SUMMER TERM 2018

ECON2007: QUANTITATIVE ECONOMICS AND ECONOMETRICS

TIME ALLOWANCE: 3 hours

Answer ALL TWO questions from Part A and answer ONE question from Part B.

Questions in Part A carry 60 *per cent of the total mark and questions in Part B carry* 40 *per cent of the total. Tables for the normal and F-distribution are at the end of the examination paper.*

In cases where a student answers more questions than requested by the examination rubric, the policy of the Economics Department is that the student's first set of answers up to the required number will be the ones that count (not the best answers). All remaining answers will be ignored.

PART A

Answer all questions from this section.

A.1 Consider the following regression model *without an intercept*,

$$y_i = \beta x_i + u_i, \tag{1}$$

where we have observed (y_i, x_i) while u_i is unobserved, $i = 1, ..., n$. We are interested in doing inference regarding the unknown parameter β.

(a) Write up the sum of squared residuals for the model, and derive the first-order condition that the OLS estimator, $\hat{\beta}$, of β has to satify. Show that $\hat{\beta}$ is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{2}$$

(b) State conditions on the model and the data under which the OLS estimator will be consistent.

(c) Derive the large-sample distribution of the OLS estimator under the conditions you provided in (b). In doing so, (i) state which asymptotic limit results you rely on and (ii) where in your derivations you employ these limit results together with the conditions you provided in (b).

(d) Provide a consistent estimator of the standard errors of the OLS estimator.

(e) Suppose that data is in fact generated by the following regression model,

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $\beta_0 \neq 0$. Will the $\hat{\beta}$ in eq. (2) be a biased estimator of β_1? If so, derive the bias of the estimator.

A.2 In Table 10.1 on the next page you find seven different estimated regressions numbered (1)-(7) for the effect of drunk driving laws on traffic deaths (reprinted from Stock and Watson, 2007, *Introdution to Econometrics*, Pearson Education). The data are for the "lower 48" U.S. states, for 1982-88. The traffic fatality rate is the number of traffic deaths in a given state in a given year, per 10,000 people living in that state in that year. The beer tax is the tax on a case of beer in $. Drinking age variables are binary indicating whether the legal drinking age is 18, 19, or 20. Drinking age 21 is the excluded dummy variable.

(a) New Jersey has a population of 8.1 million people. Suppose that New Jersey increased the tax on a case of beer by $1. Use the results in column (4) to predict the number of lives that would be saved over the next year in New Jersey. Construct a 99% confidence interval for your answer.

(b) Explain what is meant by "time effects" which are included in the regression in column (5). Do they seem to matter? Explain.

(c) A researcher conjectures that the unemployment rate has a different effect on traffic fatalities in the Western states compared to the other states. How would you test this hypothesis? (Be clear about the specification of the regression and the statistical test you would use.)

(d) Write up the two regression models whose estimates are reported in columns (1) and (2). Explain which type of omitted variable bias the model in column (2) controls for relative to the model in column (1). Which type of omitted variable bias does the model in column (2) not control for?

(e) Based on the estimates reported in columns (1) and (2), does the aforementioned omitted variable bias seem to be a concern in this application?

(f) Provide two different estimators of the regression in column (2). Is one preferable to the other?

(g) What is meant by "Clustered standard errors?"? Do they seem to matter in this application?

PART B

Answer ONE question from this section.

B.1 In "Virtual Classrooms: How Online College Courses Affect Students" (*American Economic Review*, Vol.107(9), 2017), Eric Bettinger, Lindsay Fox, Susanna Loeb and Eric Taylor explore the effects of taking a college course online, instead of in-person, on outcomes such as student achievement and progress in college. In their data, each course is offered both online and in-person, and each student enrolls in either an online section or an in-person section. Both sections are identical otherwise: they follow the same syllabus and textbook and class sizes are approximately the same. Evaluations and marking standards are also homogeneous.

TABLE 10.1 Regression Analysis of the Effect of Drunk Driving Laws on Traffic Deaths

Dependent variable: traffic fatality rate (deaths per 10,000).

