EC2260

UNIVERSITY OF WARWICK

Summer Examinations 2018/19

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** (52 marks total) and **ANY THREE** questions from **Section B** (16 marks each). Answer Section A questions in one booklet and Section B questions in a separate booklet.

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

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.

Section A: Answer ALL EIGHT Questions

---

1. A model for food expenditure for households in 2014 was estimated by OLS on 242 single occupant households (standard errors reported in brackets).

$$\widehat{ln(C)}_i = \underset{(0.055)}{0.226} + \underset{(0.045)}{0.835} ln(E)_i - \underset{(0.022)}{0.058} G_i - \underset{(0.014)}{0.022} A_i$$

RSS=202.773, TSS=219.556, where C = Food expenditure, E = Total nondurable consumption, $G = 1$ if household is female and A = Age of individual and ln is the natural log.

(a) Calculate the p-value for the test that the coefficient on A is equal to zero. **(2 marks)**

(b) At the 1% significance level, test the joint significance of the slope coefficients. What is the approximate p-value of this test? **(3 marks)**

(Question 1 continued overleaf)

(c) For the model:

$$\widehat{ln(C/E)}_i = \underset{(0.048)}{0.176} - \underset{(0.015)}{0.048}G_i - \underset{(0.013)}{0.028}A_i$$

calculate the RSS of this model. **(2 marks)**

---

2. How does class attendance impact final year 2 score? To try and answer this question the following regression was estimated, based on a sample of 152 undergraduate students (standard errors are in brackets):

$$\widehat{Score}_i = \underset{(3.41)}{51.32} + \underset{(0.34)}{1.21}Att_i + \underset{(2.34)}{3.12}M_i + \underset{(0.21)}{1.12}yr1_i$$

$R^2 = 0.35$ and where $Score_i$ is final year 2 score; Att_i is number of tutorial attended (0-30); $M_i = 1$ if male and zero otherwise; $yr1_i$ is the final year 1 score.

(a) If you removed the final year 1 score ($yr1$) variable from the regression how do you think it would change the coefficient on attendance (Att)? Briefly explain. **(2 marks)**

(b) Do you think the estimated coefficient on attendance is likely to represent the causal effect of attendance on test score? Briefly explain. **(2 marks)**

(c) A regression was estimated (standard errors are in brackets):

$$\widehat{Att}_i = \underset{(3.41)}{24.72} - \underset{(1.34)}{4.44}Early_i - \underset{(0.98)}{2.13}Late_i - \underset{(1.16)}{2.67}M_i + \underset{(0.17)}{0.55}yr1_i$$

where $Early_i = 1$ if the class is early in the day and zero otherwise; $Late_i = 1$ if the class is late in the day and zero otherwise. Are the variables for the time of day the class is run likely to be appropriate instruments for attendance? **(3 marks)**

(Continued overleaf)

3. Having looked at an analysis of test scores and class size, the Secretary of State for Education stated,"Student performance depends on class size. Students do better when class sizes are smaller; however, there is no evidence of improved performance associated with reducing class size below 20 students, nor any worsening of performance in having class size above 34".

(a) You have data on student performance and the size of their class in Key Stage 2 (age 11). Consider a regression of test score for each individual in the Key Stage 2 test, T, on the size of their class at age 11 dummy variables, CS_j for class size j. Write out the most parsimonious model that you would estimate which is consistent with the statement of the Secretary of State for Education above. Interpret the coefficient on the first CS variable in your model. **(3 marks)**

(b) A more general model of test scores, T, on class size (where no class size is smaller than 15 or in excess of 40) is written as:

$$T_i = \alpha + \sum_{j=16}^{40} \beta_j CS_{ji} + \epsilon_i \tag{1}$$

Write out the restrictions which the model in (a) imposes compared to equation (1). **(3 marks)**

---

4. A model of the school-level pass rate (as a percent) on year 8 maths tests is written as:

$$maths_i = \alpha + \beta_1 ln(expend_i) + \beta_2 ln(enroll_i) + \beta_3 ln(chprg_i) + \beta_4 / ln(expend_i) + \epsilon_i$$

where $maths$ = Maths score (out of 100), $expend$ = Expenditure per pupil (varying between 40 and 200), $enroll$=Number of students enrolled in the school, $chprg$=% on free school meals and ln is the natural log. Using UK data for 824 schools yielded the following results:

$$\widehat{maths}_i = 12.45 + 5.28 ln(expend_i) - 8.15 ln(enroll_i) + 3.674 ln(chprg_i) + 82.139 / ln(expend_i)$$

