Problem set 1: Instrumental Variables

1. *Suppose you wish to measure the impact of smoking on the weight of newborns. You are planning to use the following model,*

$$\log(bw_i) = \beta_0 + \beta_1 male_i + \beta_2 order_i + \beta_3 y_i + \beta_4 cig_i + \epsilon_i$$

where bw is the birth weight, male is a dummy variable assuming the value 1 if the baby is a boy or 0 otherwise, order is the birth order of the child, y is the log income of the family, cig is the amount of cigarettes per day smoked during pregnancy, i indexes the observation and the β's are the unknown parameters.

   (a) *What could be the problem in using OLS to estimate the above model?*

   Mothers that smoke during pregnancy might as well be less careful with other health issues that affect birth weight and that are not controlled for in the regression equation.

   (b) *Suppose you have data on the average price of cigarettes in the state of residence. Would this information help to identify the true parameters of the model?*

   If people choose the state of residence independently from the price of cigarettes, which seems a sensible assumption, then the price of cigarettes might be uncorrelated with birth weight through ways other than the amount of cigarettes smoked. In this case, the IV identification assumption, that the instrument and the error term are uncorrelated, will hold.

   However, we may as well suspect that smokers are more sensitive to health problems that may require treatment and may affect birth weight. If this is true, and price affects consumption, then the exclusion restriction ceases to hold.

   As for the rank condition, it depends on whether the instrument (price) has enough variation across states to actually affect consumption levels.

   (c) *Use data on BirthWeight.raw to estimate the model above. Use OLS and 2SLS. Discuss the results.*

   Stata commands:

```
use BirthWeight.dta
sum
```

```
gen lbw=ln( bw)
gen ly=ln( y)
regress lbw cig male order ly
ivreg lbw (cig=cigprice) male order ly
```

The OLS results show that the amount of cigarettes smoked while pregnant significantly reduce birth weight by 8%. However, the IV estimates show a non-significant impact of the amount of cigarettes on birth weight. If the IV estimates are correct, all impact seems to come from how smoking is related to other health issues or behaviour that affects birth weight, not through smoking directly.

However, the IV estimates show a worrying feature: all coefficients become insignificant, not only the one on the amount of cigarettes smoked while pregnant. We know that IV estimates have generally a higher variance than the OLS ones, but it may become a problem if the instruments are only weakly correlated with the endogenous explanatory variables. If this is so, the IV estimator becomes inconsistent and its variance becomes very large.

(d) *Estimate the reduced form for cig. Discuss.*

Stata command:

```
regress cig cigprice male order ly
```

What we predicted in the last question turns out to be true. The reduced for model for *cig* shows that the instrument *cigprice* is weak, not explaining the endogenous variables.

2. *Consider the model of earnings,*

$$\ln y_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \epsilon_i$$

where y is hourly earnings, age is age, educ is highest qualification obtained, ϵ is the unobservable component of the model, i indexes the observation and the β's are the unknown parameters.

(a) *What could be the problem in using OLS to estimate the above model?*

The explanatory variable *educ* may be endogenous through its relation with unobserved ability which simultaneously determines educational achievement and earnings.

(b) *Suppose you have data on two additional variables, distance to nearest school (dschool) and distance to nearest college (dcollege) at 16 years of age. Discuss whether these are likely to be good instruments.*

Distance to school and to college affect the cost of investing in education but we do not expect them to affect the returns to education unless parents choose where to live in response to their aspirations to their children and their children ability (which is conceivable but may be negligible). If the effects on the cost of education are sufficiently large, we expect the distance variables to affect education decisions.

However, if we expect the returns to education to vary across individuals, possibly depending on their unobserved ability, the IV procedure will not identify the (average) impact of education any longer.

(c) *Suppose you use dschool and dcollege as instruments in the estimation of the above model. Write down the reduced form for educ and state the conditions under which the parameters of the model above are identified.*

The reduced form model is

$$educ_i = \gamma_0 + \gamma_1 age_i + \gamma_2 dschool_i + \gamma_3 dcollege_i + \nu_i$$

Let $z_i = (1, age_i, dschool_i, dcollege_i)$. The identification conditions are:

- $E(z_i'\epsilon) = 0$;
- $\text{rank}\,(E(z_i'x_i)) = 3$;
- $E(z_i'z_i)$ is pd.

(d) *To test that dschool and dcollege are in fact uncorrelated with ϵ it was suggested to use OLS on the equation,*

$$\ln y_i = \beta_0 + \beta_1 age_i + \beta_2 educ_i + \alpha_1 dschool_i + \alpha_2 dcollege + \mu_i$$

and test $\alpha_1 = \alpha_2 = 0$. Would this method work? Why?

