Social Statistics Research Centre
University of Southampton

Variance Estimation for a Low-Income Proportion

Yves G. Berger and Chris J. Skinner

Abstract

Proportions below a given fraction of a quantile of an income distribution are often estimated from survey data in poverty comparisons. We consider the estimation of the variance of such a proportion, estimated from Family Expenditure Survey data. We show how a linearization method of variance estimation may be applied to this proportion, allowing for the effects of both a complex sampling design and weighting by a raking method to population controls. We show that, for 1998-99 data, the estimated variances are always increased when allowance is made for the design and raking weights, the principal effect arising from the design. We also study the properties of a simplified variance estimator and discuss extensions to a wider class of poverty measures.

SSRC Methodology Working Paper M03/03

Variance Estimation for a Low-Income Proportion

Yves G. Berger and Chris J. Skinner

University of Southampton, UK

Summary. Proportions below a given fraction of a quantile of an income distribution are often estimated from survey data in poverty comparisons. We consider the estimation of the variance of such a proportion, estimated from Family Expenditure Survey data. We show how a linearization method of variance estimation may be applied to this proportion, allowing for the effects of both a complex sampling design and weighting by a raking method to population controls. We show that, for 1998-99 data, the estimated variances are always increased when allowance is made for the design and raking weights, the principal effect arising from the design. We also study the properties of a simplified variance estimator and discuss extensions to a wider class of poverty measures.

Keywords: Calibration; Complex sampling design; Linearization; Poverty; Quantile; Raking; Survey weight.

1 Introduction

A widely used measure in poverty comparisons is the proportion falling below a fraction α of the βth quantile of a distribution. For example, Eurostat (2000) defines a low wage as one below 60% ($\alpha = 0.6$) of the national median monthly wage ($\beta = 0.5$) and compares the proportion of employees earning low wages

Address for correspondence: Yves Berger, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom. E-mail: ygb@soton.ac.uk.

in different European countries. Based upon data from the 1996 European Community Household Panel Survey they estimate, for instance, that this proportion is 21% in the United Kingdom compared with 13% in France and 17% in Germany.

When making such comparisons between countries, over time or between subgroups within countries using sample survey data, it is important to have information about the sampling variability of the estimates. The estimation of standard errors for such proportions is, however, not simply a matter of applying standard methods for proportions (e.g. Cochran, 1977,Chapter 3), since the quantile must first be estimated before estimating the proportion falling below a fraction of this estimated quantile. We shall refer to the fraction α of the βth quantile as the *low-income line* and the proportion falling below the low-income line as the *low-income proportion.* Estimation of the low-income proportion thus involves estimating the low-income line first. The term *income* is used here to denote the variable under study. For different applications, this variable will be defined in different ways and might apply to different types of units, for example individuals vs. households.

Preston (1995) considered the estimation of the sampling variance of an estimated low-income proportion. He derived exact and large sample sampling distributions and applied his results to data from the UK Family Expenditure Survey (FES). His estimator of the sampling variance is, however, derived for an unweighted point estimator under the assumption of simple random sampling. In fact, the FES employs a complex sampling scheme involving geographical clustering which may be expected to inflate standard errors. Weighting by population controls is employed and this may also be expected to affect standard errors. The aim of this paper is to show how these additional complex features of a sample survey may also be handled in the estimation of sampling variances and to consider the numerical implications in the case of the 1998-99 FES. Some

other evidence that complex sampling designs may have important effects on standard errors of (other) poverty measures is given by Howes and Lanjouw (1998). Inference about Lorenz curves and quantile shares in the presence of sampling weights is considered by Beach and Kaliski (1986).

Shao and Rao (1993) and Binder and Kovačević (1995) proposed linearization approaches to variance estimation for a low-income proportion for the case $\alpha = 0.5$, $\beta = 0.5$. They allowed for stratified multistage sampling but not for the effects of weighting by population controls. Their approaches might be considered as generalizations of the large sample method of Preston (1995). Preston(1995) provides numerical evidence in the case of simple random sampling that the asymptotic approximation of the sampling distribution, upon which the large sample method is based, is very close to the exact distribution. Shao and Rao (1993) established the consistency of both balanced repeated replication and linearisation variance estimators. Kovačević and Yung (1997) extended Binder and Kovačević (1995) in an empirical study based upon the Canadian Survey of Consumer Finance, comparing their variance estimator with some re-sampling methods, including the jackknife, the bootstrap and balanced half samples. The jackknife is known to provide inconsistent variance estimation in the case of quantiles and Kovačević and Yung found that it was subject to serious biases for the low-income proportion. Of the re-sampling methods, the bootstrap had the least bias, although it still displayed greater bias than the linearization method. Shao and Chen (1998) demonstrated the consistency of a bootstrap variance estimator when $\alpha = \beta = 0.5$ under a stratified multistage design, allowing for hot deck imputation but not weighting to population controls. Chen and Shao (1999) consider the case when the imputed values are non identifiable.

Deville (1999) also discusses the application of the linearization method to variance estimation for a low-income proportion. Moreover, he considers how

the linearization method may be extended for a general estimator via a "residual technique" to handle the effect of weighting by population controls. We apply this idea to the specific case of the low-income proportion in this paper.