$RSS = 56.28$

(a) Carefully sketch how $maths$ varies with $ln(expend)$ in the estimated model above. **(3 marks)**

(b) At the 5% significance level test the hypothesis that the response of $maths$ to $ln(expend)$ is 3.0, for a school with an expenditure of 150, given the coefficient variance-covariance matrix: **(3 mark)**

$$\begin{pmatrix} 16.466 & -4.201 & -3.119 & -1.016 & 10.01 \\ - & 2.281 & 3.321 & 4.018 & 25.909 \\ - & - & 12.468 & -6.192 & 8.447 \\ - & - & - & 5.572 & 2.811 \\ - & - & - & - & 121.54 \end{pmatrix}$$

(Continued overleaf)

5. Monthly data was collected over the period from 2006m2 to 2018m9 on the annual inflation rate in the UK, defined as $100 \times ln(p_t/p_{t-12})$, where p_t is the Consumer Price Index. Plotting the series, the autocorrelation funcation (ACF) and partial autocorrelation function (PACF) for the annual inflation rate:

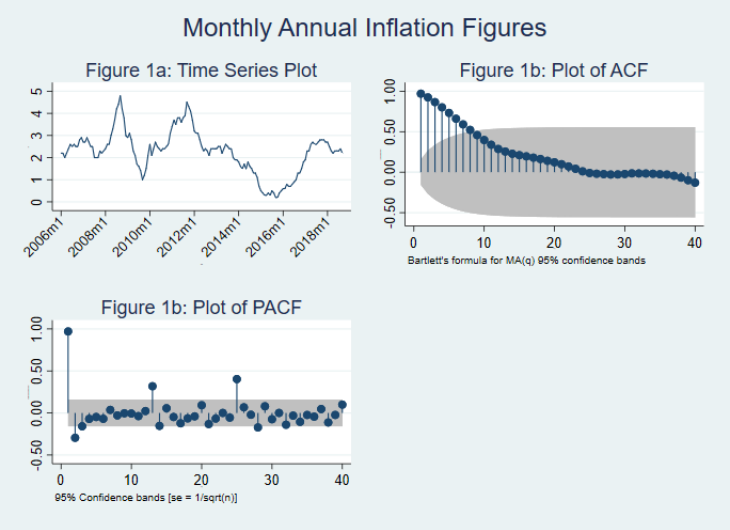
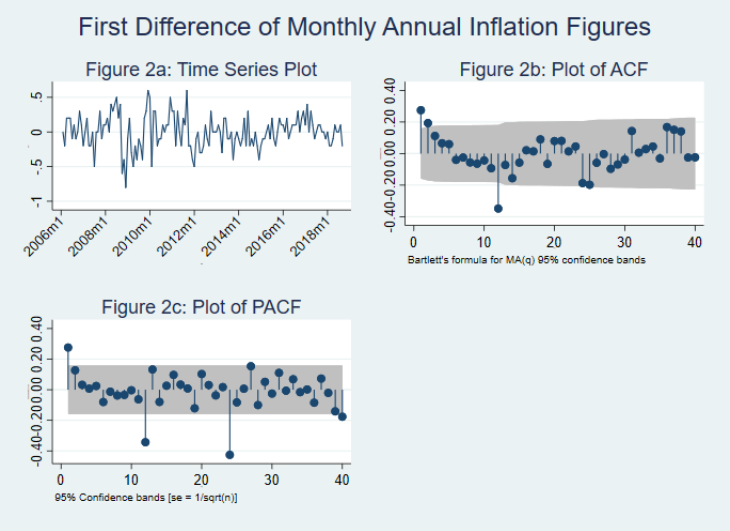


Figure 1:

Upon taking first differences of the annual inflation rate we get the plot, the ACF and the PACF as:



