UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working papers
Economics
12-19
July, 2013

Departamento de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 916249875

Measurement Error and Imputation of Consumption in Survey Data*

Rodolfo G. Campos† Iliana Reggio‡

This version: July 24, 2013

Abstract

We study how estimators used to impute consumption in survey data are inconsistent due to measurement error in consumption. Previous research suggests instrumenting consumption to overcome this problem. We show that, if additional regressors are present, then instrumenting consumption may still produce inconsistent estimators. This inconsistency arises from the correlation between additional regressors and measurement error. We propose an additional condition to be satisfied by the instrument that reduces measurement error bias. This condition is directly observable in the data. We apply our findings by revisiting recent research that imputes consumption data from the CEX to the PSID.

JEL classification: C13, C26, E21

Keywords: consumption, measurement error, instrumental variables, Consumer Expenditure Survey, Panel Study of Income Dynamics

* We thank Richard Blundell and Jesús Carro for their very useful comments on a previous draft of this paper. Magdalena Opazo provided outstanding research assistance. We gratefully acknowledge the financial support by the Spanish Ministerio de Ciencia y Tecnología (grants ECO2009-13169 and ECO2009-11165) and Ministerio de Economía y Competitividad (grants ECO2012-38134 and ECO2012-31358).

† Department of Economics, IESE Business School, Camino del Cerro del Águila 3, 28023 Madrid, Spain; email: rcampos@iese.edu.

‡ Department of Economics, Universidad Carlos III, Calle Madrid 126, 28903 Getafe (Madrid), Spain; email: ireggio@eco.uc3m.es.

1 Introduction

Even though a particular variable may be unavailable in a researcher's main data set, it may be available in a secondary data set. If both data sets share a common set of characteristics, then it is possible to resort to imputation. An imputation procedure consists of characterizing a relationship between the variable of interest and other variables that are observed in both data sets, and to use this relationship to construct the variable of interest in the main data set. In economics, where data is frequently obtained from surveys, an important question is how the imputation procedure is affected by measurement error. Measurement error affects an imputation because it impacts the estimation of the parameters that will be used to construct the imputed variable. In the classical errors-in-variables (CEV) case, it is well known that OLS estimators are inconsistent.

In this paper we address how instrumental variable (IV) techniques can be used to solve the problem of measurement error in imputation procedures. We make three contributions. First, we theoretically show that the conditions under which IV methods yield consistent estimators in the presence of measurement error are more stringent than what had been claimed in previous research. Second, we derive a specific condition which delivers consistent estimators even in the presence of measurement error. This condition can be verified in the data. Third, we use a specific example that imputes consumption from the Consumer Expenditure Survey (CEX) to the Panel Study of Income Dynamics (PSID) to illustrate how our findings may help shape future research in which imputation procedures are used. Our example also serves to examine how the influential estimations of the degree of insurance by Blundell, Pistaferri, and Preston (2008) are affected by the application of our methodology.

Our focus on the imputation of consumption data is motivated by the usual lack of good measures of total consumption despite the relevance of this variable for various strands of research. For example, consumption data at the consumer unit level is required to answer empirical questions such as the permanent income hypothesis, retirement behavior, and the analysis of consumption inequality.

For the United States there does not exist a panel survey which simultaneously provides broad and complete measures of income and consumption expenditure, and which allows to track a consumer unit for an extended period of time. Thus, there is frequently a need to merge databases which requires imputing data from one source to the other. The

PSID and the CEX are natural candidates to be merged. The PSID is the most comprehensive longitudinal data set in the United Sates containing income and socioeconomic information. The PSID does, however, not include a measure of total consumption. Comprehensive consumption data at the consumer unit level is available from the CEX. The CEX, on the other hand, follows households for at most four consecutive quarters and has less information on incomes.

A widely used method to impute consumption data from the CEX to the PSID was proposed by Skinner (1987). This method is used in several articles including, among others, Palumbo (1999), Dynan (2000), and Bernheim, Skinner, and Weinberg (2001). Skinner's imputation procedure consists in selecting several consumption expenditure categories which are common to both data sets, like food consumption (at home and away), utility payments, value of the house, and car ownership in order to run an OLS regression in the CEX of total consumption on the selected consumption categories. The parameters estimated in the regression are then used to construct artificial consumption data for the PSID.

Skinner's procedure does not address the issue of measurement error. Measurement error is a pervasive problem in consumption surveys, in particular if recall methods are used to collect consumption data, as shown in several studies. Battistin and Padula (2010) document measurement error in recall methods for the United States by comparing interview and diary data collected for food consumption in the CEX. Using Italian data, Battistin, Miniaci, and Weber (2003) find important heaping and rounding problems in recall consumption data. Ahmed, Brzozowski, and Crossley (2006) compare recall and diary data from the Canadian Food Expenditure survey and find important measurement errors in recall food consumption.

An imputation procedure that explicitly addresses the problem of measurement error was advanced by Blundell, Pistaferri, and Preston (2008). Although relatively recent, this imputation procedure has been widely used; examples include the work by Guvenen and Smith (2010), Hryshko, Luengo-Prado, and Sorensen (2010), Attanasio, Hurst, and Pistaferri (2012), and Michelacci and Ruffo (2013). Others, such as Kaplan and Violante (2010), Abraham, Koehne, and Pavoni (2012), and Broer (2012) have directly used the original imputed data from Blundell, Pistaferri, and Preston (2008), which is available online.[1]

[1] At `http://www.aeaweb.org/aer/data/dec08/20050545_data.zip`.