One complication in applying the linearization method to measures based upon estimated quantiles, such as the low-income proportion, is that it requires estimation of the probability density function of the variable. There are different approaches to this estimation problem. Deville(1999) suggests a simple approach involving "numerical differentiation" of the estimated distribution function. Preston (1995) uses kernel-based density estimation. Binder and Kovačević (1995) apply an approach proposed by Francisco and Fuller (1991) for functions of estimated quantiles, based upon the lengths of confidence intervals constructed by Woodruff's (1952) procedure.

Zheng (2001) derives asymptotic inference procedure for a wider class of poverty measures under simple random sampling assumption and obtains expressions for asymptotic variance under both stratified and cluster sampling.

The FES and its weighting scheme are described in Section 2. In Section 3, we introduce notation and define the low-income proportion and an estimator of this proportion. A method for variance estimation using linearization is introduced in Section 4. This estimator is extended to accommodate raking and to take account of a complex sampling design in Section 5. Results based on the FES data are presented in Section 6. Conclusion and extension of variance estimation to wider class of poverty measures is considered in Section 7

2 Family Expenditure Survey

The FES has a long history of being used for studies of the distribution of income (Goodman and Webb, 1994). We use data from the 1998-99 FES to produce estimates for the population of private households in the United King-

dom. The variable studied is the equivalent total weekly expenditure of the household. This is derived from total household expenditure by adjusting for the differing sizes and compositions of the households. As an approximation to the McClements scales before adjustment for housing cost (Department of Social Security, 2001), an equivalent value of 0.61 is assigned to the first adult and an equivalent value of 0.39 to each other member aged 16 or over. An equivalent value is also assigned to each child aged under 16: 0.13 for a child between 0 and 4 years old, 0.22 for a child between 5 and 9 years old and 0.26 for child between 10 and 15 years old. The equivalent total expenditure is then formed by dividing the total expenditure by the sum of these equivalent values for the household members. Using total expenditure as a measure of living standards has the advantage, compared to income variables, that it tends to be less affected by random variation in income sources, which may not reflect real changes in living standards (Blundell and Preston, 1998; Deaton, 2000 page 148).

The FES is a multi-stage stratified random sample of $n = 6630$ private households drawn from the Post Office's list of addresses. Postal sectors are the primary sample units (PSU's) and are selected by probability proportional to a measure of size, after being arranged in strata defined by standard regions, socio-economic group and ownership of cars. The Northern Ireland sample is drawn as a random sample of addresses with a larger sampling fraction than for Great Britain.

Under the FES sampling design, all households in Great Britain (GB) are selected with equal first-order inclusion probabilities. All households in Northern Ireland are likewise selected with a fixed inclusion probability, greater than that in GB. Out of the about 10,000 households selected into the target sample, about 66 per cent are contacted and cooperate fully in the survey. Response probabilities have been estimated in a study linking the target sample to the

1991 Census (Elliot, 1997; Foster, 1998). These response probabilities multiplied by the sampling inclusion probabilities generate basic survey weights d_k for each household k. These weights will be referred to as *prior weights* and will be treated as fixed, independent of the sample.

The prior weights d_k are adjusted to agree with control totals using the raking procedure proposed by Deville *et al.* (1993) and fully described in Section 5. The resulting weights are denoted w_k and termed the *raking weights*. Unlike the prior weights, these weights are sample dependent.

3 Point Estimation of Low-income Proportion

We denote the finite population of households as $U = \{1, ..., k, ..., N\}$, where N is the number of households in the population. The equivalent total expenditure for household k is denoted y_k. The distribution function of y_k is denoted $F(y)$ and defined by

$$F(y) = \frac{1}{N} \sum_{k \in U} \delta\{y_k \leq y\}, \tag{1}$$

where $\delta\{\xi\}$ takes the value 1 if ξ is true and the value 0 otherwise.

The β-th quantile of y_k is denoted Y_β and defined by

$$Y_\beta = \inf\{y : F(y) > \beta\}, \tag{2}$$

For example, $Y_{0.5}$ is the median. The low-income line is the fraction α of the β-th quantile; that is, αY_β. The finite population parameter of interest, the low-income proportion, is the proportion of households below the low-income line, denoted by $p_{\alpha\beta}$ and defined by

$$p_{\alpha\beta} = F(\alpha Y_\beta).$$

Given an estimator $\hat{F}(y)$ of $F(y)$ in (1), $p_{\alpha\beta}$ may be estimated by

$$\widehat{p}_{\alpha\beta} = \widehat{F}(\alpha \widehat{Y}_\beta), \tag{3}$$

where $\widehat{Y}_\beta$ is defined by (2) after replacing $F(y)$ by $\widehat{F}(y)$.

