Exercise 4 -Answers

1. Given 50 observations split evenly into 2 periods, you decide to estimate

$Y_i = a_1 + b_1 X_i + u_{i1}$ in period 1

and

$Y_i = a_2 + b_2 X_i + u_{i2}$ in period 2

Show how all four parameters could be obtained from a single OLS regression.

Suppose that RSS1 = .6875 and RSS2 = 2.4727 and that the RSS from the pooled regression is 6.5565.

Test the hypothesis of no structural change at the 5% and 1% level.

Proof

Above is the unrestricted form of the model (intercepts and the slopes vary in two periods)

In (partitioned) matrix form

$$y=\begin{bmatrix} y_1 \\ \dots \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}\begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = X\beta + u \qquad (1)$$

ie stack the data from the second period below that of the observations from the 1st period

in a way that allows the coefficients to differ between the periods

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N1} \\ .. \\ y_{N1+1} \\ y_{N1+2} \\ \vdots \\ y_{N1+N2} \end{bmatrix} = \begin{bmatrix} 1 & X_1 & 0 & 0 \\ 1 & X_2 & 0 & 0 \\ : & : & : & : \\ 1 & X_{N1} & 0 & 0 \\ 0 & 0 & 1 & X_{N1+1} \\ 0 & 0 & 1 & \\ : & : & : & \\ 0 & 0 & 1 & X_{N1+N2} \end{bmatrix}\begin{bmatrix} a_1 \\ b_1 \\ a_2 \\ b_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

OLS on (1) gives $\hat{\beta} = \begin{bmatrix} \hat{a}_1 \\ \hat{b}_1 \\ \hat{a}_2 \\ \hat{b}_2 \end{bmatrix} = (X'X)^{-1}X'y$

Using rules on inverse of partitioned matrices (the inverses of the elements on a diagonal partitioned matrix are just the inverses of the elements themselves)

$$\hat{\beta}= \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1}X_1'y \\ (X_2'X_2)^{-1}X_2'y \end{bmatrix}$$

which is identical to those obtained by running OLS separately on the two sub-samples

Compare this with estimates from the restricted model based on

$$y=\begin{bmatrix} y_1 \\ \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & X_1 \\ i_2 & X_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = X\beta + u$$

and use $F=\frac{(RSS_{restricted}-RSS_{unrestricted})/q}{RSS_{unrestricted}/N-k} = \frac{(6.5565-(0.6875+2.4727))/2}{(0.6875+2.4727)/50-4} \sim F[2, 50-4]$

(remember there are 4 parameters in the unrestricted model so k=4)

hence $\hat{F}= 24.72$

and from Tables $F^{.05}[2,46] = 3.2$

$\hat{F} > F_{critical}$ so reject null (or no structural change)

2. Given data combined over 2 periods, consider the pooled regression of y on a constant and a dummy variable to denote that the observation belongs to the second period

$$Y_t = b_0 + b_1 D_t + u_t \qquad (1)$$

Show that the OLS estimates of b_0 and b_1 give, respectively, the mean value of y in period 1 and the difference in mean values between period 2 and period 1.

(Hint: partition the data and use OLS matrix algebra).

Let period 1 = 1,2... N1 observations period 2= N1+1. N1+2, ... N= N2 observations

Let i_1 be an $N_1 x1$ vector of ones and i_2 be an $N_2 x1$ vector of ones

Then (1) can be written in partitioned matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1 & 0 \\ i_2 & i_2 \end{bmatrix}\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \Leftrightarrow y = X\beta + u$$

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

where X'X $= \begin{bmatrix} i_1' & i_2' \\ 0 & i_2' \end{bmatrix}\begin{bmatrix} i_1 & 0 \\ i_2 & i_2 \end{bmatrix} = \begin{bmatrix} i_1'i_1 + i_2'i_2 & i_2'i_2 \\ i_2'i_2 & i_2'i_2 \end{bmatrix}$

Since i_j is an $N_j x1$ vector then $i_1'i_1 = N_1$ and $i_2'i_2 = \begin{bmatrix} 1 & 1 & .. & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ : \\ 1 \end{bmatrix} = N_2$

and so X'X $= \begin{bmatrix} N_1 + N_2 & N_2 \\ N_2 & N_2 \end{bmatrix}$

Similarly X'y $= \begin{bmatrix} i_1' & i_2' \\ 0 & i_2' \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} i_1'y_1 + i_2'y_2 \\ i_2'y_2 \end{bmatrix}$

