# Quantitative Economics

## Tutorial 5b: Applied Micro: Returns to Schooling

1. The following dataset contains data on wages and workers' characteristics for 2,055 male individuals in the United States. In particular, it contains the following variables:

Table 3.1: Variables in the dataset

| |
|---|
| `lwage`: log(wage) in 1976 |
| `educ`: years of completed schooling in 1976 |
| `age`: in years |
| `black`: 1 if black, 0 if white |
| `south`: 1 if the individual lives in the south of the US in 1976, 0 otherwise |
| `IQ`: IQ score |
| `libcrd14`: 1 if library card in home when the individual was 14, 0 otherwise |

The descriptive statistics corresponding to these variables are:

Table 3.2: Descriptive statistics

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| `lwage` | 2055 | 6.335687 | .417847 | 4.718499 | 7.784889 |
| `educ` | 2,055 | 13.92506 | 2.274757 | 8 | 18 |
| `age` | 2,055 | 28.37616 | 2.998883 | 24 | 34 |
| `black` | 2,055 | .1435523 | .3507205 | 0 | 1 |
| `south` | 2,055 | .3411192 | .4742007 | 0 | 1 |
| `IQ` | 2,055 | 102.4637 | 15.40797 | 50 | 149 |
| `libcrd14` | 2,055 | .7586375 | .4280138 | 0 | 1 |

Table 3.3 reports the results of two OLS regressions of log wages on workers' characteristics. The difference between the first (1) and the second (2) regressions is the inclusion of `IQ` in the latter.

Table 3.3: OLS regressions

| Dependent variable: `lwage` | (1) | (2) |
|---|---|---|
| `age` | 0.043 | 0.044 |
| | (0.003) | (0.003) |
| `black` | −0.151 | −0.112 |
| | (0.025) | (0.027) |
| `south` | −0.109 | −0.102 |
| | (0.018) | (0.018) |
| `educ` | 0.033 | 0.025 |
| | (0.004) | (0.004) |
| `IQ` | – | 0.003 |
| | | (0.001) |
| `constant` | 4.714 | 4.515 |
| | (0.095) | (0.107) |
| $R^2$ | 0.18 | 0.19 |
| Number of observations | 2,055 | 2,055 |

Note: Standard errors in parentheses.

Table 3.4 reports the results of the second stage of a 2SLS regression of log wages on workers' characteristics, where the variable `educ` is instrumented with the variable `libcrd14`.

Table 3.4: 2SLS. 2nd stage regression.

| Dependent variable: `lwage` | |
|---|---|
| `educ` | 0.058 |
| | (0.027) |
| `age` | 0.043 |
| | (0.003) |
| `black` | −0.124 |
| | (0.029) |
| `south` | −0.106 |
| | (0.019) |
| `IQ` | 0.000 |
| | (0.002) |
| `constant` | 4.366 |
| | (0.162) |
| Number of observations | 2,055 |

Note: Standard errors in parentheses.

Table 3.5 reports the results of the first stage of the previous 2SLS regression, that is, the regression of years of completed schooling on workers' characteristics and the instrumental variable `libcrd14`.

Table 3.5: 2SLS. 1st stage regression.
Dependent variable: `educ`

| | |
|---|---|
| `age` | 0.042 |
| | (0.014) |
| `black` | 0.439 |
| | (0.137) |
| `south` | 0.235 |
| | (0.094) |
| `IQ` | 0.077 |
| | (0.003) |
| `libcrd14` | 0.769 |
| | (0.104) |
| `constant` | 4.140 |
| | (0.543) |
| $R^2$ | 0.29 |
| Number of observations | 2,055 |

Note: Standard errors in parentheses.

(a) Interpret the coefficient on `educ` and compute a 99% confidence interval for the coefficient on `educ` in the first (1) OLS regression and interpret it.

(b) Compare the coefficients on `educ` in the two OLS regressions and explain the differences using the omitted variables bias formula. Do you think regression (1) captures a causal relationship between education and wages? How about regression (2)? Explain.

(c) Why do you think is important to instrument `educ`? Compare the coefficient on `educ` in Table 3.3 column (2) with the coefficient on `educ` in Table 3.4. Is there any difference? Explain.

(d) What are the conditions that a valid instrumental variable must satisfy? Looking at Table 3.5, what can you conclude about the validity of `libcrd14` as an instrument for `educ`? Explain.

