WPS3543

Survey Nonresponse and the Distribution of Income

Anton Korinek, Johan A. Mistiaen, and Martin Ravallion[1]

Development Research Group, World Bank,
1818 H Street NW, Washington DC, USA

Abstract: The paper examines the distributional implications of selective compliance in sample surveys, whereby households with different incomes are not equally likely to participate. Poverty and inequality measurement implications are discussed for monotonically decreasing and inverted-U compliance-income relationships. We demonstrate that the latent income effect on the probability of compliance can be estimated from information on response rates across geographic areas. On implementing the method on the Current Population Survey for the United States we find that the compliance probability falls monotonically as income rises. Correcting for nonresponse appreciably increases mean income and inequality, but has only a small impact on poverty incidence up to poverty lines common in the United States.

Keywords: Survey nonresponse, income distribution, poverty and inequality measurement.

JEL: C42, D31, D63, I3

World Bank Policy Research Working Paper 3543, March 2005

[1] Angus Deaton's probing comments on an earlier version of this paper led to a number of improvements. Helpful comments on the paper were also received from Francesco Brindisi, Frank Cowell, Phoebus Dhrymes, Peter Lambert, Dominique van de Walle and three anonymous referees.
Address for correspondence: mravallion@worldbank.org.

1. Introduction

Errors in the incomes reported in surveys have important implications for measures of poverty and inequality based on those surveys (Van Praag et al., 1983; Chakravarty and Eichhorn, 1994; Ravallion, 1994; Cowell and Victoria-Feser, 1996; Chesher and Schluter, 2002). For example, classical (zero-mean, white noise) measurement error in the reported incomes of sampled households leads to over-estimation of standard inequality measures (Chakravarty and Eichorn, 1994).

A measurement issue that has received less attention is the fact that invariably some proportion of sampled households simply do not participate in surveys, either because they explicitly refuse to do so or nobody is at home. In the literature, this is often called "unit nonresponse" and is distinct from "item nonresponse," which occurs when some of the sampled households who agree to participate refuse to answer specific questions, such as on their incomes. Various imputation/matching methods address item nonresponse by exploiting the questions that are in fact answered (Lilard et al., 1986; Little and Rubin, 1987). However, that is not an option for unit nonresponse. Some surveys make efforts to avoid unit nonresponse, using "call-backs" to nonresponding households and fees paid to those who agree to be interviewed.[2] Nonetheless, the problem is practically unavoidable and nonresponse rates of 10% or higher are common; indeed, we know of national surveys for which 30% of those sampled did not comply.[3]

This paper studies the implications of unit nonresponse for the measurement of poverty and inequality and provides a new method of correcting survey data for selective compliance.

---

[2] On reducing bias using call-backs see Deming (1953), Van Praag et al. (1983), Alho (1990), and Nijman and Verbeek (1992).

[3] Scott and Steele (2004) report nonresponse rates for eight countries, which are as high as 26%. Holt and Elliot (1991) quote a range of 15-30% for surveys in the UK. Philipson (1997) reports a mean nonresponse rate of 21% for surveys by the National Opinion Research Center in the U.S. Cook et al. (2000) find that the nonresponse rate in Internet surveys are typically around 65% of all designated individuals, and is rising with the increasing use of this method of data collection.

To the extent that survey compliance is random, there will be no concern about biases in survey-based inferences; the final sample will still be representative of the population. However, just as income constrains almost all behavior, it undoubtedly matters to choices about compliance with sample assignments. For instance, high-income households might be less likely to participate because of a high opportunity cost of their time or concerns about intrusion in their affairs.[4] Naturally evidence on this is scarce, but in one interesting study of compliance with the "long-form" of the US Census, Groves and Couper (1998, Chapter 5) found that higher socio-economic status tended to be associated with lower compliance. It might be conjectured that the poorest are also underrepresented; some are homeless and hard to reach in standard household survey designs, and some may be physically or socially isolated and thus less easily interviewed (though the aforementioned study for the US did not find that this was the case). In section 2 of this paper we provide a simple economic model of compliance choice which can generate a range of possibilities for the relationship between compliance and income.

The presence of income-dependent compliance can create biases in survey-based estimates of measures of poverty and inequality. If compliance tends to fall as income rises then surveys will tend to overestimate the proportion of the population with incomes below any given poverty line; we prove this claim in section 3. However, if compliance tends to be lower for both the very poor and the very rich then there will be potentially offsetting effects on measures of the incidence of poverty. Potential implications for measures of inequality are even more complex. In section 3 we show that one cannot establish unambiguous orderings even when compliance falls monotonically with income. Selective compliance by income may well have an offsetting effect on measured inequality to measurement errors in reported incomes.

[4] Groves and Couper (1998) provide a useful overview of the factors thought to influence survey compliance and the available evidence.

The paper then turns to the empirical implications of selective compliance. There is a large literature in statistics on the various methods for correcting for nonresponse, both as part of survey implementation and after collecting the survey data; Korinek et al., (2004) review the methods found in this literature. Here our discussion will focus solely on how the problem has been addressed in work on measuring inequality and poverty.

One strand of the literature on measuring poverty has tried to deal with the problem of unit nonresponse and income underreporting by replacing average incomes from national accounts.[5] This approach rests on two key assumptions, namely that the national accounts give a valid estimate of mean household income and that the discrepancy between the two data sources is distribution neutral; implying one only needs to make an equi-proportionate correction at all levels. Hitherto, little or no evidence has been advanced for or against these assumptions.[6]

A second approach in the literature is based on *ex post* re-weighting of the survey data by utilizing geographic or other observable differences in survey response rates. For example, Atkinson and Micklewright (1983) use regional differences in survey response rates to correct for differential nonresponse in the U.K. Family Expenditure Survey. The Current Population Survey for the U.S. uses a similar method (Census Bureau, 2000, Chapter 10). These methods assume that the non-compliance problem is ignorable within areas. However, this assumption is essentially *ad hoc*, with no behavioral basis, and there is no *a priori* reason why it would be valid; why would compliance be non-random between areas but random within them?

---

[5] Recent examples include Bourguignon and Morrisson (2002), Sala-i-Martin (2002) and Bhalla (2002). While advocates of this approach acknowledge that they are using this method for computational convenience, they also defend the method on the grounds that it allows a correction for under-reporting and non-compliance in surveys.

[6] For further discussion (in the context of poverty measurement for India, though the point is more general) see Ravallion (2000). On the discrepancies between estimates of mean consumption from surveys versus national accounts across countries see Ravallion (2003).

Elsewhere we have shown that the ignorability assumption can be relaxed using exactly the same data used in past *ad hoc* corrections found in the literature using ex post re-weighting of the survey data (Korinek et al., 2004). We draw on the latter paper in section 4 to demonstrate that it is possible to identify the latent individual probability of survey compliance as a function of income using the empirical relationship between aggregate compliance rates across areas and the observed income distribution within areas. Our approach deals simultaneously with response bias within and between areas. We are thus able to present in section 5 the first estimates (to our knowledge) of the bias in measured income distributions due to an income effect on unit nonresponse. While we only present estimates for one country here, the minimal data requirements of our method should allow a wide range of applications in practice. Applications can also be expected to other survey data besides incomes.

2. Income-dependent survey compliance

Survey participation is a matter of individual choice; nobody is obliged to comply with the statistician's randomized assignment. There is some perceived utility gain from compliance—the satisfaction of doing one's civic duty, for example—but there is a cost as well.

Let $y \in [y_P, y_R]$ be household income per person (y_P is the income of the poorest person and y_R is for the richest) and $c(y)$ the cost to the respondent of survey participation (net of any compensation received for participation). We assume that $c'(y) \geq 0$. On possible rationalization of this assumption is by assuming that the opportunity cost of the time required to comply rises with income, while the time itself is roughly independent of income.[7]

[7] Let τ denote the time required for the survey interview and normalize total available time to unity. Full income is $y = w + \pi$ where *w* is the wage rate and π is non-wage income. The cost of survey participation is then $c(y) = \tau w = \tau(y - \pi)$ with $0 < c'(y) = \tau < 1$. Nonlinearity of *c*(*y*) can arise when τ varies with *y*.

Let utility be $u[y - c(y)d,\ d]$ where d=1 if one chooses to comply and d=0 if not. The function u is strictly increasing in both arguments. The utility gain from compliance is:

$$g(y) = u[y - c(y), 1] - u(y, 0) \tag{1}$$

with slope:

$$g'(y) = u_y[y - c(y), 1][1 - c'(y)] - u_y(y, 0) \tag{2}$$

where subscripts denote partial derivatives.

This simple model can generate a wide range of outcomes for the relationship between compliance and income. We consider some special cases.

From (2), it is evident that compliance falls monotonically with income if and only if:

$$c'(y) > 1 - \frac{u_y(y,0)}{u_y[y - c(y),1]} \text{ for all } y$$

A simple case in which this holds is when the cost of participation increases monotonically with income ($c'(y) > 0$) and the marginal utility of income is independent of survey participation, i.e., $u_y(y, 0) = u_y[y - c(y), 1]$. Then, $g'(y) = -u_y(.)c'(y) < 0$ for all y. However, in different specifications, the opposite result can be obtained, whereby compliance rises with income. For example, suppose instead that the cost of participation is independent of income ($c'(y) = 0$), implying that $g'(y) = u_y[y - c(y), 1] - u_y(y, 0)$. If there is diminishing marginal utility of income and utility is separable between income and compliance ($u_y(y, 1) = u_y(y, 0)$) then $g'(y) > 0$; the poor will be less likely to participate.

