

Name _____

HEMIS No. _____

For Internal Students of
Royal Holloway

DO NOT TURN OVER UNTIL TOLD TO BEGIN

EC5040 : ECONOMETRICS

Mid-Term Examination No. 2

Time Allowed: 1 hour

Answer All 4 questions

STATISTICAL TABLES ARE PROVIDED

Silent non-programmable calculators may be used

PRINT YOUR NAME ON THE FRONT OF THIS TEST PAPER WHERE INDICATED

WRITE ALL YOUR ANSWERS (INCLUDING ROUGH WORKING) ON THIS TEST PAPER. THERE ARE EXTRA BLANK SHEETS TOWARD THE BACK OF THE PAPER

1. Given the general linear model

$$y = X_1 B_1 + X_2 B_2 + u$$

a) Show the effect of estimating the model

$$y = X_1 B_1 + v$$

on the bias of the OLS estimate of B_1 and its variance/covariance estimate (assume σ^2 is known)

(12 marks)

$$\hat{\beta}_1 = (X_1' X_1)^{-1} X_1' y = (X_1' X_1)^{-1} X_1' (X_1 \beta_1 + X_2 \beta_2 + u)$$

$$\hat{\beta}_1 = \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 + (X_1' X_1)^{-1} X_1' u$$

taking expectations

$$E(\hat{\beta}_1) = \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2 \neq \beta_1$$

Unless X_1 and X_2 are orthogonal, $X_1' X_2 = 0$, estimates in omitted variable equation are biased

$$\text{Given } X' X = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \begin{bmatrix} X_1 & X_2 \end{bmatrix} = \begin{bmatrix} X_1' X_1 & X_1' X_2 \\ X_2' X_1 & X_2' X_2 \end{bmatrix}$$

Using rules on partitioned matrices

$$\sigma_u^2 (X_1' X_1)^{-1} = \sigma_u^2 [X_1' X_1 - X_1' X_2 (X_2' X_2)^{-1} X_2' X_1]^{-1} \quad (1)$$

compare with variance estimated in case of omitted variables

$$\sigma_u^2 (X_1' X_1)^{-1} \quad (2)$$

so (1) < (2)

and omitted variable estimates have smaller variance

The more highly correlated X_1 and X_2 the greater the difference

b) You decide to run a simple regression of log hourly pay on years of work experience

```
reg lhw exper
```

(A)

Source	SS	df	MS			
Model	52.3263364	1	52.3263364	Number of obs =	17321	
Residual	6113.83639	17319	.353013245	F(1, 17319) =	148.23	
Total	6166.16272	17320	.356014014	Prob > F =	0.0000	
				R-squared =	0.0085	
				Adj R-squared =	0.0084	
				Root MSE =	.59415	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper			12.17	0.000		
_cons	1.857851	.0092112	201.70	0.000	1.839796	1.875905

(some of the regression output has been concealed)

and then decide to include a dummy variable for whether the individual is a graduate

```
reg lhw exper grad
```

(B)

Source	SS	df	MS			
Model	854.145433	2	427.072717	Number of obs =	17321	
Residual	5312.01729	17318	.306733878	F(2, 17318) =	1392.32	
Total	6166.16272	17320	.356014014	Prob > F =	0.0000	
				R-squared =	0.1385	
				Adj R-squared =	0.1384	
				Root MSE =	.55384	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	.0079872	.0003538	22.57	0.000	.0072936	.0086807
grad	.6144253	.0120174	51.13	0.000	.59087	.6379807
_cons	1.691547	.0091817	184.23	0.000	1.67355	1.709544

Given the following regression of graduate status on experience

reg grad exper

(C)

Source	SS	df	MS			
Model	80.2003462	1	80.2003462	Number of obs =	17321	
Residual	2123.91997	17319	.122635254	F(1, 17319) =	653.97	
				Prob > F =	0.0000	
				R-squared =	0.0364	
				Adj R-squared =	0.0363	
Total	2204.12032	17320	.127258679	Root MSE =	.35019	

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0056162	.0002196	-25.57	0.000	-.0060467	-.0051857
_cons	.2706645	.0054291	49.85	0.000	.2600229	.2813061

Work out the estimated OLS coefficient on experience in the simple 2 variable model (A)

The algebra of omitted variables tells us that

$$E(\hat{\beta}_{exper}^{2\text{var model}}) = \hat{\beta}_{exper}^{3\text{var model}} + (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{grad} \quad (1)$$

so that the OLS estimate of experience in the 2 variable model equals the OLS coefficient on experience in the full model plus a correction factor which is equal to the coefficient from a regression of graduate status on experience, $(X_1'X_1)^{-1}X_1'X_2$, multiplied by the OLS coefficient on graduate in the full model

