Faculty of Economics

Part 2 Paper 3: Microeconometrics and Panel Data
Supervision 1

1. Consider a simple model to estimate the effect of personal computer (PC) ownership on college performance (CP) at a large university:

$$CP_i = \alpha + \beta PC_i + \varepsilon_i.$$

where PC is a binary variable indicating PC ownership.

a) Why might PC ownership be correlated with ε?

b) Explain why PC is likely to be related to parents' annual income. Does this mean parental income is a good IV for PC?

c) Now suppose that 4 years ago the university gave grants to buy computers to approximately one-half of the incoming students, and the students who received grants were randomly chosen. Explain how you might use this information to construct an instrumental variable for PC.

2. Suppose that you wish to estimate the effect of class attendance on student performance at a particular college. A basic model might be written as

$$stndex_i = \alpha + \beta_1 attend_i + \beta_2 priGPA_i + \beta_3 ACT_i + \varepsilon_i.$$

where $stndex$ denotes the outcome on a final exam, $attend$ represents the percentage of classes attended, $priGPA$ is the pre-college Grade Point Average, and ACT denotes the score on the college entrance exam.

a) Let $dist$ be the distance from the students lodgings to the lecture hall. Is this variable likely to be correlated with ε?

b) Assuming that $dist$ and ε are uncorrelated what other assumptions must $dist$ satisfy in order to be a valid instrument for $attend$?

c) Suppose we now add an interaction term $priGPA \cdot attend$, namely

$$stndex_i = \alpha + \beta_1 attend_i + \beta_2 priGPA_i + \beta_3 ACT_i + \beta_4 (priGPA \cdot attend)_i + \varepsilon_i.$$

What might be a good IV for $priGPA \cdot attend$?

Hint: if $E(\varepsilon | priGPA, ACT, dist) = 0$, as happens when $priGPA$, ACT, and $dist$ are exogenous, then any function of $priGPA$ and $dist$ is uncorrelated with ε.

3. Consider a simple time series model where the explanatory variable has classical measurement error

$$\begin{aligned} y_t &= \alpha + \beta x_t^* + u_t \qquad (1) \\ x_t &= x_t^* + \varepsilon_t, \end{aligned}$$

where u_t has zero mean and is uncorrelated with x_t^* and ε_t. We observe y_t and x_t only. Assume that ε_t has zero mean and is uncorrelated with x_t^*; x_t^* also has zero mean.

a) Write $x_t = x_t^* + \varepsilon_t$ and plug this into (1). Show that the composite error term, say v_t, in the new equation is negatively correlated with x_t if $\beta > 0$. What does this imply about the OLS estimator of β from the regression of y_t on x_t?

b) Now assume that in addition u_t and ε_t are uncorrelated with all past values of x_t^* and ε_t; in particular with x_{t-1}^* and ε_{t-1}. Show that $E(x_{t-1} v_t) = 0$.

(Note: this part is a little difficult but students should attempt it)

c) Are x_t and x_{t-1} likely to be correlated? Explain.

d) What do parts b) and c) suggest as a useful strategy for consistently estimating α and β?

4. Consider the simple regression model

$$y = \alpha + \beta x + u,$$

where we believe that x and u are correlated: $Cov(x, u) \neq 0$. Assume that there exists additional information in the form of an observable variable z that satisfies:

i) $Cov(z, u) = 0$; and ii) $Cov(z, x) \neq 0$. By deriving an expression for $Cov(z, y)$ show that the IV estimator for β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

5. Let y^* be a dependent variable that is measured with error, which we write as $e = y - y^*$. Given the following standard regression model

$$y^* = \alpha + \beta_1 x_1 + ... + \beta_k x_k + u,$$

does the OLS estimator deliver unbiased estimates of the model parameters? Are there any other notable properties of the OLS estimator applied to this model?

6. Consider the following wage equation that explicitly recognises that ability affects $\log(wage)$

$$\log(wage) = \alpha + \beta_1 educ + \beta_2 ability + u.$$

The above model shows explicitly that we would like to hold ability fixed when measuring the returns to education. Assuming that the primary interest is in obtaining a reliable estimate of the slope parameters β_1, and that there is no direct measurement for ability, explain how you would do this using a method based upon a proxy variable and an IV estimator. In doing so evaluate the following statement:

whilst IQ is a good candidate as a proxy variable for ability,it is not a good instrumental variable for educ.