Without separability, the outcome depends on whether compliance raises or lowers the marginal utility of income, which is not obvious on *a priori* grounds. If compliance leads to a higher marginal utility of income then again $g'(y) > 0$. If it lowers the marginal utility of

income then the income effect could go either way. Suppose that the difference in income effect on the marginal utility of income dominates at low incomes, $u_y[-c(y), 1] > u_y(0, 0)$, while the adverse effect of compliance on the marginal utility of income dominates at high y, i.e., $u_y[1-c(y), 1] < u_y(1, 0)$. Then one can again find an inverted-U pattern in which middle-income groups are more likely to participate than either tail of the distribution.

Other special cases can deliver this inverted-U relationship. For instance, assume that: (i) the cost of compliance is a non-negative and strictly increasing and convex in income, $c'(y) > 0$, $c''(y) > 0$ with $c'(y_P) = 0$; (ii) utility is separable between income and compliance and (iii) for the richest person, the cost of participation is negligibly small, i.e., $\lim_{y \to y_R} u_y[y - c(y)] \approx u_y(y)$. Then separability implies that we can re-write (2) as:

$$g'(y) = -u_y[y - c(y)]c'(y) + u_y[y - c(y)] - u_y(y) \tag{3}$$

The first term on the right-hand side is negative while the second is positive, given declining marginal utility. At low incomes the second term will dominate (since $c'(y)$ will be small) and hence $g'(y) > 0$ at low y. At high incomes, by contrast, the first term will dominate and hence $g'(y) < 0$. In other words, the gains will tend to be highest for middle-income groups.

Notice that in this model, the introduction of a fixed fee paid to those who agree to participate will increase the probability of participation, but it can also increase the likelihood of a compliance bias whereby the response rate falls with income. This will happen if the cost of compliance rises less than one-to-one with income, and there is declining marginal utility of income.

Finally, note that uncertainty can be introduced in this model by assuming that the utility gain from participation is a normally distributed random variable $v \sim N(\mu, \sigma^2)$. After

simplifying the specification by assuming that utility is separable and linear in wealth, utility can be expressed as $u(y-c(y)d,d)=y+d[v-c(y)]$. This yields a simple formula for the probability of response for each household:

$$\Pr(response)=\Pr(v>c(y))=\Pr(\mu+\sigma z>c(y))=\Pr\left(z<\frac{\mu-c(y)}{\sigma}\right)=\Phi(\alpha-\hat{c}(y)), \qquad (4)$$

where z is the transformation of v to a standard normal distribution, Φ is the cumulative distribution function of z, i.e. the probit function, $\alpha=\mu/\sigma$ and $\hat{c}(y)=c(y)/\sigma$. This equation can be readily used for estimations, e.g. by assuming $\hat{c}(y)=\beta\ln y$ or $\hat{c}(y)=\beta\ln y+\gamma(\ln y)^2$ or similar specifications for $\hat{c}(y)$. We return to this in section 5.

3. Implications for the distribution of income

In exploring the theoretical implications for the distribution of income, we confine attention to the special cases discussed above in which the compliance-income relationship is either monotonic decreasing or an inverted-U shape.

Let $F(y)$ denote the true (unobserved) cumulative distribution function of income y with continuous density function $f(y)$. The sample-based estimate is $\hat{F}(y)$ with corresponding density $\hat{f}(y)$ with $\hat{F}(y_P)=0$. The true distribution can be derived from the empirical distribution by appropriate re-weighting. The true density function is $f(y)=w(y)\hat{f}(y)$ where $w(y)=\phi[g(y)]$ are the correction factors for a strictly decreasing differentiable function ϕ. The corrected distribution function is:

$$F(y)=\int_{y_P}^{y} w(x)\hat{f}(x)dx \qquad (5)$$

The expected value of the correction factor is unity, i.e., $\int_{y_P}^{y_R} w(x)\hat{f}(x)dx=1$.

Consider first the case in which compliance falls monotonically with income, i.e., $w'(y) > 0$. Clearly the mean will be underestimated, but how will the distribution of income be affected? On integrating (5) by parts we find that:

$$H(y) \equiv F(y) - \hat{F}(y) = [w(y) - 1]\hat{F}(y) - \int_{y_P}^{y} w'(x)\hat{F}(x)dx \tag{6}$$

for which:

$$H'(y) = [w(y) - 1]\hat{f}(y) \tag{7}$$

It is evident that $H(y) < 0$ for all $y \in (y_P, w^{-1}(1))$. The possibility of $H(y) > 0$ for some $y > w^{-1}(1)$ can be ruled out by noting that $H'(y) > 0$ for all $y > w^{-1}(1)$ and that $H(y_R) = 0$ (since $F(y_R) = \hat{F}(y_R) = 1$).[8] Thus we have first-order dominance over all y, implying that the empirical distribution will overestimate the extent of income poverty for all poverty lines and all additive poverty measures satisfying standard properties (Atkinson, 1987).

Consider instead the inverted-U relationship of compliance with income. There are two support points at which no correction to the density function is needed, namely y_L and y_U with $y_L < y_U$, $w(y_L) = w(y_U) = 1$, $w(y) > 1$ for $y < y_L$ and $y > y_U$ and $w(y) < 1$ for $y_L < y < y_U$. We also assume that $w'(y) < 0$ for all $y < y_L$ and $w'(y) > 0$ for all $y > y_U$ though this can be relaxed somewhat without altering the main results. From (6):

$$F(y_L) - \hat{F}(y_L) = -\int_{y_P}^{y_L} w'(x)\hat{F}(x)dx > 0 \tag{8}$$

$$F(y_U) - \hat{F}(y_U) = -\int_{y_P}^{y_U} w'(x)\hat{F}(x)dx < 0 \tag{9}$$

[8] Suppose to the contrary that there exists an intermediate value y^* in the open interval $(w^{-1}(1), y_R)$ such that $H(y^*) = 0$; then $H'(y) < 0$ for some y in (y^*, y_R), which is a contradiction.

Intuitively, both the incidence of low-incomes ($F(y_L)$) and high incomes ($1-F(y_U)$) are underestimated, given the structure of the income effect on compliance. Recalling (6), it is evident that the impact of this pattern of income effects on compliance is as represented in Figure 1. By continuity, there must exist a point $y^* \in (y_L, y_U)$ such that $F(y^*) = \hat{F}(y^*)$. Again, for a broad class of poverty measures and all poverty lines up to y^*, the empirical distribution will underestimate the extent of income poverty.

We have seen that there is first-order dominance of the corrected distribution over the empirical distribution when the probability of survey compliance falls monotonically with income, implying that poverty will fall when we correct for this bias. Can we also establish Lorenz dominance in this case, and hence conclude that there is an unambiguous effect on all measures of inequality satisfying standard properties (including the transfer axiom)?

The analysis for inequality is easier if we work with the quantile functions: $y(p) \equiv F^{-1}(p)$ for the true distribution and $\hat{y}(p) \equiv \hat{F}^{-1}(p)$ for the empirical distribution. The Lorenz curve for the corrected and empirical distributions are $L(p) \equiv \int_0^p \frac{y(q)}{\mu} dq$ and $\hat{L}(p) \equiv \int_0^p \frac{\hat{y}(q)}{\hat{\mu}} dq$ respectively. The slopes are $L'(p) = y(p)/\mu$ and $\hat{L}'(p) = \hat{y}(p)/\hat{\mu}$. Consider first the lower bound of the Lorenz curves, at *p*=0. Note that the slopes at this limit are $L'(p) = y_P/\mu$ and $\hat{L}'(p) = y_P/\hat{\mu}$ so $L'(p) < \hat{L}'(p)$ as $p \to 0$ given that incomes have a common lower bound $y_P > 0$ and $\mu > \hat{\mu}$ (given first-order dominance). Thus it must be the case that $L(p) < \hat{L}(p)$ in the limit, as $p \to 0$. Consider next the upper bound. As $p \to 1$, the slopes approach $L'(p) = y_R/\mu$ and $\hat{L}'(p) = y_R/\hat{\mu}$ thus, $L'(p) < \hat{L}'(p)$. Therefore it must be

the case that $L(p) > \hat{L}(p)$ at sufficiently high values of p. By continuity the Lorenz curves must intersect. Without Lorenz dominance, the qualitative implications for inequality will depend on properties of the precise measure of inequality that is used (Atkinson, 1970).

With an inverted-U relationship between compliance and income, one can get an inequality reducing effect at both low and high incomes, but either sign is possible between these extremes.

4. Estimating the income effect on nonresponse

While we do not observe the individual probabilities of compliance, we do observe both the aggregate response rates by geographic area and the incomes of complying units. This allows us to develop a model to estimate the income effect on the probability of response. The method exploits the fact that incomes observed from stratified survey data are by design representative for these geographic areas. By re-weighting the observed sample accordingly, we can impute these values non-responding households. The approach presented here is a special case of the general approach developed by Korinek et al (2004).