Can test this by regressing graduate on experience

reg grad exper

Source	SS	df	MS			
Model	80.2003462	1	80.2003462	Number of obs =	17321	
Residual	2123.91997	17319	.122635254	F(1, 17319) =	653.97	
				Prob > F =	0.0000	
				R-squared =	0.0364	
				Adj R-squared =	0.0363	
Total	2204.12032	17320	.127258679	Root MSE =	.35019	

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0056162	.0002196	-25.57	0.000	-.0060467	-.0051857
_cons	.2706645	.0054291	49.85	0.000	.2600229	.2813061

and using (1) above to give

display 0.0079872+(-.0056162*.61444253)

.00453637

which is the coefficient on experience in the 2 variable model.

(1) also explains why the coefficient on experience in the unrestricted model is more positive. a) Experience and graduate status are negatively correlated (there are relatively more graduates among younger workers) - see the regression

coefficient on experience in the auxiliary regression above b) graduates earn more. The product of these two effects is negative. Not controlling for both these effects exerts a downward bias on experience in the restricted model.

Can test this by regressing graduate on experience

reg grad exper

Source	SS	df	MS			
Model	80.2003462	1	80.2003462	Number of obs =	17321	
Residual	2123.91997	17319	.122635254	F(1, 17319) =	653.97	
				Prob > F =	0.0000	
				R-squared =	0.0364	
				Adj R-squared =	0.0363	
Total	2204.12032	17320	.127258679	Root MSE =	.35019	

grad	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
exper	-.0056162	.0002196	-25.57	0.000	-.0060467	-.0051857
_cons	.2706645	.0054291	49.85	0.000	.2600229	.2813061

and using (1) above to give

display 0.0079872+(-.0056162*.61444253)

.00453637

which is the coefficient on experience in the 2 variable model.

(1) also explains why the coefficient on experience in the unrestricted model is more positive. a) Experience and graduate status are negatively correlated (there are relatively more graduates among younger workers) - see the regression coefficient on experience in the auxiliary regression above b) graduates earn more. The product of these two effects is negative. Not controlling for both these effects exerts a downward bias on experience in the restricted model.

c)

LM test

$$N \cdot R^2_{\text{auxiliary}} \sim \chi^2_{\text{no. extra variables}}$$

Where $R^2_{\text{auxiliary}}$ is R^2 from a regression of the residuals from the restricted regression $y = X_1 B_1 + u$ on a set of additional variables X_2

If estimated $\chi^2 > \chi^2_{\text{critical}}$ reject restricted regression. There are relevant variables not in the restricted model.

2. Given the general linear model

$$y = X\beta + u$$

you suspect that $\text{Var}(u_i/X_i) = E(u_i^2/X_i) \neq \text{constant}$

a) What are the consequences for OLS estimation of the B vector and its associated variance/covariance matrix? (10 marks)

Heteroskedasticity exists, so

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) = \beta + (X'X)^{-1}X'u$$

Taking expectations

$$E\left[\hat{\beta}_{OLS}\right] = \beta \quad \text{so OLS estimates remain unbiased in presence of heteroskedasticity}$$

but

$$\text{Var}\left(\hat{\beta}_{OLS}\right) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$$

$$\text{Var}\left(\hat{\beta}_{OLS}\right) = \sigma^2[(X'X)^{-1}X'\Omega X(X'X)^{-1}] \neq \sigma^2(X'X)^{-1}$$

so standard errors in OLS based on latter are biased (in an unknown direction which depends on Ω) and since t and F values also calculated using latter they are also biased id use OLS.

b) Given grouped data and a model of the form

$$y_g = X_g\beta + u_g \quad g = 1, 2, \dots, G \text{ groups}$$

write down the form of the residual variance/covariance matrix in this case and hence the form of the Feasible GLS estimator in this case

(6 marks)

Since $\text{Var}(u_g) = \sigma^2/n_g$

then the $G \times G$ residual variance covariance matrix $\sigma^2\Omega = \sigma^2$

$$\begin{bmatrix} 1/N_1 & & 0 \\ & \dots & \\ 0 & & 1/N_G \end{bmatrix}$$

and so using rules on inverse of diagonal matrix

$$\Omega^{-1} = \begin{bmatrix} N_1 & & 0 \\ & \dots & \\ 0 & & N_G \end{bmatrix}$$

Hence given

$$\hat{\beta}_{GLS} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} y = \left[\sum_{g=1}^G N_g x_g x_g' \right]^{-1} \left[\sum_{g=1}^G N_g x_g y_g \right]$$

where x_g' is the g^{th} row of X and y_g is the g^{th} element of the y vector

c) What does this imply about how to transform the data in the original model in order to obtain this Feasible GLS estimator

(3 marks)

To carry out GLS estimation all have to do is multiply each element of the x_g row of the X matrix and the y_g element by the square root of N_g (weighted least squares)

d) What is the White robustness correction?