Using the fact that $\bar{y}_j = \frac{\sum_j y_j}{N_j}$ and that $i_j'y_j = \sum_j y_j$

then X'y $= \begin{bmatrix} N_1\ \bar{y}_1\ + N_2\ \bar{y}_2 \\ N_2\ \bar{y}_2 \end{bmatrix}$

Hence $\hat{\beta}$

$$= \begin{bmatrix} N_1 + N_2 & N_2 \\ N_2 & N_2 \end{bmatrix}^{-1}\begin{bmatrix} N_1\ \bar{y}_1\ + N_2\ \bar{y}_2 \\ N_2\ \bar{y}_2 \end{bmatrix} = \frac{1}{N_1N_2}\begin{bmatrix} N_2 & -N_2 \\ -N_2 & N \end{bmatrix}^{-1}\begin{bmatrix} N_1\ \bar{y}_1\ + N_2\ \bar{y}_2 \\ N_2\ \bar{y}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{N_1}(N_1\ \bar{y}_1\ + N_2\ \bar{y}_2) - \frac{1}{N_1}(N_2\ \bar{y}_2) \\ -\frac{1}{N_1}(N_1\ \bar{y}_1\ + N_2\ \bar{y}_2) + \frac{1}{N_1}(N\ \bar{y}_2) \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{bmatrix}$$

so the coefficient on the intercept in a model with no other covariates apart from the dummy variable gives the mean value of the dependent variable in period 1 and the coefficient on the dummy variable gives the difference in the mean value of the dependent variable between period 2 and period 1

Note that if additional covariates are added to the model so that

$$y = X_1\beta_1 + X_2\beta_2 + u \qquad (2)$$

where now the data are stacked such that $X_1 = \begin{bmatrix} i_1 & 0 \\ i_2 & i_2 \end{bmatrix}$ and $X_2 = \begin{bmatrix} X_2^1 \\ X_2^2 \end{bmatrix}$

then partitioned regression tells us that the OLS on (2) gives

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$

The 1st row tells us that $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\,\hat{\beta}_2 \qquad (3)$
(see lecture notes)

where $\hat{\beta}_1 = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix}$
contains the coefficients on the intercept and on the time dummy

Now the term $(X_1'X_1)^{-1}X_1'X_2$ is a $k_1 x k_2$ matrix of OLS estimates from regressions of each of the
k_2 variables in X_2 on the set of k_1 variables in X_1

Eg if X_2 contains just 1 variable then

$$\hat{\gamma} = (X_1'X_1)^{-1}X_1'X_2 \Rightarrow \begin{bmatrix} X_2^1 \\ X_2^2 \end{bmatrix} = \begin{bmatrix} i_1 & 0 \\ i_2 & i_2 \end{bmatrix}\begin{bmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix}$$

$$\Rightarrow \qquad \hat{\gamma}_1 = \bar{X}_2^1 \qquad and\ \hat{\gamma}_2 = \bar{X}_2^2 - \bar{X}_2^1$$

so that (3) gives the adjusted OLS coefficients on the intercept and on the time dummy in a multiple regression as

$$\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 - \hat{\beta}_2 \bar{X}_2^1 \\ (\bar{y}_2 - \bar{y}_1) - \hat{\beta}_2 (\bar{X}_2^2 - \bar{X}_2^1) \end{bmatrix}$$

ie this time the correction factor equals the mean (or the difference in the mean)

of the dependent variable minus the mean (or difference in mean) of the additional explanatory variables multiplied by its own OLS regression coefficient ie the coefficients are now net of the difference in the means of the other variables

3. Given

LnQ = -3.8766+1/.4106LnL+0.4162LnK

1929-67 $R^2 = 0.9937$ $s = 0.03755$

LnQ = -4.0576+1.6167LnL+0.2197LnK

1929-48 $R^2 = 0.9759$ $s = 0.04573$

LnQ = -1.9564+0.8336LnL+0.6631LnK

1949-67 $R^2 = 0.9904$ $s = 0.02185$

To test for equality of ceofficients across the two sub-periods use the chow test

