EC3090 Mock Exam Solutions

Michael Curran

Attempt ONE question from the two questions in this section.

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

Part (a): (25 Marks)

Given that we are taking marks in 'fives', let us write the revised distribution in table 1. There are $N = 120$ students in the class, of whom we have no data on six, so $P(z = 0) = \frac{6}{120} = 0.05$ is the fraction of missing data; remember $P_N(z = 1) = \frac{1}{N}\sum_{i=1}^{N} 1[z_i = 1]$.

Score	Frequency
5	3
15	17
25	2
35	10
45	18
55	10
65	15
75	21
85	15
95	3

Table 1: Revised distribution of marks.

1. To pass, students must get at least 40. Letting B denote the set of all such marks from the revised distribution that corresponding to the students passing, to pass y must be in B:

$$y \in \{45, 55, 65, 75, 85, 95\} \equiv B$$

(1 Mark)
We want $P(y \in B)$ and can express this using the law of total probability (LTP) as

$$P(y \in B) \overset{\text{LTP}}{=} P(y \in B|z = 1)P(z = 1) + P(y \in B|z = 0)P(z = 0) \quad (1)$$

(2 Marks)
We know

$$P(z = 0) = 0.05 \Longrightarrow P(z = 1) = 1 - P(z = 0) = 1 - 0.05 = 0.95$$

(0.25 Marks for each $P(z = 0) = 0.05$ and $P(z = 1) = 0.95$)
and while $P(y \in B|z = 0)$ is the only unknown quantity in (1), because it is a probability, $P(y \in B|z = 0) \in [0, 1]$. (0.5 Marks) We need to calculate $P(y \in B|z = 1)$.

$$\begin{aligned} P_N(y \in B|z = 1) &= \frac{\sum_{i=1}^{N} 1[y_i \in B, z_i = 1]}{\sum_{i=1}^{N} 1[z_i = 1]} \\ &= \frac{18 + 10 + 15 + 21 + 15 + 3}{120 - 6} \\ &= \frac{82}{114} \\ \therefore P(y \in B) &= \frac{82}{114} \times 0.95 + [0, 1] \times 0.05 \\ &\in \left[\frac{41}{60}, \frac{11}{15}\right] \\ &\equiv H[P(y \in B)] \end{aligned}$$

which is our identification region for the probability that a student passes. (1 Mark for $P(y \in B|z=1) = \frac{82}{114}$, 1 Mark for $H[P(y \in B)] = \left[\frac{41}{60}, \frac{11}{15}\right]$)

2. The assumption of missingness at random (MAR) is

$$P(y|z=1) = P(y|z=0)$$

Observation: Observe that under MAR

$$P(y|z=1) = P(y|z=0) = P(y)$$

To see this, use the law of total probability to expand $P(y)$:

$$\begin{aligned} P(y) &\overset{\text{LTP}}{=} P(y|z=1)P(z=1) + P(y|z=0)P(z=0) \\ &\overset{\text{MAR}}{=} P(y|z=1)[\underbrace{P(z=1) + P(z=0)}_{1}] \\ &= P(y|z=1) \\ &\overset{\text{MAR}}{=} P(y|z=0) \end{aligned}$$

Note that here MAR implies that

$$P(y \in B|z=1) = P(y \in B|z=0)$$

(0.5 Marks)

$$\begin{aligned} \therefore P(y \in B) &\overset{\text{LTP}}{=} P(y \in B|z=1)P(z=1) + P(y \in B|z=0)P(z=0) \\ &\overset{\text{MAR}}{=} P(y \in B|z=1) \\ &\overset{1}{=} \frac{82}{114} \end{aligned}$$

(0.5 Marks for first line, 0.5 marks for second line, 0.5 marks for third

line). Therefore, MAR *point identifies* $P(y \in B)$. (0.5 Marks)
Certainly, the bound is tighter:(0.25 Marks)

$$H_1[P(y \in B)] = \left[\frac{41}{60}, \frac{11}{15}\right] \ni \frac{82}{114} = H_1[P(y \in b)]$$

(0.25 Marks)
It is tighter because we impose a strong, *nonrefutable* assumption on the distribution of unknown, missing data (0.25 Marks); and it is *nonrefutable* (0.25 Marks) because the assumption directly restricts the distribution $P(y \in B|z = 0)$ of missing data (0.5 Marks).

3. We want $E(y)$, so using the law of iterated expectations to expand $E(y)$, we get that

$$E(y) \overset{\text{LIE}}{=} E(y|z = 1)P(z = 1) + E(y|z = 0)P(z = 0)$$

(2 Marks)
We know

$$P(z = 1) = 0.95 \quad P(z = 0) = 0.05$$

(0.5 marks for each)
and while $E(y|z = 0)$ is unknown, marks must lie within $[0, 100]$. Actually with the assumption of 'fives', we know more:

$$5 \leq E(y|z = 0) \leq 95 \quad (1 Mark)$$

Going even further, we can write this out fully:

$$E(y|z = 0) \in \{5, 15, 25, 35, 45, 55, 65, 75, 85, 95\}$$

We need to calculate $E(y|z = 1)$ and can work this out from the revised distribution in table 1. Summing over observed i where I

denotes the number of observations:

$$\begin{aligned}
E(y|z=1) &= \frac{1}{I}\sum_i \text{score}_i \times \text{frequency}_i \\
&= \frac{(5)(3)+(15)(17)+(25)(2)+(35)(10)+(45)(18)}{114} \\
&\quad + \frac{(55)(10)+(65)(15)+(75)(21)+(85)(15)+(95)(3)}{114} \\
&= \frac{6140}{114} \\
\therefore E(y) &= \frac{6140}{114}(0.95)+[5,95](0.05) \\
&\in \left[\frac{617}{12}, \frac{671}{12}\right] \\
&[51.41\dot{6}, 55.91\dot{6}] \\
&\equiv H[E(y)]
\end{aligned}$$

which is the identification region for the average mark in the class. (1 Mark for $E(y|z=1)$, 2 Marks for $H[E(y)]$)

4. From part 3

$$\begin{aligned}
E(y|z=1) &\overset{3}{=} \frac{6140}{114} \\
&= E(y|z=0)
\end{aligned}$$

where the second equality follows by the imputation rule in this part of the question uses. (0.5 Marks) So now

$$E(y) = \frac{6140}{114}$$

(1 Mark)
and we get *point* identification rather than *partial* identification. (0.5 Marks) The imputation rule we used utilises a weaker assumption

than MAR (0.5 Marks); it restricts only means, rather than the entire probability distribution of missing data (0.5 Marks).

5. No, this statement does not accurately describe the empirical finding.

   We can only say that students having mothers who took maths in college on average scored higher than those who did not have such a mother. We cannot say that the very fact that having a mother who studyied mathematics at college increased the probability of a student scoring highly. (1 Mark)

   Asking what would happen to this $E(Y|X)$ when we vary X is akin to a hypothetical change in X, where we have no data and so the researcher has **confused correlation with causation** and has used a **counterfactual** (expressing what has not happened but what might or would happen if circumstances, i.e. data, were different). The researcher is in effect extrapolating using the assumption of external validity, which is undermined by the fact that we are only looking at *students in a particular class* and we have no data on the rest of the population of students at large. (3 Marks: 2 for pointing out, 1 for explaining)

   However, if the students were **randomly assigned** having mothers' with such backgrounds, then the researcher would be correct in saying that having mothers who studied maths in college increases the probability that a student will do better on average than a student who does not have such a mother. But since we are dealing with what actually happened (descriptive) we cannot say that having such a mother increases the probability that a student scores highly. (1 Mark)

Part (b): (25 Marks)

1. Endogenous variables $M = 4$: C, I, T, Y. 1 mark for each.
2. Yes, the system is complete (1 Mark) since the number of endogenous variables is equal to the number of equations = 4 (1 Mark).
3. Consumption equation, investment equation and tax equation are all behavioural equations. GNP identify is an 'equilibrium condition' or an 'identity'. Half a mark for each correctly labelled equation/identity.
4. Predetermined variables $K = 3$: exogenous: 1, G and lagged endogenous: Y_{t-1}. 1 mark for each.
5. The structural parameters (arranged) are

$$\begin{array}{ccccccc}
C & I & T & Y & 1 & G & Y_{t-1} \\
1 & 0 & a_2 & -a_1 & -a_0 & 0 & 0 \\
0 & 1 & 0 & 0 & -b_0 & 0 & -b_1 \\
0 & 0 & 1 & -c_1 & -c_0 & 0 & 0 \\
-1 & -1 & 0 & 1 & 0 & -1 & 0
\end{array}$$

