The Consequences of Non-Classical Measurement Error for Distributional Analysis

October 29, 2004

Abstract

This paper analyzes the consequences of non-classical measurement error for distributional analysis. We show that for a popular set of distributions negative correlation between the measurement error (u) and the true value (y^*) may reduce the bias in the estimated distribution at every value of y^*. For other distributions the impact of non-classical measurement differs throughout the support of the distribution. We illustrate the practical importance of these results using models of unemployment duration and income.

Keywords: Distribution functions, Non-classical measurement error

1 Introduction

Statistical analysis involves examining the outcomes of random experiments in order to make inferences about the distribution function underlying the true data generating process. Measurement error may lead researchers to draw incorrect inferences. The impact of specification error on means has been studied extensively (e.g. Fuller (1987), Carroll et al (1994) and Bound et al (2001)). However, less is known about the consequences of specification error for other aspects of the distribution function. Horowitz and Manski (1995) discuss circumstances in which we can use mismeasured data to bound the distribution of the true variable. They consider situations in which the variable of interest is in general well-measured though some observations may be subject to potentially large errors. In contrast the typical textbook model of measurement error reflects a situation of widespread mismeasurement (the error distribution has no mass point at zero). Chesher (1991) uses a small variance approximation to study the impact of this form of measurement error on distribution functions and argues that the sign of the bias arising from the mismeasured data can be determined by the curvature of the true underlying distribution. In particular in regions where the true underlying distribution is convex we overestimate the distribution and in regions where it is concave we underestimate. However, Chesher only considered classical measurement error, where the error term is distributed independently of the true value. In this paper we provide a simple geometric exposition of the consequences of non-classical measurement error on distribution functions. In particular we show that for a popular set of distributions, allowing for correlation between the error and true value may offset the bias that arises with classical measurement error throughout the distribution. We illustrate our results by examining the impact of measurement error on models of unemployment durations and income.

2 The Consequences of Measurement error for Distribution Functions

2.1 Theoretical Results

Let y^* be a random variable, whose cumulative distribution function is given by $F_{y^*}(y)$ with support $\left[\underline{y}, \overline{y}\right]$. However, for some reason y^* cannot be measured accurately. Instead we observe y which is defined as $y = y^* + u$, where u is measurement error with support $[\underline{u}, \overline{u}]$. The assumption that the error term is additive is less restrictive than it appears. In the case of mismeasured incomes it is often assumed that the error term enters multiplicatively (see e.g. Chesher and Schluter (2001)). To apply the results established in our paper to these models we simply consider a log transformation of the model. In this case we view y as the log of observed income and y^* as the log of true income, so that the observed level of income, I, may be written as $I = I^*V$, where $V \equiv \exp(u)$. Furthermore, since $F_I(\exp(y^0)) = F(y^0))$, our results identify the ranges of I for which which a multiplicative error process would cause over or underestimation of the true underlying distribution.

Measurement error can take the misspecified cdf outside the support of the true cdf. We denote the extended distribution function of a random variable with cdf $F_y(y)$, by $\widetilde{F}_y(y)$. We make no specific assumptions about the relationship between y^* and u. Observed data provide information on $\widetilde{F}_y(y)$. An expression for the difference between the mismeasured and true distributions $\Delta_F(y^0) = \widetilde{F}_y(y^0) - \widetilde{F}_{y^*}(y^0)$ is given in Theorem 1.

Theorem 1

The bias ($\Delta_F(y^0)$) when y^* is measured with error is given by

$$\int_{\underline{u}}^{\overline{u}} \int_{y^0}^{y^0-u} \widetilde{f}_{y^*,u}(y,u)\, dydu. \tag{1}$$

Proof. *Since $y = y^* + u$, we have $\widetilde{f}_y(y^0) = \int_{\underline{u}}^{\overline{u}} \widetilde{f}_{y^*,u}(y^0 - u, u)\, du$, such that*
$\widetilde{F}_y(y^0) = \int_{\underline{u}}^{\overline{u}} \int_{\underline{y}}^{y^0} \widetilde{f}_{y^*,u}(y-u,u)\, dydu$.
At the same time,
$\widetilde{F}_{y^*}(y^0) = \int_{\underline{u}}^{\overline{u}} \int_{\underline{y}}^{y^0} \widetilde{f}_{y^*,u}(y,u)\, dydu$.
Differencing the last two expressions results in equation (1) of the theorem. ■

Equation (1) can be given a simple graphical interpretation. This is shown in Figure 1.

