

## STATA Assignment 2

### Important instruction before you begin

The deadline to submit the assignment is **Friday February 25, 11:59 PM**. The dataset for this problem set is **analysis1.dta**. As before, we will provide a basic do file called **PS2\_yourPID.do**. Please rename with your PID so the file you turn in resembles **PS2\_A34567890**. **Only the final do file needs to be uploaded to Canvas**. You will solve this problem set by modifying the do file **only**. Ideally, your .do file should be in the same folder or directory as the data file so you do not need to use **cd** to switch to the needed directory. If you need to switch directories, put the word **cap** at the beginning of the command so it will not execute when we try to run your code.

The do file has comments indicating where your answers need to be written. For the questions which require an answer in words, write the answer where indicated in the do file. Remember that in Stata, commands beginning with a **\*** or wrapped like **/\*my command\*/** will not execute. **Make sure your code runs.**<sup>1</sup> **If it does not run, then you will have 25% subtracted from your final score.** We highly recommend running your do file before submitting it to make sure it runs.

### Background

This assignment is based on the paper “From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising” (2021) by Francesco Decarolis and Gabriele Rovigatti published in the *American Economic Review*. The authors motivate their analysis as follows:

Online advertising sales are the main fuel of all of the major digital platforms. In the internet era, advertising means capturing the attention of consumers who are browsing the web and this requires both detailed data to effectively target the ad to the right customers and algorithms to bid in the online auctions where ad space is sold. These needs have led to a major, but understudied, shift in the industry: rather than bidding individually, advertisers increasingly delegate their bidding to highly specialized intermediaries. This concentration of demand within a few large intermediaries raises the question of countervailing buyer power. Can the emergence of intermediaries counterbalance the highly concentrated supply of online ads? (pages 3299-3300)

The authors establish causality by leveraging a novel data set, natural language processing, and instrumental variables.

---

<sup>1</sup>By “run”, we mean that the TA can click “do” in the do-file editor and the whole do-file runs through and produces the desired results and produces all the requested output.

The third ingredient is an instrumental variable (IV) strategy. Instruments are needed for two reasons: measurement error in the proxy for demand concentration and potential omitted variable bias. For instance, there might be unobservable shocks to the popularity of some keywords that drive changes in both revenue and demand concentration. Similar to Dafny, Duggan, and Ramanarayanan (2012), we address this problem by exploiting the variation in intermediary concentration driven by changes in network ownership of MAs. In our sample period, there were 21 acquisitions and 2 divestments, affecting 6 out of the 7 agency networks. These merger and acquisition (MA) operations, especially the larger ones involving a multiplicity of markets, are a useful source of variation in demand concentration as the revenue dynamics in each local market are too small by themselves to cause the MA operations. We extensively discuss this empirical strategy and evaluate its robustness. (page 3301)

Now your task is to replicate some simple aspects of the author’s study and try to understand how IV helps to resolve the measurement error and omitted variable bias problems.

## Questions

1. The dataset `analysis1.dta`<sup>2</sup> contains the main data used in the paper. The outcome of interest is Google’s estimated log revenue `logr_hat` (measured in log dollars). This variable is an estimate for Google’s revenue from advertising on searches with a specific cluster of keywords. Think of a cluster of keywords as a “market” or industry.

The covariate of interest is a Herfindahl-Hirschman index (HHI) `HHI_hat`. This variable measures concentration on the demand side of the market. Firms want to advertise their products beside Google search results. If these firms are competitive, then HHI is low. If firms are oligopolistic or not competitive, then HHI is high. HHI ranges from 0 to 1.

Merger and acquisition operations mean that firms combine or break apart. When this happens, the market concentration changes. The instrument variable is `sim`, the simulated change in market concentration (HHI) that occurs from merger and acquisition operations.

Load the dataset into STATA, get an overview over the dataset (STATA: `describe` and `sum`).

2. Create a scatter plot showing the correlation between `HHI_hat` and `sim`. Include a linear prediction in your graph. Explain the intuition of this figure.
3. Now consider the model:

$$\text{logr\_hat}_{mt} = \alpha + \beta \text{HHI\_hat}_{mt} + \epsilon_{mt}, \quad (1)$$

where  $m$  indexes markets and  $t$  indexes years.

---

<sup>2</sup>Decarolis, Francesco, and Rovigatti, Gabriele. Data and Code for: From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising. Nashville, TN: American Economic Association [publisher], 2021. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-09-29. <https://doi.org/10.3886/E130502V1>

- (a) Run an OLS regression of revenue in the market-year on market concentration (HHI) using robust standard errors.
  - (b) Interpret the coefficient  $\beta$ . Is it statistically significant? Is it economically meaningful? (Answer in max. 2-3 sentences)
  - (c) According to the paper, what is the potential problem that could cause OLS to be inconsistent? (Hint: Read the Identification Strategy section on page 3316.) (Answer in max. 2-3 sentences)
4. As explained before, the authors use mergers and acquisitions as an instrument for changes in market concentration.
- (a) Run TSLS manually (i.e., run the first stage, and then run the second stage, without using the automated IV command) and compare the TSLS estimate to the OLS estimate.
  - (b) Run TSLS using the automated STATA command `ivregress 2sls`.
  - (c) According to the authors, what is the reasoning for the validity of this instrument? (Hint: Read the Identification Strategy section on page 3317.)
  - (d) Should we worry about weak instruments in this application? Conduct a formal test.
  - (e) The authors consider the more general model:

$$\begin{aligned}
 \log r_{\text{hat}}_{mt} = & \alpha + \beta \text{HHI\_hat}_{mt} \\
 & + \beta_n \text{numberofresults}_{mt} + \beta_b \text{branded}_{mt} + \beta_l \text{long\_tail}_{mt} \\
 & + \sum_m \gamma_m \mathbf{1}\{\text{market} = m\} \\
 & + \sum_t \gamma_t \mathbf{1}\{\text{year} = t\} + \epsilon_{mt},
 \end{aligned} \tag{2}$$

where `numberofresults` is the number of search results (in millions), `branded` and `long_tail` represent keywords that may be endogenously used by advertisers,  $\mathbf{1}\{\text{market} = m\}$  is a dummy variable that takes the value of one for market  $m$  and zero otherwise, and  $\mathbf{1}\{\text{year} = t\}$  is a dummy variable that takes the value of one for year  $t$  and zero otherwise. Estimate (2) (use the automated TSLS STATA command). (Hint: You need to create the dummy variables for each of the markets identified by the variable `numind` using `tabulate numind, generate(dummies_numind)` and each of the years identified by the variable `year` using `tabulate year, generate(dummies_year)`. Then include them as covariates in your estimation using `dummies*`.) Interpret your results (use a causal interpretation) and explain why the authors include these additional covariates?