1. You have data on years of work experience, EXPER, its square, EXPER2, years of education, EDUC, and the log of hourly wages, LWAGE

You estimate the following regressions:

(1) $\widehat{LWAGE} = 2.00 + 0.05*EDUC + 1.00*EXPE - 0.025*EXPER2$
(1.50) (0.25) (0.5) (0.005)

N= 104 ESS = 50 TSS = 100

(2) $\widehat{LWAGE} = 1.00 + 0.20*EDUC$
(0.50) (0.05)

N= 104 ESS = 30 TSS = 100

where the numbers in brackets are estimated standard errors

i) Comment on and interpret the results of equation (1).

(4 marks)

log-lin model so estimated coefficients are semi-elasticities
dLnw/dEduc = 0.05 so 1 extra year of education raises wages by 5%
standard error = 0.25 so t = 2 and variable is statistically significant at 5% level (given df = n-k = 104-4 =100)

experience is entered as a quadratic so effect of experience is non-linear – depends what level of experience individual has dLnw/dExp = 1 -2(0.025)Exp - both variables significant

R2 = ESS/TSS = 0.5 ie 50% of variation in log wages explained by model

ii) At how many years of experience are (the log of) wages maximised?

(3 marks)

dLnw/dExp = 1 -2(0.025)Exp

F.o.c. max = 0 = 1-.05Exp so 1=0.05Exp and Exp = 1/0.05 = 20

Ie wages maximized after 20 years of experience

ii) Test the hypothesis that the coefficients on EXPER and EXPER2 are jointly significant in the model

(5 marks)

F test of restriction that coefficients on exper and exper2 = 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

$$F = \frac{(70-50)/2}{50/100\ -4} \sim F[2{,}100]$$

F = 20 > Fcrtical at 95% level, so ***reject*** *null hypothesis that coefficients on exper & exper2 are zero*

iii) What would are the consequences for the OLS estimate on EDUC of omitting experience and experience squared from the regression?

(4 marks)

Omitted variable bias ie coefficient on education picks up (in part effect of missing variables

taking expectations (to get bias)

$$E(\hat{b}_1{}^{2\,\mathrm{var}}) = b_1 + \frac{b_2 Cov(X_1, X_2)}{\mathrm{Var}(X_1)} \neq b_1$$

So sign of bias depends on

a) the covariance between the variables, Cov(X_1, X_2)
b) the sign of the effect ß$_2$ of the extra variable, X_2, on y
(if ß$_2$= 0 shouldn't be in model in 1st place)

Also t and standard errors biased

iv) What would are the consequences for the OLS estimate on EDUC of including an irrelevant variable in (1)?

(3 marks)

In this case can show OLS estimate of ß$_1$ 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

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

so including extra irrelevant variables has a cost in terms of larger standard errors (smaller t, F values) than otherwise.

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

(6 marks)

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: $y = ß_0 + ß_1X_1 + ß_2X_2 + u$

2) save predicted (fitted) values : $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$

(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

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

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 over 100 individuals

$\text{Income}_i = b_0 + b_1 \text{Age}_i + u_i$ (1)

you suspect the presence of measurement error in the left hand side (dependent) variable on the level of income (measured in £).

ie $\text{Income}_i^{observed} = \text{Income}_i^{true} + e_i$

where e is a (random) error term

Given the following information

$Cov(Income^{true}, Age^{true}) = 5$ $Cov(Income^{true}, Age^{observed}) = 2$
$Var(Income^{true}) = 5$ $Var(Age^{true}) = 0.5$ $Var(Age^{observed}) = 1$
$Var(u) = 1$ $Var(e) = 1$
$Cov(e, u) = 0$ $E(u) = 0$ $E(e) = 0$

a) outline the consequences of this type of measurement error for OLS estimation

(4 marks)

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

Sub. (2) into (1)

$y - e = b_0 + b_1X + u$
$y = b_0 + b_1X + u + e$
$y = b_0 + b_1X + v$ *where* $v = u + e$ *(3)*

Ok to estimate (3) by OLS, since

$E(u) = E(e) = 0$
$Cov(X,u) = Cov(X,e) = 0$

(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

True: $Var(\hat{b}) = \frac{s_u^2}{NVar(X)}$ *(A)*

Estimate: $Var(\tilde{b}) = \frac{s_u^2 + s_e^2}{NVar(X)}$ *(B)*

b) given your answer to part b and the information above calculate the impact of measurement error in this example

(3 marks)

OLS estimate of variance in absence of measurement error is

$$Var(\hat{b}_{age}) = \frac{\text{var}(u)}{N * Var(Age^{true})} = \frac{1}{100 * 0.5} = 0.02$$