Blundell, Pistaferri, and Preston (2008) propose to impute consumption by estimating a demand for food equation. To deal with measurement error they instrument total consumption using wage data. Blundell, Pistaferri, and Preston (2008) argue, partly in a companion paper (Blundell, Pistaferri, and Preston, 2004), that IV regressions produce consistent estimators in the demand for food estimation under certain conditions. Moreover, they show that the variance of imputed consumption will correctly track the evolution of true consumption.

In Section 2 of this paper we show that even if consumption is instrumented, as in the proposal by Blundell, Pistaferri, and Preston (2004, 2008), the presence of additional covariates in the demand for food equation may produce inconsistent estimators because of measurement error. The reason is that measurement error in consumption biases the estimates through its correlation with the additional covariates. As a corollary to this result, in Section 2.2 we flesh out how the bias due to measurement error creates a gap between the evolution of imputed consumption and true consumption.

In Section 2.3 we quantitatively assess the magnitude of the measurement error bias through Monte Carlo simulations. We find that the bias in the estimated coefficients and in the variance of consumption may be substantial even for small correlations between measurement error and the covariates.

In Section 3 we revisit the demand for food estimation of Blundell, Pistaferri, and Preston (2008) to provide a practical example of the application of our results. Our findings in Section 2 imply that the bias attributable to measurement error can be mitigated, and even eliminated, if the additional covariates are orthogonal to the instrument. Whether this condition holds can, in principle, be verified in the data because the covariates and the instrument are observable. Changing the specification for the demand for food and using alternative instruments give rise to varying degrees of correlation between the covariates and the instruments. We experiment with different specifications and alternative instruments and report the resulting estimates.

2 Imputation with measurement error

The first step in an imputation procedure is the estimation of a relationship between the variable targeted for imputation and variables available in both surveys that are used.

In this section we show that the presence of covariates additional to those that are instrumented leads to estimation results that are, in general, inconsistent. We do so in the context of the imputation procedure by Blundell, Pistaferri, and Preston (2004, 2008), which imputes consumption data to the PSID using regression parameters estimated from a demand function using CEX data.

2.1 IV estimation in presence of an additional covariate

Blundell, Pistaferri, and Preston (2004, 2008) estimate a demand equation for food which, if augmented by an additional variable, takes the form

$$f_i = \beta_0 + \beta_1 d_i + \gamma c_i + e_i. \tag{1}$$

Demand for food is a function that relates expenditure on food f (either in levels or in logs) to total non-durable expenditure c (also measured in levels or logs) and possibly other variables. The variable denoted by d is one such variable. Think of d as any additional variable that should be included in the demand equation. For example, it may be the price of food, a price of other substitutable or complementary goods, or a characteristic of the household that acts as a demand shifter. The parameter γ measures the sensitivity of food consumption to total consumption. If variables are measured in logs, then it is called the budget elasticity. The interpretation of the parameter β_1 depends on what d is. The only other term in the equation is unobserved heterogeneity, which is represented by e. The single departure from the specification of Blundell, Pistaferri, and Preston (2004) is that the variable d is included.

Letting c_i^* denote measured nondurable consumption expenditure, c_i true nondurable consumption expenditure, and u_i an error term, measurement error is modeled as follows:

$$c_i^* = c_i + u_i, \tag{2}$$

Because true consumption c_i is unobservable, in practice, the demand for food in (1) cannot be estimated. Substituting (2) into the demand equation in (1) yields an equation in terms of c_i^*, which can be estimated:

$$f_i = \beta_0 + \beta_1 d_i + \gamma c_i^* + e_i - \gamma u_i. \tag{3}$$

Imputation proceeds by using the parameters from this equation estimated with CEX data together with observations of f and d from the PSID to obtain predicted consumption observations for all the households in the PSID.

Estimation of the parameters from the demand for food equation (3) is not straightforward. It is well known that in the presence of classical errors-in-variables (CEV), OLS estimators from the food-demand equation (3) are inconsistent.[2] IV methods may prove useful to obtain consistent estimators even under CEV if it is possible to find an observable variable z that does not belong in (3) and that is partially correlated with c^*. Blundell, Pistaferri, and Preston (2004) prove that, if total consumption c^* is the sole regressor, a valid instrument z eliminates any asymptotic bias and yields consistent estimators.

Given that the demand for food will invariably include additional regressors in practice, the question we consider is under which conditions consistency is achieved if an additional regressor d is added. To answer this question, in Proposition 1 we derive the probability limits of β_1, and γ as functions of the asymptotic theoretical biases when z is used as an instrument.[3]

Proposition 1

Let z be a valid instrument for c^*, d an exogenous regressor in (1)*, $\hat{\beta}_1$ the IV estimator of β_1, and $\hat{\gamma}$ the IV estimator of γ. Then, the IV estimation of* (3) *yields the following asymptotic results*

$$\operatorname{plim} \hat{\beta}_1 = \beta_1 - \gamma \frac{Cov(c^*, z)Cov(d, u)}{V(d)Cov(c^*, z) - Cov(c^*, d)Cov(d, z)} \tag{4}$$

$$\operatorname{plim} \hat{\gamma} = \gamma \left[1 + \frac{Cov(d, u)Cov(d, z)}{V(d)Cov(c^*, z) - Cov(c^*, d)Cov(d, z)} \right] \tag{5}$$

In contrast to what happens in the absence of an additional regressor d, both parameters are inconsistent despite z being a valid instrument for c^*. To ensure consistent estimators, the additional variable d (i.e. any additional variable that belongs in the

[2] See, for example, Wooldridge (2002, Ch. 4).

