Empirical Example: Instrumental Variable Estimation

One important topic in labor economics is how fertility affects woman labor supply. In particular we want to know how much a woman’s labor supply falls when she has an additional child. The data file labor.cvs contains a random sample of 30,000 women taken from 1980 US census. The description of the variables is below

- morekids: =1 if mom had more than 2 children
- samesex: = 1 if the first two children have same sex
- age: age of mom at census
- hispan: =1 if mom is Hispanic
- supply: mom’s weeks worked in 1979

First we can compare the average working weeks for women with fewer than 2 children (group 0) to women with more than two children (group 1). The stata commands below produce the same results:

```stata
sort morekids; 
ttest supply, by(morekids); 
reg supply morekids;
```

So on average, a woman with more than two kids works 6 weeks fewer than her counterpart with fewer than 2 kids. The result changes a little after we include age as the regressor. The question is, does this correlation imply causality?

So the key regressor is morekids. The main concern is that morekids may be endogenous, i.e., correlated with the error term, due to simultaneity or omitted variable. After controlling for age, the coefficient of morekids is -6.6 and significant at 5% level. The negative sign is expected. The number -6.6 means having more than 2 kids on average causes a woman to work 6.6 weeks (one and half month) fewer in a year than a woman with same age, but with fewer than 2 kids. Given that one year has only 48 weeks, this is a big effect.

Some people may argue that the true effect may not be that big. For instance, there is an unobserved factor $u$ of preferring staying home over working. Women with that preference tend to have more kids (so $E(u \cdot morekids) > 0$), and that preference reduces labor supply (so the coefficient of $u$ in the true model could be negative). This omitted variable results...
in that the OLS overestimates (in absolute value) the true causal effect, see Table 3.2 in the textbook.

Some people mention reverse causality, which may or may not exist. Reverse causality means that fewer labor supply causes higher fertility. Can you build a structural simultaneous model, similar to the demand-and-supply model, to show this idea? Hopefully the structural model can suggest candidates for IV.

Regardless of the reason, we can use IV to resolve the potential endogeneity issue. First we try samesex as the IV. The stata command

\texttt{ivreg supply (morekids = samesex) age, r;}

yields a new estimate of -5.46, which is insignificant at 5% level and smaller (in absolute value) than OLS estimate.

It’s instructive to compare the IV and OLS estimates, and standard errors. The OLS standard error of morekids is 0.25, much smaller than the IV standard error 3.68. This fact is typical, confirming the trade-off between bias and efficiency. The OLS estimator is precise, and maybe is biased. The IV estimator is imprecise, and maybe is unbiased. Put differently, you may get significant but biased result if using OLS. You may get insignificant but unbiased result if using (relevant) IV. The OLS is precise because it is based on the variation in \( x \); or \( E(x^2) \): The IV is imprecise because it is based on covariance \( E(\mathbf{z}x) \). Weak correlation between \( \mathbf{z} \) and \( x \) can give rise to high degree of impreciseness. The commands below

\texttt{cor morekids samesex, co;}
\texttt{dis 1/r(Var_1);}
\texttt{dis r(Var_2)/r(cov_12)^2;}

explain why IV standard error is much bigger than OLS for this problem.

Because of the trade-off between bias and efficiency, we need a formal test for the exogeneity of \( x \). The null hypothesis is that \( x \) is exogenous, i.e.,

\[ H_0: E(xu) = 0, \quad (\text{Wu-Hausman Test}). \]

If the null is true, we better use OLS other than IV, since OLS is more precise. If the null is false, we have to use IV provided that the IV is valid. Testing exogeneity is tricky since \( u \) is unobserved. Nevertheless, in two important papers of Wu (1973, Econometrica), Hausman(1978, Econometrica) the authors provide a brilliant solution. Under the alternative, \( x \)
is endogenous, so IV and OLS should differ significantly since IV is consistent and OLS is inconsistent. Under the null \( x \) is exogenous, and both IV and OLS are consistent, so should differ little. Therefore by comparing the IV and OLS estimate we can test exogeneity of \( x \). Keep in mind, the Wu-Hausman test implication assumes the IV satisfies the exclusion restriction, so is really exogenous. If IV is endogenous, then you are comparing two inconsistent estimators, which is meaningless.

We can conduct the Wu-Hausman test in two steps: first, regress \( x \) onto \( z \) and other exogenous regressor, and keep the residual \( \hat{v} \). (Note this is first-stage regression, or reduced form for \( x \)). Next, regress \( y \) onto \( x \), other exogenous regressors, and \( \hat{v} \). We reject \( H_0 : E(xu) = 0 \) if the coefficient of \( \hat{v} \) is significant at 5% level. For this problem the commands are

```
qui reg morekids samesex age;
predict vhat, res;
qui reg supply morekids age vhat, r;
test vhat;
```

we cannot reject null given the p-value 0.7538. That means morekids is exogenous and OLS estimator is consistent.

In practice you stop if Wu-Hausman test shows \( x \) is exogenous. Here, for teaching purpose we continue pretending that \( x \) is endogenous and IV is necessary. We need to defend that samesex is a good IV, satisfying both exclusion and inclusion restrictions. The inclusion restriction \( E(zx) \neq 0 \) is always testable. We run the first-stage regression in which the endogenous regressor morekids is dependent variable and IV and exogenous regressor age are independent variables, and test the significance of samesex:

```
qui reg morekids samesex age;
test samesex;
```

The reported F test is 148.11, greater than 10 (the rule of thumb), so samesex is correlated with morekids, and samesex satisfies the inclusion restriction (relevance requirement). This finding is intuitive: couple with first two kids of same sex tend to have at least one more child due to preference of diversity in gender.

