Bocconi University, Microeconometrics (cd. 20295)
Professor: Thomas Le Barbanchon and T.A.: Erick Baumgartner
Problem Set 2: Instrumental Variables

Due: Monday March 28th by 11:59pm. Submit by email to `erick.baumgartner@unibocconi.it`.

Problem Set 2

In this problem set, we will focus on two different topics related to instrumental variables.

In Exercise 1, we estimate the returns of education on health, using U.S. census data. In this exercise, we will address issues like potential omitted variables and weak instruments in a canonical 2SLS setting.

In Exercise 2, we study how import competition affects both labor and political outcomes, using different data sources. To do so, we will cover recent advancements on the shift-share instruments literature.

Instructions

No need to produce a pdf file with your answers. Save all graphics and tables requested (e.g., as `pset_2_exercise_1_question_1_a.xlsx`) and a do-file summarizing all of your work in a zipped folder identifying your group and the problem set (e.g., as `pset_X_group_Y.zip`) to `erick.baumgartner@unibocconi.it`. In the sub-questions where you are asked to write, please add your answer as a comment in your do file.

Hint: Type * to add a comment in your do-file (e.g., `* this is a comment line *`).

Hint: You can choose your preferred way of preparing tables: **(1)** one option is to use the command `outsheet` to construct the tables of summary statistics and the command `outreg2` to construct the regression tables (you can use them to export results of summary statistics and regressions to an excel file); **(2)** another option is to save results using the command `eststo` and then export these directly to a .tex (latex) file using the command `esttab`. Read carefully `help` for each command you choose and try different options, so as to have well-formatted tables.

Exercise 1

Short Summary of Discussion

Read Angrist and Krueger (1991) to understand how an instrumental variables approach solves for an omitted variable problem in their setting - that of estimating wage returns to education.

In addition, read two different references that point out fundamental problems of Angrist and Krueger (1991)'s analysis: **(a)** Bound et al. (1995) describes issues related to weak instruments and finite sample bias and **(b)** Bound and Jaeger (1996) discusses the inability of Angrist and Krueger (1991)'s quarter-of-birth IV to satisfy the exclusion restriction.

Data

We will use the 1980 US census data, the same data used by Angrist and Krueger (1991), Bound et al. (1995), and Bound and Jaeger (1996). We refer you to the data sections of these papers for more information.

Questions

1. Use `pset_2_q_1.dta`.

While doing research with instrumental variables we should start by visualizing the strength of our potential instrument before spending time estimating regressions. If we cannot visualize an identifying variation in our data, we will likely not have a strong first stage and, even if we have it, it will be very hard for us to convince our audience that we have a robust instrument.

After visualizing such an identifying variation, we should reflect on the research trade-offs of our instrument. In other words, we should carefully think about:

(i) Are exogeneity and/or exclusion restrictions violated? What bias in our IV estimates can be associated with each of these potential violations?

(ii) If our instrument satisfies both exogeneity and exclusion restrictions, what bias in our OLS estimates should we expect and why?

(iii) Which sub-population is affected by the instrument?

If we are convinced about the robustness of our instrument after answering all questions above, then we are ready to proceed to the regression estimation stage.

In this question, we will be asked to follow each step mentioned above using Angrist and Krueger (1991) as an example.

(a) Replicate figures I, II, and III from Angrist and Krueger (1991). Note that, graphics should look similar to the corresponding figures but not a perfect mirror.

Comment on the graphics generated. Are you able to observe a 1st quarter effect of lower education relative to subsequent quarters within the same year? In other words, are you able to observe what seems to be a relevant instrument?

Hint: Reflect about efficient ways to produce a variable with the mean of Education at the birth year-quarter level. If you are not able to do it, use the following code:

```
* collapsing at the year-quarter of birth level *
* while keeping year and quarter variables.*
collapse birthyear birthqtr Education (mean), by(birthdate)
```

Hint: After computing a variable with averages, use the command twoway connected with the option sort, restricting your plot to the years used in the figure you are replicating. Use birthdate to help you construct your graphic.

(b) Is **exogeneity** plausible for quarter of birth in a model about health status?

Is the **exclusion restriction** plausible?

Can you think of potential violations for the **exogeneity assumption** and the **exclusion restriction** in Angrist and Krueger (1991)'s setting?

(c) Do you expect higher or lower 2SLS estimates compared to OLS?

Why should you expected such a bias on your OLS estimates?

If the instrument is quarter of birth and the outcome health, who are the "*compliers*"?

2. Use the pset_2_q_2_and_3.dta.

(a) Compute the average of outcome Healthy and save it as a scalar named mu_y.

Compute the average of Education and save it as a scalar named mu_x.

(b) Create quarter of birth dummies named Quarter1 Quarter2 Quarter3 Quarter4.

