Regression with an Imputed Dependent Variable

Stavros Poupakis (Essex and UCL)

with Thomas F. Crossley (Essex and IFS) and Peter Levell (IFS and UCL)

July 2017

7th Conference of the European Survey Research Association
Lisbon, Portugal

Motivation

Estimate effect of income/wealth (X) on consumption (Y)
With data on both and usual assumptions we can estimate

$$Y = X\beta + \epsilon$$

with OLS

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

No data on $X'Y$, BUT

1. A dataset with food and consumption expenditure (Z_1, Y_1)
   - e.g. Consumer Expenditure Survey (CE)
2. A dataset with food and income/wealth (Z_2, X_2)
   - e.g. Panel Study of Income Dynamics (PSID)

Assume both random samples from population of interest

General Background

- With no assumptions/restrictions, the best we can get is the Fréchet bounds for ρ_{xy}

$$\rho_{xz}\rho_{yz} \pm \sqrt{1-\rho_{xz}^2}\sqrt{1-\rho_{yz}^2}$$

- Thus, for β

$$\left[\frac{\sigma_y}{\sigma_x}LB, \frac{\sigma_y}{\sigma_x}UB\right]$$

- See survey in Handbook of Econometrics chapter (Ridder and Moffitt, 2007)

Why bother?

We can invert the inter-temporal budget constraint/impute internally (Ziliak, 1998)

But

$$x_{h,t} - [w_{h,t+1} - w_{h,t}]$$

is total spending, not consumption.

Distinctions between consumption and investment spending can be important.

Combining Data

Case 1. Z is an instrument (for contrast)

- Use **Z to impute X**

Case 2. Z is a proxy (our interest)

- Use **Z to impute Y**

Case 1: Z is an instrument (for contrast)

Wish to estimate $Y = X\beta + \epsilon$ and have (Y_1, Z_1), (X_2, Z_2)

(e.g. Z is birth cohort, occupation, birth cohort x education)

Angrist and Krueger (1992) Two Sample IV (2SIV):

$$\hat{\beta}_{TSIV} = \left(Z_2' X_2\right)^{-1} Z_1' Y_1$$

Case 1: Z is an instrument (for contrast)

2SIV is not in general efficient

2S2SLS is, see Inoue and Solon (2010)

$$\hat{\beta}_{TS2SLS} = \left(\hat{X}_1'\hat{X}_1\right)^{-1}\hat{X}_1'Y_1$$

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

Case 1: Z is an instrument (for contrast)

2SIV is not in general efficient

2S2SLS is, see Inoue and Solon (2010)

$$\hat{\beta}_{TS2SLS} = \left(\hat{X}_1'\hat{X}_1\right)^{-1}\hat{X}_1'Y_1$$

$$= [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Z_1(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Y_1$$

Case 1: Z is an instrument (for contrast)

2SIV is not in general efficient

2S2SLS is, see Inoue and Solon (2010)

$$\hat{\beta}_{TS2SLS} = \left(\hat{X}_1'\hat{X}_1\right)^{-1}\hat{X}_1'Y_1$$

$$= [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Z_1(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Y_1$$

$$= (Z_2'X_2)^{-1}Z_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1$$

Case 1: Z is an instrument (for contrast)

2SIV is not in general efficient

2S2SLS is, see Inoue and Solon (2010)

$$\hat{\beta}_{TS2SLS} = \left(\hat{X}_1'\hat{X}_1\right)^{-1}\hat{X}_1'Y_1$$

$$= [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Z_1(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Y_1$$

$$= (Z_2'X_2)^{-1}Z_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1$$

Thus

$$\hat{\beta}_{TS2SLS} = (Z_2'X_2)^{-1}\ W\ Z_1'Y_1$$

Case 2: Z is a Proxy (Our interest)

Example: Z is food spending (many surveys, well-measured)

Engel curve: $Z = Y\gamma + u$

Reduced form: $Z = X\beta\gamma + \epsilon\gamma + u$

(vars usually in logs)

Case 2: Z is a Proxy (Our interest)

1. Classic paper: Skinner (1987) - inverse Engel curve
Regress Y_1 on Z_1, then predict $\hat{Y}_2 = Z_2\hat{\gamma}_r$

2. Blundell, Pistaferri and Preston (2004, 2008) - Engel curve, then invert
Regress Z_1 on Y_1, then predict $\hat{Y}_2 = Z_2\frac{1}{\hat{\hat{\gamma}}}$

3. Arellano and Meghir (1992) - Engel Curve + Reduced Form
Regress Z_1 on Y_1 to get $\hat{\gamma}$
Regress Z_2 on X_2 to get $\widehat{\beta\gamma}$
Take ratio to estimate β

