Imputing Consumptiom In The PSID Using Food Demand Estimates From The Cex

Richard Blundell Luigi Pistaferri Ian Preston

December (2006)

Amin Shirazian

EF9905 Presentation

Background

- Lack of panel data on total consumption restricted researchers to use the scanty food expenditure information in the PSID.
- Food is a necessity (i.e. the budget share for food falls as total expenditure rises), using food instead of consumption:
    1. prevent us from estimating a range of elasticity terms.
    2. generally underestimates total consumption volatility.
    3. prevent us from explaining price shocks.
- We need some type of a panel data for consumption for many researches.

Outline

- **Motivation**: Help researchers with the lack of panel data on consumption (total expenditure) , using the available data:
  1. PSID : a panel data that contains food information, but not consumption.
  2. CEX : a detailed data set on household expenditures, including food expenditure, but it is a cross sectional data.
- **Strategy**: The idea is to find a relationship between consumption and food expenditure using CEX. Then, use the relationship ro impute consumption for PSID.

Imputing Consumption

- Demand for food (notice that it captures price effect):

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

- $x \rightarrow$ observation from the CEX. (v.s. $p \rightarrow$ observation from PSID.)
- $f \rightarrow$ food expenditure (available in both CEX and PSID).
- $D \rightarrow$ prices and a set of conditioning variables (available in both data sets.)
- $c \rightarrow$ non-durable expenditure (available only in CEX)
- $e \rightarrow$ unobserved heterogeneity in the demand for food (including measurement error in food expenditure)

Imputing Consumption

- Demand for food:

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

- **Assumption 1**: functions $\tau(.)$ and $\eta(.)$ are known monotonic increasing transformations of their arguments.
- **Assumption 2**: Food is a normal good ($\gamma \geq 0$).
- **Assumption 3**: Both data sets have the same underlying population.

Imputing Consumption

- Imputed consumption in the CEX, **assuming** $\hat{\gamma} \neq 0$:

$$\widehat{c}_{i,x} = \eta^{-1}\left(\frac{\tau\left(f_{i,x}\right) - D'_{i,x}\widehat{\beta}}{\widehat{\gamma}}\right) \tag{2}$$

- Imputed measure of consumption in the PSID:

$$\widehat{c}_{i,p} = \eta^{-1}\left(\frac{\tau\left(f_{i,p}\right) - D'_{i,p}\widehat{\beta}}{\widehat{\gamma}}\right) \tag{3}$$

Imputing Consumption

- To see how well our imputations are working, we need to compare them with the data points.
- Rewriting the imputed data we get the measurment error of the form:

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

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

Simple Case

- Hence, the imputed data is simply an error riden data (with drift).
- **Assumption 4**: $\tau(x) = x$, and $\eta(x) = x$
- To see this better, unde the above assumption, we get:

$$\widehat{c}_{i,x} = \frac{(\beta - \widehat{\beta})}{\widehat{\gamma}} + \frac{\gamma}{\widehat{\gamma}} c_{i,x} + v_{i,x} \tag{5}$$

Simple case

- assume that $c_{i,x}$ is potentially measured with classical error:

$$c^*_{i,x} = c_{i,x} + u_{i,x}$$

- under assumption 4, we get:

$$f_{i,x} = \beta + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x} \tag{6}$$

- Notice that if, total expenditure decisions were made jointly with decisions on individual commodities, such as food, $\text{Cov}\left(c_x, e_x\right) \neq 0$

Estimator

- Let $\widehat{\gamma}(y) = \frac{\mathrm{Cov}(f_x, y_x)}{\mathrm{Cov}(c_x^*, y_x)}$ be an estimator of γ. We have:

$$\operatorname{plim} \widehat{\gamma}(y) = \gamma + B^e(y) + B^m(y)$$
$$\operatorname{plim} \widehat{\beta}(y) = \beta - (B^e(y) + B^m(y)) \ \operatorname{plim} \ M(c_x)$$