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/q}{RSS_{unrestricted}/N-k}$$

where $RSS_{restricted} = s^2_{restricted} * (N - k_{resttict}) = (0.03755)^2 * (39 - 3) = 0.0508$

and $RSS_{unrestrict} = RSS_{29-48} + RSS_{49-67}$

$= (0.04573)^2 * (20 - 3) + (0.02185)^2 (19 - 3)$

$= 0.0355 + 0.0076$

$= 0.0431$

and so $\hat{F} = \frac{(0.0508 - 0.0431)/3}{0.0431/39-(2*3)} \sim F[3,33]$

$= 1.96$

From Tables the 5% critical value for F[3,33] = 2.89

so $\hat{F} < F_{critical}$

and hence accept the null hypothesis (that the coefficients are the same in both periods)

4. To test for differences between the two sub-samples again use the Chow test

This time can find unrestricted RSS using the fact that

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

From information in the question

$$\hat{u}_1' \hat{u}_1 = 30 - \begin{bmatrix} 10 & 20 \end{bmatrix} \begin{bmatrix} 20 & 20 \\ 20 & 25 \end{bmatrix}^{-1} \begin{bmatrix} 10 \\ 20 \end{bmatrix}$$

$$= 30 - \begin{bmatrix} 10 & 20 \end{bmatrix} \begin{bmatrix} 25/100 & -20/100 \\ -20/100 & 20/100 \end{bmatrix} \begin{bmatrix} 10 \\ 20 \end{bmatrix}$$

$$= 30 - \begin{bmatrix} 10 & 20 \end{bmatrix} \begin{bmatrix} -3/2 \\ 2 \end{bmatrix}$$

$$= 5$$

Similarly $\hat{u}_2' \hat{u}_2 = 24 - \begin{bmatrix} 8 & 20 \end{bmatrix} \begin{bmatrix} 10 & 10 \\ 10 & 20 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 20 \end{bmatrix}$

$= 3.2$

To find the restricted RSS need to find X'X for the combined (pooled) regression

Since $X = \begin{bmatrix} X_1 \\ .. \\ X_2 \end{bmatrix}$

ie a partitioned matrix with period 1 observations stacked above those from period 2

then X'X $= \begin{bmatrix} X_1' & X_2' \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = X_1'X_1 + X_2'X_2$

$$= \begin{bmatrix} 20 & 20 \\ 20 & 25 \end{bmatrix} + \begin{bmatrix} 10 & 10 \\ 10 & 20 \end{bmatrix} = \begin{bmatrix} 30 & 30 \\ 30 & 45 \end{bmatrix}$$

Similarly X'y $= X_1'y_1 + X_2'y_2 \quad = \begin{bmatrix} 10 \\ 20 \end{bmatrix} + \begin{bmatrix} 8 \\ 20 \end{bmatrix} = \begin{bmatrix} 18 \\ 40 \end{bmatrix}$

and y'y $= y_1'y_1 + y_2'y_2 = 54$

So the restricted RSS = 54 - $\begin{bmatrix} 18 & 40 \end{bmatrix} \begin{bmatrix} 30 & 30 \\ 30 & 45 \end{bmatrix}^{-1} \begin{bmatrix} 18 \\ 40 \end{bmatrix} = 10.93$

and hence F$=\frac{(RSS_{restricted} - RSS_{unrestricted})/q}{RSS_{unrestricted}/N-k} = \frac{[10.93-(5+3.2)]/2}{(5+3.2)/30-2(2)} = 4.33$

From tables the 5% critical value for F[2,26] = 3.37

so $\hat{F}$>F$_{critical}$

and hence **reject** the null hypothesis (that the coefficients are the same in both periods)

5. You are asked to correct a simple consumption function equation for quarterly seasonal variation. Write down what the matrix of independent variables looks like for the corrected model. Now given

Ct = 6688 +	1322D2	-217D3 +	183D4 +	.638*Income		R2 = .525
(1711)	(683)	(602)	(654)	(.155)		N = 100

where the numbers in brackets are standard errors. On the basis of this regression you decide to test the hypothesis that only second quarter consumption differs from the rest, (why?). The result is that now

To seasonally adjust (quarterly) data inroduce a dummy variable for each

quarter

$D_1 = 1$ if the observation appears in quarter i

= 0 otherwise

To avoid the dummy variable trap include one less dummy variable than the total number of quarters

X =

$$\begin{bmatrix} 1 & 0 & 0 & 0 & X_1 \\ 1 & 1 & 0 & 0 & X_2 \\ : & 0 & 1 & 0 & \\ & 0 & 0 & 1 & \\ 1 & 0 & 0 & 0 & \\ 1 & 1 & 0 & 0 & \\ : & 0 & 1 & 0 & \\ 1 & : & : & : & \\ 1 & 0 & 0 & 1 & X_N \end{bmatrix} = \begin{bmatrix} const & D_2 & D_3 & D_4 & X \end{bmatrix} \qquad and \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_{D2} \\ \beta_{D3} \\ \beta_{D4} \\ \beta_X \end{bmatrix}$$

From the regerssion output above can see that the 3rd and 4th quarter dummy variables are statistically insignificant

This suggests that a more restricted model which drops these variables may be acceptable.