(3 Marks) Focusing on the investment function, the order condition is checked by:

$$K - k = 1$$
$$m - 1 = 0$$
$$\therefore \; K - k > m - 1$$

Alternatively

$$M + K - (m + k) = 4$$
$$M - 1 = 3$$
$$\therefore \; M + K - (m + k) > M - 1$$

1 Mark for either condition, 1 Mark for accurate numbers and 1 Mark for accurate signs.

In both cases, the order condition is satistfied as a strict inequality, so the investment function *may* be *over* identified. (0.5 marks for ‘over’ and 0.5 marks if had only said identified; 0 marks if say under/just/exact identified, etc.) We say *may* be since the order condition is not the sufficient condition – we will know with certainty once we have checked the rank condition, which is both necessary and sufficient as a check for identifiability of an equation in a simultaneous equation model. (1 Mark) Checking the rank condition:

$$\Lambda_I = \begin{bmatrix} 1 & a_2 & -a_1 & 0 \\ 0 & 1 & -c_1 & 0 \\ -1 & 0 & 1 & -1 \end{bmatrix}$$

$$\rho(\Lambda_I) = 3 = M - 1 = 3$$

(1.5 Marks for Λ_I, 0.5 Marks for $\rho(\Lambda_I) = M - 1$ condition, 0.25 Marks for $M - 1 = 3$, 2.25 Marks for $\rho(\Lambda_I) = 3$). Therefore, the investment function is **over** identified, given the order condition result. (1 Mark)

Extra on how we got $\rho(\Lambda_I) = 3 = M - 1 = 3$: rank cannot exceed 3, but could be 2 (look only at rows). The rank must be 3 if we can sensibly estimate the investment function. We can go about this in either of the following two ways; each way implies that the columns and rows are linearly indepenent so the rank is 3. 2.5 Marks for showing either or both of the following:

(a) Show that the determinant is non-zero:

$$det(\Lambda_I) = 1 + a_2 c_1 - a_1 \neq 0$$

(b) Consider α_1, α_2 and α_3 such that at least one is non-zero:

$$\alpha_1(1 \;\; a_2 \;\; -a_1 \;\; 0) + \alpha_2(0 \;\; 1 \;\; -c_1 \;\; 0) + \alpha_3(-1 \;\; 0 \;\; 1 \;\; -1) = 0$$

which is equal to zero if and only if

$$\alpha_1 - \alpha_3 = 0 \tag{2}$$

$$\alpha_1 a_1 + \alpha_2 = 0 \tag{3}$$

$$-\alpha_1 a_1 - \alpha_2 c_1 + \alpha_3 = 0 \tag{4}$$

$$-\alpha_3 = 0 \tag{5}$$

Equations (5) & (2) imply that

$$\alpha_1 = \alpha_3 = 0$$

Plugging $\alpha_1 = 0$ into equation (3) implies $\alpha_2 = 0$. So

$$\alpha_1 = \alpha_2 = \alpha_3 = 0$$

but this violates our assumption that at least one α_1, α_2 and α_3 is non-zero. This proves that any linear combination of the rows can only sum to zero if all coefficients α_1, α_2 and α_3 are identically zero – this is the definition of linear independence of rows of a matrix.

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)

Wing is probably not a wise choice since it is not clear if $wing = 1$ means left wing or right wing. (1 Mark) A better name would be something like $left$ or $leftwing$ where $left = 1$ if candidate is left wing politically. (1 Mark)

δ_0: differential in average growth of hourly wage between men and women (differential effect of being a woman) with same level of education. (1 Mark) The benchmark group are men. (1 Mark)

Yes, we have avoided the dummy variable trap since there are two categories in our dummy variable (male and female) and we are using 1 dummy variable (female) and an intercept. (1.5 Marks) In general, we avoid the dummy variable trap when there are g groups or categories by including at most $g - 1$ dummy variables plus an intercept, or g dummy variables and no intercept. (1.5 Marks)

The coefficient $\hat{\delta}_0 = -0.01$ in this case means that the differential effect of being a female is associated with a $100 \times \hat{\delta}_0 = -1\%$ change in wages, i.e. women earn on average one percent less than men in hourly wages for a given level of education. (1.5 Marks) To compute the exact percentage difference in predicted wages for a woman relative to a man, we calculate $100[\exp \hat{\delta}_0 - 1] \approx -0.995$ to three decimal places. (0.5 Mark)

To allow for salary differentials across categories, with married females as the base group, we would include all other categories as independent dummy variables $singmale$, $singfem$ and $marrmale$ where they correspond to single males, single females and married males:

$$\log(wage) = \beta_0 + \delta_0 singmale + \delta_1 singfem + \delta_2 marrmale + \beta_1 educ + u_2$$

(4 Marks for equation; 2 Marks without equation but with category to exclude / categories to include)

We could partition the rankings of law schools for instance into the top 10, 11-25, 26-40, 41-60 and 61-100. (2 Marks; 1 Mark extra for explaining constant partial effects; 3 Mark for reasonable explanation of test) Sample explanation of constant partial effects: Constant partial effects mean that the effect of going from a school ranked say 96 to one ranked 95 has the same effect on wages as going from a school ranked say 5 to 4. Perhaps there is a bigger effect on wages of moving from a school ranked 5 to 4 than the effect of moving from a school ranked 96 to one ranked 95. Partitioning, perhaps there is a bigger effect on wages in moving from schools ranked between 11-25 into those ranked in the top 10 than moving from schools ranked 41-60 into those ranked 26-40. Sample explanation of testing for constant partial effects: We can test for constant partial effects by having only one variable for ranking taking on values from one to 100 $ranking$:

$$\log(wage) = \beta_0 + \beta_1 ranking + \text{otherfactors} \tag{6}$$

and allowing

$$\begin{aligned}\log(wage) = \beta_0 &+ \delta_1 topten + \delta_2 secondtier + \delta_3 thirdtier + \delta_4 fourthtier \\ &+ \delta_5 bottom40 + \text{otherfactors}\end{aligned} \tag{7}$$

where $topten$ is dummy for schools ranked in top 10, $secondtier$ is dummy for schools ranked 11-25, $thirdtier$ is dummy for schools ranked 26-40,

$fourthtier$ is dummy for schools ranked 41-60 and $bottom40$ is dummy for schools ranked 61-100. Then construct an F-statistic for testing constant patial effects restriction

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n-k-1)} = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/df_{ur}}$$

where R_{ur}^2 comes from (7) and R_r^2 comes from (6), $q = 4$ here (four restrictions: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5$), df_{ur} is the degrees of freedom of the unrestricted model, which is $n - k - 1$ where n is the sample size and k are is number of coefficients other than the intercept. If we get a low p-value for the F test that $H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4$, then we reject the null hypothesis of constant partial effects. We can also check if breaking rank into different groups improves things via comparing the adjusted R^2 with that from including rank as a single variable. If R^2 is higher when we break rank into different groups, then this suggests that additional flexibility is warranted. Note that OLS depends on the random sample assumption, but each school's rank depends on the rank of other schools in the sample so data here cannot represent independent draws from the population of all law schools. This will not cause any serious problems as long as the error term is uncorrelated with explanatory variables.

The graph for $\beta_0 > 0$, $\delta_0 < 0$, $\beta_1 > 0$, $\delta_1 > 0$ and $\beta_0 + \delta_0 > 0$ is given in figure 1.[1] (2 Marks for relative intercepts, 2 Marks for relative slopes, 1 Mark extra for overall correctly specified diagram)

(ii)

The linear probability model (LPM) models dummy dependent variables

[1] Correction: wage label on vertical axis should be 'log(wage)'.

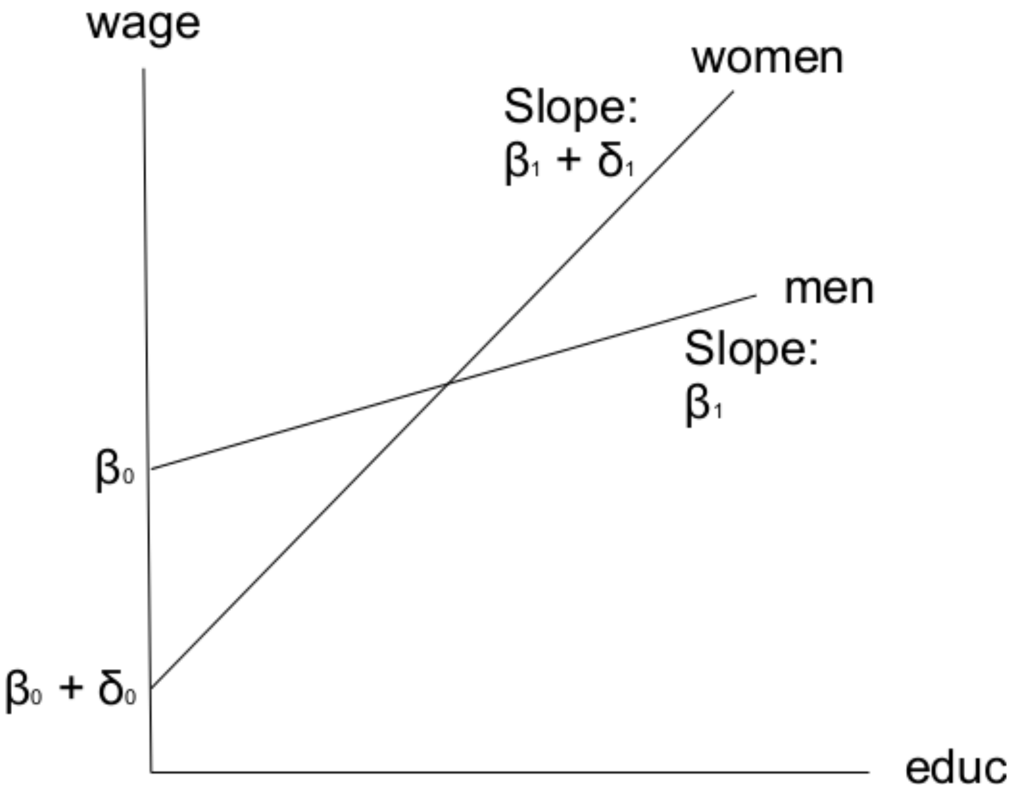

Figure 1: Differential intercept and slope for return to education between men and women.

