

Mid-Term 3 2004/05 – Outline Answers

1 a) Consider the following equations

$$i) u_i = Y_i - \beta_1 - \beta_2 X_i$$

$$ii) Y_i = \mathbf{b}_1 + \hat{\mathbf{b}}_2 X_i + \hat{u}_i$$

$$iii) \hat{Y}_i = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 X_i + u_i$$

where $\hat{}$ indicates an estimate otherwise the variable is the true value

Say, with reason, whether each equation is true or false.

i) *True: actual residual is difference between actual Y and fitted based on true value*

fitted value equals $\mathbf{b}_1 + \mathbf{b}_2 X_i$

ii) *False: Actual y value is sum of predicted value and predicted residual. In above uses combination of predicted and actual coefficients*

iii) *False: Predicted value equals $\hat{Y}_i = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 X_i$ and does not involve residual*

b) See lecture notes.

i) *Model is linear in parameters*

ie can be written in form $y = \beta_0 + \beta_1 X + u$ where coefficients appear in additive (rather than multiplicative form as eg $y = \beta_0 X^{\beta_1}$)

ii) *$E(u_i) = 0$ average value of true (unobserved) residuals is zero – nothing systematic*

iii) *$\text{Var}(u_i) = E(u_i^2/X_i) = \text{constant}$*

ie variance of residuals is constant for any value of X (homoskedasticity)

iv) *$\text{Cov}(u_i, u_j) = 0$ for all i not equal to j (zero autocorrelation)*

value of one residual gives no information about value of another

iv) *No covariance between X and true residual $\text{Cov}(X, u) = 0$*

c) “bias”

This means that we would like the expected, or average, value of the estimator to equal the true (unknown) value for the population of interest

$$E(\hat{\mathbf{b}}) = \mathbf{b}$$

Measure efficiency of any estimate by its dispersion –

- based on the variance (or more usually its square root – the standard error) Given 2 (unbiased) estimates will prefer the one whose range of estimates are more concentrated around the true value.

Gauss-Markov Theorem

Given Gauss-Markov assumptions 1-4 (see above) hold, then can prove a very important result:

that OLS estimates will have the smallest variance of all (linear) unbiased estimators

- there may be other ways of obtaining unbiased estimates, but OLS estimates will have the smallest standard errors

2. Given the following regression output

$$\hat{Investment} = 20000 - 1.0 * Interest_rates$$

$$R^2=0.90 \quad TSS = 1000 \quad ESS = 885 \quad Var(Interest_rate)=0.8$$

estimated over the period 1980-2004

i) Using the information above, find the standard error on the estimate of interest_rates in model I

$$Var(\hat{\mathbf{b}}) = \frac{s_u^2}{NVar(X)}$$

$$where \ s_u^2 = RSS/N-k$$

$$so \ s^2 = 115/25-2 = 5$$

$$Var(\hat{\mathbf{b}}) = \frac{5}{25 * 0.8} = 0.25 \quad so \ standard \ error (square \ root \ of \ variance) \ is \ 0.5$$

ii) Test the hypothesis that the true effect of interest rates on investment is zero

$$t \ value \ is \ based \ on \ null \ hypothesis \ that \ b^{null} = 0, \ so \ t = \frac{\hat{\mathbf{b}} - \mathbf{b}^0}{SE(\hat{\mathbf{b}})} \ becomes$$

$$t = -1.0 - 0 / 0.5 = -2.0$$

Critical values from t tables at 10% and 5% level given degrees of freedom

$$N-k = 25-2 = 23$$

are 1.71 and 2.07 respectively

In this case absolute value of $\hat{t} > t_{critical}^{10\%}$ but $< t_{critical}^{5\%}$ so can't reject null that coefficient is zero at 5% level ie it seems interest rates don't affect investment very strongly

95% confidence interval for each individual forecast observation given by

$$Pr[\hat{\mathbf{b}}_1 - t_{.05/2} SE(\hat{\mathbf{b}}_1) \leq \mathbf{b}_1 \leq \hat{\mathbf{b}}_1 + t_{.05/2} SE(\hat{\mathbf{b}}_1)] = 1 - .05$$

$$= \hat{\mathbf{b}}_1 \pm t_{.05/2} SE(\hat{\mathbf{b}}_1) \quad = -1.0 \pm 2.07 * 0.5$$

So can be 95% confident true slope value lies in region -2.035 to 0.035

(sample size is too small and standard errors on variables too large to make accurate estimates).