Figure 1 about here.

In this Figure u is on the x-axis and y^* is on the y-axis. The joint density $\widetilde{f}_{y^*,u}(y,u)$ is represented by the elliptical contours. Equation (1) gives us the probability between the line $y^* = y^0 - u$ and $y^* = y^0$. This involves subtracting the probability mass in S_2 from the probability mass in S_1. Since the correlation between y^* and u affects the shape of the contours, this graph provides a geometric illustration of the potential importance of this correlation in determining the size of the bias. To establish this relationship formally, let $u^1 = E(u) + \gamma$, $u^2 = E(u) - \gamma$, $y^1 = E(y) + \delta$, $y^2 = E(y) - \delta$ and Θ be the set of all distributions $\widetilde{f}_{y^*,u}(y,u)$. Consider the following definition.

Definition 1

$$T\left(\widetilde{f}_{y^*,u}(y,u), \varepsilon, \gamma, \delta\right) : \mathbf{\Theta} \times (\mathbb{R}^+)^3 \rightarrow \mathbf{\Theta} :$$
$$\left(\widetilde{f}_{y^*,u}(y,u), \varepsilon, \gamma, \delta\right) \rightarrow g_{y^*,u}(y,u)$$

is a mean preserving covariance increasing transformation of a density function $f_{y^*,u}(y,u)$ if and only if

for all $u \neq u^1, u^2$ and $y \neq y^1, y^2$: $g_{y^*,u}(y,u) = \widetilde{f}_{y^*,u}(y,u)$,
$g_{y^*,u}(y^1,u^1) = \widetilde{f}_{y^*,u}(y^1,u^1) + \varepsilon$,
$g_{y^*,u}(y^1,u^2) = \widetilde{f}_{y^*,u}(y^1,u^2) - \varepsilon$,
$g_{y^*,u}(y^2,u^1) = \widetilde{f}_{y^*,u}(y^2,u^1) - \varepsilon$,
$g_{y^*,u}(y^2,u^2) = \widetilde{f}_{y^*,u}(y^2,u^2) + \varepsilon$.

Such a transformation increases the conditional covariance between y^* and u, but does not affect either the mean or variance of the marginal distributions of u or y^*. We now establish the following Theorem.

Theorem 2 *If $g_{y^*,u}(y,u)$ can be obtained out of $\widetilde{f}_{y^*,u}(y,u)$ after a sequence of mean preserving covariance increasing transformations, then*
(a) If $E(u) \leq 0$ and $y^0 \geq E(y)$: $\widetilde{G}_y(y^0) - \widetilde{G}_{y^*}(y^0) \leq \widetilde{F}_y(y^0) - \widetilde{F}_{y^*}(y^0)$
(b) If $E(u) \geq 0$ and $y^0 \leq E(y)$: $\widetilde{G}_y(y^0) - \widetilde{G}_{y^*}(y^0) \geq \widetilde{F}_y(y^0) - \widetilde{F}_{y^*}(y^0)$

Proof. We only prove the result for case (a). The proof of case (b) is similar.
Define
$S_1 = \left\{(u,y) \in [\underline{u},\overline{u}] \times \left[\underline{y},\overline{y}\right] \mid u \leq 0, y^0 \leq y \leq y^0 - u\right\}$
$S_2 = \left\{(u,y) \in [\underline{u},\overline{u}] \times \left[\underline{y},\overline{y}\right] \mid u \geq 0, y^0 - u \leq y \leq y^0\right\}$.

These regions are illustrated in Figure 1.

A transformation $T\left(\widetilde{f}_{y^*,u,a^*}(y,u),\varepsilon,\gamma,\delta\right)$ will only affect equation (1) if it changes the probability mass in S_1 or S_2. Under the assumption that $E(u) \leq 0$ two such cases exist. First, it is possible that γ and δ are such that the point with coordinates $(u^1, y^1) \in S_2$. The probability mass in S_2 increases, such that $\Delta_G(y^0)$ is smaller than $\Delta_F(y^0)$. Second, γ and δ are such that the point with coordinates $(u^2, y^1) \in S_1$. In this case the probability mass in S_1 decreases, again resulting in $\Delta_G(y^0)$ being smaller than $\Delta_F(y^0)$. For all other values of γ and δ, the probability mass in S_1 and S_2 will not be affected, or will be affected in the same way, such that $\Delta_G(y^0)$ will equal $\Delta_F(y^0)$ for these values. ■