$y = 1$ or $y = 0$:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

For example, $y = 1$ if you have a car and $y = 0$ if not; x could be income. Note that $P(y = 1|\mathbf{x}) = E(y|\mathbf{x})$, i.e. the probability of success (e.g. owning a car, i.e. $y = 1$) is the same as the expected value of y since

$$\begin{aligned} E(Y_i|X_i) &= \sum_{j=1}^{2} Y_{ij} P(Y_{ij}|X_i) \\ &= 1P(1|X_i) + 0P(0|X_i) \\ &= P(1|X_i) \\ &= E(Y_i|X_i) \equiv P_i \\ &= \alpha + \beta X_i \end{aligned}$$

The second last equality sign shows why we call this the linear probability model. So,

$$P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{8}$$

which says that the probability of success $p(\mathbf{x}) = P(y = 1|\mathbf{x})$ is a linear function of the x_j. (8) is a binary response model and $P(y = 1|\mathbf{x})$ is the response probability. We call the model a linear probability model since the response probability is linear in the parameters β_j. (This one sentence would typically give you full / close to full marks (1.5 Marks)) The parameters β are not the change in y for a given change in x since how would it make sense to say that x changes you from owning no car to owning 0.12 of a car. Intead, we look at probabilities. In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\frac{\Delta P(y = 1|\mathbf{x})}{\Delta x_j} = \beta_j$$

(This one sentence, possibly augmented with the equation would typically give you full / close to full marks (1.5 Marks)) With multiple regressions:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

where $\hat{y}$ is the predicted probability of success so $\hat{\beta}_0$ is the predicted probability of success when each x_j is set to zero (may or may not be interesting) and the slope coefficient $\hat{\beta}_1$ measures the predicted change in the probability of success when x_1 increases by one unit.

Limitations (any two, 1.5 marks each)

1. Predicts probabilities that could be less than 0 or greater than 1. Use graph to explain.

2. Constant partial effects. Need to explain.

3. Heteroscedasticity unless probability does not depend on any of the independent variables. No bias but t and F statistics rely on homogeneity even when sample size is large. Corrections: heteroscedasticity-robust standard errors, t, F and Lagrange-Multiplier (LM) statistics and tests for heteroscedasticity plus WLS, GLS and FGLS. To see heteroscedasticity:

$$\begin{aligned} V(u) &= E[u - E(u)]^2 = E(u^2) \\ &= \sum_{j=1}^{2} u_j P(u_j) \\ &= (1 - \alpha - \beta X)^2(\alpha + \beta X) + (-\alpha - \beta X)^2(1 - \alpha - \beta X) \\ &= (1 - \alpha - \beta X)^2(\alpha + \beta X) + (\alpha + \beta X)^2(1 - \alpha - \beta) \\ &= (1 - \alpha - \beta X)(\alpha + \beta X) \\ &= P_i(1 - P_i) \end{aligned}$$

To see how to use WLS: run OLS on $Y_i = \alpha + \beta X_i + u_i$ to get $\hat{Y}_i = \hat{P}_i$

and set weights to be $w_i = \left[\hat{P}_i(1-\hat{P}_i)\right]^{\frac{1}{2}}$ and transform data as

$$Y_i^* = \frac{Y_i}{w_i} \quad X_i^* = \frac{X_i}{w_i} \quad u_i^* = \frac{u_i}{w_i}$$

Do not create a constant: do not need intercept – otherwise you are producing a new variable in place of the intercept. So

$$V(u_i^*) = V\left(\frac{u_i}{w_i}\right) = \frac{1}{w_i^2} \quad V(u_i) = \frac{w_i^2}{w_i^2} = 1$$

Run OLS on

$$Y_i^* = \alpha\frac{1}{w_i} + \beta X_i^* + u_i^*$$

to get $\hat{\alpha}$ and $\hat{\beta}$, which will be unbiased and asymptotically efficient. It turns out that in many applications, OLS statistics are not too far off and it is acceptable in applied work to present a standard OLS analysis of a LPM.

4. Binomial errors:

$$Y_i = 1 \Longrightarrow u_i = 1 - \alpha - \beta X_i$$

and

$$Y_i = 0 \Longrightarrow u_i = -\alpha - \beta X_i$$

Thus, u_i is binomial with parameter P_i and therefore errors are non-Normal, so Classical Linear Normal Regression model assumption is violated, which complicates confidence intervals, F tests, t tests, etc.

LPM:

$$P(y=1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Consider a class of binary response models of the form

$$P(y = 1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$$

where $0 < G(z) < 1$ for all $z \in \mathbb{R}$. G ensures that the estimated response probabilities are strictly between zero and one.

$$\mathbf{x}\boldsymbol{\beta} = \beta_1 x_1 + \cdots + \beta_k x_k$$

In the *logit* model, G is the logistic function

$$G(z) = \frac{\exp(z)}{1 + \exp(z)} = \Lambda(z) \in [0, 1] \; \forall z \in \mathbb{R}$$

In the *probit* model, G is the standard normal cumulative distribution function

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v) dv$$

where $\phi(z)$ is the standard normal density

$$\phi(z) = \frac{1}{(2\pi)^{-\frac{1}{2}}} \exp\left(-\frac{z^2}{2}\right)$$

(Graphs of logistic function/normal cdf are unnecessary but will be taken into account if no equations are specified; likewise discussion of odds ratio / log of odds ratio for logit is unnecessary but will be taken into account similarly.) (1.5 Marks for description of either logit/probit, 1.5 Marks for explanation of G function bounding response probabilities to the unit interval $[0, 1]$)

Interpretations are complicated by the nonlinear nature of $G()$. (1.5 marks) Need to scale β_j by adjustment factor $\partial G(\beta_0 + \mathbf{x}\boldsymbol{\beta})/\partial x_j$ to measure $P(y = 1|\mathbf{x})$. (1.5 Marks)

Assume that x_j is roughly continuous. Partial effect always has the same sign as β_j since $G()$ strictly increasing cdf in logit and probit (i.e. $g(z) > 0\ \forall z$) so partial effect of x_j on $p(\mathrm{x})$ depends on x through the positive quantity $g(\beta_0 + \mathrm{x}\boldsymbol{\beta})$:

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j$$

which shows that relative effects of any two continuous explanatory variables do not depend on x: ratio of partial effects for x_j and x_h is $\frac{\beta_j}{\beta_h}$:

$$\frac{\partial p(\mathbf{x})/\partial x_j}{\partial p(\mathbf{x})/\partial x_h} = \frac{g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_j}{g(\beta_0 + \mathbf{x}\boldsymbol{\beta})\beta_h} = \frac{\beta_j}{\beta_h}$$

In the typical case where g is symmetric about zero with unique mode at zero, the largest effect occurs when $\beta_0 + \mathrm{x}\boldsymbol{\beta} = 0$. For example, with the probit:

$$\begin{aligned} g(z) &= \phi(z) \\ g(0) &= \phi(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4 \end{aligned}$$

and for the logit:

$$\begin{aligned} g(z) &= \frac{\exp(z)}{1 + \exp(z)} \\ g(0) &= \frac{\exp(0)}{1 + \exp(0)} = \frac{1}{1+1} = 0.5 \end{aligned}$$

When x_1 is a binary explanatory variable, then the partial effect from changing x_1 from zero to one holding all other variables fixed is

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k) \qquad (9)$$

which depends on all the values of the other x_j. The sign of β_1 is sufficient to see if the program had a postive or negative effect, but to find the magnitude, we must estimate the quantity in (9). We can use (9) for other kinds of discrete variables (e.g. number of children, x_k), so the effect on the probability of x_k going from c_k to $c_k + 1$ is:

$$G[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k(c_k + 1)] - G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k c_k)$$

This measures the effect of discrete variables. Note that we must use estimates:

$$\Delta P(\widehat{y = 1}|\mathbf{x}) \approx [g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j]\Delta x_j$$

where the $\approx$ sign reflects the fact that x_j are roughly continuous. This means that for small changes $\Delta x_j = 1$ say, the effect of x_j in response probability $P(y = 1|\mathbf{x})$ will be

$$\Delta P(\widehat{y = 1}|\mathbf{x}) \approx g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})\hat{\beta}_j$$

So the cost of using logit/probit is that the partial effects here are harder to summarize since the scale factor $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ depends on $\mathbf{x}$ (all of the explanatory variables). We can plug in values for x_j like means, etc. and see how $g(\hat{\beta}_0 + \mathbf{x}\hat{\boldsymbol{\beta}})$ changes. One method in econometrics packages: replace each explanatory variable with the sample average, i.e. adjustment factor is

$$g(\hat{\beta}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}) = g(\hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k) \tag{10}$$

where $g()$ is $N(0, 1)$ (probit) and $g(z) = \frac{\exp(z)}{[1+\exp(z)]^2}$ (logit). The idea is that when we multiply (10) by $\hat{\beta}_j$, we get the partial effect of x_j for the average person in the sample.
(4 Marks for a reasonable discussion of calculating partial effects for either logit or probit)