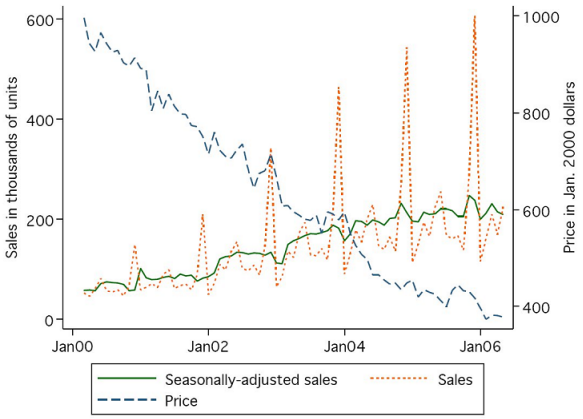
(Question 5 continued overleaf)

Using the information provided by these plots explain how you would test for the presence of a unit root in UK annual inflation over the period 2006-2018. Be careful to specify the model(s) you would estimate (noting the deterministic components and the lag length), the null and alternative hypotheses, the critical value used and any limitations of the test as applied to this series. **(7 marks)**

.....

6. Figure 3 below is from a recent research paper analyzing prices and sales of camcorders in the US in the 2000s (remember this is pre-mobile phones being able to record videos).

Figure 3: Prices and sales for camcorders



(a) Look at the long dashed line representing prices for camcorders. What are the general patterns of prices in this period? What do you think explains this pattern? **(2 marks)**

(b) Look at the dotted line representing (unadjusted) sales of camcorders. What are the general patterns of sales in this period? What do you think explains these patterns? **(2 marks)**

(c) Suppose you wanted to estimate a demand curve for camcorders. How would you account for each of the general patterns in prices and (unadjusted) sales in your econometric regression? **(2 marks)**

(Continued overleaf)

7. This question asks you about the demand for Zantac (Ranitidine), one of the leading pharmaceutical drugs used to treat ulcers and other gastro-intestinal (GI) health problems. You have data on over 70,000 uses of various GI drugs across 33,614 ulcer sufferers collected over a period of 31 months from June 1990 to December 1992. A table of summary statistics for this data is:

Variable	Obs	Mean	Std. Dev.	Min	Max
zantac	70362	.6275831	.483452	0	1
age	70362	56.21425	16.1339	15	85
female	70362	.5468008	.4978084	0	1
month	70362	381.8619	8.525145	365	395

where zantac $= 1$ when patient uses Zantac in month t and 0 when the patient uses some other GI drug, age is the age of individual at the beginning of treatment, female $=1$ if the individual is female, and month is a time trend. Estimating a Linear Probability Model (LPM) using pooled OLS (with robust standard errors) yields the following results:

zantac	Coef.	Std. Err.	t	P>|t|
age	.0039331	.0001831	21.48	0.000
female	-.0286011	.0062511	-4.58	0.000
month	.0015961	.0002652	6.02	0.000
cons	-.1873668	.1009101	-1.86	0.063

while estimating it using Fixed Effects (FE) yields the following results:

zantac	Coef.	Std. Err.	t	P>|t|
age	(omitted)			
female	(omitted)			
month	.0007903	.000309	2.56	0.011
cons	.325793	.1179865	2.76	0.006

(a) Explain why Stata cannot estimate the effects of gender and age in this specification. **(2 mark)**

(b) Suppose each fixed effect, a_i, measures "tastes for Zantac". Explain why the coefficient on month was cut almost in half in our FE results compared to the LPM results. **(2 marks)**

(Question 7 continued overleaf)

(c) What do these results suggest about the average probability a patient uses Zantac late in the sample (e.g. December 1992) compared to early in the sample (e.g. June 1990)? **(2 marks)**

---

8. Country X has 67 regions and 28 of these regions run along the border of a neighbouring country, Y (where there are almost no border controls). At the start of 2011 country Y imposed a minimum wage for individuals working on farms. A researcher wants to use this exogenous shock to see whether there was any effect on child labour supply in country X. The researcher estimates the following regression over the period 2009-2012 for children living in country X:

$$ln(L)_{it} = \underset{(1.011)}{3.036} + \underset{(0.021)}{0.056}R_{it} - \underset{(0.026)}{0.037}Y_{it} + \underset{(0.055)}{0.154}Y_{it}R_{it} + e_{it}$$

where L is hours of labour supply per month, $R = 1$ for those regions bordering country Y, 0 otherwise and $Y = 1$ if year is either 2011 or 2012. The numbers in brackets are standard errors and ln is the natural log.

