

Measurement Error

Often a data set will contain imperfect measures of the data we would ideally like.

Aggregate Data: (GDP, Consumption, Investment)

are only best guesses of theoretical counterparts

and frequently revised by government statisticians (so earlier estimates must have been subject to error)

Survey Data: (income, health, age)

Individuals often lie, forget or round to nearest large number (£102 a week or £100?)

Proxy Data: (Ability, Intelligence, Permanent Income)

Difficult to agree on definition, let alone measure

Measurement Error in Dependent Variable

True: $y^{\text{true}} = b_0 + b_1X + u$ (1)

Observe: $y = y^{\text{true}} + e$ (2)

ie dependent variable measured with error e

e is a random residual term just like u, so $E(e)=0$

Sub. (2) into (1)

$$\begin{aligned} y - e &= b_0 + b_1X + u \\ y &= b_0 + b_1X + u + e \\ y &= b_0 + b_1X + v \quad \text{where } v = u + e \end{aligned} \quad (3)$$

Ok to estimate (3) by OLS, since

$$\begin{aligned} E(u) &= E(e) = 0 \\ \text{Cov}(X,u) &= \text{Cov}(X,e) = 0 \end{aligned}$$

(nothing to suggest X variable correlated with meas. error in dependent variable)

So OLS estimates are unbiased in this case

but

standard errors are larger than would be in absence of meas. error

$$\text{True: } \text{Var}(\hat{\beta}) = \frac{\sigma_u^2}{N\text{Var}(X)} \quad (\text{A})$$

$$\text{Estimate: } \text{Var}(\tilde{\beta}) = \frac{\sigma_u^2 + \sigma_e^2}{N\text{Var}(X)} \quad (\text{B})$$

Since $\text{var}(v)=\text{var}(u+e)=\text{var}(e)+\text{var}(u)+2\text{cov}(e,u)$ and assume things that cause measurement error in y are unrelated to residual u so $\text{cov}(e,u)=0$. (B) shows that residual variance in presence of measurement error in dep. variable now also contains an additional contribution from error in y variable, σ_e^2

Measurement Error in Explanatory Variable

True: $y^{\text{true}} = b_0 + b_1 X^{\text{true}} + u$ (1)

Observe: $X = X^{\text{true}} + w$ (2)

ie rhs var. measured with error (w)

sub. (2) into (1)

$$\begin{aligned} y^{\text{true}} &= b_0 + b_1(X-w) + u \\ y^{\text{true}} &= b_0 + b_1X - b_1w + u \\ y^{\text{true}} &= b_0 + b_1X + v \end{aligned} \quad (3)$$

where now $v = -b_1w + u$

(so residual term again consists of 2 components)

Does this matter?

In true model, we know that (4)

$$\hat{b}_1 = \frac{\text{Cov}(X^{\text{true}}, y^{\text{true}})}{\text{Var}(X^{\text{true}})} = \frac{\text{Cov}(X^{\text{true}}(b_0 + b_1 X^{\text{true}} + u))}{\text{Var}(X^{\text{true}})} = b_1 + \frac{\text{Cov}(X^{\text{true}}, u)}{\text{Var}(X^{\text{true}})}$$

and by **assumption** $\text{Cov}(X^{\text{true}}, u) = 0$, so that $E(\hat{b}_1) = b_1$

but

in presence of measurement error, we estimate (3)

and so $\text{Cov}(X, v) = \text{Cov}(X^{\text{true}} + w, -b_1w + u)$

(using (2) & (3))

Expanding terms

$$= \text{Cov}(X^{\text{true}}, u) + \text{Cov}(X^{\text{true}}, -b_1w) + \text{Cov}(w, u) + \text{Cov}(w, -b_1w)$$

Since u and w are independent errors (caused by different factors), no reason to expect them to be correlated with each other **or** the **true** value of X (any error in X should not depend on the level of X), so

$$\text{Cov}(w, u) = \text{Cov}(X^{\text{true}}, u) = \text{Cov}(X^{\text{true}}, -b_1w) = 0$$

This leaves

$$\text{Cov}(X, v) \neq 0 = \text{Cov}(w, -b_1w) = -b_1 \text{Cov}(w, w) = -b_1 \text{Var}(w)$$

In other words there **is now** a correlation between the X variable and the error term in (3).

y_true	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x_true	.5999999	.0583875	10.276	0.000	.465358	.7346417
_cons	25.00002	10.47727	2.386	0.044	.8394026	49.16064

Now look at consequence of measurement error in dependent variable

```
. reg y_obs x_true
```

Source	SS	df	MS	Number of obs = 10		
Model	11880.0048	1	11880.0048	F(1, 8) =	77.65	
Residual	1223.99853	8	152.999817	Prob > F =	0.0000	
				R-squared =	0.9066	
				Adj R-squared =	0.8949	
Total	13104.0033	9	1456.00037	Root MSE =	12.369	

y_observ	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x_true	.6000001	.0680908	8.812	0.000	.4429824	.7570178
_cons	24.99999	12.21846	2.046	0.075	-3.175826	53.17581

Consequence: Coefficients virtually identical, (unbiased) but standard errors larger and hence t values smaller and confidence intervals wider.