Consider a continuum H of households of mass M that can be partitioned into I groups H_i of observationally identical households with a vector of characteristics X_i. Assume that the set can also be partitioned along geographical lines into J subsets, labeled H_j with mass M_j each. The intersection of these two partitions can be denoted as a collection of sets $H_{ij} = H_i \cap H_j$ with weights M_{ij} each. From each of the J areas, a sample $S_j \subset H_j$ of households of mass $m_j < M_j$ is selected in order to conduct a survey on the realizations of the vector X in the total population. Since we want to investigate only the effects of survey nonresponse and not of sample design, we assume that each of the J samples S_j is representative, i.e. perfectly stratified or random. We

denote the set of households with characteristics X_i in the sample in area j as $S_{ij} \subset S_j$ and its mass as m_{ij}.

We can now define a representative sample S_j of this population as one that comprises households of all groups in area j and in which the total weight m_{ij} of sampled households of each type i is proportional to M_{ij}. Clearly, the sum of all m_{ij} for a given area j must be $\sum_i m_{ij} = m_j$. For each household $\xi \in S_{ij}$, there is a Bernoulli variable $D_{ij\xi}$ with the realization $D_{ij\xi} = 1$ if the household responds to the survey and $D_{ij\xi} = 0$ in case of nonresponse. We assume that these random variables are i.i.d. within one observationally identical group i of households and independent across groups. The probability that the household responds is denoted as:

$$P(D_{ij\xi} = 1|X_i, \theta) = P_i \tag{10}$$

where θ is an unknown parameter vector from a compact parameter space. Note that our i.i.d. assumption on the random variables within a group of households implies that we can omit the subscripts j and ξ in P_i. We assume that the probability of a household to respond has a stable parametric form and is given by the following logistic function:

$$P\left(D_{ij\xi} = 1|X_i, \theta\right) = \frac{e^{X_i\theta}}{1 + e^{X_i\theta}} \tag{11}$$

Denote the mass of all respondents in group i and area j as the random variable $m_{ij}^1 \in [0, m_{ij}]$, which can be expressed as $m_{ij}^1 = \int_0^{m_{ij}} D_{ij\xi} d\xi$. Its expected value is:

$$E\left[m_{ij}^1\right] = m_{ij} \cdot P_i \tag{12}$$

Note that in this equation, the total mass m_{ij} of households in group i is unobservable. However, in order to establish an estimation method based on that equation, we can proceed as follows: First, divide (12) through the probability P_i to arrive at the expression:

$$E\left[\frac{m_{ij}^1}{P_i}\right] = m_{ij} \tag{13}$$

Then, let us denote the sum of all the fractions m_{ij}^1/P_i for a given j minus their expected value as:

$$\psi_j(\theta) = \sum_i \left\{ \frac{m_{ij}^1}{P_i} - E\left[\frac{m_{ij}^1}{P_i}\right] \right\} = \sum_i \left\{ \frac{m_{ij}^1}{P_i} - m_{ij} \right\} = \sum_i \frac{m_{ij}^1}{P_i} - m_j \tag{14}$$

where the m_j, the mass of the sample in geographical area j, is now known. By the law of iterated expectations, the expected value $E[\psi_j(\theta)] = 0$. Finally, we can stack the moment conditions $\psi_j(\theta)$ for all geographical areas j into a vector $\Psi(\theta)$, which allows us to estimate the unknown parameter θ using a minimum distance estimator of the form:

$$\hat{\theta} = \arg\min_{\theta} \Psi(\theta)' W^{-1} \Psi(\theta) \tag{15}$$

where W is the positive definite weighting matrix.[9]

The most efficient weighting matrix W is the covariance matrix of the vector $\Psi(\theta)$, or any matrix proportional to it (Hansen, 1982).[10] The GMM approach to deriving this weighting matrix would be to calculate the sample covariances of all the individual moment conditions (14). In our setup, however, the m_{ij}'s in these equations are unobservable, and we know only their aggregates. Hence we proceed as follows.

By our assumption of independence of the response decisions of all households between states we set the off-diagonal elements of the covariance matrix to zero, and concentrate on the diagonal elements. We assume that the variance of $\psi_j(\theta)$ of each state j is proportional to the mass of the sampled household population, with a factor of proportionality of σ^2:

---

[9] Korinek et al. (2004) describe the technical properties of W under which the estimator (15) will be consistent.

[10] To be precise, the described estimator does not fall into the category of GMM estimators, since the variable m_{ij} is unobservable. We can thus only use the aggregates $\psi_j(\theta)$ thereof. However, the discussion of Hansen (1982) applies analogously.

$$Var\left(\psi_j(\theta)\right) = m_j \cdot \sigma^2 \tag{16}$$

The factor of proportionality, which can also be interpreted as the variance for a sample of weight one, can be estimated using the expression:

$$\hat{\sigma}^2 = \frac{\sum \psi_j(\theta)^2}{\sum w_j} \tag{17}$$

Since the term σ^2 shows up in all the elements of our constructed variance-covariance matrix, we can simply cancel it out for the purpose of the maximization problem and use the weighting matrix $W=m_jI_J$ where I_J is a J dimensional identity matrix.

Finally, we note that the covariance matrix of $\Psi(\theta)$ is $\sigma^2 W$. Since $\Psi(\theta)$ is twice continuously differentiable, we can then express the asymptotic covariance matrix of our estimator $\hat{\theta}$ as:

$$\hat{Var}\left(\hat{\theta}\right) = \hat{\sigma}^2 \left[\frac{\partial \Psi(\theta)'}{\partial \theta} W^{-1} \frac{\partial \Psi(\theta)}{\partial \theta} \right]^{-1} \tag{18}$$

5. Application to the U.S. income distribution

Data on survey response rates across geographical areas are often available from survey producers. A case in point is the supplement of the US Current Population Survey (CPS).[11] In addition to detailed data on incomes, the CPS contains geographically referenced information on non-compliance (Census Bureau, 2002, Chapter 7). The survey contains one record for each household in the sample, i.e. for responding households as well as for "non-interview" households. It distinguishes the non-interview households by the reason for the non-interview into categories A, B, and C. Types B and C non-interviews refer to housing units that are vacant

[11] The CPS data and survey methodology details are available from the US Census Bureau and can be accessed on-line at: http://www.census.gov/hhes/www/income.html.

or that were demolished. We excluded these from our data set. Type A non-interviews comprise households that explicitly refused to be interviewed or that could not be interviewed because nobody was at home. We regard these type A households as non-responding. In the March 2004 supplement to the CPS, with a sample size of 84,116 households, the number of non-responding households totals 6,967 implying a nonresponse rate of 8.28 percent. Korinek et al., (2004) describe the survey in greater detail, including an analysis of the weights used to correct for nonresponse by the Census Bureau.

Since the CPS was designed to be representative of the US state level, we can use the 51 states as the geographical areas in our estimation methodology indexed by j. It can be seen from Table 1 that in 2004, nonresponse rates varied from 3.4% in Alabama to 15.3% in the District of Columbia. Figure 2 shows that the average state-level income is negatively correlated with response. This suggests that survey response falls with income.[12]

Table 2 gives results for various parametric forms using the 2004 CPS. We give both the parameters estimates and the corrected Gini index, as a single summary measure of inequality. The uncorrected Gini index for 2004 is 44.80% (45.20% using the CPS internal weights). Korinek et al., (2004) describe methods and results on our choice of parametric model. The specification $P = logit(\theta_1 + \theta_2 y)$ performed best by the Akaike Information Criterion for the 2004 dataset, which is our preferred parametric form in the following discussion.[13] However, it can be seen from Table 2 that our corrections to the Gini index are quite robust to the choice of

[12] Note that the simple regression shown in figure XX takes only the variation of mean income across states into account. The estimation method that we developed in the previous section incorporates the income distribution within each state.

[13] There is no reason to assume that one specific functional form performs best at explaining nonresponse across all years, since response behavior can change over time. In applying our estimation method outlined in the previous chapter, we thus suggest to perform a separate specification test on each dataset on which the methodology is applied. When we combined all data from 1998 to 2004 into one dataset (with incomes deflated to 1998 prices) we found that the specification $P = logit(\theta_1 ln(y) + \theta_2 ln(y)^2)$ performs best.

specification. The choice of specification does not alter the correction of the Gini coefficient significantly: all corrected coefficients are within one standard deviation of each other, between 49.23% to 49.76%.

A plot of the functional relationship between compliance and income in 2004 according to our estimate from our preferred specification is given in Figure 3. The implication that compliance falls monotonically with income is consistent with other evidence for the U.S. (Groves and Couper, 1998, Chapter 5, based on compliance with the long schedule of the U.S. Census administered to a random sample).

The effect on the distribution of income per capita can be seen from Figures 4a and 4b. The uppermost line in Figure 4a shows the uncorrected income distribution, i.e. the observed distribution if all individuals in a given state are assigned an equal weight, which consists of the population divided by the size of the sample in the given state. We also give results using the weights supplied with the CPS. It can be seen that both the corrected CPS weights and our estimate for a corrected income distribution first order dominate the measured distribution. For the CPS weights, this dominance seems to be particularly strong for relatively lower-income households. For our estimation methodology, the correction, and thus the first-order dominance, is stronger for higher income levels.

Our results indicate that ignoring selective compliance according to income appreciably understates the proportion of the population in the richest income quantiles and slightly overstates the population shares in lower quantiles. What is observed as the highest income percentile in the survey, for example, is estimated to comprise 1.73% of the population after correcting for its lower probability of survey compliance, and the highest observed decile makes actually up for 11.1% of the population. By contrast, the poorest observed decile and percentile

in the unadjusted data actually comprise only 9.8% and 0.98% respectively of the corrected population. The correction method of the Census Bureau assigns 1.60% and 14.74% of the population weight respectively to the top observed percentile and decile, and 0.88% and 6.95% to the bottom decile and percentile. Table 2 also gives the corrected population shares of the richest decile and richest percentile for various parametric specifications. Depending on specification, we estimate that the richest 10% in of the population based on the CPS contain 11-13%, while the richest percentile in the CPS actually represent about 2% of the population.

