IMPUTING CONSUMPTION IN THE PSID USING FOOD DEMAND ESTIMATES FROM THE CEX*

Richard Blundell
University College London and Institute for Fiscal Studies.

Luigi Pistaferri
Stanford University, SIEPR and CEPR.

Ian Preston
University College London and Institute for Fiscal Studies.

December 2006

Abstract

In this paper we discuss an empirical strategy that allows researchers to impute consumption data from the CEX to the PSID. The strategy consists of inverting a demand for food equation estimated in the CEX. We discuss the conditions under which such procedure is successful in replicating the trends of the first two moments of the consumption distribution. We argue that two factors appear to be empirically relevant: accounting for differences in the distribution of food expenditures in the two data sets, and accounting for the presence of measurement error in consumption data in the CEX.

Key words: Food expenditure, Imputation.
JEL Classification: D52; D91; I30.

*We thank Orazio Attanasio, David Johnson and Arie Kapteyn for useful comments, and Cristobal Huneeus and Sonam Sherpa for able research assistance. The paper is part of the program of research of the ESRC Centre for the Microeconomic Analysis of Public Policy at IFS. Financial support from the ESRC, the Joint Center for Poverty Research/Department of Health and Human Services and the National Science Foundation (under grant SES-0214491) is gratefully acknowledged. All errors are ours.

1 Introduction

A major difficulty faced by researchers who want to study the consumption and saving behavior of households is the lack of panel data on household expenditures. In the US, for instance, the Consumer Expenditure Survey (CEX) provides comprehensive data set on the spending habits of US households but it follows households for only four quarters at most. While a quarterly rotating panel can be constructed with these data, most of the consumption variability is likely to be of seasonal nature and hence will tend to miss the more important low frequency variability.

Unlike the CEX, the Panel Study of Income Dynamics (PSID) collects longitudinal annual data. The PSID has been following households on a consistent basis since 1968. However, its main problem for the purpose at hand is that it collects data only for a subset of consumption items, mainly food at home and food away from home (with gaps in the survey years 1973, 1988, and 1989).[1] Other panel data sets widely used by economists, such as the National Longitudinal Study (NLS) or the Health and Retirement Survey (HRS), have abundant information on income or wealth, but no information whatsoever on consumption. The lack of longitudinal data on household spending is a problem that is not limited to the US. In the UK, for example, the Family Expenditure Survey (FES) provides comprehensive data on household expenditures, but households are not followed over time. Panel data sets that collect data on income or wealth, such as the British Household Panel Survey (BHPS), typically lack consumption data.

The lack of panel data information on household consumption is puzzling given its economic relevance. Household spending accounts for as much as 65 percent of national output annually. Consumption decisions are crucial determinants of business cycles and growth. And swings in consumer confidence are good predictors of economic booms and recessions. There are certainly relevant problems in the collection of consumption data; but they do not seem worse than those faced in the collection of wealth or income data. Browning, Crossley and Weber (2003) discuss pros and cons of inserting consumption questions in general purpose surveys.

The lack of panel data on total consumption has meant that most of the tests involving individual consumption behavior have been performed using the scanty food expenditure information in the PSID.[2] Examples include Hall and Mishkin (1982), Zeldes (1989), Runkle (1991), and Shea

[1] In the early survey years the PSID collected information on additional consumption items, such as utilities and tobacco. This kind of information has later been discontinued.

[2] Notable exceptions are Jappelli and Pistaferri (2000), Hayashi (1985), and Alessie and Lusardi (1997). Jappelli and Pistaferri use Italian panel data on consumption, while Hayashi uses Japanese data which also include expected consumption information. Both papers estimate Euler equations for consumption. Alessie and Lusardi use longi-

(1995) for tests of the permanent income hypothesis, Cochrane (1991), Altug and Miller (1990) and Hayashi, Altonji and Kotlikoff (1996) for test of the consumption insurance hypothesis, Altonji (1986) and Hotz, Kydland and Sedlacek (1988) for tests of intertemporal substitution in labor supply, Cox, Ng and Waldkirch (2004) for tests of intergenerational linkages in consumption behavior, and Martin (2003) and Hurst and Stafford (2004) for tests of separability between durable and non-durable consumption. Since the dynamics of food consumption differs in important ways from the dynamics of non-durable consumption, this approach has obvious limits. First of all, food is a necessity (i.e., the budget share for food falls as total expenditure rises), which would invalidate any assumption of unit-elastic preferences and also implies that food volatility generally underestimates total consumption volatility. Furthermore, if the goal of the empirical analysis is to estimate structural parameters such as the elasticity of intertemporal substitution or the extent of substitutability/complementarity between consumption and leisure, it is not clear at all that those obtained using food consumption are indicative of the substitutability of total consumption over time or with respect to labor supply. Finally, one has to justify ignoring the influence of price variation either by making the assumption that the demand for food is separable from that of other goods or that relative prices movements are appropriately restricted, either of which might be a strong assumption. Our approach below controls for possible non-separabilities between food and other commodities by adding prices.

An alternative approach to using food is to construct pseudo-panels from repeated cross-section datasets that have a comprehensive measure of consumption, such as the CEX or the FES. In this case, one can study the dynamics of pseudo-persons rather the genuine household dynamics. Browning, Deaton and Irish (1985) and Attanasio and Weber (1993) are two noteworthy examples. While valuable, the main drawback of this strategy is that it cannot tell us anything about the idiosyncratic dynamics of total consumption. Individual heterogeneity is basically summarized by cohort heterogeneity, which may be restrictive. Blundell, Pistaferri and Preston (2005) show that certain dynamic issues in the consumption literature are better addressed with truly longitudinal data rather than pseudo-panels. Intuitively, panel data offer more identification power than repeated cross-sections.

An empirical approach that has at times surfaced in the literature is that of combining information from the CEX and the PSID to impute a measure of consumption to the PSID households.

tudinal Dutch data to test the predictions of the *saving for a rainy day* equation. They define saving as the first difference of household wealth.

The key original reference is Skinner (1987). He proposes to impute total consumption in the PSID using the estimated coefficients of a regression of total consumption on a series of consumption items (food, utilities, vehicles, etc.) that are present in both the PSID and the CEX. The regression is estimated with CEX data. From a statistical point of view, Skinner's approach can be formally justified by the idea of *matching* based on observed characteristics. Ziliak (1998) and Browning and Leth-Petersen (2003) propose as an alternative that of imputing consumption on the basis of income and the first difference of wealth (i.e., as the difference between income and savings). Browning's (1999) notion of so-called m-demand functions provides a useful conceptual framework for analyzing the behavior of one component of spending conditional on another.

Most researchers are resistant to the idea of using imputed data in the place of actual data. If the imputation data were unbiased, they would just take the form of error-ridden data. In this sense, they would not be much different from the micro data empirical economists typically use. In this paper we discuss a strategy that is similar in spirit, although different in terms of economic interpretation, to that proposed by Skinner and others. As Skinner, we also impute consumption data to all PSID households using regression parameters estimated from CEX data. Our approach differs from that of Skinner and others in that we start from a standard demand function for food (a consumption item available in both surveys); we make this depend on prices, total non durable expenditure, and a host of demographic and socio-economic characteristics of the household. Importantly we also allow the budget elasticity to shift with observable household characteristics. Under monotonicity of food demands these functions can be inverted to obtain a measure of non durable consumption in the PSID. We review the conditions that make this procedure reliable and show that it is able to reproduce remarkably well the trends in the consumption distribution, both at the mean and at the variance level. These are precisely the moments of the consumption distribution researchers care about, and so our methodology should be of potentially great interest to applied economists working with consumption data.[3]

The paper has four more sections. Section 2 discusses the imputation procedure and examines a number of extensions. The data are discussed in Section 3, and the results of our exercise in Section 4. In Section 5 we compare our approach to others in the literature. Section 6 provides a summary of the paper.

[3]We introduced our strategy in a companion paper, Blundell, Preston and Pistaferri (2004). Since the first draft of that paper, our method has been used by, among others, Ziliak, Kniesner and Holtz-Eakin (2003), Lehnert (2002), and Fisher and Johnson (2003).

2 The Imputation Procedure

2.1 Imputing Consumption

We use the subscript x to indicate an observation from the CEX (the *input* data set) and the subscript p to indicate an observation from the PSID (the *target* data set). We assume that x and p are two random samples drawn from the same underlying population. Consider the following demand for food equation in the CEX:

$$\tau\left(f_{i,x}\right) = D'_{i,x}\beta + \gamma\eta\left(c_{i,x}\right) + e_{i,x} \tag{1}$$

where f is food expenditure (available in both surveys x and p), D contains prices and a set of conditioning variables (also available in both data sets), c is total non-durable expenditure (available only in the input data set x), and e captures unobserved heterogeneity in the demand for food (including measurement error in food expenditure). The functions $\tau\left(.\right)$ and $\eta\left(.\right)$ are known monotonic increasing transformations of their arguments. For example, if $\tau\left(f\right) = \log f$ and $\eta\left(c\right) = \log c$, we have the traditional logarithmic demand function model.[4] For the remainder of the paper, we will make the (innocuous) assumption that food is a normal good ($\gamma \geq 0$).

Suppose that estimation of (1) yields estimates $\widehat{\beta}$ and $\widehat{\gamma}$. Define *imputed* consumption in the CEX by inverting assuming $\widehat{\gamma} \neq 0$:

$$\widehat{c}_{i,x} = \eta^{-1}\left(\frac{\tau\left(f_{i,x}\right) - D'_{i,x}\widehat{\beta}}{\widehat{\gamma}}\right)$$

The corresponding *imputed* measure of consumption in the PSID is similarly defined as

$$\widehat{c}_{i,p} = \eta^{-1}\left(\frac{\tau\left(f_{i,p}\right) - D'_{i,p}\widehat{\beta}}{\widehat{\gamma}}\right)$$

To understand under which conditions moments of imputed PSID consumption mirror those of "true" consumption, note that we are confronted with a (non-standard) measurement error problem of the form:[5]

[4]Note that, while extremely popular among applied microeconomists, an AIDS specification does not restrict the food budget share (f/c) to be a monotone function of total expenditure c. Nonetheless, as an empirical fact, food budget share is almost certainly monotone across ranges of c typically encountered which would be enough to justify application of our procedure.

[5]This is a non-standard characterization because there will be a drift driven by the observable characteristics D and a non-unitary slope for the true consumption value c if the estimates $\widehat{\beta}$ and $\widehat{\gamma}$ are inconsistent.

$$\eta\left(\widehat{c}_{i,x}\right) = D'_{i,x}\frac{\left(\beta - \widehat{\beta}\right)}{\widehat{\gamma}} + \frac{\gamma}{\widehat{\gamma}}\eta\left(c_{i,x}\right) + v_{i,x} \tag{2}$$

with $v_{i,x} = \frac{e_{i,x}}{\widehat{\gamma}}$.

