ESTIMATING INCOME MOBILITY WHEN INCOME IS MEASURED WITH ERROR: THE CASE OF SOUTH AFRICA

Rulof P. Burger[1,2], Stephan Klasen[3] and Asmus Zoch[1]

ABSTRACT

There are long-standing concerns that household income mobility is over-estimated due to errors in income measures, especially in developing countries where collecting reliable survey data is often difficult. We propose a GMM estimator that exploits the existence of three waves of panel data to simultaneously estimate the extent of income mobility and the reliability of the income measure. Our estimator is more efficient than 2SLS estimators used in other studies and produces over-identifying restrictions that can be used to test the validity of our identifying assumptions. We also introduce a nonparametric generalisation in which both the speed of income convergence and the reliability of the income measure varies with the initial income level. This approach is applied to a three-wave South African panel dataset. The results suggest that the conventional method over-estimates the extent of income mobility by a factor of 4 and that about 20% of variation in reported household income is due to measurement error. Nonparametric estimates show that there is relatively high (upward) income mobility for poor households, but very little (downward) income mobility for rich households, and that income is much more reliably captured for rich than for poor households.

KEYWORDS: Income mobility, inequality, longitudinal data analysis, measurement error

JEL CODES: J62, D63, C23

[1] Economics Department, University of Stellenbosch, South Africa.
[2] Centre for Studies of African Economics, University of Oxford, United Kingdom.
[3] Economics Department, University of Göttingen, Germany

1. INTRODUCTION

With the increasing availability of panel data in developing countries, studying economic mobility is now feasible and has been done in an increasing number of studies. This is often done in the setting of a so-called "micro growth regression", where income growth is regressed on initial income and some other covariates (e.g. Fields, Cichello, Freije, Menéndez, & Newhouse, 2003a; Woolard & Klasen, 2005). A robust finding of that literature is a rather large negative and highly significant coefficient on the initial income variable, indicating high mobility and suggesting a high speed of "beta-convergence". For example, Fields et al. (2003a) estimate a convergence coefficient of -0.56 (over 5 years) for South Africa, which suggests that one should expect half the income gap between the richest and poorest household to be eliminated every 4.3 years. However, there are also long-standing concerns that micro mobility is over-estimated due to errors in income measures, especially in developing countries where collecting reliable survey data is often difficult. Indeed, Fields (2008b) acknowledges that "a task for the future is to estimate empirically the effect of measurement error on estimates of … micro-mobility".

In this paper we develop an alternative approach that exploits the existence of three waves of panel data to simultaneously estimate the speed of convergence and the extent of measurement error. Our estimates are more efficient than the two-stage least squares (2SLS) estimator that have been used in other studies and can be generalised to allow both the speed of income convergence and the reliability of the income measure to vary with the initial level of income. This approach is applied to a three-wave South African household panel dataset. The results suggest that previous studies have over-estimated the extent of income mobility by a factor of 4 and that about 20% of variation in reported household income is due to measurement error. Nonparametric estimates show that there is relatively high (upward) income mobility for poor households, but very little (downward) income mobility for rich households, and that the income is much more reliably captured for rich than for poor households.

2. INCOME MOBILITY AND MEASUREMENT ERROR

There are various ways to measure the mobility of households in the income distribution[4]. In this paper we will restrict our attention to the concept of weak unconditional beta convergence. Accordingly, the log per capita household income, y_t^*, is characterised as an AR(1) process:

$$y_t^* = \mu + \rho y_{t-1}^* + u_t$$

Current income depends on past income as well as a stochastic income shock, u_t, which is often assumed to be $iid(0, \sigma_u^2)$ (Fields, 2008a, p. 5). The proportional change in income between two periods can then be expressed as

[4] See Jäntti and Jenkins (2015) for a recent overview of this literature.

$$\Delta y_t^* = y_t^* - y_{t-1}^* = \mu + \beta y_{t-1}^* + u_t \quad [1]$$

where $\beta \equiv \rho - 1$ reflects the extent of income mobility in the economy. This specification is deliberately parsimonious[5], since the object of interest is the speed of income convergence rather than the causal mechanism that determines household income. Most studies (e.g. Jarvis and Jenkins (1998), Fields et al. (2003a), Antman and McKenzie (2007b), Fields, Duval-Hernández, Freije, and Puerta (2014)) have focussed on testing $\beta = 0$ against the alternative hypothesis $\beta < 0$. If $\beta = 0$ then there is no tendency for rich and poor household to experience different growth rates, whereas if $\beta < 0$ then poor households tend to grow more rapidly than rich ones. The empirical literature on unconditional convergence in developing countries produces a "virtual consensus" (Fields, 2008a, p. 6) that poorer households grow more quickly than richer ones. Evidence of weak unconditional income convergence has been established, amongst others, for Indonesia, Venezuela, South Africa (Fields et al., 2003a; Woolard & Klasen, 2005), Vietnam (Glewwe, 2012), Argentina, Mexico (Fields et al., 2014) and China (Heng, Shi, & Quheng, 2006; Khor & Pencavel, 2006).

Most empirical studies only test whether there is any evidence of income convergence, but the point estimate of β can also be used to understand the speed at which this convergence occurs. Equation [1] and the assumption that income shocks u_t are i.i.d. imply that the expected change in the relative income gap between any two households (denoted A and B) can be expressed as

$$\frac{\left(y_{A,t-1}^* - y_{B,t-1}^*\right) - E\left(y_{A,t}^* - y_{B,t}^*\right)}{y_{A,t-1}^* - y_{B,t-1}^*} = -\beta$$

In other words, if $\beta < 0$ then $-\beta$ represents the share of any income gap that we expect to be eliminated between periods $t-1$ and t. This convergence parameter can be used to calculate the expected half-life of an income gap (i.e. the expected duration required for half any income gap to be eliminated) as $t \cong \frac{0.7}{\log(\beta+1)}$ periods. For example, Fields et al. (2003a) find convergence coefficients of -0.56 (over 5 years for South Africa), -0.53 (over 4 years for Indonesia), -0.52 (over 1 year for Spain) and -0.64 (over 1 year for Venezuela). These coefficients imply that the expected half-life of the income gap between the richest and poorest households is 4.3 years (South Africa), 3.7 years (Indonesia), 1 year (Spain) and 0.7 years (Venezuela), respectively.