Measurement error in explanatory variable:

```
. reg y_true x_obs
```

Source	SS	df	MS	Number of obs = 10		
Model	11658.3625	1	11658.3625	F(1, 8) =	83.15	
Residual	1121.63494	8	140.204367	Prob > F =	0.0000	
				R-squared =	0.9122	
				Adj R-squared =	0.9013	
Total	12779.9974	9	1419.99971	Root MSE =	11.841	

y_true	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x_observ	.4522529	.0495956	9.12	0.000	.3378852	.5666206
_cons	50.11701	9.225319	5.43	0.001	28.84338	71.39063

Consequence: **both** coefficients biased.

and slope coefficient is biased toward zero
(0.45 compared with true 0.60 ie underestimate effect by 25%)

Intercept is biased upward (compare 50.1 with 25.0)

Problem is that $Cov(X,u) \neq 0$
ie residual and right hand side variable are correlated
In such situations the X variable is said to be **endogenous**

Solution?

- Get better data

If that is not possible do something to get round the problem.

- replace the variable causing the correlation with the residual with one that is not but that at the same time is still related to the original variable

Any variable that has these 2 properties is called an **Instrumental Variable**

More formally, an instrument Z for the variable of concern X satisfies

1) $Cov(X,Z) \neq 0$

correlated with the problem variable

2) $Cov(Z,u) = 0$

but uncorrelated with the residual (so does not suffer from measurement error and also is not correlated with any unobservable factors influencing the dependent variable)

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \quad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

$$\begin{aligned} \text{So } Cov(Z,y) &= Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u) \\ &= 0 + b_1Cov(Z,X) + 0 \end{aligned}$$

(using rules on covariance of a constant and assumption 1 above)

$$\text{So } b_1^{IV} = \frac{Cov(Z, y)}{Cov(Z, X)} \quad \left(\text{compare with } b_1^{OLS} = \frac{Cov(X, y)}{Var(X)} \right)$$

The IV estimate is **unbiased** in large samples – consistent - (can prove this using similar steps to above) which makes it a useful estimation technique to employ

However can show that (in the 2 variable case)

$$Var(\hat{\beta}_1^{IV}) = \frac{s^2}{N * Var(X)} * \frac{1}{r_{XZ}^2}$$

where r_{xz}^2 is the square of the correlation coefficient between endogenous variable and instrument

$$\left(\text{compared with OLS } Var(\hat{\beta}_1^{OLS}) = \frac{s^2}{N * Var(X)} \right)$$

So IV estimation is less precise (efficient) than OLS estimation when the explanatory variables are exogenous, but

the greater the correlation between X and Z the *smaller* is $\text{Var}(\hat{\beta}^{IV})$.

ie IV estimates are generally more precise if the correlation between instrument and endogenous variable is large.

However if X and Z are perfectly correlated then Z must also be correlated with u and so problem is not solved.

Even though IV estimation is unbiased this property is *asymptotic* ie only holds when sample size is very large

Since can always write the IV estimator as

$$b_1^{IV} = \frac{\text{Cov}(Z, y)}{\text{Cov}(Z, X)} = b_1 + \frac{\text{Cov}(Z, u)}{\text{Cov}(Z, X)}$$

(just sub. in for y from (1))

In small samples poor correlation between X and Z can lead to “small sample bias”

So: always check extent of correlation between X and Z before any IV estimation

You can have as many instruments as you like – though finding good ones is a different matter. In large samples more is better. In small samples a minimum number of instruments is preferred.

Where to find good instruments?

- difficult. The appropriate instrument will vary depending on the issue under study.
- In the case of measurement error, could use the *rank* of X as an instrument (ie order the variable X by size and use the number of the order rather than the actual value. Though this assumes that the measurement error is not so large as to affect the (true) ordering of the X variable)

```
egen rankx=rank(x_obs) /* stata command to create the ranking of x_observ */
```

```
. list x_obs rankx
      x_observ      rankx
  1.         60         1
  2.         80         2
  3.        100         3
  4.        120         4
  5.        140         5
  6.        200         6
  7.        220         7
  8.        240         8
  9.        260         9
 10.        280        10
```

