Problem Set 5

This week in lecture we discussed instrumental variables (IV), their uses, and issues that may arise. This problem set considers the theory underlying IV estimation and in a particular application: the relationship between institutions and growth. While it is not required, the following paper may be helpful as you consider the questions in this exercise: Acemoglu, Johnson and Robinson (2001) "Colonial Origins of Comparative Development: An Empirical Investigation" American Economic Review vol 91 no. 5

Question 1 - IV Estimates Common Pitfalls

In class we discussed the system of equation:

$$s_i = \pi z_i + \epsilon_i \quad (1)$$
$$y_i = \rho s_i + \eta_i \quad (2)$$

A) **Question:** *Suppose we were estimating equations (1) and (2) in two steps. That is you estimate equation (1), obtain the fitted values of s_i, substitute those into equation (2) and estimate equation (2). Could you use the second stage residual to estimate your standard errors?*

Answer: You run the **first stage** regression

$$s_i = \pi z_i + \epsilon_i$$

and obtain the fitted values $\hat{s}_i$. The true **second stage** equation is then[1]

$$y_i = \rho \hat{s}_i + \underbrace{\eta_i + \rho(s_i - \hat{s}_i)}_{\nu_i}$$

But if you just run in Stata a simple OLS of y_i on x_i' and $\hat{s}_i$ this fails to realize the structure of the error ν_i and would just calculate $s^2(X'X)^{-1}$. This would be the correct formula under **A4GM** but since here **A4Ω** is correct this formula is wrong.

B) **Question:** *As it was presented in class, the instrument, z_i, must have two properties $Cov(z_i, s_i) \neq 0$ and $Cov(z_i, \eta_i) = 0$. Suppose that equation (2) was modified to include the covariates X so that in the second stage of 2SLS you estimate $y_i = x_i'\alpha + \rho s_i + \eta_i$ but in the first stage still estimate equation (1). Would your 2SLS estimate be consistent? Please explain/show why.*

Answer: Suppose you run the **correct first stage** regression

$$s_i = x_i'\delta + z_i\pi + \epsilon_i.$$

Then *by construction* the OLS residuals $(s_i - \hat{s}_i)$ will be uncorrelated from x_i' and z_i. Hence in the second stage x_i' from the error term. This is because *by construction* it is uncorrelated from $(s_i - \hat{s}_i)$ and *by assumption* it is uncorrelated from η_i (otherwise we would need an instrument for x_i, too).

But suppose **you forget to include all** x_i' in the first stage regression and run instead

$$s_i = w_i'\delta + \pi z_i + \epsilon_i$$

where w_i' is a subset of the covariates x_i'. The $(s_i - \hat{s}_i)$ from this is still uncorrelated from w_i' and z_i, **but most likely not from the remaining variables in** x_i'. Hence in the second stage

[1] Note that by construction s_i is uncorrelated from $s_i - \hat{s}_i$ and by assumption s_i and hence $\hat{s}_i$ are uncorrelated from η_i. Hence A3Rsru is satisfied in this equation!

$$y_i = x_i'\gamma + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i)]$$

x_i' is likely correlated with $(s_i - \hat{s}_i)$ and hence **A3Rsru** does not hold.

Question 2 - Bias in 2SLS and Weak Instruments

We saw that the IV estimator is only consistent, but not unbiased. A condition for this was that our instrument z_i - to actually 'shock' the endogenous variable s_i - needs to be correlated with it. We will now analyze how the bias of the IV estimator depends on how big this correlation is. ***Let us derive a formulation of the bias of IV which we can interpret.***
Forget about the x_i' for this question and suppose we have a matrix of instruments Z. Then the 2SLS estimator is

$$\hat{\rho}_{2SLS} = \rho + (s'P^Z s)^{-1} s'P^z \eta$$

where $P^Z = Z(Z'Z)^{-1}Z$ and its bias is

$$E[\hat{\rho}_{2SLS} - \rho] = E[(s'P^Z s)^{-1} s'P^Z \eta].$$

Substituting the first stage relation $s = Z\pi + \epsilon$ we get[2]

$$\begin{aligned} E[\hat{\rho}_{2SLS} - \rho] &= E[(s'P^Z s)^{-1}[Z\pi + \epsilon]'P^Z \eta] \\ &= E[(s'P^Z s)^{-1}(\pi' Z'\eta)] + E[(s'P^Z s)^{-1}(\epsilon' P^Z \eta)] \end{aligned}$$