(3 marks)

Since exact form of heteroskedasticity is often unknown may be better to fix up OLS standard errors to presence of heteroskedasticity

Given OLS residuals \hat{u} then can show $S = \sum_i \hat{u}_i x_i x_i'$ is a consistent estimator of

$$\sigma^2 (X' \Omega X) = \sum_i \sigma_i^2 x_i x_i', \text{ so replace}$$

$$\text{Var} \left(\hat{\beta}_{OLS} \right) = \sigma^2 [(X' X)^{-1} X' \Omega X (X' X)^{-1}]$$

with

$$\text{Var} \left(\hat{\beta}_{OLS}^{robust} \right) = \sigma^2 (X' X)^{-1} S (X' X)^{-1}$$

3. Given the following model, (in mean deviation form), you suspect that the variable, x_1 , is measured with error and so the OLS estimate of b suffers from attenuation bias

$$y_i = b x_1 + u_i \quad (1)$$

Given a possible instrument x_2 , let $W = [y : x_1 : x_2]$ and

$$W'W = \begin{bmatrix} 175 & 9 & 10 \\ 9 & 3 & 2 \\ 10 & 2 & 5 \end{bmatrix}$$

the sample size is 100

a) Find the OLS and IV estimates of the coefficient on x_1

(6 marks)

OLS estimates given by $(x'x)^{-1}x'y$ where $x = [x_1]$

IV estimates given by $(z'x)^{-1}z'y$ where $x = [x_1]$ and $z = [x_2]$ so in this case

Given

$$W'W = \begin{bmatrix} y'y & y'x_1 & y'x_2 \\ x_1'y & x_1'x_1 & x_1'x_2 \\ x_2'y & x_2'x_1 & x_2'x_2 \end{bmatrix}$$

So $b_{ols} = 9/3 = 3$

And $b_{iv} = 10/2 = 5$

(as would expect in presence of measurement error OLS estimates suffer from attenuation bias (closer to zero))

b) The estimated residual variance from the IV estimation and hence the standard error and statistical significance of the IV estimate of b

(10 marks)

$$\begin{aligned} \text{Need } s^2_{IV} &= u_{IV}' u_{IV} / n \\ &= (y - x_1 b_{IV})' (y - x_1 b_{IV}) / n \\ &= (y'y - 2b_{IV} y'x_1 + b_{IV}' x_1' x_1 b_{IV}) / n \end{aligned}$$

From matrix and answer to above

$$s^2_{IV} = (175 - (2*5*9) + 5(3)5) / 100 = 160/100 = 1.6$$

$$\begin{aligned} \text{So } \text{Var}(b_{IV}) &= s^2_{IV} (Z'X)^{-1} (Z'Z) (Z'X)^{-1} \\ &= s^2_{IV} (x_1'x_2)^{-1} x_2'x_2 (x_1'x_2)^{-1} \\ &= 1.6(1/2)5(1/2) = 2 \end{aligned}$$

$$\text{So } SE(b_{IV}) = \sqrt{2} = 1.41$$

(Hence IV estimate of b is statistically significant at 5% level, since $t=5/1.41 = 3.54$)

c) What do you understand by the term reliability ratio?

(3 marks)

The reliability ratio = $\frac{\sigma_{xt}^2}{\sigma_{xt}^2 + \sigma_u^2} = 1 - \frac{\sigma_u^2}{\sigma_{xt}^2 + \sigma_u^2}$ = which can

be derived from the equation for the degree of attenuation bias of OLS in the presence of measurement error is the proportion of the variation in the unobserved variable that can be explained by the variation in the observed variables. The closer this ratio to 1 the less measurement error (noise) and the greater the "signal" in the observed variable.

d) Show how you could estimate the reliability ratio if you had 2 independent measures of the variable x_1

(6 marks)