OLS estimate of variance in absence of measurement error is

$$Var(\hat{b}_{age}) = \frac{\text{var}(u) + \text{var}(e)}{N * Var(Age^{true})} = \frac{1+1}{100 * 0.5} = 0.04$$

You are given new information that says that it is the right hand side variable (age) that is instead measured with error
ie $Age^{observed} = Age^{true} + w$

where w is a random error

Find
c) the true (unobserved) OLS estimate of the effect of age on income expenditure and income in the absence of measurement error

(3 marks)

OLS estimate of slope effect in absence of measurement error

$$\hat{b}_{age}{}^{true} = \frac{Cov(Age^{true}, Income^{true})}{Var(Age^{true})} = \frac{5}{0.5} = 10$$

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

(3 marks)

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

$$\hat{b}_{age}{}^{observed} = \frac{Cov(Age^{observed}, Income^{true})}{Var(Age^{observed})} = \frac{2}{1} = 2$$

e) Why do the results change like this?

(4 marks)

OLS estimates are always biased toward zero (<u>Attenuation Bias</u>)

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

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

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

f) If measurement error is a problem among right hand side variables outline the details of a technique that could solve the problem.

(8 marks)

- 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) ? 0
2) Cov(Z,u) = 0

Instrumental variable (IV) estimation proceeds as follows:

Given a model

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

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

So $Cov(Z,y) = Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u)$
$= 0 + b_1Cov(Z,X) + 0$

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

So $b_1^{IV} = \dfrac{Cov(Z,y)}{Cov(Z,X)}$ *(compare with* $b_1^{OLS} = \dfrac{Cov(X,y)}{Var(X)}$ *)*

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

3. Given the following advertising expenditure (advert) and total sales (sales) equations estimated over 240 monthly observations

Sales: $Sales_t = b_0 + b_1Price_t + Advert_t + u_t$ (1)
Advertising: $Advert_t = a_0 + a_1Profits_t + a_2Sales_t + a_3Elasticity_t + e_t$ (2)

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

sales and advert appear on both sides of respective equations and are ***interdependent*** *since*

Any shock, represented by Du $\rightarrow DS$ *in (1)*
but $DS \rightarrow DA$ *from (2)*
and $DA \rightarrow DS$ *from (1)*

so changes in S lead to changes in A ***and*** *changes in A lead to changes in S*

but the fact that $Du \rightarrow DA$ *means* $Cov(X,u) = Cov(A,u) \neq 0$ *in (1)*

which given OLS implies

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

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)

(8 marks)

*"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."*

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

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 = 3 (price, profits, elasticity)
k= 1 (price)
m = 2 (sales, advert)

so $3 - 1 > 2 - 1$

equation is (over)identified – can find an instrument for endogenous rhs variable

In (1)
K = 3 (price, profits, elasticity)
k= 2 (profits, elasticity)
m = 2 (sales, advert)

so $3 - 2 = 2 - 1$

equation is just identified – can find an instrument for endogenous rhs variable

c) What instruments, if any, could you use for IV estimation of equation (1) ? Which would be the most efficient solution?

(5 marks)

Since 1 equation is (over)identified – can find an instrument for endogenous rhs variable advert exogenous variables that appear in equation (2) ie profits, elasticity - since correlated with advert but uncorrelated with sales since don't appear in (1)

Since 2 possible instruments, mist efficient solution is to use both and estimate by 2sls if sample size is large enough – it is (Otherwise better to use just 1 instrument).

d) Outline the form of the test to use to check on the validity (exogeneity) of any extra instruments you may have in

(8 marks)

One way to do this would be to compute two different 2SLS estimates, one using one instrument and another using the other instrument (rather like in the above example on prices, wages productivity and unemployment). If these estimates are radically different you might conclude that one (or both) of the instruments was invalid (not exogenous).

An implicit test of this – that avoids having to compute all of the possible IV estimates is based on the following idea

Given $y = b_0 + b_1X + u$ *and* $Cov(X,u) ? 0$

If an instrument Z is valid (exogenous) it is uncorrelated with u

To test this simply regress u on ***all*** *the possible instruments.*

$$u = d_0 + d_1Z_1 + d_2Z_2 + \ldots. d_kZ_k + v$$

If the instruments are exogenous they should be uncorrelated with u and so the coefficients $d_1 .. d_k$ *should all be zero (ie the Z variables have no explanatory power)*

Since u is never observed have to use a proxy for this which turns out to be the residual from the 2SLS estimation

$$\hat{u}^{2sls} = y - \hat{b}_0^{2sls} - \hat{b}_1^{2sls} X$$

(since this is a consistent estimate of the true unknown residuals)