[3] All proofs are in the Appendix.

demand for food) would need to be instrumented as well. Because it is not, asymptotic bias due to measurement error sneaks back into the estimates through the covariance between the additional variable and measurement error $Cov(d, u)$, which is potentially non-zero.

2.2 The variance of consumption

A reason for imputing consumption to the PSID is to track the evolution of consumption inequality through time. Since a measure of consumption inequality that is commonly used is the variance of consumption, the question is how well the variance of imputed consumption tracks the variance of true consumption in the presence of measurement error. Blundell, Pistaferri, and Preston (2004) show that, in the absence of an additional regressor d, the sample variance of imputed consumption converges in probability to the same limit as the variance of true consumption, up to an additive term. In their setting the variance of imputed consumption is just an upward translated version of the variance of true consumption with the same time trends. In this section we address how measurement error affects the relationship between the variances of imputed and true consumption in the presence of the additional regressor d.

Imputed consumption $\hat{c}$ is obtained from the estimated value of the budget share $\hat{\gamma}$ and the other estimated parameters: $\hat{\beta}_0$ and $\hat{\beta}_1$. After inverting the demand for food, imputed consumption is calculated as

$$\hat{c_i} = \frac{1}{\hat{\gamma}} \left[f_i - \hat{\beta}_0 - \hat{\beta}_1 d_i \right]. \tag{6}$$

By using the demand for food (1) to replace f_i in the above equation, we obtain an equation involving $\hat{c_i}$ and c_i:

$$\hat{c_i} = \frac{1}{\hat{\gamma}} \left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) d_i + \gamma c_i + e_i \right]. \tag{7}$$

Because in this equation c_i is multiplied by the ratio of the true budget elasticity to the estimated budget elasticity $\frac{\gamma}{\hat{\gamma}}$, the relationship between the variances of imputed and true consumption will be impacted by its square. The complete expression for the probability limit of consumption is derived in Proposition 2.

Proposition 2

The probability limit of the variance of predicted consumption is

$$\begin{aligned}\operatorname{plim} V(\hat{c}) = \left(\frac{1}{1+\frac{Cov(d,u)Cov(d,z)}{V(d)Cov(c^*,z)-Cov(c^*,d)Cov(d,z)}}\right)^2 \Bigg[& \operatorname{plim} V(c) \\ & + \frac{1}{\gamma^2}\operatorname{plim} V(e) + \frac{2}{\gamma}\operatorname{plim} Cov(e,c) \\ & + \left(\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z)-Cov(c^*,d)Cov(d,z)}\right)^2 \operatorname{plim} V(d) \\ & + 2\left(\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z)-Cov(c^*,d)Cov(d,z)}\right)\operatorname{plim} Cov(c,d)\Bigg] \end{aligned} \tag{8}$$

In comparison to the result by Blundell, Pistaferri, and Preston (2004), the presence of the additional regressor d augments the expression for the variance of imputed consumption by two additive terms. Because of measurement error, the variance of imputed consumption is additively impacted by the variance of the additional regressor $V(d)$ and the covariance $Cov(c,d)$. More importantly, the multiplicative term in front of $\operatorname{plim} V(c)$ is different from one because $\hat{\gamma}$ is inconsistent even if consumption is instrumented. Thus, the variance of predicted consumption does not move in lockstep with the variance of true consumption, implying that the evolution of the variance of true consumption over time is not tracked by the variance of imputed consumption.

2.3 An orthogonality condition

Consistency of estimates cannot be achieved by instrumenting consumption when the measurement error is correlated with d. Proposition 1 shows that the resulting asymptotic bias of both estimates is proportional to $Cov(d,u)$, which is unobservable. There is, however, an observable orthogonality condition that removes the bias in the estimate of one of the parameters: the budget elasticity. If the condition $Cov(d,z)=0$ is satisfied, implying that the instrument for consumption expenditure is orthogonal to the additional regressor, then the estimate $\hat{\gamma}$ can be shown to be consistent. In turn, the consistency of $\hat{\gamma}$ implies that the slope coefficient in the expression for the variance in

Proposition 2 is one, and that the variance of true consumption is tracked by the variance of imputed consumption. Thus, the orthogonality condition provides an indication that can be verified in the data of whether the evolution of the variance of imputed consumption over time is representative of that of true consumption. This result is formally stated in Proposition 3.

Proposition 3

Let z be a valid instrument for c^* and d an exogenous regressor. If the instrument z is uncorrelated to the additional regressor, i.e. $Cov(d, z) = 0$, then the IV estimator for γ in (3) is consistent and the slope coefficient multiplying the variance of true consumption in Proposition 2 is one.

Strict fulfillment of the orthogonality condition $Cov(d, z) = 0$ will likely be impossible in practice. Whether the condition comes close to be fulfilled will depend on the particular data set, and on which additional variables are included in the demand for food. For example, the range of correlations between regressors and the instrument in the data of Blundell, Pistaferri, and Preston (2008) goes from close to zero (correlation with having three or more kids) to 0.627 (correlation with being a high school graduate).