From now on we will consider for simplicity the univariate regression case with $\eta(c) = c$ and $\tau(f) = f$. We will discuss more general cases later. Thus in this case, the demand for food equation is

$$f_{i,x} = \beta + \gamma c_{i,x} + e_{i,x} \tag{3}$$

and the measurement error representation is:

$$\widehat{c}_{i,x} = \frac{\left(\beta - \widehat{\beta}\right)}{\widehat{\gamma}} + \frac{\gamma}{\widehat{\gamma}}c_{i,x} + v_{i,x} \tag{4}$$

Let us define $M(a) = \frac{\sum_{i=1}^{N} a_i}{N}$and $V(a) = \frac{\sum_{i=1}^{N}(a_i - M(a))^2}{N}$, the sample cross-sectional mean and variance of the variable a. Let us also define $Cov(a,b) = \frac{\sum_{i=1}^{N}(a_i - M(a))(b_i - M(b))}{N}$ the sample cross-sectional covariance of the variables a and b.We will now consider conditions under which the mean and variance of imputed PSID consumption (obtained from the demand estimation procedure described above) converge to the true but unknown population mean and variance of consumption. We focus on the first two moments because most of the interest of applied economists is precisely on these. For example, the empirical analyses of Bernheim, Skinner and Weinberg (2001) and Palumbo (1999) require that the mean of imputed consumption converges to the true population mean, while other studies are more interested in the performance of the second moment of imputed consumption (Blundell, Pistaferri and Preston (2004), Krueger and Perri (2003)).

2.2 Sample Moments

To start with a general case, assume that $c_{i,x}$ is potentially measured with classical error: $c^*_{i,x} = c_{i,x} + u_{i,x}$, and thus rewrite (3) as:

$$f_{i,x} = \beta + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x}$$

Assume also that total expenditure can be potentially endogenous (i.e., $Cov(c_x, e_x) \neq 0$) even in the absence of measurement error. This would be the case if, for example, total expenditure decisions were made jointly with decisions on individual commodities, such as food.

It follows that certain estimates of β and γ in (3) may be inconsistent. Let $\widehat{\gamma}(y) = \frac{Cov(f_x, y_x)}{Cov(c_x^*, y_x)}$ be an estimator of γ. The choice of instrument y is crucial. If $y_x = c_x^*$, then we have the traditional OLS estimator, while $y_x = z_x \neq c_x^*$ defines an (exactly identified) IV estimator. Note that:

$$\text{p}\lim\widehat{\gamma}(y) = \gamma + B^e(y) + B^m(y)$$

where $B^e(y) = \frac{\text{plim } Cov(e_x, y_x)}{\text{plim } Cov(c_x^*, y_x)}$ is the "endogeneity bias", and $B^m(y) = -\gamma\frac{\text{plim } Cov(u_x, y_x)}{\text{plim } Cov(c_x^*, y_x)}$ is the "measurement error bias". If z_x is a valid instrument, i.e., if it satisfies the restrictions $\text{plim}Cov(c_x^*, z_x) \neq 0$, $\text{plim}Cov(e_x, z_x) = 0$ and $\text{plim}Cov(u_x, z_x) = 0$, then $B^e(z) = B^m(z) = 0$ and $\text{plim}\widehat{\gamma}(z) = \gamma$. In the OLS case, in contrast, $y_x = c_x^*$ and generally $\text{plim}\widehat{\gamma}(c^*) \neq \gamma$, although the sign of the bias is ambiguous.[6] As for the intercept, it is easy to show that $\text{plim}\widehat{\beta}(y) = \beta - (B^e(y) + B^m(y))\text{plim } M(c_x)$, and so inconsistency in the slope propagates to the intercept.

Let $\widehat{c}_x(y)$ denote the prediction obtained using estimates $\widehat{\beta}(y)$ and $\widehat{\gamma}(y)$. It can be shown that:

$$\text{plim } M(\widehat{c}_x(y)) = \text{plim } M(c_x) \tag{5}$$

Thus according to (5) the sample mean of predicted CEX consumption $M(\widehat{c}_x(y))$ converges in probability to the same limit as the sample mean of true consumption, $M(c_x)$, regardless of whether consumption is measured with error or whether it is correlated with heterogeneity in food spending (and thus plim $M(\widehat{c}_x(c^*))$ =plim $M(\widehat{c}_x(z))$, showing that if interest centers on just the mean of consumption, inconsistent estimators work as well as consistent ones).

As for the sample variance of predicted CEX consumption, one can prove that:

$$\text{plim } V(\widehat{c}_x(y)) = \left(\frac{\gamma}{\gamma + B^e(y) + B^m(y)}\right)^2 \left(\text{plim } V(c_x) + \frac{1}{\gamma^2}\text{plim } V(e_x) + \frac{2}{\gamma}\text{ plim } Cov(c_x, e_x)\right) \tag{6}$$

Consider first the case in which γ is consistently estimated (i.e., an IV estimator is available with valid instrument $y_x = z_x$ and thus $B^e(.) = B^m(.) = 0$). In this scenario

$$\text{plim } V(\widehat{c}_x(z)) = \text{plim } V(c_x) + \frac{1}{\gamma^2}\text{plim } V(e_x) + \frac{2}{\gamma}\text{ plim } Cov(c_x, e_x), \tag{7}$$

[6] Of course, the OLS estimator could be consistent even if $B^e(c^*) \neq 0$ and $B^m(c^*) \neq 0$ in the fortuitous case of balancing biases $B^e(c^*) + B^m(c^*) = 0$.

and therefore the sample variance of predicted CEX consumption converges in probability to the same limit as the variance of true consumption, $V(c_x)$, up to an additive term, $\frac{1}{\gamma^2}\text{p}\lim V(e_x) + \frac{2}{\gamma}\text{p}\lim Cov(c_x, e_x)$. This term generally decreases with the value of the budget elasticity γ.[7] If the demand for food is relatively inelastic ($\gamma \to 0$) this additive term may become potentially quite large. Thus $V(\widehat{c}_x(z))$ is an inconsistent estimator of the variance of true consumption. However, if the asymptotic bias is stationary, the growth of $V(\widehat{c}_x(z))$ is a consistent estimator for the growth of $V(c_x)$ (i.e., $V(\widehat{c}_x(z))$ can be used to understand how $V(c_x)$ trends over time; the two measures will move in lock-step and will differ only by a constant term).

Consider now the case $y_x = c_x^*$ corresponding to the OLS estimator. In general, both $B^e(c^*) \neq 0$ and $B^m(c^*) \neq 0$. Consider first the case in which consumption is measured with error but $\text{plim}Cov(c_x, e_x) = 0$. Here $B^e(y) = 0$ and $B^m(y) < 0$, and therefore

$$\text{plim}V(\widehat{c}_x(c^*)) = \left(\frac{\gamma}{\gamma + B^m(c^*)}\right)^2 \left(\text{plim } V(c_x) + \frac{1}{\gamma^2}\text{plim } V(e_x)\right)$$

In this case the variance of predicted consumption is also an inconsistent estimator of the variance of true consumption (the additive asymptotic bias term is higher than in (7), however). What is worse, $V(\widehat{c}_x(c^*))$ will grow more rapidly than $V(c_x)$, and thus one will have the impression that the consumption variance is growing more than it actually is.

Consider next the case in which consumption is free from error, but plim $Cov(c, e) \neq 0$, so that $B^e(y) \neq 0$ and $B^m(y) = 0$. Again, $V(\widehat{c}_x(c^*))$ is an inconsistent estimate of the *level* and the *growth* of the consumption variance (the magnitude of the inconsistency depends, of course, on the sign of $Cov(c_x, e_x)$).

The case in which consumption is measured with error and it is also endogenous is a combination of the two cases just discussed. The main conclusion to be drawn here is that an IV estimator with a valid instrument ensures that the slope of the asymptotic relationship between $V(\widehat{c}_x)$ and $V(c_x)$ is unity. An OLS estimator will not guarantee that. If consumption is measured with error or it is endogenous, the slope of the asymptotic relationship between $V(\widehat{c}_x)$ and $V(c_x)$ is no longer unity, and thus -when plotted against time- $V(\widehat{c}_x)$ and $V(c_x)$ will diverge even if plim $V(e_x)$ and plim $Cov(c_x, e_x)$ were constant over time.

Figure 1 gives a simple example.[8] Call $\lambda = \frac{\gamma}{\gamma + B^e(y) + B^m(y)}$. We plot the variance of consumption $V(c_x)$ and the variance of imputed consumption $V(\widehat{c}_x)$ against time, for three cases of interest:

[7] The term unambiguously decreases with γ if $Cov(c_x, e_x) > 0$.

[8] The figure is derived assuming that $V(c_x)$ grows linearly over time ($\Delta V(c_x)_t = 0.01$), and that $V(e_x) = 0.005$, $Cov(c_x, e_x) = 0.0025$, and $\gamma = 0.8$.

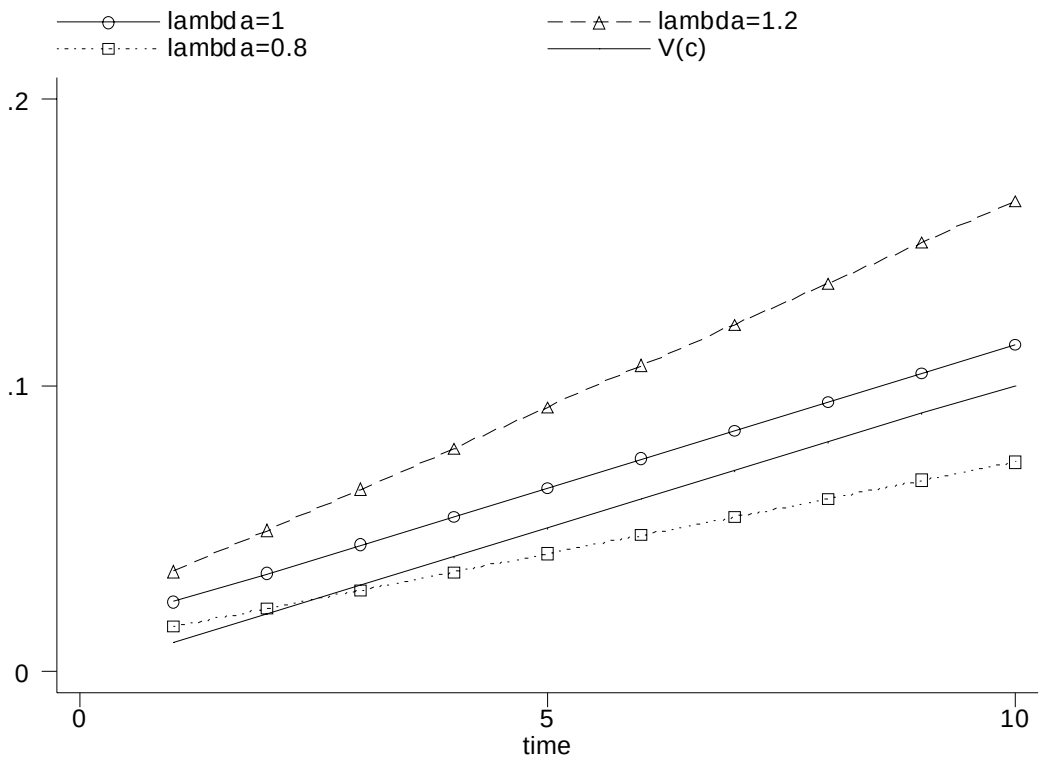

Figure 1: The relationship between $V(\widehat{c}_x)$ and $V(c_x)$ under three different assumptions on λ.

