

1. Given the general linear model $y = XB + u$ where y is $k \times 1$ and X is $n \times k$ and $u \sim N(0, \sigma^2)$,

show that

a) the OLS estimate of B , $\hat{\beta} = (X'X)^{-1}X'y$ (7 marks)

b) $X'u = 0$ (8 marks)

c) $(y'y - \bar{ny}^2) = (\beta'X'X\beta - \bar{ny}^2) + u'u$ (10 marks)

Give a short verbal description of what these results mean in your answers

- Answers to all these can be found in your lecture notes

$$a) u'u = (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

f.o.c. for minimum requires

$$\frac{\partial u'u}{\partial \hat{\beta}} = 0 = -2X'y + 2X'X\hat{\beta}$$

so $X'y = X'X\hat{\beta}$ and $\hat{\beta} = (X'X)^{-1}X'y$

minimising the sum of squared residuals gives this result

b) the X variables are uncorrelated with the OLS residuals

c) start from

$$y'y = (y+u)'(y+u) = y'y + u'u \quad \text{- using answer to part b)}$$

$$y'y = \beta'X'X\beta + u'u$$

and subtract \bar{ny}^2 from both sides

the total sum of squares equals the explained sum of squares plus the residual sum of squares

2. The following is taken from a regression of the log of hourly wages on the number of years of work experience, (exper) the square of experience (exper2) and a dummy variable (female) that takes the value 1 if the individual is female and 0 otherwise log of labour input and log of capital input. Some of the information has been concealed.

$$\hat{\ln}(\text{hourwage}) = 2.00 + 0.036 * \text{exper} - 0.001 * \text{exper}^2 - 0.30 \text{female}$$

$$N=1004 \quad R^2 = 0.090 \quad \bar{R}^2 = 0.080$$

The variance/co-variance matrix of the parameter estimates, $\text{Var}(\hat{\beta})$, is given by

	exper	exper2	female
exper	0.029		
exper2	-0.005	0.001	
female	-0.090	0.002	.015625

a) Interpret the coefficient on the female dummy variable (3 marks)

since this is a log-lin model the coefficients are semi-elasticities,

$$\delta \ln W / \delta x_i = \hat{\beta}_i = (\delta W / W) / \delta x_i \quad \text{so } \% \text{ change in wage} = (\hat{\beta}_i * \delta x_i) * 100$$

however this is a dummy variable so need to transform by $(\exp(\hat{\beta}_i) - 1)$

$$\hat{\beta}_i = 0.3 \text{ so the effect of female is } \exp(0.3) - 1 = 0.35 * 100 = 35\%$$

b) Find the standard error of the estimate on the female variable
the estimated t value
the 95% confidence interval around this estimate

(10 marks)

variance covariance matrix of parameter estimates gives variance of coefficients down main diagonal and covariances in off diagonal (matrix is also symmetric)

$$\text{so } \text{Var}(\hat{\beta}_{\text{female}}) = 0.015625, \text{ so } S.E.(\hat{\beta}_{\text{female}}) = \sqrt{0.015625} = \sqrt{1/64} = 1/8 = 0.125$$

$$\text{hence } t = \frac{\hat{\beta}_{\text{female}} - \beta_{\text{female}}^0}{s.e.(\hat{\beta}_{\text{female}})} = (0.3 - 0) / 0.125 = 2.4$$

95% confidence interval for each individual forecast observation given by

$$\Pr[\hat{\beta}_1 - t_{n-k}^{\alpha/2} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{n-k}^{\alpha/2} SE(\hat{\beta}_1)] = 1 - .05$$

$$\text{critical value for } t_{n-k}^{\alpha/2} = t_{1004-4}^{.05/2} = 1.96$$

$$= \hat{\beta}_1 \pm t_{.05/2} SE(\hat{\beta}_1) = -0.3 \pm 1.96 * 0.125$$

So can be 95% confident true slope value lies in region- 0.055 to -0.545

b) Test the hypothesis that the coefficients on exper and exper2 add to zero
(5 marks)

