UNIVERSITY OF WARWICK

Summer Examinations 2015/16

Econometrics 1

Time Allowed: 3 Hours, plus 15 minutes reading time during which notes may be made (on the question paper) BUT NO ANSWERS MAY BE BEGUN.

Answer ALL EIGHT questions in Section A and ANY THREE questions from Section B. Section A carries 52 marks in total. Each of the questions in Section B is worth 16 marks. Answer Section A questions in one booklet and Section B questions in a separate booklet.

Statistical Tables and a Formula Sheet are provided. Approved pocket calculators are allowed.

Read carefully the instructions on the answer book provided and make sure that the particulars required are entered on each answer book. If you answer more questions than are required and do not indicate which answers should be ignored, we will mark the requisite number of answers in the order in which they appear in the answer book(s): answers beyond that number will not be considered.

(Continued overleaf)

Section A: Answer ALL EIGHT Questions

1. An equation for earnings for individuals born in 1970 at the age of 30 is estimated by OLS as:

Source	*SS*	*df*	*MS*		
Model	85.497	4	**(c)**	No. of obs=	2339
Residual	645.367	2334	.2765	F(4,**(a)**)=	**(b)**
Total	730.865	**(d)**	.3126	Prob >F=	0.0000
				R-squared=	0.1170
				Adj R-squared=	0.1155
				Root MSE=	**(e)**

$lw08$	$Coef.$	$Std.Err.$	t	$P>\|t\|$	$95\%Conf.Interval$	
$edu3$	.1202	.02898	4.15	0.000	**(f)**	.17704
$edu4$	.3783	.02931	12.91	0.000	.32087	.43586
$edu5$	.2675	0.1277	2.09	**(g)**	.24944	.40567
sex	-0.2289	0.0218	-10.48	0.000	-.27186	-.18618
$cons$	2.8609	0.0381	75.03	0.000	2.7861	2.9357

where $lw08$= ln(wages) when individuals were aged 38, $edu3$=1 if has a degree as the highest qualification, $edu4$=1 if has an MA or MSc as highest qualification and $edu5$=1 if has a PhD as highest qualification, sex=1 if female.

Fill in each of the answers **(a) through to (g)** **(7 marks)**

2. A researcher was interested in the effect of attractiveness on performance in exams at age 14. Data is collected on 516 boys whose facial features have been rated on a scale of 1-100 in terms of attractiveness. The following model is estimated by OLS:

$$ln(P_i) = \beta_0 + \beta_1 A_i + \beta_2 ln(Q_i) + \epsilon_i$$

where $ln(P)$ = performance in exams, A = attractiveness, Q = ability measure and ln is the natural log.

Variable	Coefficient	Standard error
CONSTANT	15.473	5.3559
A	1.6652	1.0283
$ln(Q)$	0.2745	0.1449

(Question 2 continued overleaf)

(a) At the 1% significance level test the hypothesis (i) $\beta_1 = 0$ and (ii) $\beta_2 = 0$. **(2 marks)**

(b) At the 1% significance level test the hypothesis $\beta_1 = \beta_2 = 0$, given $cov(b_1, b_2) = 0$. How do you reconcile this answer with that in (a)? **(2 marks)**

(c) Determine the effect on performance of a 100% increase in Q. **(2 marks)**

---

3. (a) You have collected cross-sectional data over three years 2008-2010 for a total of 5028 individuals living in the UK (this is called pooled cross-sectional data) and you estimate the following model:

$$ln(h_i) = \alpha + \beta_1 ln(I_i) + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 U_i + \epsilon_i$$

where h = happiness, I = income, $D_2 = 1$ if year is 2009, $D_3 = 1$ if year is 2010, and U is the UK unemployment rate. What are the problems of estimating this model by OLS? **(2 marks)**

(b) Rather than using the above model you estimate the equation

$$ln(h_i) = \alpha + \beta_1 ln(I_i) + \beta_2 t_i + \epsilon_i$$

where t=1 in 2008, 2 in 2009 and 3 in 2010. Write out the restriction(s) imposed by this model compared to a model which excludes t_i, but includes D_{2i} and D_{3i}, as defined above. **(2 marks)**

(c) A model for the performance of university degree students (who belong to one of 3 faculties Science, Humanities or Social Science) in a universal exam at the end of year 1 is specified as:

$$ln(P_i) = \alpha + \beta_1 ln(A_i) + \beta_2 Sci_i + \beta_3 Hum_i + \beta_4 Bio_i + \epsilon_i$$

where P = performance in the exam, A = A-level scores, and Sci = 1 if Science student, Hum = 1 if Humanities student, and Bio=1 if Biology student (where Biology is in the Science faculty). Interpret the coefficients of β_2 and β_4. **(3 marks)**

