

1. You have data on the natural log of hourly wages, (LWAGE), age, measured in years, (AGE), a dummy variable, (FEMALE), that takes the value 1 if female 0 otherwise, the product of FEMALE and AGE, FEMALE\*AGE, years of work in the current job, TENURE, and its square, TENURE2.

You estimate the following regressions:

(1) reg lhw age female femage tenure tenure2

Source	SS	df	MS			
Model	100.000000	5	16.808223	Number of obs =	3006	
Residual	600.000000	3000	.277909507	F( 5, 2299) =	60.48	
Total	700.000000	3005	.313782583	Prob > F =	0.0000	
				R-squared =	0.1429	
				Adj R-squared =	0.1443	
				Root MSE =	.52717	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0042398	.0014985	2.83	0.005	.0013012	.0071784
female	-.0200000	.0500000	-0.40	0.745	-.1860868	.1331442
femage	-.0100000	.0020022	-5.00	0.000	-.0098752	-.0020224
tenure	.0060000	.0006000	10.00	0.000	.0021002	.0033886
tenure2	-.0005000	.0001000	-5.00	0.000	-.0007000	-.0003000
_cons	1.764674	.0596079	29.60	0.000	1.647783	1.881564

(2) reg lhw age female femage

Source	SS	df	MS			
Model	50.0000000	3	18.8903724	Number of obs =	3006	
Residual	650.000000	3002	.289562779	F( 3, 2301) =	65.24	
Total	700.000000	3005	.313782583	Prob > F =	0.0000	
				R-squared =	0.0714	
				Adj R-squared =	0.0772	
				Root MSE =	.53811	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0084819	.0014144	6.00	0.000	.0057082	.0112556
female	-.049775	.082544	-0.60	0.547	-.2116433	.1120934
femage	-.0056949	.0020203	-2.82	0.005	-.0096568	-.0017331
_cons	1.762608	.0586584	30.05	0.000	1.647579	1.877637

i) After how many years of work in the current job will the (log of) wages be maximised?

$$d\ln w/dTENURE = .006 - 2(0.0005)TENURE$$

$$F.o.c. \max = 0 = .006 - .001TENURE$$

$$\text{so } .006 = 0.001TENURE \text{ and } TENURE = .006/0.001 = 6$$

*ie wages maximized after 6 years of experience*

ii) Test the hypothesis that the coefficients on TENURE and TENURE2 are jointly significant in the model

*F test of restriction that coefficients on tenure and tenure2 = 0 is given by*

$$F = \frac{RSS_{restrict} - RSS_{unrestrict} / j}{RSS_{unrestrict} / N - k_{unrestrict}} \sim F[j, N - k_{unrestrict}]$$

where  $j$  is number of restricted coefficients (IN THIS CASE  $J=2$ )  
so

$$F = \frac{(650 - 600) / 2}{600 / 3006 - 6} \sim F[2, 3000]$$

$F = 125 > F_{critical}$  at 95% level, (3.00),  
so **reject** null hypothesis that coefficients on tenure & tenure2 are zero

iii) What would be the effect for the OLS estimates in equation (1) of omitting the variable FEMALE from the regression?

*In this case can show OLS estimates of other coefficients will not be biased (since true effect is zero would expect on average the estimate to equal zero. If it does not then it is only the result of chance. Its presence in the model does not affect the bias of the other variables)*

*but will be inefficient, since in 3 variable model*

$$Var(\hat{\mathbf{b}}_1) = \frac{s^2}{N * Var(X)} * \frac{1}{1 - r_{X_1 X_2}^2} \neq \frac{s^2}{N * Var(X)}$$

*so including extra irrelevant variables has a cost in terms of larger standard errors (smaller  $t$ ,  $F$  values) than otherwise. – unless the variables are orthogonal in which case  $r_{x_1 x_2}^2 = 0$ . Unlikely in this particular example.*

iv) Outline how you would test the hypothesis that the specification of the variables on the right hand side of (1) were correct

*To test whether should have included extra variables (strictly higher order terms of the included variables) then do the Ramsey Regression Specification Error Test (RESET)*

