

Regression with an Imputed Dependent Variable

New Perspectives on Consumption Measures, STICERD

Thomas F. Crossley (Essex and IFS)

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

December 2016

Motivation

- Wish to estimate $C = X\beta + \epsilon$
- β is main object of interest
- OLS would be fine if we had complete data $E[X\epsilon] = 0$,
 $\text{plim}(X\epsilon) = 0$
- But no data on $\frac{1}{N} \sum CX$
- Have some data on $(C_1, Z_1), (X_2, Z_2)$
 - these are both random samples from the population of interest
 - subscript indexes data set (or sample), absence of subscript means population

Motivating Example

- Estimating effect of income or wealth (shocks) on consumption with
 - a data set with food and income/wealth
 - a data set with food and consumption expenditure
 - e.g., PSID and CE

Why bother?

- We can invert the inter-temporal budget constraint/impute internally/Ziliak method
- But

$$y_{h,t} - [(1 + r_{h,t})^{-1} w_{h,t+1} - w_{h,t}] \approx y_{h,t} - [w_{h,t+1} - w_{h,t}]$$

is total spending, not consumption, or even consumption expenditure

- for example, evidence that house price shocks lead to *investment* spending
- measurement error in wealth or income will be on the left and the right (Browning, Crossley & Winter, 2014)

Combining Data

- 1 Z is in instrument (for contrast)
- 2 Z is a proxy (our interest)

Case 1

Case 1: Z is an instrument (for contrast)

- Wish to estimate $C = X\beta + \epsilon$, have some data on $(C_1, Z_1), (X_2, Z_2)$
- 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; see also Arellano and Meghir, 1992)

$$\hat{\beta}_{TSIV} = \left(\frac{Z_2' X_2}{n_2} \right)^{-1} \left(\frac{Z_1' C_1}{n_1} \right)$$

- key assumption: $Z \perp \epsilon$ (Z affects C only through X)
- Measurement error in C is no problem as long as ME uncorrelated with Z

Case 1

Case 1: Z is an instrument (for contrast)

- 2SIV is not in general efficient because it does not take account of fact that Z_1 and Z_2 are different in finite samples
- Two-Sample-two-stage Least Squares improvement due to Inoue & Solon (2010)

$$\hat{\beta}_{TS2SLS} = \left(\hat{X}_1' \hat{X}_1 \right)^{-1} \hat{X}_1' C_1$$

- where $\hat{X}_1 = Z_1(Z_2'Z_2)^{-1}Z_2'X_2$

$$\hat{\beta}_{TS2SLS} = \left(\frac{Z_2'X_2}{n_2} \right)^{-1} W_{12} \left(\frac{Z_1'C_1}{n_1} \right)$$

- where $W_{12} = (Z_2'Z_2/n_2)(Z_1'Z_1/n_1)^{-1}$.

Case 2

Case 2: Z is a Proxy (Our interest)

- Example: Z is food spending.
 - many surveys, thought well-measured
- Engel curve: $Z = C\gamma + u = X\beta\gamma + \epsilon\gamma + u$ (variables might be logs)
- Note that Z **must** depend on ϵ (Z has some information about C that is not contained in X)
 - this is the **opposite assumption to Case 1**
- In this case **impute C using Z**
- If Z contains only one good then is just a vector, if several goods, $Z^A, Z^B, \dots$ then Z is a matrix
- $C \perp u$ or need an additional instrument (e.g. if measurement error in C). More below.

How to impute C using Z ?

- Classic paper: Skinner (1987 *Economic Letters*) - regress C_1 on Z_1 to predict $\hat{C}_2$
 - estimating “inverse Engel curve”
 - Browning, Crossley, Weber (*EJ* 2003)
 - Attanasio and Pistaferri (*AER(P&P)*, 2014), Arrondel, Lamarche, Savignac (*ECB*, 2015),
- Blundell, Preston, Pistaferri (BPP, 2004, 2008 *AER*) - regress Z_1 on C_1 predict $\hat{C}_2 = Z_1^{\frac{1}{\hat{\gamma}}}$
 - estimate Engel curve and invert
 - also Attansio, Hurst and Pistaferri (*CRIW Volume*, 2015)
- Engel Curve + Reduced Form:
 - regress Z_1 on C_1 to get $\hat{\gamma}$, regress Z_2 on X_2 to get $\hat{\gamma}\beta$, take ratio to estimate β

Case 2

Skinner (estimate inverse Engel curve)

- $E[\hat{C}] = E[C]$
- BUT neither unbiased nor consistent for β

$$\hat{\beta}_{Skinner} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1$$

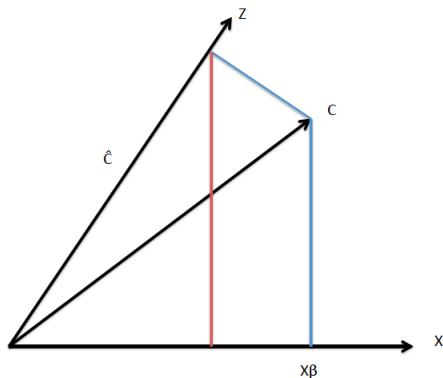
$$plim(\hat{\beta}_{Skinner}) = \beta - plim((X'X)^{-1}X'M_Z C) = R^2\beta$$