To gauge how estimators are affected by deviations from the orthogonality condition $Cov(d, z) = 0$ we conduct a Monte Carlo simulation. Our simulation exercise measures how $Cov(d, z)$ affects measurement error bias for different values of $Cov(u, d)$ and provides a sensitivity analysis of the estimated parameters if the orthogonality condition $Cov(d, z)$ is not satisfied.

The range of sample correlations in the data of Blundell, Pistaferri, and Preston (2008) goes from close to 0 to 0.6. Thus, in our simulations we consider four values for the correlations between d and z: 0, 0.1, 0.3, and 0.6 (and call them *zero*, *low*, *medium*, and *high*). We use the same range of values for the unobservable correlation of the additional regressor with measurement error. For these correlations, we assume a hypothetical true value of $\gamma = 0.5$ and estimate $\hat{\gamma}$ in an IV regression.[4]

We present three types of results. Table 1 reports the estimates of $\hat{\gamma}$. Table 2 shows the percentage of rejections of the null hypothesis that the estimated coefficient is equal to

[4]Random variables are taken from standard Normal distributions. We do 10,000 simulations with sample size 5,000.

the true value of γ for a significance level of 5 percent. A consistent estimator would yield 5 percent of rejections in all cases. Finally, Table 3 addresses whether the variance of imputed consumption tracks the evolution of the variance of true consumption over time —it reports $\left(\frac{\gamma}{\hat{\gamma}}\right)^2$.

Table 1: *Estimates of $\hat{\gamma}$ for different correlations between the additional regressor and the measurement error and the instrument. The labels (Zero, Low, Medium, High) correspond to correlations of (0, 0.1, 0.3, 0.6).*

		$Corr(u,d)$			
		Zero	Low	Medium	High
$Corr(z,d)$	Zero	0.500	0.500	0.500	0.500
	Low	0.500	0.507	0.520	0.540
	Medium	0.500	0.522	0.566	0.632
	High	0.500	0.562	0.688	0.876

Table 2: *Percentage of rejections of the null hypothesis that the coefficient is equal to the true value, for different correlations between the additional regressor and the measurement error and the instrument. The labels (Zero, Low, Medium, High) correspond to correlations of (0, 0.1, 0.3, 0.6).*

		$Corr(u,d)$			
		Zero	Low	Medium	High
$Corr(z,d)$	Zero	0.05	0.05	0.05	0.05
	Low	0.05	0.08	0.29	0.85
	Medium	0.05	0.30	0.99	1.00
	High	0.05	0.93	1.00	1.00

Bias due to measurement error disappears if at least one of the correlations is zero, meaning that either the additional regressor is uncorrelated with measurement error or the instrument is uncorrelated with the additional regressor. This follows from the result in Proposition 1 that measurement error produces an asymptotic bias equal to

$$B \equiv \operatorname{plim} \hat{\gamma} - \gamma = \gamma \frac{Cov(u,d)Cov(d,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}. \tag{9}$$

Thus, the first line and the first column of Table 1 have $\hat{\gamma} = \gamma$ if either $Cov(u,d) = 0$ or $Cov(d,z) = 0$. As expected, the empirical rejection rate in Table 2 is 5 percent.

In contrast, if both correlations are positive, biased estimates are obtained. It only takes medium-sized correlations to obtain estimates that exhibit significant bias. For example,

Table 3: *Sensitiveness of the ratio* $\left(\frac{\gamma}{\hat{\gamma}}\right)^2$ *to different correlations between the additional regressor and the measurement error and the instrument. The labels (Zero, Low, Medium, High) correspond to correlations of (0, 0.1, 0.3, 0.6).*

		$Corr(u,d)$			
		Zero	Low	Medium	High
$Corr(z,d)$	Zero	1.000	1.000	1.000	1.000
	Low	1.000	0.975	0.925	0.855
	Medium	1.000	0.918	0.781	0.626
	High	1.000	0.791	0.529	0.326

when both correlations are medium-sized, $\hat{\gamma} = 0.566$ (Table 1) and the null hypothesis that it is equal to the true value of 0.500 is rejected in 99 percent of the simulations (Table 2). The ratio $(\frac{\gamma}{\hat{\gamma}})^2$, which governs the effect of marginal increase in the variance of true consumption on the variance of imputed consumption, drops by more than 20 percent, to 0.781, when both correlations are medium-sized (Table 3).

3 Application

In this section we consider a concrete application that reproduces the estimations of Blundell, Pistaferri, and Preston (2008) using specifications with different degrees of correlation between instruments and regressors. The application illustrates the kind of changes that can be made to mitigate measurement error bias in the estimation used for imputation. In addition, it serves as a sensitivity analysis of the results by Blundell, Pistaferri, and Preston (2008).

To implement the imputation procedure, Blundell, Pistaferri, and Preston (2008) estimate a demand equation for food using CEX data from 1980 to 1992 of the form

$$\ln f_{i,t} = W'_{i,t}\mu + p'_{i,t}\theta + \beta(D_i)\ln c_{i,t} + \varepsilon_{i,t}, \tag{10}$$

where $\ln f$ stands for the log of real food consumption, W contains demographic variables that are available in both the CEX and the PSID, p contains relative prices, $\ln c$ stands for the log of nondurable expenditure (available only in the CEX), and ε captures unobserved heterogeneity in the demand for food and measurement error in food expenditure. The coefficient on nondurable expenditure is allowed to vary with demographic characteristics

(D) and over time. Nondurable consumption is instrumented using the average of the hourly wage of the husband (by cohort, year, and education) and the average of the hourly wage of the wife (also by cohort, year, and education).

