Regression with an Imputed Dependent Variable

CEMMAP-UKHLS Missing Data Workshop

Thomas F. Crossley (Essex, IFS and ESCoE)

with Peter Levell (IFS and UCL) and Stavros Poupakis (UCL)

May 2019

Motivation

- Wish to estimate $y = X\beta + \epsilon$.
- β is the object of interest.
- OLS would be fine if we had complete data: $plim(X\epsilon) = 0$.
- But no data on $\frac{1}{N}\sum yX$.
- Have some data on $(y_1, Z_1), (X_2, Z_2)$.
  - these are both random samples from the population of interest
  - subscript indexes data set (or sample), absence of subscript means population.
- Z is a proxy or proxies for y.

Motivating Example

- Estimating effect of income or wealth (shocks) on consumption expenditure with:
  - a data set with food exp. and income/wealth.
  - a data set with food exp. and consumption exp.
  - e.g., PSID and CE, UKHLS and LCF, HFCS and National Budget Surveys

Set up

$$y = X\beta + \epsilon$$

$$plim(\frac{1}{n_j} X_j' X_j) = \Sigma_{xx}$$

and

$$plim(\frac{1}{n_j} X_j' \epsilon_j) = 0$$

- (de-meaned variables)

Set up

- data on (y_1, Z_1) and (X_2, Z_2).

$$z = y\gamma + u.$$

$$z = X\beta\gamma + \epsilon\gamma + u.$$

- Z *must* depend on ϵ

$$plim(\frac{1}{n_j} X_j' u_j) = 0$$

$$plim(\frac{1}{n_j} y_j' u_j) = 0 \, (can \, relax)$$

Alternative Imputation/Data Combination Strategies (1)

- Skinner (1987) suggested regressing y_1 on Z_1 in the CE and using the resulting coefficients to predict $\hat{y}_2$ in the PSID
- Then regressing $\hat{y}_2$ on X_2.
- With a single spending category as the proxy, the first stage is an "inverse" Engel curve.
- **RP** procedure for "Regression Prediction".
- Advocated by Browning, Crossley, Weber (2003).
- Employed by Attanasio and Pistaferri (2014), Arrondel et al., (2015).
- Add a first-stage residual: **RP+**.

Alternative Imputation/Data Combination Strategies (2)

- Blundell, Pistaferri, Preston (2004,2008) regress Z_1 on y_1 then predict $\hat{y}_2 = Z_2 \frac{1}{\hat{\gamma}}$: **BPP** procedure.
- Again using the CE and PSID, proxy, z, is food expenditure.
- Estimate an Engel curve and then invert it to predict consumption.
- Recently been employed Attanasio, Hurst and Pistaferri (2012).

Alternative Imputation/Data Combination Strategies (3)

- Do not impute y at the unit level at all.
- Recover β from a combination of moments taken from the two surveys.
- Arellano and Meghir (1992): **AM** procedure.
- Here:
  - Regress Z_1 on y_1 to get $\hat{\gamma}$.
  - Regress Z_2 on X_2 to get $\hat{\gamma\beta}$
  - Take ratio of the two to estimate β.

RP inconsistent for β

- **RP** does not consistently estimate β: $plim(\hat{\beta}^{RP}) = \beta R^2_{y_1,Z_1}$

Proof.