Of course, the estimate of β is only informative about the extent of income mobility if such an estimate is reliable. In practice, income measures obtained from surveys are usually only noisy approximations of true household income, especially in developing countries where collecting reliable survey data can be difficult. Many of the above-mentioned empirical studies mention the issue of measurement error as a

[5] A different strand of the literature focuses on conditional income convergence: the speed at which households converge on their own expected income level, as determined by their observable covariates or household fixed effects.

potential confounding factor that may lead to an over-estimation of the extent of income mobility. In order to formally investigate the effect of measurement error, suppose the available income measure, y_t, suffers from classical measurement error, so that $e_t \equiv y_t - y_t^* \sim iid(0, \sigma_e^2)$. Rewriting equation [1] in terms of the observed but noisy income measure gives:

$$\Delta y_t = \mu + \beta y_{t-1} + u_t + e_t - (\beta + 1)e_{t-1} \quad [2]$$

The econometric problem is that initial income y_{t-1} is negatively correlated with the model error term via the period $t-1$ measurement error term e_{t-1}, which will downwardly bias the OLS estimate of the convergence parameter β. Under the maintained assumptions that both e_t and u_t are i.i.d., the expected value of the OLS slope coefficient obtained from regressing Δy_t on y_{t-1} (which we denote as θ_1) can be expressed as:

$$E(\theta_1) = \frac{\text{Cov}(\Delta y_t, y_{t-1})}{\text{Var}(y_{t-1})} = \frac{\beta \text{Var}(y_{t-1}) - (\beta+1)\text{Cov}(e_{t-1}, y_{t-1})}{\text{Var}(y_{t-1})} = \beta - \frac{-(\beta+1)\sigma_e^2}{\text{Var}(y_{t-1})} = (\beta + 1)\alpha - 1 \quad [3]$$

where $\alpha \equiv \frac{\text{Var}(y_{t-1}^*)}{\text{Var}(y_{t-1})} = \frac{\text{Var}(y_{t-1}^*)}{\text{Var}(y_{t-1}^*) + \sigma_e^2}$ is the share of the total variation in the initial income measure that is due to variation in actual initial income, y_{t-1}^*, rather than measurement error, e_{t-1}. This parameter is restricted to lie within the unit interval and represents the reliability of the observed measure of initial income y_{t-1}. A value of $\alpha = 1$ represents the case of income measured without error, whereas $\alpha = 0$ would indicate that the income measure is all noise and contains no information about actual household income. In the case of no measurement error it follows from equation [3] that $E(\theta_1 | \alpha = 1) = \beta$, so the OLS estimator will provide an unbiased estimate of the extent of income mobility. However, whenever income is measured with some error equation [3] indicates that $E(\theta_1 | \alpha < 1) < \beta$. This will create the appearance of income mobility, even where none exist. Intuitively, if household income is reported with error (and this error is uncorrelated over time), then we would expect households who under-reported their income in period $t-1$ to report a higher income in period t, and vice versa, even if their actual household income was unchanged.

The most common way of addressing measurement error in income mobility studies is to use instrumental variables to obtain a predicted value of lagged income in equation [2] (Fields, Cichello, Freije, Menéndez, & Newhouse, 2003b; Newhouse, 2005; Lee, 2009; Glewwe, 2012; Fields et al., 2014). Glewwe (2012) finds that that at least 15%, and perhaps as much as 42%, of estimated mobility in Vietnam is due to measurement error bias. Turning to previous studies on South African income dynamics, Agüero, Carter, and May (2007) instrument for initial income using household health measures and find that measurement error accounts for between 14% and 60% of all mobility between two successive waves of the South African Kwazulu-Natal Income Dynamics Study (KIDS) panel dataset. Woolard and Klasen (2005) apply a similar approach to the same dataset but find that their results are largely unaffected by measurement error. Lechtenfeld and Zoch (2014) use a three wave panel dataset to

instrument for initial income with previous period income and conclude that conditional income convergence is over-estimated by 39% in the KIDS panel and by 77% in the South African National Income Dynamics Study (NIDS) panel dataset.

The assumption of classical measurement error is convenient to work with, but might not be entirely appropriate when working with self-reported household income. A number of studies have used validation data from different sources, like administrative records, to investigate directly the reliability of self-reported *earnings* data (Bound & Krueger, 1991; Bound, Brown, & Mathiowetz, 2001; Gottschalk & Huynh, 2010; Akee, 2011). Such studies typically find that the measurement error in earnings is serially correlated over time and negatively correlated to true earnings. In this more general case, the expected value of θ_1 can be more accurately expressed as:

$$E(\theta_1) = (\beta + 1)\alpha - 1 + \frac{\text{Cov}(e_t, y^*_{t-1}) + \text{Cov}(e_t, e_{t-1}) - (\beta+1)\text{Cov}(e_{t-1}, y^*_{t-1})}{\text{Var}(y_{t-1})} \quad [4]$$

Antman and McKenzie (2007b) argue, based on the insights from the validation studies, that the last term on the RHS of equation [4] will be positive and so the tendency of classical measurement error to overstate earnings mobility could be partly been offset by the non-classical features of this error. Bound and Krueger (1991) also estimate the ratio of variance of the signal to the true earnings at .82 for men and .92 for women. This could serve as a useful benchmark for our estimate of the corresponding ratio for per capita household income, which we defined as α above.

The finding that the measurement error in total self-reported earnings reveals some non-classical features should make us wary of uncritically applying this classical measurement error assumption to an analysis of per capita household income dynamics. However, validation data is rarely available for developing country surveys, or even for per capita household income in developed economies, so we are limited in what we can do with the available data. Furthermore, if the household size variable that is used to scale total household income also suffers from mean-reverting measurement error – as is typical for categorical variables (Bound et al., 2001, p. 3725) – then this will tend to reduce the correlation between per capita income and its measurement error.