Since $\widetilde{G}_{y^*}(y^0) = \widetilde{F}_{y^*}(y^0)$ Theorem 2 directly relates mean preserving covariance increasing transformations to the size of the bias in the mismeasured distribution function. If the transformation increases the covariance then the bias becomes less positive (or more negative) provided y^0 is greater than the mean of y and $E(u) \leq 0$. If y^0 is less than the mean and $E(u) \geq 0$ the opposite occurs. It is easy to see that the results are reversed when the transformation decreases the covariance. In many applications it may be reasonable to assume that $E(u) = 0$. In this case Theorem 2 allows us to establish the impact of non-classical measurement error over the entire range of y.

2.2 Examples

Example 1. Consider a simple case where y^* and u are independent and u is symmetric around zero. If $f_{y^*}(y^0)$ is normal then Chesher (1991) shows that we overestimate (underestimate) $F_{y^*}(y^0)$ at points below (above) $E(y)$, while the bias is zero at $E(y)$.[1]

[1] Chesher's results are based on approximations. O'Neill et al (2004) establish exact results for cdfs under which the sign the bias can be established. These depend on modified curvature conditions of the true underlying distribution. For the normal distribution with symmetric measurement error Chesher's approximate conditions are equivalent to the exact curvature conditions.

Theorem 2 shows that, in this case, introducing negative correlation between y^* and u (which is what we tend to see in earnings data (Bound et al. (1994)) may in fact reduce the extent of overestimation (underestimation) throughout the distribution. Indeed this is true for the wider class of distributions with cdfs that are convex below the mean and concave above the mean. These include distributions such as the t-distribution and the logistic distribution .

Example 2. Consider the power distribution $F(y, \theta) = y^\theta$ for $0 \leq y \leq 1, \theta > 0$. This distribution is convex for all y between zero and 1 provided $\theta > 1$. Chesher's (1991) results imply that classical measurement error will lead us to overestimate the distribution at each of the points in the original support. Theorem 2 shows that the consequences of correlated measurement error differs depending on the value of y. For values of y below the mean, negatively correlated measurement error may result in the bias becoming smaller, while for values of y above the mean the bias must increase.

Example 3a. Consider the exponential distribution $F(y, \lambda) = 1 - e^{-\lambda y}$, $y > 0$, $\lambda > 0$. The exponential distribution arises naturally in many statistical problems associated with waiting times. For instance, if the occurrence of an event is governed by a Poisson process then it can be shown that the sequence of inter-arrival times are independent identically distributed exponential random variables. In applied research the exponential distribution is widely used as a starting point for the analysis of unemployment duration and strike duration data (Kiefer (1988)). It is easy to show that the exponential distribution function is concave for all y. In this case classical measurement error will lead us to underestimate the distribution throughout the original support.[2] However, there is some evidence that longer spells of unemployment are more likely to be subject to underreporting, implying a negative correlation between the true level of unemployment duration and measurement error (Torelli and Trivellato (1989)). As in Example 2 the consequences of correlated measurement error differs depending on the value of y, though in this case the effect goes in the opposite direction. For values of y below the mean, negatively correlated measurement error will accentuate the bias (the bias becomes more negative); however for values of y above the mean the bias may fall in absolute value (become less negative), though if the correlation is sufficiently negative the mismeasured distribution could move above the true distribution, thus inducing a positive bias. Thus measurement error in unemployment durations, that is negatively correlated with the truth, may be preferable to independent measurement error if our focus is on long unemployment spells but will compound the problem of independent measurement error when considering shorter spells.[3]

Example 3b. The exponential distribution is restrictive in that the implied hazard rate is constant. However the conclusions from Example 3a generalise to less restrictive cases with non-constant hazards. The Weibull distribution, given by

[2] For a detailed discussion of classical measurement error in duration reponse data see Chesher et al (2002).

[3] Since we are basing the sign of the bias with independent measurement error on small-variance approximations this preference ranking over types of measurement error need not apply to very short or very long durations.