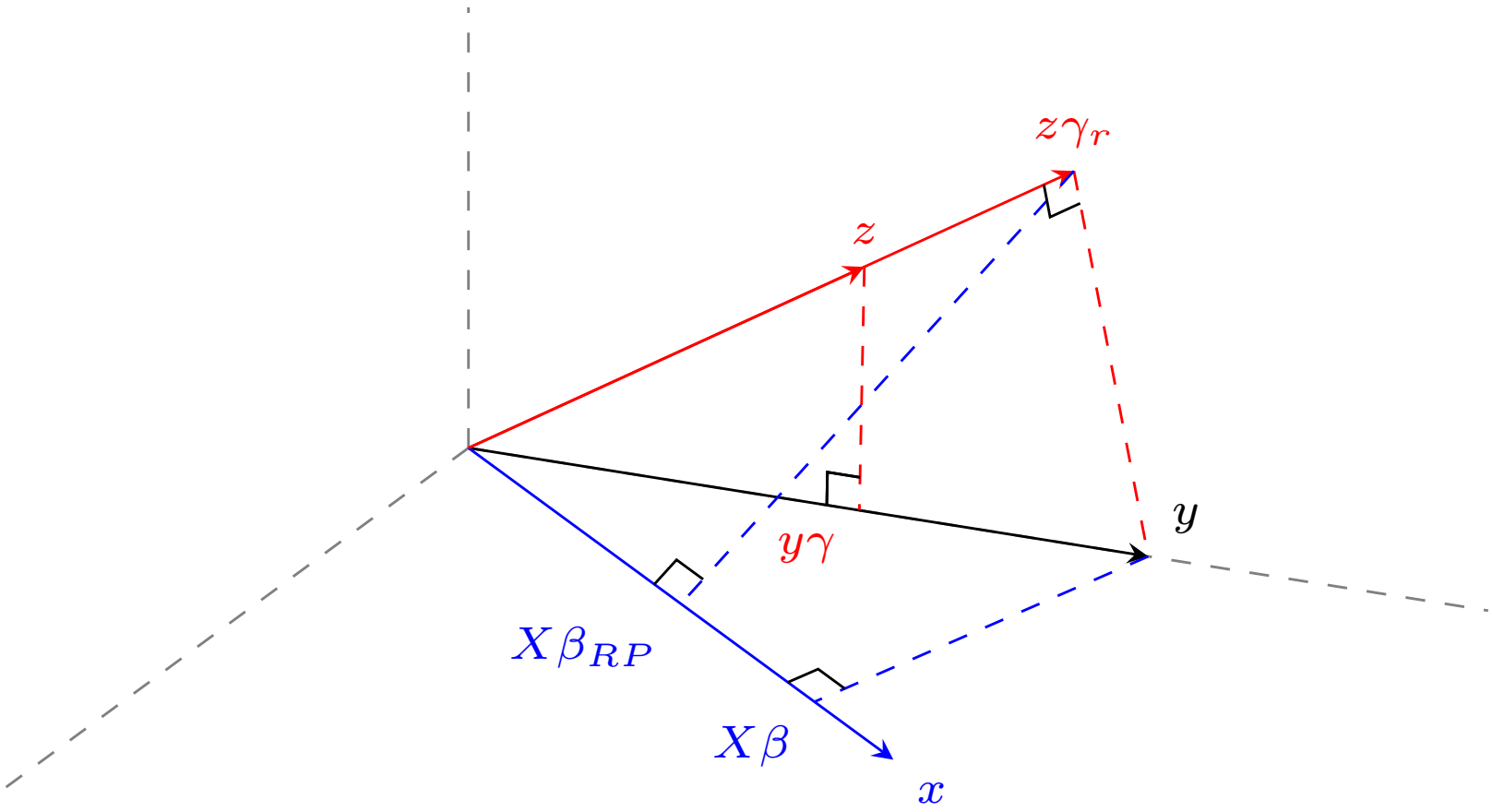
$$plim(\hat{\beta}^{RRP}) = plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\right\}$$

$$= plim\left\{\left(\frac{X_2'X_2}{n_2}\right)^{-1}\frac{X_2'Z_2}{n_2}\left(\frac{Z_1'Z_1}{n_1}\right)^{-1}\frac{Z_1'y_1}{n_1}\frac{1}{R^2_{y_1,Z_1}}R^2_{y_1,Z_1}\right\}$$

$$= \beta\gamma Q_{ZZ}^{-1}\gamma'\Sigma_{yy}\left[\Sigma_{yy}\gamma Q_{zz}^{-1}\gamma'\Sigma_{yy}\right]^{-1}\Sigma_{yy}\phi_{y,Z}$$

$$= \beta\gamma Q_{ZZ}^{-1}\gamma'\left[\gamma Q_{zz}^{-1}\gamma'\right]^{-1} = \beta\phi_{y,Z}$$

Geometric Intuition



Magnitude

- First stage R^2 for food on total consumption $50 - 70\%$.
- Downward bias $30 - 50\%$.

RP+

- Similar problem with **RP+** (intuition below).

RRP

- As the bias in the *RP* procedure is an estimable quantity, it can be corrected.
- Rescale $\hat{\beta}^{RP}$ by the estimated first stage (centered) $R^2_{y,Z}$: "Re-scaled Regression Prediction" (**RRP**, $\hat{\beta}^{RRP}$).

$$plim(\hat{\beta}^{RRP}) = plim\left(\frac{\hat{\beta}^{RP}}{R^2_{y_1,Z_1}}\right) = \beta$$

- Proof follows immediately from Proposition 1.
- Rescaling of $\hat{\beta}^{RP}$ is equivalent to rescaling the predicted consumption vector $\hat{y}_2^{RP}$ by $1/R^2_{y,Z}$.