From here it is hard to go on since in general $E[a \cdot b] \neq E[a] \cdot E[b]$. But in this special case there is a fairly complicated proof that we can rewrite this expression approximately as

$$E[\hat{\rho}_{2SLS} - \rho] \approx (E[s'P^Z s])^{-1} E[\pi' Z'\eta] + (E[s'P^Z s])^{-1} E[\epsilon' P^Z \eta]$$

A) **Question:** *What property of the instrument Z allows us to simplifies this to:*

$$E[\hat{\rho}_{2SLS} - \rho] \approx (E[\pi' Z' Z\pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta]$$

Answer: By condition 2 for a valid instrument $E[\pi' Z'\eta] = 0$ and if specified the first stage correctly $E[\pi' Z'\epsilon] = 0$. Then

$$\begin{aligned} E[\hat{\rho}_{2SLS} - \rho] &\approx (E[s'P^Z s])^{-1} E[\epsilon' P^Z \eta] \\ &= (E[(\pi' Z' + \epsilon')P^Z (Z\pi + \epsilon)])^{-1} E[\epsilon' P^z \eta] \\ &= (E[\pi' Z' Z\pi + \pi' Z'\epsilon + \epsilon' Z\pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta] \\ &= (E[\pi' Z' Z\pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta] \end{aligned}$$

B) **Question:** *In class you will be given an outline of how the expression in A can be approximately rewritten as*

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z\pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1} \quad (3)$$

where $Q = rank(P^Z)$ and $\sigma_\epsilon^2 = Var(\epsilon|Z)$ and $\sigma_{\eta\epsilon} = Cov(\eta, \epsilon)|Z)$.

[2]Using that $Z'P^Z = (P^Z Z)' = Z'$.

Outline: Using Vassilis' usual trace-trick[3] you can show that

$$E[\epsilon' P^Z \epsilon] = \sigma_\epsilon^2 Q$$

and

$$E[\epsilon' P^Z \eta] = \sigma_{\eta\epsilon} Q$$

and hence

$$\begin{aligned} E[\hat{\rho}_{2SLS} - \rho] &\approx (E[\pi' Z' Z \pi + \epsilon' P^Z \epsilon])^{-1} E[\epsilon' P^Z \eta] \\ &= (E[\pi' Z' Z \pi] + \sigma_\epsilon^2 Q)^{-1} \sigma_{\eta\epsilon} Q \\ &= \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1} \end{aligned}$$

Question: *How does the expression of the 2SLS estimate in part B help you understand why there is bias in the 2SLS estimate? Why is this less of a worry if the correlation between the instruments Z and the variable of interest s is large? [HINT: Think about how to interpret the term $\sigma_{\eta\epsilon}$.]*

Answer: Now we see what creates the bias in 2SLS: $\sigma_{\eta\epsilon}$. Intuitively, since $\hat{s}_i$ is *estimated* it will be fitted towards very high and low errors ϵ_i. But if these are correlated with η_i, then $\hat{s}_i$ is still correlated with η_i.

Secondly, we will see how 'strong' instruments help. We can realize that $\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2}$ is the population explained sum of squares of the **first stage** over the population error sum of squares of the first stage, so it is the **population F-statistic of the first stage**.[4] Hence

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \left[\frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2} + 1 \right]^{-1} = \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \frac{1}{F+1}$$

and hence as $F \to \infty$, bias $\to 0$. So when the instruments are jointly highly significant in the first stage , which is what we call 'strong' instruments, the bias vanishes.