7. *Part A 2003 Exam Question*

In a given industry firms relate their stocks of finished goods, y, to their expected sales, X^e, according to the linear relationship

$$y = \beta_1 + \beta_2 X^e.$$

Actual sales, X, differ from X^e by a random quantity, u :

$$X = X^e + u,$$

where $u \sim (0, \sigma^2)$, and $E(u|X^e) = 0$.

(a) If an analyst has data on y and X for a cross-section of firms in the industry, describe the problems encountered if OLS were used to estimate β_2.

(b) The amount of labour, L, employed by firms is also a linear function of expected sales:

$$L = \eta_1 + \eta_2 X^e.$$

Show how this relationship might be exploited to counter the problem of measurement error bias.

8. *Part A 2004 Exam Question*

Consider the following regression model

$$Y = \alpha + \mathbf{X}\boldsymbol{\beta} + Y_1\theta_1 + Y_2\theta_2 + ... + Y_l\theta_l + \varepsilon, \tag{2}$$

where $\mathbf{X}$ is a $1 \times k$ vector, $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters. $Y_1, ..., Y_l$ are endogenous variables, and ε is a stochastic error term.

(a) Assuming that the analyst has at her disposal a vector $\mathbf{Q}$ which contains both the elements of $\mathbf{X}$ plus an additional number of exogenous variables in $\mathbf{Z}$, where $\mathbf{Z}$ is a $1 \times l$ vector. Write down the necessary condition for identification of the parameters $\theta_1, \theta_2, ..., \theta_l$.

(b) In what sense may the instrumental variable estimator be considered a two stage least squares estimator?

(c) In the reduced form regression equation

$$Y_1 = \varpi + X_1\delta_1 + X_2\delta_2 + ... + X_k\delta_k + Z_1\lambda_1 + Z_2\lambda_2 + ... + Z_p\lambda_p + \varepsilon, \tag{3}$$

write down the form of the null hypothesis required to test for identification of the parameter θ_1. What test statistic would you use?

B1 *Part B 2006 Exam Question*

Consider the following linear model of log wages (w) explained using years of schooling (S), years of experience and its square (E, E^2), and

3 dummy variables indicating whether the individual was black (B), lived in the south (Sth), and lived in a metropolitan area (Sm).

$$w_i = \alpha + \beta_1 S_i + \beta_2 E_i + \beta_3 E_i^2 + \beta_4 B_i + \beta_5 Sth_i + \beta_6 Sm_i + \varepsilon_i \quad (4)$$

Given that experience is not directly measurable we use the proxy $E_i = age_i - S_i - 6$, assuming that people start school at the age of 6. ε_i is an unobserved error term, and i indexes individuals.

A reduced form model for schooling is written as

$$S_i = \delta + \mathbf{z}_i' \boldsymbol{\pi} + v_i, \quad (5)$$

where $\mathbf{z}_i$ is a $L \times 1$ vector including all the exogenous variables in (4), $\boldsymbol{\pi}$ is a $L \times 1$ vector of unknown parameters, δ is an unknown scalar parameter, and v_i is an error term.

a) Why might $Cov(v_i, \varepsilon_i)$ be non-zero?

b) Assuming that S_i is endogeneous how might you use the reduced form equation in conjunction with (4) to identify the parameter β_1?

c) In Tables 1.1 and 1.2 we report results from estimating two models based on (4). Data is a sample of 3010 men taken from the US National Longitudinal Survey of Young Men; the year is 1976.

   Provide a careful comparison of the results based on the OLS and IV estimators. Discuss the choice of instruments.

d) If instead of a single year, data was available for the same individuals for 1976-1980, discuss how one might use a fixed effects estimator to provide inference on the effects of schooling on wages.

Table 1.1 Wage Equation Estimated by OLS

Dependent: log(wage)

Variable	Estimate	Std. Error	t-ratio
constant	4.734	0.068	70.02
S	0.074	0.004	21.11
E	0.084	0.007	12.58
E^2	-0.002	0.000	-7.05
B	-0.189	0.018	-10.76
Sm	0.181	0.016	10.37
Sth	-0.125	0.015	-8.26

$R^2 = 0.291$, F=204.93

Table 1.2 Wage Equation Estimated by IV

Dependent Variable: log(wage)

Variable	Estimate	Std. Error	t-ratio
constant	4.066	0.609	6.68
S	0.133	0.051	2.58
E	0.056	0.026	2.15
E^2	-0.008	0.001	-0.06
B	-0.103	0.077	-1.33
Sm	0.108	0.005	2.17
Sth	-0.098	0.028	-3.41

Instruments: age, age^2,lived near college
used for: E, E^2, and S
R^2 for reduced form for S: 0.119

Dr. M. Weeks
Lent 2008