H0: D3=D4=0

To test this formally use

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/q}{RSS_{unrestricted}/N-k} = \frac{(R^2_{unrestricted} - R^2{}_{restricted})/q}{1-R^2_{unrestricted}/N-k} = \frac{(0.525-0.515)/2}{(1-0.525)/100-5} \sim F[2,95]$$

(using $R^2 = 1 - (RSS/TSS)$)

Hence $\hat{F} = 4.21$

From tables the 5% critical value for F[2,95] = 3.10

so $\hat{F} > F_{critical}$

and hence **reject** the null hypothesis (that the 3rd and 4th quarter dummy variables have no explanatory value)

6. Given $y = X_1\beta_1 + X_2\beta_2 + v$ (1)

and $y = X_1\beta_1 + u$ (2)

then the Frisch-Waugh theorem tells us that the OLS estimate of β_1 can be obtained from the alternative regression

$$M_{X2}y = M_{X2}X_1\beta_1 + \epsilon \qquad (3)$$

where $M_{X2} = I - X_2(X_2'X_2)^{-1}X_2$ is the idempotent "residual maker" matrix

and $M_{X2}y$ are the residuals from a regression of y on X_2 alone
and $M_{X2}X_1$ are the residuals from a regression of X_1 on X_2 alone

OLS on (3) gives $\tilde{\beta}_1 = (X_1' M_{X2}' M_{X2} X_1)^{-1} X_1' M_{X2}' M_{X2} y$
$= (X_1' M_{X2} X_1)^{-1} X_1' M_{X2} y$
sub. in true y from (2)

$$= (X_1' M_{X2} X_1)^{-1} X_1' M_{X2} (X_1 \beta_1 + u)$$

$$\tilde{\beta}_1 = \beta_1 + (X_1' M_{X2} X_1)^{-1} X_1' M_{X2} u$$

Taking expectations $E(\tilde{\beta}_1) = E[\beta_1 + (X_1' M_{X2} X_1)^{-1} X_1' M_{X2} u]$
$= \beta_1 + E[(X_1' M_{X2} X_1)^{-1} X_1' M_{X2} u]$
$= \beta_1$
So estimates on relevant variables from a model that includes irrelevant

variables are unbiased

Now consider the estimates on the irrelevant variables.
Again the Frisch-Waugh theorem tells us that the OLS estimate of β_2 can be obtained from the alternative regression

$$M_{X1} y = M_{X1} X_2 \beta_2 + \eta \qquad (4)$$

where now $M_{X1} = I - X_1 (X_1' X_1)^{-1} X_1$
so $\tilde{\beta}_2 = (X_2' M_{X1}' M_{X1} X_2)^{-1} X_2' M_{X1}' M_{X1} y$

$$= (X_2' M_{X1} X_2)^{-1} X_2' M_{X1} y$$

sub. in for true y from (2) $\tilde{\beta}_2 = (X_2' M_{X1} X_2)^{-1} X_2' M_{X1} (X_1 \beta_1 + u)$

$$= 0 + (X_2' M_{X1} X_2)^{-1} X_2' M_{X1} u$$

(since $M_{X1} X_1 = 0$)

taking expectations $E\left[\tilde{\beta}_2\right] = E[(X_2' M_{X1} X_2)^{-1} X_2' M_{X1} u] = 0$

so expected values of irrelevant variables (assuming model (2) is correct) are zero

7.
Given

True: $y = X_1\beta_1 + X_2\beta_2 + u$ (1)

Estimate: $y = X_1\beta_1 + v$ (2)

then OLS estimate of the residual variance based on (2) gives
$s^2 = \frac{\hat{v}'\hat{v}}{N-k_1}$

($k_1 =$

no. of parameters
where $\hat{v} = y - X_1\hat{\beta}_1 = \left[I - X_1(X_1'X_1)^{-1}X_1'\right]y$