*Given chosen model*

1) Estimate:  $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

2) save predicted (fitted) values :  $y = \hat{\mathbf{b}}_0 + \hat{\mathbf{b}}_1 X_1 + \hat{\mathbf{b}}_2 X_2$

*(predicted value is a weighted average of all the right hand side variables with weights given by size of coefficients)*

3) Add higher order powers of this predicted variable to the original equation

$$y = b_0 + b_1 X_1 + b_2 X_2 + y^{\wedge 2} + y^{\wedge 3} + \dots + y^{\wedge k} + v$$

higher orders of predicted value are weighted averages of higher orders of all the right hand side variables

(number of extra terms is arbitrary – should check robustness of result to variation in number)

4) F test for inclusion of these extra variables

5) Reject null of **no** functional form mis-specification if estimated  $F > F_{critical}$

2. Given the following model estimated with annual data over the period 1950-1999

$$GDP_t = b_0 + b_1 Investment_t + u_t \quad (1)$$

you suspect the presence of measurement error in the estimate of the annual level of GDP (measured in £billion).

$$ie \text{ } GDP_t^{observed} = GDP_t^{true} + e_t$$

where e is a (random) error term

i) Outline the consequences of this type of measurement error for OLS estimation

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

Sub. (2) into (1)

$$\begin{aligned} y - e &= b_0 + b_1 X + u \\ y &= b_0 + b_1 X + u + e \\ y &= b_0 + b_1 X + v \quad \text{where } v = u + e \end{aligned} \quad (3)$$

Ok to estimate (3) by OLS, since

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

(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

$$\text{True: } Var(\hat{\mathbf{b}}) = \frac{\mathbf{s}_u^2}{NVar(X)} \quad (A)$$

$$\text{Estimate: } \text{Var}(\tilde{\mathbf{b}}) = \frac{\mathbf{s}_u^2 + \mathbf{s}_e^2}{N\text{Var}(X)} \quad (B)$$

ii) given your answer to part i) and the following information, calculate the impact of measurement error in this case

$$\begin{aligned} \text{Var}(u) &= 2 & \text{Var}(e) &= 2 \\ \text{Var}(\text{GDP}^{\text{true}}) &= 0.2 & \text{Var}(\text{GDP}^{\text{observed}}) &= 0.5 \\ \text{Var}(\text{Investment}^{\text{true}}) &= 0.1 & \text{Var}(\text{Investment}^{\text{observed}}) &= 2 \\ \text{Cov}(\text{GDP}^{\text{true}}, \text{Investment}^{\text{true}}) &= 0.3 & \text{Cov}(\text{GDP}^{\text{true}}, \text{Investment}^{\text{observed}}) &= 0.2 \\ \text{Cov}(e, u) &= 0 & E(u) &= 0 & E(e) &= 0 \end{aligned}$$

OLS estimate of variance of coefficient estimate in absence of measurement error is

$$\hat{\text{Var}}(b_{\text{investment}}) = \frac{\text{var}(u)}{N * \text{Var}(\text{Investment}^{\text{true}})} = \frac{2}{50 * 0.1} = 0.4$$

OLS estimate of variance in absence of measurement error is

$$\hat{\text{Var}}(b_{\text{investment}}) = \frac{\text{var}(u) + \text{var}(e)}{N * \text{Var}(\text{Investment}^{\text{true}})} = \frac{2 + 2}{50 * 0.1} = 0.8$$

You are now given new information that says that it is the right hand side variable (Investment) that is instead measured with error

$$\text{ie} \quad \begin{aligned} \text{Investment}^{\text{observed}} &= \text{Investment}^{\text{true}} + w \\ \text{GDP}_t^{\text{observed}} &= \text{GDP}_t^{\text{true}} \end{aligned}$$

where w is a random error

iii) Find the true (unobserved) OLS estimate of the effect of investment on the level of GDP in the absence of measurement error

OLS estimate of slope effect in absence of measurement error

$$\hat{b}_{\text{Investment}^{\text{true}}} = \frac{\text{Cov}(\text{Investment}^{\text{true}}, \text{GDP}^{\text{true}})}{\text{Var}(\text{Investment}^{\text{true}})} = \frac{0.3}{0.1} = 3$$

d) the actual OLS estimate given this type of measurement error

OLS estimate of slope effect in presence of measurement error in age

$$\hat{b}_{\text{investment}^{\text{observed}}} = \frac{\text{Cov}(\text{Investment}^{\text{observed}}, \text{GDP}^{\text{true}})}{\text{Var}(\text{Investment}^{\text{observed}})} = \frac{0.2}{0.1} = 2$$

e) Why do the results change like this?