The size of a test (α) is the level of probability $0 < \alpha < 1$ used to construct statistical tests such that can use this to construct confidence intervals of the form

$$Pr[-d \leq z \leq d] = 1 - \alpha$$

which says that the probability of a value drawn from a distribution lying between the values $\pm d$ is $1 - \alpha$ %

By **reducing** the size of the test we **increase** the acceptance region for a given t estimate (and reduce the range of estimate that fall in the rejection region)

The danger of this is that increase the chance that the null hypothesis could be false, but that we accept it.

This called Type II error

By **increasing** the size of the test (α) we **reduce** the acceptance region for a given t estimate (and increase the range of estimate that fall in the rejection region)

The danger of this is that increase the chance that reject the null hypothesis even though it is true.

This called Type I error

vi) From tables must lie between 1.71 and 2.07 which are 10% and 50% critical values so p is around 6%

3. Interpret the meaning of the OLS estimates of the slope in each of the following time series regressions of consumption on income, (where both consumption and income are measured in £million). Some of the regression output has been hidden.

i) reg cons income

Source	SS	df	MS	Number of obs = 45		
Model	12000.0					
Residual	3000.0			R-squared = 0.80		
Total	15000.0			Adj R-squared = 0.82		
				Root MSE =		
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.9172948	.0118722	77.26	0.000	.8933523	.9412372
_cons	13.49616	4.025455	3.35	0.002	5.378055	21.61426

*Answer: Linear model so slope coefficient $\beta_{\text{income}} = d\text{Cons}/d\text{Income}$
 = change in cons unit change (ie £1 million in this case) in income.
 (or equivalently change in $d\text{Cons} = \beta_{\text{income}} * d\text{Income}$
 $= 0.917 * 1$*

*So £1million extra aggregate income leads to £917,000 **increase** in aggregate consumption, on average.*

. reg LnCons LnIncome

Source	SS	df	MS	Number of obs = 45		
Model	4.95869491		4.95869491	F(,) = 8575.43		
Residual	.024864517		.000578245	Prob > F = 0.0000		
Total	4.98355943		.113262714	R-squared = 0.9950		
				Adj R-squared = 0.9949		
				Root MSE = .02405		
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LnIncome	.9448621	.0102033	92.60	0.000	.9242852	.965439
_cons	.2790579	.0583503	4.78	0.000	.1613834	.3967325

*Log-linear (double log) – don't confuse with log-lin: so slope coefficient is an elasticity:
 $d\text{Ln}(\text{cons})/d\text{Ln}(\text{Income})$*

*= % change in consumption wrt % change in income.
 or equivalently*

$$(\% \text{ change in cons}/100) = \beta_{\log(\text{income})} * (\% \text{ change in income}/100)$$

$$= 0.2 * 1$$

(Since both sides divide by 100, this cancels)

So 1% increase in aggregate income increases consumption by 0.95%.

iii) reg LnCons income

Source	SS	df	MS	Number of obs = 45		
Model	4.86481426		4.86481426	F(,) = 1761.65		
				Prob > F = 0.00		

Residual		.118745167	.002761516		R-squared	=	0.97

Total		4.98355943	.113262714		Adj R-squared	=	0.97

					Root MSE	=	.052

LnCons		Coef.	Std.8Err.	t	P> t	[95% Conf. Interval]	

income		.0029489	.0000703	41.97	0.000	.0028072	.0030906
_cons		4.728025	.0238226	198.47	0.000	4.679983	4.776068

Log-lin model, so slope coefficient $\beta_{\text{income}} = d\ln(\text{cons})/d\text{Income}$

= (% change in consumption/100) / unit change in income (ie £1million in this case)

*(or equivalently (% change in consumption/100) = $\beta_{\text{income}} * \text{unit change in income}$
= 0.003*1*

*So £1million extra income leads to a **0.3 percent increase** in aggregate consumption, on average.*

b)

$$\text{Use } F = \frac{R^2/k - 1}{(1 - R^2)/(N - k)} \sim F[k-1, N-k] = (0.80/1)/(1-0.80)/45-2 \sim F[1, 43]$$

$$= 172$$

^

So $F > F_{\text{critical}} = 4.06$

Hence reject null that model as a whole has no explanatory power (R^2 so large unlikely to have arisen by chance)

c) Do the Chow forecast test of the hypothesis that the OLS coefficients estimated over the shorter period predict well out of sample.