Maximum likelihood estimation due to nonlinear nature of $E(y|\mathbf{x})$, which renders OLS and WLS inapplicable. (1 Mark for mentioning MLE, 1 Mark for mentioning nonlinear nature compilcates OLS/WLS; 2 Marks for elaborating) Sample elaboration: Could use NLLS/NWLWLS. With MLE, heteroscedasticity is acounted for:

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y} \quad y = 0, 1 \tag{11}$$

where the intercept is in the vector $\mathbf{x}$. When $y = 1$, we have $G(\mathbf{x}_i\boldsymbol{\beta})$ and when $y = 0$, we have $1-G(\mathbf{x}_i\boldsymbol{\beta})$. The log-likelihood function for observation i is a function of the parameters and the data ($\mathbf{x}_i, y_i$) and is simply the log of (11):

$$\ell_i(\boldsymbol{\beta}) = y_i \log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1 - y_i)\log[1 - G(\mathbf{x}_i\boldsymbol{\beta})]$$

Since $G()$ is strictly between 0 and 1 for logit and probit, $\ell_i(\boldsymbol{\beta})$ is well-defined for all values of $\boldsymbol{\beta}$

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \tag{12}$$

MLE of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}$ maximises (12). If $G()$ is the standard logit/normal cdf, then $\hat{\boldsymbol{\beta}}$ is the logit/probit estimator.

An alternative attempt at this question might be the following. Suppose we have data $Y = 1, 1, 0$, $X = X_1, X_2, X_3$ and $P = P_1, P_2, P_3$. If OLS/WLS are inappropriate, we can use MLE, which maximises the probability of the observed sample of data. The problem is to choose α, β to maximise $L(\alpha, \beta|data)$ where the likelihood equation for the logit say is

$$\begin{aligned} L &= \Pi_i P_i \Pi_j (1 - P_j) \\ &= \Pi_i \frac{e^{\alpha+\beta X_i}}{1 + e^{\alpha+\beta X_i}} \Pi_j \frac{e^{\alpha+\beta X_j}}{1 + e^{\alpha+\beta X_j}} \end{aligned}$$

We can differentiate this with respect to z and put this first order conditions equal to zero and solve for z. This would be done numerically, not analytically. Maximum likelihood calculations are difficult, but done routinely in econometrics packages.

The LR test is the same concept as F in a linear model. While F measures the increase in SSR when variables are dropped from a model, the LR test is based on differences in log-likelihood functions for unrestricted and restricted models. The idea is that because MLE maximises log-likelihood, dropping variables generally leads to a *smaller* or at least no larger log-likelihood.

$$LR = 2(\ell_{ur} - \ell_r)$$

Since $\ell_{ur} \geq \ell_r$, LR is nonnegative and usually strictly positive. LR has approximately chi-square distribution under H_0 and if we are testing q exclusion restrictions

$$LR \overset{a}{\sim} \chi^2_q$$

For example, to test the hypothesis that $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, consider $LR = \frac{L_0}{L_1}$, which will be the opposite of R^2. When LR is low, we might have a situation where R^2 for model one is high since the deviations are low and R^2 for model zero is low because the deviations are high. When LR is high, there may be a very small difference. See figure 2. We can estimate the model with variables and without variables. If the restrictions are valid, then $\mathbf{x}$ has no business in the model, which implies that we can impose restreictions that should not affect the likelihood, i.e. $LR \approx 1$ if restrictions are valid; else, $L_0 \approx 0$ so $LR \approx 0$, i.e. we need $\mathbf{x}$. This is the intuitive interpretation of LR. The test statistic is $-2ln(LR) \overset{a}{\sim} \chi^2_k$ where k is the number of exclusion restrictions and ln is the natural logarithm. At a signficance level of $\alpha = 5\%$, we reject H_0 if $-2ln(LR) > \chi^2_{k,0.05}$ or if the p-value is less than 0.05; else we fail to reject the null hypothesis.
(Up to 4 Marks depending on elaboration and clarity)

1. Alternatives to R^2 include McFadden's $R^2 = 1 - \frac{ln(L_1)}{ln(L_0)}$ where $ln(L)$

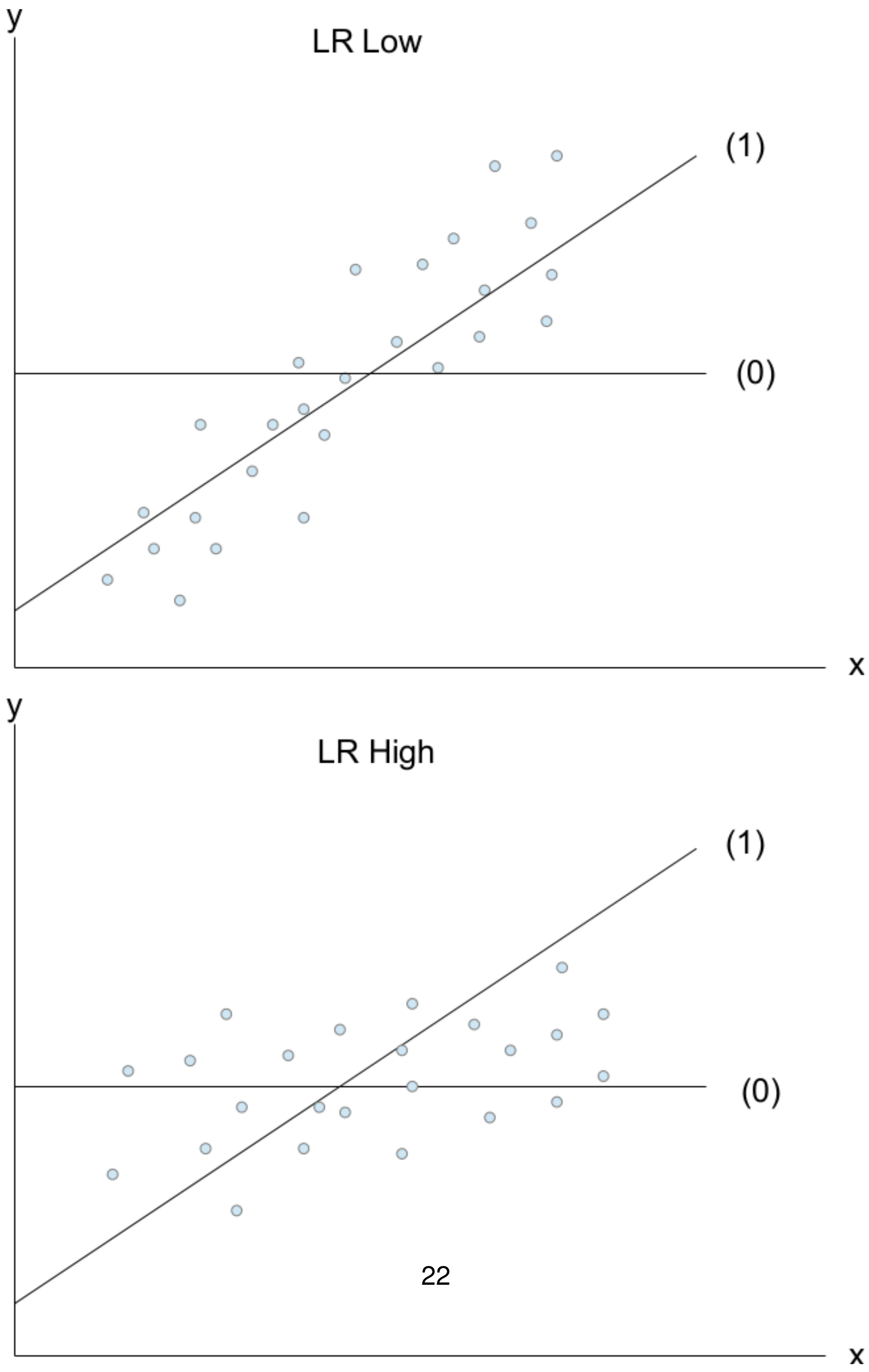

Figure 2: LR test.

is the log likelihood and subscripts 1 and 0 denote alternative and null model, respectively (null is where $\beta = 0$ and alternative is where $\beta \neq 0$); this will lie in $[0, 1]$.

2. Another alternative is Aldrich & Nelson's 'pseudo' R^2:

$$R^2 = \frac{D}{N + D}$$

where $D = -2ln(LR)$ and N is the sample size.

3. Yet another alternative is the count R^2, which is also known as the proportion of accurate predictions. If the model predicts that $P(y_i = 1) = P_i > \frac{1}{2}$ and $y_i = 1$, then we have a 'correct' prediction and if the model predicts that $P(y_i = 1) = P_i < \frac{1}{2}$ and $y_i = 0$, then we have a 'correct' predictrion.