One recently popular approach that attempts to address non-classical measurement error with the available data makes use of pseudo panels (Antman & McKenzie, 2007a; Antman & McKenzie, 2007b; Cuesta, Ñopo, & Pizzolitto, 2011). Successive cross-sectional datasets are used to track the average income for households with heads from the same birth cohort over time. The benefit of this approach is that the within-cohort averaging procedure will remove the effects of income measurement error in sufficiently large cohorts, even where this error is non-classical in nature. Unfortunately, it also averages away all of the highly informative within-cohort variation in household income, which will dramatically reduce the estimator precision and make the estimates highly vulnerable to any deviations in its

identifying assumptions. Fields and Viollaz (2013) apply pseudo-panel estimators to actual panel data, and find that these methods perform poorly in predicting the actual income mobility pattern.

3. REGRESSION COEFFICIENTS IN A THREE WAVE PANEL DATA SET

Although validation data is rarely available in developing countries, it is increasingly common to have three consecutive waves of panel data with which to study household income dynamics. In such cases there is additional information that we can use distinguish between true income mobility and measurement error. The remainder of this paper will develop one approach to do exactly that.

As soon as we have more than two waves of panel data, we are required to make additional assumptions about how income mobility and measurement error changes between waves time. The income dynamics equation [1] can be generalised as

$$\Delta y_t^* = \mu_t + \beta_t y_{t-1}^* + u_t \qquad [5]$$

in which both the intercept and the slope of the first-order autoregressive income process are time-varying. Our proposed approach requires assuming that $-2 < \beta_t = \beta < 0$ and $u_t \sim iid(0, \sigma_u^2)$. The income convergence coefficient is therefore assumed to be constant over the period under consideration. The income intercept term, on the other hand, is completely unrestricted over time. These assumptions imply that income is a trend stationary process[6], i.e. a stationary process after removing the underlying deterministic but potentially non-linear time trend represented by the μ_t parameters. We also maintain the assumption that income measurement error is classical: $e_t \equiv y_t - y_t^* \sim nid(0, \sigma_e^2)$. It is possible to relax some of these assumptions (e.g. allowing β, σ_u^2 or σ_e^2 to change between waves) at the cost of losing over-identifying restrictions, but we will start with the simplest and most restrictive version of the model..

In this case there are at least seven regression coefficients that can be used to inform our estimates of the convergence and income measure reliability parameters, β and α. These coefficients are all easy to estimate and straightforward to interpret. Let $L(y|x) = \theta x$ be the linear projection of y on x. The seven regression coefficients are defined in the first column of Table 1[7] and discussed below.

Table 1: Regression coefficients and population moments

Parameter		Population mean	
		No measurement error	**Classical measurement error**
θ_1	$L(y_2 - y_1 \mid y_1) = \theta_1 y_1$	β	$(\beta+1)\alpha - 1$
θ_2	$L(y_3 - y_2 \mid y_2) = \theta_2 y_2$	β	$(\beta+1)\alpha - 1$
θ_3	$L(y_3 - y_2 \mid y_1) = \theta_3 y_1$	$\beta(\beta+1)$	$\alpha\beta(\beta+1)$
θ_4	$L(y_3 - y_1 \mid y_1) = \theta_4 y_1$	$\beta(\beta+2)$	$\alpha(\beta+1)^2 - 1$

[6] This requires additional assumptions: $-2 < \beta < 0$, and the time invariance of σ_e, σ_u and β.

[7] Assuming that y_t is demeaned allows us to omit the intercept term and is without loss of generality.

θ_5	$L(y_3 - y_2 \mid y_1, y_2) = \theta_5 y_1 + \theta_6 y_2$	0	$\frac{(\beta+1)^2(\alpha-1)\alpha}{\alpha^2(\beta+1)^2-1}$
θ_6	$L(y_3 - y_2 \mid y_1, y_2) = \theta_5 y_1 + \theta_6 y_2$	β	$\frac{1-\alpha(\beta+1)+\alpha^2\beta(\beta+1)^2}{\alpha^2(\beta+1)^2-1}$
θ_7	$L(y_3 - y_2 \mid y_2 - y_1) = \theta_7(y_2 - y_1)$	$\frac{1}{2}\beta$	$-\frac{1-\alpha+\alpha\beta^2}{2(1-\alpha-\alpha\beta)}$

The first coefficient, θ_1, represents the effect of wave 1 income, y_1, on subsequent income growth between waves 1 and 2, Δy_2. We define θ_2 as the same relationship between wave 2 income, y_2, and income growth between waves 2 and 3, Δy_3. These two coefficients represent the conventionally reported estimates of the convergence parameter in a two-wave panel dataset, and either coefficient should provide a consistent estimate of β if this parameter is time invariant and income is measured without error. Comparing these coefficients allows a test of the assumption that $\beta_t = \beta$ (regardless of whether or not income is measured with error).

As discussed in section 2 and elsewhere in this literature, measurement error will tend to bias the estimates of θ_1 and θ_2 away from β and towards -1, since $E(\hat{\theta}_1|\beta,\alpha) = E(\hat{\theta}_2|\beta,\alpha) = (\beta+1)\alpha - 1$. One way to gauge the reliability of $\hat{\theta}_1$ or $\hat{\theta}_2$ as estimates of β is to compare them to regression coefficients $\hat{\theta}_3$ (the regression coefficient obtained from regressing Δy_2 on y_1) and $\hat{\theta}_4$ (obtained from regressing $y_3 - y_1$ on y_1. In the absence of measurement error, a stationary AR(1) process that eliminates in expectation $-\beta$ of income gaps between waves 1 and 2 should eliminate a smaller proportion $-\beta(\beta+1)$ of the initial income gaps between waves 2 and 3. Over two periods the total proportional income convergence should therefore be $-\beta(\beta+2)$. In the absence of measurement error, regression coefficients θ_3 and θ_4 provide estimates of these two quantities: $E(\hat{\theta}_3|\beta,\alpha=1) = \beta(\beta+1)$ and $E(\hat{\theta}_4|\beta,\alpha=1) = \beta(\beta+2)$.