$F(y,\alpha,\gamma) = 1 - e^{-\gamma y^{\alpha}}$, $y > 0, \alpha > 0, \gamma > 0$, is a two parameter generalisation of the exponential distribution, which allows for a non-constant hazard. It is easy to show that the conclusions reached in Example 3a regarding non-classical measurement error remain valid provided $\alpha < 1$. However, $\alpha < 1$ is equivalent to specifying negative duration dependence, which is typical in many studies of unemployment.[4]

Example 4. When modelling the consequences of measurement error in non-negative variables, such as unemployment duration, one may prefer to adopt a multiplicative form for the error process. As noted earlier this is easily incorporated within our specification. To see this reconsider the exponential distribution. Assume that $F(t,\lambda) = 1 - e^{-\lambda t}$, $t > 0, \lambda > 0$ and denote the measurement error by V. In this case we may wish to model observed duration as $S = TV$. To apply our framework to this model we simply take a log transformation of the multiplicative model, so that $\ln(S) = \ln(T) + \ln(V)$.[5] This model is now in the format specified in our earlier theorems. To establish the impact of measurement error in this model we need to be able to describe the distribution of $\ln(T)$. However, if T is exponentially distributed then $\ln(T)$ has a Type 1 extreme value distribution with density given by $g(y) = \lambda \exp(y) \exp(-\lambda \exp(y))$. The mean of this random variable is given by $E(\ln(T)) = -\ln(\lambda) - \gamma$, where γ is Euler's constant (approximately .5772). Furthermore it is easy to show that this distribution is convex provided $\ln(T) < \ln(1/\lambda) = -\ln(\lambda)$ and concave otherwise. Using Chesher's results we conclude that with classical measurement error we overestimate provided $\ln(T) < \ln(1/\lambda)$ and underestimate otherwise. In terms of the actual unemployment durations, T, this implies that the distribution of S overestimates the distribution of T provided $T < 1/\lambda \equiv E(T)$ and underestimates provided $T > E(T)$.

From Theorem 2 we can deduce that allowing $\ln(T)$ and $\ln(V)$ to be negatively correlated will cause the bias to become less positive provided $\ln(T) < E(\ln(T)) = -\ln(\lambda) - \gamma$, and causes the bias to become less negative when $\ln(T) > -\ln(\lambda) - \gamma$. Combining this with our earlier analysis we see that for short unemployment durations, specifically those such that $\ln(T) < -\ln(\lambda) - \gamma$, negative correlation may help offset the original *positive* bias resulting from classical measurement error. For long unemployment durations, such that $\ln(T) > -\ln(\lambda)$, non-classical measurement error may help offset the *negative* bias introduced by uncorrelated measurement error. However, since $E(\ln(T)) < \ln(E(T))$, there is now also an intermediate range of log durations, from $[-\ln(\lambda) - \gamma, -\ln(\lambda)]$, for which the original tendency to overestimate with classical measurement error is compounded by correlated measurement error. In terms of the raw durations T, the range for which non-classical error compounds the original biases is given by $[E(T)/1.78, E(T)]$.

This shows how the framework we have introduced can be easily extended so as to yield practical insights into the consequences of non-classical measurement error with alternative error structures.[6] A similar analysis can also be conducted in cases

[4] For a recent overview of the literature on duration dependence in unemployment see Serneels (2002).

[5] See Kiefer (1988), Section IV, for a more detailed discussion of the potential use of log-linear models for duration analyses.

[6] In this example we have assumed that $E(\ln(V)) = 0$. This need not imply that $E(V) = 1$ which

where the original distribution of durations is Weibull, since the natural logarithm of a random variable with a Weibull distribution also has a Type 1 extreme value distribution.

2.3 Consequences for Estimated Poverty Rates

To explore the magnitude of non-classical measurement error in practice we consider a calibrated model of the distribution of income for white couples in the U.S in the early 1990's. For simplicity we assume that income is distributed as log normal[7]. Letting y^* denote the log of income we assume that $y^* \sim N(10.72,\ .24)$.[8] We assume that $u \sim N(0, \sigma_u^2)$ where σ_u^2 is chosen so that $\frac{\sigma_u^2}{\sigma_{y^*}^2} = .33$. This corresponds to a reliability ratio of .75 when measurement error is classical. This is within the range of estimates presented in recent studies (Zimmerman (1992), Angrist and Krueger (1999)). We consider two cases. First we assume that y^* and u are independent. We then compare this to the case where y^* and u have a correlation equal to -.3.[9,10]