In order to consider possible estimators $\hat{F}(y)$ of $F(y)$, let the sample of responding households for which values of y_k are available be denoted s, a subset of U. Given a set of survey weights w_k ($k \in s$), the usual weighted estimator of $F(y)$ and the one considered here is given by

$$\widehat{F}(y) = \frac{1}{\widehat{N}_w} \sum_{k \in s} w_k \delta\{y_k \leq t\},$$

where $\widehat{N}_w = \sum_{k \in s} w_k$. Note that $\hat{F}(y)$ is invariant to multiplication of the weights by a constant and that we assume a scaling of the w_k for which it is natural to view $\widehat{N}_w$ as an estimator of N. Some alternative estimators of $F(y)$ which make use of auxiliary information are discussed, for example, by Nascimento Silva and Skinner (1995). A simple unweighted estimator of $F(y)$ and hence $p_{\alpha\beta}$, as considered by Preston (1995), is obtained by setting each w_k in $\hat{F}(y)$ equal to a constant. We shall suppose that the weights w_k are the ones described in Section 2.

4 Variance Estimation by Linearization

We now consider estimating the variance of $\widehat{p}_{\alpha\beta}$, defined by (3), with respect to the sampling design. We treat non-response as part of the sampling process and assume that the probability π_k that household k is included in s is inversely proportional to d_k, the prior weight (see Section 2).

In this section we treat the weights w_k as fixed. In the following section, we allow for the fact that this is not the case and show how to to include this in the estimation of variance. The basic idea of the linearization method (Campbell and Little, 1980; Deville, 1999) is to find a "pseudo-variable", taking value z_k for household k, such that

$$var(\widehat{p}_{\alpha\beta}) \approx var(\widehat{t}_z), \tag{4}$$

where

$$\widehat{t}_z = \sum_{k \in s} w_k z_k$$

and the approximation $\approx$ is justified by some large sample argument. The variance of the linear statistic $\widehat{t}_z$ may then be estimated by standard survey sampling techniques, which allow for the actual sampling design used. This is considered in Section 5.

The form of the pseudo-variable may be illustrated in the simplest case when the sampling variation in $\widehat{Y}_\beta$ is ignored and $\widehat{Y}_\beta$ is treated as equal to Y_β. In this case, $\widehat{p}_{\alpha\beta}$ is a ratio and a simple pseudo-variable is given by

$$z_k = \frac{1}{N} \left[\delta\{y_k \leq \alpha Y_\beta\} - p_{\alpha\beta} \right]. \tag{5}$$

The variance of $\widehat{t}_z$ may then just be estimated using standard survey software for variance estimation for ratios. Deville(1999) shows that, in order to reflect the sampling variation in $\widehat{Y}_\beta$, we need to include an additional term in the pseudo-variable, that is set

$$z_k = \frac{1}{N} \left\{ \delta\{y_k \leq \alpha Y_\beta\} - p_{\alpha\beta} - \alpha \; R_{\alpha\beta} \left[\delta\{y_k \leq Y_\beta\} - \beta \right] \right\}, \tag{6}$$

where

$$R_{\alpha\beta} = \frac{f(\alpha Y_\beta)}{f(Y_\beta)}$$

and $f(\cdot)$ is the density function corresponding to the distribution function $F(y)$. As $F(y)$ is a step function, the definition of f requires reference to a super-population model (Franscisco and Fuller, 1991) or some other construction (Campbell and Little, 1980; Deville, 1999).

In the case when $\alpha = \beta = 1/2$, these z_k are the same as those proposed by Shao and Rao (1993) Binder and Kovačević (1995). In Section 5, it will be shown that for simple random sampling, these z_k generate a variance $var(\widehat{t}_z)$ which is the same as the large sample variance formula given by Preston (1995).

As the pseudo-variable depends on population parameters, the z_k cannot be computed in practice. The natural solution is to replace the pseudo-variable by its sample estimate

$$\widehat{z}_k = \frac{1}{\widehat{N}_w}\left\{\delta\{y_k \leq \alpha\widehat{Y}_\beta\} - \widehat{p}_{\alpha\beta} - \alpha\ \widehat{R}_{\alpha\beta}\left[\delta\{y_k \leq \widehat{Y}_\beta\} - \beta\right]\right\}, \tag{7}$$

where

$$\widehat{R}_{\alpha\beta} = \frac{\widehat{f}(\alpha\widehat{Y}_\beta)}{\widehat{f}(\widehat{Y}_\beta)}$$

and where $\widehat{f}$ is an estimate of f. To estimate f, we follow Preston (1995) in using a kernel-based estimator of f; that is

$$\widehat{f}(y) = \frac{1}{\widehat{N}\ b}\sum_{k\in s} w_k K\left(\frac{y - y_k}{b}\right),$$

where $K(x) = (2\pi)^{-1/2}exp(-x^2/2)$ is the Gaussian kernel function with a bandwidth $b = 0.79(\widehat{Y}_{0.75} - \widehat{Y}_{0.25})\widehat{N}^{-1/5}$ given by Silverman (1986, page 45-47).

5 Allowing for the Effects of the Complex Sample Design and Weighting to Population Controls

In the previous section we treated the survey weights w_k as fixed. In fact, these weights are sample-dependent and this dependence affects the variance. In this section we first show how the approach of Deville(1999) may be used to modify the pseudo-variable to accommodate the sample dependence of the w_k.