Skinner (estimate inverse Engel curve)

Inconsistent

$$\hat{\beta}_{Skinner} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1$$

$$plim(\hat{\beta}_{Skinner}) = R^2 \beta$$

where R^2 is from the (population) regression of Y on Z

Suggestion: Modified Skinner

However, we can fix it

$$\hat{\beta}_{SkinnerR^2} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1 / R^2$$
$$\hat{\beta}_{SkinnerR^2} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1 [Y_1' Z_1 (Z_1' Z_1)^{-1} Z_1' Y_1]^{-1} Y_1' Y_1$$

Suggestion: Modified Skinner

However, we can fix it

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1/R^2$$
$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1[Y_1'Z_1(Z_1'Z_1)^{-1}Z_1'Y_1]^{-1}Y_1'Y_1$$

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

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Y_1'Z_1)^{-1}Y_1'Y_1$$

Consistent:

$$plim\hat{\beta}_{SkinnerR^2} = \beta$$

Equivalent to rescaling $\hat{y}_2$ by $1/R^2$ (or rescaling $\hat{\beta}$ by $1/R^2$)

BPP (estimate Engel curve and invert)

Numerically identical to R^2-rescaled Skinner if one proxy

$$\hat{\beta}_{BPP} = (X_2' X_2)^{-1} X_2' Z_2 (Y_1' Z_1)^{-1} Y_1' Y_1$$

Intuition: In simple regression, product of coefficients from regression and reverse regression is the R^2.

$$\hat{\gamma} \times \hat{\gamma}_r = R^2 \quad \Leftrightarrow \quad \frac{1}{\hat{\gamma}} = \frac{\hat{\gamma}_r}{R^2}$$

AM (Engel Curve + Reduced Form)

Reduced form: $Z = C\gamma + u = X\beta\gamma + \epsilon\gamma + u$

$$\hat{\beta\gamma} = (X_2' X_2)^{-1} X_2' Z_2$$

$$\hat{\gamma} = (Y_1' Y_1)^{-1} Y_1' Z_1$$

Ratio $\frac{\hat{\beta\gamma}}{\hat{\gamma}}$ identical to previous estimators

$$\frac{\hat{\beta\gamma}}{\hat{\gamma}} = (X_2' X_2)^{-1} X_2' Z_2 (Y_1' Z_1)^{-1} Y_1' Y_1$$

AM (Engel Curve + Reduced Form)

Reduced form: $Z = C\gamma + u = X\beta\gamma + \epsilon\gamma + u$

$$\hat{\beta\gamma} = (X_2' X_2)^{-1} X_2' Z_2$$

$$\hat{\gamma} = (Y_1' Y_1)^{-1} Y_1' Z_1$$

Ratio $\frac{\hat{\beta\gamma}}{\hat{\gamma}}$ identical to previous estimators

$$\frac{\hat{\beta\gamma}}{\hat{\gamma}} = (X_2' X_2)^{-1} X_2' Z_2 (Y_1' Z_1)^{-1} Y_1' Y_1$$

AM (Engel Curve + Reduced Form)

Reduced form: $Z = C\gamma + u = X\beta\gamma + \epsilon\gamma + u$

$$\hat{\beta\gamma} = (X_2' X_2)^{-1} X_2' Z_2$$

$$\hat{\gamma} = (Y_1' Y_1)^{-1} Y_1' Z_1$$

Ratio $\frac{\hat{\beta\gamma}}{\hat{\gamma}}$ identical to previous estimators

$$\frac{\hat{\beta\gamma}}{\hat{\gamma}} = (X_2' X_2)^{-1} X_2' Z_2 (Y_1' Z_1)^{-1} Y_1' Y_1$$

Link of Case 1 and Case 2

$$\hat{\beta}_{SkinnerR^2} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1 / R^2_{Z,Y}$$

Link of Case 1 and Case 2

$$\hat{\beta}_{SkinnerR^2} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1 / R^2_{Z,Y}$$

$$= (X_2' X_2)^{-1} X_2' Z_2 (Z_2' Z_2')^{-1} Z_2' X_2 [(Z_2' Z_2')^{-1} Z_2' X_2]^{-1} (Z_1' Z_1)^{-1} Z_1' Y_1 / R^2_{Z,Y}$$