$\lambda = 0.8$ (corresponding to an OLS estimator in a endogenous consumption case, featuring $B^e(y) > 0$ and $B^m(y) = 0$), $\lambda = 1$ (corresponding to an IV estimator with valid instrument which ensures $B^e(y) = B^m(y) = 0$), and $\lambda = 1.2$ (corresponding to an OLS estimator in a measurement error case, featuring $B^e(y) = 0$ and $B^m(y) < 0$). The graph shows that when $\lambda \neq 1$ the two variances $V(c_x^*)$ and $V(\widehat{c}_x)$ grow progressively apart, while in the case $\lambda = 1$, $V(\widehat{c}_x)$ is just an upward translated version of $V(c_x)$ picking up exactly the same time trends.

Consider now the asymptotic behavior of the moments of PSID imputed consumption.[9] This is defined as:

[9]It will be useful to note that:

$$M(\widehat{c}_p(y)) = M(\widehat{c}_x(y)) + \frac{1}{\widehat{\gamma}(y)}[M(f_p) - M(f_x)]$$

and that:

$$V(\widehat{c}_p(y)) = V(\widehat{c}_x(y)) + \frac{1}{\widehat{\gamma}(y)^2}[V(f_p) - V(f_x)]$$

$$\widehat{c}_{i,p}(y) = \frac{f_{i,p} - \widehat{\beta}(y)}{\widehat{\gamma}(y)}$$

and thus one can prove that:

$$\text{plim } M(\widehat{c}_p(y)) = \text{plim } M(c_x) + \frac{1}{\gamma + B^e(y) + B^m(y)} [\text{plim } M(f_p) - \text{plim } M(f_x)] \tag{8}$$

This shows that the sample mean of imputed PSID consumption converges to the same limit as the sample mean of true consumption, up to an additive term. If food consumption is on average the same in the two data sets, the second term on the right hand side vanishes, and the sample mean of imputed PSID consumption converges to the same limit as the sample mean of true consumption. Otherwise, the sample mean of imputed PSID consumption may underestimate or overestimate the sample mean of true consumption depending on whether $(\text{plim } M(f_p) - \text{plim } M(f_x)) \lesseqgtr 0$. If $y_x = z_x$ (and therefore $B^e(.) = B^m(.) = 0$), it is possible to correct for this discrepancy using, e.g., $\frac{M(f_p) - M(f_x)}{\widehat{\gamma}}$ as a "correction factor" in small samples. Note that there are various reasons why $\text{plim } M(f_p) - \text{plim } M(f_x) \neq 0$. One possibility is that the two samples from x and p are not random samples drawn from the same underlying population (even after accounting for differences in stratification and other data collection issues).

Consider now the sample variance of PSID imputed consumption, which has the following expression

$$\text{plim } V(\widehat{c}_p(y)) = \left(\frac{\gamma}{\gamma + B^e(y) + B^m(y)}\right)^2 \left(\begin{array}{c} \text{plim } V(c_x) + \frac{1}{\gamma^2}\text{plim } V(e_x) + \frac{2}{\gamma} \text{ plim } Cov(c_x, e_x) \\ + \frac{1}{\gamma^2}(\text{plim } V(f_p) - \text{plim } V(f_x)) \end{array} \right)$$

This expression is the same as (6), with the exception that there is an additional reason for $V(\widehat{c}_p(y))$ to differ asymptotically from $V(c_x)$, namely the fact that food as measured in the PSID has different variance than that measured in the CEX. Apart from this, the discussion above applies here: the variance of imputed PSID consumption follows the same trend as the variance of CEX consumption provided $B^e(y) = B^m(y) = 0$ (as when parameters are estimated by IV with valid instruments).

2.3 Covariates

Assume now that the food demand equation (allowing for measurement error in $c_{i,x}$) is

$$\begin{aligned} f_{i,x} &= D'_{i,x}\beta + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x} \\ D_{i,x} &= \left(\begin{array}{cccccc} 1 & {}_2d_{i,x} & {}_3d_{i,x} & \dots & {}_{k-1}d_{i,x} \end{array} \right)' \end{aligned}$$

Then we can write the expression of imputed consumption in the CEX as:

$$\widehat{c}_{i,x} = D'_{i,x}\frac{\left(\beta - \widehat{\beta}\right)}{\widehat{\gamma}} + \frac{\gamma}{\widehat{\gamma}}c_{i,x} + v_{i,x}$$

where $v_{i,x} = \frac{e_{i,x}}{\gamma}$ and β is now a vector. To fix ideas, let us consider a simple case with just one covariate $d_{i,x}$:

$$f_{i,x} = \beta_0 + \beta_1 d_{i,x} + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x}$$

It is easy to show that:

$$\begin{aligned} \operatorname{p\lim} \widehat{\beta}_0 (d,y) &= \beta_0 - \left(B^e(d,y) + B^m(d,y)\right)\left(\operatorname{p\lim} M(c) - \rho \operatorname{p\lim} M(d)\right) \\ \operatorname{p\lim} \widehat{\beta}_1 (d,y) &= \beta_1 - \rho\left(B^e(d,y) + B^m(d,y)\right) \\ \operatorname{p\lim} \widehat{\gamma} (d,y) &= \gamma + B^e(d,y) + B^m(d,y) \end{aligned}$$

where $\rho = \frac{p\lim Cov(c^*,d)}{p\lim V(d)}$ is the coefficient of a linear projection of c^* onto d. As before,

$$B^e(d,y) = \frac{Cov(e,y)\,V(d)}{V(d)\,Cov(c^*,y) - Cov(c^*,d)\,Cov(d,y)}$$

denotes the "endogeneity bias", and

$$B^m(d,y) = -\gamma\frac{Cov(u,y)\,V(d)}{V(d)\,Cov(c^*,y) - Cov(c^*,d)\,Cov(d,y)}$$

is the "measurement error bias". Note that $B^e(d,y) = B^e(y)$ and $B^m(d,y) = B^m(y)$ of the previous section if c^* and d are orthogonal, i.e., $\rho = 0$.

The implications of measurement error are complicated by the propagation of possible inconsistency to coefficients of other covariates. However we can pursue similar reasoning to that pursued in the previous sections to generalize the earlier expressions. Specifically[10]

[10]The results below follow from:

$$\widehat{c}_{i,p} = \widehat{c}_{i,x} + \frac{1}{\widehat{\gamma}}\left[(f_{i,p} - f_{i,x}) - \widehat{\beta}_1(d_{i,p} - d_{i,x})\right].$$

$$\text{plim } M\left(\widehat{c}_x\left(d,y\right)\right) = \text{plim } M\left(c_x\right) \tag{9}$$

$$\text{plim } M\left(\widehat{c}_p\left(d,y\right)\right) = \text{plim } M\left(c_x\right) + \frac{1}{\gamma + B^e\left(d,y\right) + B^m\left(d,y\right)} \begin{bmatrix} \text{plim } M\left(f_p - \widehat{\beta}_1\left(d,y\right) d_p\right) \\ -\text{plim } M\left(f_x - \widehat{\beta}_1\left(d,y\right) d_x\right) \end{bmatrix} \tag{10}$$

Hence the convergence of the sample mean of imputed consumption to the population mean in the CEX is assured, as earlier, by the fact that estimates pass through sample means. The probability limit of the imputed sample mean in the PSID may converge to a different value either because of discrepancy in the mean of the input variable (f) or the mean of the covariates (d) in the two data sets. The discrepancy from the first source is greater for less elastic demand ($\gamma \to 0$) and it may be amplified by inconsistency in estimation of the demand function parameters.

As regards the variance, we have in the CEX

$$\text{plim}V\left(\widehat{c}_x\left(d,y\right)\right) = \left(\frac{\gamma}{\gamma + B^e\left(d,y\right) + B^m\left(d,y\right)}\right)^2 \begin{bmatrix} \text{plim}V\left(c_x\right) + \frac{1}{\gamma^2}\text{plim}V\left(e_x\right) \\ +\frac{2}{\gamma}\text{plim}Cov\left(c_x, e_x\right) \\ +\left(\frac{\rho(B^e(d,y)+B^m(d,y))}{\gamma}\right)^2 \text{plim}V\left(d_x\right) \\ +\frac{2\rho(B^e(d,y)+B^m(d,y))}{\gamma}\text{plim}Cov\left(c_x, d_x\right) \end{bmatrix}$$

This expression shows that the asymptotic discrepancy between the sample variance of predicted consumption and the variance of true consumption depends on both consistency in estimating parameters and the extent of the correlation between consumption and the covariates of the food demand equation. Of course, when the coefficients of covariates are consistently estimated (as for example when estimation is by OLS and they are uncorrelated with total consumption, $\rho = 0$) the expression collapses to the one examined in the previous section. Note also that if $y = z$ is a valid instrument, then covariates have no role in determining the asymptotic expression of $V\left(\widehat{c}_x\left(d,y\right)\right)$ regardless of their correlation with total expenditure.

Finally, one can prove that

$$\text{plim}V\left(\widehat{c}_p\left(d,y\right)\right) = \text{plim}V\left(\widehat{c}_x\left(d,y\right)\right) + \left(\frac{1}{\gamma + B^e\left(d,y\right) + B^m\left(d,y\right)}\right)^2 \begin{bmatrix} \text{plim } V\left(f_p - \widehat{\beta}_1\left(y\right) d_p\right) \\ -\text{plim } V\left(f_x - \widehat{\beta}_1\left(y\right) d_x\right) \end{bmatrix} \tag{11}$$

Any difference between the variance of covariates in the two datasets may therefore be a further contribution to the additive bias.

2.4 Budget elasticity heterogeneity

Most demand functions have the property that the parameters vary with observable household characteristics (such as the number of children or education). Suppose that there is one such characteristics, $q_i = \{0, 1\}$, a binary variable. Then one can write the demand function for food in the univariate case as:

$$f_{i,x} = \beta + \gamma c_{i,x} + \delta q_{i,x} c_{i,x} + e_{i,x} = \begin{cases} \beta + \gamma c_{i,x} + e_{i,x} & \text{if } q_{i,x} = 0 \\ \beta + (\gamma + \delta) c_{i,x} + e_{i,x} & \text{if } q_{i,x} = 1 \end{cases}$$

The measurement error interpretation is such that:

$$\widehat{c}_{i,x} = \begin{cases} \widehat{\gamma}^{-1} \left(\left(\beta - \widehat{\beta} \right) + \gamma c_{i,x} + e_{i,x} \right) & \text{if } q_{i,x} = 0 \\ \left(\widehat{\gamma} + \widehat{\delta} \right)^{-1} \left(\left(\beta - \widehat{\beta} \right) + (\gamma + \delta) c_{i,x} + e_{i,x} \right) & \text{if } q_{i,x} = 1 \end{cases}$$

The asymptotic expressions of the moments of imputed consumption are a simple extension of those reported in the previous sections. In particular, assume that there is no estimation inconsistency problems ($B^e(y) = B^m(y) = 0$), so that $\widehat{\beta}$, $\widehat{\gamma}$ and $\widehat{\delta}$ are all consistent. Then one can prove that

$$\text{plim } M(\widehat{c}_x) = \text{plim } M(c_x) - \frac{\delta}{\gamma(\gamma + \delta)} \text{plim } Cov(q_x, e_x)$$

If $q_{i,x}$ is an exogenous characteristic with which one partitions the sample, i.e. $E(e_{i,x}|q_{i,x}) = 0$, then plim $M(\widehat{c}_x)$ =plim $M(c_x)$. As for the variance, one can make the same assumption and derive an extension to (7),

$$\text{plim } V(\widehat{c}_x) = \text{plim } V(c_x) + \left(\eta \frac{1}{\gamma} \right)^2 \text{plim } V(e_x) + 2\eta \frac{1}{\gamma} \text{plim } Cov(c_x, e_x)$$

where the multiplicative factor $\eta = \frac{\gamma - \delta}{\gamma + \delta}$, and so the intercept can be higher or lower than in the case of no budget elasticity heterogeneity (in particular, $\eta \gtreqless 1$ if $\delta \lesseqgtr 0$). Intuitively, when $\delta > 0$ the budget elasticity is higher and this reduces the intercept (vis-à-vis a case where no interactions are allowed). In deriving this expression we have assumed $E\left(e_{i,x}^2 | q_{i,x} = 1\right) = E\left(e_{i,x}^2 | q_{i,x} = 0\right)$ and $E(c_{i,x} e_{i,x} | q_{i,x} = 1) = E(c_{i,x} e_{i,x} | q_{i,x} = 0)$. The expressions for plim $M(\widehat{c}_p)$ and plim $V(\widehat{c}_p)$ are simple extensions of those derived in the previous sections. Also extensions are the relevant expressions when the demand function admits covariates.

