Attempt TWO question from the four questions in this section.

Question 1 (100 Marks) – Identification & Simultaneous Equations Models.

Part (a): (50 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

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.

data on six, so $P(z = 0) = \frac{6}{120} = 0.05$ is the fraction of missing data; remember $P_N(z = 1) = \frac{1}{N}\sum_{i=1}^{N} 1[z_i = 1]$.

i. 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:

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

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

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

We know

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

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

$$\begin{aligned} P_N(y \in B|z=1) &= \frac{\sum_{i=1}^{N} 1[y_i \in B, z_i = 1]}{\sum_{i=1}^{N} 1[z_i = 1]} \\ &= \frac{18+10+15+21+15+3}{120-6} \\ &= \frac{82}{114} \\ \therefore P(y \in B) &= \frac{92}{114} \times 0.95 + [0,1] \times 0.05 \\ &\in \left[\frac{41}{60}, \frac{11}{15}\right] \\ &\equiv H[P(y \in B)] \end{aligned}$$

which is our identification region for the probability that a student passes.

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

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

We know

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

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

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

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.

iii. No, this statement does not accurately describe the empirical finding. We can only say that students who are members of water-polo clubs on average scored higher than those who are not. We cannot say that the very fact of such membership increased the probability of a

student scoring highly.
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.
However, if the students were **randomly assigned** with membership or non-membership of water-polo clubs, then the researcher would be correct in saying that membership of water-polo clubs increases the probability that a student will do better on average than a student who is not a member of a water-polo club. But since we are dealing with what actually happened (descriptive) we cannot say that having such membership increases the probability that a student scores highly.

Part (b): (50 Marks)

i. Endogenous variables (determined within model) $M = 4$: C, I, T, Y. Predetermined variables (exogenous [determined outside the model] and lagged endogenous) $K = 3$: exogenous: 1, G and lagged endogenous: Y_{t-1}. The tax equation is a *behavioural* equation – it attempts to explain the behaviour of economic agents.

ii. The structural parameters (arranged) are

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

Focusing on the tax function, the order condition is checked by:

$$\begin{aligned}
K - k &= 2 \\
m - 1 &= 1 \\
\therefore\ K - k &> m - 1
\end{aligned}$$

Alternatively

$$\begin{aligned}
M + K - (m + k) &= 4 \\
M - 1 &= 3 \\
\therefore\ M + K - (m + k) &> M - 1
\end{aligned}$$

In both cases, the order condition is satistfied as a strict inequality, so the tax function *may* be *over* identified. 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. Checking the rank condition:

$$\Lambda_T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -b_1 \\ -1 & -1 & -1 & 0 \end{bmatrix}$$
$$\rho(\Lambda_T) = 3 = M - 1 = 3$$

Therefore, the tax function is **over** identified, given the order condition result.

Extra on how we got $\rho(\Lambda_T) = 3 = M - 1 = 3$: rank cannot exceed 3, but could be 2 (look only at rows). The rank must be 3 if we can sensibly estimate the tax 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.

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

$$det(\Lambda_T) = -1 - b_1 \neq 0$$

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

$$\alpha_1(1\ \ 0\ \ 0\ \ 0) + \alpha_2(0\ \ 1\ \ 0\ \ -b_1) + \alpha_3(-1\ \ -1\ \ -1\ \ 0) = 0$$

which is equal to zero if and only if

$$\alpha_1 - \alpha_3 = 0 \tag{2}$$
$$\alpha_2 - \alpha_3 = 0 \tag{3}$$
$$-\alpha_3 = 0 \tag{4}$$
$$-b_1\alpha_2 = 0 \tag{5}$$

Equations (2)–(4) imply that

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

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.

As the tax equation is overidentified, a single equation method such as 2SLS would be an appropriate estimation technique; system methods include 3SLS and FIML. As for other single equation methods, OLS leads to simultaneity bias and ILS only works for just-identified case, so both would be inappropriate in this case.

(a) Estimate reduced form equations for endogenous variables on right-hand side of structural form equations of interest via OLS and obtain reduced form estimates of these endogenous variables.

(b) Substitute for right-hand side endogenous variables in structural equation using reduced form estimates from stage 1 as proxies (instruments) and estimate structural form equation by OLS. This produces 2SLS estimates of structural form parameters.

Question 2 (100 Marks) – Limited Dependent Variables & Instrument Variables.

Part (a): (50 Marks)

i. 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. 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. The coefficient $\hat{\delta}_0 = -0.02$ in this case means that the differential effect of being a female is associated with a $100 \times \hat{\delta}_0 = -2\%$ change in wages, i.e. women earn on average two percent less than men in hourly wages for a given level of education. To compute the exact percentage difference in predicted wages for a woman relative to a man, we calculate $100[\exp \hat{\delta}_0 - 1] \approx -1.98$ to two decimal places; a woman earns on average 1.98$\%$ less than a man with the same level of education. 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]

ii. Limitations: (any two)

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

(b) Constant partial effects. Need to explain.