However, if the data is measured with classical error, then $E(\hat{\theta}_3|\beta,\alpha) = \alpha\beta(\beta+1)$ and $E(\hat{\theta}_4|\beta,\alpha) = \alpha(\beta+1)^2 - 1$. Whereas classical measurement error downwardly biases θ_1 and θ_2 it will upwardly bias θ_3. The measurement error of the regressor used to produce θ_3, y_1, is uncorrelated to the measurement error of the regressand, and hence it only suffers the usual attenuation bias. Relative to wave 1 income, classical measurement error therefore leads to an over-estimation of income convergence between waves 1 and 2, and an underestimation of convergence between waves 2 and 3. The effect on total income convergence between waves 1 and 3 is dominated by the former effect, so coefficient θ_4 tends to over-estimate income mobility, but less so than coefficient θ_1. This offers a natural way of using the coefficient estimate[8] of θ_3 or θ_4 to test for the presence of measurement error: check whether there is surprisingly

[8] Note that regardless of the values of β and α, $\theta_1 + \theta_3 = \theta_4$. This means that these two regressions coefficients only add one additional linearly independent population moment that can be used to test hypotheses about or estimate the values of β and α.

little additional income convergence between waves 2 and 3, given the income mobility between waves 1 and 2. In fact, under our maintained assumptions we can obtain consistent estimates of (α, β) from the estimates of θ_1 and θ_4 as

$$\hat{\beta} = \frac{\hat{\theta}_4+1}{\hat{\theta}_1+1} - 1 \text{ and } \hat{\alpha} = \frac{(\hat{\theta}_1+1)^2}{\hat{\theta}_4+1} \quad [6]$$

Additional information is contained in regression coefficients θ_5 and θ_6, the coefficients on y_1 and y_2 when simultaneously included in a regression of Δy_3. If income is measured without error then y_1 should have no effect on Δy_3 after we control for y_2, and the effect of y_2 is simply the convergence parameter β: $E(\hat{\theta}_5|\beta, \alpha = 1) = 0$ and $E(\hat{\theta}_6|\beta, \alpha = 1) = \beta$.As before, the expected values of these regression coefficients are affected by measurement error, and in a way that provides us with useful information about income mobility and the reliability of the income measure.

Let us start by considering the estimated effect of y_2 on Δy_3. If we do not control for y_1 then this effect is represented by θ_2, which is a downwardly biased estimate of actual income convergence. Now, y_1 contains some information about the true value of wave 2 income (as long as $\beta \neq -1$), but no information about the wave 2 measurement error term, which means that controlling for y_1 will exacerbate the attenuation bias in the coefficient on y_2. Coefficient θ_6 will therefore be more downwardly biased than θ_2: $E(\hat{\theta}_6|\beta, \alpha) = \frac{1-\alpha(\beta+1)+\alpha^2\beta(\beta+1)^2}{\alpha^2(\beta+1)^2-1}$. In order to compensate for the downwardly biased θ_6 coefficient estimate, the θ_5 coefficient will be upwardly biased in the presence of classical measurement error: $E(\hat{\theta}_5|\beta, \alpha) = \frac{(\beta+1)^2(\alpha-1)\alpha}{\alpha^2(\beta+1)^2-1}$. Measurement error will therefore make an AR(1) process seem like an AR(2) process in which income growth depends negatively on initial income and positively on previous period income. This provides another test of the validity of our model[9].

Finally, θ_7 is defined as the slope coefficient obtained from regressing Δy_3 on Δy_2. In the absence of measurement error this coefficient estimate has expected value $E(\hat{\theta}_7|\beta, \alpha = 1) = \frac{1}{2}\beta$, which captures the fact that households that experienced more rapid income growth between waves 1 and 2 should expect to experience slower subsequent income growth. If we allow for measurement error then the expected value of this coefficient estimate becomes $E(\hat{\theta}_7|\beta, \alpha) = -\frac{1-\alpha+\alpha\beta^2}{2(1-\alpha-\alpha\beta)}$. If income is measured with error, then the negative correlation between Δy_2 and Δy_3 should be larger than expected.

These regression coefficients can be used to provide more efficient estimates of the model parameters (α, β), while also providing a test of the internal consistency of the identifying assumptions. There are at least two hypotheses of interest. First, is income measured without error? Secondly, do our maintained

[9] Coefficients θ_1, θ_3 and θ_5 are linearly dependent, so only θ_6 provides new information that can be used to test or estimate the model parameters.

assumptions of classical measurement error and a trend stationary first-order autoregressive income process produce an internally consistent set of regression coefficients?

There are various ways to use the regression coefficients to explore the validity of these two hypotheses. We start with the first hypothesis: $\alpha = 1$. The sample estimates of θ_1 and θ_4 can be used to calculate an estimate of α according to equation [6]. The standard error of this estimate can be approximated with the delta method, which would allow us to directly test the hypothesis of interest. However, this test only uses two regression coefficients and is therefore not efficient under the maintained assumptions. An straightforward but informal test[10] of this hypothesis that makes use of all the regression coefficients would be to combine the assumption that $\alpha = 1$ with the sample estimate $\hat{\theta}_1$ (as an estimate of β) to calculate the predicted values of the other six regression coefficients using the formulae in column 3 of Table 1. If income is measured without error, then the predicted and estimated values should only differ due to sampling variation. If the regression coefficient estimates are very different from those implied by $\hat{\theta}_1$ and the assumption of no measurement error, then may suggest that this assumption is invalid.