Create one dummy for each region in region.

Create a local named Controls with: Central, Married and the regional dummies.

Create a local named Birth_Year_FEs with dummies for each year of birth.

(c) Estimate an OLS regression of Healthy on Education.

Estimate an OLS regression of Healthy on Education and Controls.

Estimate an OLS regression of Healthy on Education, Controls, and Birth_Year_FEs.

(d) Export each OLS regression to an excel table named `TABLE_Q_2`.

Only show the coefficient of `Education` and the usual regression statistics.

Add two lines to the table, one named `Controls` and another named `Year of birth FEs`, with legends `YES` or `NO` to communicate to your reader when a regression model has controls or birth year dummies, respectively.

Add a line to the table with the average of `Healthy`, naming it `Mean y`.

Add a line to the table with the average of `Education`, naming it `Mean x`.

(e) Estimate an IV regression of `Healthy` on `Education` using `Quarter1 Quarter2 Quarter3` as instrumental variables for `Education`.

Estimate an IV regression of `Healthy` on `Education` using `Quarter1 Quarter2 Quarter3` as instruments for `Education`, and `Controls` as controls.

Estimate an IV regression of `Healthy` on `Education` using `Quarter1 Quarter2 Quarter3` as instruments for `Education`, and `Controls` and `Birth_Year_FEs` as controls.

(f) Append each 3 IV regression to `TABLE_Q_2`.

Only show the coefficient of `Education` and the usual regression statistics.

Add two lines to the table, one named `Controls` and another named `Year of birth FEs`, with legends `YES` or `NO` to communicate to your reader when a regression model has controls or birth year dummies, respectively.

Add a line with the F-statistic of the excluded instruments, naming it `F-statistic IVs`.

3. Use the `pset_2_q_2_and_3.dta`.

Sometimes we have limited space to present our IV estimates. Questions **3.(a)** to **3.(d)** are intended to prep us into summarizing our most relevant output in one given IV specification in one concise table. While answering these questions we will estimate regressions previously handled in the current problem set.

(a) Estimate an **OLS** regression of `Healthy` on `Education` using `Controls` and `Birth_Year _FEs` as controls, and export it to an excel table named `TABLE_Q_3`.

Only show the coefficient of `Education` and the usual regression statistics.

(b) Estimate a **first stage** for an IV regression of `Healthy` on `Education` using `Quarter1 Quarter2 Quarter3` as instruments for `Education`, and `Controls` and `Birth_Year_FEs` as controls, and append such a first stage regression to `TABLE_Q_3`.

Only show the coefficient of `Quarter1 Quarter2 Quarter3` and the usual regression statistics.

Add a line with the F-statistic of the excluded instruments, naming it `F-statistic IVs`.

(c) Estimate a **reduced form** regression of `Healthy` on instruments `Quarter1 Quarter2 Quarter3` and `Controls` and `Birth_Year_FEs` as controls.

Append such a reduced form regression to `TABLE_Q_3`.

Based on the results of question **2.(e)**, what are the expected signs of the coefficients of `Quarter1 Quarter2 Quarter3`?

Are these reduced form coefficients in line with your expectations?

(d) Estimate a **second stage** for an IV regression of `Healthy` on `Education` using `Quarter1 Quarter2 Quarter3` as instruments for `Education`, and `Controls` and `Birth_Year_FEs` as controls.

Append such a second stage regression to `TABLE_Q_3`.

Only show the coefficient of `Education` and the usual regression statistics.

(e) Bound et al. (1995) discuss a number of issues that arise when confronted with both weak instruments and finite sample bias, in particular, when there exists a weak correlation between the instrument and the outcome variable.

Discuss how these issues can generate a bias in the IV regression you have estimated in item **(d)**.

Can you reject the the null hypothesis of the test of joint significance of the instruments?

Based on the size the F-statistic, can you say if finite sample bias is likely or not to be an issue in this case?

(f) Create a local named `State_FEs` with dummies for each state of birth, except Wyoming, which is intended to be the reference category.

Create a local named `Year_Quarter_FEs` with dummies for each year-quarter births, except 1939-4, which is intended to be the reference category.

Create a local named `State_Quarter_FEs` with dummies for each state-quarter births, except Wyoming-4, intended to be the reference category.

(g) Estimate an IV regression of `Healthy` on `Education` using `Year_Quarter_FEs` as instruments for `Education`, `Controls` and `Birth_Year_FEs` as controls.

Estimate an IV regression of `Healthy` on `Education` using `State_Quarter_FEs` as instruments for `Education`, and `Controls`, `Birth_Year_FEs`, and `State_FEs` as controls.

Use Wyoming as omitted category for state of birth and include Washington DC.

(h) Compute the F-statistic for the excluded instruments in point **(g)**.

Can you say if both regressions are likely to suffer from finite sample bias?