ranks from smallest observed x to largest

Now do instrumental variable estimates using rankx as the instrument for x_obs

```
ivreg y_t (x_ob=rankx)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS			
Model	11654.5184	1	11654.5184	Number of obs =	10	
Residual	1125.47895	8	140.684869	F(1, 8) =	84.44	
Total	12779.9974	9	1419.99971	Prob > F =	0.0000	
				R-squared =	0.9119	
				Adj R-squared =	0.9009	
				Root MSE =	11.861	

y_true	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x_observ	.460465	.0501086	9.19	0.000	.3449144	.5760156
_cons	48.72095	9.307667	5.23	0.001	27.25743	70.18447

Instrumented: x_observ
Instruments: rankx

Can see both estimated coefficients are a little closer to their true values than estimates from regression with measurement error (but not much)
In this case the rank of X is not a very good instrument

Note that standard error in instrumented regression is larger than standard error in regression of y_true on x_observed as expected with IV estimation

Testing for Endogeneity

It is good practice to compare OLS and IV estimates. If estimates are very different this may be a sign that things are amiss.
Using the idea that IV estimation will always be (asymptotically) unbiased whereas OLS will only be unbiased if $Cov(X,u) = 0$ then can do the following:

Wu-Hausman Test for Endogeneity

$$1. \text{ Given } y = b_0 + b_1X + u \quad (A)$$

Regress X on the instrument(s) Z

$$X = d_0 + d_1Z + v \quad (B)$$

Save the residuals \hat{v}

$$2. \text{ Include this residual as an extra term in the original model ie estimate}$$

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

3. If $b_2 = 0$ conclude there is no correlation between X and u
 If $b_2 \neq 0$ conclude there is correlation between X and u

(intuitively, since assume Z is uncorrelated with u – the Z variables are exogenous - , only way X could be correlated with u is through v (in (B) and $u = b_2v + e$)

Example:

The data set *ivdat.dta* contains information on the number of GCSE passes of a sample of 16 year olds and the total income of the household in which they live.

Income tends to be measured with error. Individuals tend to mis-report incomes, particularly third-party incomes and non-labour income. The following regression may therefore be subject to measurement error in one of the right hand side variables, (the gender dummy variable is less subject to error).

```
. reg nqfede incl female
```

Source	SS	df	MS			
Model	274.029395	2	137.014698	Number of obs = 252		
Residual	2344.9706	249	9.41755263	F(2, 249) = 14.55		
				Prob > F = 0.0000		
				R-squared = 0.1046		
				Adj R-squared = 0.0974		
Total	2619.00	251	10.4342629	Root MSE = 3.0688		

nqfede	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
incl	.0396859	.0087786	4.52	0.000	.022396	.0569758
female	1.172351	.387686	3.02	0.003	.4087896	1.935913
_cons	4.929297	.4028493	12.24	0.000	4.13587	5.722723

To test endogeneity first regress the suspect variable on the instrument and any exogenous variables in the original regression

```
reg incl ranki female
```

Source	SS	df	MS			
Model	81379.4112	2	40689.7056	Number of obs = 252		
Residual	40863.626	249	164.110948	F(2, 249) = 247.94		
				Prob > F = 0.0000		
				R-squared = 0.6657		
				Adj R-squared = 0.6630		
Total	122243.037	251	487.024053	Root MSE = 12.811		

incl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ranki	.2470712	.0110979	22.26	0.000	.2252136	.2689289
female	.2342779	1.618777	0.14	0.885	-2.953962	3.422518
_cons	.7722511	1.855748	0.42	0.678	-2.882712	4.427214

1. save the residuals

```
. predict uhat, resid
```

2. include residuals as additional regressor in the original equation

```
. reg nqfede incl female uhat
```

Source	SS	df	MS	Number of obs = 252		
Model	281.121189	3	93.7070629	F(3, 248)	=	9.94
Residual	2337.87881	248	9.42693069	Prob > F	=	0.0000
-----				R-squared	=	0.1073
Total	2619.00	251	10.4342629	Adj R-squared	=	0.0965
-----				Root MSE	=	3.0703
ngfede	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
incl	.0450854	.0107655	4.19	0.000	.0238819	.0662888
female	1.176652	.3879107	3.03	0.003	.4126329	1.940672
uhat	-.0161473	.0186169	-0.87	0.387	-.0528147	.0205201
_cons	4.753386	.4512015	10.53	0.000	3.864711	5.642062

Now added residual is not statistically significantly different from zero, so conclude that there is no endogeneity bias in the OLS estimates. Hence no need to instrument.

Note you can also get this result by typing the following command after the ivreg command

```
ivendog
```

```
Tests of endogeneity of: incl
```

```
H0: Regressor is exogenous
```

```
Wu-Hausman F test: 0.75229 F(1,248) P-value = 0.38659
```

```
Durbin-Wu-Hausman chi-sq test: 0.76211 Chi-sq(1) P-value = 0.38267
```

the first test is simply the square of the t value on uhat in the last regression (since $t^2 = F$)

N.B. This test is only as good as the instruments used and is only valid asymptotically. This may be a problem in small samples and so you should generally use this test only with sample sizes well above 100.