Given two alternative proxies for x_1 : x_a and x_b measured with error such that

$$x_a = x^{\text{true}} + e \quad \text{and} \quad x_b = x^{\text{true}} + v$$

Then the correlation coefficient

$$r(x_a, x_b) = \frac{\text{Cov}(x_a, x_b)}{\sqrt{\text{Var}(x_a)\text{Var}(x_b)}} = \frac{\text{Cov}(x^{\text{true}} + e, x^{\text{true}} + v)}{\sqrt{\text{Var}(x^{\text{true}} + e)\text{Var}(x^{\text{true}} + v)}}$$

which since e and v are independent implies

$$r(x_a, x_b) = \frac{\text{Var}(x^{\text{true}})}{\sqrt{\text{Var}(x^{\text{true}}) + \text{var}(e) + \text{Var}(x^{\text{true}}) + \text{var}(v)}} \cong \frac{\sigma_{xt}^2}{\sigma_{xt}^2 + \sigma_u^2}$$

and identical if $\text{var}(e) = \text{var}(v) = \sigma_u^2$

so correlation coefficient between the two mismeasured proxies equals the reliability ratio

4.

a) Show that the IV estimator is a consistent estimator of β in

$$y = X\beta + u$$

$$\text{if } \text{plim}[(1/n)(X'u)] \neq 0$$

(10 marks)

$$\hat{\beta}_{IV} = [(X'Z)(Z'Z)^{-1}(Z'X)]^{-1} X'Z(Z'Z)^{-1} Z'y = (X'P_Z X)^{-1} X'P_Z y$$

where $P_Z = Z(Z'Z)^{-1}Z'$

Hence

$$p \lim(\hat{\beta}_{IV}) = p \lim\left(\frac{(X'P_Z X)^{-1} X'P_Z y}{N}\right) = p \lim\left(\frac{(X'P_Z X)^{-1} X'P_Z (X\beta + u)}{N}\right)$$

$$p \lim(\hat{\beta}_{IV}) = \beta + p \lim\left(\frac{(X'P_Z X)^{-1}}{N}\right) p \lim\left(\frac{X'P_Z u}{N}\right)$$

since

$$p \lim\left(\frac{(X'P_Z X)^{-1}}{N}\right) = \left[p \lim\left(\frac{X'Z}{N}\right) p \lim\left(\frac{Z'Z}{N}\right)^{-1} p \lim\left(\frac{Z'X}{N}\right) \right]^{-1}$$

all these terms (sample averages) are finite ie will not converge to zero as sample size increases

However

$$p \lim\left(\frac{X'P_Z u}{N}\right) = \left[p \lim\left(\frac{X'Z}{N}\right) p \lim\left(\frac{Z'Z}{N}\right)^{-1} p \lim\left(\frac{Z'u}{N}\right) \right]$$

and final term will converge to zero as $N \rightarrow \infty$ by assumption needed for instrumental variables

Hence

$$p \lim(\hat{\beta}_{IV}) = \beta + p \lim\left(\frac{(X'P_Z X)^{-1}}{N}\right) * 0 = \beta$$

and IV is a consistent estimator

b) What are the problems caused by weak instruments for IV estimation? (6 marks)

In 2 variable model

$$\hat{\beta}_{IV} = \frac{\sum zy}{\sum zx} = \beta + \frac{\sum zu}{\sum zx} = \beta + \frac{(1/N)\sum zu}{(1/N)\sum zx}$$

so

$$p \lim(\hat{\beta}_{IV}) = \beta + \frac{Cov(zu)}{Cov(zx)} = \beta + \frac{Corr(zu) * se(u)}{Corr(zx) * se(x)}$$

weak instrument means low correlation and hence IV estimate may be long way from true value even in large samples

compare with

$$p \lim(\hat{\beta}_{OLS}) = \beta + \frac{Cov(xu)}{Var(x)} = \beta + \frac{Corr(xu) * se(u)}{se(x)}$$

so not necessarily better to use IV rather than OLS if

$$\frac{Corr(zu)}{Corr(zx)} > Corr(xu)$$

c) How could you test for the presence of a weak instrument in your data? (4 marks)

1st stage of 2sls will automatically give significance of instrument. In case of single endogenous variable can show

$$p \lim(\hat{\beta}_{IV}) = \beta + \frac{p \lim(\hat{\beta}_{OLS}) - \beta}{E(F) - 1}$$

where $E(F)$ is the F value of goodness of fit in the 1st stage of the 2sls regression

so if $E(F) = 10$ then size of bias of IV wrt OLS is $1/9$ which is deemed small enough to be acceptable (and compensate for higher standard errors in IV)

d) Outline the form of a test of the validity of the overidentifying restrictions in your instrument set.

(5 marks)

either Hausman test of subset of instruments against full instrument set

or

- 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)*

\hat{u}_{2sls}

$$u = d_0 + b_1X_1 + d_1Z_1 + d_2Z_2 + \dots d_1Z_1 + v$$

and save the R^2

- 3. Compute $N \cdot R^2$*
- 4. Under the null that all the instruments are uncorrelated then $N \cdot R^2 \sim \chi^2$ with $L-k$ degrees of freedom*

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