2.5 Non-linear case

The final extension we consider is when we have the general non-linear form:

$$\tau\left(f_{i,x}\right)=\beta+\gamma\eta\left(c_{i,x}\right)+e_{i,x}$$

This does not pose any new problem if what we are interested in are moments of $\eta\left(c_{i,x}\right)$. In fact, the predicted value is in this case

$$\eta\left(\widehat{c}_{i,x}\right)=\frac{\tau\left(f_{i,x}\right)-\widehat{\beta}}{\widehat{\gamma}}$$

and therefore one can simply interpret the expressions derived above for the moments of $\widehat{c}_x$ (or $\widehat{c}_p$) as representing those for the moments of $\eta\left(\widehat{c}_x\right)$ (or $\eta\left(\widehat{c}_p\right)$).

More generally, the expressions derived in the previous sections give us

$$\begin{aligned}\text{plim } M\left(\eta\left(\widehat{c}_p\right)\right) &= \theta_0+\theta_1\text{plim } M\left(\eta\left(c_x\right)\right)\\ \text{plim } V\left(\eta\left(\widehat{c}_p\right)\right) &= \phi_0+\phi_1\text{plim } V\left(\eta\left(c_x\right)\right)\end{aligned}$$

Consider the case when we are interested in moments of a function of consumption that is different than $\eta\left(c\right)$. In such a case we can make use of Taylor expansions. We will limit here to the most relevant case, namely the one that occurs when we are interested in moments of $\widehat{c}$ and our estimated demand equation involves $\eta\left(c\right)=\log c$. In this case, all we have to do is to use the approximations:

$$M\left(\widehat{c}\right) \simeq e^{M(\log\widehat{c})+\frac{1}{2}V(\log\widehat{c})} \tag{12}$$

$$V\left(\widehat{c}\right) \simeq e^{2M(\log\widehat{c})+2V(\log\widehat{c})}-e^{2M(\log\widehat{c})+V(\log\widehat{c})} \tag{13}$$

to derive all the asymptotic expressions of interest.

3 The data

Our empirical analysis is conducted on two microeconomic data sources: the 1978-1992 PSID and the 1980-1992 CEX. We describe their main features and our sample selection procedures in turn. Blundell, Pistaferri and Preston (2004) is a companion paper that uses the technique described in this paper to study consumption inequality and partial insurance of income shocks.

3.1 The PSID

Since the PSID has been widely used for microeconometric research, we shall only sketch the description of its structure in this section.[11]

The PSID started in 1968 collecting information on a sample of roughly 5,000 households. Of these, about 3,000 were representative of the US population as a whole (the core sample), and about 2,000 were low-income families (the Census Bureau's Survey of Economic Opportunities, or SEO sample). Thereafter, both the original families and their split-offs (children of the original family forming a family of their own) have been followed.

The PSID includes a variety of socio-economic characteristics of the household, including age, education, labor supply, and income of household members. Questions referring to income and wages are retrospective; thus, those asked in 1993, say, refer to the 1992 calendar year. In contrast, many researchers have argued that the timing of the survey questions on food expenditure is much less clear [see Hall and Mishkin, 1982, and Altonji and Siow, 1987, for two alternative views]. Typically, the PSID asks how much is spent on food in an average week. Since interviews are usually conducted around March, it has been argued that people report their food expenditure for an average week around that period, rather than for the previous calendar year as is the case for family income. We assume that food expenditure reported in survey year t refers to the previous calendar year.

The hourly wage measure used below is given by the ratio of annual earnings and annual hours and is deflated using the CPI (1982-84). Education level is computed using the PSID variable "grades of school finished". Individuals who changed their education level during the sample period are allocated to the highest grade achieved.

Since CEX data are available on a consistent basis since 1980, we construct an unbalanced PSID panel using data from 1980 to 1992. Due to attrition, changes in family composition, and various other reasons, household heads in the 1980-1992 PSID may be present from a minimum of one year to a maximum of 13 years. We thus create unbalanced panel data sets of various length. The longest panel includes individuals present from 1980 to 1992; the shortest, individuals present for two consecutive years only (1980-81, 1981-82, up to 1991-92).

The objective of our sample selection is to focus on a sample of continuously married couples headed by a male (with or without children). The step-by-step selection of our PSID sample is

[11] See Hill [1992] for more details about the PSID.

illustrated in Table I. We eliminate households facing some dramatic family composition change over the sample period. In particular, we keep only those with no change, and those experiencing changes in members other than the head or the wife (children leaving parental home, say). We next eliminate households headed by a female. We also eliminate households with missing report on education and region,[12] and those with topcoded income. We keep continuously married couples and drop some income outliers.[13] We then drop those born before 1920 or after 1959.

As noted above, the initial 1967 PSID contains two groups of households. The first is representative of the US population (61 percent of the original sample); the second is a supplementary low income subsample (also known as SEO subsample, representing 39 percent of the original 1967 sample). To account for the changing demographic structure of the US population, starting in 1990 a representative national sample of 2,000 Latino households has been added to the PSID database. We exclude both Latino and SEO households and their split-offs. Finally, we drop those aged less than 30 or more than 65. This is to avoid problems related to changes in family composition and education, in the first case, and retirement, in the second. The final sample used in the exercise below is composed of 17,788 observations and 1,788 households.

3.2 The CEX

The Consumer Expenditure Survey provides a continuous and comprehensive flow of data on the buying habits of American consumers. The data are collected by the Bureau of Labor Statistics and used primarily for revising the CPI. Consumer units are defined as members of a household related by blood, marriage, adoption, or other legal arrangement, single person living alone or sharing a household with others, or two or more persons living together who are financially dependent. The definition of the head of the household in the CEX is the person or one of the persons who owns or rents the unit; this definition is slightly different from the one adopted in the PSID, where the head is always the husband in a couple. We make the two definitions compatible.

The CEX is based on two components, the Diary, or record keeping survey and the Interview survey. The Diary sample interviews households for two consecutive weeks, and it is designed to obtain detailed expenditures data on small and frequently purchased items, such as food, personal care, and household supplies. The Interview sample follows survey households for a maximum of 5 quarters, although only inventory and basic sample data are collected in the first quarter. The

[12] When possible, we impute values for education and region of residence using adjacent records on these variables.

[13] An income outlier is defined as a household with an income growth above 500 percent, below −80 percent, or with a level of income below $100 a year or below the amount spent on food.

data base covers about 95% of all expenditure, with the exclusion of expenditures for housekeeping supplies, personal care products, and non-prescription drugs. Following most previous research, our analysis below uses only the Interview sample.

The CEX collects information on a variety of socio-demographic variables, including characteristics of members, characteristics of housing unit, geographic information, inventory of household appliances, work experience and earnings of members, unearned income, taxes, and other receipts of consumer unit, credit balances, assets and liabilities, occupational expenses and cash contributions of consumer unit. Expenditure is reported in each quarter and refers to the previous quarter; income is reported in the second and fifth interview (with some exceptions), and refers to the previous twelve months. For consistency with the timing of consumption, fifth-quarter income data are used.

We select a CEX sample that can be made comparable, to the extent that this is possible, to the PSID sample. Our initial 1980-1998 CEX sample includes 1,249,329 monthly observations, corresponding to 141,289 households. We drop those with missing record on food and/or zero total nondurable expenditure, and those who completed less than 12 month interviews. This is to obtain a sample where a measure of annual consumption can be obtained. A problem is that many households report their consumption for overlapping years, i.e. there are people interviewed partly in year t and partly in year $t+1$. Pragmatically, we assume that if the household is interviewed for at least 6 months at $t+1$, then the reference year is $t+1$, and it is t otherwise. Prices are adjusted accordingly. We then sum food at home, food away from home and other nondurable expenditure over the 12 interview months. This gives annual expenditures. For consistency with the timing of the PSID data, we drop households interviewed after 1992. We also drop those with zero before-tax income, those with missing region or education records, single households and those with changes in family composition. Finally, we eliminate households where the head is born before 1920 or after 1959, those aged less than 30 or more than 65, those with outlier income (defined as a level of income below the amount spent on food), and those with incomplete income responses. Our final sample contains 15,137 households. Table II details the sample selection process in the CEX.

The definition of total non durable consumption is similar to Attanasio and Weber [1995]. It includes food (at home and away from home), alcoholic beverages and tobacco, services, heating fuel, transports (including gasoline), personal care, clothing and footwear, and rents. It excludes expenditure on various durables, housing (furniture, appliances, etc.), health, and education.

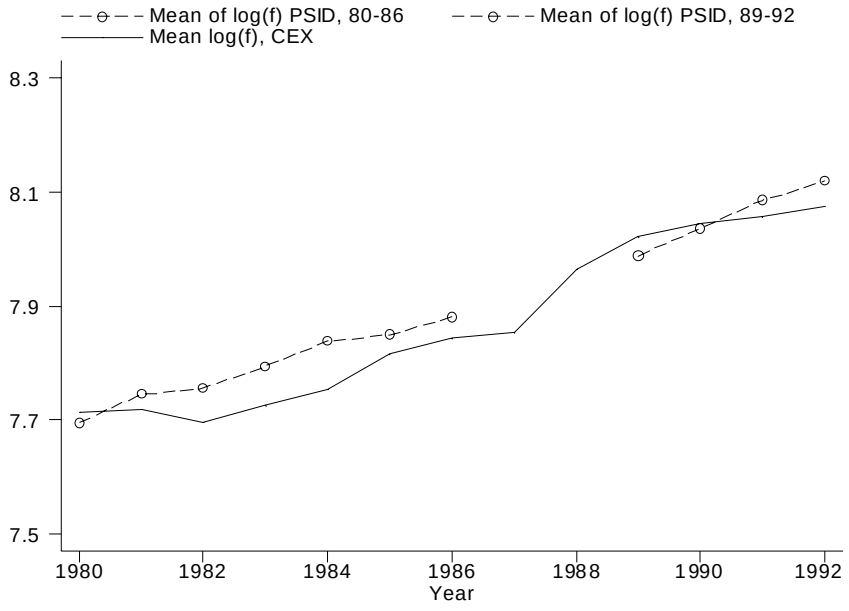

Figure 2: Trends in food expenditure, PSID and CEX.