In light of our results, a potential problem with the estimation arises if instruments are correlated with the additional regressors in equation (10). Our simulations showed that medium-sized correlations were enough to induce a significant departure of the estimate from the true value of the budget elasticity. In the data there are two groups of variables with high sample correlations with the instruments: education and prices. Their correlations with the average of the hourly wage of the husband are presented in the first column of Table 4. The correlation between hourly wage of the husband and education dummies are around 0.6 (what we called a high correlation); the correlations with prices are a somewhat lower.

Table 4: *Correlations between the hourly wage of the husband and selected regressors.*

	BPP	Alt. IV	BPP - Real	Alt. IV - Real
Prices				
- Food	0.467	0.753	-0.072	-0.150
- Alcohol and Tobacco	0.469	0.758	-0.069	-0.134
- Fuel and Utilities	0.427	0.696	-0.066	-0.117
- Transports	0.455	0.738	-0.078	-0.148
Education				
- Elementary	-0.567	-0.058	-0.607	-0.013
- HS Graduate	0.627	0.055	0.674	0.010

Correlations computed for the sample used in the estimation of the food equation. We select regressors with the largest correlations with the average hourly wage of the husband. Column show the correlations with the different instruments used in the analysis.

We address the high correlation between education and prices and the instruments considering two complementary ways of reducing measurement error bias. One way is to drop the problematic regressors in the estimation of the budget elasticity, and the other is to use an alternative instrument that is less correlated with the regressors.

As a benchmark, we replicate the specification of Blundell, Pistaferri, and Preston (2008) and obtain the same results as they do. The budget elasticity is estimated to be 0.850 (Table 5, Col. 1). In contrast, if education dummies are excluded from the baseline specification, then a lower budget elasticity of 0.799 is obtained (Col. 3). It can be

argued that this is not a fair comparison because the exclusion of education dummies also implies dropping the interactions between education and $\ln c$. Thus, we also re-estimate the specification of Blundell, Pistaferri, and Preston (2008) without the interactions and report the results in Column 2. The result is stronger. Comparing columns 2 and 3, we find that the estimated budget elasticity drops from 1.081 to 0.799 if education dummies are excluded.

The difference in the results when education is removed suggests that measurement error could be biasing the estimate of the budget elasticity upward. It is far from a definitive proof; if education dummies are deemed necessary in the demand equation, then their removal may generate omitted variable bias. On the other hand, the demand function contains total consumption expenditure that is instrumented by wage rates. This reduces the role of education as a proxy for income. In any case, the sensitivity of the estimate of the budget elasticity to the removal of education dummies should at least cast doubt on the exact value of the estimate.

The second approach does not require to drop any variables from the demand equation. The difference is in the construction of the instruments. To achieve less correlation with education we calculate the average hourly wage of the husband and the average hourly wage of the wife by cohort and year but without conditioning on education. Doing so lowers the correlation between the instrument and education dummies to close to zero (Table 4, Col. 2).

Using these alternative instruments, which by construction are less correlated with education, the point estimate of the budget elasticity drops to 0.718 (Table 5, Col. 4). In this case, the relevant comparison is with the original estimate of 0.850. The estimate is less precise and does not allow to statistically distinguish between these values at the usual probability thresholds. Nevertheless, if the difference in the point estimates is attributed to measurement error, then the evidence indicates that the budget elasticity is biased upward, as before.

The other group of variables correlated with the instruments are prices. Their correlation with the instrument is not removed by the alternative definition of the instrument; in fact, correlations with prices are higher (Table 4, Col. 2). The reason behind the large correlation with prices is that total consumption expenditure enters the food demand equation in nominal terms. Wages used to instrument consumption are also nominal. Wages and prices are linked by inflation.

Table 5: *Sensitivity of the budget elasticity to different specifications and to the use of alternative instruments.*

VARIABLES	(1) BPP	(2) No Interactions	(3) No Education	(4) Alt. IV
$\ln c$	0.850***	1.081***	0.799***	0.718***
	(0.151)	(0.112)	(0.032)	(0.203)
$\ln c$ x HS	0.073			-0.004
	(0.072)			(0.076)
$\ln c$ x College	0.083			0.058
	(0.089)			(0.108)
Observations	14,430	14,430	14,430	14,430
R-squared	0.671	0.619	0.687	0.682
RMSE	0.249	0.268	0.243	0.245

*Standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). In the first three columns the instruments are the average (by cohort, year, and education) of the hourly wage of the husband and the average (also by cohort, year, and education) of the hourly wage of the wife. In column (4) the instruments are the average (by cohort and year) of the hourly wage of the husband and the average (also by cohort and year) of the hourly wage of the wife.*

A more flexible specification for the demand for food breaks this link. We separate nominal expenditures into a real component and a price index, and do the same with the instrument. We do so by deflating nominal values using the Consumer Price Index and add this index as an additional regressor. This change in the specification reduces the correlation between the instrument and additional regressors. Correlations with prices are lower both for the original instrument in Blundell, Pistaferri, and Preston (2008) (Table 4, Col. 3) and for the alternative definition of the instrument (Table 4, Col. 4).

We repeat our previous regressions using real consumption expenditure instrumented by real wages. Results are shown in Table 6. The columns are analogous to those in Table 5. The specification of Blundell, Pistaferri, and Preston (2008) with real expenditures and real wages produces an estimate of 0.937 (Table 5, Col. 1). Again, lower point estimates for the budget elasticity are obtained when education dummies are dropped (Table 5, Col. 3) and when alternative definition is used for the instruments (Table 5, Col.4).