Using our correction method, median income per person rises from an uncorrected $16,096 to $16,410, while the mean increases from an uncorrected $22,039 to $24,722 per capita. Using the weights provided by the Census Bureau, median income rises to $19,333, and mean income to $26,958.

Figure 4b shows a magnification of the lower 30% of the distribution. It can be seen that using our correction method, the impact on poverty incidence is small for poverty lines commonly used in the U.S., giving poverty rates around 12% (Census Bureau, 2001). However, since there is first-order dominance, poverty measures using the uncorrected, equally-weighted distribution of incomes unambiguously overestimate poverty. Note that the correction methodology of the Census Bureau leads to a significant increase in the estimated level of poverty, however.

Figure 5a depicts the Lorenz curves for the uncorrected income distribution, the distribution according to the Census Bureau's weights, and according to our correction method. The effect of our correction for selective response is a marked downward shift in the Lorenz curve, implying higher inequality. Using the Census Bureau's correction weights, in contrast, hardly has an effect on the Gini coefficient. As our theoretical analysis predicts, we do not find

Lorenz dominance. However, the intersection occurs at the extreme upper end, and only a strong magnification as in Figure 5b makes it visible. The Gini index increases by 3.66% from 44.80% to 49.56% (std. dev. 0.62%) on correcting for our estimates of the income effect on compliance. In contrast, using the CPS weights increase the Gini coefficient to only 45.20% in our dataset, i.e. the Census Bureau's correction method does not seem to significantly affect measured inequality.

By inverting the CDF to obtain the quantile function for the original distribution, the correction according to Census Bureau weights, and the corrected distribution according to our method, we can calculate the income correction at each percentile of income. The results are given in Figure 6. For the Census Bureau's correction, income at any given percentile shifts up almost uniformly by about 20%. This implies that the Census Bureau's correction method affects the national average, but is almost distribution neutral, as we found already in our comparison of Gini coefficients. For our method, the correction is quite low (around +2 to +3%) for the bottom 9 deciles and then rises sharply, to reach almost +100% for the uppermost percentile.

Figure 7 depicts the weight correction of each observed income percentile. This figure reveals why the Census Bureau's correction method has almost no effect on the Gini coefficient: their methodology heavily reduces the weights of low-income individuals (by almost 40% for some of the bottom percentiles) and attributes this weight the uppermost third of the income distribution. Our method, in contrast, reduces the weight of bottom four-fifth only by up to 3%, and redistributes this weight mainly to the top percentiles.

Table 3 gives results for various augmented specifications on the 2004 data. The additional household characteristics we considered were household size (*hhsize*), a dummy variables for whether the interviewed household is located in a metropolitan area (I_{MSA}) and

whether the household owned the house/apartment in which it lives ($I_{homeownership}$). In addition, we included various characteristics of the household head, such as sex (dummy variable I_{female}) race ($I_{caucasian}$), employment status (dummy variables for $I_{working}$ and $I_{unemployed}$), education (measured by an index that indicates the years of schooling, i.e. *edu*; and alternatively by dummy variables for attaining different levels of education, of which attaining a graduate degree was most significant, i.e. $I_{edu \geq master}$), and *age*, which we use both as a level and squared. Our estimated correction to the Gini index turns out to be quite robust to these changes.

So far we have focused on the 2004 CPS. Table 4 gives estimates of the correction in the Gini index for all CPSs from 1998 to 2004. The implied upward corrections of the Gini index (as compared to the Gini calculated from the uncorrected distribution) are of similar magnitudes, ranging from 3.39% points in 2000 to 5.74% in 1998. There is no clear pattern over time.

6. Conclusions

We have argued that there is likely to be an income effect on survey compliance, with implications for measures of poverty or inequality. Past empirical work has either ignored the problem of selective compliance in surveys or made essentially *ad hoc* corrections, often involving the assumption that nonresponse is negligible within certain subgroups. We have demonstrated that these methods will generally result in a bias. Consequently, we have shown how the latent income effect on compliance can be estimated consistently with the available data on average response rates and the measured distribution of income across geographic areas. This allowed us to re-weight the raw data to arrive at a corrected income distribution. Our correction method for nonresponse should not be thought of as a replacement for the methods currently employed by statistical agencies, such as various poststratification steps used by the U.S. Census Bureau, but it should be seen rather as complementary. Indeed, it should be easy to extend the

econometric model that we presented in section 4 so that $\Psi(\theta)$ can also include moment conditions that perform the poststratification.

On implementing our method using U.S. data, we find that the problem is not ignorable. We can also reject the assumptions made in past *ad hoc* correction methods. We find a highly significant negative income effect on survey compliance. While we do not find strict Lorenz dominance, measured inequality is considerably higher after correcting for selective compliance. Thus we find that unit nonresponse has the opposite impact on inequality to the problem of classical measurement error in reported incomes that has been studied in past work in the literature. An upward revision to the overall mean is also called for to correct for selective compliance. In terms of the impact on measures of poverty, the downward bias in the mean tends to offset the downward bias in measured inequality. The tendency for low-income groups to be over-represented (because of their higher compliance probabilities) still means that the poverty rate tends to be over-estimated, though the impact is small up to poverty lines normally used in the U.S.

References

Alho, J.M. (1990) "Adjusting for Nonresponse Bias using Logistic Regression," *Biometrica*, 77(3): 617-24.

Atkinson, A. B., (1970) "On the Measurement of Inequality," *Journal of Economic Theory*, 2: 244-263.

Atkinson, A.B. (1987) "On the Measurement of Poverty," *Econometrica* 55: 749-764.

Atkinson, A.B. and J. Micklewright (1983) "On the Reliability of Income Data in the Family Expenditure Survey 1970-1977," *Journal of the Royal Statistical Society Series A*, 146(1): 33-61.

Bhalla, Surjit (2002) Imagine There's No Country: Poverty, Inequality and Growth in the Era of Globalization, Washington DC.: Institute for International Economics.

Bourguignon, Francois and Christian Morrisson (2002) "Inequality Among World Citizens: 1820-1992," *American Economic Review* 92(4): 727-744.

Bureau of Labor Statistics (1981), "Urban Family Budgets and Comparative Indexes for Selected Urban Areas," U.S. Department of Labor.

Census Bureau (2001) "Poverty in the United States: 2001" Current Population Report P60-219, Washington, D.C: U.S. Department of Commerce.

_____ (2002) "Current Population Survey – Design and Methodology," Technical Paper 63RV. Washington, D.C: U.S. Department of Commerce.

Chakravarty, S.R., and W. Eichhorn (1994) "Measurement of Income Inequality: Observed versus True Data," in W. Eichhorn (ed.) *Models and Measurement of Welfare and Inequality*, Heidelberg: Springer-Verlag.

Chesher, A., and C. Schluter (2002) "Welfare Measurement and Measurement Error," *Review of Economic Studies*, 69: 357-378.

Cowell, F.A., and M. Victoria-Feser (1996) "Robustness of Inequality Measures," *Econometrica,* 64: 77-101.

Deming, W.E. (1953) "On a Probability Mechanism to Attain an Economic Balance between the Resultant Error of Response and the Bias of Nonresponse," *Journal of the American Statistical Association*, 48: 743-72.

Eichhorn, W., H. Funke, and W.F. Richter (1984) "Tax Progression and Inequality of Income Distribution," *Journal of Mathematical Economics* 13: 127-131.

Friar, Monica E. and Herman B. Leonard (1998), "Variations in Cost of living Across States," Taubman Center for State and Local Government, John F. Kennedy School, Harvard University.

Groves, Robert E., and Mick P. Couper (1998) *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons.

Hansen, L. P. (1982) "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica 50(4): 1029-54.

Holt, D., and D. Elliot (1991) "Methods of Weighting for Unit Nonresponse," *The Statistician,* 40: 333-342.

Korinek, Anton, Johan Mistiaen and Martin Ravallion (2004) "An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys," Policy Research Working Paper, World Bank. forthcoming.

Lillard, L., Smith, J.P. and F. Welch (1986) "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy*, 94(3):489-506.

Little, R.J.A. and D.B. Rubin (1987) *Statistical Analysis with Missing Data*. New York: Wiley.

Nijman, T. and M. Verbeek (1992) "Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function," *Journal of Applied Econometrics*, 7:243-57.

Peracchi, F. (2001) "Econometrics," New York: Wiley.

Ravallion, M. (1994) "Poverty Rankings Using Noisy Data on Living Standards," *Economics Letters*, 45: 481-485.

_____. 2000, "Should Poverty Measures be Anchored to the National Accounts?" *Economic and Political Weekly* 34 (August 26): 3245-3252.

_____, 2002, "Measuring Aggregate Welfare in Developing Countries: How Well do National Accounts and Surveys Agree?," *Review of Economics and Statistics*, 85: 645-652.

Sala-i-Martin, Xavier (2002), "The World Distribution of Income (Estimated from Individual Country Distributions)," NBER Working Paper No. W8933.