- where

$$B^e(y) = \frac{\operatorname{plim} \mathrm{Cov}(e_x, y_x)}{\operatorname{plim} \mathrm{Cov}(c_x^*, y_x)}$$

$$B^m(y) = -\gamma \frac{\operatorname{plim} \mathrm{Cov}(u_x, y_x)}{\operatorname{plim} \mathrm{Cov}(c_x^*, y_x)}$$

Sample moments - CEX

- Let $\widehat{c}_x(y)$ denote the imputation with $\widehat{\beta}(y)$ and $\widehat{\gamma}(y)$, we have:

$$\text{plim } M\left(\widehat{c}_x(y)\right) = \text{ plim } M\left(c_x\right) \tag{7}$$

- And the variance term becomes:

$$\text{plim } V\left(\widehat{c}_x(y)\right) = \left(\frac{\gamma}{\gamma + B^e(y) + B^m(y)}\right)^2 \left(\text{plim } V\left(c_x\right) + \frac{1}{\gamma^2} \text{ plim } V\left(e_x\right) + \frac{2}{\gamma} \text{ plim Cov}\left(c_x, e_x\right)\right) \tag{8}$$

Sample moments - CEX

- $M(\hat{c}_x(y))$ converges in probability to $M(c_x)$, regardless of **measurement error** or **heterogeneity in food spending**.
- Keeping the assumption that expenditure decision is made jointly, notice the last two terms on the variance term would bias our variance estimation.
- Also notice that the trend in variance is independent of those terms.

IV case

- a valid instrument z_x satisfies:

$$\begin{aligned}
\operatorname{plim} \operatorname{Cov}\left(c_x^*, z_x\right) &\neq 0 \\
\operatorname{plim} \operatorname{Cov}\left(e_x, z_x\right) &= 0 \\
\operatorname{plim} \operatorname{Cov}\left(u_x, z_x\right) &= 0
\end{aligned}$$

- then $B^e(z) = B^m(z) = 0$:

$$\begin{aligned}
&\operatorname{plim} \widehat{\gamma}(z) = \gamma \\
&\operatorname{plim} \widehat{\beta}(z) = \beta \\
&\operatorname{plim} M\left(\widehat{c}_x(z)\right) = \operatorname{plim} M\left(c_x\right) \\
&\operatorname{plim} V\left(\widehat{c}_x(z)\right) = \operatorname{plim} V\left(c_x\right) + \frac{1}{\gamma^2} \operatorname{plim} V\left(e_x\right) + \frac{2}{\gamma} \operatorname{plim} \operatorname{Cov}\left(c_x, e_x\right)
\end{aligned}$$

OLS case

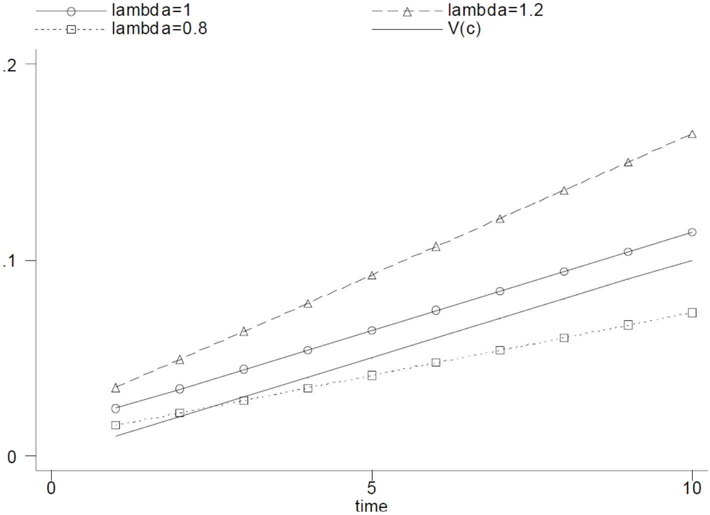
- In OLS case, $c^\star$ satisfies:

$$\begin{array}{ll} \text{Inseparability:} & \operatorname{plim} \operatorname{Cov}(e_x, c_x^*) \neq 0 \\ \text{Measurement:} & \operatorname{plim} \operatorname{Cov}(u_x, c_x^*) \neq 0 \end{array}$$