If a model forecasts well out of sample then we would expect all the out-of-sample residuals to be close to zero. It can be shown that the joint test of all the out-of-sample-residuals being close to zero is given by:

$$F = \frac{RSS_{\text{in+out}} - RSS_{\text{in}} / N_o}{RSS_{\text{in}} / N - k} \sim F[N_o, N - k]$$

where N_o is the number of out-of-sample observations

N is the number of in-sample observations

k is the number of RHS coefficients

Use fact that $TSS = ESS + RSS$ so $RSS = TSS - ESS$

$$F = \frac{3000 - 1000/10}{1000/35 - 2} \sim F[10, 33] = 6.6$$

From tables $F_{critical}^{.05} [10,33] = 2.14$

^

So $F > F_{critical}$ and therefore **reject** null that model predicts well out of sample.
(Difference between RSS in 2 regressions is so large that unlikely to have occurred by chance. Large residuals mean model is not a good predictor)

d) We know the formula for the correlation coefficient

$$r_{Y, \hat{Y}} = \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}}$$

Which since $y = \hat{y} + u$

$$r_{Y, \hat{Y}} = \frac{\text{Cov}([\hat{Y} + u], \hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}}$$

Multiplying through terms

$$r_{Y, \hat{Y}} = \frac{\text{Cov}(\hat{Y}, \hat{Y}) + \text{Cov}(u, \hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}}$$

Since $\text{Cov}(y, u) = 0$

and $\text{Cov}(y, y) = \text{Var}(y)$

$$r_{Y, \hat{Y}} = \frac{\text{Var}(\hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}}$$

$$\text{Now since } \frac{\text{Var}(\hat{Y})}{\sqrt{\text{Var}(Y) \text{Var}(\hat{Y})}} = \sqrt{\frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}}$$

then

$$\begin{aligned} r_{Y, \hat{Y}} &= \frac{\sqrt{\text{Var}(\hat{Y})}}{\sqrt{\text{Var}(Y)}} \\ &= \sqrt{R^2} \end{aligned}$$

4. The following output is taken from regressions of a) the annual employment growth rate (measured in % points) b) the log of annual employment growth rate on annual gdp growth rate (measured in % points) using a sample of 20 countries

a) reg empl gdp

Source	SS	df	MS	Number of obs = 20		
Model	8.31647759	1	8.31647759	F(1, 18)	=	25.56
Residual	5.85581708	18	.325323171	Prob > F	=	0.0001
-----				R-squared	=	0.5868
Total	14.1722947	19	.745910246	Adj R-squared	=	0.5639
-----				Root MSE	=	.57037
empl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.3970308	.0785257	5.06	0.000	.2320544	.5620073
_cons	-.1091787	.272871	-0.40	0.694	-.6824595	.464102

b) reg lnempl gdp

Source	SS	df	MS	Number of obs = 20		
Model	6.92550263	1	6.92550263	F(1, 18)	=	5.71
Residual	21.8251244	18	1.21250691	Prob > F	=	0.0280
-----				R-squared	=	0.2409
Total	28.750627	19	1.5131909	Adj R-squared	=	0.1987
-----				Root MSE	=	1.1011
lnempl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.3623099	.1515991	2.39	0.028	.043812	.6808078
_cons	-1.467879	.5267954	-2.79	0.012	-2.574635	-.3611229

i) Interpret the estimated effect of gdp in both regressions

In 1st equation 1%point (not 1%) rise in gdp growth rate raises employment growth rate by 0.39 % points

2nd equation is log-lin (semi-log model) so 1%point rise in gdp raises employment growth rate by 36%

ii) Write down the equations needed to calculate a) the arithmetic mean b) the geometric mean

$$\text{arithmetic mean} = \frac{\sum_{i=1}^N x_i}{N}$$

$$\text{geometric mean} = (y_1 * y_2 * \dots * y_N)^{1/N}$$

iii) The following is output from regressions of employment growth and the log of employment growth divided by the geometric mean, (empladj and lempladj respectively). Do the Box-Cox test of the null hypothesis that the RSS from the two regressions are the same. Which specification do you prefer?

