EC3090 Mock Exam

Michael Curran

Attempt ONE question from the two questions in this section.

Question 1 (50 Marks) – Identification & Simultaneous Equations Models.

Part (a): (25 Marks)

Suppose we have data on how well a class of Leaving Cert students perform in their Higher Level Leaving Cert Maths exam. Consider the case where for some reason, we are not able to access the scores on a few students – perhaps we have data from a questionnaire and 6 out of 120 students decline to answer. Let y denote result in percentage terms where students need at least 40 to pass and z denote whether we observe the result or not. The distribution of marks for students we have data on is given in table 1. Take scores in fives for each interval students score in (so $5, 15, 25, \ldots$). You can take 5 as the minimum and 95 as the maximum score obtainable.

1. What bounds does the probability that a student taken at random from the class passes lie between? (6 Marks)
2. What if we assume that the data was missing at random? Is the bound tighter? Why or why not? Is it refutable? (4 Marks)

Score	Frequency
0-10	3
11-20	17
21-30	2
31-40	10
41-50	18
51-60	10
61-70	15
71-80	21
81-90	15
91-100	3

Table 1: Distribution of marks.

3. What bounds can we place on the average mark in the entire population? (7 Marks)

4. Compare your answer from part 3 with an imputation rule that students whose results are missing are given the average mark of the observed results. (3 Marks)

5. Finally, assume that we no longer have any missing data – we have the distribution of marks for all 120 students. Let $mothmath$ be a dummy variable equal to one if a student has a mother who studied maths in university and zero otherwise. Suppose upon discovering that $P(y \geq 75|x = 1) = 0.5$ and $P(y \geq 75|x = 0) = 0.1$ an examiner states the following:

   > Data indicate that having a mother who studied maths in college increases the probability a student will score highly in Higher Level Leaving Cert maths. The effect of having a mother who studied maths in college is to increase the probability of scoring at least 75% from 0.1 to 0.5.

Does this statement accurately describe the empirical findings? Explain. (5 Marks)

Part (b): (25 Marks)

Consider the following macroeconomic model in structural form:

$$\begin{aligned}\text{Consumption: } C_t &= a_0 + a_1 Y_t - a_2 T_t + u_t \\ \text{Investment: } I_t &= b_0 + b_1 Y_{t-1} + v_t \\ \text{Tax: } T_t &= c_0 + c_1 Y_t + w_t \\ \text{GNP identity: } Y_t &= C_t + I_t + G_t\end{aligned}$$

i) What are the endogenous variables? (3 Marks)

ii) Is the system complete? (2 Marks)

iii) Label the equations. (2 Marks)

iv) What are the predetermined variables? (3 Marks)

v) Check the identifiability of the investment equation. (15 Marks)

Question 2 (50 Marks) – Limited Dependent Variables & Instrument Variables.

Part (a): (25 Marks)

Attempt one of the following two questions.

1)

Attempt either (i) OR (ii).

(i)

Suppose that in a study comparing election outcomes between left wing candidates and right wing candidates, you wish to indicate the political persuasion of each candidate. Is a name *wing* a wise choice for a binary variable in this case? What would be a better name? (2 Marks)

Consider the following regression where $wage$ is the hourly wage in Euros, $female$ is a binary variable equal to 1 if female and 0 if male and $educ$ is the years of education:

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + u_1 \tag{1}$$

Interpret δ_0 and identify the benchmark group. (2 Marks) Have we avoided the dummy variable trap? Explain. (3 Marks) Estimating (1) by OLS, we get $\hat{\delta}_0 = -0.01$. Interpret this coefficient. (2 Marks) Now suppose we want to include single men, single women, married men and married women. Consider allowing salary differentials across these four categories with married females as the base group. Write out the corresponding regression equation. (4 Marks) Now suppose we wanted instead to focus on wages for individuals who have attended law schools. Suppose there are 100 law schools ranked from 1 (best) to 100 (worst). How might we – in a limited amount of time – construct a model and test whether the ranking of law school has a constant partial effect? (6 Marks) Suppose estimating

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u$$

yielded $\hat{\beta}_0 > 0$, $\hat{\delta}_0 < 0$, $\hat{\beta}_1 > 0$, $\hat{\delta}_1 > 0$ and $\hat{\beta}_0 + \hat{\delta}_0 > 0$. Draw a (very) rough sketch of the return to education. (5 Marks)

(ii)

		Predicted $y = 0$	Predicted y=1	Total
Actual	$y = 0$	471	16	487
Actual	$y = 1$	183	20	203
	Total	654	36	690

Table 2: Migration Study

Why is the linear probability model called the linear probability model and how do we interpret the estimates? (3 Marks) Describe any two limitations of this model. (3 Marks) Describe EITHER the logit OR the probit model and how they ensure response probabilities are strictly between zero and one. (3 Marks) Explain how we interpret coefficients in EITHER the logit OR the probit model. (3 Marks) How might we calculate partial effects (CHOOSE the logit OR the probit model)? (4 Marks) How do we estimate these models and why can we not use OLS? (4 Marks) EITHER describe the LR test OR two alternatives to R^2. (4 Marks) Consider the migration study by Tumali (1986) between European cities via a logit model on how certain circumstances and factors can lead to the migration of families. Table 2 shows some results. Calculate count R^2. (1 Mark)

OR

2)

Attempt either (i) OR (ii) OR (iii).