Scott, Kinnon and Diane Steele (2004), "Measuring Welfare in Developing Countries: Living Standards Measurement Study Surveys," in UN Statistical Division, *Surveys in Developing and Transition Countries*, forthcoming.

Van Praag, B., A. Hagenaars and W. Van Eck (1983) "The Influence of Classification and Observation Errors on the Measurement of Income Inequality," *Econometrica*, 51: 1093-1108.

Table 1: Summary data by state for the March 2004 CPS, sorted by response rate

State	Response rate (%)	Sample size		Income per capita ($)	Median income ($)
		Households	Individuals		
Alabama	96.47	1,189	2,981	19,915	15,183
North Dakota	96.03	1,082	2,725	18,925	15,415
Indiana	95.73	1,500	3,927	21,585	16,667
South Dakota	95.53	1,164	3,087	19,676	14,763
Utah	95.35	1,010	3,334	19,040	14,205
Wisconsin	95.29	1,528	4,083	21,653	17,294
Arkansas	95.29	976	2,442	17,401	12,704
Montana	94.60	871	2,120	17,886	13,013
Georgia	94.55	1,175	2,943	20,609	16,049
Iowa	93.69	1,379	3,487	20,940	16,904
Louisiana	93.67	979	2,477	17,209	12,550
Florida	93.51	3,680	8,936	21,545	15,400
Kansas	93.41	1,441	3,743	20,898	16,085
Wyoming	93.35	1,128	2,841	20,567	15,561
Illinois	93.28	2,945	7,703	22,846	16,898
Arizona	93.23	1,167	3,144	21,118	13,750
Nevada	93.23	1,594	4,191	21,664	15,999
Delaware	93.16	1,082	2,741	23,637	18,039
Oklahoma	93.12	1,047	2,451	19,006	13,667
West Virginia	92.91	1,170	2,836	17,026	13,150
Mississippi	92.81	904	2,281	17,530	13,440
Idaho	92.81	973	2,670	18,079	12,494
Minnesota	92.51	1,535	4,078	25,282	19,194
Nebraska	92.47	1,302	3,445	21,040	16,086
Kentucky	92.18	1,138	2,828	19,270	14,700
Pennsylvania	92.14	2,964	7,522	22,813	17,385
Missouri	92.04	1,269	3,158	21,571	16,251
Virginia	92.04	1,470	3,699	24,789	19,322
Tennessee	91.62	1,014	2,483	18,703	14,167
Texas	91.51	3,864	10,250	18,932	12,547
Colorado	91.50	1,788	4,579	23,864	17,816
Massachusetts	91.49	1,540	3,816	26,888	19,856
Michigan	91.46	2,319	5,908	22,206	16,700
Rhode Island	91.44	1,518	3,831	22,611	17,018
Maine	91.44	1,366	3,304	19,393	15,098
Connecticut	91.36	1,574	3,976	27,367	20,779
Ohio	91.34	2,517	6,208	22,128	17,102
North Carolina	90.78	1,811	4,323	19,794	14,251
South Carolina	90.53	1,162	2,721	19,785	14,904
Hawaii	90.53	1,193	3,297	23,126	17,377
New Mexico	90.46	1,090	2,684	18,171	12,000
Washington	90.19	1,509	3,658	23,251	17,751
California	90.06	5,984	16,269	21,915	14,908
Oregon	89.99	1,289	3,074	20,873	15,442
Vermont	89.04	1,277	3,017	23,174	17,710
Alaska	88.64	1,206	3,190	21,622	16,523
New Hampshire	88.50	1,400	3,602	26,602	20,367
New Jersey	88.50	2,200	5,558	27,250	20,208
Maryland	88.00	1,408	3,294	28,177	20,255
New York	87.56	4,245	10,257	22,657	16,141
District of Columbia	84.66	1,180	2,069	30,014	17,210

Table 2: Estimation results of various specifications for 2004 data

Specification	θ_1	θ_2	θ_3	AIC	Gini	Top 10%	Top 1%
3: $z = \theta_1 + \theta_2\, ln(y)$	19.112 (1.708)	-1.613 (0.155)		-23.881	49.23 (0.92)	12.95%	2.22%
5: $z = \theta_1 + \theta_2\, ln(y)^2$	10.108 (0.747)	-0.072 (0.006)		-24.717	49.41 (0.90)	12.81%	2.22%
6: $z = \theta_1\, ln(y) + \theta_2\, ln(y)^2$	1.809 (0.116)	-0.152 (0.010)		-25.690	49.60 (0.87)	12.63%	2.22%
7: $z = \theta_1 + \theta_2\, ln(y) + \theta_3\, ln(y)^2$	-1.157 (9.791)	2.017 (1.766)	-0.161 (0.079)	-23.738	49.63 (0.93)	12.60%	2.21%
9: $z = \theta_1 + \theta_2\, y$	2.900 (0.055)	$-1.232*10^{-5}$ $(4.368*10^{-7})$		-26.785	49.56 (0.62)	11.07%	1.73%
10: $z = \theta_1\, ln(y) + \theta_2\, y$	0.300 (0.005)	$-1.448*10^{-5}$ $(4.361*10^{-7})$		-24.339	49.60 (0.60)	10.69%	1.72%
11: $z = \theta_1 + \theta_2\, ln(y) + \theta_3\, y$	7.968 (3.878)	-0.511 (0.386)	$-8.704*10^{-6}$ $(2.755*10^{-6})$	-26.157	49.62 (0.69)	11.82%	1.92%
13: $z = \theta_1 + \theta_2\, ln(y)^2 + \theta_3\, y$	5.396 (1.896)	-0.025 (0.019)	$-8.221*10^{-6}$ $(3.072*10^{-6})$	-26.212	49.66 (0.69)	11.82%	1.93%
14: $z = \theta_1\, ln(y) + \theta_2\, ln(y)^2 + \theta_3\, y$	1.075 (0.361)	-0.079 (0.036)	$-7.199*10^{-6}$ $(3.328*10^{-6})$	-26.229	49.76 (0.70)	11.82%	1.95%

Note: Standard errors are in brackets. The probability of response is modeled as $P = logit(z)$ for all given models. We calculated the Akaike Information Coefficient (AIC) calculated as $AIC = J \cdot \log\left(\sum \psi_j^2\left(\hat{\theta}\right)\Big/ J\right) + 2m$, where J is the number of states, i.e. 51 here, and m is the number of estimated parameters. The lowest value (i.e. here the highest absolute value) indicates the specification that best fits the data for a given year. In the table above we have underlined these AIC. Note that the uncorrected Gini coefficient for 2004 data (equally weighted within states) is 44.80%, and using the official CPS weights it is 45.20%. The columns "top 10%" and "top 1%" indicate the fraction of the population that is estimated to be located in what is the observed top income decile and percentile in each specification. The corresponding shares in the unadjusted data are 9.80% and 0.98%.

Table 3: Augmented specifications for 2004 data

Specification	θ_1	θ_2	θ_3	θ_4	θ_5	AIC	Gini
$z = \theta_1 + \theta_2 y$ [baseline]	2.900 (0.055)	$-1.232*10^{-5}$ $(4.368*10^{-7})$				-26.785	49.56 (0.62)
$z = \theta_1 + \theta_2 y + \theta_3 hhsize$	2.589 (0.611)	$-1.197*10^{-5}$ $(8.156*10^{-7})$	0.1183 (0.2416)			-25.048	49.54 (0.62)
$z = \theta_1 + \theta_2 y + \theta_3 I_{MSA}$	2.974 (0.111)	$-1.214*10^{-5}$ $(5.067*10^{-7})$	-0.1174 (0.1463)			-25.515	49.39 (0.67)
$z = \theta_1 + \theta_2 y + \theta_3 I_{homeownership}$	4.544 (4.866)	$-1.223*10^{-5}$ $(4.627*10^{-7})$	-1.686 (4.914)			-25.596	49.51 (0.64)
$z = \theta_1 + \theta_2 y + \theta_3 I_{female}$	3.365 (0.460)	$-1.172*10^{-5}$ $(6.097*10^{-7})$	-0.8995 (0.7214)			-27.040	48.94 (0.73)
$z = \theta_1 + \theta_2 y + \theta_3 I_{caucasian}$	2.750 (0.174)	$-1.204*10^{-5}$ $(5.657*10^{-7})$	0.1834 (0.2065)			-25.593	49.26 (0.73)
$z = \theta_1 + \theta_2 y + \theta_3 I_{working}$	2.694 (0.545)	$-1.252*10^{-5}$ $(7.934*10^{-7})$	0.3160 (0.8917)			-24.933	49.55 (0.62)
$z = \theta_1 + \theta_2 y + \theta_3 I_{unemployed}$	3.043 (0.204)	$-1.253*10^{-5}$ $(5.494*10^{-7})$	-1.793 (1.191)			-25.535	49.56 (0.64)
$z = \theta_1 + \theta_2 y + \theta_3 edu$	6.154 (1.419)	$-1.010*10^{-5}$ $(1.061*10^{-6})$	-0.3207 (0.1289)			-31.771	48.94 (0.67)
$z = \theta_1 + \theta_2 y + \theta_3 I_{edu \geq master}$	3.145 (0.174)	$-9.374*10^{-6}$ $(1.309*10^{-6})$	-1.645 (0.567)			-31.926	48.24 (0.71)
$z = \theta_1 + \theta_2 y + \theta_3 age$	3.115 (0.724)	$-1.232*10^{-5}$ $(4.367*10^{-7})$	-0.00441 (0.01465)			-24.922	49.61 (0.65)
$z = \theta_1 + \theta_2 y + \theta_3 age + \theta_4 age^2$	29.626 (12.776)	$-1.213*10^{-5}$ $(6.840*10^{-7})$	-1.048 (0.493)	0.00989 (0.00470)		-27.711	49.47 (0.63)
$z = \theta_1 + \theta_2 y + \theta_3 age + \theta_4 age^2 + \theta_5 edu$	60.906 (26.147)	$-9.196*10^{-6}$ $1.164*10^{-6}$	-1.569 (0.932)	0.0146 (0.0086)	-0.410 0.111	-38.065	49.19 (0.86)
$z = \theta_1 + \theta_2 y + \theta_3 age + \theta_4 age^2 + \theta_5 I_{edu \geq master}$	40.496 (22.731)	$-9.122*10^{-6}$ $(1.344*10^{-6})$	-1.451 (0.853)	0.0137 (0.0079)	-1.716 (0.435)	-38.456	48.74 (0.80)