- then $B^e(z) \neq 0$, $B^m(z) \neq 0$:

$$\begin{aligned} \operatorname{plim} \widehat{\gamma}(c^*) &= \gamma + B^e(c^*) + B^m(c^*) \\ \operatorname{plim} \widehat{\beta}(c^*) &= \beta - (B^e(c^*) + B^m(c^*)) \operatorname{plim} \ M(c_x) \\ \operatorname{plim} M(\widehat{c}_x(c^*)) &= \operatorname{plim} M(c_x) \\ \operatorname{plim} V(\widehat{c}_x(c^*)) &= \left(\tfrac{\gamma}{\gamma + B^e(y) + B^m(y)}\right)^2 \operatorname{plim} V(\widehat{c}_x(z)) \end{aligned}$$

OLS-IV comparison



Sample moments - PSID

- We want to use CEX sample moments to compute PSID's, for the mean, we have:

$$\text{plim}\, M\left(\widehat{c}_p(y)\right) = \mathrm{p}\lim M\left(c_x\right) + \frac{1}{\gamma + B^e(y) + B^m(y)}\left[\mathrm{p}\lim M\left(f_p\right) - \mathrm{p}\lim M\left(f_x\right)\right] \tag{9}$$

- If food consumption is on average the same in the two data sets, the second term on the right hand side vanishes.
- Otherwise, the sample mean of imputed PSID consumption is potentially biased. (e.g.the two samples are not random samples drawn from the same underlying population)

Sample moments - PSID

- For the Variance, we have:

$$\text{plim}\, V\left(\widehat{c}_p(y)\right) = \begin{array}{l} \left(\frac{\gamma}{\gamma+B^e(y)+B^m(y)}\right)^2 \\ \left(\begin{array}{l} \text{plim}\, V\left(c_x\right) + \frac{1}{\gamma^2}\,\text{plim}\, V\left(e_x\right) + \\ \frac{2}{\gamma}\,\text{plim}\,\text{Cov}\left(c_x, e_x\right) + \\ \frac{1}{\gamma^2}\left(\text{plim}\, V\left(f_p\right) - \text{plim}\, V\left(f_x\right)\right) \end{array}\right) \end{array} \quad (10)$$

- Notice that the slope term is not affected comparing to (8).
- There is an additional reason for $V(\hat{c}_p(y))$ be different from the population variance.

Covariates

- Adding covariates to the simple model, we get:

$$f_{i,x} = D'_{i,x}\beta + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x}$$
$$D_{i,x} = \begin{pmatrix} 1 & {}_2d_{i,x} & {}_3d_{i,x} & \dots & {}_{k-1}d_{i,x} \end{pmatrix}'$$

- To fix ideas, let us consider a simple case:

$$f_{i,x} = \beta_0 + \beta_1 d_{i,x} + \gamma c^*_{i,x} + e_{i,x} - \gamma u_{i,x} \tag{11}$$

Covariates

- It is easy to show:

$$\mathrm{p}\lim \widehat{\beta}_0(d,y) = \beta_0 - (B^e(d,y) + B^m(d,y))\,(\mathrm{p}\lim M(c) - \rho\,\mathrm{p}\lim M(d))$$
$$\mathrm{p}\lim \widehat{\beta}_1(d,y) = \beta_1 - \rho\,(B^e(d,y) + B^m(d,y))$$
$$\mathrm{p}\lim \hat{\gamma}(d,y) = \gamma + B^e(d,y) + B^m(d,y)$$

- Where the bias terms are adjusted so that they can take the covariate into account:

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

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

- and ρ is defined as:

$$\rho = \frac{p\lim \mathrm{Cov}\,(c^*,d)}{p\lim V(d)}$$

Covariates

- Sample means become:

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

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

- Convergence of the sample mean in CEX is assured.
- In PSID, sample mean 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.