3.3 Comparing the two data sets

How similar are the two data sets in terms of average demographic and socio-economic characteristics? This is important as differences in the underlying demographics may explain part of the discrepancy between PSID-imputed moments and CEX moments. Mean comparisons are reported in Table III for selected years: 1980, 1983, 1986, 1989, and 1992.

PSID respondents are slightly younger than their CEX counterparts; there is, however, little difference in terms of family size and composition. The percentage of whites is slightly higher in the PSID. The distribution of the sample by schooling levels is quite similar, while the PSID tends to under-represent the proportion of people living in the West.

Trends in food expenditure (the sum of food at home and food away from home) are quite similar across the two data sets (the mean difference is hardly greater than \$200), but some differences emerge. See Figure 2.[14] For example, the mean of food expenditure is higher in the PSID than in the CEX in the early 1980s. As we shall see, accounting for differences in the way the food question is asked in the two data sets helps replicating the trends in average consumption.

In Table IV we present the results of three tests. First, we test that the cross-dataset difference

[14]The definition of consumption that we use in the figures is per adult equivalent, $x = c/\sqrt{n}$, where n is family size.

$M(d_p) - M(d_x)$ is constant over time.[15] This difference appears on the right hand side of (10), along with $M(f_p) - M(f_x)$, as a shifter of the (asymptotic) relationship between $M(\widehat{c}_p)$ and $M(c_x)$. Second, we test that the cross-data set differences $V(d_p) - V(d_x)$ and $Cov(d_p, f_p) - Cov(d_x, f_x)$ are constant over time. These two terms appear (in their asymptotic representation) in an expansion of the right hand side of (11), along with $V(f_p) - V(f_x)$. If all these terms are constant over time, then the slope of the relationship between $V(\widehat{c}_p)$ and $V(c_x)$ is unaffected.

The table shows that we almost never reject the null hypotheses. The only exception is for the "Born in 1955-59" dummy. Given our sample selection (we restrict attention to people aged 30-65), people born in 1955-59 appear in our sample only after 1985. This may be problematic in the PSID, where - given the absence of sample "refreshing"- very young households may appear only because of the split-off mechanism described above (indeed, the test passes barely for "Born in 195-54"). For this reason, we do not think this rejection tells us that there is something fundamentally different between the two data sets.

More interestingly, we note that the test that $M(f_p) - M(f_x)$ is constant over time is rejected. This was apparent from figure 2 as well. It implies that we will be able to reconcile cross-dataset differences in consumption only by accounting for differences in the way the food question is asked in the two data sets. Also interestingly, we note that in contrast we cannot reject the null that the difference $V(f_p) - V(f_x)$ is constant over time.

4 Empirical Results

After some experimentation, we selected a loglinear functional form. The main advantage of the loglinear demand function is that it provides "ready-to-use" predictions for total nondurable expenditure, avoiding, for instance, the problem of negative predicted values faced when using the linear expenditure demand function. The loglinear demand function has also a series of shortcomings, however. In particular, it cannot capture zero expenditures, it does not satisfy adding up if applied to all goods in a demand system, and it does not capture apparent non-linearities in Engel curve relationships. Nevertheless, these shortcomings do not appear particularly relevant here. There are no zeros in food spending, the specification below is applied to just one good, and the Engel curve

[15] The test is constructed as follows. We start by pooling CEX and PSID data, and generate a dummy for whether an observation comes from the CEX. To test that $M(d_p) - M(d_x)$ is constant over time we regress d_i (the covariate of interest: age, family size, etc.) on the CEX dummy, year dummies, and the interaction of the CEX dummy with year dummies. A test that $M(d_p) - M(d_x)$ is constant over time is a test that the coefficients on the interactions are all the same. We follow a similar procedure to test that $V(d_p) - V(d_x)$ and $Cov(d_p, f_p) - Cov(d_x, f_x)$ are constant over time.

for food is not far from being log linear.[16]

We sum food at home and food away from home to obtain total food expenditure f. To estimate the demand function for food, we pool all the CEX data from 1980 to 1992. Our specification includes the log of the price of food, the log of the price of alcohol, the log of the price of fuel and utilities, and the log of the price of transport.[17] Our specification also includes the log of total nondurable expenditure and its interaction with education dummies, indicators for number of children (no children, one child, two or three children, four children or more), and year dummies.[18] Finally, we include a vector of demographics meant to capture heterogeneity in the demand for food: a quadratic in age, dummies for education, region of residence, year of birth dummies, indicators for number of children (as above), family size, and a dummy for whites.[19]

We estimate the demand equation for food by an instrumental variables (IV) procedure. As argued above, using an IV procedure is important if one wants to eliminate the bias induced by measurement error in consumption expenditure. The IV estimation procedure uses 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 as instruments for consumption expenditure (and interactions with year, children dummies, and education dummies as instruments for the interaction of consumption with the same variables). The IV estimates of the food demand equation are reported in Table V. The budget elasticity is 0.88 (0.81 in the OLS case). The price elasticity is -0.96. The prices of other goods are very imprecisely measured. We test the overidentifying restrictions and fail to reject the null hypothesis (p-value of 56 percent). We also report statistics for judging the power of excluded instruments. They are all acceptable. Most of the demographics have the expected sign.

Armed with the estimated demand parameters, we invert the demand equation for food and obtain a measure of total nondurable expenditure in the PSID matching on observable characteristics that are common to the two data sets. As explained previously, a good inversion procedure should

[16]See Lehnert (2002). He uses our procedure to test the monotonicity assumption. While he formally rejects it, he also shows that rejection is almost entirely accounted for by the top part of the consumption distribution and so little bias is to be expected from imposing monotonicity over the relevant part of the distribution.

[17]These prices are indexes obtained from the BLS. The indexes are US city averages, not seasonally adjusted, and are expressed in 1982-84 dollars. We omit prices of other commodities due to multicollinearity problems.

[18]The interactions with year dummies attempt to capture the fact that as income grows (over time), the food budget elasticity declines. This is exactly the kind of pattern we find in the data.

[19]There are, of course, alternative ways of using the information on the two food components for the purpose of predicting total expenditure. One could, for example, estimate a demand system for food at home and food away from home imposing theoretical restrictions, such as symmetry. One could also estimate a demand function just for food at home (conditioning on food away from home or on its price). The strategy presented in the text is the most successful one in terms of predictive power, at least for this sample and for this period.

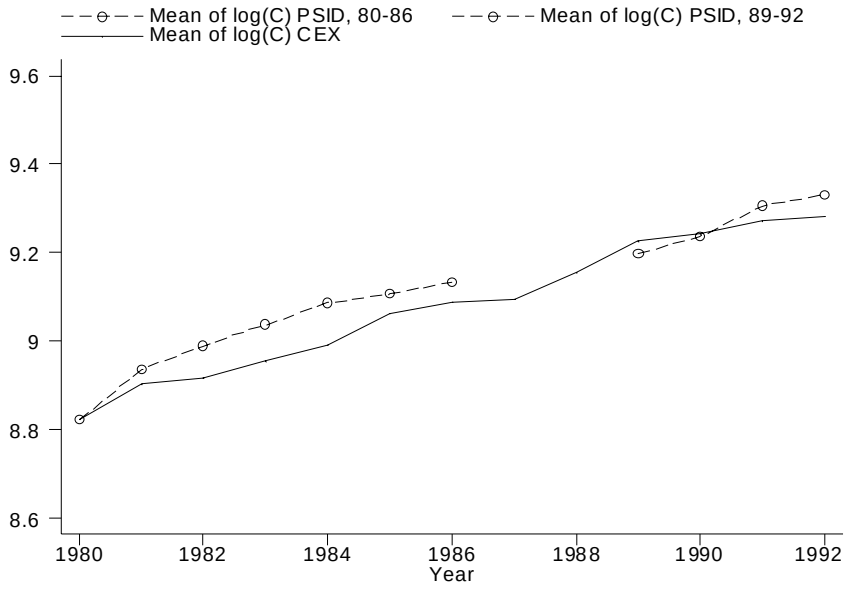

Figure 3: Imputed PSID consumption and CEX consumption.

have two defining properties: (a) average (imputed) consumption in the PSID should coincide with average consumption in the CEX, and (b) the variance of (imputed) consumption in the PSID should exceed the variance of consumption in the CEX by an additive factor (the variance of the error term of the demand equation scaled by the square of the expenditure elasticity). If this factor is constant over time the *trends* in the two variances should be identical.

Figure 3 shows that average consumption differs quite substantially in the first half of the 1980s.

One can intuitively reason on the origin of the discrepancy by looking at the expressions (10). If one is prepared to accept that plim $M(d_p) - \text{plim } M(d_x) = 0$ as the results in Table IV would suggest, and that $B^e(y) = B^m(y) = 0$ if our IV procedure has eliminated the bias induced by endogeneity and/or measurement error in consumption, the difference between $M(\widehat{c}_p)$ and $M(c_x)$ can only be attributed to cross-dataset differences in the mean of food spending (a difference that has an exacerbating effect because $\gamma < 1$). This is a plausible explanation given the evidence on $M(f_p) - M(f_x)$ presented in Table IV and evident from figure 2.

In figure 4 we plot $(M(\widehat{c}_p) - M(c_x))$ (the continuous line) and $\frac{M(f_p)-M(f_x)}{\widehat{\gamma}}$ (the dashed line) against time. If our reasoning is correct, the difference between the two should be close to zero. This is indeed the case.

As a further descriptive characterization, figure 5 plots $M(c_x)$ and the "corrected" PSID mean

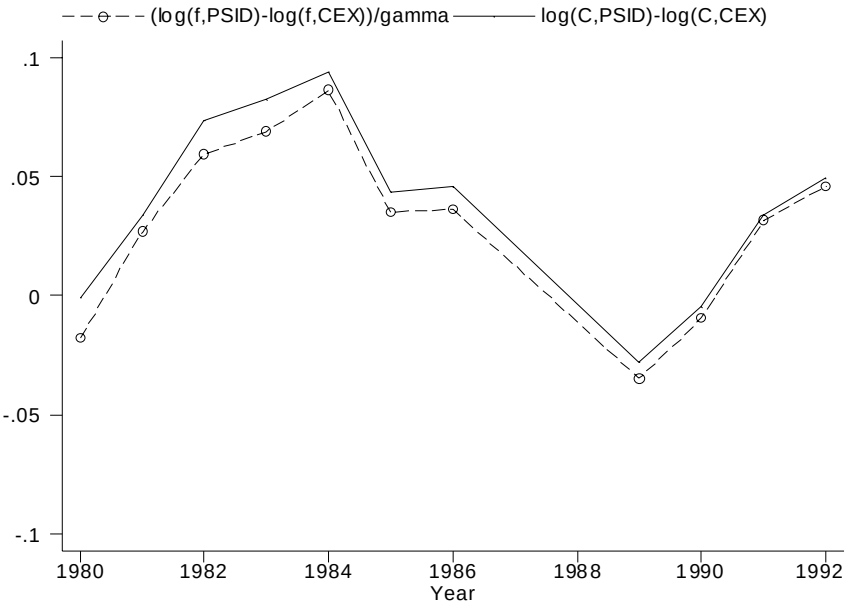

Figure 4: Plotting $(M(\widehat{c}_p) - M(c_x))$ and $\frac{M(f_p)-M(f_x)}{\widehat{\gamma}}$ against time.