Note: standard errors in brackets

Table 4: Specification 3, $P = logit(\theta_1 + \theta_2 ln(y))$, over time

Year	θ_1	θ_2	Gini (%)			Correction to Gini index (%)
			Uncorrected	CPS	Corrected	
1998	19.904	-1.696	45.52	45.32	50.96	5.74
	(2.071)	(0.188)			(1.25)	(1.25)
1999	18.099	-1.528	45.22	44.87	48.98	3.76
	(2.420)	(0.223)			(1.17)	(1.17)
2000	22.207	-1.889	44.43	44.31	47.82	3.39
	(2.545)	(0.230)			(0.79)	(0.79)
2001	20.111	-1.702	45.14	45.01	49.61	4.47
	(1.728)	(0.156)			(0.81)	(0.81)
2002	17.807	-1.489	44.98	45.51	48.69	3.71
	(1.920)	(0.176)			(0.92)	(0.92)
2003	17.388	-1.454	44.79	45.25	48.88	4.09
	(2.100)	(0.193)			(1.20)	(1.20)
2004	19.113	-1.613	44.80	45.20	49.23	4.43
	(1.708)	(0.155)			(0.92)	(0.92)
All	18.838	-1.599	44.99	45.07	49.14	4.15
	(0.793)	(0.073)			(0.40)	(0.40)

Note: standard errors in brackets. This specification best fit the data (based on AIC) in 3 of the 7 years.

Figure 1: Pattern of bias for an inverted-U relationship between compliance and income

$F(y) - \hat{F}(y)$

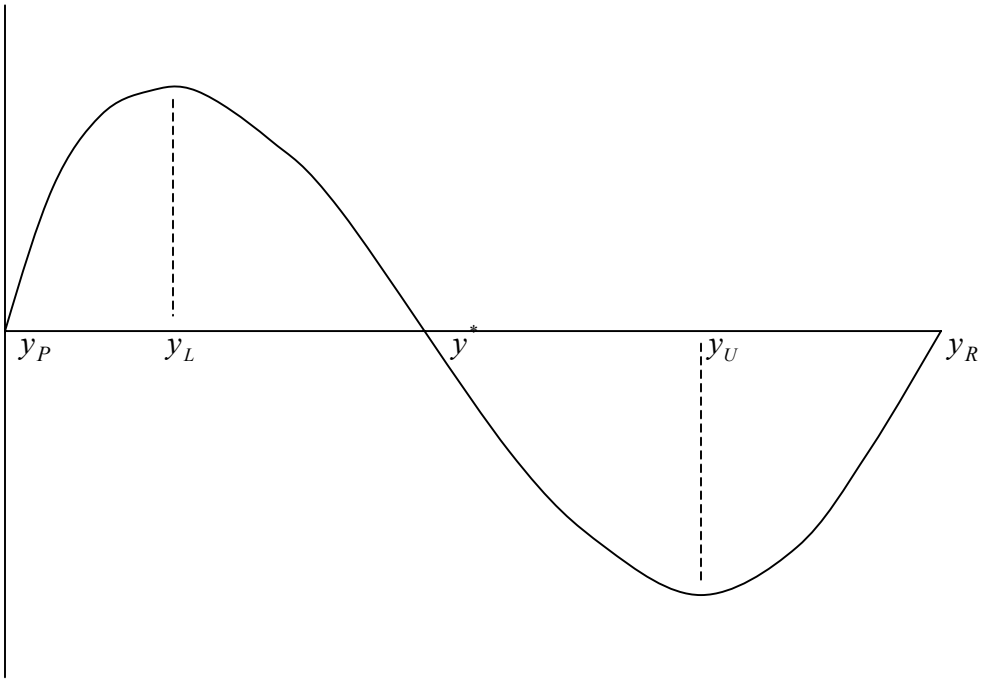


Figure 2: Probability of response against per-capita income by state

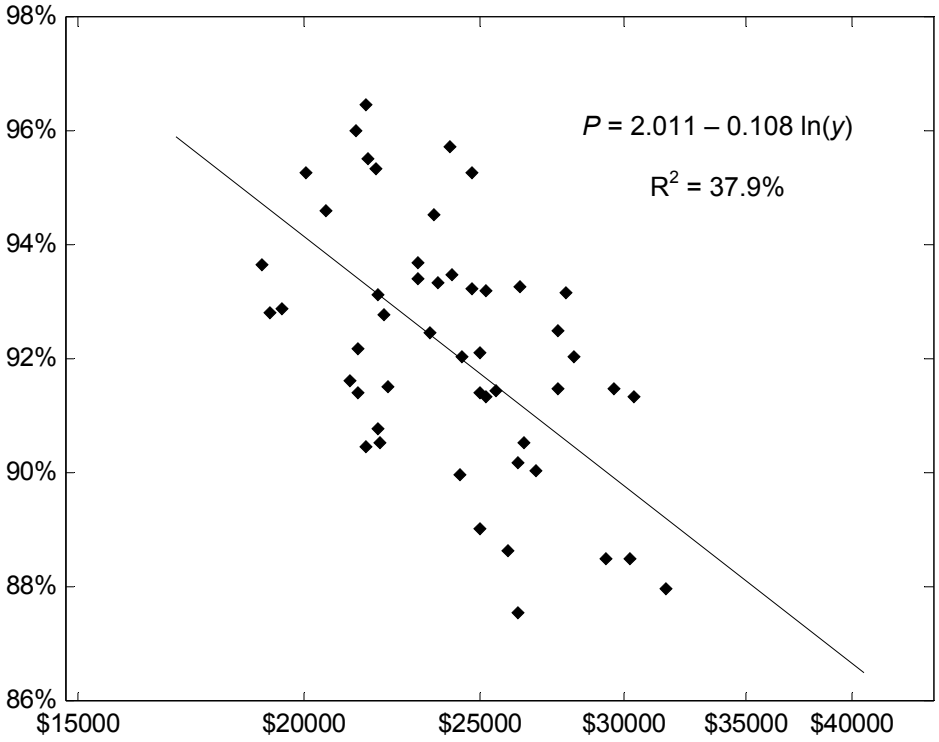
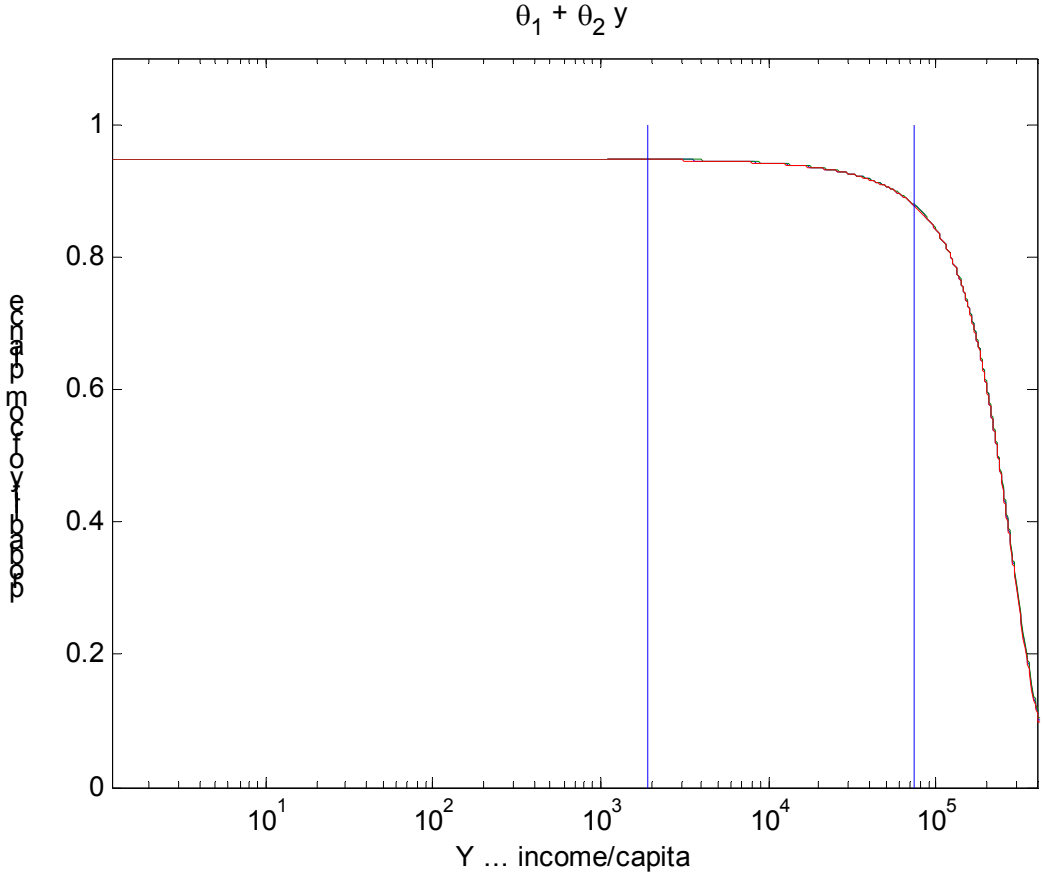
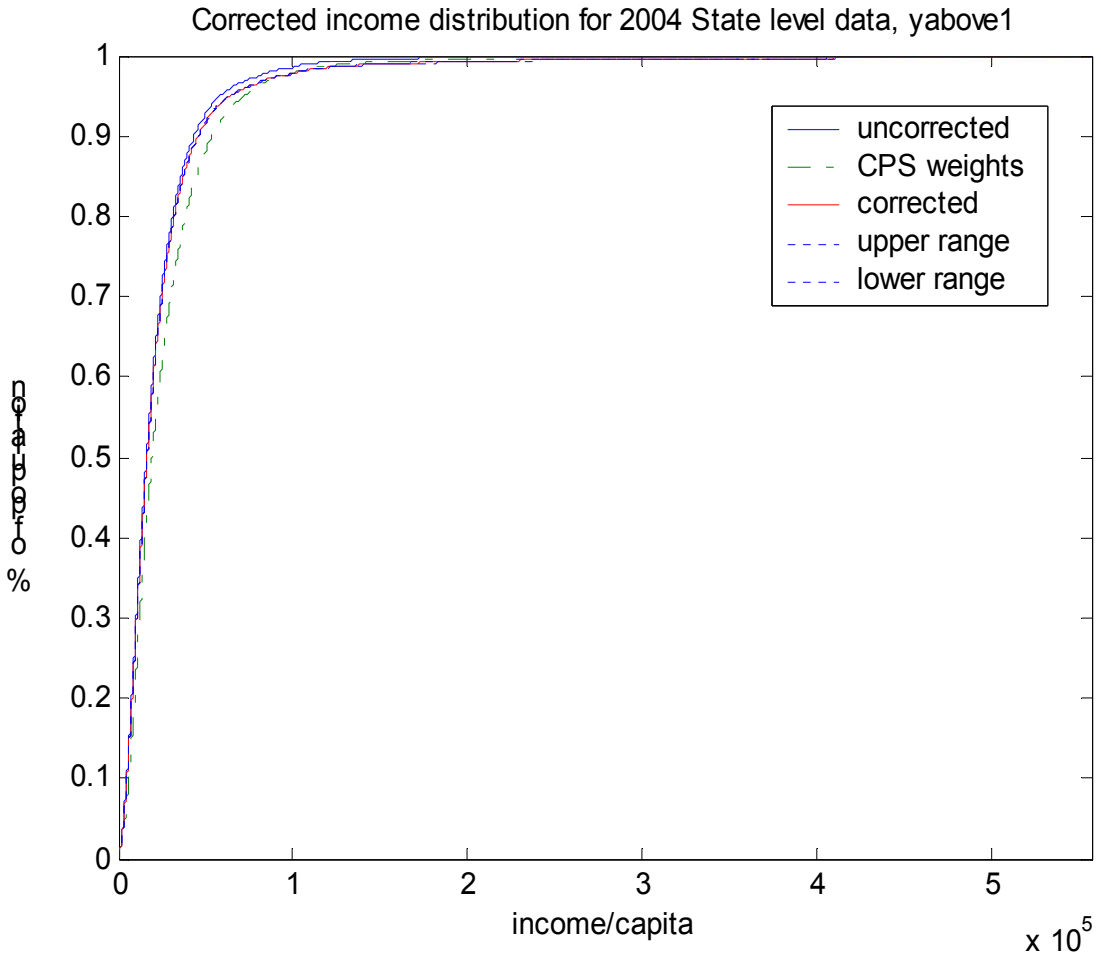