(Any two of last three, 1 Mark each: half a mark for name, half a mark for description)
So, the proportion of correct predictions is given by the ratio

$$\frac{\text{number correct predictions}}{\text{number total observations}} = \frac{491}{690} = 71.2\% = \text{Count } R^2$$

(1 Mark: $\frac{1}{2}$ for correct equation, $\frac{1}{2}$ for correct answer)

OR

2)

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

(i)

Yes, since Tobit models describe corner solutions (lots of zeros in the distribution and then positive values away from zero) and since many wives will have no extramarial affairs and then some will have positive numbers of extramarital affairs, possibly even up to tens or hundreds, the Tobit model is apt. (Need explanation for full marks; 0.5 Marks if say yes with no explanation)

We estimate Tobit models by **maximum likelihood** (2 Marks). (Reasonable explanation of MLE: 3 Marks)
Sample explanation of MLE: Since y^* is Normal, y has continuous distribution over strictly positive values. In particular, the density of y given $\mathbf{x}$ is the same as the density of y^* given $\mathbf{x}$ for positive values. Further:

$$\begin{aligned} P(y=0|\mathbf{x}) &= P(y^*<0|\mathbf{x}) = P(u<-\mathbf{x}\boldsymbol{\beta}|\mathbf{x}) \\ &= P\left(\frac{u}{\sigma}<-\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}|\mathbf{x}\right) = \Phi\left(-\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) = 1-\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

since $\frac{u}{\sigma}$ is $N(0,1)$ and is independent of $\mathbf{x}$; for notationaly simplicity, the intercept is subsumed within $\mathbf{x}$. If ($\mathbf{x}_i, y_i$) is a random draw from the population, then the density of y_i given $\mathbf{x}_i$ is given by

$$(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[\frac{-(y_i-\mathbf{x}_i\boldsymbol{\beta})^2}{(2\sigma)^2}\right] = \frac{1}{\sigma}\phi\left[\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] \quad y_i>0 \tag{13}$$

$$P(y_i=0|\mathbf{x}_i) = 1-\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) \tag{14}$$

From (13) & (14), we can obtain the log-likelihood function for each observation i:

$$\ell_i(\boldsymbol{\beta},\sigma) = 1(y_i=0)\log\left[1-\Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right] + 1(y_i>0)\log\left\{\frac{1}{\sigma}\phi\left[\frac{y_i-\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right]\right\} \tag{15}$$

The log-likelihood for random sample of size n is obtained by summing (15) across all i. Note the dependence on σ (standard deviation of u) and β_j. MLE of $\boldsymbol{\beta}$ and σ require numerical methods (mostly done easily on a packaged routine).

We cannot interpret $\hat{\beta}_j$ from Tobit MLE the same as from linear model OLS. Computers produce Tobit MLE with not much more pain than OLS for linear models, so it is tempting to interpret β_j similarly, though we should not do this. Note that

$$\beta_j = \frac{\partial E(y^*|\mathbf{x})}{\partial x_j}$$

so β_j measures the partial effects of x_j on $E(y^*|\mathbf{x})$.
(2 Marks for equation or in words; 1 Mark only if discuss pain of interpreting without specifying clear interpretation)
We may want to explain y (observed outcome, e.g. hours worked). We can estimate $P(y = 0|\mathbf{x})$ from (14) and this allows us to estimate $P(y > 0|\mathbf{x})$, but we may want to estimate the expected value of y as a function of $\mathbf{x}$. Given $E(y|y > 0, \mathbf{x})$, we can find $E(y|\mathbf{x})$ through the law of iterated expectations (LIE):

$$E(y|\mathbf{x}) \overset{LIE}{=} P(y > 0|\mathbf{x}) \cdot E(y|y > 0, \mathbf{x}) = \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) \cdot E(y|y > 0, \mathbf{x}) \qquad (16)$$

To get $E(y|y > 0, \mathbf{x})$ use the result that $z \sim N(0, 1)$ implies that $E(\mathbf{z}|\mathbf{z} > c) = \frac{\phi(c)}{1-\Phi(c)}$ for any constant c. But:

$$\begin{aligned} E(y|y > 0, \mathbf{x}) &= \mathbf{x}\boldsymbol{\beta} + E(u|u > -\mathbf{x}\boldsymbol{\beta}) \\ &= \mathbf{x}\boldsymbol{\beta} + \sigma E\left[\frac{u}{\sigma}|\frac{u}{\sigma} > -\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right] \\ &= \mathbf{x}\boldsymbol{\beta} + \frac{\sigma\phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)} \end{aligned}$$

since $\phi(-c) = \phi(c)$, $1 - \Phi(-c) = \Phi(c)$ and $\frac{u}{c}$ has standard Normal distribution independent of $\mathbf{x}$. So

$$E(y|y>0,\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \sigma\lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) \tag{17}$$

where $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$ is called the **inverse Mills ratio**. Since $\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) = \phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$, combining (16) & (17) gives

$$E(y|\mathbf{x}) = \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\left[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\right] = \Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}\boldsymbol{\beta} + \sigma\phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right) \tag{18}$$

Equation (18) shows when y follows a Tobit model, $E(y|\mathbf{x})$ will be a nonlinear function of $\mathbf{x}$ and $\boldsymbol{\beta}$ and the right-hand side will be strictly positive for all values of $\mathbf{x}$ and $\boldsymbol{\beta}$. With $\boldsymbol{\beta}$ estimates, we can be sure that predicted values for y (estimates of $E(y|\mathbf{x})$) are positive, but this comes at a cost: equation (18) is more complicated than a linear model. (Tobit is a nonlinear model, which is complicated, but it ensures that $E(y|\mathbf{x}) > 0$).
(4 Marks for conceptual / intuitive explanation of above)
The partial effects of x_j on $E(y|y>0,x)$ and $E(y|\mathbf{x})$ have the same sign as the coefficient β_j but the magnitude of the effects depend on values of *all* explanatory variables and parameters. Since σ appears in (18), the partial effects depend on σ also. This is an extra complication of the Tobit relative to the logit/probit models. Let x_j be continuous and assume x_j is not related to other regressors. Then

$$\frac{\partial E(y|y>0,\mathbf{x})}{\partial x_j} = \beta_j + \beta_j \cdot \frac{d\lambda}{dc}\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$$

Differentiating $\lambda(c) = \frac{\phi(c)}{\Phi(c)}$ using $\frac{d\Phi}{dc} = \phi(c)$ and $\frac{d\phi}{dc} = -c\phi(c)$, we get that

$$\begin{aligned}\lambda'(c) &= \frac{-c\Phi(c)\phi(c) - \phi^2(c)}{\Phi^2(c)} \\ &= -c\lambda(c) - \lambda^2(c) = -\lambda(c)[c + \lambda(c)]\end{aligned}$$

and so

$$\frac{\partial E(y|y>0,\mathbf{x})}{\partial x_j}=\beta_j\left\{1-\lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\left[\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}+\lambda\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)\right]\right\} \tag{19}$$

Estimate (19) by plugging in MLEs of β_j and σ. The subtlety is that σ appears in the partial effects directly so it is crucial to estimate this for estimatig partial effects; σ is called the 'ancillary' parameter. Equation (19) shows the partial effect of x_j on $E(y|y>0,\mathbf{x})$ is not determined just by β_j – there is an adjustment factor in brackets that depends on a linear function of $\mathbf{x}$.

(2.5 Marks for conceptual/intuitive description of above; 2.5 Marks for conceptual/intuitive description of below) Also note that

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j}=\frac{\partial P(y>0|\mathbf{x})}{\partial x_j}\cdot E(y|y>0,\mathbf{x})+P(y>0|\mathbf{x})\cdot\frac{\partial E(y|y>0,\mathbf{X})}{\partial x_j} \tag{20}$$

and since $P(y>0|\mathbf{x})=\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$

$$\frac{\partial P(y>0|\mathbf{x})}{\partial x_j}=\frac{\beta_j}{\sigma}\phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$$

We can estimate each term in (20) once we plug in MLEs of β_j and σ and particular values of the x_j.

We can roughly compare OLS and Tobit estimates from

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j}=\beta_j\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$$

We simply multiply Tobit estimates by the adjustment factor $\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$ to compare Tobit estimates with OLS estimates. (1.5 Marks) OLS slope coefficients (e.g. $\hat{\gamma}_j$ from regressing y_i on $x_{i1},x_{i2},\dots,x_{ik}$ where $i=1,\dots,n$, i.e. using all of the data) are direct estimates of $\partial E(y|\mathbf{x})/\partial x_j$. To make Tobit coefficients $\hat{\beta}_j$ comparable to $\hat{\gamma}_j$, we multiply $\hat{\beta}_j$ by an adjustment factor, $\Phi\left(\frac{\mathbf{x}\boldsymbol{\beta}}{\sigma}\right)$, which we can compute following two approaches:

1. Evaluate $\Phi\left(\frac{\mathbf{x}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)$ at sample averages to obtain $\Phi\left(\frac{\bar{\mathbf{x}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)$.
2. Average the individual adjustment factors $n^{-1}\sum_{i=1}^{n}\Phi\left(\frac{\bar{\mathbf{x}}_i\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)$.