2. Read the article "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," by Joshua Angrist and Alan Krueger, available here, and answer the following questions:

   (a) What do the authors mean by *identification*? Explain briefly.

   (b) The authors write on p. 69 "If the demand and supply curves shift over time, the observed data on quantities and prices reflect a set of equilibrium points on both curves. Consequently, an ordinary least squares regression of quantities on prices fails to identify–that is, trace out–either the supply or demand relationship." (quotes added) Provide a graphical illustration of the identification problem and show (graphically) how certain "curve shifters" can be used to address the identification problem at stake.

   (c) Even though IV estimates are not always useful, it is always useful to look at the reduced form. True or false? Why?

   (d) Geographical variables typically make excellent instruments because they are clearly and uncontroversially exogenous. True or false? Why?

   (e) The authors write on p. 77 "One difficulty in interpreting instrumental variables estimates is that not every observation's behavior is affected by the instrument." (quotes added) Explain.

3. **Siblings, Education and Earnings.** In the lectures we discussed running regressions with (identical) twins. This question is about running regressions with (non-twin) siblings. Consider the following long regression

$$Y_{i,f} = \alpha^L + \beta^L S_{i,f} + \gamma A_{i,f} + \epsilon^L_{i,f}$$

   where $Y_{i,f}$ is the log of annual earnings of individual $i$ in family $f$, $S_{i,f}$ is the number of completed years of schooling by individual $i$ in family $f$, $A_{i,f}$ is a measure of genetically transmitted ability of individual $i$ within family $f$ and $\epsilon^L_{i,f}$ is a regression residual. Unfortunately, you do *not* observe $A_{i,f}$.

   (a) Suppose that you run a regression of $Y_{i,f}$ on $S_{i,f}$ and let $\beta^S$ be the slope of this short regression. Find an expression for $\beta^S$. Does $\beta^S$ equal $\beta^L$? Explain.

   (b) Suppose that for each family $f$, you have information on *two* siblings $i = \{1, 2\}$, and you run a regression of $\Delta Y_f$ on $\Delta S_f$, where $\Delta Y_f = Y_{1,f} - Y_{2,f}$ and $\Delta S_f = S_{1,f} - S_{2,f}$. Let $\beta^{S'}$ be the slope of this short regression. Find an expression for $\beta^{S'}$. Does $\beta^{S'}$ equal $\beta^L$? Explain.

   (c) Compare $\beta^S$ to $\beta^{S'}$.

   (d) What do you think are the pros and cons of using (non-twin) siblings versus (identical) twins to recover the return to schooling? Explain.

4. **Is Our Financial Future In Our Chromosomes?(Jon Entine, 14 October 2012: Science 2.0)** "Genoeconomics" studies the genetic determinants of socioeconomic outcomes. Roughly speaking, if genes are randomly assigned at birth, one may think of studying the effects of genes on socioeconomic outcomes (e.g., earnings). Suppose that an economist is interested in estimating the impact of a particular gene $G = \{0, 1\}$ on earnings $Y$.

   (a) Does a regression of $Y$ on $G$ have a causal interpretation? Explain using the potential outcomes framework.

(b) How about running a regression of $Y$ on $G$ and $C$, where $C = 1$ denotes having a college degree, and 0 otherwise? Would this longer regression be more likely to provide a causal answer than the shorter one? Why? Explain using the potential outcomes framework.

5. **(Optional) "Monotonicity: With or without you."** The following question allows us to provide an alternative interpretation of the findings in Angrist and Krueger (1991). Let $Q4_i$ be a fourth quarter of birth indicator of individual $i$ (1 if the individual $i$ was born in the fourth quarter, 0 otherwise), $D_i$ be a high-school graduation indicator of individual $i$ (1 if the individual $i$ has at least high-school, 0 otherwise) and $Y_i$ be the log of annual earnings of individual $i$.

   (a) Show that if $Q4_i$ has *heterogeneous* causal effects on high-school graduation $D_i$, that is,

   $$D_i = D_i(0) + (D_i(1) - D_i(0))Q4_i$$

   but the causal effect of high-school graduation on annual earnings is the *same* for all individuals

   $$Y_i(1) = Y_i(0) + \rho$$

   then $LATE = \rho$, *with or without* monotonicity.

   (b) Under this alternative framework, and assuming that $Q4_i$ is a valid instrument, what would be the interpretation of the IV estimate of the return to education in Angrist and Krueger (1991)?