Figure 3: Probability of response as a function of income (2004)



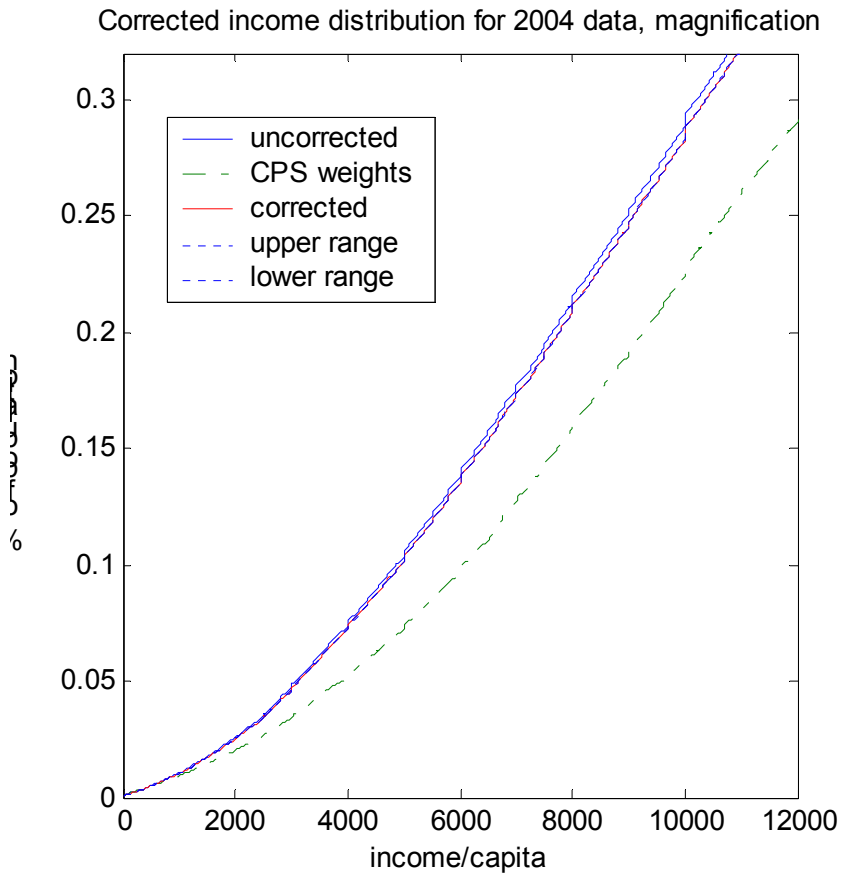
Note: A 95% confidence interval around the estimated probability of response function is included, but almost coincides with the function shown in the figure. The two vertical lines indicate the interval in which the median 95% of income observations are located. This is a plot of specification 9 in table 2.

Figure 4a: Empirical and compliance corrected cumulative income distribution



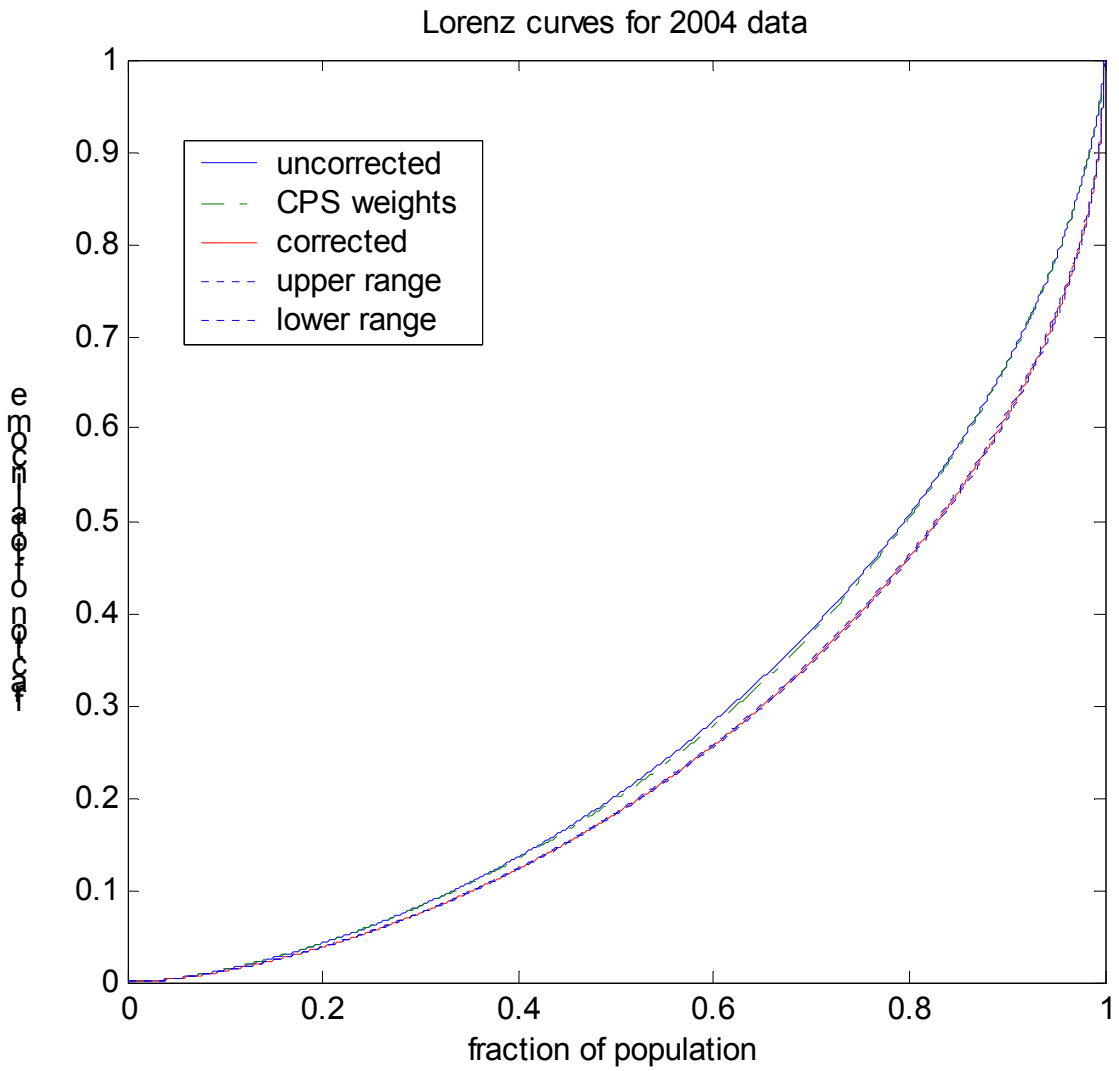
Note: Two dotted lines around the original income distribution curve depict a 95% confidence interval, but these lines almost coincide with the correction distribution curve.