So to Test Overidentifying Restrictions

1. *Estimate model by 2SLS and save the residuals*
2. *Regress these residuals on all the exogenous variables (including those X variables in the original equation that are not suspect) and save the* R^2
3. *Compute* $N*R^2$
4. *Under the null that all the instruments are uncorrelated then*
   $N*R^2 \sim ?^2$ *with L-k degrees of freedom*

(L is the number of instruments and k is the number of endogenous right hand side variables)

Note that can only do this test if there are more instruments than endogenous right hand side variables

4. You have quarterly data on an index of sterling's exchange rate against a basket of world currencies, e_t , and the level of interest rates, r_t, over the period 1960-1999 and fit the following:

(1) an ordinary least squares (OLS) regression of e_t on r_t
(2) An OLS regression of e_t on r_t and r lagged by one year, r_{t-1}
(3) A Prais-Winsten Feasible Generalised Least Squares (FGLS) regression of e_t on r_t

The table gives the estimated regression coefficients. Standard errors are given in brackets.

	OLS 1	OLS 2	FGLS
r_t	6	4	5.80
	(2)	(2)	(3.00)
r_{t-1}		1	
		(0.5)	
Durbin Watson	1.05	1.70	1.80

a) What is autocorrelation, what might cause autocorrelation and what are the consequences for OLS estimation?

(6 marks)

autocorrelation as signifying a systematic relationship between the residuals measured at different points in time

$u_t = r u_{t-1} + e_t$ *-1<= r <=1*

We know OLS estimation gives

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

and hence bias in OLS depends on whether Cov(X,u) = 0

but autocorrelation means $Cov(u_t, u_{t-1}) \neq 0$*, so OLS remains unbiased in presence of autocorrelation ie* $E(\hat{b}) = b$

but standard errors are biased so in general OLS will **underestimate** the true variance so the t values on the OLS estimates will be **larger** than should be so might conclude variables are statistically significant when they are not (type I error)

b) Does equation (1) suffer from autocorrelation?

(3 marks)

$DW < DW_{lower} = 1.44$ for k'= 1 T= 40
So conclude (positive) 1st order autocorrelation exists

c) Has specification (2) solved the problem? Why might you be suspicious of the test results from this equation?

(5 marks)

Now *$DW > DW_{upper} = 1.54$ for k'= 1 T= 40*
So appears (positive) 1st order autocorrelation no longer exists

d) Has specification (3) solved the problem? Why might you be suspicious of the test results from this equation?

(5 marks)

FGLS suggests $DW > DW_{upper} = 1.54$ for k'= 1 T= 40
So again appears (positive) 1st order autocorrelation no longer exists

But

Test is not valid in presence of endogenous rhs variables and it is highly likely that the level of interest rates is affected by the exchange rate – so there is interdependence between variables ie endogeneity.

e) Outline the form of a test for autocorrelation that is not affected by this problem

(6 marks)

Breusch-Godfrey Test for AR(q)
*This is in fact a general test for autocorrelation of **any** order (ie residuals may be correlated over more than one period)*

$$u_t = r_1 u_{t-1} + r_2 u_{t-2} + r_3 u_{t-3} + \ldots . r_q u_{t-q} + e_t$$

So test for no autocorrelation of order q amounts to test

$$H0: r_1 = r_2 = r_3 = \ldots r_q = 0$$

Do this as follows:
1. *Estimate original model*

$$Y_t = b_0 + b_1 X_t + u_t$$

Save residuals

2. *Regress estimated residuals on lagged values up to lag q* ***and*** *all the original RHS X variables*

$$\hat{u}_t = r_1 \hat{u}_{t-1} + r_2 \hat{u}_{t-2} + \ldots r_q \hat{u}_{t-1} + g_1 X_t + e_t$$

3. *Either compute the F test for the joint significance of the residuals*

$\hat{u}_{t-1} \ldots . \hat{u}_{t-q}$

and if $\hat{F} > F_{critical}$ ***reject*** *null of no q order autocorrelation*

or

compute $(N-q)*R^2_{auxillary} \sim c^2_{(q)}$

if estimated $c^2 > c^2_{critical}$ *again reject null of no q order A/c.*
(intuitively if lagged residuals are significant this gives a high R^2*)*

Useful test since

a) generalises to any order autocorrelation wish to test
b) is robust to inclusion of lagged dep. variables

But

1. *Since this is a test of joint significance may not be able to distinguish which lagged residual is important*
2. *Test is only valid asymptotic ally (ie in large samples)*