These weights are formed using $M = 49$ population control totals, defined by age- group, sex and region. The m-th control total is denoted by

$$t_{x;m} = \sum_{k\in U} x_{km}$$

where x_{km} is the value for the k-th household of the m-th raking variable such as the number of males aged between 25 and 30 or a region indicator variable.

The raking weights w_k are constructed to agree with the M control totals; that is, for each m the w_k satisfy

$$\sum_{k \in s} w_k x_{km} = t_{x;m}. \tag{8}$$

The method used for the FES to satisfy these constraints is to choose the w_k to minimise the following measure of distance

$$\sum_{k \in s} d_k D\left(\frac{w_k}{d_k}\right), \tag{9}$$

between the prior weights d_k and the weights w_k subject to the set of constraints (8); where D is the logit function defined by

$$D(x) = \frac{(1-\ell)(L-1)}{L-\ell} \log\left[\left(\frac{x-\ell}{1-\ell}\right)^{(x-\ell)} \left(\frac{L-x}{L-1}\right)^{(L-x)}\right]$$

if $\ell < x < L$ and $D(x) = \infty$ otherwise, where ℓ and L are two constants. This method imposes an upper limit L and lower limit ℓ on the weight ratio w_k/d_k. This is often desirable to avoid negative and very large weights. The values used for the FES are $\ell = 0.7$ and $L = 1.4$. The weights that minimise (9) subject to (8) are computed using the CALMAR (Deville *et al.*, 1993) macro in SAS.

Deville (1999) shows that effect of the sample-dependence of the w_k's on variance estimation can be allowed for by replacing the pseudo-variable $\widehat{z}_k$ by residuals $\widetilde{z}_k$ defined by

$$\widetilde{z}_k = \widehat{z}_k - \sum_{m=1}^{M} \widehat{\beta}_m x_{km}, \tag{10}$$

where

$$\widehat{\beta}_m = \sum_{k \in s} d_k \widehat{z}_k x_{km} \left(\sum_{k \in s} d_k x_{km}^2\right)^{-1}.$$

The $\widetilde{z}_k$ are the residuals of the regression of the pseudo-variable (7) on the raking variables x_{km}.

We now show how to use the $\widetilde{z}_k$, obtained in the previous section, to estimate the variance in (4), taking into account the complex nature of the sampling design.

The FES involves a two-stage sampling design. At the first stage the PSUs are stratified into H strata and a sample s_{Ih} of n_{Ih} PSUs is selected from the h-th stratum ($h = 1, ..., H$). Within the ith sampled PSU, a sample s_{ih} of n_{ih} households is selected.

Under two-stage sampling, the variance (4) involves both within and between PSU components. The variance could be estimated by estimating these components separately, for example using the method of Raj (1968), but the resulting calculations can be computationally intensive (Särndal *et al.*, 1992 page 137). A widely used alternative variance estimator, which is computationally simpler and which may be expected to exhibit only very minor upward bias for the small sampling fractions employed in the FES, is given by Särndal *et al.* (1992 page 154):

$$\widehat{var}(\widehat{p}_{\alpha\beta}) = \sum_{h=1}^{H} \frac{n_{Ih}}{(n_{Ih}-1)} \sum_{i \in s_{Ih}} \left(\breve{z}_{Ihi} - \frac{\widehat{t}_{h;z}}{n_{Ih}} \right)^2 ; \tag{11}$$

where

$$\begin{aligned} \breve{z}_{Ihi} &= \sum_{k \in s_{ih}} d_k \widetilde{z}_k, \\ \widehat{t}_{h;z} &= \sum_{i \in s_{Ih}} \breve{z}_{Ihi}. \end{aligned} \tag{12}$$

This estimator may be considered as a generalization of the estimator in Preston (1995). For, in the case of simple random sampling with no survey and no raking weighting, $H = 1$, $n_{Ih} = n$ and $s_{Ih} = s$, so that the variance estimator in (11) above reduces to

$$\widehat{var}(\widehat{p}_{\alpha\beta}) = \frac{n}{n-1} \widehat{var}_{srs}(\widehat{p}_{\alpha\beta}), \tag{13}$$

where

$$\widehat{var}_{srs}(\widehat{p}_{\alpha\beta}) = \frac{1}{n} \left[\widehat{p}_{\alpha\beta} \left(1 - \widehat{p}_{\alpha\beta}\right) + \beta(1-\beta)\alpha^2 \widehat{R}^2_{\alpha\beta} - 2\widehat{p}_{\alpha\beta}(1-\beta)\alpha \widehat{R}_{\alpha\beta} \right] \tag{14}$$

is the variance estimator proposed by Preston (1995). The proof of (13) is given in the Appendix.

6 Results

In this section, we compute values of the point estimator $\widehat{p}_{\alpha\beta}$ as well as estimates of its variance for different values of α and β using the 1998-99 FES data.

First, we study the effect of weighting. We compute the value of $\widehat{p}_{\alpha\beta}$ for different values of α and β, using different methods of weighting: "equal weights" with each household having the same weight, "prior weights" d_k and "raking weights" w_k. The results are presented in Table 1. We see that the effect of using the prior weights or raking weights on $\widehat{p}_{\alpha\beta}$ is relatively minor. Raking tends to increase $\widehat{p}_{\alpha\beta}$ slightly for all the values of α and β considered, which appears to reflect the fact that age-sex groups with lower incomes tend to be under-represented among the respondents.