(4 marks)

```
. reg empladj gdp
```

Source	SS	df	MS	Number of obs = 16		
Model	16.2996276	1	16.2996276	F(1, 14)	=	23.25
Residual	9.81551482	14	.701108202	Prob > F	=	0.0003
-----				R-squared	=	0.6241
Total	26.1151424	15	1.7410095	Adj R-squared	=	0.5973
-----				Root MSE	=	.83732
empladj	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.5962491	.1236606	4.82	0.000	.3310234	.8614747
_cons	-.3921257	.450953	-0.87	0.399	-1.359324	.5750723

```
. reg lempladj gdp
```

Source	SS	df	MS	Number of obs = 16		
Model	7.77622121	1	7.77622121	F(1, 14)	=	5.73
Residual	19.0038339	14	1.35741671	Prob > F	=	0.0313
-----				R-squared	=	0.2904
Total	28.7500552	15	1.78533701	Adj R-squared	=	0.2397
-----				Root MSE	=	1.1651
lemladj	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gdp	.4118349	.1720662	2.39	0.031	.0427897	.7808801
_cons	-1.448807	.6274735	-2.31	0.037	-2.794603	-.1030099

Problem is that TSS from original 2 equations are not comparable ($Var(\log(\text{empl}))$) ?
 $Var(\text{empl})$ so can't use R^2 Idea is to transform data – by dividing through by the geometric mean

Formal test of significant difference between the 2 specifications

$$= N/2 \log(RSS_{\text{largest}}/RSS_{\text{smallest}}) \sim \chi^2_{(1)}$$

$$= (16/2) * \log(19.0/9.8) = 5.30$$

$$\hat{c}^2 > c^2$$

Given test is Chi-Squared with 1 degree of freedom. Estimated $\hat{c}^2 > c^2$ critical (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) so models are significantly different in terms of goodness of fit.

Since levels has lowest RSS prefer 1st equation

iv) The following output is taken from the distribution of OLS residuals from the two regressions, resa and resb respectively

```
. su resa, det
```

Percentiles		Residuals	
		Smallest	
1%	-.8867749	-.8867749	
5%	-.800374	-.7139732	
10%	-.6933984	-.6728235	Obs 20
25%	-.4225267	-.6685373	Sum of Wgt. 20
50%	-.0230367		Mean -2.00e-09

		Largest	Std. Dev.	.5551584
75%	.471398	.5822049		
90%	.7580767	.6624828	Variance	.3082009
95%	.8717977	.8536705	Skewness	.600000
99%	.8899248	.8899248	Kurtosis	1.200000

. su resb, det

		Residuals			
Percentiles		Smallest			
1%	-3.17601	-3.17601		Obs	20
5%	-2.60786	-2.039709		Sum of Wgt.	20
10%	-1.345669	-.6516281		Mean	1.12e-08
25%	-.4761055	-.6334555		Std. Dev.	1.07177
50%	.3739764			Variance	1.148691
		Largest		Skewness	-3.000000
75%	.7360109	.8852508		Kurtosis	6.000000
90%	.9456291	.8893937			
95%	1.032405	1.001864			
99%	1.062945	1.062945			

What is skewness?

*Symmetry in a distribution is represented by a value of zero for the skewness coefficient
Right Right skewness gives a value > 0 (more values clustered to close to left of mean
and a few values a long way to the right of the mean tend to make the value >0)*

Left skewness gives a value < 0

What is kurtosis?

*A distribution is said to display kurtosis if the height of the distribution is unusual (suggests observations more bunched or more spread out than should be). Measure this by
A normal distribution should have a kurtosis value of 3*

**Do the Jarque-Bera test for normality in the OLS residuals for both equations.
Which model do you prefer?**

Residuals in levels show little signs of skewness but some kurtosis (platykurtic – since peak of distribution lower than expected for a normal distribution)

Residuals from log-lin specification is much more left skewed and higher kurtosis

To test more formally

For levels

Construct Jarque-Bera test

$$JB = N * \left[\frac{Skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right]$$

$$jb = 16 * [((0.6^2/6) + (((1.2-3^2)/24)]$$

$$= 3.12$$

The statistic has a Ch^2 distribution with 2 degrees of freedom, (one for skewness one for kurtosis).

From tables critical value at 5% level for 2 degrees of freedom is 5.99

So $JB < c^2_{critical}$, so **accept** null that residuals are normally distributed.

For log-lin model

$$jb = 16 * [((-3^2/6) + (((6-3^2)/24))] = 30.0$$

So now $JB > c^2_{critical}$, so **reject** null that residuals are normally distributed.

v) If a model fails this test what does this mean for OLS estimation?

Suggests should try another functional form to try and make residuals normal, otherwise t and F stats which are all based around assumption that residuals are normal may be invalid.