AM, BPP and RRP

- IFF there is a single proxy z, $\hat{\beta}^{RRP}$, $\hat{\beta}^{BPP}$ and $\hat{\beta}^{AM}$ are numerically identical.

Proof.

$$\hat{\beta}^{BPP} = (X_2' X_2)^{-1} X_2' z_2 (y_1' z_1)^{-1} y_1' y_1 = \hat{\beta}^{RRP}$$

$$\begin{aligned}\hat{\beta}^{AM} = \widehat{\beta\gamma}/\hat{\gamma} &= (X_2' X_2)^{-1} X_2' z_2 \left[(y_1' y_1)^{-1} y_1' z_1\right]^{-1} \\ &= (X_2' X_2)^{-1} X_2' z_2 (y_1' z_1)^{-1} y_1' y_1 = \hat{\beta}^{RRP} = \hat{\beta}^{BPP}\end{aligned}$$

∎

- Therefore **BPP**, **AM** consistently estimate β.

Other Moments: Variances and Covariances

$$Asymp\ Var(\hat{y}^{RP}) = Asymp\ Var(y) \times \phi_{y,Z}.$$

$$Asymp\ Cov(\hat{y}^{RP}, X) = Asymp\ Cov(y, X) \times \phi_{y,Z}.$$

$$Asymp\ Var(\hat{y}^{RRP}) = Asymp\ Var(\hat{y}^{BPP}) = Asymp\ Var(y)/\phi_{y,Z}.$$

$$Asymp\ Cov(\hat{y}^{RRP}, X) = Asymp\ Cov(\hat{y}^{BPP}, X) = Asymp\ Cov(y, X)$$

Other Moments: Means

- de-meaning not necessary
- $plim(\frac{\sum \hat{y}^{RP}}{n_2}) = \mu_y$,
- But then $plim(\frac{\sum \hat{y}^{RP}/R^2_{y,Z}}{n_2}) \neq \mu_y$.
- Statistical agency cannot release a single imputed $\hat{y}$.

Other Moments

Table: Summary of imputation methods (consistency)

	μ_y	σ_{yy}	β
Regression Prediction (**RP**)	✓	×	×
Regression Prediction + $\hat{e}$ (**RP+**)	✓	✓	×
Rescaled Regression Prediction (**RRP**)	×	×	✓
Blundell et al., 2004; 2008 (**BPP**)	✓	×	✓
Arellano and Meghir, 1992 (**AM**)	-	-	✓

Hot-Deck Imputation and Item Non-response

- y drawn from a matched cell is a regression prediction plus a residual (Lillard et al., 1983; David et al., 1986). .
  - Saturated regression on categorical variables
- If one or more matching variables are excluded from X this maps into **RP+**.
- Suppose fraction $\frac{n_a}{N}$ of cases missing.

$$plim\left(\hat{\beta}\right) = \frac{N - n_a}{N}\beta + \frac{n_a}{N}\beta\phi_{y,z} = \beta\left(1 + \frac{n_a}{N}\left(\phi_{y,z} - 1\right)\right).$$

Practicalities: Additional Covariates

- Residualize or add to both stages (Frisch-Waugh-Lovell).
- first-stage partial R^2.

Practicalities: Measurement error in y

- Easy to see that with **AM** we need an instrument for y because we need consistent estimate of γ.
- True for **BPP** and **RRP** too.
- For **RRP**, sample R^2 will not be consistent estimate of population R^2 in presence of ME (but can estimate).

