Problem Set 2 - IV

Professor: Chiara Monfardini;
Teaching assistant: Vito Stefano Bramante

- This problem set was sent on **March 11, 2022**
- Due date: **March 18, 2022** at **midnight**
- E-mail your assignment at vitostefano.bramant2@unibo.it no later than **midnight**. Assignments sent after the deadline will not be graded (no exceptions).
- You must attach a single zip file containing: (i) a pdf answer sheet; (ii) the Stata log-file; (iii) the Stata do-file. The pdf file should be **no longer than 5 pages (w/out tables).**
- Each table and graph in the pdf file should be fully reproducible. **By simply running the .do file** I should be able to reproduce the **exact** table or graph you are showing in the pdf, including title, variable names and numbers. In the .do file you have to signal clearly which chunk of code reproduces which table or which chart. **Doing so will guarantee you up to 2 bonus points on each problem set**.
- Name the zip file (and each file) as *surname*1_*surname*2_*surname*3.*zip*; remember to write the name, surname and student number of each student in the answer sheet.
- The grade for the 3 assignments will be 40% of the final grade.
- Please follow carefully the instructions detailed above. **Any misconduct will negatively impact the grading of the assignment.**

Consider the data set *yogurt.dta*, which contains data about sales of products in the yogurt market across a sample of grocery stores in the Italian territory. Ignore the time/within component, as the problem set is focused on instrumental variables techniques in the cross-section. The aim of this problem set is to **estimate a demand function** accounting for possible endogeneity. The unit of observation is the product, for which information about intrinsic characteristics and commercial performance have been collected. The outcome variable is *sales_volume*, the quantity sold (1 unit = 1 serving = 125 g) of product i in week t while *price* is the main explanatory variable: average weekly prices in € per serving, without promotions or discounts and expressed in September 2012.

$$\log sales\_volume_{i,t} = \alpha + \gamma \log price_{i,t} + \mathbf{x}'_{\mathbf{i,t}}\beta_x + \mathbf{s}'_{\mathbf{i}}\beta_s + \lambda t + \epsilon_{i,t} \tag{1}$$

where $\mathbf{x}'_{\mathbf{i,t}}$ is a vector of explanatory variables related to intrinsic characteristics of the product (and β_x is the corresponding vector of coefficients). Variables that must be included in $\mathbf{x}'_{\mathbf{i,t}}$ are: *carbo, energy, fat, protein, cram, drink, flav, firm* and *pri_label*. Please see the **Codebook** for a detailed description of the variables. $\mathbf{s}'_{\mathbf{i,t}}$ is a vector of explanatory variables related to characteristics of the store where the product is sold (and β_s is the corresponding vector of coefficients). The variable t is a time trend which might account for seasonality.

During the analysis:

- consider a 10% level of statistical significance when testing hypotheses
- Discuss the estimation of the results commenting signs, magnitudes and statistical significance of the estimated coefficients
- Choose carefully the variables to include in the vector **s**; once you have decided them, stick with your specification and do not change it throughout.
- These data are confidential. Please do not use them for your own personal research.

Part 1: Theoretical Analysis

Consider product i's market demand and supply functions:

$$q_i^d = a + bp_i + u_i \tag{2}$$

$$q_i^s = c + dp_i + v_i \tag{3}$$

where $q_i^d(q_i^s)$ is the quantity demanded (supplied) and p_i is the per-unit price. The error term u_i represents shocks affecting product i's demand such as price of substitutes/complements or consumers' tastes; these shocks are *demand-shifter* because they move the demand function in the (q,p) plane.The error term v_i represents shocks affecting product i's supply such as input prices; these shocks are *supply-shifters* because they move the supply function in the (q,p) plane. Assume that $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ and $Cov(u_i, v_i) = 0$. Let $\sigma_u^2 = Var(u_i)$ and $\sigma_v^2 = Var(v_i)$; both are different from zero.

Now consider the market-clearing equilibrium condition: $q_i = q_i^d = q_i^s$. Equation 2 and 3 become:

$$q_i = a + bp_i + u_i \tag{4}$$

$$q_i = c + dp_i + v_i \tag{5}$$

1. What are the risks associated with estimating Equation 4 or equation 5 ? Do you think we can trust an OLS estimator applied to either of the two equations?
2. Prove that the orthogonality condition fails in both equation 4 and equation 5. *[Hint: solve the system for q_i and p_i and use the explicit equations derived to verify whether $Cov(p_i, u_i) = 0$ and $Cov(p_i, v_i) = 0$, that is plug-in the resulting equation for p_i.]*
3. Imagine to have the opportunity to observe a sudden increase in the price of milk, a crucial production input for the yogurt market. Do you think it would be an appropriate instrument? Justify your answer through the lenses of the properties that constitute a relevant instrument, discussing both those that could be tested and those that could not.

