Data Analysis Test

1 Main Instructions

A large part of the work as a Research Assistant consists in manipulating and analyzing survey and administrative data. We will use this test to evaluate your skills in working with data and communicating your analysis's main results. You can complete the test using any software among Stata, Python, R, Matlab, Julia (or a combination of those). You may consult any resources you like, but you must complete the test without help from other people. Recall that this is an RA responsibility for you, not for the person you might want to consult with.

The test contains two main exercises. In the first exercise, we ask you to implement an instrumental variable strategy. In the second exercise, we ask you to take transaction- and firm-level data to compute the direct and total shares of sales to Multinational Enterprises. Both questions have equal weight and each item within a question has the same weight.

The test is designed to take less than five hours from the moment you receive this document. We value speed, but we value correctness even more. A correct answer sent after slightly more than five hours has more merit than an incorrect answer sent after one hour. We won't evaluate submissions that take more than five hours and a half. All necessary files for completing this test are included.

Please send your .log file, .do / markdown file, etc., the codes, the final dataset and any deliverable in a zip file labeled with your first and last name. Please separate into two different subfolders the outputs of the two exercises. Be careful and methodical with your work. Ensure your code is well formatted and annotated. If you become stuck, explain (in comments in your code) what you would have done if you had more time or knew the correct commands. References are not needed, although, if you want to need to add them, a hyperlink is acceptable.

Good Luck and Have Fun!

2 First Exercise: IV strategy

2.1 Data

You'll find in your .zip file a dataset called "INEGI_employment.dta". This file corresponds to a panel dataset of Mexican commuting zones (CZ), which contains the following variables:

- *year:* Year of the Economic Census.
- *CZ:* Commuting zone (the union of several municipalities).
- *country_code:* The country of ownership of establishments employing workers in a given CZ and year.
- *workers:* The total number of employees working for an establishment whose ownership comes from *country_code* in a given CZ and year.

2.2 Context

We want to understand the effect of expanding foreign establishments' employment (those with a country of ownership/origin different from MEX) on domestic employment (employment of Mexican firms).

Let t be a calendar year. Our main outcome of interest is

$$\Delta \ell_{cz,t} = \log(L^D_{cz,t}) - \log(L^D_{cz,t-5}),$$

where $L^D_{cz,t}$ is the total employment in Mexican owned firms in commuting zone cz and year t. Our main explanatory variable is defined as the growth in foreign employment as a share of the total employment of the CZ:

$$\hat{X}_{cz,t} \equiv \frac{L^F_{cz,t} - L^F_{cz,t-5}}{L_{cz,t-5}},$$

where $L^F_{cz,t-5}$ is the total foreign employment and $L_{cz,t-5} = L^F_{cz,t-5} + L^D_{cz,t-5}$ is the total employment in cz, t. The total foreign employment in cz,t is the sum of foreign employment from different countries of ownership/origin o. Therefore $L^F_{cz,t} = \sum_o L^{F_o}_{cz,t}$, where $L^{F_o}_{cz,t}$ is the total foreign employment from origin o in cz, t. Note that one can rewrite the previous

expression as:

$$\begin{aligned}\hat{X}_{cz,t} \equiv & \frac{\sum_o \left(L^{F_o}_{cz,t} - L^{F_o}_{cz,t-5}\right)}{L_{cz,t-5}} \\ = & \sum_o \frac{L^{F_o}_{cz,t} - L^{F_o}_{cz,t-5}}{L^{F_o}_{cz,t-5}} \frac{L^{F_o}_{cz,t-5}}{L_{cz,t-5}} \\ = & \sum_o \frac{L^{F_o}_{cz,t} - L^{F_o}_{cz,t-5}}{L^{F_o}_{cz,t-5}} S^o_{cz,t-5} \\ = & \sum_o \Delta L^{F_o}_{cz,t} \times S^o_{cz,t-5}\end{aligned}$$

where $S^o_{cz,t-5} \equiv \frac{L^{F_o}_{cz,t-5}}{L_{cz,t-5}}$, and $\Delta L^{F_o}_{cz,t} \equiv \frac{L^{F_o}_{cz,t} - L^{F_o}_{cz,t-5}}{L^{F_o}_{cz,t-5}}$ is the percentage change in origin-o employment in the commuting zone cz from Census year $t-5$ to t. Note that the last equation has a shift-share structure.