$$H_0: \hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2} = 0$$

Use $F =$

$$\frac{\left(\hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2} - 0 \right)^2}{\text{Var}(\hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2})} = \frac{\left(\hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2} \right)^2}{\text{Var}(\hat{\beta}_{\text{exper}}) + \text{var}(\hat{\beta}_{\text{exper}^2}) + 2\text{Cov}(\hat{\beta}_{\text{exper}}, \hat{\beta}_{\text{exper}^2})}$$

$$= (.036 - 0.001)^2 / (.029 + .001 + 2 * -0.005) = .001225 / (.02) = 0.061 \sim F[1, 1000]$$

From F tables $F_{\text{critical}}^{5\% [1, 1000]} = 3.84$

$\hat{F} < F_{\text{critical}}^{5\%}$ so CANNOT reject null that coefficients sum to zero.

N.B. could use fact that $t^2 = F$ and hence equivalent test is

$$t = \frac{\hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2} - 0}{\text{S.E.}(\hat{\beta}_{\text{exper}} + \hat{\beta}_{\text{exper}^2})} = \frac{0.035}{\sqrt{0.2}} = 0.25 \sim t_{1000}$$

c) Test the significance of the goodness of fit of the model as a whole (4 marks)

F test for goodness of fit of model given by

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

$$\text{So } \hat{F} = \frac{0.09 / 4 - 1}{1 - 0.09 / 1004 - 4} \sim F[4-1, 1004-4]$$

$$\hat{F} = 33.0 \sim F[3, 1000]$$

From F tables $F_{\text{critical}}^{5\% [3, 1000]} = 2.60$

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

d) Intuitively the distance between the null and hypothesised values is so large even allowing for sampling variation that the null is rejected.

3. Given the model

$$y = b_1 + b_2 X_2 + b_3 X_3 + u$$

and the following information

$$X'X = \begin{bmatrix} 103 & 0 & 0 \\ 0 & 10 & 5 \\ 0 & 5 & 3 \end{bmatrix} \quad X'y = \begin{bmatrix} 10 \\ 5 \\ 1 \end{bmatrix} \quad \sum_i (Y_i - \bar{Y})^2 = 17$$

Since the model contains a constant we know

$$X = [i \ X_2 \ X_3] \quad \text{so } X'X = \begin{bmatrix} N & \sum X_2 & \sum X_3 \\ \sum X_2 & \sum X_2^2 & \sum X_2 X_3 \\ \sum X_3 & \sum X_2 X_3 & \sum X_3^2 \end{bmatrix}$$

So the sample size is given by the top left hand element of $X'X = 103$

Since $\sum X_3 = \sum X_2 = 0$, it follows that $\bar{X}_2 = \bar{X}_3 = 0$. Hence lower right-hand sub-matrix of $X'X$ is equivalent to mean deviation form, as is lower 2×1 sub-matrix of $X'y$ (since $\sum xy = \sum xY = \sum XY$)

so can estimate slope parameters according to $\hat{\beta} = (x'x)^{-1} x'y$

$$= \begin{bmatrix} 10 & 5 \\ 5 & 3 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \frac{1}{30-25} \begin{bmatrix} 3 & -5 \\ -5 & 10 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

b) to find the estimated standard errors on X_2 and X_3 need $\text{var}(\hat{\beta}_i) = s^2 (x'x)^{-1}_{ii}$

where $(x'x)^{-1}_{ii}$ is the i th element on the main diagonal of $(x'x)^{-1}$

and $s^2 = \frac{u'u}{n-k}$ is the unbiased ols estimator of the true residual variance σ^2

first find s^2 use $u'u = y'y - \hat{\beta}'x'y = 17 - [2 \ -3] \begin{bmatrix} 5 \\ 1 \end{bmatrix} = 17 - 7 = 10$

and $s^2 = 10/103 - 3 = 0.1$

so $\text{var}(\hat{\beta}) = 0.1 \begin{bmatrix} 0.6 & -1 \\ -1 & 2 \end{bmatrix}$