Figure 2 presents the true distribution and both the misspecified distributions (with and without correlation)[11]. The findings with independent measurement error are consistent with Chesher (1991); we overestimate in the region where the true distribution is convex, underestimate where it is concave and the bias is zero at the mean. Given the calibration of our model the size of the bias arising from independent measurement error is relatively small. As predicted by Theorem 2, introducing negative correlation between the error and the true value causes the bias to become less positive for values of y below the mean and less negative for values above the mean. Indeed, for our calibrated model the distribution with correlated measurement is virtually identical to the true model. To summarise the impact of correlated error terms we consider measures of the poverty rate based on mismeasured data, both with and without correlation between the error and the true income. We choose 1/2 median income as the measure of poverty. Under our assumptions the poverty line is constant across all 3 distributions. The estimated poverty rate is 8% for both the true and correlated models and 11% for the independent specification. It is worth emphasising that if the correlation becomes more negative (i.e less than -.3) the correlated income distribution falls *below* the true distribution for $y's$ below the mean and rises *above* it for $y's$ above the mean. The corresponding poverty rate with correlated errors would then underestimate the true poverty levels. For instance if we pick a correlation of -.69 (Coder (1992) as referenced by Bound et al (2001) Table 1) the

may be desirable in multiplicative error models. However, this will be approximately true given the small variance approximations adopted in this paper.

[7] For a discussion of the suitability of this specification see Cowell (1995).

[8] See Altonji and Doraszelski (2005).

[9] This is within the range of estimates reported by Bound et al (Section 6).

[10] For simplicitly in this latter case we also assume that u and y^* are bivariate normal. This allows us to obtain analytical expressions for the distributions of concern. Theorem 2 does not require any such parametric assumptions. More general distributions could be incorporated into our example using Monte-Carlo methods.

[11] There are actually 3 curves in Figure 2. However, given the values used in calibrating our model the true distribution and the distribution with correlated measurement error are indistinguishable.

estimated poverty rates are 8% for the true model, 11% for the independent case and 3% for the correlated case.

3 Conclusion

In this paper we present a simple geometric exposition of the impact of non-classical measurement error for the derivation of distribution functions. For a popular set of distributions we show that positively correlated errors will unambiguously worsen the bias throughout the distribution, while negatively correlated error may help offset the bias that arises with independent errors. For other distributions the consequence of correlated errors differs throughout the distribution in a way that depends on the curvature of the true underlying distribution.

References

Altonji, J. and Doraszelski, U (2005), The Role of Permanent Income and Demographics in Black-White Differences in Wealth, forthcoming Journal of Human Resources.

Angrist, J. and A. Krueger (1999), Empirical Methods in Labor Economics, in O. Ashenfelter and D. Card (eds.), Handbook of Labor Economics Vol.3A, Elsevier.

Bound, J, C. Brown and G. Duncan (1994), Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data, Journal of Labor Economics, Vol. 12, No. 3, 345-368

Bound, J and C. Brown and N. Mathiowetz (2001), Measurement Error in Survey Data, in Handbook of Econometrics (editors J.J. Heckman and E. Leamer), Vol. 5, 3707-3843.

Carroll, R.J. , D. Ruppert and L. Stefanski (1994), Measurement Error in Nonlinear Models, Chapman and Hall, London.

Chesher, A. (1991), The Effect of Measurement Error, Biometrika, Vol.78, 451-462.

Chesher, A. and Schluter, C., (2001), Welfare Measurement and Measurement Error, Review of Economic Studies, 69(2), 357-378.

Chesher, A. Dumangane, M. and Smith, R., (2002). Duration Response Measurement Error, Journal of Econometrics, 111(2), 169-194.

Cowell, F. (1995), Measuring Inequality, Prenctice-Hall.

Fuller, W (1987), Measurement Error Models, New York: Wiley and sons.

Horowitz, J. and C. Manski (1995), Identification and Robustness with Contaminated and Corrupted data, Econometrica, Vol. 63, pp. 281-302.

Kiefer, N. (1988), Economic Duration Data and Hazard Functions, Journal of Economic Literature, Vol. XXVI, June, pp. 646-679.

Lehmann, E.L. (1955), Ordered Families of Distributions, Annals of Mathematical Statistics, Vol. 26, 399-419.