β	α	Equal weights	Prior weights	Raking weights	β	α	Equal weights	Prior weights	Raking weights
0.3	0.3	0.008	0.008	0.009	0.4	0.6	0.155	0.154	0.157
0.3	0.4	0.025	0.025	0.028	0.4	0.7	0.213	0.215	0.218
0.3	0.5	0.056	0.055	0.058	0.5	0.3	0.033	0.034	0.035
0.3	0.6	0.098	0.099	0.100	0.5	0.4	0.081	0.083	0.086
0.3	0.7	0.146	0.146	0.148	0.5	0.5	0.148	0.149	0.151
0.4	0.3	0.017	0.017	0.018	0.5	0.6	0.216	0.219	0.222
0.4	0.4	0.049	0.049	0.052	0.5	0.7	0.292	0.294	0.298
0.4	0.5	0.098	0.099	0.101	0.5	0.8	0.363	0.364	0.366

Table 1: Values of $\widehat{p}_{\alpha\beta}$ for different values of β and α and for different weighting schemes.

We next consider the values of alternative variance estimators defined by (11) and (14). To standardise the results for different values of α and β, we consider the relative variance (RV) given by

$$\widehat{RV} = 100\,\frac{\widehat{var}(\widehat{p}_{\alpha\beta})}{\widehat{p}_{\alpha\beta}^{(r)\,2}},$$

where $\widehat{p}_{\alpha\beta}^{(r)}$ is the low-income proportion computed using the raking weights, i.e. the last column of Table 1. In order to assess the impact of raking and

complex sampling, we compute three alternative variance estimates and associated estimates of the relative variance, as shown in Table 2. The first estimator ignores the weighting and complex design and thus effectively makes simple random sampling assumptions, as in Preston (1995) and is denoted $\widehat{RV}_{(srs)}$. This variance estimator is given by (14). The second estimator allows for the prior weighting and the complex design but ignores the effect of raking - it is denoted $\widehat{RV}_{(design)}$. This estimator is given by (11) where the $\widetilde{z}_k$'s are replaced by the $\widehat{z}_k$'s defined in (7). The third estimator allows for the full survey weighting, complex design and effect of raking and is denoted $\widehat{RV}_{(full)}$ and given by (11).

	Raking	Complex Sampling	Weighting
$\widehat{RV}_{(srs)}$	No	No	Equal
$\widehat{RV}_{(design)}$	No	yes	Prior Weights
$\widehat{RV}_{(full)}$	Yes	Yes	Raking Weights

Table 2: Definition of estimators of the relative variances considered.

The values of these three estimated relative variances for different values of β and α are given in Table 3 . In addition, we present values of the misspecification effects, $meff_{(raking)}$ and $meff_{(full)}$, which are obtained by dividing $\widehat{RV}_{(full)}$ by $\widehat{RV}_{(design)}$ and $\widehat{RV}_{(full)}$ by $\widehat{RV}_{(srs)}$, respectively. These measure the effect of misspecifying the variance estimator by ignoring the raking effect or by ignoring both the effect of raking and complex sampling, respectively (Skinner, Holt and Smith, 1989,Ch.2). There is a strong inverse relationship between the variance and the estimated value of the low income proportion, just as for the binomial variance of a proportion. Comparing $\widehat{RV}_{(srs)}$ with $\widehat{RV}_{(design)}$, we see that the variance is almost always underestimated if the complex design is ignored. The values of $meff_{(raking)}$ indicate that ignoring raking tends to lead to a slight underestimation of the variance, but not always. Overall, the effect of raking and the complex design, as measured by $meff_{(full)}$, is consistently to increase

the variance, but never by more than 17% for the values of α and β considered. There is no evident strong dependence of these values on α or β.

The sampling variation in the estimated low-income proportion arises from two sources: sampling variation in the estimated low-income line and sampling variation in the estimated low-income proportion given this estimated line. An interesting finding of Preston (1995) is that these two sources can be mutually compensating "in a manner that is typically helpful to the estimation of relative poverty incidence" (Preston, 1995, page 95). As a result, if the variance of the estimated low-income proportion is estimated under the simplifying assumption that the low-income line is fixed, the resulting estimated variance may actually be conservative, whereas one might have expected it to be an underestimate since it ignores a source of sampling variation. Preston(1995) finds that the variance under the simplifying assumption is larger than the actual large sample variance particularly for large values of α. Indeed, if α is large $p_{\alpha\beta}$ is close to the constant β.