For comparing scaled Tobit coefficients to OLS coefficients, the second scale facgtor generally is more appropriate. Both scale factors tend to be closer to one when there are relatively few observations with $y_i = 0$. In the extreme cases that all $y_i > 0$, Tobit and OLS estimates are identical. With discrete explanatory variables, comparing OLS and Tobit estimates is not so easy; however, scale factor for continuous explanatory variables is often a useful approximation. (1.5 Marks for reasonable elaboration on adjustment factor)

Regarding how we can informally evaluate whether Tobit is appropriate, we should first estimate a probit where

$$w = \begin{cases} 1 & y > 0 \\ 0 & y = 0 \end{cases}$$

Then from (14), i.e.

$$P(y_i = 0|\mathbf{x}_i) = 1 - \Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

w follows a probit model where the coefficient on x_j is $\gamma_j = \frac{\beta_j}{\sigma}$ so we can estimate the ratio of β_j to σ by probit for each j. If the Tobit model is adequate, then the probit estimate $\hat{\gamma}_j$ should be close to $\hat{\beta}_j/\hat{\sigma}$ where these are Tobit estimates. They need never be identical due to sampling error but warnging signs include different signs or same signs and much different magnitude differences; sign changes or magnitude differences on

explanatory variables that are insignificant in both models should not be a cause for concern.
(3 Marks; 2 Marks if no discussion)

<u>(ii)</u>

A count variable is a variable that can be zero or take on a few (and only a few) different values. Examples may include the number of times an individual is arrested in a year, the number of times someone was sick in a given month, the number of chilren ever born to a woman, the number of times a person was ever married, etc. Since the logarithm of zero does not exist (the exponential of zero is one) and count variables are designed to account for corner solutions, i.e. where there are a lot of zeros, we cannot use logarithms. (1 mark for definition of count variable, 2 marks for any appropriate example and 2 mark for explaining why we cannot use logarithms)

Observe that since

$$E(y|x_1, x_2, \ldots, x_k) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \tag{21}$$

the log of expected value will be linear. (3 Marks)

$100\beta_j$ is the approximate change in $E(y|\mathbf{x})$ given a one-unit increase in x_j, i.e.

$$\%\Delta E(y|\mathbf{x}) \approx (100\beta_j)\Delta x_j$$

When a more accurate estimate is needed, we can look at discrete changes in the expected value. Keep all explanatory variables except x_k fixed and let x_k^0 be initial value and x_k^1 be subsequent value. Proportionate change in expected value is

$$\left[\frac{\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^2)}{\exp(\beta_0 + \mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} + \beta_k x_k^0)}\right] - 1 = \exp(\beta_k \Delta x_k) - 1$$

where $\mathbf{x}_{k-1}\boldsymbol{\beta}_{k-1} = \beta_1 x_1 + \cdots + \beta_{k-1}x_{k-1}$ and $\Delta_k = x_k^1 - x_k^0$. If $\Delta x_k = 1$ (e.g. x_k takes two values, either 0 or 1 and this time we go from 0 to 1), then the change in the expected value of $y|\mathbf{x}$ is $\exp(\beta_k) - 1$. Given $\hat{\beta}_k$, we obtain $\exp(\hat{\beta}_k) - 1$ and multiply by 100 to turn proportionate change into percentage change. If β_j multiplies $log(x_j)$, then β_j is an elasticity. (4 Marks for reasonable discussion of interpretting β_j)

Since (21) is nonlinear in parameters due to the exponential function, we cannot use linear regression methods to estimate the Poisson model. We could use nonlinear least squares, but all standard count distributions are heteroscedastic and nonlinear least squares does not exploit this so we rely on **maximum likelihood estimation** (2 Marks just for mentioning MLE) and an important related method of quasi-maximum likelihood estimation. A count variable cannot have a Normal distribution because the Normal distribution is for continuous variables that can take on all (or a large range approximately) of values and as a count variable takes on very few values, the distribution is very different from a Normal; instead, we use the Poisson distribution for count data. Writing the right-hand side of (21) as $\exp(\mathbf{x}\boldsymbol{\beta})$, the probability that $y = h|\mathbf{x}$ is

$$P(y = h|\mathbf{x}) = \frac{1}{h!}e^{-e^{-\mathbf{x}\boldsymbol{\beta}}}\left(e^{\mathbf{x}\boldsymbol{\beta}}\right)^h \quad h = 0, 1, \ldots$$

Given a random sample $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$, likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}) &= \Pi_i P(y_i) \\ &= \Pi_i \frac{1}{y_1!}e^{-e^{\mathbf{x}_i\boldsymbol{\beta}}}\left(e^{\mathbf{x}_i\boldsymbol{\beta}}\right)^{y_i} \end{aligned}$$

and so the log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} y_i \mathbf{x}_i \boldsymbol{\beta} - e^{\mathbf{x}_i \boldsymbol{\beta}} - \log(y_i!)\end{aligned}$$

but note that $\log(y_i!)$ does not contain $\boldsymbol{\beta}$ so it does not affect the maximisation of the log-likelihood, so we can concentrate on

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \mathbf{x}_i \boldsymbol{\beta} - \exp(\mathbf{x}_i \boldsymbol{\beta})$$

Poisson MLEs are not obtained in closed form but econometric packages (e.g. Stata) can estimate these models via maximum likelihood with ease. (3 Marks for elaboration)

While Poisson MLE is a natural first step for count data, it's often too restrictive since all probabilities and higher moments of Poisson distributions are determined entirely by the mean, e.g.

$$Var(y|\mathbf{x}) = E(y|\mathbf{x})$$

This has been shown to be violated in many applications. (4 Marks)

Fortunately, the Poisson distribution has a nice robustness property: whether or not the Poisson distribution holds, we still get consistent, asymptotically Normal estimates of β_j like the Normality assumption of OLS (consistent and asymptotically Normal irrespective of the Normality assumption). (2 Marks) When we use Poisson MLE but do not assume that the Poisson distribution is entirely correct, we call the analysis *quasi-maximum likelihood estimation* (QMLE); most econometric packages do this. (2 Marks)

(iii)

When a significant fraction of the population under investigation has been censored, censored models may be useful. Censored models take account of data observability. An example of data censoring is when we know individuals' income up to a point and thereafter we only know that they earn over the threshold. Censoring can arise from survey design. Perhaps there is a box that individuals can tick if they earn over $500,000$ Euros but if they earn less than this, they write in the exact figure. All other covariates, e.g. male/female, level of education, etc. that they have to respond to may be observed, but we only observe income up to $500,000$ and we only know whether an individual earns more than $500,000$ – we do not observe how much exactly. Focusing on the censored normal regression model, y follows the classical linear model and letting i emphasise a random draw from the population

$$\begin{aligned} y_i &= \beta_0 + \mathbf{x}_i\boldsymbol{\beta} + u_i \quad u_i|\mathbf{x}_i \;\; c_i \sim N(0,\sigma^2) \\ w_i &= \min(y_i, c_i) \end{aligned}$$

When we only observe y_i if it is less than a censoring value c_i, we say that we have censoring from above (right censoring); top coding is an example of right data censoring – we know its value only up to a certain threshold; for responses greater than the threshold, we only know the variable is at least as large as the threshold; e.g. family wealth in some surveys is top coded: can respond 'more than $500,000$ Euros and in this case c_i censoring threshold is the same for all i; in many cases c_i changes with individual/family characteristics.

(1.5 Marks for reasonable discussion)

We can estimate $\boldsymbol{\beta}$ and σ^2 by **maximum likelihood** given a random sample $(\mathbf{x}_i, w_i)$. (1.5 Marks for mentioning maximum likelihood; extra 1 Mark for elaborating) We need the density of w_i given ($\mathbf{x}_i, c_i$). For uncensored observatios $w_i = y_i$ and the density of w_i is the same as y_i. For censored observations, we need:

$$P(w_i = c_i|\mathbf{x}_i) = P(y_i \geq c_i|\mathbf{x}_i) = P(u_i \geq c_i - \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi\left[\frac{c_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right]$$

Combine these two parts to obtain density of w_i given $\mathbf{x}_i$ and c_i:

$$f(w|\mathbf{x}_i, c_i) = \begin{cases} 1 - \Phi\left[\frac{c_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] & w = c_i \\ \frac{1}{\sigma}\phi\left[\frac{w - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right] & w < c_i \end{cases}$$

We can obtain the log-likelihood for observation i by taking the natural log of the density for each i. We can maximise the sum of these across i with respect to β_j and σ to get MLEs.

We can interpret β_j just as in linear regression under random sampling – this is different to Tobit say where expectations of interest are nonlinear functions of the β_j.