- where R^2 is from the (population) imputation regression of C on Z

Case 2

Skinner

- $\hat{\beta}_{Skinner} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1$
- $plim(\hat{\beta}_{Skinner}) = R^2\beta$



Relationship to Literature

- classical measurement error on the right:
 $\tilde{X} = X + \tilde{v}$, $\tilde{v} \perp X$, $C = \tilde{X}\beta - \tilde{v}\beta + \epsilon$
- classical measurement error on the left:
 $\tilde{C} = C + \tilde{v}$, $\tilde{v} \perp C$, $\tilde{C} = X\beta + \epsilon + \tilde{v}$
- Berkson measurement error (prediction error) on the right:
 $X = \hat{X} + \hat{v}$, $\hat{v} \perp \hat{X}$, $C = \hat{X}\beta - \hat{v}\beta + \epsilon$
- Here: prediction error (Berkson) on the left.
- Hyslop and Imbens (2001) show attenuation bias in a regression of $\hat{C}$ on X where $\hat{C}$ is an optimal linear predictor.
- Key differences:
 - assume prediction by respondent and respondent knows Z , β and $E[X]$
 - also assume $Z = C + u$; ($\gamma = 1$)

Suggestion: Modified Skinner

- However, we can fix it

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1/R^2$$

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1[C_1'Z_1(Z_1'Z_1)^{-1}Z_1'C_1]^{-1}C_1'C_1$$

- Where in the case of a single proxy, this reduces to

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(C_1'Z_1)^{-1}C_1'C_1$$

- Consistent:

$$plim\hat{\beta}_{SkinnerR^2} = \beta \frac{\gamma\beta\sigma_{X'X}^2 + \gamma\sigma_{\epsilon}^2}{\gamma\beta\sigma_{X'X}^2 + \gamma\sigma_{\epsilon}^2} = \beta$$

- Equivalent to rescaling $\hat{C}_2$ by $1/R^2$ (or rescaling $\hat{\beta}$ by $1/R^2$)
 - but note that $\frac{1}{R^2}E[\hat{C}_2] \neq E[C]$ (and $E[\frac{1}{R^2}\hat{C}_2] \neq E[C]$)

Case 2

BPP (estimate Engel curve and invert)

- Numerically identical to R^2 -rescaled Skinner in the case of 1 proxy

$$\hat{\beta}_{BPP} = (X_2'X_2)^{-1}X_2'Z_2(C_1'Z_1)^{-1}C_1'C_1$$

- Therefore also consistent
- Intuition: product of coefficients from simple regression and reverse regression are the R^2 .

Engel Curve + Reduced Form

- Reduced form: $Z = C\gamma + u = X\beta\gamma + \epsilon\gamma + u$
- $\hat{\beta}\gamma = (X_2'X_2)^{-1}X_2'Z_2$
- $\hat{\gamma} = (C_1'C_1)^{-1}C_1'Z_2$
- Ratio $\hat{\beta}\gamma/\hat{\gamma}$ identical to previous estimators
- Consistency direct from Slutsky (as well as by numerical equivalence to BPP and R^2 -rescaled Skinner)

Case 2

How much does it matter?

- Skinner inverse Engel curve $R^2 \approx 0.7$
 - multiple proxies: food, food at home, rent
- Suggest using unadjusted Skinner imputation leads to underestimation of β by 30%

Variances and Covariances

- Skinner:
 - $Var[\hat{C}] = Var[C]R^2$;
 - $Cov[\hat{C}, X] = Cov[C, X]R^2$
 - Attanasio-Pistaferri show trends in $Var[\hat{C}]$ and $Var[C]$ similar.
- BPP:
 - $Var[\hat{C}] = \frac{Var[C]}{R^2}$
 - $Cov[\hat{C}, X] = Cov[C, X]$

Extensions

- 1 Precision: correcting for multiple samples
- 2 Precision: multiple proxies
- 3 Measurement error in C
- 4 Panel Case

Precision: correcting for multiple samples

- Just like the Inoue & Solon (2010) refinement to 2SIV, we can improve precision by accounting for the fact that Z_1 and Z_2 are different in finite samples

$$\hat{\beta} = (X_2'X_2)^{-1}X_2'Z_2W(Y_1'Z_1)^{-1}C_1'C_1$$

- where $W = (Z_2'Z_2)^{-1}(Z_1'Z_1)$ is a correction matrix for differences between the two samples (just as in TS2SLS (Inoue & Solon, 2010))

Back to Case 1 (Two Sample IV)

