

Mid-Term 3 – Outline Answers

1. Interpret the meaning of the OLS estimates of the constant and the slope in the following regressions of hourly pay on work experience (where wage is measured in £ an hour and work experience is measured in years)

a)	Source	SS	df	MS	Number of obs = 485		
	Model	264.138348	1	264.138348	F(1, 483) =	6.98	
	Residual	18278.5824	483	37.8438559	Prob > F =	0.0085	
	Total	18542.7207	484	38.3114065	R-squared =	0.0142	
					Adj R-squared =	0.0122	
					Root MSE =	6.1517	

	hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	experience	.1000	.050	2.00	0.009	.0155704	.1559488
	_cons	8.78	.559	15.70	0.000	7.683295	9.882285

(4 marks)

b)	Source	SS	df	MS	Number of obs = 485		
	Model	4.8040887	1	4.8040887	F(1, 483) =	16.11	
	Residual	144.032368	483	.29820366	Prob > F =	0.0001	
	Total	148.836457	484	.30751334	R-squared =	0.0323	
					Adj R-squared =	0.0303	
					Root MSE =	.54608	

	log(hourpay)	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	experience	.100	.002	4.00	0.000	.0041828	.0142055
	_cons	1.98	.049	39.89	0.000	1.883596	2.078797

(4 marks)

c)	Source	SS	df	MS	Number of obs = 485		
	Model	9.13700444	1	9.13700444	F(1, 483) =	32.85	
	Residual	139.661589	483	.278116538	Prob > F =	0.0000	
	Total	148.836457	484	.296649777	R-squared =	0.0644	
					Adj R-squared =	0.0625	
					Root MSE =	.52737	

	log(hourpay)	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	log(experince)	.200	.030	6.67	0.000	.1123737	.3296125
	_cons	1.68	.087	19.20	0.000	1.511381	1.856046

(4 marks)

d)	Source	SS	df	MS	Number of obs = 485		
	Model	871.632709	1	871.632709	F(1, 483) =	23.82	
	Residual	17671.088	483	36.5861036	Prob > F =	0.0000	
	Total	18542.7207	484	38.3114065	R-squared =	0.0470	
					Adj R-squared =	0.0450	
					Root MSE =	6.0486	

	hourpay	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	log(experince)	1.500	.250	5.00	0.000	.8149897	1.913277
	_cons	6.294	.819	7.68	0.000	4.683757	7.904829

(4 marks)

$$\hat{a) Wage} = 8.78 + 0.10 * Experience$$

*Answer: Linear model so slope coefficient $dWage/dExperience$
 = change in wage wrt unit change (ie 1 year in this case) in experience.
 (or equivalently change in wage = $\beta_{experience} * \text{unit change in experience}$
 $= 0.1 * 1$)*

*So 1 year of experience means a **10 pence (0.1 pounds) increase** in wages, on average. Intercept gives theoretical hourly wage if experience were zero, (in this case £8.78).*

$$\hat{b) Ln(Wage)} = 1.98 + 0.10 * Experience$$

*Log-lin model, so slope coefficient $dLn(wage)/dExperience$
 = (% change in wage/100) / unit change in experience (ie 1 year in this case)
 (or equivalently (% change in wage/100) = $\beta_{experience} * \text{unit change in experience}$
 $= 0.1 * 1$)*

*So 1 year of experience leads to a **10 percent increase** in wages, on average.
 Intercept gives theoretical log hourly wage (not the wage) if experience were zero,
 (in this case 1.98).*

$$\hat{c) Ln(wage)} = 1.68 + 0.20 * Ln(Experience)$$

*Log-linear (double log) – don't confuse with log-lin: so slope coefficient is an elasticity:
 $dLn(wage)/dLn(Experience)$*

*= % change in food expenditure wrt % change in income.
 or equivalently*

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

$$= 0.2 * 1$$

(Since both sides divide by 100, this cancels)

So 1% increase in experience increases wages by 0.2%.

Intercept gives theoretical level of Log wages when log experience (not the level of experience) is zero ie when experience is 1 year.

$$\hat{d) Wage} = 6.29 + 1.50 * Ln(Experience)$$

*Lin-log model, so slope coefficient is a semi-elasticity
 $dWage/dLn(Experience) = \text{change in wage wrt } \% \text{ change in experience}/100.$*

or equivalently

$$\begin{aligned} \text{unit change in wage}/100 &= \beta_{\log(\text{experience})} * (\% \text{ change in experience}/100) \\ &= 1.5 * 1/100 \\ &= 0.015 \end{aligned}$$

So 1% increase in experience raises hourly wage, on average, by £0.015.