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Beer tax	0.36** (0.05)	−0.66** (0.20)	−0.64* (0.25)	−0.45* (0.22)	−0.70** (0.25)	−0.46* (0.22)	−0.45 (0.32)
Drinking age 18				0.028 (0.066)	−0.011 (0.064)		0.028 (0.076)
Drinking age 19				−0.019 (0.040)	−0.078 (0.049)		−0.019 (0.054)
Drinking age 20				0.031 (0.046)	−0.102* (0.046)		0.031 (0.055)
Drinking age						−0.002 (0.017)	
Mandatory jail?				0.013 (0.032)	−0.026 (0.065)		0.013 (0.018)
Mandatory community service?				0.033 (0.115)	0.147 (0.137)		0.033 (0.144)
Mandatory jail or community service?						0.039 (0.084)	
Average vehicle miles per driver				0.008 (0.008)	0.017 (0.010)	0.009 (0.008)	0.008 (0.007)
Unemployment rate				−0.063** (0.012)		−0.063** (0.012)	−0.063** (0.014)
Real income per capita (logarithm)				1.81** (0.47)		1.79** (0.45)	1.81* (0.69)
State effects?	no	yes	yes	yes	yes	yes	yes
Time effects?	no	no	yes	yes	yes	yes	yes
Clustered standard errors?	no	no	no	no	no	no	yes
***F*-statistics and *p*-values testing exclusion of groups of variables:**							
Time effects = 0			2.47 (0.024)	11.44 (< 0.001)	2.28 (0.037)	11.62 (< 0.001)	8.64 (< 0.001)
Drinking age coefficients = 0				0.48 (0.696)	2.09 (0.102)		0.30 (0.825)
Jail, community service coefficients = 0				0.17 (0.845)	0.59 (0.557)		0.28 (0.758)
Unemployment rate, income per capita = 0				38.29 (< 0.001)		40.15 (< 0.001)	25.88 (< 0.001)
$\overline{R}^2$	0.090	0.889	0.891	0.926	0.893	0.926	0.926

These regressions were estimated using panel data for 48 U.S. states from 1982 to 1988 (336 observations total), described in Appendix 10.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The individual coefficient is statistically significant at the *5% level or **1% significance level.

(a) Let D_i denote whether a student enrols in an online course ($D_i = 1$) or in an in-person section ($D_i = 0$). The authors have information on whether the in-person section is offered at the student's home campus (Avail_i) and how far the student is to his/her local campus (Dist_i). The former variable is a dummy $= 1$ if the student's local campus offers an in-person section and $= 0$, otherwise. The latter variable records distances in tens of miles. They allow those variables to affect a student's decision to take an online or in-person section. More precisely, the authors consider the following (*Linear Probability*) model for the decision to take an online or in-person section:

$$\mathbb{P}(D_i = 1|\text{Avail}, \text{Dist}) = \beta_0 + \beta_{\text{Avail}}\text{Avail}_i + \beta_{\text{Dist}}\text{Dist}_i + \beta_{\text{Avail}\times\text{Dist}_i}\text{Avail}_i \times \text{Dist}_i. \quad (3)$$

(The authors also allow other student characteristics to influence their decision, but I omit those here for simplicity.) Write down the expression for the Partial Average Effect (PAE) for a marginal change in Dist_i. Explain how the computation of the Partial Average Effect (PAE) compares with the computation of the Partial Effect at the Average (PEA) for this particular model.

(b) The model above can be represented as:

$$\begin{aligned} D_i &= 1, \text{ if } \beta_0 + \beta_{\text{Avail}}\text{Avail}_i + \beta_{\text{Dist}}\text{Dist}_i + \beta_{\text{Avail}\times\text{Dist}_i}\text{Avail}_i \times \text{Dist}_i - U_i \geq 0 \\ &= 0, \text{ otherwise.} \end{aligned}$$

If the error term U_i is uniformly distributed on the interval $[0, 1]$, one obtains the Linear Probability Model above. Suppose instead that the error term U_i follows a normal distribution with zero mean and unit variance (i.e., $U_i \sim \mathcal{N}(0, 1)$). Write down the (log-)likelihood for this model and the expression for the PAE of a marginal change in Dist_i.
HINT: Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the CDF and PDF of a standard normal random variable, respectively.

(c) The main goal of the paper is to estimate the effect of online courses on student outcomes (Y_i) (for example, achievement or progress in college). To address this question, the authors focus on a model similar to:

$$Y_i = \delta_0 + \delta_D D_i + \delta_{\text{Avail}}\text{Avail}_i + \delta_{\text{Dist}}\text{Dist}_i + V_i,$$

where D_i, Avail_i and Dist_i are defined as above and V_i is an unobservable error term. (Again, the authors also allow other student characteristics to influence the outcome Y_i, but I omit those here for simplicity.) The authors worry that the decision to enrol in an online course may be correlated with unobservable factors encoded in V_i and use the

interaction $\text{Avail}_i \times \text{Dist}_i$ as an instrumental variable for D_i. Describe how you would implement the TSLS estimator in this context. The table below presents OLS and TSLS estimates (and their standard errors) for δ_D. (Y_i here is the Course Grade ranging from $A = 4$ to $F = 0$ and standard errors are in parenthesis.)