Covariates

- For sample variance, we have:

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

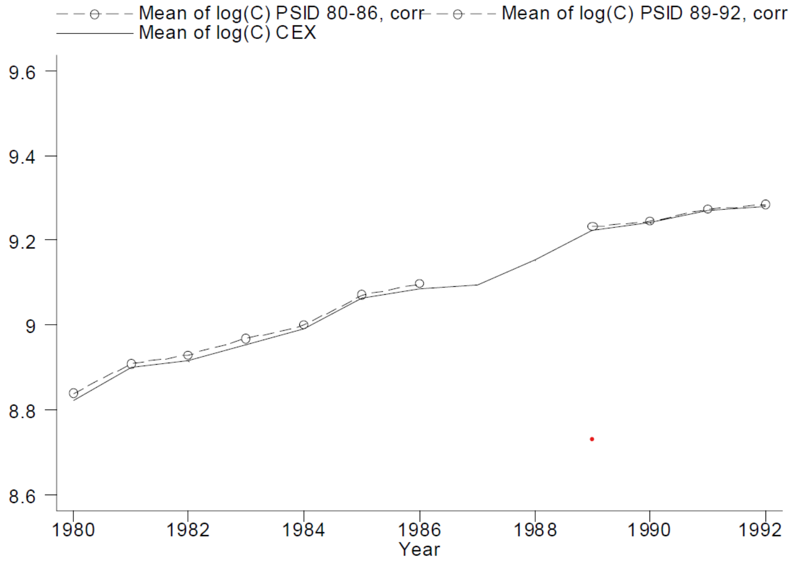
$$\text{plim } V\left(\widehat{c}_p(d,y)\right) = \text{plim } V\left(\widehat{c}_x(d,y)\right) + \left(\frac{1}{\gamma + B^e(d,y) + B^m(d,y)}\right)^2 \begin{bmatrix} \text{plim } V\left(f_p - \widehat{\beta}_1(y) d_p\right) \\ - \text{plim } V\left(f_x - \widehat{\beta}_1(y) d_x\right) \end{bmatrix}$$

- Note also that if $y = z$ is a valid instrument, then covariates have no role in determining the asymptotic expression $V\left(\widehat{c}_x(d,y)\right)$.

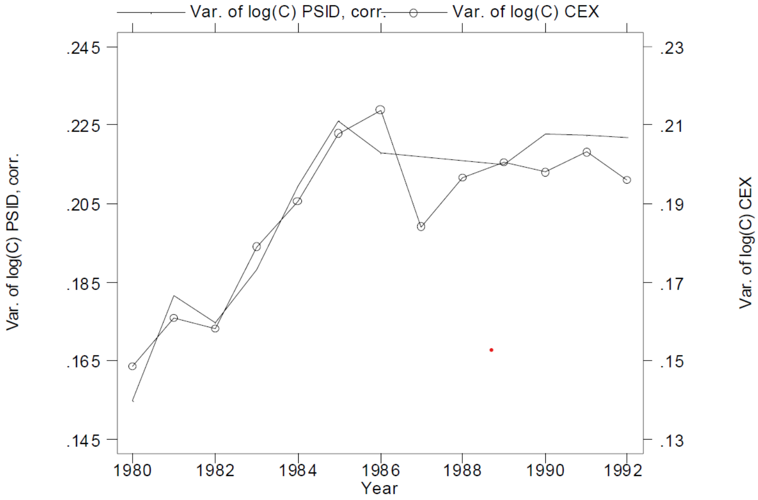
Results

- The paper expands for the case of having non-linear functional forms for $\tau(.)$ and $\eta(.)$, and having budget heterogeneity.
- Hourly wages are considered as instruments for total expenditures.
- After estimating the parameters of the model, figure 1 describes $\left(M\left(\widehat{c}_p\right) - M\left(c_x\right)\right)$, and $\frac{M(f_p)-M(f_x)}{\hat{\gamma}}$.
- The main source of difference in the imputed and the true consumption is for the latter term.
- After correcting for the difference in mean food expenditure, we see the model performs well.

Results-Corrected Mean



Results-Corrected Variance



Discussion

- We saw how once can use a combination of panel-Cross data to impute one variable from one to another.
- The idea is applicable to other datasets. (e.g. MEPS and HRS)
- Figuring if there is any other way that we can contribute to precision of imputation.