Microeconometrics
Supervision 1

Samuel Lee

Question 1

Of the variables that are mentioned in this question, "crmrte" is the crimes committed per person, "polpc" is the police per capita, "west", "central" are dummy variables equal to 1 if the county is in western or central N.C. (presumably North Carolina), "urban" is a dummy variable equal to 1 if the county is in a standard metropolitan statistical area (SMSA), and "taxpc" is the tax revenue per capita.

Table 1: Summary statistics for CRIME.dta

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
county	630	100.600	58.036	1	51	151	197
year	630	84.000	2.002	81	82	86	87
crmrte	630	0.032	0.018	0.002	0.018	0.038	0.164
prbarr	630	0.307	0.171	0.059	0.218	0.353	2.750
prbconv	630	0.689	1.690	0.068	0.348	0.636	37.000
prbpris	630	0.426	0.087	0.149	0.374	0.483	0.679
avgsen	630	8.955	2.658	4.220	7.160	10.197	25.830
polpc	630	0.002	0.003	0.0005	0.001	0.002	0.036
density	630	1.386	1.440	0.198	0.533	1.508	8.828
taxpc	630	30.239	11.455	14.303	23.426	33.271	119.761
west	630	0.233	0.423	0	0	0	1
central	630	0.378	0.485	0	0	1	1
urban	630	0.089	0.285	0	0	0	1

Table 1 shows the summary statistics for some of the variables. There doesn't seem to be anything out of the ordinary, so we proceed to the linear regression of "crmrte" on "polpc", "west", "central", and "urban".

Table 2 shows the summary statistics of the OLS regression. The coefficient on "west" is negative and significant at the 1% level, so this roughly means that counties in Western N.C. are on average associated with a crime rate that is lower by 1.5 percentage points relative to counties not in Western or Central N.C. (since the dummy for Central N.C. is also included in the regression), after taking into account the police per capita and whether the county is in an MSMA. So far this says nothing about causality.

If we take this 'naïve' model at face value, the positive and siginifcant coefficient on "polpc" suggests that counties with higher police per capita tend to have a higher crime rate. Realistically,

Table 2: OLS estimates of 'naïve' linear model

	Dependent variable:
	crmrte
polpc	1.291^{***}
	(0.196)
west	-0.015^{***}
	(0.001)
central	-0.002^{*}
	(0.001)
urban	0.034^{***}
	(0.002)
Constant	0.030^{***}
	(0.001)
Observations	630
R^2	0.454
Adjusted R^2	0.451
Residual Std. Error	0.013 (df = 625)
F Statistic	130.017^{***} (df = 4; 625)
Note:	*p<0.1; **p<0.05; ***p<0.01

any positive association probably runs in the other direction: a higher crime rate induces local authorities to retain more police per capita. In any case this is at best an association for now.

It seems plausible to impose this simultaneous equations model:

$$crmrte_i = \beta_0 + \beta_{pol} polpc_i + \beta_{west} west_i + \beta_{cen} central_i + \beta_{urb} urban_i + \varepsilon_i \quad (1)$$

$$polpc_i = \gamma_0 + \gamma_{crm} crmrate_i + \gamma_{west} west_i + \gamma_{cen} central_i + \gamma_{urb} urban_i + \gamma_{tax} taxpc_i + \nu_i \quad (2)$$

We've included location-specific fixed effects in both the equations that determine crime rate and police per capita; this should be defensible since local institutions can differ in a way that affects both. There may be reasons for a positive correlation between tax revenue per capita and the crime rate. For instance, "taxpc" could be proxying for the income per capita in a county. For now, we abstract from these possibilities and work with the assumption that "taxpc" only has an effect on "polpc", perhaps through the effects on the funding of law enforcement departments. If we are using "taxpc" as an IV for "polpc", it means we are trying to estimate β_{pol} in the first equation. If we are trying to identify equation (1), there must be some exogenous variables in equation (2) but not in equation (1). "taxpc" satisfies this criterion. The reason why a 'naive' linear regression

