading institutions: An experimental study.

329 M. Søberg (2002): The Duhem-Quine thesis and experimental economics: A reinterpretation.

330 A. Raknerud (2002): Identification, Estimation and Testing in Panel Data Models with Attrition: The Role of the Missing at Random Assumption

331 M.W. Arneberg, J.K. Dagsvik and Z. Jia (2002): Labor Market Modeling Recognizing Latent Job Attributes and Opportunity Constraints. An Empirical Analysis of Labor Market Behavior of Eritrean Women

332 M. Greaker (2002): Eco-labels, Production Related Externalities and Trade

333 J. T. Lind (2002): Small continuous surveys and the Kalman filter

334 B. Halvorsen and T. Willumsen (2002): Willingness to Pay for Dental Fear Treatment. Is Supplying Fear Treatment Social Beneficial?

335 T. O. Thoresen (2002): Reduced Tax Progressivity in Norway in the Nineties. The Effect from Tax Changes

336 M. Søberg (2002): Price formation in monopolistic markets with endogenous diffusion of trading information: An experimental approach

337 A. Bruvoll og B.M. Larsen (2002): Greenhouse gas emissions in Norway. Do carbon taxes work?

338 B. Halvorsen and R. Nesbakken (2002): A conflict of interests in electricity taxation? A micro econometric analysis of household behaviour

339 R. Aaberge and A. Langørgen (2003): Measuring the Benefits from Public Services: The Effects of Local Government Spending on the Distribution of Income in Norway

340 H. C. Bjørnland and H. Hungnes (2003): The importance of interest rates for forecasting the exchange rate

341 A. Bruvoll, T.Fæhn and Birger Strøm (2003): Quantifying Central Hypotheses on Environmental Kuznets Curves for a Rich Economy: A Computable General Equilibrium Study

342 E. Biørn, T. Skjerpen and K.R. Wangen (2003): Parametric Aggregation of Random Coefficient Cobb-Douglas Production Functions: Evidence from Manufacturing Industries

343 B. Bye, B. Strøm and T. Åvitsland (2003): Welfare effects of VAT reforms: A general equilibrium analysis

344 J.K. Dagsvik and S. Strøm (2003): Analyzing Labor Supply Behavior with Latent Job Opportunity Sets and Institutional Choice Constraints

345 A. Raknerud, T. Skjerpen and A. Rygh Swensen (2003): A linear demand system within a Seemingly Unrelated Time Series Equation framework

346 B.M. Larsen and R.Nesbakken (2003): How to quantify household electricity end-use consumption

347 B. Halvorsen, B. M. Larsen and R. Nesbakken (2003): Possibility for hedging from price increases in residential energy demand

348 S. Johansen and A. R. Swensen (2003): More on Testing Exact Rational Expectations in Cointegrated Vector Autoregressive Models: Restricted Drift Terms

349 B. Holtsmark (2003): The Kyoto Protocol without USA and Australia - with the Russian Federation as a strategic permit seller

350 J. Larsson (2003): Testing the Multiproduct Hypothesis on Norwegian Aluminium Industry Plants

351 T. Bye (2003): On the Price and Volume Effects from Green Certificates in the Energy Market

352 E. Holmøy (2003): Aggregate Industry Behaviour in a Monopolistic Competition Model with Heterogeneous Firms

353 A. O. Ervik, E.Holmøy and T. Hægeland (2003): A Theory-Based Measure of the Output of the Education Sector

354 E. Halvorsen (2003): A Cohort Analysis of Household Saving in Norway

355 I. Aslaksen and T. Synnestvedt (2003): Corporate environmental protection under uncertainty

356 S. Glomsrød and W. Taoyuan (2003): Coal cleaning: A viable strategy for reduced carbon emissions and improved environment in China?

357 A. Bruvoll T. Bye, J. Larsson og K. Telle (2003): Technological changes in the pulp and paper industry and the role of uniform versus selective environmental policy.

358 J.K. Dagsvik, S. Strøm and Z. Jia (2003): A Stochastic Model for the Utility of Income.

359 M. Rege and K. Telle (2003): Indirect Social Sanctions from Monetarily Unaffected Strangers in a Public Good Game.

360 R. Aaberge (2003): Mean-Spread-Preserving Transformation.

361 E. Halvorsen (2003): Financial Deregulation and Household Saving. The Norwegian Experience Revisited

362 E. Røed Larsen (2003): Are Rich Countries Immune to the Resource Curse? Evidence from Norway's Management of Its Oil Riches

363 E. Røed Larsen and Dag Einar Sommervoll (2003): Rising Inequality of Housing? Evidence from Segmented Housing Price Indices

364 R. Bjørnstad and T. Skjerpen (2003): Technology, Trade and Inequality

365 A. Raknerud, D. Rønningen and T. Skjerpen (2003): A method for improved capital measurement by combining accounts and firm investment data

366 B.J. Holtsmark and K.H. Alfsen (2004): PPP-correction of the IPCC emission scenarios - does it matter?

367 R. Aaberge, U. Colombino, E. Holmøy, B. Strøm and T. Wennemo (2004): Population ageing and fiscal sustainability: An integrated micro-macro analysis of required tax changes

368 E. Røed Larsen (2004): Does the CPI Mirror Costs.of.Living? Engel's Law Suggests Not in Norway

369 T. Skjerpen (2004): The dynamic factor model revisited: the identification problem remains

370 J.K. Dagsvik and A.L. Mathiassen (2004): Agricultural Production with Uncertain Water Supply

371 M. Greaker (2004): Industrial Competitiveness and Diffusion of New Pollution Abatement Technology – a new look at the Porter-hypothesis

372 G. Børnes Ringlund, K.E. Rosendahl and T. Skjerpen (2004): Does oilrig activity react to oil price changes? An empirical investigation

373 G. Liu (2004) Estimating Energy Demand Elasticities for OECD Countries. A Dynamic Panel Data Approach

374 K. Telle and J. Larsson (2004): Do environmental regulations hamper productivity growth? How accounting for improvements of firms' environmental performance can change the conclusion

375 K.R. Wangen (2004): Some Fundamental Problems in Becker, Grossman and Murphy's Implementation of Rational Addiction Theory

376 B.J. Holtsmark and K.H. Alfsen (2004): Implementation of the Kyoto Protocol without Russian participation

377 E. Røed Larsen (2004): Escaping the Resource Curse and the Dutch Disease? When and Why Norway Caught up with and Forged ahead of Its Neughbors

378 L. Andreassen (2004): Mortality, fertility and old age care in a two-sex growth model

379 E. Lund Sagen and F. R. Aune (2004): The Future European Natural Gas Market - are lower gas prices attainable?

380 A. Langørgen and D. Rønningen (2004): Local government preferences, individual needs, and the allocation of social assistance

381 K. Telle (2004): Effects of inspections on plants' regulatory and environmental performance - evidence from Norwegian manufacturing industries

382 T. A. Galloway (2004): To What Extent Is a Transition into Employment Associated with an Exit from Poverty

383 J. F. Bjørnstad and E.Ytterstad (2004): Two-Stage Sampling from a Prediction Point of View