(2 Marks)

When the data is truncated, we should use a truncated regression model. Truncation means that we throw out or do not use data under some given rule. For example, we may decide to only study individuals who earn up to 1.5 times the income at the poverty line. We do not include any data on any covariate on any individual who earns more. (2 Marks) Unlike a censored regression model, with truncated regression models, we do not observe any information about a certain segment of the population. The truncated regression model differs from the censored regression model since there we observe $\mathbf{x}_i$ for any randomly drawn observation from the population; in the truncated model, we only observe $\mathbf{x}_i$ if $y_i \leq c_i$. (2 Marks)

The truncated normal regression model begins with an underlying population model that satisfies Classical linear regression model assumptions:

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

Given a random sample from the population, OLS is the most efficient estimation procedure. However, we do not observe a random sample from the population – there is a clear deterministic rule truncating the data. A random draw $(\mathbf{x}_i, y_i)$ is only observed if $y_i \leq c_i$ where c_i is the truncated threshold that can depend on exogenous variables, in particular $\mathbf{x}_i$ so if $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ is our observed sample, then y_i is necessarily less than or equal to c_i. So to estimate β_j and σ we need the distribution of $y_i | y_i \leq c_i, \mathbf{x}_i$

$$g(y|\mathbf{x}_i, c_i) = \frac{f(y|\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)}{F(c_i|\mathbf{x}_i\boldsymbol{\beta}, \sigma^2)} \quad y \leq c_i \tag{22}$$

where $f \sim N(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}, \sigma^2)$ and F is a Normal cumulative distribution function with the same mean and variance evaluated at c_i. We are renormalising the density by dividing the area under $f(\cdot|\mathbf{x}_i\boldsymbol{\beta})$ that is to the left of c_i. We take logarithms of (22), sum across i and maximise the result with respect to β_j and σ^2 to get the **maximum likelihood** estimates, leading to consistent, approximately Normal estimators; inference is standard including standard errors.
(2 Marks for mentioning maximum likelihood estimation and an extra 2 marks for elaborating)
OLS applied to a sample truncated from above generally gives estimates biased towards zero. (1 Mark) See figure 3. (1 Mark for graph)

If the sample is selected on the basis of the dependent variable, we have *endogenous* sample selection. (1 Mark) If sample selection is determined solely by an exogenous explanatory variable, then we have *endogenous* sample selection. (1 Mark)

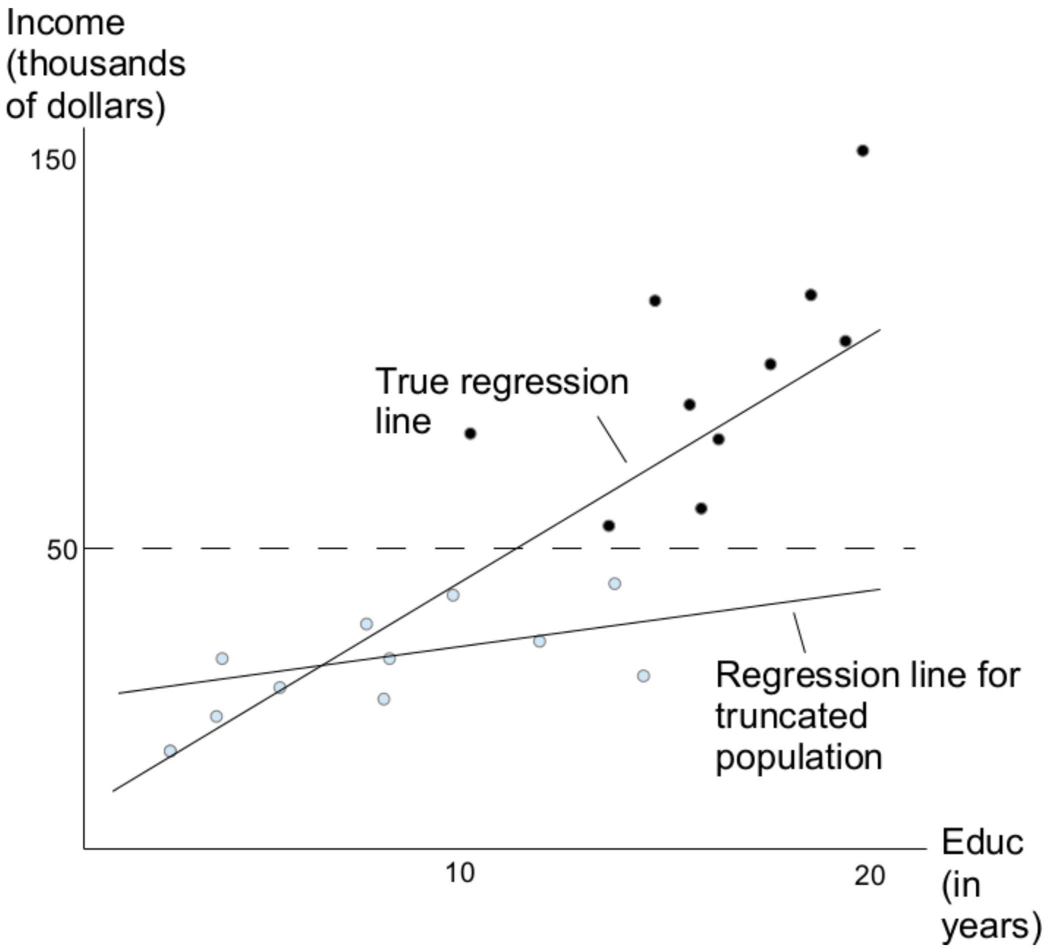

Figure 3: Truncated models: OLS biased towards zero.

Incidental truncation: can only observe income for those who work — wage offer is assumed to be the observed wage so truncation of wage offer is *incidental* because it depends on another variable, *viz*. labour force participation. Usual approach to incidental trunation is to add an explicit *selection equation* to the population model of interest

$$y = \mathbf{x}\boldsymbol{\beta} + u \quad E(u|\mathbf{x}) = 0 \tag{23}$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0] \tag{24}$$

where $s = 1$ if observe y and 0 otherwise. Assume elements of $\mathbf{x}$ and $\mathbf{z}$ are always observed. Our main interest is (23) and we can estimate $\boldsymbol{\beta}$ by OLS given a random sample. The selection equation (24) depends on observed variables z_h and an unobserved error v. The standard assumption is that $\mathbf{z}$ is exogenous in (23):

$$E(u|\mathbf{x}, \mathbf{z}) = 0$$

For the following proposed methods to work well, we will require $\mathbf{x}$ to be a strict subset of $\mathbf{z}$: any x_j is also an element of $\mathbf{z}$ and we have some elements of $\mathbf{z}$ that are not also in $\mathbf{x}$. Assume that v is independent of $\mathbf{z}$ (and thus v is independent of $\mathbf{x}$) and that $v \sim N(0, 1)$. The correlation between u and v generally causes a sample selection problem. Assume u, v are independent of $\mathbf{z}$. Using the fact that $\mathbf{x}$ is a strict subset of $\mathbf{z}$, take

$$E(y|\mathbf{z}, v) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{z}, v) \overset{(u,v)\perp\!\!\!\perp\mathbf{z}}{=} \mathbf{x}\boldsymbol{\beta} + E(u|v)$$

If (u, v) are jointly Normal with zero mean, then $E(u|v) = \rho v$ for some parameter ρ. So

$$E(y|\mathbf{z}, v) = \mathbf{x}\boldsymbol{\beta} + \rho v$$

We do not observe v, but we can use this equation to compute $E(y|\mathbf{z}, s)$

and then specialise this to $s = 1$ to get

$$E(y|\mathbf{z}, s) = \mathbf{x}\boldsymbol{\beta} + \rho E(v|\mathbf{z}, s)$$

Since s and v are related by (24) and $v \sim N(0, 1)$, we can show that $E(v|\mathbf{z}, s)$ is simply the inverse Mills ratio $\lambda(\mathbf{z}\boldsymbol{\gamma})$ when $s = 1$, so

$$E(y|\mathbf{z}, s = 1) = \mathbf{x}\boldsymbol{\beta} + \rho\lambda(\mathbf{z}\boldsymbol{\gamma})$$

Remember that we want to estimate $\boldsymbol{\beta}$ and this equatio shows that we can using only the selected sample as long as we include the term $\lambda(\mathbf{z}\boldsymbol{\gamma})$ as an additional regressor. If $\rho = 0$, $\lambda(\mathbf{z}\boldsymbol{\gamma})$ does not appear and OLS of y on $\mathbf{x}$ using the selected sample consistently estimates $\boldsymbol{\beta}$. Else we have an omitted variable $\lambda(\mathbf{z}\boldsymbol{\gamma})$, which is genneraly correlated with $\mathbf{x}$. When does $\rho = 0$? When u and v are correlated. Since $\boldsymbol{\gamma}$ is unknown, we cannot evaluate $\lambda(\mathbf{z}_i\boldsymbol{\gamma})$ for each i, but from assumptions so far, s given z follows the probit:

$$P(s = 1|\mathbf{z}) = \Phi(\mathbf{z}\boldsymbol{\gamma})$$

This follows because since $v \sim N(0, 1)$ and v is independent of $\mathbf{z}$

$$\begin{aligned} P(s = 1|\mathbf{z}) &= P(\mathbf{z}\boldsymbol{\gamma} + v \geq 0|\mathbf{z}) \\ &= P(v \geq -\mathbf{z}\boldsymbol{\gamma}|\mathbf{z}) \\ &= 1 - \Phi(-\mathbf{z}\boldsymbol{\gamma}) \\ &= \Phi(\mathbf{z}\boldsymbol{\gamma}) \end{aligned}$$

We can estimate $\boldsymbol{\gamma}$ by a probit of s_i on $\mathbf{z}_i$ using the *entire* sample. In the second step, we can estimate $\boldsymbol{\beta}$. This is called the *Heckit method* after Heckman (1976) – James Heckman of Chicago University – who won the Nobel prize in 2000 for this (sample selection correction). The Heckit method for sample selection corrections in the case of incidental truncation is composed of two steps:

1. Using all n observations, estimate probit of s_i on $\mathbf{z}_i$ and obtain estimates $\hat{\boldsymbol{\gamma}}_h$. Compute the inverse Mills ratio $\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\boldsymbol{\gamma}})$ for each i. Actually, we only need these for i such that $s_i = 1$.
2. Using the selected sample, i.e. observations for which $s_i = 1$ (say n_1 of them), run the regression of y_i on $\mathbf{x}_i$, $\hat{\boldsymbol{\lambda}}_i$.