Practicalities: Panel Case

- Often wish to estimate $\Delta y = \Delta X\beta + \epsilon$
- OLS would be fine if we had complete data $E[X\epsilon] = 0$, $plim(X\epsilon) = 0$
- But no data on $\frac{1}{N}\sum \Delta y \Delta X$
- Have some data on $(y_1^1, Z_1), (y_2^0, Z_2), (\Delta X_3, Z_3)$ where $\Delta y = y^1 - y^0$
  - e.g. cross sectional budget survey and panel income/wealth survey

Practicalities: Panel Case

- **BPP** and **RRP** are identical (with one proxy) and consistent.

$$\hat{\beta} = (\Delta X' \Delta X)^{-1} \Delta X' \Delta \hat{y}$$

where $\Delta \hat{y} = \hat{y}^1 - \hat{y}^0$

Related Literature: Berkson Measurement Error

- classical measurement error on the right:
  $\tilde{X} = X + \tilde{v},\ \tilde{v} \perp X,\ y = \tilde{X}\beta - \tilde{v}\beta + \epsilon$
- classical measurement error on the left:
  $\tilde{y} = y + \tilde{v},\ \tilde{v} \perp y,\ \tilde{y} = X\beta + \epsilon + \tilde{v}$
- Berkson measurement error (prediction error) on the right:
  $X = \hat{X} + \hat{v},\ \hat{v} \perp \hat{X},\ y = \hat{X}\beta - \hat{v}\beta + \epsilon$
- Here: prediction error (Berkson) on the left.

Related Literature: Berkson Measurement Error

- Hyslop and Imbens (2001) show attenuation bias in a regression of $\hat{y}$ on X where $\hat{y}$ is an optimal linear predictor (see also Hoderlien and Winter (2010))
- Key differences:
  - assume prediction by respondent and respondent knows Z, β and $E[X]$
  - also assume $Z = y + u$; $(\gamma = 1)$

Related Literature: 2SIV

- Wish to estimate $y = X\beta + \epsilon$, have some data on $(y_1, Z_1), (X_2, Z_2)$.
- Z is a grouping variable (e.g.. birth cohort, occupation, birth cohort x education).
- In this case we effectively use **Z to impute X**.
- Two Sample IV (2SIV, Angrist & Krueger, 1992).

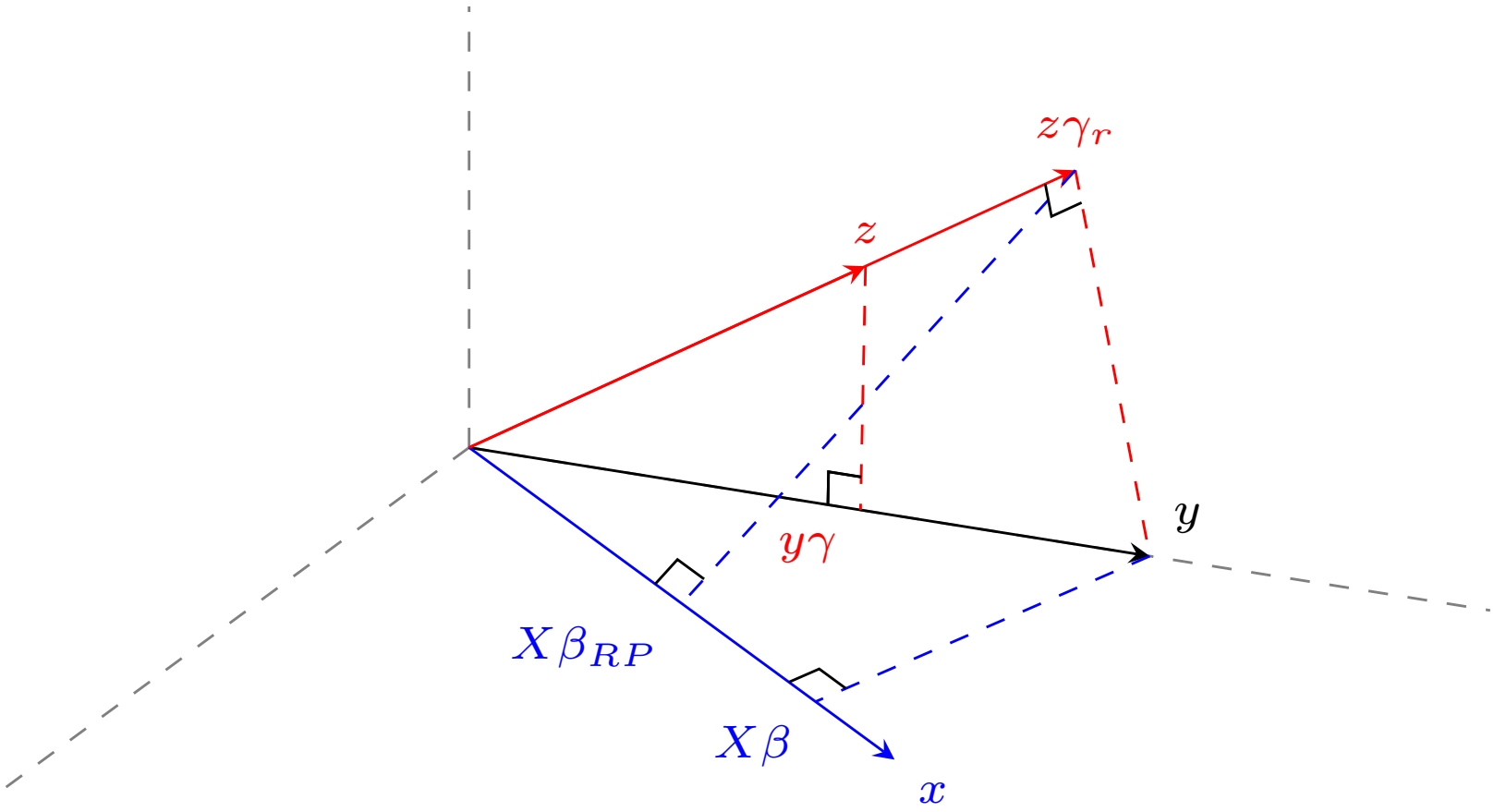
$$\hat{\beta}_{TSIV} = \left(\frac{Z_2' X_2}{n_2}\right)^{-1} \left(\frac{Z_1' y_1}{n_1}\right)$$