No, this method would not work because *educ* is not exogenous in the above equation and this problem affects the consistency of the estimator of all the parameters, not only that of *educ*.

(e) *How can the assumption rank$(Z'X) = k$ be tested?*

This assumption can be tested using the reduced form for *educ* above and testing whether $\gamma_2 = \gamma_3 = 0$.

3. *Consider the following model,*

$$y = z_1\beta + w\alpha + \epsilon$$

where $E(z\epsilon) = 0$ and $z = (z_1, z_2)$ is vector of exogenous variables. The variable w is endogenous: $E(w\epsilon) \neq 0$.

Suppose we use the following procedure to estimate (β, α):

Step 1: *Regress w on z_2 and obtain the fitted values, $\widehat{w}$.*

Step 2: *Regress y on $(z_1, \widehat{w})$ and obtain $(\widehat{\beta}, \widehat{\alpha})$.*

(a) *Will $(\widehat{\beta}, \widehat{\alpha})$ be generally consistent? Show.*

Suppose we use the reduced form for w

$$w = z_2\gamma + \nu$$

and estimate γ through OLS. Then $E\left(z_2\nu\left(\widehat{\gamma}\right)\right) = 0$ by construction but it is possible that $E\left(z_1\nu\left(\widehat{\gamma}\right)\right) \neq 0$.

If the inequality holds, then the estimates will be inconsistent. To see why, notice that

$$\begin{aligned} y &= z_1\beta + w\alpha + \epsilon \\ &= z_1\beta + \left(z_2\widehat{\gamma} + \widehat{\nu}\right)\alpha + \epsilon \\ &= z_1\beta + \widehat{w}\alpha + \epsilon + \widehat{\nu}\alpha \\ &= z_1\beta + \widehat{w}\alpha + v \end{aligned}$$

where $\widehat{\nu} = w - \widehat{w}$ and $\widehat{w} = z_2\widehat{\gamma}$. The problem is that $E\left(z_1'\nu\right) \neq 0$ and thus $E\left(z_1'v\right) \neq 0$, which leads to inconsistency.

(b) *When will $(\widehat{\beta}, \widehat{\alpha})$ be consistent?*

If $E\left(w|z_1, z_2\right) = E\left(w|z_2\right)$, implying that $E\left(z_1'\nu\right) = 0$, then the estimates will be consistent.

4. *Consider the regression model*

$$y = \beta_0 + \beta_1 x + \epsilon$$

where x is endogenous and z is a binary instrument for x.

(a) *Show that the IV estimator for β_1 is,*

$$\beta_1^{IV} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0}$$

where $\overline{y}_d$ and $\overline{x}_d$ and the averages of y and x, respectively, when $z = d$.

Start by centering the model,

$$\begin{aligned} \widetilde{y}_i = y_i - \overline{y} &= \beta_1 (x_i - \overline{x}) + u_i \\ &= \beta_1 \widetilde{x}_i + u_i \end{aligned}$$

Define $I_1 = \{i : z_i = 1\}$. Then we have,

$$\begin{aligned} \beta_1^{IV} &= (z'\widetilde{x})^{-1} z'\widetilde{y} \\ &= \frac{\sum_{i \in I_1} (y_i - \overline{y})}{\sum_{i \in I_1} (x_i - \overline{x})} \\ &= \frac{\overline{y}_1 - \overline{y}}{\overline{x}_1 - \overline{x}} \\ &= \frac{\overline{y}_1 - \left(\frac{N_1}{N}\overline{y}_1 - \frac{N_0}{N}\overline{y}_0\right)}{\overline{x}_1 - \left(\frac{N_1}{N}\overline{x}_1 - \frac{N_0}{N}\overline{x}_0\right)} \\ &= \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0} \end{aligned}$$

where N, N_0 and N_1 are the sizes of the whole sample and the sub-samples with $z = 0$ and $z = 1$, respectively.

(b) *What is the interpretation of β_1^{IV} if x is also binary (say, if it represents participation in treatment)?*

In this case β_1^{IV} represents the estimated average impact of treatment on the individuals that change their treatment status in response to a change in z (form 0 to 1 or conversely from 1 to 0).

5. *Suppose you wish to estimate β in*

$$y_i = \alpha + x_i\beta + u_i$$

(a) *Derive the consequences for the OLS estimator of y being measured with error which is independent of x.*

The OLS estimator of β is,

$$\beta^{OLS} = \frac{\text{cov}(x_i, y_i)}{\text{var}(x_i)}$$

If y is measured with error, so that we observe y^* where

$$y_t^* = y_i + \epsilon_i$$

where x is independent of ϵ. Then the OLS estimator will be,

$$\begin{aligned}\beta^{OLS} &= \frac{\text{cov}(x_i, y_i^*)}{\text{var}(x_i)} \\ &= \frac{\text{cov}(x_i, \alpha + \beta x_i + u_i + \epsilon_i)}{\text{var}(x_i)} \\ &= \beta\frac{\text{var}(x_i)}{\text{var}(x_i)} + \frac{\text{cov}(x_i, u_i)}{\text{var}(x_i)}\end{aligned}$$

That is, for as long as the OLS estimator is consistent for the estimation of β in the absence of measurement error in y, it will be consistent with measurement error in y if we can guarantee that the measurement error is independent of x.