---

4. Consider the following equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

In each of the following cases comment on the implications for the OLS estimator of the parameter β_1:

(a) $E[\epsilon_i] = 2$. **(2 marks)**

(Question 4 continued overleaf)

(b) $cov(x_{1i}, x_{2i}) = 0.46$. **(2 marks)**

(c) $V(\epsilon_i) = \alpha_0 + \alpha_1(1/x_{1i})$. **(2 marks)**

---

5. (a) Explain the concepts of stationarity and weak dependence. Explain why stationarity and weak dependence holding are important when undertaking time series analysis. **(3 marks)**

(b) Consider the following time series process

$$y_t = \phi y_{t-1} + \epsilon_t$$

where ϵ_t is a white noise process. Discuss whether stationarity and weak dependence hold for the cases when $|\phi| < 1$ and $\phi = 1$. **(3 marks)**

---

6. Consider the following estimated relationship between two time series y and x

$$ln(y_t) = 0.546 + 0.565\Delta ln(y_{t-1}) + 0.198\Delta ln(x_t) - 0.222(ln(y_{t-1}) - 1.237 ln(x_{t-1})) + \epsilon_t$$

(a) Calculate the long run elasticity of y with respect to x. **(2 marks)**

(b) Calculate the total 0-period, 1-period, and 2-period elasticity of y with respect to x. **(5 marks)**

---

7. Consider the following panel data model

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + a_i + u_{i,t},\ t = 1, ..., T,\ i = 1, ..., N.$$

(a) What are the implications of using the random effects estimator, if a_i is **correlated** with x_i? **(2 marks)**

(b) What are the implications of using the fixed effects estimator if a_i is **uncorrelated** with x_i? **(2 marks)**

(c) Based on your answer to (a) and (b), which estimator should you use if a_i is **uncorrelated** with x_i? Explain how always using fixed effects estimation can be viewed as a conservative strategy. **(2 marks)**

8. A researcher is interested in modelling the probability of going to a private school, using data on 983 Indian boys collected in 2010:

$$Pr(private = 1) = \Phi(-2.17 + 0.031ability + 0.043wealth - 0.523lowcaste)$$

where $\Phi()$ is the standard normal cumulative distribution function, $private = 1$ if an Indian boy goes to a private school, $ability$ is a meaure of ability of the boy, $wealth$ is a meaure of household wealth, and $lowcaste = 1$ if the boy comes from a low caste family. The mean values of $private$, $ability$, $wealth$ and $lowcaste$ are 0.400, 30.5, 24.3 and 0.133, respectively.

(a) Calculate the probability that a boy goes to a private school at average characteristics of all explanatory variables. **(2 marks)**

(b) Calculate the marginal effect of an increase in wealth on the probability of a boy going to a private school. Interpret the marginal effect **(3 marks)**

(c) Calculate the marginal effect of being from a low caste family on the probability of a boy going to a private school. **(2 marks)**

Section B: Answer THREE Questions

9. In a report about the New York housing market, researchers looked at sales of 19,490 properties in New York city over the period 2008-2012. This table shows results of a regression where the dependent variable is the natural log of house prices in 2010 $ US.

Variable	Coeff	Std. err.
$Constant$	7.3093	0.0570
$Pre-war$	0.2675	0.0110
$ln(sq)$	0.7776	0.0054
$Terrace$	0.1016	0.0048

where $Pre-war$=1 if built prior to 1939, $ln(sq)$=ln(size of flat measured in square feet), $Terrace$=1 if has a terrace and RSS 984.5.

(a) What proportionate increase in square feet would it take to compensate, in terms of price, for the lack of a terrace? **(3 marks)**

(Question 9 continued overleaf)

(b) You receive data for 2013. Fitting the model to the enlarged sample of 21,456 observations gives a RSS of 1055.5. At the 1% significance level test this model against that specified in the table above. **(3 marks)**

(c) New York city has 5 boroughs (Bronx, Brooklyn, Manhattan, Queens and Staten Island) and separate dummy variables are included to control for these boroughs in addition to the variables in the table above. The RSS of this model is 982.32. At the 1% significance level test the joint significance of the borough dummies. **(2 marks)**

(d) Including a dummy for a property in either Manhattan or Staten Island yielded a RSS of 982.78. Which of the models in (c) or (d) is the most appropriate for controlling for borough effects? **(3 marks)**