Table 6: *Sensitivity of the budget elasticity to different specifications and the use of alternative instruments using real expenditures instrumented by real wages.*

	(1)	(2)	(3)	(4)
VARIABLES	BPP	No Interactions	No Education	Alt.IV
$\ln c$	0.937***	1.025***	0.786***	0.772***
	(0.119)	(0.100)	(0.032)	(0.211)
$\ln c$ x HS	0.112			-0.101
	(0.129)			(0.129)
$\ln c$ x College	0.018			-0.151
	(0.121)			(0.126)
Observations	14,430	14,430	14,430	14,430
R-squared	0.655	0.635	0.686	0.682
RMSE	0.255	0.262	0.243	0.245

*Standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). All regressions use consumption and wages in real terms. In the first three columns the instruments are the average (by cohort, year, and education) of the hourly wage of the husband and the average (also by cohort, year, and education) of the hourly wage of the wife. In column (4) the instruments are the average (by cohort and year) of the hourly wage of the husband and the average (also by cohort and year) of the hourly wage of the wife.*

In conclusion, our results in this section indicate that in the estimation of the demand for food, the instruments used by Blundell, Pistaferri, and Preston (2008) are highly correlated with two sets of additional regressors, education and prices. Modifying the specification and using alternative instruments are two approaches that lead to lower correlations between those instruments and the additional regressors. Estimated coefficients of the budget elasticity were lower in all cases, suggesting that the unobserved correlation between additional regressors and measurement error is leading to an overestimation of this elasticity.

As a final point, we consider how our results affect the main results by Blundell, Pistaferri, and Preston (2008). Their article uses the imputation procedure only as an intermediate step. The final objective is to estimate the response of household consumption to permanent and transitory shocks. In Table 7 we show how their answers are influenced by the different demands for food we estimated.[5]

Blundell, Pistaferri, and Preston (2008) estimate the response of consumption to permanent and transitory income shocks (denoted by ϕ and ψ). The response of consumption to transitory income shocks is not substantially affected by the different imputation procedures. It remains low and is not significantly different from zero. In contrast, either dropping education dummies or using the alternative definition of the instruments, leads to a rise in ϕ, the response of consumption to permanent shocks. This happens regardless of whether nominal or real expenditures are used for the imputation. This suggests that the likely bias in the budget elasticity implies an underestimation of the impact of permanent income shocks on consumption.[6]

4 Conclusion

We show that the presence of measurement error produces inconsistent estimates in procedures used to impute consumption. Inconsistency is not easily removed using instrumental variables, an approach used in the previous literature. The source of the

[5]For details on the estimation of the response to permanent and transitory income shocks consult Blundell, Pistaferri, and Preston (2008). Blundell, Pistaferri, and Preston (2008) have made their data and code publicly available. We use their data and adapt their code to obtain our results.

[6]On the other hand, the values estimated for ϕ are inside the range of values considered by Blundell, Pistaferri, and Preston (2008) in their robustness checks.

Table 7: *Robustness of the response to permanent and transitory income shocks.*

	BPP	No education	Alt. IV
Nominal Imputation			
ϕ	0.6423	0.7882	0.8186
	(0.0945)	(0.1153)	(0.1191)
ψ	0.0533	0.0558	0.0601
	(0.0435)	(0.0523)	(0.0584)
Real Imputation			
ϕ	0.5988	0.7871	0.7668
	(0.0877)	(0.1150)	(0.1106)
ψ	0.0453	0.0545	0.0501
	(0.0396)	(0.0519)	(0.0553)

inconsistency is the presence of additional regressors that are likely to be correlated with consumption measurement error.

The theoretical contribution of this paper is to show the existence of the asymptotic bias due to measurement error and to derive a mathematical expression for it. Using this expression we further show that if the interest lies in the evolution of the variance of consumption, then IV estimation may prove useful, provided an additional orthogonality condition between the instrument and the additional regressors is satisfied. Specifically, the variance of imputed consumption tracks the evolution of the variance of true consumption if additional regressors are uncorrelated with the instrument used in the imputation. In a Monte Carlo simulation we quantify the bias due to measurement error when the orthogonality condition is not exactly satisfied and find that correlations of around 0.3 may already pose a significant problem.

We use the imputation of the influential article by Blundell, Pistaferri, and Preston (2008) as a concrete application to illustrate techniques that may mitigate measurement error bias. We consider two different approaches to lower the correlation between instruments and regressors: changes in the specification (such as the exclusion of problematic variables) and changes in the definition of instruments. In our results, we find lower

estimates for the budget elasticity of food consumption. In turn, in the context of the model by Blundell, Pistaferri, and Preston (2008), these revised estimates imply a larger role of permanent income shocks in driving consumption.

Because there is frequently a need to impute consumption data across databases, and because measurement error is a pervasive problem in survey data, our findings should prove useful to researchers who require the imputation of consumption to address a larger set of questions, such as the permanent income hypothesis, retirement behavior, and the analysis of consumption inequality.

References

Abraham, A., S. Koehne, and N. Pavoni (2012): "Optimal income taxation with asset accumulation," MPRA Paper 38629, University Library of Munich, Germany.

Ahmed, N., M. Brzozowski, and T. F. Crossley (2006): "Measurement Errors in Recall Food Consumption Data," *The Institute for Fiscal Studies*, 06/21.

