4 Instrumental Variables

1. Bias of the IV estimator.

   The IV estimator is unusual in that it is generally biased, but consistent.

   (a) Show that the IV estimator can be written as follows:

   $$\hat{\beta}_{\text{IV}} = \frac{\sum_{i=1}^{n}(z_i - \overline{z})y_i}{\sum_{i=1}^{n}(z_i - \overline{z})x_i} \tag{1}$$

   (b) Show that $\hat{\beta}_{\text{IV}}$ is generally a biased estimator for the population parameter β_1.
   Hint: Derive the expected value of $\hat{\beta}_{IV}$ conditional on x and z, and then note that the structural model is given by $y_i = \beta_0 + \beta_1 x_i + u_i$.

   (c) Show that $\hat{\beta}_{\text{IV}}$ is a consistent estimator of β_1.
   Hint: Derive the probability limit of $\hat{\beta}_{IV}$.

   (d) What does this result imply for practical empirical work?

2. Derivation of the Wald estimator.

   Assume we have the following structural model with a constant treatment effect ($= \beta_1$):

   $$y_i = \beta_0 + \beta_1 d_i + u_i \tag{2}$$

   We suspect d to be endogenous. However, we also observe a (potential) binary instrument z_i which we can use to estimate β_1.

   (a) Derive the Wald estimator.

   (b) State the necessary assumptions that are needed to identify β_1 using the instrument. Which ones of them you can test empirically?

3. Self selection revisited.

   You are studying the efficacy of a specific program on the outcome variable Y_i. Let $D_i = 1$ denote participation in the program, and $D_i = 0$ non-participation. The outcome given non-participation and given participation, respectively, are given by:

   $$Y_{0i} = \beta_0 + u_{0i} \tag{3}$$
   $$Y_{1i} = \beta_0 + \beta_1 + u_{1i} \tag{4}$$

   Assume that the error terms u_0, u_1 have zero expectation and are uncorrelated. Further assume that $\beta_1 > 0$. Individuals self-select into the program if the program has a positive effect on their outcome. That is:

   $$D_i = \mathbf{1}(Y_{1i} - Y_{0i} > 0) \tag{5}$$

   (a) Determine the ATE and the ATT. Which of the two is larger?

   (b) You want to estimate the ATT, using data on realized outcome Y_i and D_i. Will you over- or underestimate the true ATT? Give an analytical argument for your answer.
   Hint: Build your argument starting from the naive estimator's formula.

   (c) You induce participation in the program by an instrument, which is randomly assigned across individuals.

Assume that the participation decision (conditional on $Z_i = 1$) is given by $D_{1i} = \mathbf{1}(Y_{1i} - Y_{0i} + Z_i > 0)$, and that the decision rule given $Z_i = 0$ is $D_{0i} = \mathbf{1}(Y_{1i} - Y_{0i} > 0)$. Determine the LATE.

(d) Is the LATE larger or smaller than the ATE? Explain.

4. Application: Angrist's (1990) study on military service.

   (a) A simple OLS regression yields the following result:

$$\widehat{\ln(\text{earnings}_i)} = 6.4364 - 0.0255 \cdot \text{veteran}_i \tag{6}$$

   Explain why the simple OLS estimate of the effect of veteran status may be biased.

   (b) The following table shows the number of observations, indexed by actual instrument and treatment status (i.e. $Z_i = 1$ indicates a low lottery number).

	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	5,928	1,875
$D_i = 1$	1,400	863

   Estimate the fraction of (i) compliers, (ii) never-takes, and (iii) always-takers.

   (c) The following table shows estimated average outcomes by treatment and instrument status.

	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	$\widehat{\mathbb{E}(Y)} = 6.4472$	$\widehat{\mathbb{E}(Y)} = 6.4028$
$D_i = 1$	$\widehat{\mathbb{E}(Y)} = 6.4076$	$\widehat{\mathbb{E}(Y)} = 6.4289$

   Determine average potential outcomes for the four different compliance types. Compare outcomes across compliance types. What do you conclude?

   (d) Estimate the local average treatment effect.

5. IV in action.

   Use the data contained in `mortality.dta`. The data contains part of data used in the analysis of Kuhn, Wuellrich and Zweimüller (2010). Specifically, the file contains a 70% random sample of the overall female sample (with a restricted set of control variables).

   The outcome of interest is a binary variable indicating death before age 67 (death_i^{b67}). The (endogenous) regressor of interest is the number of years "spent" in early retirement (dist65_ageATend4emp). The instrument (Zd_during) denotes whether a worker has been eligible to a program that granted access to early retirement. The data also contains a set of control variables (work experience, birth cohort, region).[1]

   (a) First, describe the interesting variables (i.e. outcome, endogenous variable, instrument).

   (b) Estimate a simple OLS regression of the mortality indicator on the endogenous variable with and without a small set of control variables:[2]

$$\text{death}_i^{b67} = \beta_0 + \beta_1 \text{early retirement}_i + x_i\beta + \epsilon_i \tag{7}$$

[1] The control variables are labelled as follows: the variables for past work experience are `czeit·yATage50` (with $\cdot = 1, 2, 5, 10, 25$), birth cohort is denoted by `halfyearOFbirth`, and `nutsATage50` contains NUTS-3 regions.

[2] Include the following control variables: Dummy indicators for birth cohort (biannual frequency), controls for past work experience and past work experience spared (i.e. work experience in the last 1, 2, 5, 10, 25 years before age 50), and a set of regional dummies (at NUTS-3 level).

(c) Describe your results (e.g. assess through Stata or R the magnitude of your effects). Does the inclusion of the controls make any difference?

(d) Explain why this simple OLS estimate is probably biased. Explain the likely direction of the bias.
Hint: Use an omitted-variable bias argument with true health-status being the unobserved variable.

The variable `Zd_during` denotes whether a worker has been eligible to a program that granted access to early retirement.

(e) Plot the density of the endogenous variable by instrument status.
Hint: The `kdensity` *command in Stata might me useful. Also remember to plot the distributions only for individuals who already retired.*

(f) Estimate the following first-stage regression:

$$\text{early retirement}_i = \alpha_0 + \sum_j (Z_i \cdot C_{ij})\alpha_{Z_j} + x_i\beta + \varepsilon_i \tag{8}$$

where C_{ij} is the set of cohort dummies. We thus use a set of instruments (given by the interactions between Z_i and cohort dummies). The remaining control variables are the same as in the OLS regression.

(g) Estimate equation (7) using 2SLS (i.e. using the first-stage as given by (8) above).
Hint: Use the `ivregress` *command or the* `ivreg2` *package in Stata. However, also check that doing the two steps in sequence yields the same point estimate (but not the same standard errors).*

(h) Compare the OLS and 2SLS results. What do you conclude?