Exercise 2

Short Summary of Discussion

Read Autor et al. (2013) to understand which kind of endogeneity does Bartik instruments address in a setting in which local labor market outcomes are being related to import competition from abroad (more specifically, from China).

Afterwards, read Goldsmith-Pinkham et al. (2020) to both (a) have a sense of how should one interpret a Bartik instrument in a setting such as Autor et al. (2013) and (2) have a better understanding of which general identifying assumptions must hold in order for you to say that Bartik IV estimates are causal.

Lastly, read Autor et al. (2020) to understand how an IV identical to that used in Autor et al. (2013) can be applied in a study focused on understanding how globalization affected political outcomes as broad as ideology of campaign donors, congressional election outcomes, presidential election votes, among others.

Data

To replicate Autor et al. (2013), we will partially follow Goldsmith-Pinkham et al. (2020). For more information on which datasets are used and how are variables defined, we refer you to Goldsmith-Pinkham et al. (2020)'s online appendix (`here`).[1] Instead, when turning to Autor et al. (2020), we will take reference on Autor et al. (2020)'s replication files (`here`).

Questions

1. Read Autor et al. (2013) and Goldsmith-Pinkham et al. (2020). Focus on those sections that address identification - i.e., Section I.B in Autor et al. (2013); Sections II.A, IV.A and A1.4 in Goldsmith-Pinkham et al. (2020).

[1] This replication takes reference from Adão et al. (2019) and Borusyak et al. (2022).

(a) Assume that import competition affects the rate of manufacturing employment homogeneously. State which identification assumptions must hold under this setting for those IV estimates presented in Autor et al. (2013) to be consistent.

(b) Which additional assumptions would be necessary to hold if import competition would affect manufacturing employment differently, depending on a set of factors?

2. Study carefully `make_rotemberg_summary_ADH.do` (available on Blackboard). Use those `.dta` files referenced in `make_rotemberg_summary_ADH.do` (available on BB as well).

   In addition, read Goldsmith-Pinkham et al. (2020). Focus on section A1.5 which discusses whether the Bartik IV used in Autor et al. (2013) suffers from misspecification.

   (a) Plot the distribution of Rotemberg weights associated to Autor et al. (2013) in a parametric and a non-parametric manner (overlay both graphs in a single plot).

   `Hint:` Search in `make_rotemberg_summary_ADH.do` for a variable titled `alpha1`.

   (b) Compile a LaTeX table with 2 panels: panel A, identical to panel E in table A1 of Goldsmith-Pinkham et al. (2020); panel B, inspired in panel D of table A1 - the rows of the table should include those industries belonging to the top 5 industries in terms of Rotemberg weights; the columns should include information about: (i) $\hat{\alpha}_k$, (ii) g_k, (iii) $\hat{\beta}_k$, (iv) 95% CIs, (v) Ind. Share and (vi) Share of overall β.

   Hint: Perform minor adjustments on `make_rotemberg_summary_ADH.do`.

   (c) Replicate Figures A2 and A3 from Section A of Goldsmith-Pinkham et al. (2020). Having both these figures into account, which type of TE heterogeneity seems to be present in Autor et al. (2013)? Does it preclude you from interpreting the IV estimates present in the paper as a LATE?

3. Read Autor et al. (2020). Focus in particular on those sections that (i) discuss how is import competition instrumented and (ii) relate import competition with voting decisions in presidential elections - Section II.A and Section V accordingly.

   (a) State which are the identifying assumptions necessary for those IV estimates presented in Section V to be consistent. Discuss whether these assumptions are more or less plausible in this setting (relative to Autor et al., 2013).

References

Adão, R., Kolesár, M., and Morales, E. (2019). Shift-Share Designs: Theory and Inference. *The Quarterly Journal of Economics*, 134(4):1949–2010.

Angrist, J. D. and Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics*, 106(4):979–1014.

Autor, D. H., Dorn, D., Hanson, G. H., and Majlesi, K. (2020). Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure. *American Economic Review*, 110(10):3139–83.

Autor, David, H., Dorn, D., and Hanson, G. H. (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review*, 103(6):2121–68.

Borusyak, K., Hull, P., and Jaravel, X. (2022). Quasi-Experimental Shift-Share Research Designs. *The Review of Economic Studies*, 89(1):181–213.

Bound, J. and Jaeger, D. A. (1996). On the Validity of Season of Birth as an Instrument in Wage Equations: A Comment on Angrist & Krueger's "Does Compulsory School Attendance Affect Schooling and Earnings?". *NBER Working Paper Series.* No. 5835.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430):443–450.

Goldsmith-Pinkham, P., Sorkin, I., and Swift, H. (2020). Bartik Instruments: What, When, Why, and How. *American Economic Review*, 110(8):2586–2624.