C) **Question:** *Defining $F = \frac{E[\pi' Z' Z \pi]/Q}{\sigma_\epsilon^2}$, how does the expression in part B compare to the bias in the OLS estimate if we estimate equation (2) directly? What happens as π gets close to zero? Note that F is the F-statistic for joint significance of the first-stage regression. [HINT: It will be helpful to think of the omitted variable in the OLS estimate as ϵ, the component of the first stage uncorrelated with the instrument.]*

Answer: Remember that the bias in the OLS estimate was $\frac{\sigma_{\eta s}}{\sigma_s^2}$. From the formula derived we see that

$$E[\hat{\rho}_{2SLS} - \rho] \approx \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2} \frac{1}{F+1} \to \frac{\sigma_{\eta\epsilon}}{\sigma_\epsilon^2}, \text{ as } F \to 0$$

And how will F be 0? By the instrument having no influence on s_i, or $\pi = 0$. But then $s_i = \epsilon_i$ and hence $\sigma_s^2 = \sigma_\epsilon^2$. Thus as the instrument becomes weaker, the IV bias approaches the OLS bias. (Intuition: The instrument doesn't help at all.)

[3] To begin note that $\epsilon' P^Z \eta$ is a scalar and therefore equal to its trace. Also remember that $trace(P^Z) = rank(PZ) = Q$ because P^Z is idempotent. Then $E[\epsilon' P^Z \eta | Z] = E[trace(\epsilon' P^Z \eta)|Z] = E[trace(P^Z \eta \epsilon')|Z] = trace(P^Z E[\eta\epsilon'|Z]) = trace(P^Z \sigma_{\eta\epsilon} I) = \sigma_{\eta\epsilon} trace(P^Z)$. This works similarly for the other expression.

[4] The actual F-Stat would be $[\hat{\pi}' Z' Z \hat{\pi}/Q] \cdot (1/\hat{\sigma}_\epsilon^2)$. This is the population analog to which the sample F-statistic will tend if the sample gets very big.

D) **Question:** *Use your results from part D to explain why the F-statistic is useful in testing if your instrument is sufficiently "strong"?*

Answer: If we have a small sample and want to see whether our instrument(s) are strong, calculating the F-Statistic of the excluded instruments from the first stage is informative.

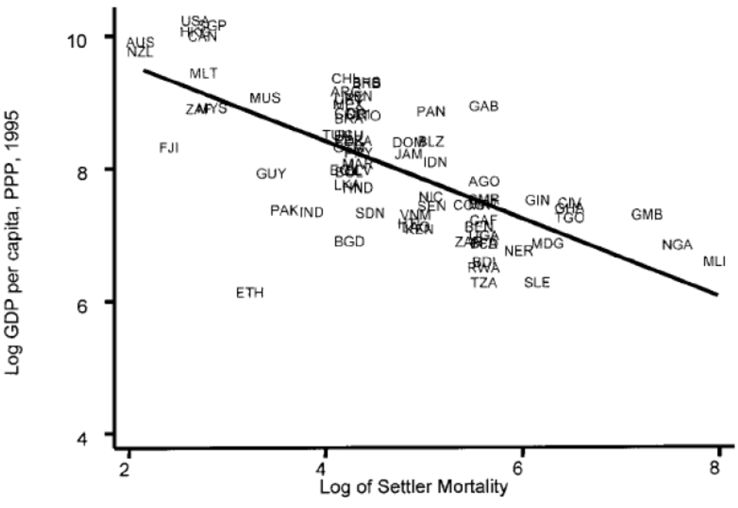
Question 3 - Application: Colonial Origins of Development

