EC711 PS1. Many and Weak Instruments

Iván Fernández-Val

Boston University

Due on February 2

Question 1: Returns to Schooling (Angrist and Krueger, 1991) In this question you are going to replicate the empirical analysis of Angrist and Krueger (1991) and the simulation exercise of Hansen, Hausman and Newey (2007) using the dataset `hhn.dta`. The dataset contains a sample of $329,509$ observations including males born in 1930–1939 from the 1980 U.S. Census. The model includes the log of weekly earning as the outcome; the number of years of education as the endogenous regressors; the quarter, year and state of birth as the instruments; and indicators for married, black and SMSA as the exogenous controls (i.e., variables that should be included in both the first and second stage).

1. Estimate the returns of schooling by 2SLS and LIML using three sets of instruments: 3 quarter of birth indicators (S1); S1 plus interactions of quarter of birth indicators with year of birth indicators (S2); and S2 plus interactions of quarter of birth indicators with state of birth indicators (S3).

2. Calibrate the parameters of the following model using the data:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2' W_i + \sigma_u U_i, \\ X_i &= \pi_0 + \pi_1' Z_i + \pi_2' W_i + \sigma_v V_i, \quad i = 1, \dots, n, \end{aligned}$$

   where Y_i is the outcome, X_i is the endogenous regressor, W_i are the exogenous controls, Z_i is the set of controls, and U_i and V_i are zero mean errors independent of W_i and Z_i. Calibrate also $\rho_{u,v}$, the correlation between U_i and V_i. Here, you can follow Hansen, Hausman and Newey (2007) and use LIML with S1 as the set of instruments.

3. Generate $1,000$ synthetic datasets using the model fixing Z_i and W_i to the values in the data, $n = 329,509$, and generating U_i and V_i from a standard bivariate normal distribution with correlation $\rho_{u,v}$ independent across i. For each dataset, obtain the 2SLS and LIML estimates together with 95% confidence intervals obtained using the traditional and Bekker's standard errors for each of the 3 sets of instruments: S1, S2 and S3.

4. Report a table similar to the one in the lecture notes including bias, RMSE and empirical coverage probability of 95% confidence intervals using the traditional and Bekker's standard errors.

[Hint: the package `ivmodel` implements the 2SLS, LIML, and the Bekker's standard errors.]

Question 2: Weak Instrument Distribution (Staiger and Stock, 1997) In this question you are going to replicate the comparison between the normal and weak instrument approximation to the distribution of the 2SLS estimator of the lecture notes.

1. Generate $200,000$ datasets from the model

$$\begin{aligned} Y_i &= \beta X_i + U_i, \\ X_i &= (\Delta/\sqrt{n}) Z_i + V_i, \quad i = 1, \ldots, n, \end{aligned}$$

   where $\beta = 0$, $\Delta = .5$, $n = 100$, $Z_i \sim N(0,1)$ independent across i, and (U_i, V_i) is standard bivariate normal with correlation .5 independent across i. Compute the 2SLS estimator in each dataset.

2. Compare an estimate of the density of the 2SLS estimator with the normal density approximation and the weak instrument density approximation obtained by $200,000$ simulations.

3. Repeat the exercise for the LIML estimator. What do you find? Why?

[Hint: the R command `density` produces a kernel density estimate of a vector of observations.]

Question 3: Impact of Institutions on Economic Growth (Acemoglu, Johnson and Robinson, 2001) Acemoglu et al. (2001) examined the effect of institutions on economic performance using mortality rates among European colonists as an instrument for current institutions. In this exercise you are going to check the robustness of their results to the weak instrument problem. The file `ajr.txt` contains data on 64 countries on the following variables: the log of PPP adjusted GDP per capita in 1995 as the outcome Y; the average protection against expropriation risk from 1985 to 1995, which provides a measure of institutions and well-enforced property rights, as the endogenous regressor X; the log of the early European settler mortality rates as the instrument Z; and a normalized measure of distance from the equator (latitude) as the exogenous control W. Acemoglu, Johnson and Robinson argued that setter mortality is likely to be relevant for current institutions because settlers set up better institutions in colonies where they are more likely to establish long-term settlements, settler mortality was a relevant factor to decide to stay for a long term at the time of the initial colonization, and institutions are highly persistent. The exclusion restriction is motivated by the argument that GDP, while persistent, is unlikely to be directly influenced by mortality in the previous century, after conditioning on geography that is highly persistent and related to both GDP and mortality.

1. Estimate the effect of institutions X on log GDP per capita Y controlling for distance from the equator W by OLS and 2SLS using settler mortality as an instrument.

2. Compute the Anderson-Rubin (AR) statistic at a grid of values between .10 and 2.25 and plot it in a figure.

3. Report a table including 95% confidence intervals using the normal approximation, 95% AR confidence intervals that are robust to weak instruments, OLS estimates and standard errors, 2SLS estimates and standard errors, and AR estimates and standard errors. The AR estimates can be constructed as the center of the AR confidence interval and the AR as the length of the AR confidence interval divided by $2 * 1.96$. What is the logic behind these AR estimates and standard errors?

4. Compare the results for 2SLS and AR. Are the results of Acemoglu et al. (2001) robust to the weak instrument problem?