OLS estimates in the presence of measurement error on the right hand side are always biased toward zero (Attenuation Bias)

$$\hat{b}_1 = \frac{\text{Cov}(X^{\text{observed}}, y^{\text{true}})}{\text{Var}(X)} = b_1 + \frac{-b_1 \text{Var}(w)}{\text{Var}(X)} \neq b_1$$

if true  $b_1 > 0$  then  $\hat{b}_1^{\text{ols}} < b_1$

if true  $b_1 < 0$  then  $\hat{b}_1^{\text{ols}} > b_1$

ie closer to zero in both cases (means harder to reject any test that coefficient is zero)

3. Given the following house price and inflation equations

$$\text{House\_Prices}_t = a_0 + a_1 \text{Inflation}_t + u_t \quad (1)$$

$$\text{Inflation}_t = b_0 + b_1 \text{House\_Prices}_t + b_2 \text{Interest\_Rates}_t + b_3 \text{Money\_Supply}_t + e_t \quad (2)$$

a) What would happen if you estimated (1) or (2) by OLS and why?

house prices and inflation appear on both sides of respective equations and are **interdependent** since

Any shock, represented by  $Du$   $\otimes DH$  in (1)  
 but  $DH$   $\otimes DI$  from (2)  
 and  $DI$   $\otimes DH$  from (1)

so changes in H lead to changes in I **and** changes in I lead to changes in H

but the fact that  $Du$   $\otimes DI$  means  $\text{Cov}(X, u) = \text{Cov}(\text{Inflation}, u) \neq 0$  in (1)

which given OLS implies

$$\hat{b} = \frac{\text{Cov}(X, y)}{\text{Var}(X)} = b + \frac{\text{Cov}(X, u)}{\text{Var}(X)}$$

means  $E(\hat{b}) \neq b$

So OLS in the presence of interdependent variables gives biased estimates.

b) Find the order condition for identification of equations (1) and (2) and say which, if any, instruments, you would use in each case

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous variables excluded** from that equation is greater than or equal to the total number of **endogenous variables** in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

where  $K$  = Total no. of exogenous variables in the system

$k$  = No. of exogenous variables included in the equation

$m$  = No. of endogenous variables included in the equation

In (1)

$K = 2$  (Interest\_Rates, Money\_Supply)  
 $k = 0$   
 $m = 2$  (House\_Prices, Inflation)

so  $2 - 0 > 2 - 1$

equation is (over)identified – can find an instrument (Interest rates or money supply) for endogenous rhs variable

In (2)  
 $K = 2$  (Interest\_Rates, Money\_Supply)  
 $k = 2$  (Interest\_Rates, Money\_Supply)  
 $m = 2$  (House\_Prices, Inflation)

so  $2 - 2 < 2 - 1$

equation is NOT identified – can't find an instrument for endogenous rhs variable

c) What would be the most efficient solution in equation (1) ?

Since 1 equation is (over)identified – could use either instrument. However in large samples we know it is more efficient to use ALL the instruments, but in small samples better to use MINIMUM number of instruments. So answer depends on sample size

d) Outline the form of the test to use to check on the endogeneity of any one of the right hand side variables in an equation

*Wu-Hausman Test for Endogeneity*

1. Given  $y = b_0 + b_1X + u$  (A)

Regress  $X$  on the instrument(s)  $Z$

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

Save the residuals  $v$

2. Include this residual as an extra term in the original model ie estimate

$$y = b_0 + b_1X + b_2v + e$$

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  $u = b_2v + e$  )

4.

i) What do you understand by the term autocorrelation?

$Cov(u_t, u_{t-j}) \neq 0$  for  $j \neq 0$

Systematic relationship between residuals over time

ii) What can cause it?

Incorrect functional form; Interpolation of Data, Revisions to Data; Inertia in Economic Data (eg multiplier effects)

iii) What are the consequences for OLS estimation?