$$\text{and s.e.}(\beta_2) = (0.06)^{1/2} = 0.24$$

$$\text{and s.e.}(\beta_2) = (0.2)^{1/2} = 0.45$$

Find the OLS estimates of

- the slope coefficients (6 marks)
- the residual sum of squares (RSS) (4 marks)
- the standard error of the slope coefficients, b_2 and b_3 (5 marks)
- What do you understand by the term “encompassing principle” (4 marks)
- Outline the form of an encompassing test that you might use to choose between y and $\log(y)$ as a dependent variable (6 marks)

4. The following regression output is based on estimates from a data set containing 30 yearly observations on total consumption and income

. reg cons income if year<97

Source	SS	df	MS			
Model	1.3386e+11	1	1.3386e+11	Number of obs = 27		
Residual	2.6125e+09	25	104501857	F(1, 25) = 1280.91		
Total	1.3647e+11	26	5.2488e+09	Prob > F = 0.0000		
				R-squared = 0.9809		
				Adj R-squared = 0.9801		
				Root MSE = 10223		
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.9016052	.0251917	35.79	0.000	.8497219	.9534886
_cons	17898.81	9314.261	1.92	0.066	-1284.271	37081.89

. reg cons income if year>=97

Source	SS	df	MS			
Model	706995633	1	706995633	Number of obs = 3		
Residual	192948516	1	192948516	F(1, 1) = 3.66		
Total	899944149	2	449972074	Prob > F = 0.3065		
				R-squared = 0.7856		
				Adj R-squared = 0.5712		
				Root MSE = 13891		
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	1.75263	.9155935	1.91	0.306	-9.881088	13.38635
_cons	-422003.4	486938.4	-0.87	0.545	-6609142	5765135

reg cons income

Source	SS	df	MS			
Model	2.0574e+11	1	2.0574e+11	Number of obs = 30		
Residual	3.9618e+09	28	141494469	F(1, 28) = 1454.04		
Total	2.0970e+11	29	7.2311e+09	Prob > F = 0.0000		
				R-squared = 0.9811		
				Adj R-squared = 0.9804		
				Root MSE = 11895		
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

income	.9196488	.0241176	38.13	0.000	.8702462	.9690514
_cons	11536.83	9381.629	1.23	0.229	-7680.561	30754.23

You can however test the forecast accuracy of the 1st regression by using the Chow forecast test

$$F = \frac{RSS_{restricted} - RSS_{unrestricted}/N_0}{RSS_{unrestricted}/N_1 - k} \sim F[N_0, N_1 - k]$$

where $RSS_{unrestricted}$ is equal to the RSS from the model estimated over the 1st N_1 observations and $RSS_{restrict}$ is equal to the RSS estimated over all $N_1 + N_0$ observations

A regression over the entire sample period gives

$$\text{So } F = \frac{3.952 - 2.613/3}{2.613/27-2} \sim F[3, 27-2]$$

$$F = 4.30$$

From tables F critical 5% level for 3 and 25 degrees of freedom is 2.99
Hence estimated $F > F_{critical}$ so reject null that parameters are constant in and out of sample

- a) Write down the restricted and unrestricted models in matrix form (5 marks)

from lecture notes we know that unrestricted model in stacked matrix form is given by

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ X_2 & I_{N_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

and restricted model can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

- a) Do the Chow forecast test of parameter constancy (5 marks)

You can test the forecast accuracy of the 1st regression by using the Chow forecast test

$$F = \frac{RSS_{restricted} - RSS_{unrestricted}/N_2}{RSS_{unrestricted}/N_1 - k} \sim F[N_2, N_1 - k]$$

where $RSS_{unrestricted}$ is equal to the RSS from the model estimated over the 1st N_1 observations and $RSS_{restrict}$ is equal to the RSS estimated over all $N_1 + N_2$ observations

```
. reg cons income if year<97
```