(i)

Let y be the number of extramarital affairs for a married woman from the US population. The goal is to explain this variable (y) in terms of other characteristics of the woman, especially whether she works outside of the home, her husband and her family. Is this a good candidate for a Tobit model? Explain. (2 Marks) Now consider

$$\begin{aligned} y^* &= \beta_0 + \mathbf{x}\boldsymbol{\beta} + u \quad u|\mathbf{x} \sim N(0, \sigma^2) \\ y &= \max(0, y^*) \end{aligned}$$

Assume that the latent variable y^* satisfies Classical Linear Regression Model assumptions, in particular it is Normally distributed with homoscedastic variance and a linear conditional mean. How do we estimate such a Tobit model? (5 Marks) How do we interpret β_j? (3 Marks) What if we are interested in $E(y|\mathrm{x})$? (4 Marks) Discuss the partial effects of x_j on $E(y|y > 0, \mathrm{x})$ and $E(y|\mathrm{x})$. (5 Marks) How do we compare OLS and Tobit estimates? (3 Mark) How can we informally evaluate whether Tobit is appropriate? (3 Marks)

(ii)

Suggest an example of a count variable. (3 Marks) Why can we not use logarithms for count variables? (2 Marks) Instead we will model the expected value as an exponential function:

$$E(y|x_1, x_2, \ldots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

How can this model be made linear? (3 Marks) Interpret β_j. (4 Marks) How do we estimate Poisson models? (5 Marks) Why might we consider Poisson models to be restrictive? (4 Marks) Why might we still use Poisson models and if we do not assume the Poisson distribution, how can we estimate these models? (4 Marks)

(iii)

Under what circumstances might one entertain the possibility of employing a censored model and how do we estimate such models and interpret β_j? (6 Marks) Under what circumstances might one entertain the possibility of employing a truncated model? (2 Marks) Distinguish a truncated model from a censored model. (2 Marks) How might we estimate a truncated model and if we used OLS, how might you expect the estimates to be affected by truncation? (6 Marks) Distinguish between exogenous and endogenous sample selection. (2 Marks) Describe the Heckit method. (4 Marks) How might we test for selection bias? (3 Marks)

Part (b): (25 Marks)

Consider the simple wage regression

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

where the dependent variable is the log hourly wage for working adults and $educ$ is years of education.

- Why might education be endogenous in this model? (3 Marks)
- Suggest a possibly omitted variable. (1 Mark)
- How would our estimates be affected by an omitted variable? (2 Marks)

Suppose that you are interested in estimating the following expanded wage regression model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1 \tag{2}$$

where $exper$ is years of experience and we assume that $exper$ and $exper^2$

are uncorrelated with u_1.[1] Let us suppose that you have two excluded exogenous regressors (instruments) $z_3 = motheduc$ and $z_4 = fatheduc$ where $motheduc$ is the years of mother's education and $fatheduc$ is years of father's education.

- Why might z_3 and z_4 be good instrumental variables? (3 Marks)
- Suggest an alternative instrument that might be better and explain why. (2 Marks)
- Why might we want to test for endogeneity? (1 Mark)
- Describe the Hausman test for endogeneity of $y_2 = educ$. (6 Marks)

We wish to test if at least one of $motheduc$ and $fatheduc$ are correlated with $educ$ in a regression of $educ$ on $exper$, $exper^2$, $motheduc$ and $fatheduc$. We get $F = 55.40$ and a p-value of approximately $.0000$. Interpret this finding. (2 Marks)
Estimating equation (2) by 2SLS in Stata, we obtain the following results

$$\log\widehat{(wage)} = .048 + .061 educ + .044 exper - .0009 exper^2$$
$$(.400) \; (.031) \quad (.013) \quad (.0004)$$
$$n = 428, R^2 = .136$$

- Interpret the coefficient on $educ$. (2 Marks)
- Why might the 2SLS estimate be barely significant at the 5% level against a two-sided alternative? (1 Mark)
- Is $R^2 = .136$ a cause for concern? Explain. (2 Marks)

[1] We include the quadratic term $exper^2$ because we may think that experience has a diminishing effect on wage say (e.g. we might think $\beta_2 > 0$ but $\beta_3 < 0$).