Since long time economists believe all kind of things to be important for development, e.g. schooling policy, health policy or macro-economic policies. Only relatively recently the debate focused largely on ***institutions****, e.g. the protection of property rights and many more. The empirical question is: Are institutions really important for economic growth and if so, how much?*
Unfortunately, this is one of the harder questions around, since institutions normally do not fall from the sky but are themselves the consequence of or determined jointly with economic outcomes.
Acemoglu, Johnson and Robinson's paper became famous (in our world: very famous) for coming up with an instrumental variable: Settler Mortality. So they estimate the system of equations

$$\begin{aligned} log(y_i) &= \mu + \alpha R_i + x_i'\gamma + \epsilon_i \\ R_i &= \zeta + \beta log(M_i) + x_i'\delta + \nu_i \end{aligned}$$

where y_i is income per capita, R_i is modern day property rights, X_i is a vector of covariates and M_i is early settler mortality. For this to be a valid instrument, they need to argue that settler mortality indeed is correlated with modern day institutions and that the exclusionary restriction holds.

A) **Question:** *They present the following graph*



Does this prove that settler mortality is a valid instrument?

Answer: No. It only shows what is called the 'reduced form' - and proves that mortality is correlated with GDP. A valid instrument has two properties: Correlated with the variable of interest and uncorrelated with unobservable factors. The variable of interest is modern day institutions. To be valid, early settler mortality must be correlated with modern day property rights. The argument in the paper is that this occurs because Europeans invested in good institutions only in areas where they intended to stay - which where areas with low mortality. The exclusion restriction requires that early settler mortality is uncorrelated with any other determinants of log GDP.

B) **Question:** *Some argue that diseases prevalent in Africa and Latin America, such as malaria, are important causes of poverty and low GDP. Why might this be a problem for the instrument in equation (4)? How could the authors solve this?*

Answer: If colonial malaria prevalence is correlated both with settler mortality and modern malaria prevalence, which in turn might influence modern day growth, then the instrument M is correlated with ϵ_i and does not satisfy the exclusion restriction. Note: the exclusionary restriction requires that $Cov(M_i, \epsilon_i) = 0$, not that $Cov(M_i, x_i) = 0$! So if the authors included a control for modern prevalence of malaria this would solve the problem. Cause then modern day malaria prevalence is no longer part of the error term (which captures everything that influences y_i we do not control for) and thus returning $Cov(M, \epsilon) = 0$.

C) **Question:** *The authors find that the 2SLS estimates are larger than the OLS estimates. Is this what you would have expected? What could explain this?*

Answer: We would have expected the OLS estimate to be upward biased because we expect unobserved factors which improve institutions and improve growth rates. The lower estimates suggest either (1) an omitted variable that is negatively correlated with either growth or institutions or (2) substantial measurement error.

D) **Question:** *A recent criticism of the Acemoglu Johnson Robinson paper suggests that much of the correlation between early settler mortality and later institutions comes from soldiers deaths during military campaigns. When you adjust for this, there is not a robust correlation between early settler mortality and current property rights. If this is correct, why might this be a problem for the paper? Should the authors adjust for soldiers deaths?*

Answer: So the criticism says that the variable labeled 'settler mortality' in the AJR paper in fact should be called 'settler and soldier mortality'. But if this was done then you might have doubts about the instrument, since 'soldier mortality' might be driven by something which as well drives modern day conflicts and hence growth. Then those who wrote the critique say: 'Ok, we can try to tackle this problem by just calculating a correct measure of settler mortality'. But when they do so, the correlation between early settler mortality and modern institutions is close to zero. This runs the risk of a weak instruments problem, which will make the coefficients unstable, biased, and have an ill-defined distribution. Alternatively the authors could again control for soldiers deaths if you believe that the factors that generated higher soldier deaths were relevant in affecting modern day GDP. If these deaths are independent of other correlates of GDP then the authors can use any form of mortality, including soldiers on military campaigns.

Question 4

Discussed in class.