	Y_i = Course Grade ($A = 4, \ldots, F = 0$)	
	OLS	TSLS
Took course online (D_i)	-0.381	-0.440
	(0.012)	(0.049)
Sample mean for dependent variable	2.821	
Observations	2,323,023	

How do you interpret the results? Describe how you would test whether D_i is endogenous.

(d) The F-statistic for the hypothesis that $\beta_{\text{Avail}\times\text{Dist}_i} = 0$ in the Linear Probability Model represented by equation (5) is approximately 100. Explain why this is important in this context. The authors also write that it is important that "(i) any mechanism through which students' distance from campus affects course grades (i.e., Y_i) is constant across terms with and without an in-person class option; and (ii) any mechanism causing grades (i.e., Y_i) to differ between terms with and without an in-person class option affects students homogeneously with respect to their distance from campus." Why is this important?

(e) Let $\overline{Y}_t$ now denote the average Course Grade in the college for year t. Suppose you are interested in modelling the dynamic evolution of $\overline{Y}_t$ using the following autoregressive model:

$$\overline{Y}_t = \rho_0 + \rho_1 \overline{Y}_{t-1} + \eta_t, \qquad |\rho_1| < 1.$$

If you estimate the model by OLS, is the estimator unbiased and consistent if η_t is serially correlated? Explain.

(f) How would you test whether there is serial correlation in η_t?

B.2 This question is based on "Capital Accumulation and Growth: A New Look at the Empirical Evidence", by Steve Bond, Asli Leblebicioglu and Fabio Schiantarelli (*Journal of Applied Econometrics*, Vol.25, 2010). In this article, the authors are interested in a regression model for the (logarithm of) output-per-capita y_t in a given country and time period t similar to:

$$y_t = \alpha + \rho y_{t-1} + \beta x_t + \gamma t + \epsilon_t, \tag{4}$$

where x_t is (the logarithm of) investment-per-output. Assume for items (a)-(c) that ϵ_t is *not* serially correlated.

(a) Imagine that investment rates are also affected by current output-per-capita, so that

$$x_t = \psi_0 + \psi_y y_t + \eta_t. \tag{5}$$

Assume η_t is *not* serially correlated. Equations (6) and (7) then form a simultaneous equation system. Obtain the reduced form equations for y_t and x_t. Is equation (6) identified?

(b) The second equation has two endogenous variables (y_t and x_t) and two (contemporaneously) exogenous variables (y_{t-1} and t). It can be identified using y_{t-1} and t as instrumental variables for y_t. Explain how you would implement the TSLS estimator in this context. How would you assess relevance?

(c) A classmate suggests that you can perform an over-identification test since there are two instrumental variables and only one endogenous regressor. Describe how to implement such a test (under homoskedasticity). Would you be able to perform this test if there were only one instrumental variable?

(d) Assume that $x_t = \psi_0 + \eta_t$ ($\psi_y = 0$), $\rho = 0$ and $\gamma = 0$. Furthermore, suppose you only observe y_t if $y_t \geq 10$ and whether this occurs or not. Under the assumption that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, write down the (log-)likelihood for the model.
HINT: Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the CDF and PDF of a standard normal random variable, respectively. The CDF for U_i is $F(u) = \Phi(u/\sigma)$ and its PDF is $f(u) = \phi(u/\sigma)/\sigma$.

(e) Assume that $x_t = \psi_0 + \eta_t$ (i.e., $\psi_y = 0$), $\rho = 0$ and $\gamma = 0$. How would you test whether ϵ_t is serially correlated? Explain.

(f) Assume that $x_t = \psi_0 + \eta_t$ (i.e., $\psi_y = 0$), $\beta = 0$ and $\gamma = 0$. Take a country for which y_t is stationary (i.e., $|\rho| < 1$), so that

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t.$$

If $\epsilon_t = \lambda \epsilon_{t-1} + \nu_t$, is the OLS estimator for α and ρ consistent?

	5 % Critical values for the F_{ν_1,ν_2} distribution														
$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	161	199.	216.	225.	230.	234.	237.	239.	242.	244.	246.	248.	250.	252.	254.
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
10	4.96	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.23	2.16	2.07	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.20	2.12	2.04	1.97	1.84
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	2.01	1.93	1.84	1.76	1.62
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.84	1.75	1.65	1.56	1.39
80	3.97	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.79	1.70	1.60	1.51	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.77	1.68	1.57	1.48	1.28
120	3.91	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.91	1.83	1.75	1.66	1.55	1.46	1.25
∞	3.85	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.67	1.57	1.46	1.35	1.00

NORMAL CUMULATIVE DISTRIBUTION FUNCTION ($Prob(z < z_a)$ where $z \sim N(0,1)$)

z_a	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995