Of course, we may be concerned that such a difference arises because one of the other maintained assumptions is invalid. For example, if $\beta_2 \neq \beta_3$ then there would have been a different speed of convergence between waves 1 and 2 than between waves 2 and 3. In this case we would not expect $\hat{\theta}_1$ to be a good predictor of $\hat{\theta}_2$, but this does not imply anything about income measurement error. This is the basis of our second hypothesis: that the maintained assumptions of classical measurement error and a trend stationary income process produce an internally consistent set of regression coefficients. A simple and similarly informal way to test this is to use the estimates of β and α obtained from equation [6] to obtain predicted values of the remaining five regression coefficients (all those except $\hat{\theta}_1$ and $\hat{\theta}_4$, which are used to calculate $\hat{\beta}$ and $\hat{\alpha}$) using the formulae in column 4 of Table 1. If the predicted regression coefficients obtained in this way are similar to the estimated regression coefficients, whereas those obtained using only $\hat{\theta}_1$ and the assumption of no measurement error are not, then this provides evidence against the hypothesis of classical measurement while providing no evidence against our maintained hypotheses of classical measurement error and a trend stationary first-order autoregressive income process. In section 6 below, we apply this method to explore the validity of these hypotheses for South African household income data. The two sets of predicted coefficient values are reported in rows 2 and 3 of Table 3.

Another approach, similar in spirit to the one outlined above, would be to use the seven estimated regression coefficients to derive implied estimates of β. If we assume that income is measured without error then the formulae in column 3 of Table 1 can be used to calculate the value of β implied by each of the regression coefficients (with the exception of θ_5). If these estimates all lie within a relatively narrow

[10] A more formal test of this hypothesis is discussed in section 4 below.

range, then this would provide evidence in favour of the assumption of no measurement error. Alternatively, if income is measured with error then we could use each of the regression coefficients and the estimated value of α (from equation [6]) to obtain implied values of β. If these estimates all lie within a narrow range, but those obtained under the no measurement error assumption, then it provides evidence against the assumption of measurement error, but no evidence against our maintained assumptions of classical measurement error and a trend stationary first-order autoregressive income process. These two sets of estimates for South African income dynamics are reported in rows 4 and 5 of Table 3 below.

4. SYSTEM ESTIMATORS AND OVER-IDENTIFICATION TESTS

A system estimator offers a more efficient approach to estimating the model parameters and testing the over-identifying restrictions than the informal approach outlined above. In the presence of classical measurement error there are five linearly independent coefficients that depend on two unknown parameters. The generalised method of moments (GMM) estimator provides a (possibly asymptotically efficient) way of estimating the values of β and α. Given the relationship between our parameters of interest (β, α) and the vector of regression coefficients $\boldsymbol{\theta}$, we can construct a vector of sample moments:

$$g\big(y_{it}, \boldsymbol{\theta}(\beta,\alpha)\big) = \begin{bmatrix} (y_{2i} - y_{1i} - \theta_1 y_{1i}) y_{1i} \\ (y_{3i} - y_{2i} - \theta_2 y_{1i}) y_{1i} \\ (y_{3i} - y_{2i} - \theta_4 y_{2i}) y_{2i} \\ (y_{3i} - y_{2i} - \theta_5 y_{1i} - \theta_6 y_{2i}) y_{2i} \\ \big(y_{3i} - y_{2i} - \theta_7 (y_{2i} - y_{1i})\big)(y_{2i} - y_{1i}) \end{bmatrix}$$

The identifying assumption $E\big[g\big(y_{it}, \boldsymbol{\theta}(\beta_0, \alpha_0)\big)\big] = 0$ follows directly from the assumptions that both u_t and e_t are iid processes. The GMM estimator can then be expressed as:

$$(\hat{\beta}, \hat{\alpha}) = \arg\min_{\beta,\alpha} \left(\frac{1}{N}\sum\nolimits_{it} g\big(y_{it}, \boldsymbol{\theta}(\beta,\alpha)\big)\right)' \widehat{W} \left(\frac{1}{N}\sum\nolimits_{it} g\big(y_{it}, \boldsymbol{\theta}(\beta,\alpha)\big)\right)$$

where $\widehat{W}$ is the weighting matrix. Assuming that this is an optimal weighting matrix and that our over-identification restrictions are valid implies that

$$J = \left(\frac{1}{\sqrt{N}}\sum\nolimits_{it} g\big(y_{it}, \boldsymbol{\theta}(\beta,\alpha)\big)\right)' \widehat{W} \left(\frac{1}{\sqrt{N}}\sum\nolimits_{it} g\big(y_{it}, \boldsymbol{\theta}(\beta,\alpha)\big)\right)$$

has a chi-squared limiting distribution with three degrees of freedom. This provides a straightforward test of the validity of our identifying assumptions. We can also estimate the value of β under the assumption that $\alpha = 1$ to test whether this is consistent with the sample regression coefficients. Even in the case of

no measurement error, this approach should provide more efficient estimates of the convergence parameter β than estimates of regression coefficients θ_1 and θ_2.

5. NONPARAMETRIC EXTENSION: INITIAL-INCOME DEPENDENT CONVERGENCE AND MEASUREMENT ERROR VARIANCE

There are numerous ways in which income can be incorrectly measured that do not conform to the assumptions of classical measurement error. One such a deviation is when the reliability of the income measure varies by the level of initial income: $\alpha(y^*_{t-1})$. In this case respondents still provide noisy but unbiased estimates of their household income, but the variance of the measurement error may be larger or smaller for households with higher incomes. Once we relax the assumption that α is constant, it is straightforward to also allow the convergence parameter to vary by initial income: $\beta(y^*_{t-1})$. In this case we can rewrite equation [2] as

$$\Delta y_t = \mu + \beta(y^*_{t-1})y_{t-1} + u_t + \sigma(y^*_t)e_t - (\beta(y^*_{t-1}) + 1)\sigma(y^*_{t-1})e_{t-1} \quad [7]$$

where $e_t - 1$ and e_t have now been standardised to have a standard deviation of 1 and $\sigma(y^*_{t-1})$ reflects the effect of initial income on the standard deviation of the income measurement error. It follows[11] that

$$\frac{\partial E(\Delta y_2|y_1)}{\partial y_1} \cong (\beta(y_1) + 1)\alpha(y_1) - 1 \equiv \theta_1(y_1)$$

$$\frac{\partial E(y_3 - y_1|y_1)}{\partial y_1} = \alpha(y_1)(\beta(y_1) + 1)^2 - 1 \equiv \theta_4(y_1)$$