(b) *Instead of measuring x you measure x^* where $x_i^* = x_i + \epsilon_i$ and ϵ_i is a measurement error which is independent across individuals and independent of x. Show that the OLS estimator converges asymptotically to $\delta\beta$ where $0 \leq \delta \leq 1$. Explain the implication of this result for estimating the elasticity of hours worked with respect to wages when wages are measured with iid errors.*

If x is measured with error, so that we observe x^* where

$$x_i^* = x_i + \epsilon_i$$

where x is independent of ϵ.

The data model is,

$$\begin{aligned} y_i &= \alpha + \beta x_i^* + (u_i - \beta\epsilon_i) \\ &= \alpha + \beta x_i^* + v_i \end{aligned}$$

Then the OLS estimator

$$\begin{aligned} \beta^{OLS} &= \frac{\text{cov}(x_i^*, y_i)}{\text{var}(x_i^*)} \\ &= \frac{\text{cov}(x_i^*, \alpha + \beta x_i^* + (u_i - \beta\epsilon_i))}{\text{var}(x_i^*)} \\ &= \frac{\beta\text{var}(x_i^*) - \beta\text{cov}(x_i^*, \epsilon_i)}{\text{var}(x_i^*)} \\ &= \beta\left(1 - \frac{\text{cov}(x_i + \epsilon_i, \epsilon_i)}{\text{var}(x_i + \epsilon_i)}\right) \\ &= \beta\left(1 - \frac{\text{var}(\epsilon_i)}{\text{var}(x_i) + \text{var}(\epsilon_i)}\right) \\ &= \beta\left(1 - \frac{\sigma_\epsilon^2}{\sigma_x^2 + \sigma_\epsilon^2}\right) \end{aligned}$$

Thus, OLS will be downward biased - attenuation bias: the elasticity of hours worked with respect to wages will be under-estimated.

6. *Show that* $W = E\left[u_i^2 z_i' z_i\right]^{-1}$ *is the optimal weighting matrix for the GMM estimator.*

Let $\Sigma = E\left[u_i^2 z_i' z_i\right]$. The general form of the variance of the GMM estimator is,

$$\Omega = \left(M_{xz} W M_{xz}'\right)^{-1} M_{xz} W \Sigma W M_{xz}' \left(M_{xz} W M_{xz}'\right)^{-1}$$

while if $W = \Sigma^{-1}$ the variance is,

$$\Omega^* = \left(M_{xz} \Sigma^{-1} M_{xz}'\right)^{-1}$$

The optimal weighting matrix is the one that minimises the variance of the GMM estimator. If $W = \Sigma^{-1}$ is the optimal weighting matrix, then it must be that $\Omega - \Omega^*$ is positive semi-definite. This implies that $\Omega^{*^{-1}} - \Omega^{*^{-1}}$ is positive semi-definite for any alternative choice of W where

$$\Omega^{*^{-1}} - \Omega^{*^{-1}} = M_{xz}\Sigma^{-1}M_{xz}' - M_{xz}WM_{xz}'\left(M_{xz}W\Sigma WM_{xz}'\right)^{-1}M_{xz}WM_{xz}'$$

We can write,

$$\begin{aligned}
\Omega^{*^{-1}} - \Omega^{*^{-1}} &= M_{xz}\left[\Sigma^{-1} - WM'_{xz}\left(M_{xz}W\Sigma WM'_{xz}\right)^{-1}M_{xz}W\right]M'_{xz} \\
&= M_{xz}\Sigma^{-1/2}\left[I - \Sigma^{1/2}WM'_{xz}\left(M_{xz}W\Sigma^{1/2}\Sigma^{1/2}WM'_{xz}\right)^{-1}M_{xz}W\Sigma^{1/2}\right]\Sigma^{-1/2}M'_{xz} \\
&= M_{xz}\Sigma^{-1/2}\left[I - D\left(D'D\right)^{-1}D'\right]\Sigma^{-1/2}M'_{xz}
\end{aligned}$$

where $D = \Sigma^{1/2}WM'_{xz}$.

The matrix $\left[I - D\left(D'D\right)^{-1}D\right]$ is idempotent and symmetric, and therefore positive semi-definite. But then, the above expression defines a positive semi-definite matrix since it is a quadratic form with a positive semi-definite matrix.