James Wacker
R0483229

Application and Limitation of 2SLS

Introduction / the problem of endogeneity

Vietnam war veterans faced many difficulties after returning to civilian life, the effects of which were of frequent study in the years following the war. One example was income, did being a veteran have a negative impact on one's income? Here we illustrate the tools that were available at the time (1970s) for one trying to answer this question.

Initially one might be tempted to run OLS on a linear regression of hourly wage (y) with various regressors, including a binary indicator of whether someone is a veteran:

$$y = bo + b1 * z1 + b2 * z2 + \; b3 * vet + u$$

With hourly wage, *y;* education, *z1*; work experience, *z2*; veteran status, *vet*; and *u*, the error term, i.e. other determinants of hourly wage. The problem with *vet* is one of self-selection – although some people were drafted (i.e. their veteran status is due to randomness), others voluntarily joined the military. Volunteers may have different characteristics from draftees, e.g. perhaps they joined with the goal of getting the most out of their service. Since this non-randomness may correlate with their post-war wage, we would expect for *vet* to correlate with the error term, *u*. With *vet* being endogenous, we expect OLS to produce biased coefficient estimates.

To demonstrate this scenario, we simulate the variables used in the above linear model according to **Table A** in the appendix, with *vet* at least partially dependent on *u* to simulate endogeneity*.* As expected, OLS produces biased coefficient estimates as shown in **Table B**. Note that although the endogenous variable *vet* has the highest bias, the other coefficients (including the intercept) are biased as well, thus we find that one endogenous variable is enough to contaminates all the coefficient estimates.

Two-stage least squares

An interesting way to work around this is to make use of the military draft lottery data. All young men were given draft lottery numbers, but their subsequent selection turned out to *not* be completely random- smaller lottery numbers had a higher chance of being drafted. This means we would observe negative correlation between lottery number, *x1,* and whether someone is a veteran, *vet*, but we would not observe any correlation between *x1* and *u*. Thus, *x1* may serve as a valid instrumental variable for *vet*.

Using the two-stage least squares (2SLS) estimation method, *x1* is used to help estimate *vet*, which is then used in the estimation of salary. As shown in **Table C,** we observe much better coefficient estimates than we did with OLS. Additionally, when we increase the sample size by a factor of 10 (to 10,000) the estimates are nearly perfect, demonstrating that 2SLS is a consistent estimator. We note that the standard errors are a bit higher than with OLS, this is because we are running two simultaneous regressions, and some of the uncertainty (regarding the estimates) from the first regression is carried over to the second.

Importance of a strong instrument

In the above simulation, the correlation between *x1* and *vet* was ~-0.4. Suppose the likelihood that someone was drafted was instead only slightly negatively correlated with their draft number. If a researcher was not aware of this, he might still choose to use two-stage least squares, thinking that he might as well use whatever information is available to try and improve his results. However, when the correlation between the instrument and endogenous variable is low, 2SLS becomes biased and highly inconsistent, to the point where regular OLS becomes the better option.

Here we demonstrate this by changing how *vet* is generated – in **Table A** the coefficient for *x1* was -1, here we lower it to -0.2. The result is that corr(x1,vet) = ~ -0.1.

When repeating the OLS and 2SLS estimations, we obtain the results in **Table D.** Not only is 2SLS more biased than OLS, but the much larger standard errors show how unreliable 2SLS is in this setting. This demonstrates the problem of weak instrumental variables, and how they can be harmful to 2SLS regression. If high standard errors are observed in 2SLS, the instrumental variable(s) are probably not helpful, and the researcher should attempt to find other processes that would be uncorrelated with *u,* but that may have a relation to the endogenous variable.

A side note on the draft lottery numbers

The draft lottery uses its own algorithm that is actually based on the birthdays of the potential draftees. The process of generating the lottery numbers is therefore not entirely random, but if we consider each person's birthday as random, we can then consider the lottery data as random as well.

Appendix:

Table A: Variable and error term simulations

Variable	Description	Distribution	Notes
z1	Education	N(1,1)	Simple distribution used for purpose of the simulation, *x1* for example is very complex in reality, but for the simulation it just needs to consist of random values that are (in part) used to generate *vet*
z2	Work experience	N(1,1)	
x1	Lottery #	N(1,1)	
u	Eq. 1.1 error	N(0,1)	Other determinants of salary
v	Eq. 1.2 error	N(0,1)	Other determinants of joining the military
vet	Veteran status (0 or 1)	B(n,~0.20)	Generated as a linear combination of all the above: *vet* = - 0.75*z1 - 0.75*z2 - 1*x1 + 0.5*u + 0.75*v This results in a unitless score. A cut-off value was then chosen to simulate a realistic proportion of Vietnam veterans. z1 and z2 are negatively correlated since not serving means more time for work/education
y	Response variable: hourly wage	~N(6,3)	Generated as y = 5 + 0.75*z1 + 0.75*z2 - 2*vet + u the intercept is not particularly meaningful, it is just to demonstrate how endogeneity causes bias in the intercept estimate as well. It does help add a bit of realism to the distribution of y (average hourly wage for that time period)

Table B: Results of OLS

Variable	True B	B_hat$_{OLS}$	SE	p-value
Intercept	5	4.65	0.07	0.00
z1	0.75	0.87	0.03	0.00
z2	0.75	0.84	0.03	0.00
vet	-2	-1.42	0.08	0.00

*B_hat$_{OLS}$ is shown for n = 1,000. For n=10,000 it is virtually unaffected

Table C: Results of 2SLS

Variable	True B	B_hat$_{2SLS}$	SE	P-value	B_hat$_{2SLS}$ for n=10,000
Intercept	5	4.84	0.10	0.00	4.96
z1	0.75	0.82	0.04	0.00	0.77
z2	0.75	0.79	0.04	0.00	0.76
vet	-2	-1.87	0.17	0.00	-1.97

Table D: Weak instrument: OLS vs 2SLS results when corr(x1,vet) is small

Variable	True B	B_hat$_{OLS}$	SE$_{OLS}$	B_hat$_{2SLS}$	SE$_{2SLS}$
Intercept	5	4.61	0.07	4.49	0.33
z1	0.75	0.84	0.03	0.92	0.10
z2	0.75	0.86	0.03	0.92	0.09
vet	-2	-1.19	0.09	-0.71	0.75