Source	SS	df	MS	Number of obs = 27		
Model	1.3386e+11	1	1.3386e+11	F(1, 25) =	1280.91	
Residual	2.6125e+09	25	104501857	Prob > F =	0.0000	
Total	1.3647e+11	26	5.2488e+09	R-squared =	0.9809	
				Adj R-squared =	0.9801	
				Root MSE =	10223	
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.9016052	.0251917	35.79	0.000	.8497219	.9534886
_cons	17898.81	9314.261	1.92	0.066	-1284.271	37081.89

```
reg cons income
```

Source	SS	df	MS	Number of obs = 30		
Model	2.0574e+11	1	2.0574e+11	F(1, 28)	=	1454.04
Residual	3.9618e+09	28	141494469	Prob > F	=	0.0000
-----				R-squared	=	0.9811
Total	2.0970e+11	29	7.2311e+09	Adj R-squared	=	0.9804
-----				Root MSE	=	11895
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.9196488	.0241176	38.13	0.000	.8702462	.9690514
_cons	11536.83	9381.629	1.23	0.229	-7680.561	30754.23

Note degrees of freedom in numerator = $N_2 =$ number of observations (restrictions) in the 2nd sub-sample

$$\text{So } F = \frac{3.961 - 2.613/3}{2.613/27} \sim F[3, 27-2]$$

$$F = 4.64$$

From tables F critical 5% level for 3 and 25 degrees of freedom is 2.99

Hence estimated $F > F_{\text{critical}}$ so reject null that parameters are constant in and out of sample

c) A dummy variable that takes the value 1 if the observation belongs to the year 1991 and 0 otherwise is now added to the regression.

```
reg cons income d91
```

Source	SS	df	MS	Number of obs = 30		
Model	2.0646e+11	2	1.0323e+11	F(2, 27)	=	859.61
Residual	3.2424e+09	27	120088034	Prob > F	=	0.0000
-----				R-squared	=	0.9845
Total	2.0970e+11	29	7.2311e+09	Adj R-squared	=	0.9834
-----				Root MSE	=	10958
cons	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
income	.9273132	.022438	41.33	0.000	.8812742	.9733522
d91	-27551.04	11255.95	-2.45	0.021	-50646.34	-4455.744
_cons	9554.774	8680.729	1.10	0.281	-8256.611	27366.16

Explain, (without giving formal proof), what the addition of the dummy variable does to the residual sum of squares in the model and to the estimated coefficient on income (6 marks)

Since the unrestricted stacked matrix model contains an identity matrix this is equivalent to including a set of N_2 dummy variables alongside the original X variables which takes the value 1 for the $N_1 + i$ th observation and 0 otherwise. We know this has the effect of making the RSS identical to that from a regression on the 1st N_1 observations and so the regression coefficients on the original X are the same whether the $N_1 + i$ th observation is included or not.

d) Now suppose the regression equation in part c were projected beyond the year 1999 on the following data

year	consumption	income
2000	55000	56000
2001	57000	60000

find the forecast values of consumption for these two periods (to the nearest whole number)

(3 marks)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \text{Income}$$

$$\text{in 2000} \quad = 9554.8 + 0.927 * 56000 = 61467$$

$$\text{in 2001} \quad = 9554.8 + 0.927 * 60000 = 65175$$

write down the formula that would enable you to calculate the 95% confidence interval for the true value of consumption in each year based around these forecasts (6 mark)

the 95% confidence interval around a forecast is based around the fact that given by

$$t = \frac{\hat{y}_O - y_O}{\hat{s.e.}(y_O - y_O)} = \frac{\hat{y}_O - y_O}{\hat{s.e.}(u_O)} \sim t_{n-k}$$

and given by

$$\hat{y}_O \pm t_{n-k}^{\alpha/2} \hat{s.e.}(u_O)$$

the standard error of a forecast is given by

$$\hat{s.e.}(u_O) = s \sqrt{x_O' (X'X)^{-1} x_O}$$

$$\text{where } s = \sqrt{\frac{RSS}{N-k}}$$

and x_O' is the o^{th} row of X