Estimates of these slope parameters can be obtained from local polynomial regressions, and used to estimate the model parameters $\big(\alpha(y_1), \beta(y_1)\big)$ using a generalisation of equation [3]:

$$\left(\frac{\hat{\theta}_3(y_1)+1}{\hat{\theta}_1(y_1)+1} - 1, \frac{(\hat{\theta}_1(y_1)+1)^2}{\hat{\theta}_3(y_1)+1}\right) \quad [8]$$

6. INCOME CONVERGENCE IN SOUTH AFRICA BETWEEN 2008 AND 2012

This approach is now applied to the three waves of the South African National Income Dynamics Study (NIDS) panel dataset[12]. The three waves were collected in 2008, 2010 and 2012, respectively. NIDS is a large, nationally representative dataset. In order to circumvent unbalanced panel issues, we only use the

[11] This follows from the maintained assumptions that $E(u_3|y_1) = E(u_2|y_1) = E(e_3|y_1) = E(e_2|y_1) = 0$, the implication that $E(e_1\sigma_e(y^*_1)|y_1) = \big(1 - E(\alpha(y^*_1)|y_1)\big)y_1$ and the approximations that $E(\alpha(y^*_1)|y_1) \cong \alpha(y_1)$, $E(\beta(y^*_1)|y_1) \cong \beta(y_1)$ and $\beta(y_1) \cong \beta(y_2)$.

[12] See Finn and Leibbrandt (2013) for a detailed survey description.

households that were captured and had the same household head in all three waves. Balanced panel weights are used to adjust the sample for attrition across all waves, as explained in Finn and Leibbrandt (2013). This provides us with a sample of 2770 households. Our chosen measure of income, y, is real per capita household income from all sources and with imputations for any missing values.

Table 2 reports the estimates for the seven regression coefficients of interest. The regression coefficients in columns 1 and 2 (that correspond to θ_1 and θ_2) are both slightly above -0.25, suggesting that approximately 25% income gaps are eliminated in the two-year periods between surveys. At this rate of convergence we could expect half of the income gap between the richest and poorest South African household to be eliminated every 4.8 years. This seems like a surprisingly high degree of income mobility, but is consistent with the estimate obtained by Fields et al. (2003a) for South Africa using the KIDS panel, which implies that 56% of income gaps should be eliminated over 5 years. It is also in line with the convergence coefficient that (Fields et al. (2003a)) estimate for Indonesia (53% over 4 years), Spain (52% over 1 year) and Venezuela (64% over 1 year). Of course, all of these estimates are vulnerable to the presence of measurement error.

Table 2: Regression coefficients for South African income regressions

	(1)	(2)	(3)	(4)	(5)	(6)
	Δy_2	Δy_3	Δy_3	$y_3 - y_1$	Δy_3	Δy_3
y_1	-0.249***		-0.0427**	-0.292***	0.329***	
	(0.0251)		(0.0196)	(0.0254)	(0.0295)	
y_2		-0.243***			-0.495***	
		(0.0227)			(0.0267)	
Δy_2						-0.409***
						(0.0280)
Constant	1.825***	1.911***	0.471***	2.296***	1.375***	0.189***
	(0.174)	(0.156)	(0.139)	(0.176)	(0.134)	(0.0211)
Observations	2,770	2,770	2,770	2,770	2,770	2,770
R-squared	0.129	0.141	0.004	0.170	0.252	0.194

Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1

In order to investigate the internal consistency of these estimates Table 3 compares these coefficients and their standard errors (in row 1) along with (in the second row) the expected values of these coefficients if the estimated regression coefficient θ_1 is equal to the convergence parameter β and income is measured without error. Apart from θ_2, none of the regression coefficients is near its predicted value. The effect of wave 1 income on income growth between waves 2 and 3 (represented by θ_3), and total income growth between waves 1 and 3 (represented by θ_4) are both much smaller than we would have expected given the

rapid income growth that occurred between waves 1 and 2, and waves 2 and 3. The coefficient estimates obtained from regressing Δy_3 on y_1 and y_2 (θ_5 and θ_6) are also very different than what we would expect in the absence of measurement error. Instead of values close to zero and θ_1, we observe estimates that are significantly positive and significantly more negative than θ_1. Finally, θ_7 reveals a stronger negative correlation between Δy_3 and Δy_2 than we can be explain without measurement error.

Another way of testing for the existence of measurement error is to calculate the values of β implied by each regression coefficient estimate (apart from θ_5, which does not depend on β in the no measurement error case). The implied parameter values (reported in the fourth row of Table 3) vary from very small (-0.045) to very large (-0.495), and seem unlikely to be the result of sampling variation in the data. Our inspection method of testing whether income is measured without error therefore provides evidence against this hypothesis, although we cannot attach a p-value to this test.

Equation 3 showed how we can obtain point estimates of β and α using the estimates of θ_1 and θ_4. This produces estimates of $\hat{\beta} = -0.057$ and $\hat{\alpha} = 0.80$. These estimates are used to predict the values of the other five regression coefficients (column three in Table 3). We observe that these predictions are very close to the estimated values in column 1. Column five also reports the values of β that are implied by each of the regression coefficients if $\alpha = 0.8$. These values are distributed in a narrow range between -0.061 and -0.05. Allowing for classical measurement error makes it is possible to provide an internally consistent explanation of the estimated regression coefficients, while it is impossible to do so while maintaining the assumption that income is measured without error.