Intercept gives theoretical level of hourly pay when log experience (not experience) is zero (ie experience=1 year).

What would happen to the estimates of a) the slope effect b) the intercept in model a if both wages and experience were multiplied by 100?

- Problem set 2 shows that rescaling both dependent and independent variable by same amount has following effects

$$\hat{\mathbf{b}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(100X, 100Y)}{\text{Var}(100X)} = \frac{100^2 \text{Cov}(X, Y)}{100^2 \text{Var}(X)} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

so no effect on slope estimate

$$\text{OLS estimate of intercept is } \hat{\mathbf{a}} = \bar{Y} - \hat{\mathbf{b}} \bar{X} = 100\bar{Y} - \hat{\mathbf{b}} 100\bar{X} = 100(\bar{Y} - \hat{\mathbf{b}} \bar{X}) = 100\hat{\mathbf{a}}_{\text{original}}$$

So new intercept equals original times constant of multiplication (in this case 100)

2.

a) What are the 5 main assumptions that underlie the OLS estimation technique? (10 marks)

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

b) Prove that the OLS estimate of the slope in the 2 variable model will give an unbiased estimate of the true slope coefficient

(8 marks)

see lecture notes

$$\hat{\mathbf{b}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(X, (\mathbf{b}_0 + \mathbf{b}_1 X + u))}{\text{Var}(X)} = \frac{0 + \mathbf{b}_1 \text{Cov}(X, X) + \text{Cov}(X, u)}{\text{Var}(X)} = \mathbf{b}_1 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}$$

taking expectations to get expression for bias

$$E(\hat{\mathbf{b}}) = E\left[\mathbf{b}_1 + \frac{\text{Cov}(X, u)}{\text{Var}(X)}\right] = \mathbf{b}_1 + E\left[\frac{\text{Cov}(X, u)}{\text{Var}(X)}\right]$$

which since assume $\text{Cov}(X, u) = 0$ gives

$$E(\hat{\mathbf{b}}) = \mathbf{b}_1 \quad \text{ie OLS gives an unbiased estimate of slope}$$

c) Consider the following equations

$$\text{a) } Y_t = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 X_t + \hat{u}_t$$

$$\text{b) } \hat{Y}_t = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 X_t$$

$$\text{c) } u_t = \hat{Y}_t - \beta_1 + \beta_2 X_t$$

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

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

a) *True: predicted value + predicted residual must equal true value*

b) *True; Predicted value equals $\hat{Y}_t = \hat{\mathbf{b}}_1 + \hat{\mathbf{b}}_2 X_t$*

c) *False: Actual residual is difference between actual Y and prediction based on true value of $u_t = Y_t - \beta_1 + \beta_2 X_t$*

3. Given the following regression output

$$\hat{\text{Consumption}} = 5000 + 0.90 * \text{Income}$$

$$R^2=0.89 \quad \text{TSS} = 1000 \quad \text{RSS} = 102 \quad \text{Var}(\text{Income})=1$$

estimated over the period 1990-2001

and

i) What factors influence the precision of the OLS estimate of the slope and why?

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

So precision of estimate increases with

1) sample size

1) fit of model (s_u^2)

2) variance of X variable

ii) Hence find the standard error on the estimate of income in model I

$$\text{Var}(\hat{\mathbf{b}}) = \frac{s_u^2}{N\text{Var}(X)} \quad s_u^2 = \text{RSS}/N-k$$

$$\text{so } s^2 = 102/12-2 = 10.2$$

$$\text{Var}(\hat{\mathbf{b}}) = \frac{10.2}{12*1} = 0.85 \quad \text{so standard error (square root of variance) is } 0.92$$

ii) Test the hypothesis that the true effect of income on consumption in model I is

a) Zero

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

$$\hat{t} = 0.9-0/0.92 = 0.98$$

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

$$N-k = 12-2 = 10$$

are 1.81 and 2.28 respectively

In this case absolute value of $\hat{t} < t_{\text{critical}}^{10\%} < t_{\text{critical}}^{5\%}$ so can't reject null that coefficient is zero ie income doesn't affect consumption (sample size is too small and standard errors on variables too large to make accurate estimates).

Find the 95% confidence interval for the estimated slope coefficient

95% confidence interval for each individual forecast observation given by

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

$$= \hat{\mathbf{b}}_1 \pm t_{.05/2} SE(\hat{\mathbf{b}}_1) = 0.9 \pm 2.23 * 1$$

So can be 95% confident true slope value lies in region -1.33 to 3.13

v) What do you understand by the term *p* value?