would not work and an IV is needed can be seen by substituting (1) into (2), yielding

$$\begin{aligned} polpc_i &= \gamma_0 + \gamma_{crm}(\beta_0 + \beta_{pol}polpc_i + \beta_{west}west_i + \beta_{cen}central_i + \beta_{urb}urban_i + \varepsilon_i) \\ &+\gamma_{west}west_i + \gamma_{cen}central_i + \gamma_{urb}urban_i + \gamma_{tax}taxpc_i + \nu_i \\ &= \frac{1}{1-\gamma_{crm}\beta_{pol}}[\gamma_0 + \gamma_{crm}\beta_0 + (\gamma_{crm}\beta_{west} + \gamma_{west})west_i + (\gamma_{crm}\beta_{cen} + \gamma_{cen})central_i \\ &+(\gamma_{crm}\beta_{urb} + \gamma_{urb})urban_i + \gamma_{tax}taxpc_i + \gamma_{crm}\varepsilon_i + \nu_i \end{aligned}$$

and it is immediately apparent that $\text{Cov}[polpc_i, \varepsilon_i] \neq 0$ which leads to biased and inconsistent estimates of β_{pol} in a 'naïve' linear regression of equation (1). For "taxpc" to be a valid IV for "polpc", γ_{tax} should be non-zero (the relevance condition) and there shouldn't be any non-zero β_{tax} (the exclusion restriction), hence no such coefficient appears in equation (1). And of course, "taxpc" cannot be correlated with ε_i.

Table 3: IV and OLS estimates of (1)

	Dependent variable:	
	crmrte	
	OLS	2SLS
	(1)	(2)
polpc	1.291***	2.315
	(0.196)	(1.716)
west	−0.015***	−0.015***
	(0.001)	(0.002)
central	−0.002*	−0.002
	(0.001)	(0.001)
urban	0.034***	0.034***
	(0.002)	(0.002)
Constant	0.030***	0.029***
	(0.001)	(0.003)
Observations	630	630
R^2	0.454	0.430
Adjusted R^2	0.451	0.427
Residual Std. Error (df = 625)	0.013	0.014
F Statistic	130.017*** (df = 4; 625)	
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 3 shows the summary statistics of the 2SLS regression alongside those of the OLS regression. The estimated coefficients are remarkably similar between the two models other than the one for "polpc". Even more surprisingly, the 2SLS estimate for "polpc" is still positive and

in fact almost twice that of the OLS estimate, although the larger standard error makes the 2SLS result more obscured by statistical noise. Although we may be less confident in the precision of the 2SLS estimate, it is still the best estimate we have of the coefficient on "polpc" provided the simultaneous equations model was appropriately specified. It suggests that increasing the number of police per capita will *increase* the crime rate in a county. If we take the description of "crmrte" literally ("crimes committed per person"), this may be difficult to reconcile, but it is more likely that "crmrte" captures crimes reported or identified. It is plausible that having more police around makes it easier to identify and report crime, hence an increase in "crmrte". Otherwise, it could simply be because "taxpc" doesn't satisfy the exclusion restriction for the reasons mentioned before. It is difficult to extract any conclusions of the effect of crime on police employment. Even though we know that the 2SLS estimate could suggest the OLS estimate is biased downwards, with 4 regressors it is not easy to derive which coefficients contribute to this downward bias, and subsequently it is difficult to extract any information about the sign of γ_{crm}.

As mentioned before, a reason why "taxpc" may not be a valid instrument is that it could simply be proxying for a higher income per capita, and there is an empirical assocation between higher material standards of living and lower crime rates.