We cannot statistically test the exogeneity of \( z \) since the model is just identified. i.e., the number of instruments equals the number of endogenous regressors. We can test the exogeneity of \( z \) if the number of instruments is greater than the number of endogenous regressors. In
that case the model is over-identified, and the test for exogeneity of \( z \) essentially compares one IV estimate to the other. Big difference implies that at least one IV is endogenous.

Here, intuition tells us same sex should be exogenous, since it is virtually randomly assigned, not self-selected by the human being. Moreover, same sex should be excluded from the true model since it has no direct effect on labor supply.

Next I want to explain where the name 2SLS comes from. The following commands save the predicted value of the first stage regression

\[
\texttt{qui reg morekids same sex age; predict morekidshat, xb;}
\]

Then in the second stage we replace the endogenous regressor with its fitted value.

\[
\texttt{reg supply morekidshat age, r;}
\]

It cannot be emphasized enough that, though the coefficient estimate is correct (and the same as that reported by \texttt{ivreg}), the reported standard error, t-value and p-value are all wrong since they do not account for the fact that a fitted value is used as regressor. So my advice is to running the two stage regression manually for purposes other than obtaining the final results. Use \texttt{ivreg} to get the final results.

Next we try hispan as IV, which produces an even smaller estimate of -3.7. The difference between the two IV estimates (-3.75 and -5.46) cannot be taken seriously since both are insignificant. The fact that IV same sex has a t-value (-1.48) bigger than IV hispan (-1.22) may suggest same sex is better IV. The F test for hispan in the first-stage regression is 206.69, greater than 10. So hispan is relevant. It is plausible that hispan is exogenous since it is not chosen by human being. We report the Wu-Hausman test with IV hispan. Again the morekids is shown to be exogenous.

Smart guy A may say, gee, age is exogenous and correlated more kids, so why don’t we use it as IV? Smart guy B may say, gee, hispan should not be IV. Instead, we should use it as exogenous regressor and explicitly control for it. Basically they are questioning about model specification. Suppose the true structural model, which has causal interpretation, is

\[
y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x + u
\]

where \( y_2 \) denotes endogenous regressor. For this problem, \( y_1 = \text{supply}, y_2 = \text{morekids} \). The key question is, what is the exogenous regressor \( x \)? Does \( x \) include age, or hispan?
First of all, $x$ should have direct effect on $y_1$. That is why it is included in the model in the first place. We cannot use included exogenous variable alone as IV; otherwise, the fitted value from the first-stage regression, $\hat{y}_2$, would be perfectly correlated with $x$ in the second stage regression. This answers A’s question. Age has direct effect on supply, so is regressor. IV cannot be age alone.

In order to avoid perfect multicollinearity in the second-stage regression, we need at least one exogenous variable which is excluded from the true model. We call that excluded exogenous variable the instrumental variable. So the exogeneity requirement is also called exclusion restriction. The fact that IV variable is excluded from the model does not imply $z$ is uncorrelated with the dependent variable. Actually we hope $z$ and $y_1$ are highly correlated, but only through $y_2$. In other words, $z$ should have indirect, not direct, effect on $y_1$. That is why $z$ is excluded from the true model in the first place.

Now we can answer the second guy’s question. We don’t believe hispan has direct effect on supply. Put differently, we believe the observed correlation between hispan and supply is due to the latent variable of morekids. As another way to put it, after morekids has been controlled for, hispan has no effect on supply. So hispan should be excluded from the true model, and cannot be used as regressor.

We can run the following regression to check that hispan has no direct effect on supply:

```
reg supply hispan morekids age, r;
```

which yields p-value of 0.349 for hispan (unreported in the log file). This fact indicates that hispan should be excluded.

Since both samesax and hispan are good IVs, why not use them all? So next we use both as IV, and run the stata command

```
ivreg supply (morekids = samesex hispan) age, r;
```

The new estimate is -4.46, and significant at 10% level. We note that using two IV leads to reduced standard error.

With two IV and one endogenous regression, now the model becomes over-identified. Then we can test the exogeneity of instruments by conducting the so called over-identification test (or J test). The null hypothesis is that all IV are exogenous. The stata command is

```
ivreg supply (morekids = samesex hispan) age, r;
predict uhat, resid;
* Overidentification (J) test;
```
qui reg uhat samesex hispan age;
dis "over-identification test is " e(N)*e(r2);
dis "p-value of over-identification test is " chiprob(1, e(N)*e(r2));

Note that the degree of freedom for the \(\chi^2\) distribution is the number of instruments minus the number of endogenous regressors. See section 15.5 in textbook for more details about the J test. Basically, the idea for the over-identification test is, under the null hypothesis all instruments should have coefficient zero when dependent variable is 2SLS residual. This is because under the null the instruments are supposed to be uncorrelated with \(u\), which is estimated by 2SLS residual. For this problem, the p-value of the over-identification test is 0.723, so we cannot reject the null. So hispan and samesex are indeed exogenous.

Finally we check the relevance requirement by running the first-stage regression

qui reg morekids samesex hispan age;
test samesex hispan;

The F test is 179.49, greater than 10.

To summarize, morekids is exogenous; samesex and hispan are both valid IVs. The OLS estimate -6.6 is the final result.