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1/R_{Z,C}^2$$

$$\hat{\beta}_{SkinnerR^2} = (Z_2'X_2)^{-1}Z_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1R_{Z,X}^2/R_{Z,C}^2$$

$$\hat{\beta}_{SkinnerR^2} = \hat{\beta}_{TS2SLS}R_{Z,X}^2/R_{Z,C}^2$$

Precision: multiple proxies

- Precision can be improved with multiple Z s
- BPP, “Engel curve + Reduced Form”: combine proxies (but note: logs), or minimum distance.
- But we can still do R^2 -rescaled Skinner

$$\hat{\beta} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'C_1[C_1'Z_1(Z_1'Z_1)^{-1}Z_1'C_1]^{-1}C_1'C_1$$

- Equivalence of methods fails because $(Z'Z)$, $(Z'C)$ no longer scalars
- R^2 -rescaled Skinner consistent, and as efficient as minimum distance

Measurement error in C

- Easy to see that with “Engel curve + Reduced Form” we need an instrument for C because we need consistent estimate of γ
- True for Blundell and R^2 -rescaled Skinner too.
- For R^2 -rescaled Skinner, sample R^2 will not be consistent estimate of population R^2 in presence of ME (But can estimate the population R^2)
- Need to bear in mind Campos (2014 *EL*): one instrument insufficient if ME correlated with other covariates

Panel Case

- Often wish to estimate $\Delta C^* = \Delta X\beta + \epsilon$
- β is main object of interest
- OLS would be fine if we had complete data $E[X\epsilon] = 0$, $\text{plim}(X\epsilon) = 0$
- But no data on $\frac{1}{N} \sum \Delta C \Delta X$
- Have some data on $(C_1^1, Z_1), (C_2^0, Z_2), (\Delta X_3, Z_3)$ where $\Delta C = C^1 - C^0$
 - e.g.. cross sectional budget survey and panel income/wealth survey

Panel Case

- Blundell and R^2 -rescaled Skinner are identical (with one proxy) and consistent.

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

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

- Can do Inoue-Solon-type correction for each $\hat{C}$

Monte Carlo evidence

Design:

$$X_2 \sim U(-2, 2)$$

$$C_1 = 1.0X_2 + \epsilon \text{ with } \sigma_\epsilon^2 = 2,$$

$$Z_{A1} = 0.5C_1 + u_1 \text{ \& } Z_{B1} = 0.5C_1 + u_1 \text{ with } u_1 \sim MVN(0, \Sigma)$$

$$Z_{A2} = 0.5C_1 + u_2 \text{ \& } Z_{B2} = 0.5C_1 + u_2 \text{ with } u_2 \sim MVN(0, \Sigma)$$

$$\text{where } \Sigma = \begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix} \text{ with } \sigma_{AB} = 0.6 \text{ and } \sigma_B^2 = 3.$$

Monte Carlo evidence

Results:

<i>n</i>	$\sigma_A^2 = 2$				$\sigma_A^2 = 4$			
	250		1000		250		1000	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Full	1.000	(0.111)	1.000	(0.055)	1.000	(0.111)	1.000	(0.055)
Skinner1Z	0.401	(0.090)	0.400	(0.044)	0.250	(0.071)	0.250	(0.035)
Skinner2Zs	0.476	(0.096)	0.474	(0.048)	0.403	(0.092)	0.400	(0.045)
BPP1Z	1.004	(0.211)	1.001	(0.103)	1.010	(0.275)	1.002	(0.136)
Skinner1ZR²	1.004	(0.211)	1.001	(0.103)	1.010	(0.275)	1.002	(0.136)
BPP1ZC	1.000	(0.193)	1.001	(0.095)	1.005	(0.256)	1.001	(0.127)
Skinner1ZR²C	1.000	(0.193)	1.001	(0.095)	1.005	(0.256)	1.001	(0.127)
Skinner2ZsR²C	0.995	(0.175)	1.000	(0.086)	0.998	(0.192)	0.999	(0.096)

Empirical Example

- PSID (79-92) + CE (80-92) (From BPP)
- z - food at home
- $\Delta \ln \hat{c} = \beta \Delta \ln y + \epsilon$, instrument with $\Delta \ln y_{-1}$
 - c - nondurable consumption, y - net income (including transfers)
 - in the spirit of Ziliak (JME, 1998)

Empirical Example

Results:

	Skinner	BPP	BPP-C
OLS 2nd Stage			
β	0.0615	0.1019	0.0886
(SE)	(0.0092)	(0.0154)	(0.0134))
IV 2nd Stage			
β	0.0323	0.0421	0.366
(SE)	(0.0350)	(0.0588)	(0.0511)

Key points

- Imputed dependent variable can bias regression coefficients
 - Berkson measurement error in dependent variable a problem
- First stage R^2 is a critical quantity
- Method of imputation can matter

What do do

- If you have instruments, use the Inoue-Solon refinement to Angrist-Krueger (TS-2SLS)
- If you have proxies, R^2 -rescaled Skinner method for multiple proxies with the Inoue-Solon-type refinement is a possible approach.
 - even with a single proxy, you want the Inoue-Solon-type refinement.
- Stability of the relationship between $Var[\hat{C}]$ and $Var[C]$ easy to check