$\left(M(\widehat{c}_p) - \frac{(M(f_p)-M(f_x))}{\widehat{\gamma}}\right)$. The two series are now indistinguishable: this proves that our imputation procedure is capable of replicating quite well trends in mean spending once account is made for differences in the mean of the input variable (food spending) in the two data sets.

What about the variance of consumption? Trends in the variance are interesting in their own right, and they are plotted in Figure 6. Note that the range of variation of $V(\widehat{c}_p)$ is on the left-hand side; that of $V(c_x)$, on the right hand side. Trends in the variance of consumption are remarkably similar in the two data sets (note that $V(\widehat{c}_p)$ and $V(c_x)$ have different ranges of variation, but similar scale). In fact $V(\widehat{c}_p)$ appears to be just an upward-translated version of $V(c_x)$, as the theoretical Figure 1 surmised. Both series suggest that in the first half of the 1980s the variance of consumption grows quite substantially. Afterwards, the variance is flat. The levels differ by a common factor as expected if the imputation procedure is reliable.

To see this, use the expression (11) derived above. Assuming from Table IV that plim $V(d_p)$ =plim $V(d_x)$, plim $Cov(f_p, d_p)$ =plim $Cov(f_x, d_x)$, and plim $V(f_p)$ =plim $V(f_x)$, and that $B^e(y) = B^m(y) = 0$ from our IV procedure, one should find that $V(\widehat{c}_p)$ follows the same trend over time as $V(c_x)$ up to the term $\frac{1}{\gamma^2}\text{plim}V(e_x) + \frac{2}{\gamma}\text{plim}Cov(c_x, e_x)$. If this term is time-stationary (which requires that there is no time trend in the dispersion of food spending heterogeneity), then the two series should be moving at the same rate and pace. This is precisely what the graph is showing.

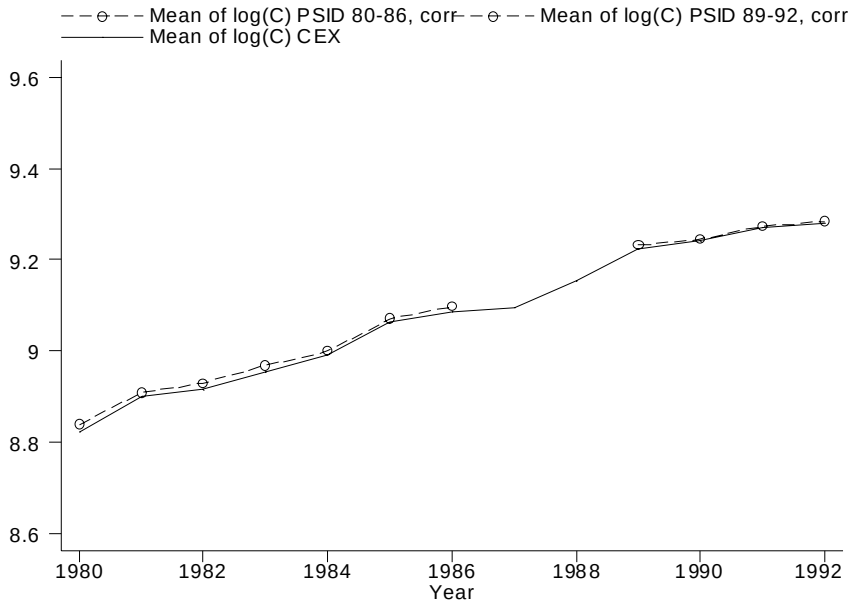

Figure 5: The mean of CEX consumption and the mean of "corrected" PSID imputed consumption.

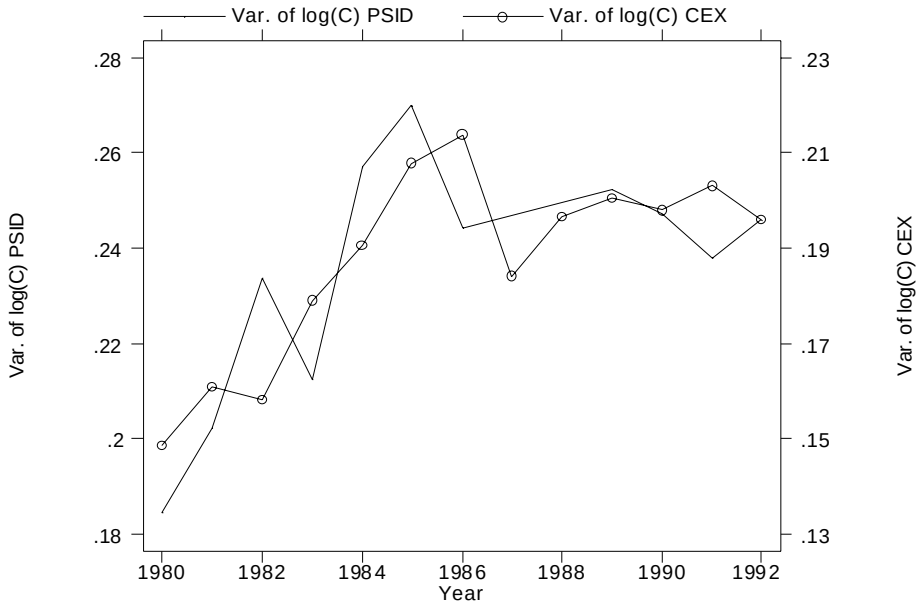

Figure 6: The variance of PSID-imputed consumption and the variance of CEX consumption.

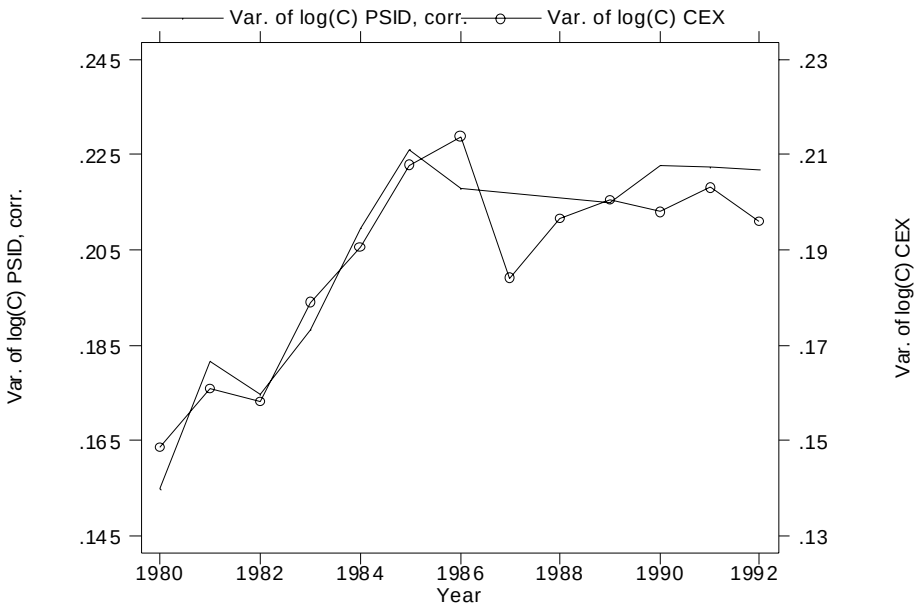

Figure 7: The variance of CEX consumption and the variance of PSID-imputed consumption (corrected for cross-dataset differences in food spending).

One can also use $V(f_p) - V(f_x)$ to reduce the vertical distance between the two series. This also makes treatment of the different moments look more symmetric. The results of correcting the difference $V(\widehat{c}_p) - V(c_x)$ by the correction factor $\frac{V(f_p)-V(f_x)}{\widehat{\gamma}^2}$ are shown in figure 7. The PSID series is now much smoother.

5 Comparison with other approaches

As mentioned in the Introduction, there are many consumption imputation methodologies that have been proposed in the literature, some based on using multiple data sets (as ours), and some based on using extra information contained in the survey of interest. Skinner (1987) is an example of the first type. He uses CEX data for 1972-73 and 1983 to regress total consumption on food at home, food away from home, market value of the home if homeowner, rent, utilities, and the number of vehicles owned. The estimated coefficients are used to impute total consumption in the PSID. Fisher and Johnson (2004) use Skinner's approach but enrich it by adding a wealth of demographic characteristics available in both data sets. Palumbo (1999) compares three different imputation methods: that originally suggested by Skinner, a flexible Engel curve adaption of it, and one based on a structural model of household expenditure (based on unpublished AIDS estimates for food

provided by Attanasio and Weber and obtained from CEX data). Ziliak (1998) and Browning and Leth-Petersen (2003) are examples of the second type of approach. They propose imputing consumption on the basis of income and the first difference of wealth (i.e., as the difference between income and savings). In this Section, we compare our procedure with Skinner's.

The main problem with replicating Skinner's procedure over the sample period of interest of our data (1980-1992) is that collection of key consumption items in the PSID was discontinued in the mid-1980s. For example, utility payments are missing after 1986, the number of vehicles owned is missing after 1985, and actual rent data is missing in 1987-88. To this, one needs to add the fact that both our procedure and Skinner's procedure must deal with the absence of information on food (at home and away from home) in 1987-88. This means that the "extended" Skinner's approach can only be implemented for 1980-1985. Our procedure's advantage is that it is data-consistent throughout the sample period. In what follows, we compare our methodology with both an "extended" and a "reduced" Skinner's methodology. The "extended" Skinner specification is one in which total consumption is regressed onto expenditure on food at home, food away from home, utilities, rent payments (set to zero for those who own), value of the house (set to zero for those who rent), and number of cars owned (0, 1, and 2+). Regressions are run separately for each year. The "reduced" specification omits expenditure on utilities and the number of cars owned.[20]

Figure 8 shows that the variance of imputed PSID log consumption obtained using the "extended" Skinner's specification matches reasonably well the trends of the variance of "true" log consumption in the CEX for the 1980-85 period. Figure 9 shows the results using the "reduced" Skinner's specification. Again, the match is reasonable. In both cases, there is a level effect: the variance of imputed PSID log consumption is lower that the variance of "true" log consumption in the CEX. Figure 10 uses again the "reduced" Skinner's specification and shows that the Skinner's specification matches the trends in mean log consumption reasonably well (again, not the level: there is a somewhat constant downward bias in the Skinner's imputed measure). Correction factors could be constructed in the same way as we constructed them for our methodology. However, the point to be made here is not to show the performance of the Skinner's approach in absolute terms, but rather in a comparative sense, i.e., to show that we have a coherent approach which does at least as well (if not better) than Skinner's.