Link of Case 1 and Case 2

$$\hat{\beta}_{SkinnerR^2} = (X_2^{'}X_2)^{-1}X_2^{'}Z_2(Z_1^{'}Z_1)^{-1}Z_1^{'}Y_1/R^2_{Z,Y}$$

$$= (X_2^{'}X_2)^{-1}X_2^{'}Z_2(Z_2^{'}Z_2^{'})^{-1}Z_2^{'}X_2[(Z_2^{'}Z_2^{'})^{-1}Z_2^{'}X_2]^{-1}(Z_1^{'}Z_1)^{-1}Z_1^{'}Y_1/R^2_{Z,Y}$$

$$= (Z_2^{'}X_2)^{-1}Z_2^{'}Z_2(Z_1^{'}Z_1)^{-1}Z_1^{'}Y_1R^2_{Z,X}/R^2_{Z,Y}$$

Link of Case 1 and Case 2

$$\hat{\beta}_{SkinnerR^2} = (X_2'X_2)^{-1}X_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1/R^2_{Z,Y}$$

$$= (X_2'X_2)^{-1}X_2'Z_2(Z_2'Z_2')^{-1}Z_2'X_2[(Z_2'Z_2')^{-1}Z_2'X_2]^{-1}(Z_1'Z_1)^{-1}Z_1'Y_1/R^2_{Z,Y}$$

$$= (Z_2'X_2)^{-1}Z_2'Z_2(Z_1'Z_1)^{-1}Z_1'Y_1R^2_{Z,X}/R^2_{Z,Y}$$

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

Ways to improve Precision

1. Correcting for multiple samples
2. Using Multiple proxies

Precision: correcting for multiple samples

Just like Inoue and Solon (2010) refinement, account for the fact that Z_1 and Z_2 are different in finite samples

$$\hat{\beta} = (X_2'X_2)^{-1}X_2'Z_2 W_{12}(Y_1'Z_1)^{-1}Y_1'Y_1$$

where $W_{12} = (Z_2'Z_2)^{-1}(Z_1'Z_1)$ is a correction matrix for differences between the two samples (just as in TS2SLS)

Precision: multiple proxies

- Precision can be improved with multiple Zs
- Numerical Equivalence breaks
- But we can still do R^2-rescaled Skinner

$$\hat{\beta} = (X_2' X_2)^{-1} X_2' Z_2 (Z_1' Z_1)^{-1} Z_1' Y_1 [Y_1' Z_1 (Z_1' Z_1)^{-1} Z_1' Y_1]^{-1} Y_1' Y_1$$

- In essence, choose the proxies that have the higher partial R^2, as with many instruments (Shea, 1997)

Monte Carlo evidence

Design:

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

$Y_1 = 1.0X_2 + \epsilon$ with $\sigma_\epsilon^2 = 2$,

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

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

where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{bmatrix}$ with $\sigma_{AB} = 0.6$, $\sigma_B^2 = 3$ and $\sigma_A^2 = 4$.

10,000 replications

Monte Carlo evidence

Results:

n	250		1000	
	Mean	SD	Mean	SD
Full	1.000	(0.111)	1.000	(0.055)
Skinner1Z	0.250	(0.071)	0.250	(0.035)
Skinner2Zs	0.403	(0.092)	0.400	(0.045)
BPP1Z	1.010	(0.275)	1.002	(0.136)
Skinner1ZR^2	1.010	(0.275)	1.002	(0.136)
BPP1ZC	1.005	(0.256)	1.001	(0.127)
Skinner1ZR^{2}C	1.005	(0.256)	1.001	(0.127)
Skinner2ZsR^{2}C	1.002	(0.192)	1.001	(0.096)

Conclusion

Our method

- R^2-rescaled method is easy to use
- Asymptotics and finite sample properties
- Include multiple proxies based on partial R^2
- Use Inoue and Solon (2010) correction.
- Standard Errors need correction (as usual in TS estimators)

Generalises to cases:

- with measurement error
- with additional covariates
- with panel data

References

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. Journal of the American Statistical Association, 87(418):328–336.

Arellano, M. and Meghir, C. (1992). Female labour supply and on-the-job search: an empirical model estimated using complementary data sets. The Review of Economic Studies, 59(3):537–559.

Blundell, Pistaferri and Preston (2004). Imputing consumption in the PSID using food demand estimates from the CEX. IFS Working Papers.

Blundell, Pistaferri and Preston (2008). Consumption inequality and partial insurance. American Economic Review, pages 1887–1921.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. The Review of Economics and Statistics, 92(3):557–561.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. Handbook of Econometrics, 6:5469–5547.

Shea, J. (1997). Instrument relevance in multivariate linear models: A simple measure. The Review of Economics and Statistics, 79(2):348–352.

Skinner, J. (1987). A superior measure of consumption from the panel study of income dynamics. Economics Letters, 23(2):213–216.

Ziliak, J. P. (1998). Does the choice of consumption measure matter? an application to the permanent-income hypothesis. Journal of Monetary Economics, 41(1):201–216.