Part 2: OLS and IV estimation

1. Estimate your model by OLS and comment the results. Adopt robust standard errors at your discretion, explaining the reason behind your decision. In light of the theoretical investigation of part 1, can you trust your results? Discuss.
2. *"The most popular identifying assumption used to deal with the above endogeneity problem is to [use] the observed product characteristics (excluding price and other potentially endogenous variables), the sums of the values of the same characteristics of other products offered by that firm (if the firm produces more than one product), and the sums of the values of the same characteristics of products offered by other firms"* (Nevo 2000, pp. 534-535) [1]. Comment why such approach can work in the estimation of the market demand. But also, through the lenses of the properties that make instruments relevant, *discuss the reasons why this approach might fail.*
3. Consider *carbo*1 as an instrument for log *price*. Test its relevance. Might this be a good instrument given the result of your test and the discussion above?
4. Adopt a two-stage least squares approach to estimate Equation 1 considering *carbo*1 as the instrument. Estimate $\hat{\gamma}_{IV}$ both manually (with two separate OLS regressions) and using the command *ivreg2*. Comment your results. Are the two procedures generating some differences?
5. Test the endogeneity of log *price* both manually and using the output of the *ivreg2* command. Are your conjectures correct?

Part 3: Multiple instruments

1. Now consider the whole set of instruments you have at your disposal. Test the joint relevance of these instruments.
2. Estimate Equation 1 using *ivreg2* and the whole set of instruments. Then, test for overidentifying restrictions and comment about the conclusion of your test. Should we trust the price elasticity just found? What about the results at part 2? *[We are interested in your reasoning.]*

[1] Nevo, A. (2000) "A Practitioner's Guide to Estimation of Random-Coefficients Logit Models of Demand". *Journal of Economics and Management Strategy*, 9(4), 513-548

Codebook

Time variables

- *week*: time unit
- *yweek*: mod(week,52)
- *periodo1*: cos(yweek/52*2*3.14159)
- *periodo2*: sin(yweek/52*2*3.14159)

Product variables

- *firm*: id number of the producer
- *sales_volume*: volume of sales for the product i in week t
- *category*: product i's category; there are 10 of them (Flavor, Creamy Flavor, Probiotic Flavor, Functional Flavor, Biologic Flavor, Probiotic Drinkable Flavor, Natural, Natural Probiotic, Natural Probiotic Drinkable, Topping)
- *subcat*: product i's subcategory; there are 5 of them (Low Fat, Whole, Healthy, Healthy Whole, Healthy Low Fat)
- *rank*: the position of product i in the ranking of most sold products within store in each week
- *pri_labe*: 1 if product i is a private label
- *energy*: calories per 100g of yogurt
- *carbo*: grams of sugar per 100g of yogurt
- *fat*: grams of fat per 100g of yogurt
- *protein*: grams of protein per 100 g of yogurt
- *flav*: 1 if the brand is flavoured; 0 if it is natural or white
- *cream*: 1 if the yogurt has a creamy texture
- *drink*: 1 if the brand is sold only in bottles (instead of pots or jars)

Store variables (constant within stores)

- *hyper*: 1 if store is a superstore (floor space above 2,500 squared meters)
- *poptot*: population in the market of store
- *hhtot*: number of households in the market of store
- *incometot*: total income (e) in the market of store
- *constot*: value of consumption (e) in the market of store
- *mtot (wtot)*: number of men (women) in the population in the market of store
- *age_pop*: average age of population in the market of store
- *sqmtot*: total squared meters of stores in the market of store
- *sqm_own*: squared meters of store

Variables that can be used as instruments

- *energy1* : average of the value of energy for the other products (offered by the same firm)
- *carbo1* : average of the value of carbohydrates for the other products (offered by the same firm)
- *fat1* : average of the value of fat for the other products (offered by the same firm)
- *protein1* : average of the value of protein for the other products (offered by the same firm)
- *energy2* : average of the value of energy for all the other products (offered by the same and the other firms)
- *carbo2* : average of the value of carbohydrat for all the other products (offered by the same and the other firms)
- *fat2* : average of the value of fat for all the other products (offered by the same and the other firms)
- *protein2* : average of the value of protein for all the other products (offered by the same and the other firms)