(c) 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,

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

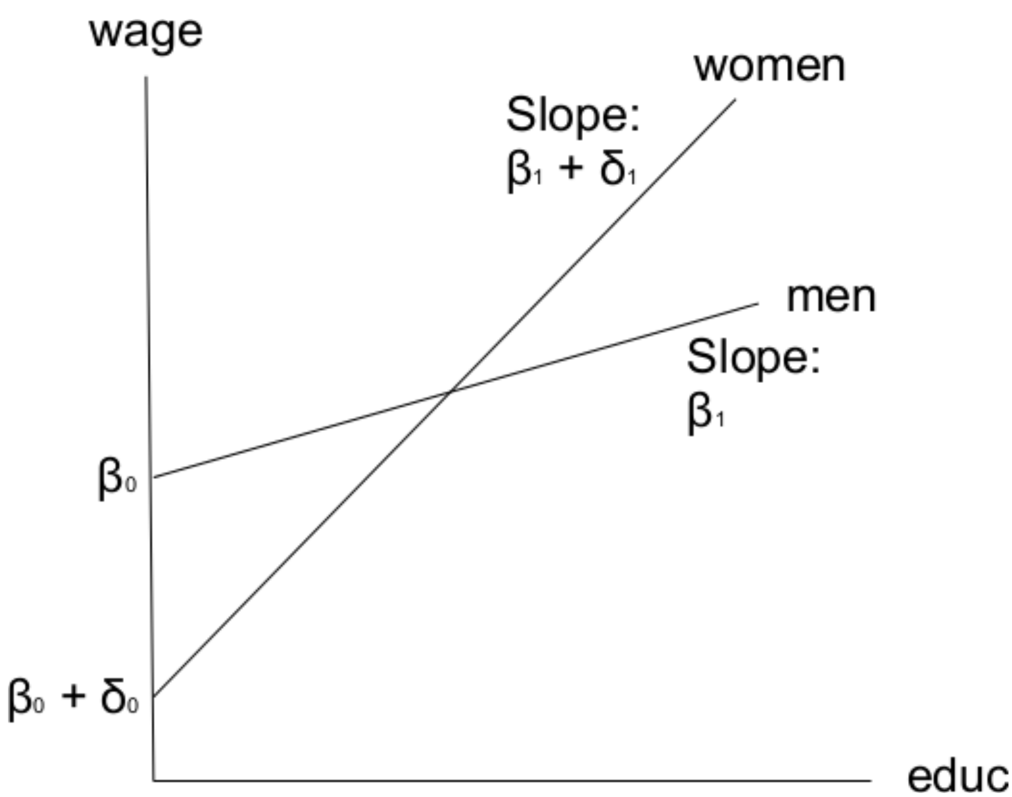

Figure 1: Differential intercept and slope for return to education between men and women.

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.

(d) 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.

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

$$f(y|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^y[1 - G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y} \quad y = 0, 1 \tag{6}$$

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 (6):

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

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

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \tag{7}$$

MLE of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}$ maximises (7). 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.

iii. An alternative to R^2 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' prediction. So, the proportion of correct predictions is given by

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

Part (b): (50 Marks)

i. 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 ($Corr(x, u) = 0$).
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 edu-

cation 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). If we leave ability out of our model, our estimates will be subject to omitted-variable bias and also be inconsistent (sampling distribution does not collapse to the true value of the parameter we want to estimate).

Definition: An instrumental variable (IV) z is such that

(a) $Corr(z, x) \neq 0$ – z is **valid** instrument for x.

(b) $Corr(z, u) = 0$ – z is a **predetermined** variable.

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 ia 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. The number of siblings an individual has might be a better alternative. 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 the first assumption, 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 ib would not hold.

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

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

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u_1 \quad (8)$$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \quad (9)$$

where $z_1 = exper$, $z_2 = tenure$, $z_3 = motheduc$, $z_4 = fatheduc$ and equation (9) 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 (8) 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 (9) and doing a t-test; however, we do not observe v_2. Hausman proposed a neat solution. We estimate the reduced form (9) 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 (9). 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 (8) to get

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + \text{error} \tag{10}$$

Run OLS on (10) and do a t-test of $H_0 : \delta_1 = 0$. Rejection implies endogeneity since $Corr(v_2, u_1) \neq 0$. The reduced equation for $educ$ is

$$educ = \pi_0 + \pi_1 exper + \pi_2 tenure + \pi_3 motheduc + \pi_4 fatheduc + v_2 \tag{11}$$

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 (11) using an F-test, we get $F = 55.40$ and p-value $= .0000$. As expected, $educ$ is (partially) correlated with parents' education.

iii. The estimated return to education is about 6.1%.
No, $R^2 = .136$ is not a cause for concern. 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.