- key assumption: $Z \perp \epsilon$ (Z affects y *only* through X).
- Lusardi (1996): CE and PSID.

Related Literature: Partial Imputation

- Hirsch and Schumacher (2004), Bollinger and Hirsch (2006)
  - Partial imputation of the dependent variable (item non-response).
  - Hot-deck, matching on *subset* of X (no proxies).
  - Bias.
- common with us: limit case where Z spans X

Geometric Intuition



Inference

- OLS with full data: asymptotic variance for $\hat{\beta}$ of $(\Sigma_{XX})^{-1}\sigma_{\epsilon}^{2}$.
- **two** losses of precision: imputation, data combination.
- One proxy case:
  - $\hat{\beta}^{AM}$, $\hat{\beta}^{RRP}$ and $\hat{\beta}^{BPP}$ are numerically identical
  - Start from $\hat{\beta}^{AM}$
- General case in paper.

Imputation on Full Data

- Asymptotic variance-covariance matrix for moments:

$$F = \begin{bmatrix} \sigma_u^2 \Sigma_{yy} & \beta\sigma_u^2 \Sigma_{XX} \\ \beta\sigma_u^2 \Sigma_{XX} & \left(\gamma^2\sigma_u^2 + \sigma_\epsilon^2\right) \Sigma_{XX} \end{bmatrix}$$

- The asymptotic variance covariance matrix of (β, γ) is $\left(G' F^{-1} G\right)^{-1}$.
- The asymptotic variance of $\hat{\beta}$ is:

$$Asymp.Var(\hat{\beta}) = \frac{\left(\Sigma_{XX}\right)^{-1} \sigma_\epsilon^2}{\phi_{yZ}}$$

- Loss of asymptotic precision proportional to the first stage R_{yZ}^2.
  - Note similarity to linear IV (Shea, 1997).

Imputation and Data Combination

- Asymptotic variance-covariance matrix for moments:

$$F = \begin{bmatrix} \sigma_u^2 \Sigma_{yy} & 0 \\ 0 & \left(\gamma^2 \sigma_u^2 + \sigma_\epsilon^2\right) \Sigma_{xx} \end{bmatrix}$$

- The asymptotic variance of $\hat{\beta}$ is:

$$\left(\Sigma_{xx}\right)^{-1} \left(\sigma_\epsilon^2 + \gamma^{-2} \sigma_u^2\right) + \left(\Sigma_{yy}\right)^{-1} \beta^2 \gamma^{-2} \sigma_u^2.$$

$$= \frac{\left(\Sigma_{xx}\right)^{-1} \sigma_\epsilon^2}{\phi_{yz}} + 2\beta^2 \left(\frac{1 - \phi_{yz}}{\phi_{yz}}\right)$$

- The second term is a *second* loss of precision.