2.3 Questions

The main output of this exercise is a pdf with the tables indicated below in addition to your log file / Rmarkdown / Jupyter note, etc., where you show your code, calculations and results of all the subparts of this exercise. Please complete the following questions:

1. We want to construct six instrumental variables denoted as $\hat{Z}^i_{cz,t}$ with $i \in \{1, \ldots, 6\}$. These are:

$$\hat{Z}^1_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz'} L^{F_o}_{cz',1994}} S^o_{cz,1994}$$

$$\hat{Z}^2_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz' \neq cz} L^{F_o}_{cz',1994}} S^o_{cz,1994}$$

$$\hat{Z}^3_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz'} L^{F_o}_{cz',t-5}} S^o_{cz,t-5}$$

$$\hat{Z}^4_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz' \neq cz} L^{F_o}_{cz',t-5}} S^o_{cz,t-5}$$

$$\hat{Z}^5_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz'} L^{F_o}_{cz',t-5}} S^o_{cz,1994}$$

$$\hat{Z}^6_{cz,t} \equiv \sum_o \frac{\sum_{cz' \neq cz} \left(L^{F_o}_{cz',t} - L^{F_o}_{cz',t-5}\right)}{\sum_{cz' \neq cz} L^{F_o}_{cz',t-5}} S^o_{cz,1994}$$

You will create three of these variables. Pick one from each of these tuples: ($\hat{Z}^1_{cz,t}$, $\hat{Z}^2_{cz,t}$), ($\hat{Z}^3_{cz,t}$, $\hat{Z}^4_{cz,t}$), and ($\hat{Z}^5_{cz,t}$, $\hat{Z}^6_{cz,t}$). Please write the numbers of the variables you intend to create. In addition, compute the variables $\Delta\ell_{cz,t}$ and $\hat{X}_{cz,t}$.

2. Present a summary statistics table for these five variables by Census. Include the 1^{st} and 99^{th} percentiles, median, mean, and standard deviation for each variable.

3. We want to implement an IV strategy to identify the causal effect of $\hat{X}_{cz,t}$ on $\Delta\ell_{cz,t}$. $\hat{Z}^i_{cz,t}$ for $i \in \{1, \ldots, 6\}$ are candidates for an IV. Which of them could be written with a shift-share structure? Write the number of the instruments that correspond to the shift-share structure and complete the algebra to obtain the $\Delta L^{Fo}_{cz,t}$ of the instruments that follow the shift-share structure.

4. Compute the estimates of the parameters of the following regression.

$$\Delta\ell_{cz,t} = \beta\hat{X}_{cz,t} + \gamma_{cz} + \gamma_t + \varepsilon_{cz,t},$$

where γ_{cz} are CZ fixed effects, γ_t are year fixed effects, and $\varepsilon_{cz,t}$ is an error term. Pick any $\hat{Z}^i_{cz,t}$ to instrument $\hat{X}_{cz,t}$. Use the OLS and IV estimators with fixed effects. Cluster the standard errors at the CZ-year level. Present a table with four columns: the OLS, the First Stage, the Reduced Form, and the IV regressions. Ensure to include the coefficient of interest, standard error, number of observations, first stage F statistic, and R^2.

5. Discuss in less than 150 words the main concerns to interpret β as a causal estimate with the OLS estimator.

6. Considering the results of the regressions, discuss in less than 300 words the assumptions for the IV estimator to causally estimate β and how you can assess them with the results you obtained.

7. Given your last response, discuss in less than 300 words the relevance of the instrument interpreting the different tests' statistics and if the IV strategy works in this context.

3 Second Exercise: Indirect sales to Multinational Enterprises

3.1 Data description

For this question you will use two administrative datasets provided to you:

1. "firms.dta": This file presents information for 12,000 fictional firms. It contains the following variables:

   - *ID*: Identifier for the firm.