β	α	$\widehat{p}_{\alpha\beta}^{(r)}$	$\widehat{RV}_{(srs)}$	$\widehat{RV}_{(design)}$	$\widehat{RV}_{(full)}$	$meff_{(raking)}$	$meff_{(full)}$
0.3	0.3	0.009	1.645	1.712	1.793	1.047	1.090
0.3	0.4	0.028	0.479	0.463	0.492	1.064	1.028
0.3	0.5	0.058	0.221	0.238	0.245	1.030	1.111
0.3	0.6	0.100	0.107	0.119	0.124	1.044	1.153
0.3	0.7	0.148	0.055	0.055	0.056	1.018	1.018
0.4	0.3	0.018	0.801	0.838	0.834	0.995	1.040
0.4	0.4	0.052	0.267	0.302	0.311	1.031	1.165
0.4	0.5	0.101	0.120	0.138	0.138	1.000	1.144
0.4	0.6	0.157	0.063	0.065	0.067	1.025	1.052
0.4	0.7	0.218	0.034	0.037	0.037	1.002	1.095
0.5	0.3	0.035	0.413	0.440	0.441	1.003	1.068
0.5	0.4	0.086	0.155	0.173	0.172	0.994	1.110
0.5	0.5	0.151	0.077	0.084	0.085	1.007	1.101
0.5	0.6	0.222	0.042	0.047	0.048	1.024	1.129
0.5	0.7	0.298	0.023	0.026	0.025	0.986	1.110

Table 3: Values of the relative variances (%) for different values of β and α. $meff_{(raking)}$ is the effect of raking and $meff_{(full)}$ is the effect of the design and raking.

We now extend this comparison to include allowance for the complex design and weighting in the FES. To do this we consider the pseudo-variable (5) instead of (6), that is we use $\widehat{z}_k = [\delta\{y_k \leq \alpha\widehat{Y}_\beta\} - \widehat{p}_{\alpha\beta}]N_w^{-1}$. The resulting "naive" variance estimator based upon this binary pseudo-variable is easy to compute and, indeed, may be obtained from standard software for survey variance estimation by treating the low-income proportion as a standard estimated proportion. In particular, this variance estimator does not require the estimation of the density function. Table 4 gives these naive estimates of variance.

β	α	$\widehat{p}_{\alpha\beta}^{(r)}$	$\widehat{RV}_{(srs)}$	$\widehat{RV}_{(design)}$	$\widehat{RV}_{(full)}$	$meff_{(raking)}$	$meff_{(full)}$
0.3	0.3	0.009	1.525	1.562	1.644	1.053	1.078
0.3	0.4	0.028	0.465	0.499	0.523	1.047	1.123
0.3	0.5	0.058	0.236	0.300	0.288	0.957	1.219
0.3	0.6	0.100	0.135	0.196	0.182	0.927	1.350
0.3	0.7	0.148	0.086	0.128	0.114	0.890	1.321
0.4	0.3	0.018	0.744	0.823	0.812	0.987	1.091
0.4	0.4	0.052	0.262	0.347	0.337	0.972	1.285
0.4	0.5	0.101	0.132	0.194	0.180	0.926	1.368
0.4	0.6	0.157	0.080	0.121	0.108	0.895	1.346
0.4	0.7	0.218	0.053	0.081	0.071	0.877	1.334
0.5	0.3	0.035	0.388	0.456	0.430	0.944	1.109
0.5	0.4	0.086	0.152	0.211	0.199	0.943	1.312
0.5	0.5	0.151	0.083	0.126	0.111	0.884	1.336
0.5	0.6	0.222	0.052	0.079	0.069	0.873	1.327
0.5	0.7	0.298	0.035	0.054	0.047	0.872	1.352

Table 4: Values of the relative variances (%) for different values of β and α, ignoring sampling variation in the low income line.

Comparing Table 3 and 4, we see, as in Preston (1995), that $\widehat{RV}_{(srs)}$ is larger for the naive estimator if α is sufficiently large, for each value of β. The same finding applies to $\widehat{RV}_{(full)}$ for markedly wider ranges of α. Thus, the variance estimator that takes the raking and the sample design into account is conservative for all cases between $\alpha = 0.4$ and $\alpha = 0.7$. The effect of the design and the raking adjustment tends to be more marked for this estimator, as measured by the difference between the misspecification effects and 1.

7 Conclusion and Extension

We have shown how both complex sampling schemes and raking adjustments may be handled in variance estimation for low income proportions. The approach is straightforward and could be handled with standard survey software for variance estimation together with software which enables the calculation of the pseudo-variable in (6) and the regression residuals in (10). Using data from the 1998-99 FES, the impact of complex sampling and raking tends to increase the estimated standard errors for all values of α and β considered, although the inflation of the variance never exceeds 17%. We have also considered the use of a simpler 'naive' approach, which ignores the sampling variation in the low income line and treats the low income proportion just like a standard proportion. As in Preston (1995), this approach appears to be conservative so long as α is not too small.

As a measure of poverty, the low income proportion considered in this paper is crude, since it takes no account of how far an income falls below the low-income line. The shortfall of an income y_k below a low income line θ may be taken account of in the wide class of "decomposable" measures, considered by Zheng (2001),

$$p = \frac{1}{N} \sum_{k \in U} h(y_k, \theta)$$

where $h(y_k, \theta)$ is a "poverty deprivation function" with $h(y_k, \theta) = 0$ if $y_k > \theta$. An important sub-class arises when $h(y_k, \theta) = [(\theta - y_k)/\theta]^\gamma \ \delta\{y_k \leq \theta\}$ and γ is a specified non-negative constant (Foster *et al.* 1984). The low-income proportion is the special case where $\gamma = 0$ and $\theta = \alpha Y_\beta$. Measure with $\gamma = 1$ or 2 also have natural interpretation (Foster *et al.* 1984).