Attanasio, O., E. Hurst, and L. Pistaferri (2012): "The Evolution of Income, Consumption, and Leisure Inequality in The US, 1980-2010," NBER Working Papers 17982, National Bureau of Economic Research, Inc.

Battistin, E., R. Miniaci, and G. Weber (2003): "What Do We Learn from Recall Consumption Data?," *The Journal of Human Resources*, Vol. 38, No. 2, 354–385.

Battistin, E., and M. Padula (2010): "Survey Instruments and the Reports of Consumption Expenditures: Evidence from the Consumer Expenditure Surveys," CSEF Working Papers 259, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.

Bernheim, D., J. Skinner, and S. Weinberg (2001): "What Accounts for the Variation in Retirement Wealth Among US Households?," *American Economic Review*, 91(4), 832–857.

Blundell, R., L. Pistaferri, and I. Preston (2004): "Imputing Consumption in the PSID using food demand estimates from the CEX.," *Institute for Fiscal Studies Working Paper*, 04/27.

——— (2008): "Consumption Inequality and Partial Insurance," *American Economic Review*, 98(5), 1887–1921.

Broer, T. (2012): "The wrong shape of insurance? What cross-sectional distributions tell us about models of consumption-smoothing," mimeo.

Dynan, K. E. (2000): "Habit Formation in Consumer Preferences: Evidence from Panel Data," *American Economic Review*, 90(3), 391–406.

Guvenen, F., and A. Smith (2010): "Inferring Labor Income Risk from Economic Choices: An Indirect Inference Approach," NBER Working Papers 16327, National Bureau of Economic Research, Inc.

Hryshko, D., M. J. Luengo-Prado, and B. E. Sorensen (2010): "House prices and risk sharing," *Journal of Monetary Economics*, 57(8), 975–987.

Kaplan, G., and G. L. Violante (2010): "How Much Consumption Insurance beyond Self-Insurance?," *American Economic Journal: Macroeconomics*, 2(4), 53–87.

Michelacci, C., and H. Ruffo (2013): "Optimal Life Cycle Unemployment Insurance," mimeo, CEMFI.

Palumbo, M. G. (1999): "Uncertain Medical Expenses and Precautionary Saving Near the End of the Life Cycle," *The Review of Economic Studies*, 66(2), 395–421.

Skinner, J. (1987): "A superior measure of consumption from the Panel Study of Income Dynamics," *Economics Letters*, 23, 213–216.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

A Proofs

Proof of Proposition 1:

Start by stating a basic result of the IV framework.

The general IV framework

The general model consists of K regressors where one of them (X_k) suffers from measurement error. It is instrumented with a instrumental variable denoted by W.

$$Y = X\theta + \varepsilon \tag{A.11}$$

where $X = [X_1, ..., X_K]$

Let W be a valid instrument for X_k and define $Z = [1\ X_1\ ...X_{k-1}\ \ W\ X_{k+1}\ ...X_K]$.

The IV formula for the estimators of the parameters in equation (A.11) is:

$$\hat{\theta} = (Z^\top X)^{-1} Z^\top y \tag{A.12}$$

Derivation of the probability limits

Apply the formula in (A.12), and properties of convergence in probabilities, to obtain the probability limit of both estimators: $\hat{\beta}_1$ and $\hat{\gamma}$. The plim of the estimator of the parameter of the variable measured with error is

$$\operatorname{plim} \hat{\gamma} = \frac{Cov(f,z)V(d) - Cov(d,z)Cov(d,y)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)} \tag{A.13}$$

Replacing equation (3) in (A.13) yields

$$\begin{aligned}\operatorname{plim} \hat{\gamma} = \frac{1}{\Phi} \{ & V(d) \left[\beta_1 Cov(d,z) + \gamma Cov(c^*,z) + Cov(e,z) - \gamma Cov(u,z)\right] \\ & -Cov(d,z) \left[\beta_1 V(d) + \gamma Cov(d,c^*) + Cov(d,e) - \gamma Cov(d,u)\right]\}\end{aligned} \tag{A.14}$$

where $\Phi \equiv V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)$. After some algebra the probability limit of $\hat{\gamma}$ is

$$\begin{aligned}\operatorname{plim} \hat{\gamma} = \gamma &+ \frac{Cov(e,z)V(d)}{\Phi} - \frac{\gamma Cov(u,z)V(d)}{\Phi} \\ &- \frac{Cov(d,z)Cov(d,e)}{\Phi} + \frac{\gamma Cov(d,z)Cov(d,u)}{\Phi}\end{aligned} \tag{A.15}$$

Define

$$
\begin{aligned}
B^e &= \frac{Cov(e,z)V(d)}{\Phi} \\
B^m &= -\gamma\frac{Cov(u,z)V(d)}{\Phi} \\
B^{ed} &= -\frac{Cov(d,e)Cov(d,z)}{\Phi} \\
B^{md} &= \gamma\frac{Cov(d,u)Cov(d,z)}{\Phi}
\end{aligned}
$$

The probability limit of $\hat{\gamma}$ can then be written as

$$\operatorname{plim}\hat{\gamma} = \gamma + B^e + B^m + B^{ed} + B^{md} \tag{A.16}$$