OLS SE are biased

- With some algebra, can show that:

$$plim\left[\hat{V}^{OLS}(\hat{\beta})\right] = \left[\frac{(\Sigma_{XX})^{-1}\sigma_{\epsilon}^{2}}{\phi_{yZ}} + \beta^{2}\left(\frac{1-\phi_{yZ}}{\phi_{yZ}}\right)\right]$$

- Too small by factor $\beta^2\left(\frac{1-\phi_{yZ}}{\phi_{yZ}}\right)$.
- Can be corrected using the available consistent estimates of β and ϕ_{yZ}.
- STATA command available from authors.

Monte Carlo: Design

- $x \sim N(0, 2)$
- $y = 1 + \beta x + \epsilon$ with $\sigma_\epsilon = 1$, $\beta = 1$
- $z_j = 1 + 0.5 y_j + u_j$ with $\sigma_u = 1$
- First stage $R^2 = 0.56$
- Simulate population, draw samples (500) of (y_1, z_1) and (x_2, z_2).
- 10,000 replications.

Monte Carlo Results: One Proxy

	FULL	RP	RP+	RRP	BPP	AM
$\hat{\beta}$ $(\beta = 1)$	1.000	0.556	0.555	1.002	1.002	1.002
Std. Dev. $\hat{\beta}$	0.022	0.036	0.049	0.065	0.065	0.065
SE($\hat{\beta}$)	0.022	0.028	0.043	0.050	0.050	
Corrected SE($\hat{\beta}$)				0.064		
$E(\hat{y})$ $(E[y] = 1)$	1.000	1.000	0.999	1.805	1.000	
$V(\hat{y})$ $(V[y] = 5)$	4.999	2.784	5.000	9.048	9.048	

Monte Carlo Results: Two Proxies

	FULL	RP	RP+	RRP	AM
$\hat{\beta}$	1.000	0.712	0.712	1.000	1.001
Std. Dev. $\hat{\beta}$	0.023	0.034	0.044	0.048	0.048
SE($\hat{\beta}$)	0.022	0.028	0.039	0.039	
Corrected SE($\hat{\beta}$)				0.048	

Monte Carlo Results: Hot-deck Imputation (1 proxy, 10 bins)

	FULL	Hot-deck (HD)	RHD
$\hat{\beta}$ $(\beta = 1)$	1.000	0.532	0.986
Std. Dev. $\hat{\beta}$	0.022	0.049	0.088
$E(\hat{y})$ $(E[y] = 1)$	1.000	1.001	1.858
$V(\hat{y})$ $(V[y] = 5)$	4.999	4.990	17.22

Empirical Examples

- Impute consumption.
- PSID 2005-2013: "full" data.
- Also impute consumption from CE (Attanasio and Pistaferri, 2014).
  - Food consumption as proxy.
  - Discarding "full" data.

Housing Wealth Effects

- Elasticity of nondurable consumption wrt house value.
- Following Skinner (1989).
- Homeowners only.

Housing Wealth Effects: First Stage

Table: Imputing nondurable consumption spending using the CE

	(1) 2005	(2) 2007	(3) 2009	(4) 2011	(5) 2013
log Food	0.562***	0.545***	0.541***	0.549***	0.555***
	(0.012)	(0.008)	(0.008)	(0.008)	(0.008)
log Utilities	0.377***	0.382***	0.410***	0.384***	0.389***
	(0.015)	(0.010)	(0.011)	(0.010)	(0.011)
Cars	0.035***	0.036***	0.027***	0.042***	0.031***
	(0.005)	(0.004)	(0.003)	(0.004)	(0.004)
Partial R^2	0.728	0.755	0.751	0.761	0.753
N	1,590	2,896	2,759	2,668	2,470