The measure p may be estimated by

$$\widehat{p} = \frac{1}{\widehat{N}_w} \sum_{k \in s} w_k \ h(y_k, \theta) \tag{15}$$

if θ is known, or by substituting an estimator $\widehat{\theta}$ for θ otherwise. If θ is a given constant then $\widehat{p}$ is simply a ratio of linear statistics and the approach in the paper to handle complex sampling and raking (see Section 5) may be followed by replacing z_k in (5) by

$$z_k = \frac{1}{N}\left[h(y_k, \theta) - p\right]$$

If θ is estimated from the sample s then further linearisation is required. If $\theta = \alpha Y_\beta$ as in this paper with $\widehat{\theta} = \alpha \widehat{Y}_\beta$ then, following the argument of Zheng (2001), under regularity conditions on the function $h(.,.)$ given by Zheng, the pseudo-variable in (6) may be replaced by (see Zheng, 2001, page 343 and Deville, 1999, page 197)

$$z_k = \frac{1}{N}\left\{h(y_k, \theta) - p - \alpha\ R^{(h)}_{\alpha\beta}\left[\delta\{y_k \leq Y_\beta\} - \beta\right]\right\} \tag{16}$$

where

$$\begin{aligned} R^{(h)}_{\alpha\beta} &= \left[a + h(\theta, \theta)\ f(\theta)\right]\ f(Y_\beta)^{-1} \\ a &= \frac{1}{N}\sum_{k \in U} h_\theta(y_k, \theta) \end{aligned}$$

where $h_\theta(y_k, \theta) = \partial h(y_k, \theta)/\partial\theta$ and $f(\cdot)$ is the density function considered earlier.

A linearisation variance estimator may then be determined by replacing N, θ, p, Y_β, $f(Y_\beta)$ and a by $\widehat{N}_w$, $\widehat{\theta}$, $\widehat{p}$, $\widehat{Y}_\beta$, $\widehat{f}(\widehat{Y}_\beta)$ and $\widehat{a} = \widehat{N}_w^{-1}\sum_{k \in s} w_k h_\theta(y_k, \widehat{\theta})$. Note that in the case of the low income proportion, we have $a = 0$ and $h(\theta, \theta) = 1$ so (16) reduces to (6). Note that for the case when $h(y_k, \theta) = [(\theta - y_k)/\theta]^\gamma$ $\delta\{y_k \leq \theta\}$ and $\gamma > 0$ we have $h(\theta, \theta) = 0$.

Another common choice of the low-income line is $\theta = \alpha\ \mu$, where μ is the mean income $\mu = N^{-1}\sum_{k \in U} y_k$ and α is a given fraction. If $\widehat{p}$ in (15) is defined with θ replaced by $\widehat{p} = \alpha \widehat{N}_w^{-1}\sum_{k \in s} w_k\ y_k$, then, following Zheng (2001), the pseudo-variable becomes (see Zheng, 2001, page 343 and Deville, 1999, page

197)

$$z_k = \frac{1}{N}\left\{h(y_k,\theta) - p + (\alpha\ y_k - \theta)[a + h(\theta,\theta)\ f(\theta)]\right\}$$

and the linearization variance estimator may be determined again by replacing N, θ, p, a, $f(\theta)$ and a by $\widehat{N}_w$, $\widehat{\theta}$, $\widehat{p}$, $\widehat{a}$, $\widehat{f}(\widehat{\theta})$. Note that for the class of measures $h(y_k,\theta) = [(\theta - y_k)/\theta]^\gamma\ \delta\{y_k \leq \theta\}$ with $\gamma > 0$, the computation of this variance estimator is simplified since $h(\theta,\theta) = 0$ and it is not necessary to estimate the density function $f(\cdot)$.

Acknowledgements

This work was supported by grant R000223397 of the Economic and Social Research Council. We are grateful to Ian Crawford for advice on the choice of variable and to the referees for helpfull comments.

8 Appendix - Proof of (13)

In the case of simple random sampling with no survey weighting, $H = 1$, $n_{Ih} = n$, $s_{Ih} = s$, $\widetilde{z}_k = \widehat{z}_k$ and $d_k = Nn^{-1}$, so that the variance estimator in (11) reduces to

$$\widehat{var}(\widehat{t}_z) = \frac{N^2}{n(n-1)}\sum_{k\in s}(\widehat{z}_k - \overline{z}_s)^2$$

where $\overline{z}_s$ is the sample mean of the $\widehat{z}_k$'s. It can be easily shown that $\overline{z}_s = 0$. Thus,

$$\begin{aligned}\widehat{var}(\widehat{t}_z) \quad = \quad & \frac{1}{n(n-1)}\left\{\sum_{k\in s}\left[\delta\{y_k \leq \alpha\widehat{Y}_\beta\} - \widehat{p}_{\alpha\beta}\right]^2\right. \\ & +\alpha^2\widehat{R}^2_{\alpha\beta}\sum_{k\in s}\left[\delta\{y_k \leq \widehat{Y}_\beta\} - \beta\right]^2 \\ & \left.-2\alpha\widehat{R}_{\alpha\beta}\sum_{k\in s}\left[\delta\{y_k \leq \alpha\widehat{Y}_\beta\} - \widehat{p}_{\alpha\beta}\right]\left[\delta\{y_k \leq \widehat{Y}_\beta\} - \beta\right]\right\},\end{aligned}$$