Assuming that z is a valid instrument: $Cov(z,e) = Cov(z,u) = 0$, and that d is exogenous in (1): $Cov(d,e) = 0$ then:

$$
\begin{aligned}
\operatorname{plim}\hat{\gamma} &= \gamma + B^{md} \\
&= \gamma\left[1 + \frac{Cov(u,d)Cov(d,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}\right]
\end{aligned} \tag{A.17}
$$

A similar derivation is done for β_1; the IV formula for $\hat{\beta}_1$ implies that the probability limit is

$$\operatorname{plim}\hat{\beta}_1 = \frac{Cov(d,f)Cov(c^*,z) - Cov(c^*,x_1)Cov(f,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)} \tag{A.18}$$

Replace f using equation (3) to obtain

$$
\begin{aligned}
\operatorname{plim}\hat{\beta}_1 = \frac{1}{\Phi}\{&Cov(c^*,z)\left[\beta_1 V(d) + \gamma Cov(c^*,d) + Cov(d,e) - \gamma Cov(d,u)\right] \\
&-Cov(c^*,d)\left[\beta_1 Cov(d,z) + \gamma Cov(c^*,z) + Cov(e,z) - \gamma Cov(u,z)\right]\}
\end{aligned} \tag{A.19}
$$

After some algebra, the probability limit of $\hat{\beta}_1$ is

$$
\begin{aligned}
\operatorname{plim}\hat{\beta}_1 = \beta_1 &- \frac{Cov(c^*,d)\left[Cov(e,z) - \gamma Cov(u,z)\right]}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)} \\
&+ \frac{Cov(c^*,z)\left[Cov(d,e) - \gamma Cov(d,u)\right]}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}
\end{aligned} \tag{A.20}
$$

If z is a valid instrument, then $Cov(e,z) = Cov(u,z) = 0$. If d is exogenous in (1), then $Cov(d,e) = 0$. Therefore,

$$\operatorname{plim}\hat{\beta}_1 = \beta_1 - \gamma\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)} \tag{A.21}$$

Proof of Proposition 2:

Start from

$$\hat{c}_i = \frac{1}{\hat{\gamma}}\left[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)d_i + \gamma c_i + e_i\right]. \tag{A.22}$$

Applying properties of convergence in probability, write the probability limit of the sample variance of predicted consumption as

$$\begin{aligned}
\operatorname{plim} V(\hat{c}_i) = &\left(\frac{\gamma}{\operatorname{plim}\hat{\gamma}}\right)^2 \operatorname{plim} V(c) + \left(\frac{\beta_1 - \operatorname{plim}\hat{\beta}_1}{\operatorname{plim}\hat{\gamma}}\right)^2 \operatorname{plim} V(d) \\
&+ \left(\frac{1}{\operatorname{plim}\hat{\gamma}}\right)^2 \operatorname{plim} V(e) + 2\left(\frac{\gamma}{(\operatorname{plim}\hat{\gamma})^2}\right) \operatorname{plim} Cov(e,c) \\
&+ 2\gamma\left(\frac{\beta_1 - \operatorname{plim}\hat{\beta}_1}{(\operatorname{plim}\hat{\gamma})^2}\right) \operatorname{plim} Cov(c,d) \\
&+ 2\left(\frac{\beta_1 - \operatorname{plim}\hat{\beta}_1}{(\operatorname{plim}\hat{\gamma})^2}\right) \operatorname{plim} Cov(e,d)
\end{aligned} \tag{A.23}$$

From Proposition 1, if z is a valid instrument and d is an exogenous regressor, then

$$\begin{aligned}
\operatorname{plim}\hat{\beta}_1 &= \beta_1 - \gamma\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)} \\
\operatorname{plim}\hat{\gamma} &= \gamma\left[1 + \frac{Cov(u,d)Cov(d,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}\right]
\end{aligned} \tag{A.24}$$

Finally, replace $\operatorname{plim}\hat{\gamma}$ and $\beta_1 - \operatorname{plim}\hat{\beta}_1$ in (A.23) to obtain the expression for the probability limit of the sample variance of predicted consumption:

$$\begin{aligned}
\operatorname{plim} V(\hat{c}_i) = &\left(\frac{1}{1 + \frac{Cov(u,d)Cov(d,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}}\right)^2 \left[\operatorname{plim} V(c) + \frac{1}{\gamma^2}\operatorname{plim} V(e)\right. \\
&+ \left(\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}\right)^2 \operatorname{plim} V(d) \\
&+ 2\left(\frac{Cov(c^*,z)Cov(d,u)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}\right) \operatorname{plim} Cov(c,d) \\
&\left.+ \frac{2}{\gamma}\operatorname{plim} Cov(e,c)\right]
\end{aligned} \tag{A.25}$$

Proof of Proposition 3:

If z is a valid instrument then $Cov(e, z) = Cov(u, z) = 0$ and $Cov(c^*, z) \neq 0$. If d is exogenous then $Cov(e, d) = 0$. Coupled with $Cov(d, z) = 0$, this implies that

$$\frac{Cov(u, d)Cov(d, z)}{V(d)Cov(c^*, z) - Cov(c^*, d)Cov(d, z)} = 0. \tag{A.26}$$

Therefore, from Proposition 1

$$\operatorname{plim} \hat{\gamma} = \gamma, \tag{A.27}$$

and from Proposition 2

$$\left(\frac{1}{1 + \frac{Cov(u,d)Cov(d,z)}{V(d)Cov(c^*,z) - Cov(c^*,d)Cov(d,z)}} \right)^2 = 1. \tag{A.28}$$