(a) Interpret ALL the coefficient estimates in the equation above. **(5 marks)**

(b) What is the average number of hours of labour supply by children in (i) regions bordering country Y pre-2011, (ii) regions bordering country Y post-2010, (iii) regions NOT bordering country Y pre-2011, (iv) regions NOT bordering country Y post-2010? **(2 marks)**

(Continued overleaf)

Section B: Answer THREE Questions
Please use a separate booklet

9. You have a dataset containing information on 1089 births, indexed by i. The dataset contains the following variables: $lbwt$ = natural log of birth weight; $male = 1$ if the baby is male, otherwise = 0; nch = number of children the mother has (including this one); $cigspreg$ = average number of daily cigarettes smoked during pregnancy; $faminc$ = Family income (in US$000s); $mothedu$ = number of years in full time education of child's mother.

The following table shows the regression output from a linear regression of $lbwt$ on some of the explanatory variables listed above.

Some information is missing from the Table:

Dependent Variable: $lbwt$

Variable	Coeff	Std. err.	t-ratio	P $>\|t\|$
$Cons$	4.7280	0.0163	290.06	0.0000
nch	0.0141	0.0061	2.298	[B]
$faminc$	0.0006	[A]	[E]	0.0404
$male$	[C]	0.0307	5.451	0.0000
$cigspreg$	-0.0051	0.0010	-5.376	0.0000
$R^2 =$	0.0429	$\bar{R}^2 =$	0.0394	
$RSS =$	33.5303			
$F - stat. =$	12.15	$Prob(F) =$	[D]	

(a) Fill in the values for the terms [A] to [E]. **(5 marks)**

(b) Interpret the estimated value for $faminc$. Does this seem to be an important effect? **(2 marks)**

(c) You have heard that on average female babies weigh about 3,500 grams. How much would you expect male babies to be at birth? **(2 marks)**

(d) You now estimate the following model:

$$lbwt_i = \beta_0 + \beta_1 nch_i + \beta_2 male_i + \beta_3 cigspreg_i + \epsilon_i \quad (2)$$

$R^2 = 0.0325$

(Question 9 continued overleaf)

(i) You want to test the residuals in equation (2) for heteroskedasticity (at the 1% significance level). You suspect that the variables $cigspreg$ and nch are the variables that cause any heteroskedasticity. Detail the testing procedure you would undertake to test for the presence of heteroskedasticity and then test for the presence of heteroskedasticity. Assume that the auxiliary regression you estimate produces the following regression statistics: $R^2 = 0.0017236$. **(3 marks)**

(ii) Considering your finding in part d(i), what can you say about the distribution of the coefficient estimator for nch? Ensure you state any extra assumptions you make. **(2 marks)**

(iii) You want to test for incorrect functional form for equation (2) (at the 1% significance level). Detail the testing procedure you would undertake to do a RESET test with up to the fourth power of the fitted values and test for incorrect functional form. Assume that the auxiliary regression you estimate produces the following regression statistics: $RSS = 33.3938$. **(2 marks)**

(Continued overleaf)

10. A researcher is interested in migration intentions and asks a random sample of 1268 individuals whether they would consider emigrating from the UK in the next 5 years. The variables used are Emig = 1 if individual would consider emigrating, 0 otherwise; Female = 1 if individual is female, 0 otherwise; Child = 1 if individual has at least 1 child at home; Age = Age of individual; Income = Annual income, with summary statistics:

Variable	Mean	Std. Dev.	Minimum	Maximum
Emig	0.210	0.407	0.00	1.00
Female	0.432	0.495	0.00	1.00
Child	0.322	0.467	0.00	1.00
Age	35.23	10.34	25.0	50.0
Age*Age	1550.0	262.8	625.0	2500.0
ln(Income)	4.182	1.38	1.07	6.86

The PROBIT estimation results are:

Iteration 0: log likelihood = -2139.7712 (obtained assuming slope coefficients are zero)

Iteration 1: log likelihood = -2037.5784

Iteration 2: log likelihood = -2037.2528

Iteration 3: log likelihood = -2037.2528

Variable	Coefficient	Std. Error
Female	-0.132	0.038
Child	0.075	0.027
Age	0.105	0.033
ln(Income)	0.013	0.002
Age*Age	-0.0016	0.0006
Constant	-1.997	0.562