(e) A new term $Terrace \times Pre-war$ is added into the model in the table above. Interpret the coefficient on this variable. **(2 marks)**

(f) There are few pre-war properties with terraces and those that do exist are highly sought after. How do you think including the variable $Terrace \times Pre-war$ will affect the coefficients on the variables $Pre-war$ and $Terrace$? **(3 marks)**

---

10. There is a large empirical literature that tries to estimate and understand the gap in educational attainment between white and black students in the US. Using data from the Early Childhood Longitudinal Study kindergarten cohort (ECLS), a model was estimated by OLS on 3290 children with an average age of 67 months (robust standard errors are reported in brackets).

$$\begin{array}{lllll} \widehat{math}_i = & -\ 0.638B_i & -\ 0.722H_i & +\ 0.150A_i & -0.503O_i \\ & (0.022) & (0.022) & (0.026) & (0.041) \end{array}$$

and $RSS = 6.543$. $math$= math test score that has been normalized to have a mean zero and a standard deviation of one; W=1 when the child is White; B=1 when the child is Black; H=1 when the child is Hispanic; A=1 when the child is Asian; and O=1 when the child is of any other race.

(a) Why is there no intercept and no dummy variable W in the model? **(3 marks)**

(b) Write out the OLS estimates if O were the excluded dummy variable instead of W. **(2 marks)**

(c) Make a careful economic and statistical interpretation of the coefficient on B. **(2 marks)**

(Question 10 continued overleaf)

(d) In order to get a better understanding of why there exists a gap in educational attainment, further variables are added to the model.

$$\begin{array}{llll} \widehat{math}_i = & \text{-}0.098B_i & \text{-}0.152H_i & +0.160A_i \quad \text{-}\ 0.11O_i \\ & (0.073) & (0.064) & (0.071) \quad\ (0.076) \\ & + 0.005Bk & + 0.032(Bk^2/1000) & \\ & (0.001) & (0.003) & \end{array}$$

and $RSS = 6.032$. Bk are the number of children's books the child has at home at the moment of the survey.

(i) Explain precisely, stating appropriate conditions, why the coefficient on B might be different in this model compared to the earlier model. **(2 marks)**

(ii) For how many books is the math score associated the highest? **(2 marks)**

(iii) Discuss clearly what, if any, policy implications there are given your answer to question (d)(ii). **(2 marks)**

(iv) Estimating the model in (d) above for families with books and those without books yields RSS of 4.799 and 1.144, respectively. At the 1% significance level, test this model against that in (d). **(3 marks)**

---

11. In recent years there has been a lot of discussion about how to tackle the problem of smoking. One proposed solution is to increase the tax rate per pack of cigarettes. Critics of that idea, however, point out that this will not work as smoking is addictive, implying that people's consumption is unresponsive to prices. To explore this question empirically, you are given some results from a research paper that is devoted to this question. The estimated model is:

$$ln(Q_i) = \beta_0 + \beta_1 ln(P_i) + \epsilon_i \tag{1}$$

where Q is the quantity of cigarette packs purchased and P is the price per pack of cigarettes and ln is the natural log. They find that the estimated coefficient on β_1 is -0.661 with a standard error of 0.193.

(a) The claim that the demand for cigarettes is unresponsive to prices implies a prediction on β_1. Specify the null hypothesis and test this hypothesis. **(2 marks)**

(b) Another criticism about increasing taxes on cigarettes is that it would be unfair towards the poor: smoking is much more prevalent among the poor; increasing taxes would increase their expenditure and make them even poorer. However, a counter-argument is that poor people are much more responsive to prices than rich people.

(Question 11 continued overleaf)

Specify a new model that can shed light about these arguments. Be specific about the variables you need and make clear predictions, wherever possible, about the sign of the coefficients, and the implied price responsiveness for rich and poor people. **(3 marks)**

(c) Why might the authors be worried about using OLS estimation of equation (1)? **(2 marks)**

(d) Suppose that two instruments, z_1 and z_2, are available. Writing out the appropriate equations, show how you would test for endogeneity of $ln(P)$ in equation (1) using these instruments. **(3 marks)**

(e) The test statistic for the test of exogeneity of the instruments z_1 and z_2 is 4.93. Writing out the appropriate equations explain clearly how you test for exogeneity. At the 5% level what do you conclude? **(3 marks)**

(f) Sales tax and cigarette-specific tax in the different States in the US are used as instruments. Comment on the appropriateness of these instruments. **(3 marks)**

---

12. A researcher is interested in the effect of international aid payments on GDP. They collect data sampled annually from 2000-2014 on 52 countries. They begin by estimating the following equation by pooled OLS

$$ln(\widehat{GDP})_{i,t} = 3.14 - 0.546 ln(aid)_{i,t} - 0.214 africa_{i,t} - 0.364 asia_{i,t} \tag{2}$$

where $ln(GDP)_{i,t}$ is the value of the log of GDP in country i at time t, $ln(aid)$ is the value of the log of aid payments (measured in \$m) made to country i in time t and $africa$ and $asia$ are two dummy variables that take a value of 1 if a country is in Africa or Asia, respectively.