- *exports*: Exports in USD.
- *consumption*: Sales to consumers in USD.
- *firm_type*: A categorical variable indicating the firm's ownership type. It takes the value of 1 if the firm is domestic informal, 2 if it is domestic formal, and 3 if it's a Multinational Corporation (MNC).

2. "transactions.dta": This file corresponds to the transactions (sales) register between the firms described in the previous dataset. It contains the ID for both the *seller* and the *buyer*, as well as the *trans* (amount) sold. This file is already collapsed by seller-buyer, meaning there is at most one observation per pair of firms.

3.2 Context and notation

In this exercise, you will calculate and graph the share of direct and total sales from domestic firms to multinationals (MNC).

Denote x_{ij} as the total transactions from seller i to buyer j. Define the total sales of firm i $X_i \equiv x_{iC} + x_{iX} + \sum_j x_{ij}$, where x_{iC} are the sales to final consumption, x_{iX} are exports, and $\sum_j x_{ij}$ are the transactions of firm i to other firms.

We denote as $s_{i,j}$ the element on the row i, column j of a matrix $\mathbf{S}$. These entries are given by:

$$s_{i,j} = \frac{x_{ij}}{X_i}, \forall i \notin \mathcal{M}.$$

Denote $\mathcal{M} = \{i \mid i \text{ is an MNC}\}$ as the set of MNC firms. Denote $s_{i,M}$ as the share of direct sales of firm i to MNCs. Note that $s_{i,M} \equiv \frac{\sum_{j \in \mathcal{M}} x_{ij}}{X_i}$.

Finally, denote the **total** share of sales to MNCs as $S_{i,M}$. This refers to the direct plus indirect sales share of domestic seller i's sales that reach MNCs through all possible input-output linkages (e.g., through direct sales plus sales to the suppliers plus sales to the suppliers of the suppliers and so on). Note that $S_{i,M}$ follow this recursive equation:

$$S_{i,M} = s_{i,M} + \sum_{j \notin \mathcal{M}} s_{i,j} S_{j,M}, \forall i \notin \mathcal{M}, \tag{1}$$

which would allow you to compute the vector $\mathbf{S_M} = (S_{1,M}, S_{2,M}, \ldots, S_{N,M})$, with N being the total number of firms in the economy.

3.3 Questions

1. Write the system of equations 1 in matrix form.

2. Prepare a dataset that merges the firms.dta and the transactions.dta datasets. The resulting dataset should include the transactions, total sales, sales to final consumers, and exports for each buyer and each seller.

3. Compute $s_{i,j}$, $s_{i,M}$. What is the mean and variance across all firms i?

4. Solve the system in 1 following the system you proposed in point 1 to obtain $S_{i,M}$.

5. Plot two histograms: One for $S_{i,M}$ and one for $S_{i,M} \cdot X_i$. Include all $i \notin \mathcal{M}$

6. Assume you must solve the system in 1 using a large dataset. Inverting a matrix of hundreds of thousands of firms involves time-consuming computational power and memory. Denote (with some abuse of notation) Where $S^k = \underbrace{S \cdot S \cdot S \cdot ...}_{k\text{-times}}$. . You suggest approximating $\mathbf{S_M}$ in the following way:

$$\mathbf{S_M} \approx (\mathbf{I} + \mathbf{S}) \left(\mathbf{I} + \mathbf{S}^2\right) \left(\mathbf{I} + \mathbf{S}^4\right) \left(\mathbf{I} + \mathbf{S}^8\right) \cdots \left(\mathbf{I} + \mathbf{S}^L\right) \mathbf{s_M},$$

where I is the identity matrix, $\mathbf{s_M} = (s_{1,M}, s_{2,M}, \ldots, s_{N,M})$ and $\mathbf{S_M} = (S_{1,M}, S_{2,M}, \ldots, S_{N,M})$, and N is the total number of firms in the economy. Propose an algorithm to implement this approximation. What value of the exponent L would be the minimum value to produce a good enough approximation for $\mathbf{S_M}$?