If "polpc" is exogenous (along with "west", "central", and "urban"), and "taxpc" is a valid instrument, then in the following model

$$polpc_i = \delta_0 + \delta_{west} west_i + \delta_{cen} central_i + \delta_{urb} urban_i + \delta_{tax} taxpc_i + \eta_i \tag{3}$$

"polpc" can only be exogenous if η_i is uncorrelated with ε_i (γ_{crm} cannot appear in the above or it is endogenous by default). A similar way of expressing this condition is that in the following model

$$crmrte_i = \beta_0 + \beta_{pol} polpc_i + \beta_{west} west_i + \beta_{cen} central_i + \beta_{urb} urban_i + \alpha\eta_i + \varepsilon_i \tag{4}$$

"polpc" is only exogenous if $\alpha = 0$. We can check this by estimating the parameters above using OLS, where we use the residuals from the OLS estimates of equation (3) as an estimator for η_i.

Table 4 shows the results of the estimation of equation (4) using the estimated residuals from equation (3). Although the estimate of α is negative and the magnitude seems large, the standard errors are also large enough that we cannot reject the hypothesis that $\alpha = 0$. Under these terms we fail to reject the hypothesis that "polpc" may be exogenous.

Table 4: Results of endogeneity test

	Dependent variable:
	crmrte
polpc	2.315
	(1.681)
west	-0.015^{***}
	(0.002)
central	-0.002^{*}
	(0.001)
urban	0.034^{***}
	(0.002)
resid	-1.038
	(1.693)
Constant	0.029^{***}
	(0.003)
Observations	630
R^2	0.455
Adjusted R^2	0.450
Residual Std. Error	0.013 (df = 624)
F Statistic	103.985^{***} (df = 5; 624)
Note:	*p<0.1; **p<0.05; ***p<0.01

Question 2

The first equation is probably a labour supply equation. Ignoring income effects, we should expect b1 to be positive, in that a higher wage makes it more worthwhile to work, thus increasing the desired hours of work. We should expect b2 to be negative, especially for female workers, given that empirically women with more children tend to work fewer hours. This could reflect the higher opportunity cost of work as the number of children increases; the marginal return to household production and care work increases (even if such work is usually unpaid). The second equation seems to be a labour demand equation. We might expect c1 to be negative; the more number of hours women are offering in the labour market, the lower the wage needed for firms to be willing to provide that many hours of employment. We might also expect c2 to be negative if employers discriminate against female workers with children.

In much the same reason as before, a 'naïve' OLS estimate of hours on wage and kids leads to

biased and inconsistent estimates of b1. Again, this can be seen by substituting (1) into (2):

$$\begin{aligned} wage &= c0 + c1 \cdot (b0 + b1 \cdot wage + b2 \cdot kids + u) + c2 \cdot kids + v \\ &= \frac{1}{1 - b1 \cdot c1}[c0 + c1 \cdot b0 + (c1 \cdot b2 + c2) \cdot kids + c1 \cdot u + v] \\ \text{Cov}[wage, u] &= \frac{c1}{1 - b1 \cdot c1} \cdot \sigma_u^2 \neq 0 \end{aligned}$$

Wages are correlated with the error term in (1), and we know by now that this leads to biased and inconsistent estimates of b1.

The expression for wages in terms of kids, u, and v has already been derived above. To get the expression for hours in terms of kids, u, and v, we substitute (2) into (1) instead:

$$\begin{aligned} hours &= b0 + b1 \cdot (c0 + c1 \cdot hours + c2 \cdot kids + v) + b2 \cdot kids + u \\ &= \frac{1}{1 - b1 \cdot c1}[b0 + b1 \cdot c0 + (b1 \cdot c2 + b2) \cdot kids + b1 \cdot v + u] \end{aligned}$$

One of the Gauss-Markov assumptions is that $\text{E}[u|x_i] = E[u] = 0$, where x_i is a regressor and u is the error term. With $\text{Cov}[wage, u] \neq 0$, this is not satisfied except in the uninteresting case where c1 is 0.

Just like before, using "kids" as an IV to identify equation (1) requires that b2 = 0 and c2 $\neq$ 0. Also, $\text{Cov}(\text{kids}, u)$ should be zero. The same thing follows: we perform a 2SLS regression with equation (2) as the first stage and equation (1) as the second stage.