[20] Despite slightly different sample selections, our results are comparable to those obtained by Skinner (1987, see his Table 1). For example, using data from the 1983 CEX, we find, in the "extended" specification, regression coefficient estimates of 1.47 for food at home (as opposed to Skinner's 1.54), 2.54 for food away from home (3.02), 0.02 for the value of the house (0.09), 1.49 for utilities (1.95), 1.24 for rent (1.31), and 995 for the number of vehicles owned (1339). The R^2 of our regression is 0.76 (as opposed to Skinner's 0.73).

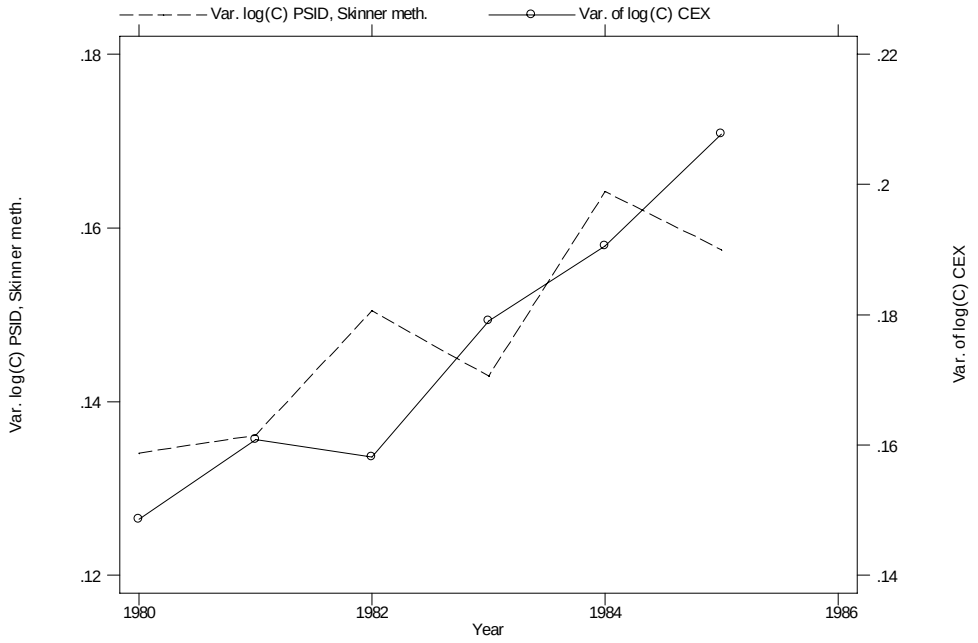

Figure 8: The variance of imputed log consumption in the PSID and the variance of log consumption in the CEX using the "extended" Skinner specification, 1980-1985.

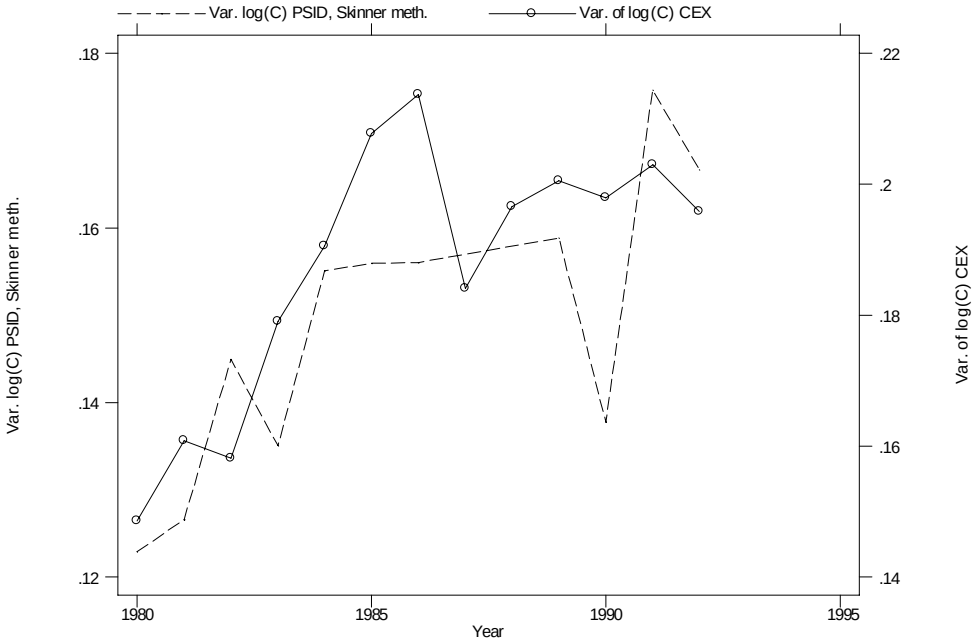

Figure 9: The variance of imputed log consumption in the PSID and the variance of log consumption in the CEX using the "reduced" Skinner specification, 1980-1992.

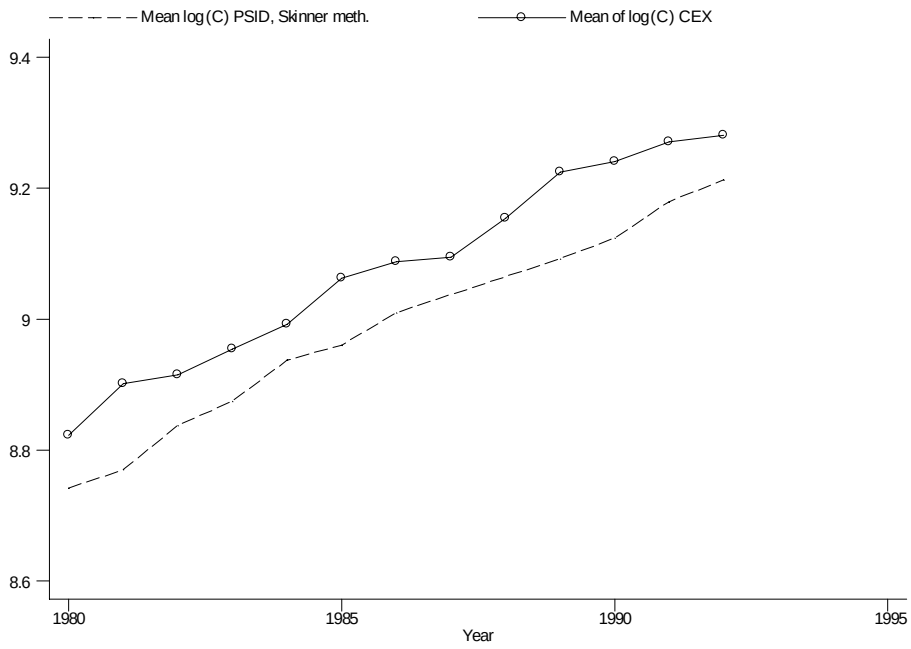

Figure 10: The mean of imputed log consumption in the PSID and the variance of log consumption in the CEX using the "reduced" Skinner specification, 1980-1992.

6 Summary

The major contribution of this paper has been to develop a regular demand equation framework in order to combine panel data on income from the PSID with consumption data from repeated CEX cross-sections. This allowed for important heterogeneity in the income elasticities. Our analysis shows show that the success of an imputation procedure (in the sense of matching the first two moments of the consumption distribution (actual consumption in the input data set, and imputed consumption in the target data set) depends on two points: (a) control is made for initial differences in the input and target data sets, (b) an instrumental variable procedure is used to avoid the bias induced by measurement error in consumption and possible endogeneity of consumption.

Our approach has been to estimate the demand equation for food by an instrumental variables procedure and to use this as a basis for imputation. This was shown to produce reliable results even if there is bias induced by measurement error in consumption expenditure. Our imputation procedure was shown to be capable of replicating the trends in mean spending once account is made for differences in the mean of the input variable (food spending) in the two data sets. We also showed that our procedure is likely to work well for estimating second moments. For example,

the trends in the variance in the CEX and that from the imputed data from the PSID were remarkably similar in the two data sets. The levels differed by a common factor as expected if the imputation procedure is reliable.

References

[1] Alessie, Rob and Annamaria Lusardi (1997), "Saving and Income Smoothing: Evidence from Panel Data", *European Economic Review*, 41(7), 1251-79.

[2] Altonji, Joseph G. (1986), "Intertemporal Substitution in Labor Supply: Evidence from Micro Data", *Journal of Political Economy*, Part 2, 94(3), S176-S215.

[3] Altonji, J., and A. Siow (1987), "Testing the response of consumption to income changes with (noisy) panel data", *Quarterly Journal of Economics*, 102, 293-28.

[4] Altug, Sumru and Robert A. Miller (1990), "Household Choices in Equilibrium", *Econometrica*, 58(3), 543-70.

[5] Attanasio, Orazio P., and Guglielmo Weber (1993), "Consumption Growth, the Interest Rate and Aggregation", *Review of Economic Studies*, 60(3), 631-49.

[6] Banks, J., Blundell, R. and A. Lewbel (1999), "Quadratic Engel curves and consumer demand", *Review of Economics and Statistics*, 79, 527-39.

[7] Battistin, E. (2003), "Errors in survey reports of consumption expenditures", IFS Working Paper 03/07.

[8] Bernheim, D., J. Skinner, and S. Weinberg (2002), "What accounts for the variation in retirement wealth among U.S. households?", *American Economic Review*, 91, 832-57.

[9] Blundell, R. Luigi Pistaferri and Ian Preston (2004), "Consumption inequality and partial insurance", IFS Working Papers, W04/28, November.

[10] Browning, Martin, Angus Deaton, and Margareth Irish(1985), "A Profitable Approach to Labor Supply and Commodity Demands over the Life-Cycle", *Econometrica*, 53(3), 503-43.

[11] Browning, Martin, and Soren Leth-Petersen (2003), "Imputing Consumption from Income and Wealth Information", *Economic Journal*, 113(4), F282-301.

[12] Cochrane, John H. (1991) ,"A Simple Test of Consumption Insurance", *Journal of Political Economy*, 99(5), 957-76.

[13] Cox, Donald, Serena Ng and Andrew Waldkirch (2004), "Intergenerational Linkages in Consumption Behavior", *Journal of Human Resources* (forthcoming).

[14] Deaton, A. and J. Muellbauer (1980), "An Almost Ideal Demand System", *American Economic Review*, 70, 312-26.

[15] Hall, Robert E. and Frederic S. Mishkin (1982), "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households", *Econometrica*, 50(2), 461-81.

[16] Hayashi, Fumio (1985), "The Permanent Income Hypothesis and Consumption Durability: Analysis Based on Japanese Panel Data", *Quarterly Journal of Economics*, 100(4), 1083-1113.

[17] Hayashi, Fumio, Joseph Altonji, and Laurence Kotlikoff (1996), "Risk-Sharing between and within Families", *Econometrica*, 64(2), 261-94.

[18] Hill, M. (1992), *The Panel Study of Income Dynamics: A user's guide*, Newbury Park, California: Sage Publications.

[19] Hotz, V. Joseph, Finn E. Kydland, and Guilherme L Sedlacek, . (1988), "Intertemporal Preferences and Labor Supply", *Econometrica*, 56(2), 335-60.