$$= \frac{1}{n-1}\Big\{\widehat{p}_{\alpha\beta}\,(1-\widehat{p}_{\alpha\beta}) + \beta(1-\beta)\alpha^2\widehat{R}^2_{\alpha\beta} \\ -2\alpha\widehat{R}_{\alpha\beta}\left[\frac{1}{n}\sum_{k\in s}\Big[\delta\{y_k \le \alpha\widehat{Y}_\beta\}\delta\{y_k \le \widehat{Y}_\beta\}\Big] - \beta\widehat{R}_{\alpha\beta}\right]. \qquad (17)$$

It is clear that for $\alpha < 1$,

$$\frac{1}{n}\sum_{k\in s}\Big[\delta\{y_k \le \alpha\widehat{Y}_\beta\}\delta\{y_k \le \widehat{Y}_\beta\}\Big] = \widehat{p}_{\alpha\beta}. \qquad (18)$$

Thus by replacing (18) in (17), we obtain (13). □

References

Beach, C.M. and Kaliski, S.F. (1986) Lorenz Curve Inference with Sample Weights: An Application to the Distribution of Unemployment Experience. *Appl. Statist.*, **35**, 38-45.

Binder, D.A. and Kovačević, M.S. (1995) Estimating some measures of income inequality from survey data: an application of the estimating equations approach. *Survey Methodology*, **21**, 137-145.

Blundell, R. and Preston, I.. (1998) Consumption Inequality and Income Uncertainty. *Quart. J. Econom.*, **113**, 603-640.

Campbell, C. and Little, D. (1980) A different view of finite population estimation. *Proceeding Survey Research Methods Section,* ASA, 319-324.

Chen, Y. and Shao, J (1999) Inference with survey data imputed by hot deck when imputed values are nonidentifiable. *Statistica Sinica,* **9**, 361-384.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd. Ed. New York: Wiley.

Deaton, A. (2000) *The Analysis of Household Surveys*, 3rd. Ed. Baltimore: The John Hopkins University Press.

Department of Social Security (2001) *Households Below Average Income 1999/00*, Appendix 2: Methodology.

Deville, J.C. (1999) Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology,* **25**, 193-203.

Deville, J.C. and Särndal, C.E. (1992) Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

Deville, J.C., Särndal, C.E. and Sautory, O.(1993) Generalised raking procedure in survey sampling. *J. Amer. Statist. Assoc.*, **88**, 1013-1020.

Elliot, D. (1997) Software to Weight and Gross Survey Data. *GSS Methodology Series no*, **1**, Office for National Statistics, UK.

Eurostat (2000) Low-wage employees in EU countries. Statistics in Focus: Population and Social Conditions. Theme 3-11/2000. *Office for Official Publications of the EC*, Luxembourg.

Foster, K. (1998) Evaluating non-response on household surveys. *GSS Methodology Series no*, **8**, Office for National Statistics, UK.

Foster, K., Greer, J. and Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrika*, **52**, 761-766.

Francisco, C.A. and Fuller, W.A. (1991) Quantile estimation with complex survey design. *Ann. Statist.*, **19**, 454-469.

Goodman, A. and Webb, S. (1994) For richer, for poorer: the changing distribution of income in the UK, 1961-91. *Fiscal Studies*, **15**, 29-62.

Howes, S. and Lanjouw, J.O.(1998) Does sample design matter for poverty rate comparisons? *Review of Income and Wealth*, **44**, 99-109.

Kovačević, M.S. and Yung, W. (1997) Variance estimation for measures of income inequality and polarization - an empirical study *Survey Methodology*, **23**, 41-52.

Nascimento Silva, P.L.D. and Skinner, C.J. (1995) Estimating distribution functions with auxiliary information using poststratification. *J. Off. Statist.*, **11**, 277-294.

Preston, I. (1995) Sampling distributions of relative poverty statistics. *Appl. Statist.*, **44**, 91-99.

Raj, D., (1968) *Sampling Theory*, New York, McGraw Hill.

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

Shao, J. and Rao, J.N.K. (1993) Standard errors for low income proportions estimated from stratified multistage samples. *Sankhya*, Ser. B, **55**, 393-414.

Shao, J. and Chen, Y. (1998) bootstrapping sample quantiles based on complex survey data under hot deck imputation. *Statistica Sinica*, **8**, 1071-1085.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Skinner, C.J., Holt, D. and Smith, T.M.F. eds. (1989) *Analysis of Complex Surveys*. Chichester: Wiley.

Woodruff, R.S. (1952) Confidence intervals for medians and other position measures. *J. Amer. Statist. Assoc.*, **47**, 635-646.

Zheng, B. (2001) Statistical inference for poverty measure with relative poverty lines. *J. Econometrics*, **101**, 337-356.