Housing Wealth Effects: Results

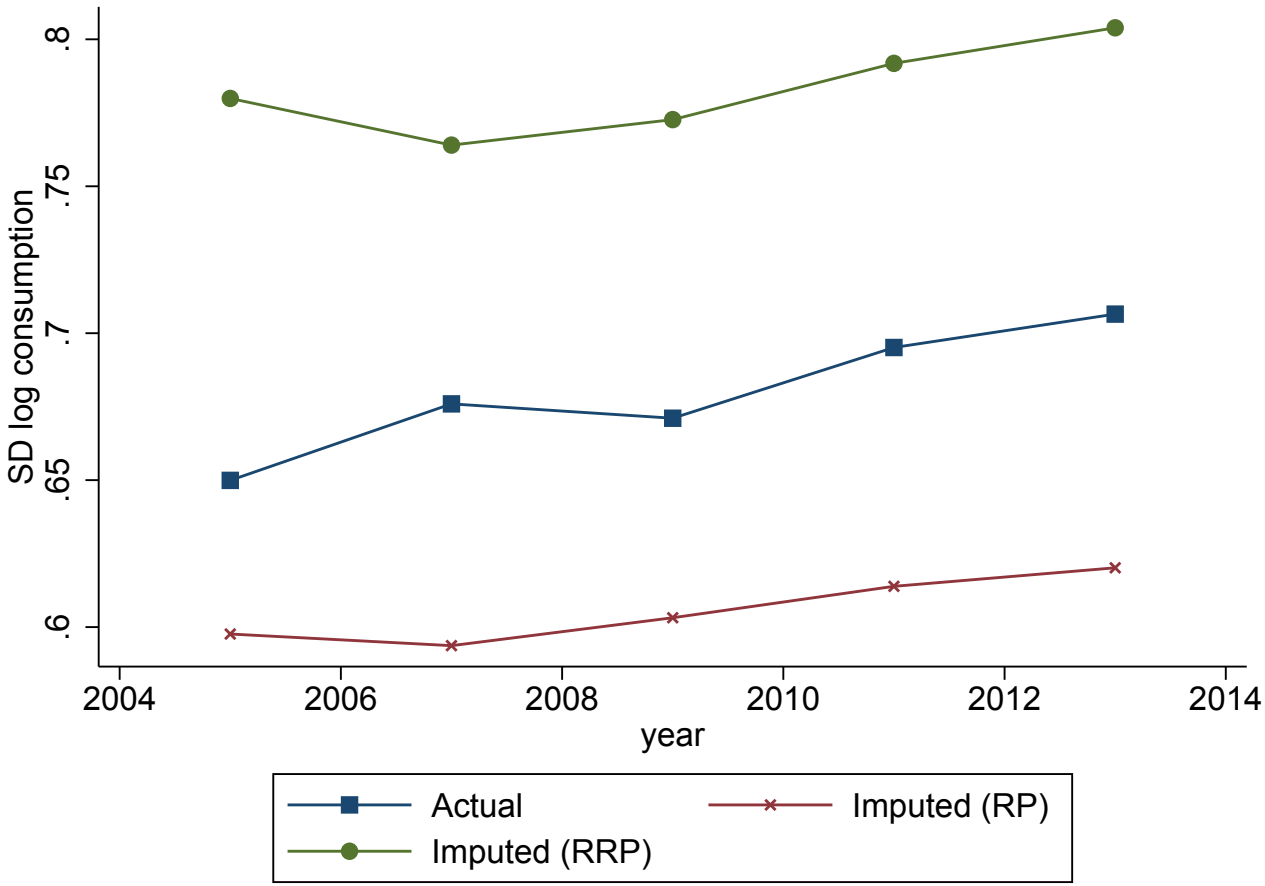
	(1) PSID	(2) CE (RP)	(3) CE (RRP)
log Income$_{t-3}$	0.047**	0.036*	0.048*
	(0.016)	(0.015)	(0.020)
log Income$_{t-2}$	0.064***	0.043*	0.057*
	(0.018)	(0.017)	(0.022)
log Income$_{t-1}$	0.040*	0.024	0.032
	(0.017)	(0.016)	(0.021)
log Income$_{t}$	0.109***	0.080***	0.106***
	(0.020)	(0.019)	(0.025)
log Income$_{t+1}$	0.105***	0.075***	0.100***
	(0.010)	(0.010)	(0.013)
log House value	0.114***	0.083***	0.111***
	(0.010)	(0.010)	(0.013)
N	5,406	5,406	5,406

Consumption Inequality

- In the spirit of Attanasio and Pistaferri (2014).
- Homeowners and non-homeowners.

Consumption Inequality: Results

Figure: Standard deviation of log consumption



Key Points (1)

- Imputed dependent variable can bias regression coefficients.
  - Berkson measurement error in *dependent* variable a problem.
- Bias related to first-stage R^2 under plausible assumptions.
- Method of imputation matters (e.g. **BPP**).
- Imputation of consumption.
- More generally: hot-deck imputation.

Key Points (2)

- Solution depends on structure (but: sometimes theory helps)
  - A good proxy can't be a good instrument (and vice versa).
- Best imputation strategy also depends on objects of interest.
- Approach to data combination and imputation should depend on object of interest and plausibility of alternative assumptions.
  - Challenge to data providers.