Introduction to Econometrics
PS # 7 - Instrument Variables

1 Endogeneity: Revisited

1.1 Omitted Variable Bias

The classic case of omitted variable bias occurs when certain variables are hard to measure. Consider the following regression:

$$earnings_i = \beta_0 + \beta_1 educ_i + \beta_2 abil_i + u_i$$

Where $\beta_0 \neq 0, \beta_1, \beta_2 > 0$ and $E[u_i|educ_i, abil_i] = 0$. Ability of individuals is often difficult to precisely measure, suppose we drop it and run a new regression:

$$earnings_i = \alpha_0 + \alpha_1 educ_i + z_i$$

1. Explain why we can expect omitted variable bias in the new regression
2. What is the expected sign of bias?

1.2 Simultaneity

Suppose you have data on health scores and incomes from a random sample of individuals. Both income and health are simultaneously determined, as follows:

$$income_i = \beta_0 + \beta_1 health_i + \epsilon_i$$

$$health_i = \alpha_0 + \alpha_1 income_i + u_i$$

1. Show that this simultaneity breaks the ZCM condition *i.e.*, prove that $E[\epsilon_i|health_i] \neq 0$
2. Explain the intution behind this result [**Hint**: Think of how ϵ_i affects $health_i$]

2 Two-Stage Least Squares: Theory

Suppose we are interested in the causal effect of education on earnings.

$$earnings_i = \alpha_0 + \alpha_1 educ_i + u_i$$

As seen before in $\mathcal{Q}1.1$, education is a potentially endogenous variable due to unobserved factors like ability that affect both education and earnings. If earnings in turn determines education, then we could also have a simultaneity issue. In both situations, the ZCM condition will be violated.

1. Define $dist_i$ as the distance to the university. Argue why $cov(dist_i, educ_i) \neq 0$ and $cov(dist_i, u_i) = 0$
2. Suppose we regress education on the distance to university, as follows:

   $$educ_i = \pi_0 + \pi_1 dist_i + v_i$$

   and get the estimated values: $\widehat{educ}_i = \widehat{\pi}_0 + \widehat{\pi}_1 dist_i$

   (a) Explain why $\widehat{educ}_i$ is uncorrelated with factors in u_i

   (b) What kind of factors affect the residuals? ($\widehat{v_i} = educ_i - \widehat{educ}_i$)

3. Suppose we now regress

   $$earnings_i = \alpha_0 + \alpha_1 \widehat{educ}_i + u_i$$

   Argue why $\widehat{\alpha}_1$ provides a consistent estimate of α_1

3 Two-Stage Least Squares: Application

How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise, you will estimate this effect using data for married women from the 1980 U.S. Census using the fertility.csv dataset.

1. (Use **R**) Regress $weeksm1$ (number of weeks worked per year) on the indicator variable $morekids$ (=1 if mom had more than 2 children), using OLS. On average, do women with more than two children work less than women with two or less children?

$$weeksm1_i = \beta_0 + \beta_1 morekids_i + u_i$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.06843    0.05466   385.4   <2e-16 ***
morekids    -5.38700    0.08861   -60.8   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Do you expect the above regression to suffer from endogeneity?

3. The dataset also contains a variable called $samesex_i$, defined as follows:

$$samesex_i = \begin{cases} 1 & , \text{ first two children are of same sex (boy-boy or girl-girl)} \\ 0 & , \text{ otherwise} \end{cases}$$

   Argue why this variable is a valid instrument for $morekids_i$. (Use **R**) Run the first stage regression:

$$morekids_i = \pi_0 + \pi_1 samesex_i + v_i$$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.346425   0.001365  253.79   <2e-16 ***
samesex     0.067525   0.001920   35.17   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. (Use **R**) Extract the predicted values $\widehat{morekids}_i$ from the first-stage regression and run the second-stage regression

$$weeksm1_i = \beta_0 + \beta_1 \widehat{morekids}_i + u_i$$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   21.4211     0.4904  43.684  < 2e-16 ***
morekids_hat  -6.3137     1.2835  -4.919  8.7e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```