(a) Calculate the probability of expressing a desire to emigrate at mean values for all explanatory variables. **(2 marks)**

(b) Calculate the marginal effect of having children on the probability of expressing an interest in emigrating (at mean values of other variables). **(3 marks)**

(c) Sketch the marginal effects for age on the probability of expressing an interest in emigrating at average values of the other variables. At approximately what age does this marginal effect change sign? **(5 marks)**

(d) At the 1% significance level, test the null hypothesis that all slope coefficients are zero, being careful to specify the null hypothesis, alternative hypothesis and the distribution of the test statistic. **(3 marks)**

(Question 10 continued overleaf)

(e) How might you evaluate the appropriateness of this model reported above? **(3 marks)**

.....

11. A university is trialling online courses on its 3 year degree programmes and is interested in understanding teacher evaluations which students complete at the end of the course. Each teacher is trained in the same way for these on-line courses and is then responsible for around 30 students.

The course is delivered to many thousands of students and in total there are around 800 tutors. Let $E_i \in [0, 100]$ represent the evaluation of teacher i where 0 means completely unsatisfied and 100 is completely satisfied. All course material is online: there are no lectures or classes and the only correspondence is via non-anonymous forums between the students and the teacher. The following regression was estimated on an independent random sample of 200 tutors.

$$\hat{E}_i = \underset{(4.32)}{65.51} + \underset{(3.41)}{10.11} M_i + \underset{(3.32)}{6.21} EXP_i + \underset{(1.32)}{5.40} YR2_i - \underset{(3.32)}{4.51} YR3_i$$

where $M_i = 1$ if the teacher is male and zero otherwise. $EXP_i = 1$ if the teacher has 2 or more years experience at university level teaching and zero otherwise; $YR2_i = 1$ for year 2 courses and zero otherwise; $YR3_i = 1$ for year 3 courses and zero otherwise.

(a) Give an interpretation to ALL coefficients in the above model. **(4 marks)**

(b) Diagrammatically represent the relationship between evaluation and year of study for males with less than two years of experience. On the same diagram show the same relationship for females with less than two years experience. **(3 marks)**

(c) The university was surprised to see the positive and significant sign on the male coefficient. The regression was re-estimated with a new variable: $Q_i = 1$ if the course the teacher was associated with was a quantitative course and zero otherwise. The coefficient estimates on Q and M in the new regression are 5.12 and 5.06, respectively. What does this imply about the relationship between male and quantitative courses? **(3 marks)**

In the equation described in (c) the coefficient on male was still statistically significant at the 1% significance level. To get experimental variation in gender in the next year of the courses it was decided NOT to reveal the REAL name of the teacher. Instead the name of the teacher the students saw for their group was a randomised (male/female) name. Two models were then specified:

$$E_i = \beta_0 + \beta_1 M_i + \epsilon_i$$

$$E_i = \delta_0 + \delta_1 M_i + \delta_2 EXP_i + \delta_3 YR2_i + \delta_4 YR3_i + \epsilon_i$$

(Question 11 continued overleaf)

(d) If the randomisation worked would you expect the OLS estimate of the coefficient associated with unknown parameter δ_1 to be greater than, less than, or equal to the unknown parameter β_1? Briefly explain. **(3 marks)**

(d) If the randomisation worked would you expect the OLS estimate of the standard error on the coefficient associated with unknown parameter δ_1 to be greater than, less than, or equal to the OLS estimate of the standard error on the coefficient associated with unknown parameter β_1? Briefly explain. **(3 marks)**

.....

12. Tables 1 and 2 contain the results for an imports equation for Germany using quarterly data over the period 1980.2-1993.3,

TABLE 1 Dependent Variable: LM Sample(adjusted): 1980:2 1993.3

Variable	Coeff.	Std. Err.	t-Stat	P-value
C	1.6304	0.2973	5.483	0.000
LM(-1)	0.7642	0.0512	14.92	0.000
LRY	0.5087	0.1022	4.975	0.000
LRY(-1)	0.2896	0.1122	2.581	0.013
LRP	-1.0848	0.1670	-6.496	0.000
LRP(-1)	0.2567	0.2836	0.905	0.307

R-squared	0.8505	Mean dep. var	9.83587
R-bar-squared	0.8349	S.D. dep. var	0.04898
S.E. of regression	0.0199	Sum squared resid	0.0191
F-statistic	54.40	Prob(F-statistic)	0.0000

where, LM=Natural log of imports, LRY=Natural log of real GDP, LRP=Natural log of relative prices. All variables are seasonally adjusted.