(a) What is the estimated elasticity of GDP with respect to aid payments? Does this elasticity make sense economically? Explain why pooled OLS may not be an appropriate technique when dealing with panel data sets. **(3 marks)**

(b) The researcher believes that unobserved corruption may play an important role in how aid spending translates into GDP growth. As such she performs a fixed effects estimation yielding the following results.

$$ln(\widehat{GDP})_{i,t} = 1.96 + 1.014 ln(aid)_{i,t} + a_i \tag{3}$$

Why have the variables $africa$ and $asia$ not been included in equation (3)? **(2 marks)**

(Question 12 continued overleaf)

(c) Assuming increased aid payments provide additional incentive for corruption, which is detrimental to a country's economy, explain why the coefficient on $ln(aid)$ has changed in equation (3) compared to that in equation (2). **(3 marks)**

(d) Do you think it is sensible for the researcher to use fixed effects estimation? Explain the Hausman test and how it is used to select between fixed effects and random effects estimation. Suppose the researcher runs the Hausman test and fails to reject the null hypothesis. Which estimation technique should she use? **(4 marks)**

(e) The researcher realises that due to the business cycle a more sensible model includes the first lag of the dependent variable in the model as:

$$ln(GDP)_{i,t} = \alpha + \beta_1 ln(aid)_{i,t} + \beta_2 ln(GDP)_{it-1} + a_i + \epsilon_{i,t}. \tag{4}$$

Explain what are the potential problems of estimating this equation. **(4 marks)**

---

13. You are interested in modelling the relationship between GDP per capita and life expectancy in a developing nation. Suppose you have 78 quarterly observations for both life expectancy ($LIFE$) and GDP per capita (GDP).

(a) You estimate the following regression by OLS

$$LIFE_t = \beta_0 + \beta_1 GDP_t + \beta_2 GDP_{t-1} + \beta_3 GDP_{t-2} + \beta_4 LIFE_{t-1} + \epsilon_t. \tag{5}$$

You are worried about the potential presence of serial correlation in the error term ϵ_t. Explain how you would undertake the Breusch-Godfrey test for the presence of serial correlation of order 4 (F-test version) at the 5% significance level. State the form of any regressions you need to do the test, the null and alternative hypothesis and the distribution of the test statistic under the null. You obtain an F-test of 3.25, what do you conclude? **(4 marks)**

(b) You remember that it is important to check the order of integration of time series data before undertaking any form of regression analysis. Explain how you would use the Augmented Dickey-Fuller (ADF) test to determine the order of integration of $LIFE$. Be sure to specify any regression equations you estimate, null and alternative hypotheses, decision rule and the form of the equation used. **(4 marks)**

(c) Suppose you perform the ADF test with a constant and trend for the two series GDP and $LIFE$ and obtain test statistics of -2.86 and -3.36, respectively. What do you conclude about the order of integration of these series? **(2 marks)**

(d) You test the residuals of the solved long-run version of equation (5) for stationarity and find an ADF test statistic value of -3.969. Discuss the form and appropriateness of the ADF test on the residual to test whether there is any evidence of cointegration in between the two series. **(3 marks)**

(Question 13 continued overleaf)

(e) Irrespective of your result in (d) you estimate an ECM model for both $LIFE$ and GDP as:

$$
\begin{array}{lllll}
\Delta \widehat{LIFE}_t = & 0.013 & + 0.122\Delta GDP_{t-1} & + 0.150\Delta LIFE_{t-1} & - 0.322\hat{\eta}_{t-1} \\
 & (0.009) & (0.042) & (0.076) & (0.141) \\
\Delta \widehat{GDP}_t = & 0.023 & + 0.472\Delta GDP_{t-1} & + 0.217\Delta LIFE_{t-1} & + 0.103\hat{\eta}_{t-1} \\
 & (0.011) & (0.149) & (0.162) & (0.141)
\end{array}
$$

where $\hat{\eta}_t$ are the long-run residuals from equation (5). For both models interpret the coefficient on the error correction term, $\hat{\eta}_{t-1}$. **(3 marks)**

(End)