Table 3: Regression coefficients and implied parameter values

θ_k	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7
Estimated values of θ_k	-0.249*** (0.0251)	-0.243*** (0.0227)	-0.0427** (0.0196)	-0.292*** (0.0254)	0.329*** (0.0295)	-0.495*** (0.0267)	-0.409*** (0.0280)
Predicted values of θ_k if $\beta = -0.249$ and $\alpha = 1$	-0.249	-0.249	-0.187	-0.436	0.000	-0.249	-0.125
Predicted values of $\theta_k$if $\beta = -0.057$ and $\alpha = 0.8$	-0.249	-0.249	-0.043	-0.292	0.330	-0.497	-0.414
Value of β implied by θ_k (if $\alpha = 1$)	-0.249	-0.243	-0.045	-0.159	NA	-0.495	-0.818
Value of β implied by θ_k (if $\alpha = 0.8$)	-0.057	-0.050	-0.057	-0.057	-0.058	-0.060	-0.061

Next, we proceed to estimate the model parameters using the system GMM estimator. The results are shown in Table 4. If we place no restriction on the value of α, then β is estimated to be -0.0590. This is similar to the point estimate obtained using equation [3] and implies that only about 6% of income gaps are expected to be eliminated during the two years between survey waves. This is much lower than the estimates obtained by either regressing Δy_2 on y_1 or Δy_3 on y_2. In fact, this estimate suggests that the conventional approach over-estimates South African income mobility by a factor of between 4 and 5. The

implied expected half-life of any income gap is now approximately 27 years, not 5, which means that South Africa has considerably less economic mobility than previous studies may have led us to believe. The GMM estimate of α indicates that this discrepancy arises because only 80% of the variation in log household income is due to variation in actual incomes, whereas the remaining 20% is due to measurement error.

Apart from allowing us to simultaneously estimate the income convergence and data reliability parameters, the GMM estimator has the added advantage of providing estimates that are highly efficient, as can be observed by the small standard errors. Furthermore, it also allows us to formally test the validity of the over-identifying restrictions. The J-test indicates that the GMM estimates can explain all five linearly independent regression coefficients in a way that is internally consistent.

Column 2 of Table 4 estimates the convergence parameter under the assumption of no measurement error (by restricting $\alpha = 1$). The point estimate of β obtained from using all the regression coefficients is much smaller than suggested by the estimates of either θ_1 or θ_2. However, the associated J-test also strongly rejects that validity of the over-identifying restrictions, which confirms that the assumptions of no measurement error is inconsistent with the observed covariance pattern in the data.

Table 4: GMM estimates for South African income dynamics

	(1)	(2)
β	-0.0590***	-0.0886***
	-0.0174	-0.00455
α	0.801***	1
	-0.0195	.
Observations	2,770	2,770
J-test statistic	0.249	73.2
p-value	0.969	0

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Finally, we use local linear regressions to estimate the nonparametric generalisations of regression coefficients θ_1 and θ_4. These estimates are plotted (against demeaned y_1 on the x-axis) in Figure 1, and used to calculate nonparametric estimates of $\beta(y_1)$ and $\alpha(y_1)$ according to equation [8]. The resulting estimates of the β and α functions are graphed in Figure 2. The income measure reliability parameter varies between 0.6 for low initial income values and 0.95 for high initial incomes. Instead of 20% of the variation in all household incomes being due to measurement error, this share is as high as 40% for poor households and as low as 5% for rich ones. The income convergence parameter varies between -0.14 for poor households and -0.03 for rich households, which reveals that income mobility also depends on initial income. Whereas our parametric estimates indicated that all household could expect to move 6%

towards the mean income level, the nonparametric estimate reveals that poor households can expect to experience more upward mobility, whereas rich household should experience relatively little downward mobility on average.

Figure 1: Local linear regression estimates of $\theta_1(y_1)$ and $\theta_4(y_1)$

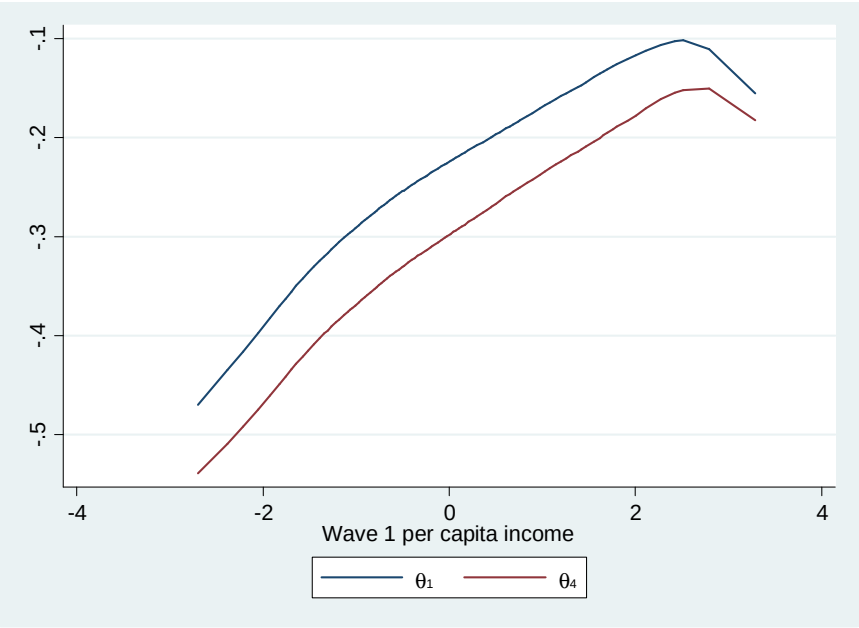
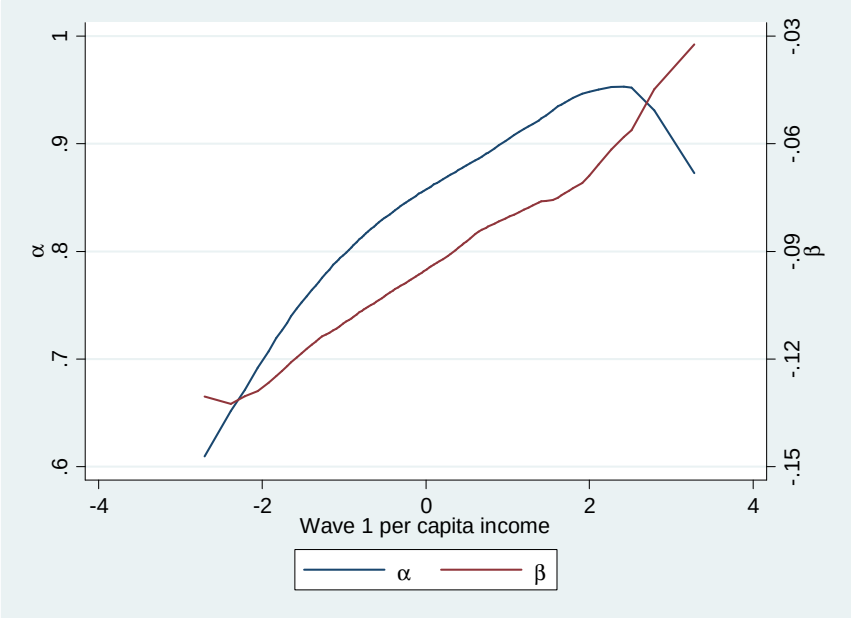