(Question 12 continued overleaf)

TABLE 2 Dependent Variable: LM Sample(adjusted): 1980:2 1993.3

Variable	Coeff.	Std. Err.	t-Stat	P-value
C	1.0366	0.2013	5.149	0.000
LM(-1)	0.2644	0.1052	8.213	0.000
LRY	0.5197	0.1141	4.555	0.000
LRY(-1)	0.2191	0.1027	2.133	0.039
LRP	-1.2848	0.2040	-3.191	0.003
LRP(-1)	0.5577	0.2111	2.642	0.011
D1	0.1026	0.0578	2.640	0.011
D2	0.1534	0.0618	2.482	0.017
D3	0.1945	0.0609	3.194	0.003
D4	0.1923	0.0619	3.107	0.003
D5	0.2011	0.0677	2.970	0.005
D6	0.2127	0.0601	3.539	0.001

where R-squared=0.9012 and D1=1 for 1991:2, D2=1 for 1991:3, D3=1 for 1992:1, D4=1 for 1992:4, D5=1 for 1993:1, D6=1 for 1993:3.

TABLE 3

RESET(4th order) Test	F-stat	1.789	P-value	0.1881
Breusch-Godfrey Serial Correlation Test (4th order)	F-stat	2.021	P-value	0.1135
ARCH Test (1-4)	F-stat	1.758	P-value	0.1588

(a) At the 5% significance level, test for the joint significance of the additional variables included in Table 2. What test are you undertaking? **(2 marks)**

(b) Using the results in Table 2, calculate (i) the contemporaneous, (ii) 1-period, (ii) 2-period, (iii) long-run elasticity of imports with respect to real GDP. **(4 marks)**

(c) Write out the long-run model for imports and comment on the appropriateness of coefficient estimates. **(2 marks)**

(d) Give a brief description of the diagnostic tests reported in the Table 3 (which relate to the results from Table 2) and hence comment on the appropriateness of the model. **(4 mark)**

(e) Explain how would you test for the presence of a cointegrating equation for imports using the model in Table 2. **(4 marks)**

(Continued overleaf)

13. You are interested in the causal effect of X on Y for a sample of $i = 1, ..., n$ individuals and for $t = 1, ..., T$ time periods. Consider the following equations:

$$Y_{it} = \gamma_0 + \gamma_1 X_{it} + \gamma_2 Z_{1it} + \mu_{1i} + \epsilon_{1it} \tag{3}$$

$$X_{it} = \pi_0 + \pi_1 Z_{1it} + \pi_2 Z_{2it} + \mu_{2i} + \epsilon_{2it} \tag{4}$$

In all parts of the question below, assume that all variables Y_{it}, X_{it}, Z_{1it} and Z_{2it} are continuous variables and, further, assume Z_{1it} and Z_{2it} are exogenous.

(a) If $E[\mu_{1i} + \epsilon_{1it}|X_{it}] = 0$, briefly explain how would you estimate equation (3). **(3 marks)**

(b) If $E[\mu_{1i}|X_{it}] \neq 0$, but $E[\epsilon_{1it}|X_{it}] = 0$, briefly explain how would you estimate equation (3). **(3 marks)**

(c) If $E[\mu_{1i}|X_{it}] \neq 0$, but $E[\epsilon_{1it}|X_{it}] \neq 0$, briefly explain how would you estimate equation (3). **(3 marks)**

(d) If $E[\mu_{1i} + \epsilon_{1it}|X_{it}] = 0$ and an interaction term $(X_{it} \times Z_{1it})$ was added to equation (3). How would your interpretation of γ_1 and γ_2 change? **(3 marks)**

(e) Discuss the Hausman test used to compare Fixed Effects and Random Effects models of equation (3). Explain how this test can be carried out in practice. **(4 marks)**

(End)