OLS remains unbiased (not affect  $Cov(X, u) = 0$  assumption) but standard errors on estimates are biased (upwards if positive autocorrelation)

Given the following information from a regression of the model

$$\text{Investment}_t = b_0 + b_1 \text{Investment}_{t-1} + b_2 \text{GNP}_t + b_3 \text{Interest\_rates}_t + u_t$$

```
. reg invest invest1 GNP interest
```

Source	SS	df	MS			
Model	1324.46636	3	441.488785	Number of obs =	29	
Residual	226.700416	25	9.06801664	F( 3, 25) =	48.69	
				Prob > F =	0.0000	
				R-squared =	0.8539	
				Adj R-squared =	0.8363	
Total	1551.16677	28	55.3988133	Root MSE =	3.0113	

  

invest	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
invest1	.4749651	.1736102	2.74	0.011	.1174081	.8325221
GNP	.4141392	.1509537	2.74	0.011	.1032442	.7250342
interest	-.2460615	.1169471	-2.10	0.046	-.4869186	-.0052043
_cons	5.412312	2.29867	2.35	0.027	.6781121	10.14651

Durbin-Watson Statistic = 1.589473.

Durbin-Watson h-statistic: .3989117 t = 1.405329

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df
1	2.298	1

H0: no serial correlation

test for the presence of 1<sup>st</sup> order autocorrelation in the residuals

From Tables, given  $T=29$  and  $K'=3$ ,  $DW_{low} = 1.20$  and  $DW_{high} = 1.65$   
( $k'$ =no. rhs variables **excluding** the constant)

So estimated value is in inconclusive region can **not** say anything about whether autocorrelation exists.

$h \sim \text{Normal}(0,1)$

So that  $\Pr[-1.96 \leq h \leq 1.96] = 0.95$  ie 95% chance that value of  $h$  will lie between  $-1.96$  and  $+1.96$  if null of no autocorrelation is true

In this case estimated  $h$  lies WITHIN confidence interval so can NOT reject null of no autocorrelation

Breusch-Godfrey test is  $(N-q) \cdot R^2_{aux} = 2.3$

Since this is a test of one lag then this statistic has a chi-squared distribution with 1 degree of freedom (equal to the number of lags tested)

From tables  $c^2_{critical}$  at 5% level = 3.84

So estimated  $c^2 < c^2_{critical}$ , so as before, can **NOT** null that residuals are correlated over one year to the next.

Which of the test statistics do you prefer and why?

Durbin Watson is biased toward 2 in presence of lagged dependent variable.

Breusch\_Pagan is only valid asymptotically. So in this case Durbin's  $h$  test is the only one can rely on.

Outline the Feasible GLS solution to the problem of autocorrelation

1. Can, in principle, manipulate the data to remove autocorrelation from the residuals.

Suppose you had

$$Y_t = b_0 + b_1 X_t + u_t \quad (1)$$

and **assumed** AR(1) behaviour in the residuals

$$u_t = r u_{t-1} + e_t \quad (2)$$

$$(1) \text{ } \mathcal{D} \quad Y_{t-1} = b_0 + b_1 X_{t-1} + u_{t-1} \quad (3)$$

(ie relationship holds in any time period)

Multiplying (3) by  $r$

$$r Y_{t-1} = r b_0 + r b_1 X_{t-1} + r u_{t-1} \quad (4)$$

(1) - (4)

$$Y_t - r Y_{t-1} = b_0 - r b_0 + b_1 X_t - r b_1 X_{t-1} + u_t - r u_{t-1}$$

or

$$Y_t = b_0 - r b_0 + r Y_{t-1} + b_1 X_t - r b_1 X_{t-1} + u_t - r u_{t-1}$$

or

$$Y_t = (b_0 - r b_0) + r Y_{t-1} + b_1 X_t - r b_1 X_{t-1} + e_t \quad (5)$$

$$Y_t - r Y_{t-1} = (b_0 - r b_0) + b_1 (X_t - r X_{t-1}) + e_t \quad (6)$$

Since  $e_t = u_t - r u_{t-1}$  from (2), then if estimate (5) by OLS there should be no autocorrelation. This is called Feasible Generalised Least Squares (FGLS)