[20] Hurst, Erik, and Frank Stafford (2004), "Home Is Where the Equity Is: Mortgage Refinancing and Household Consumption", *Journal of Money, Credit and Banking* (forthcoming).

[21] Jappelli, Tullio and Luigi Pistaferri (2000), "Using Subjective Income Expectations to Test for Excess Sensitivity of Consumption to Predicted Income Growth", *European Economic Review*, 44(2), 337-58.

[22] Lehnert, Andreas (2002), "Imputing Consumption From the ConsumerExpenditure Survey to the Panel Study of Income Dynamics", mimeo.

[23] Martin, Robert (2003), "Consumption, Durable Goods, and Transaction Costs", IF Discussion Paper 756 Board of Governors.

[24] Palumbo, Michael G. (1999), "Uncertain Medical Expenses and Precautionary Saving Near the End of the Life Cycle", *Review of Economic Studies*, 66(2), 395-421.

[25] Runkle, David E. (1991), "Liquidity Constraints and the Permanent-Income Hypothesis: Evidence from Panel Data", *Journal of Monetary Economics*, 27(1), 73-98.

[26] Shea, John (1995), "Union Contracts and the Life-Cycle/Permanent-Income Hypothesis", *American Economic Review*, 85(1), 186-200.

[27] Skinner, J. (1987), "A superior measure of consumption from the Panel Study of Income Dynamics", *Economic Letters*, 23, 213-16.

[28] Zeldes, Stephen P. (1989), "Consumption and Liquidity Constraints: An Empirical Investigation", *Journal of Political Economy*, 97(2), 305-46.

[29] Ziliak, James P. (1998), "Does the Choice of Consumption Measure Matter? An Application to the Permanent-Income Hypothesis", *Journal of Monetary Economics*, 41(1), 201-16.

[30] Ziliak, James P. Thomas J. Kniesner, and Douglas Holtz-Eakin (2003), "The Effect of Income Taxation on Consumption and Labor Supply: New Implications for the Optimal Income Tax", mimeo.

Table I
Sample selection in the PSID

	# dropped	# remain
Initial sample (1968-1992)	0	145,940
Interviewed prior to 1978	52,408	93,532
Change in family composition	18,570	74,962
Female head	23,779	51,183
Missing values and topcoding	308	50,875
Change in marital status	5,882	44,993
Income outliers	2,407	42,586
Born before 1920 or after 1959	8,510	34,076
Poverty subsample	12,600	21,476
Aged less than 30 or more than 65	3,674	17,778

Table II
Sample selection in the CEX

	# dropped	# remain
Initial sample	0	141,289
Missing expenditure data	1,351	139,938
Present for less than 12 months	76,773	63,165
Observed after 1992	19,310	43,855
Zero before-tax income	1,308	42,547
Missing region or education	14,029	28,418
Marital status	5,848	22,570
Born before 1920 or after 1959	4,648	17,922
Aged less than 30 or more than 65	1,843	16,079
Income outliers and incomplete income response	942	15,137

Table III
Comparison of means, PSID and CEX

	1980		1983		1986		1989		1992	
	PSID	CEX	PSID	CEX	PSID	CEX	PSID	CEX	PSID	CEX
Age	42.97	43.58	43.36	44.90	43.84	46.01	44.00	45.26	45.89	47.01
Family size	3.61	3.98	3.52	3.74	3.48	3.64	3.44	3.60	3.42	3.55
# of children	1.31	1.49	1.25	1.28	1.21	1.19	1.18	1.17	1.14	1.15
White	0.91	0.89	0.92	0.88	0.92	0.89	0.93	0.89	0.93	0.88
HS dropout	0.21	0.20	0.17	0.19	0.16	0.18	0.14	0.14	0.13	0.15
HS graduate	0.30	0.33	0.31	0.33	0.32	0.30	0.32	0.30	0.31	0.30
College dropout	0.49	0.47	0.52	0.48	0.53	0.53	0.54	0.56	0.56	0.55
Northeast	0.21	0.20	0.21	0.24	0.22	0.21	0.22	0.23	0.22	0.22
Midwest	0.33	0.28	0.31	0.26	0.30	0.27	0.30	0.28	0.30	0.29
South	0.31	0.28	0.31	0.28	0.30	0.27	0.31	0.27	0.30	0.26
West	0.15	0.24	0.17	0.21	0.18	0.25	0.18	0.23	0.18	0.23
Husband working	0.96	0.97	0.94	0.92	0.93	0.90	0.94	0.92	0.93	0.88
Wife working	0.69	0.67	0.71	0.66	0.74	0.71	0.78	0.72	0.77	0.73
Family income	32,759	29,078	37,907	35,923	45,035	43,630	52,919	51,205	61,911	56,520
Food expenditure	4,449	4,656	4,858	4,617	5,306	5,199	5,864	6,135	6,620	6,431

Table IV
Statistical significance of cross-dataset differences

Covariate (d)	Test that $M(d_p) - M(d_x)$ is constant over time	Test that $V(d_p) - V(d_x)$ is constant over time	Test that $C(f_p, d_p) - C(f_x, d_x)$ is constant over time
Family size	$F = 1.27$, p-value $= 23\%$	$F = 1.05$, p-value $= 40\%$	$F = 0.76$, p-value $= 67\%$
Age	$F = 1.05$, p-value $= 40\%$	$F = 0.64$, p-value $= 80\%$	$F = 0.69$, p-value $= 73\%$
White	$F = 0.62$, p-value $= 81\%$	$F = 0.60$, p-value $= 83\%$	$F = 0.68$, p-value $= 74\%$
High school dropout	$F = 0.74$, p-value $= 71\%$	$F = 0.75$, p-value $= 69\%$	$F = 1.57$, p-value $= 11\%$
High school graduate	$F = 0.95$, p-value $= 49\%$	$F = 0.97$, p-value $= 47\%$	$F = 1.25$, p-value $= 26\%$
College graduate	$F = 1.33$, p-value $= 20\%$	$F = 1.05$, p-value $= 40\%$	$F = 1.23$, p-value $= 27\%$
Northeast	$F = 0.68$, p-value $= 76\%$	$F = 0.67$, p-value $= 77\%$	$F = 0.28$, p-value $= 98\%$
Midwest	$F = 0.59$, p-value $= 84\%$	$F = 0.67$, p-value $= 76\%$	$F = 0.61$, p-value $= 81\%$
South	$F = 0.19$, p-value $= 99\%$	$F = 0.24$, p-value $= 99\%$	$F = 1.16$, p-value $= 31\%$
West	$F = 0.53$, p-value $= 89\%$	$F = 0.40$, p-value $= 96\%$	$F = 0.56$, p-value $= 84\%$
Born 1955-59	$F = 2.39$, p-value $= 1\%$	$F = 2.35$, p-value $= 1\%$	$F = 0.15$, p-value $= 99\%$
Born 1950-54	$F = 1.29$, p-value $= 12\%$	$F = 1.26$, p-value $= 24\%$	$F = 1.31$, p-value $= 22\%$
Born 1945-29	$F = 1.16$, p-value $= 22\%$	$F = 1.04$, p-value $= 41\%$	$F = 1.87$, p-value $= 4\%$
Born 1940-24	$F = 0.27$, p-value $= 99\%$	$F = 0.27$, p-value $= 99\%$	$F = 1.02$, p-value $= 43\%$
Born 1935-39	$F = 0.62$, p-value $= 81\%$	$F = 0.65$, p-value $= 78\%$	$F = 0.21$, p-value $= 99\%$
Born 1930-34	$F = 0.71$, p-value $= 73\%$	$F = 0.72$, p-value $= 72\%$	$F = 1.74$, p-value $= 7\%$
Born 1925-29	$F = 0.53$, p-value $= 88\%$	$F = 0.53$, p-value $= 88\%$	$F = 0.88$, p-value $= 55\%$
Number of children	$F = 0.71$, p-value $= 73\%$	$F = 1.09$, p-value $= 36\%$	$F = 0.22$, p-value $= 99\%$
$\ln(f)$	$F = 3.97$, p-value $= 0\%$	$F = 0.38$, p-value $= 96\%$	-.-

Table V
The demand for food in the CEX

This table reports IV estimates of the demand equation for (the logarithm of) food spending in the CEX. We instrument the log of total nondurable expenditure (and its interaction with age, time education dummies) with the cohort-education-year specific average of the log of the husband's hourly wage and the cohort-education-year specific average of the log of the wife's hourly wage (and their interactions with age, time and education dummies). Standard errors are in round parenthesis; the Shea's partial R^2 for the relevence of instruments in square brackets. In all cases, the p-value of the F-test on the excluded instrument is <0.01 percent.

Variable	Estimate	Variable	Estimate	Variable	Estimate
$\ln c$	0.8503 (0.1511) [0.012]	$\ln c * 1992$	0.0037 (0.0056) [0.083]	Family size	0.0272 (0.0090)
$\ln c$*High School dropout	0.0730 (0.0718) [0.050]	$\ln c$*One child	0.0202 (0.0336) [0.150]	$\ln p_{food}$	−0.9784 (0.2160)
$\ln c$*High School graduate	0.0827 (0.0890) [0.027]	$\ln c$*Two children	−0.0250 (0.0383) [0.120]	$\ln p_{transports}$	5.5376 (8.0500)
$\ln c * 1981$	0.1151 (0.1123) [0.053]	$\ln c$*Three children+	0.0087 (0.0340) [0.197]	$\ln p_{fuel+utils}$	−0.6670 (4.7351)
$\ln c * 1982$	0.0630 (0.0837) [0.052]	One child	−0.1568 (0.3215)	$\ln p_{alcohol+tobacco}$	−1.8684 (4.1425)
$\ln c * 1983$	0.0508 (0.0704) [0.048]	Two children	0.3214 (0.3650)	Born 1955-59	−0.0385 (0.0554)
$\ln c * 1984$	0.0478 (0.0662) [0.051]	Three children+	0.0132 (0.3259)	Born 1950-54	−0.0085 (0.0477)
$\ln c * 1985$	0.0304 (0.0638) [0.064]	High school dropout	−0.7030 (0.6741)	Born 1945-49	−0.0060 (0.0406)
$\ln c * 1986$	0.0223 (0.0587) [0.068]	High school graduate	−0.8458 (0.8298)	Born 1940-44	−0.0051 (0.0348)
$\ln c * 1987$	0.0528 (0.0599) [0.065]	Age	0.0122 (0.0085)	Born 1935-39	−0.0044 (0.0273)
$\ln c * 1988$	0.0416 (0.0458) [0.049]	Age^2	−0.0001 (0.0001)	Born 1930-34	0.0032 (0.0193)
$\ln c * 1989$	0.0370 (0.0373) [0.046]	Northeast	0.0087 (0.0065)	Born 1925-29	−0.0051 (0.0140)
$\ln c * 1990$	0.0187 (0.0295) [0.060]	Midwest	−0.0213 (0.0105)	White	0.0769 (0.0129)
$\ln c * 1991$	−0.0004 (0.0318) [0.111]	South	−0.0269 (0.0096)	Constant	−0.6404 (0.9266)
OID test				20.92 (d.f. 18; χ^2p-value 28%)	
Test that income elasticity does not vary over time				27.69 (d.f. 12; χ^2p-value 0.6%)	