Figure 2: Nonparametric estimates of $\beta(y_1)$ and $\alpha(y_1)$



7. CONCLUSION

This study proposed a new approach that uses three-wave panel data to estimate income mobility when incomes are measured with error. This approach is applied to a three-wave South African panel dataset and finds that the conventional method over-estimates the extent of income mobility by a factor of between 4 and 5. This occurs because about 20% of variation in reported household income is due to measurement error. Nonparametric estimates show that there is relatively high (upward) income mobility for poor households, but very little (downward) income mobility for rich households, and that the income is much more reliably captured for rich than for poor households.

8. REFERENCES

Agüero, J., Carter, M. R., & May, J. (2007). Poverty and inequality in the first decade of South Africa's democracy: What can be learnt from panel data from KwaZulu-Natal? *Journal of African Economies, 16*(5), 782-812.

Akee, R. (2011). Errors in self-reported earnings: The role of previous earnings volatility and individual characteristics. *Journal of Development Economics, 96*(2), 409-421.

Antman, F., & McKenzie, D. (2007a). Poverty traps and nonlinear income dynamics with measurement error and individual heterogeneity. *The journal of development studies, 43*(6), 1057-1083.

Antman, F., & McKenzie, D. J. (2007b). Earnings Mobility and Measurement Error: A Pseudo-Panel Approach. *Economic Development and Cultural Change, 56*(1), 125-161.

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement Error in Survey Data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics, Vol 5*. New York: Elsevier Science B.V.

Bound, J., & Krueger, A. B. (1991). The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right? *Journal of Labor Economics*, 1-24.

Cuesta, J., Ñopo, H., & Pizzolitto, G. (2011). Using Pseudo-Panels To Measure Income Mobility In Latin America. *Review of Income and Wealth, 57*(2), 224-246.

Fields, G. S. (2008a). A brief review of the literature on earnings mobility in developing countries *ILR Working Paper 2-20-2008*. Ithaca: Cornell University.

Fields, G. S. (2008b). Income mobility. In L. Blume & S. N. Durlauf (Eds.), *The new Palgrave dictionary of economics*. New York, NY: Palgrave Macmillan.

Fields, G. S., Cichello, P. L., Freije, S., Menéndez, M., & Newhouse, D. (2003a). For richer or for poorer? Evidence from Indonesia, South Africa, Spain, and Venezuela. *The Journal of Economic Inequality, 1*(1), 67-99.

Fields, G. S., Cichello, P. L., Freije, S., Menéndez, M., & Newhouse, D. (2003b). Household income dynamics: a four-country story. *The journal of development studies, 40*(2), 30-54.

Fields, G. S., Duval-Hernández, R., Freije, S., & Puerta, M. L. S. (2014). Earnings mobility, inequality, and economic growth in Argentina, Mexico, and Venezuela. *The Journal of Economic Inequality*, 1-26.

Fields, G. S., & Viollaz, M. (2013). Can the Limitations of Panel Datasets be Overcome by Using Pseudo-Panels to Estimate Income Mobility? *Universidad Cornell-CEDLAS*.

Finn, A., & Leibbrandt, M. (2013). Mobility and Inequality in the First Three Waves of NIDS: Southern Africa Labour and Development Research Unit, University of Cape Town.

Glewwe, P. (2012). How much of observed economic mobility is measurement error? IV methods to reduce measurement error bias, with an application to Vietnam. *The World Bank Economic Review, 26*(2), 236-264.

Gottschalk, P., & Huynh, M. (2010). Are earnings inequality and mobility overstated? The impact of nonclassical measurement error. *The Review of Economics and Statistics, 92*(2), 302-315.

Heng, Y., Shi, L., & Quheng, D. (2006). Income Mobility in Urban China. *Economic Research Journal, 10*, 002.

Jäntti, M., & Jenkins, S. P. (2015). Chapter 10 - Income Mobility. In B. A. Anthony & B. François (Eds.), *Handbook of Income Distribution* (Vol. Volume 2, pp. 807-935): Elsevier.

Jarvis, S., & Jenkins, S. (1998). How much income mobility is there in Britain? *The Economic Journal, 108*(447), 428-443.

Khor, N., & Pencavel, J. (2006). Income mobility of individuals in China and the United States. *Economics of Transition, 14*(3), 417-458.

Lechtenfeld, T., & Zoch, A. (2014). Income Convergence in South Africa: Fact or Measurement Error?

Lee, N. (2009). Measurement error and its impact on estimates of income and consumption dynamics. *Available at SSRN 1299330*.

Newhouse, D. (2005). The persistence of income shocks: Evidence from rural Indonesia. *Review of development Economics, 9*(3), 415-433.

Woolard, I., & Klasen, S. (2005). Determinants of income mobility and household poverty dynamics in South Africa. *Journal of Development Studies, 41*(5), 865-897.