The $\hat{\beta}_j$ are consistent and approximately Normally distributed.
(2 Marks for intuitive explanation of Heckit method; 2 Marks for elaboration and specifics)

A simple test of selection bias is available from regressing y_i on $\mathbf{x}_i$, $\hat{\boldsymbol{\lambda}}_i$: use the usual t statistic on $\hat{\boldsymbol{\lambda}}_i$ as a test of $H_0 : \rho = 0$. Under H_0, there is no sample selection problem. If $\rho \neq 0$, then OLS standard error from this regression are not exactly correct since they do not account for estimation of γ, which uses the same observations in this regression and more. Some econometric packages compute the corrected standard errors, which are not as simple as the heteroscedasticity-adjusted standard errors. In many cases, adjustments do not lead to important differences but it is hard to know that beforehand unless $\hat{\rho}$ is small and insignificant.
(3 Marks – unnecessary to elaborate as long as answer is specific)

Part (b): (25 Marks)

- First part: 1 mark definition, up to 2 marks for explanation / discussion.
  A regressor is endogenous if it is correlated with the error term in the structural equation: $Corr(educ, u) \neq 0$, violating the Classical assumption that regressors are exogenous ($Corr(x, u) = 0$). (1 Mark)

Sample discussion: It could be that there is simultaneity between wages and education in that a higher return to education may inspire people to invest in human capital. If parents earn high incomes, they might emphasise the importance of education to their children by investing their money in their child's education. As individuals start earning higher wages, they may decide to invest more in their education or increase their skills through further education. In this case, a higher wage enables individuals to return to school or college to take extra courses. This is quite common in the medical profession in terms of specialists. As they become more specialised, in certain fields like pediatrics, doctors must constantly educate themselves about the latest developments and often do so through part-time special masters courses. Perhaps individuals with high ability tend to also have more years of education. Ability might positively affect wage, *cet. par.*. Endogeneity can result from omitted variables, measurement errors and simultaneity, among other things.

- It might be that we have an omitted variable (subsumed in the error term) that education is correlated with (e.g. ability (1 Mark)).
- If we leave ability out of our model, our estimates will be subject to omitted-variable bias (1 Mark) and also be inconsistent (sampling distribution does not collapse to the true value of the parameter we want to estimate) (1 Mark).
- Definition: An instrumental variable (IV) z is such that

  1. $Corr(z, x) \neq 0$ – z is **valid** instrument for x. (1 Mark)
  2. $Corr(z, u) = 0$ – z is a **predetermined** variable. (1 Mark)

  So, an IV z for $educ$ must be (i) correlated with education and (ii) uncorrelated with u (ability and any other unobserved factors affecting

wage).
Typically, when parents have a lot of education, there children have a lot education and *vice-versa*; hence z_3 and z_4 are valid instruments as assumption 1 holds. We can test $Cov(z,x) \neq 0$ by estimating

$$x = \pi_0 + \pi_1 z + v$$

and since $\pi_1 = Cov(z,x)/Var(x)$, $Cov(z,x) \neq 0$ if and only if $\pi_1 \neq 0,$ thus we should be able to reject the null hypothesis

$$H_0 : \pi_1 = 0$$

against two-sided alternative $H_0 : \pi_1 \neq 0$ at a sufficiently small significance level (say 5% or 1%). If this is the case, then we can be fairly confident that $Cov(z,x) \neq 0$.
Whether parents' education is correlated with any unobserved factors affecting wage is generally untestable as we do not observe these factors; we usually appeal to economic behaviour or introspection or simply assume there is no such correlation. (1 Mark for a reasonable discussion of why education might be correlated with education and possibly uncorrelated with u)

- The number of siblings an individual has might be a better alternative. (1 Mark)
  1 Mark for discussion of why better, requires arguing number of siblings satisfies instrumental variables assumptions and why parents' education may not.
  Sample discussion: Typically, it is observed that education and the number of siblings are inversely related – individuals from large families tend to have less education relative to individuals from smaller families. So, number of siblings would seem to satisfy assumption 1, i.e. ($Cov(z,x) \neq 0$). As to why $Cov(z,u) = 0$, the number of siblings

has probably nothing to do with the ability of an individual and possibly other unobserved factors that affect wage. However, perhaps mother's / father's education is correlated with ability to the extent that if a mother has a lot of education, it may be that she has/had a high level of ability and this was passed on to her child. If so, mother's education would be correlated with ability of the individual, which is unobserved and part of the error term; hence, we could have $Cov(z, u) \neq 0$ and so assumption 2 would not hold.

- 2SLS is less efficient than OLS if explanatory variables are exogenous – large standard errors – so we might want to test for endogeneity. (1 Mark)

- Hausman (1978) test for endogeneity of $y_2 = educ$:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1 \qquad (25)$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \qquad (26)$$

  where $z_1 = exper$, $z_2 = exper^2$, $z_3 = motheduc$, $z_4 = fatheduc$ and equation (26) is the reduced form equation for y_2 (reduced form since all regressors on right-hand side are exogenous/predetermined (in this case all are exogenous – no lagged endogenous variables)). Equation (25) is the structural equation. We want to test for possible endogeneity of $y_2 = educ$. If all z's are exogenous, then they are all uncorrelated with u_1, the disturbance term from the structural equation, i.e. $Corr(z_j, u_1) = 0$ for $j = 1, 2, 3, 4$. This implies that the only way y_2 is endogenous, i.e. $Corr(y_2, u_1) \neq 0$ is if $Corr(v_2, u_1) = 0$. We want to test whether in fact $Corr(v_2, u_1) = 0$. Let $u_1 = \delta_1 v_2 + e_1$ where $Corr(e_1, v_2) = 0$ and $E(e_1) = 0$; so, we have expressed the structural disturbance term u_1 as a linear function of v_2; if u_1 is a linear function of v_2, then clearly v_2 and u_1 are related, i.e. $Corr(v_2, u_1) \neq 0$. We want to test this, i.e. we want to test whether $\delta_1 = 0$ since if $\delta_1 = 0$, then $u_1 = e_1$, i.e. u_1 is not a function of / related to v_2. We could test this by putting v_2 as an extra regressor in (26) and doing a t-test; however, we do not observe v_2. Hausman proposed a neat solution. We estimate the reduced form (26) by OLS to get

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 + \hat{\pi}_4 z_4$$

and then simply recognize that $y_2 - \hat{y}_2 = \hat{v}_2$; and so we get $\hat{v}_2$. Next we use $\hat{v}_2$ as an additional regressor in (26). To do this, recall that we let $u_1 = \delta_1 v_2 + e_1$, so simply replace v_2 by $\hat{v_2}$ and substitute this for u_1 in (25) to get

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + \text{error} \qquad (27)$$

Run OLS on (27) and do a t-test of $H_0 : \delta_1 = 0$. Rejection implies endogeneity since $Corr(v_2, u_1) \neq 0$. (6 Marks)

The reduced equation for $educ$ is

$$educ = \pi_0 + \pi_1 exper + \pi_2 exper^2 + \pi_3 motheduc + \pi_4 fatheduc + v_2 \qquad (28)$$

and identification requires that at least one of π_3 and π_4 is non-zero, i.e. $\pi_3 \neq 0$ or $\pi_4 \neq 0$ or both. Testing $H_0 : \pi_3 = 0, \pi_4 = 0$ in (28) using an F-test, we get $F = 55.40$ and p-value $= .0000$. As expected, $educ$ is (partially) correlated with parents' education. (2 Marks)

STATA PART:

- The estimated return to education is about 6.1%. (2 Marks)
- The 2SLS estimate is barely statistically significant at the 5% level against a two-sided alternative due to its relatively large standard error. (1 Mark)

- No, $R^2 = .136$ is not a cause for concern. (1 Mark) 2SLS R^2 can be negative since $R^2 = 1 - SSR/SST$ is negative if $SSR > SST$ so not very useful to report R^2 for 2SLS estimation. When regressors are endogenous $Corr(x, u) \neq 0$ say, we cannot decompose the variance of y into $\beta_1^2 Var(x) + Var(u)$ so R^2 has no natural interpretation. Goodness of fit is not a factor – goal of IV / 2SLS is to provide better estimates of *cet. par.* effect of x on y when x and u are correlated. If goal is to produce the largest R^2, always use OLS but high R^2 from OLS is of little comfort if we cannot consistently estimate β_1. (1 Mark: only need one of these sentences)