$= M_1 y$
and M_1 is an idempotent matrix $M_1'M_1 = M_1$

sub. in for true y from (1)
$\hat{v} = M_1[X_1\beta_1 + X_2\beta_2 + u] = M_1X_2\beta_2 + M_1u$ $(\sin ceM_1X_1 = 0)$

$\therefore \hat{v}'\hat{v} = (M_1X_2\beta_2 + M_1u)'(M_1X_2\beta_2 + M_1u) = \beta_2'X_2'M_1X_2\beta_2 + u'M_1X_2\beta_2 + \beta_2'X_2'M_1u + u'M$
and $E(\hat{v}'\hat{v}) = E[\beta_2'X_2'M_1X_2\beta_2 + u'M_1X_2\beta_2 + \beta_2'X_2'M_1u + u'M_1u]$
$= E[\beta_2'X_2'M_1X_2\beta_2] + E[u'M_1X_2\beta_2] + E[\beta_2'X_2'M_1u] + E[u'M_1u]$

middle two terms are zero since E(u)=0
$\therefore E(\hat{v}'\hat{v}) = \beta_2'X_2'M_1X_2\beta_2 + E[u'M_1u]$

Now $u'M_1u$ is a scalar

and can use the fact that u'u = tr(u'u)=tr(uu') when u'u is a scalar
where the trace is the sum of the diagonal elements = sum of the characteristic roots = rank of idempotent matrix
$\therefore E(u'M_1u) = E[tr(u'M_1u)] = E[trM_1 u')]$

(law of cyclic permutations; see eg Greene appendix A)

$= tr[M_1E(uu')]$
$tr(M_1)\sigma^2 I_N$

Now tr(M_1) $= tr\left[I_N - X_1(X_1'X_1)^{-1}X_1'\right] = tr\left[I_N - X_1'X_1(X_1'X_1)^{-1}\right]$
$=$ tr$[I_N - I_{k1}]$
$=$N-k$_1$

Hence E($\hat{v}'\hat{v}$) $= \beta_2'X_2'M_1X_2\beta_2 + \sigma^2(N-k1)$
$> \sigma^2(N-k1)$

ie OLS estimate of residual variance no longer unbiased. It is biased **upward** by an amount equal to the increase in the RSS when X_2 is excluded from the model

8. Given the formula for omitted variable bias

$$\hat{\beta}_i^{omit} = \hat{\beta}_i^{true} + (X_1'X_1)^{-1}X_1'X_2\,\hat{\beta}_2$$

where $X_1 = age$ $\quad X_2 = tenure$

Hence bias in OLS estimate of age in omitted variable model A depends on

a) $\hat{\beta}_2 = \hat{\beta}_{tenure} = \frac{\delta LnW}{\delta Tenure} = \frac{\%\Delta HourlyWage \div 100}{Unit\Delta Tenure}$ so

$\hat{\beta}_{tenure} * Unit\Delta Tenure * 100 = \%\Delta HourlyWage$

in this case $\hat{\beta}_2 = 0.017 =$

A unit (1 year) increase in job tenure raises wages by 1.7%
this would tend to raise the OLS estimate on age in the omitted variable model

A relative to that in the true model B

b) The covariance between X_1 and X_2 $\equiv$OLS coefficient from regression of X_2 on X_1

This information is not given in the question, but can work out its effect since

$$\frac{\hat{\beta}_i^{omit} - \hat{\beta}_i^{true}}{\hat{\beta}_2} = (X_1'X_1)^{-1}X_1'X_2 = \frac{.0085-.0032}{.017} = 0.311$$

which is just the coefficient from a regression of tenure (X_2) on $X_1(age)$

This is confirmed by the regression
reg tenure age

Source	SS	df	MS
Model	81512.3518	1	81512.3518
Residual	333694.698	6223	53.6228022
Total	415207.05	6224	66.7106443

Number of obs	= 6225
F(1, 6223)	= 1520.11
Prob > F	= 0.0000
R-squared	= 0.1963
Adj R-squared	= 0.1962
Root MSE	= 7.3228

tenure	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
age	.3093224	.0079337	38.99	0.000	.2937697	.3248751
_cons	-3.197329	.3189538	-10.02	0.000	-3.822588	-2.572069

and because the correlation between age and job tenure is positive (older workers tend to stay i in jobs longer)
the influence of this component on the bias is also upward