O'Neill, D, O.Sweetman and D. Van de gaer (2004), The Consequences of Specification Error for Distributional Analysis, Economics Dept., National University of Ireland Maynooth.

Serneels, P. (2002), Explaining Non-Negative Duration Dependence Among the Unemployed, The Centre for the Study of African Economies Working Paper Series. Working Paper 172.

Torelli, N and U. Trivellato (1989), Youth Unemployment Duration from the Italian Labour Force Survey, European Economic Review, Vol. 33, no. 407-415.

Zimmerman, D.J. (1992), Regression Towards Mediocrity in Economic Stature, American Economic Review, Vol.82, 409-429.

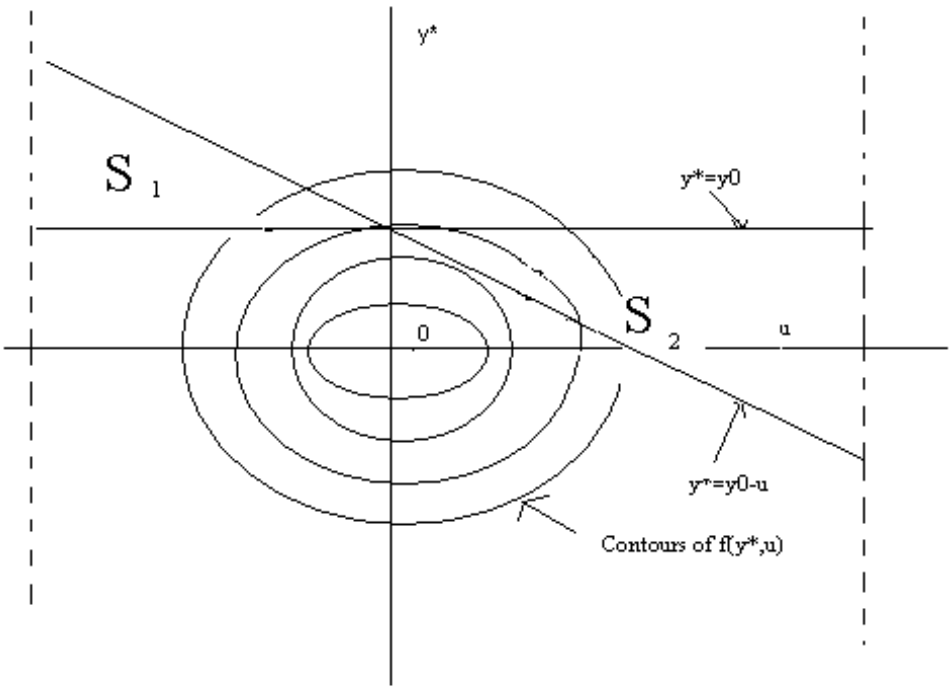

Figure 1: Graphical Derivation of the Bias when the Dependent Variable is Measured with Error.

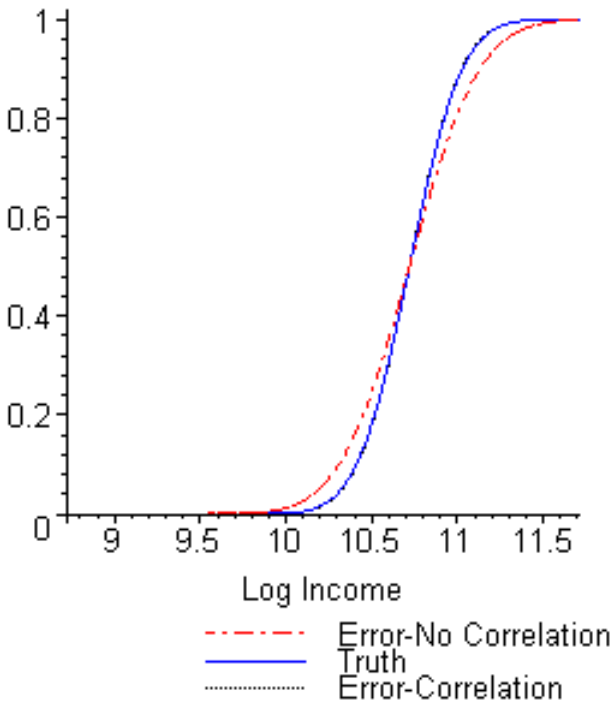

Figure 2: Calibrated Distributions of Income with and without Measurement Error