Figure 4b: Lower segment of cumulative income distribution from Figure 4a



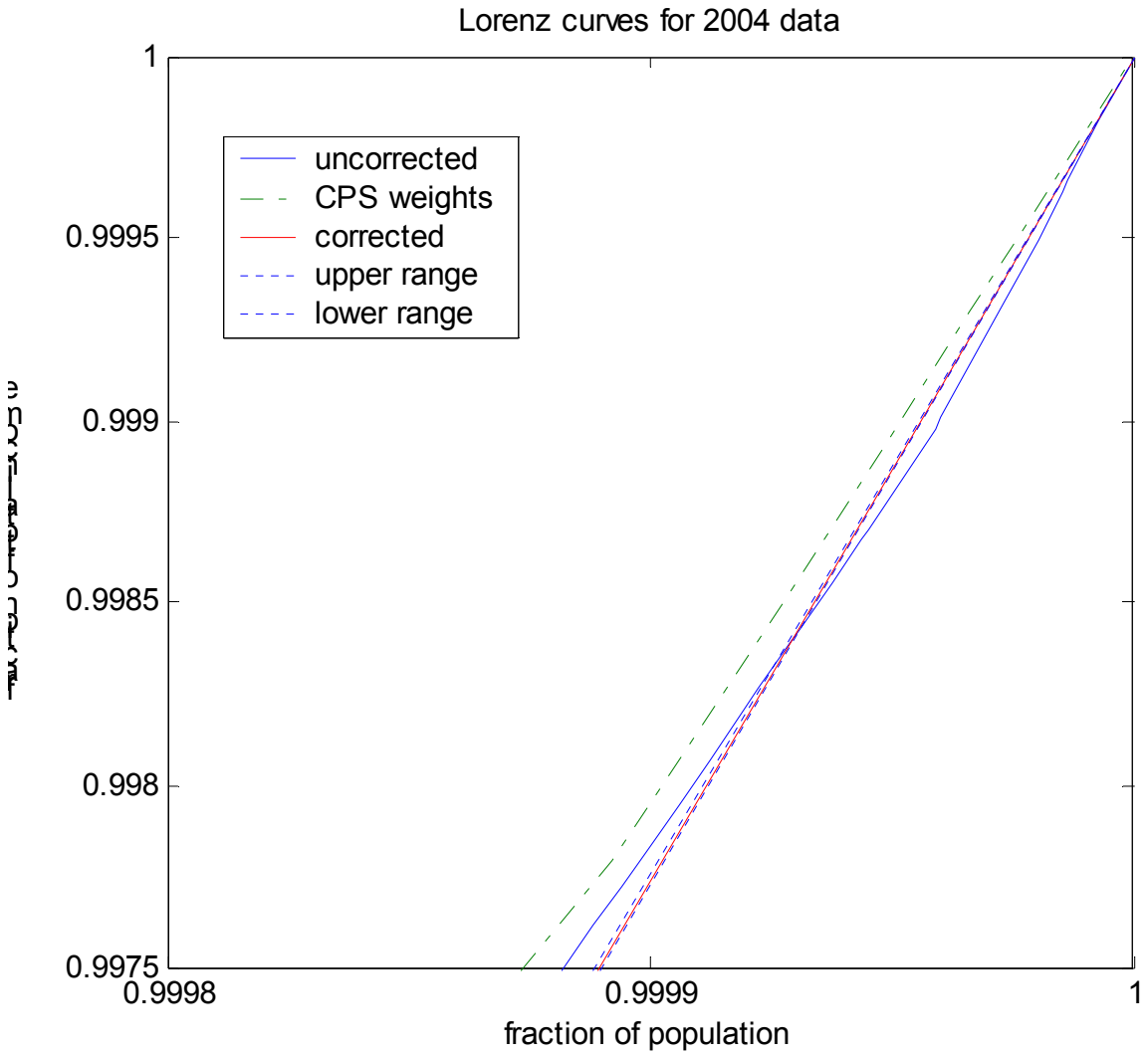
Note: The two dotted lines around the original income distribution curve depict a 95% confidence interval.

Figure 5a: Observed and corrected Lorenz curves



Note: The dotted lines around the corrected Lorenz curve depict a 95% confidence interval.

Figure 5b: Magnification of upper right part: Lorenz curves intersect



Note: The dotted lines around the corrected Lorenz curve depict a 95% confidence interval.

Figure 6: Percentage correction of income by percentile of income distribution

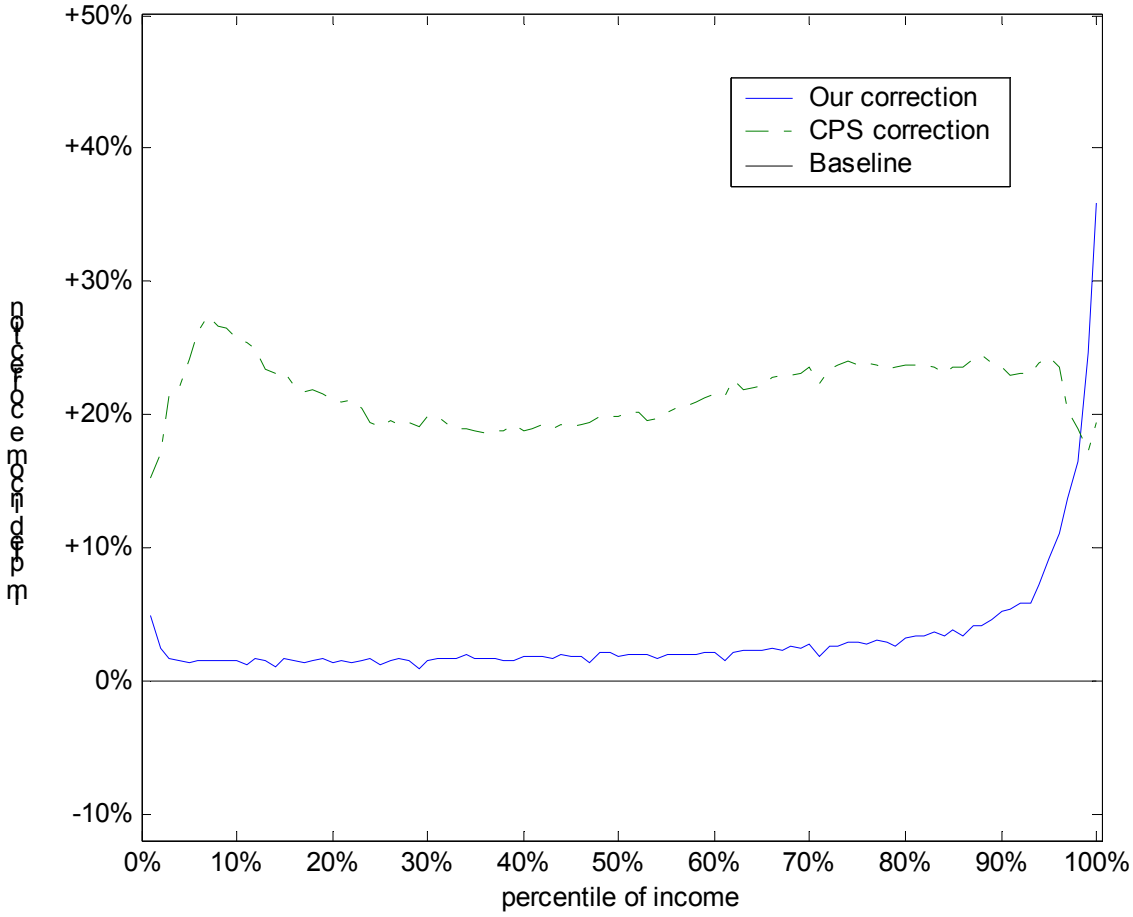


Figure 7: Weight correction for each observed percentile