The *p* value is the lowest significance level at which the null hypothesis can be rejected (the exact probability of committing Type I error)

In practice this amounts to finding the significance level *a* which, given sample size *N* and no. right hand side coefficients *k*, equates the critical and estimated *t* values

$$\hat{t} = t_{N-k}^{a/2}$$

Intuitively if can't *a* a lot (and in so doing reduce the acceptance region) and still accept the null hypothesis, this suggests the null hypothesis is likely to be true.

So a **high** *p* value (high *a*) is evidence **in favour** of the null and a **low** *p* value is evidence **in against** the null

vi) What is the approximate p value for the null hypothesis that the true coefficient on income is zero?

From tables must lie between .7 and 1.372 which are 50% and 20% critical values so p is Around 40%

4. The following output is taken from regressions of a) the level of house prices (measured in £ thousands) b) the log of house prices on the level of interest rates (measured in percentage points)

a)	Source	SS	df	MS	Number of obs =	102
	Model		1	100.0	F(,) =	
	Residual	300.0	200	1.500	R-squared =	
	Total	500.0	201	2.487		

	houseprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
interestrate		-10000.000	200.000	5.00	0.000	
_cons		20000.000	200.000	10.00	0.000	

b)	Source	SS	df	MS	Number of obs =	102
	Model		1	10.0	F(,) =	
	Residual	40.0	200	0.200	R-squared =	
	Total	50.0	201	.2487		

	loghouseprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
interestrate		-0.1000	0.250	4.00	0.000	
_cons		16000.000	160.000	10.00	0.000	

i) Do the F test for goodness of fit of model a

1st Find the value of the R^2 of the model

Use fact that $R^2 = 1 - (RSS/TSS) = 1 - 300/500 = 0.4$

F test for goodness of fit of model given by

$$\frac{R^2 / k - 1}{1 - R^2 / N - k} \sim F[k-1, N-k]$$

$$\text{So } \hat{F} = \frac{0.4/2-1}{1-0.4/102-2} \sim F[2-1, 102-2]$$

$$\hat{F} = 66.6 \sim F[1, 100]$$

From F tables $F_{critical}^{5\% [1, 100]} = 3.94$

$\hat{F} > F_{critical}^{5\%}$ so reject null that model as a whole has zero explanatory power.

ii) How would you test whether the fit of the two models is significantly different?

The basic idea behind testing for the appropriate functional form of the dependent variable is to transform the data so as to make the RSS comparable

Do this by dividing each observation by the geometric mean to give $Y^* = y/G.\text{mean}$ and $\log y^*$ and use these as alternative dependent variables and applying the formula

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

If estimated value exceeds critical value (from tables Chi-squared at 5% level with 1 degree of freedom is 3.84) reject the null hypothesis that the models are the same (ie there is a significantly different in terms of goodness of fit).

iii) The following output is taken from the distribution of OLS residuals from the two regressions

from model a

Residuals				
	Percentiles	Smallest		
1%	-6.323713	-6.448029		
5%	-5.227664	-6.414042		
10%	-4.487434	-6.233384	Obs	102
25%	-3.299453	-6.18045	Sum of Wgt.	102
50%	-.9010892		Mean	-1.44e-10
		Largest	Std. Dev.	4.585193
75%	1.951943	16.25964		
90%	5.31512	17.23212	Variance	21.02399
95%	9.095832	17.61862	Skewness	2.00000
99%	17.42537	20.11699	Kurtosis	7.00000

From model b

Residuals				
	Percentiles	Smallest		
1%	-1.169948	-1.427379		
5%	-.743401	-1.216464		
10%	-.6195743	-1.123431	Obs	102
25%	-.3532593	-1.113894	Sum of Wgt.	102
50%	-.0013355		Mean	-1.52e-09
		Largest	Std. Dev.	.491033
75%	.3477869	1.147901		
90%	.5949472	1.218912	Variance	.2411134
95%	.8182689	1.228457	Skewness	0.100000
99%	1.223685	1.29072	Kurtosis	3.600000

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

(6 marks)

$$JB = \frac{N}{6} \left[\text{skewness}^2 + \frac{(\text{kurtosis} - 3)^2}{4} \right] \sim \chi^2(2)$$

$$\text{Model a} = (102/6)*((2^2)+((7-3)^2 /4)) = 17(4+4) = 136$$

$$\text{Model b} = (102/6)*((0.1^2)+((3.6-3)^2 /4)) = 17(0.01+0.09) = 1.7$$

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

*So $JB > \mathcal{C}_{critical}^2$ so **reject** null that residuals are normally distributed in model a*

*So $JB < \mathcal{C}_{critical}^2$ so **accept** null that residuals are normally distributed in model b*

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

If JB test is rejected, OLS remains unbiased but t values, F tests and confidence intervals – all of which are based on assumption of normality in residuals are biased