README: “Market Structure, Oligopsony Power, and Productivity” by Michael Rubens

Overview

This file documents the programs and datasets used to generate the results in the paper. The programs were run using STATA 17, but should equally work with older versions of STATA. The package ‘estout’ should be installed by the user prior to running the code. The code was last run on a 4-core Intel-based laptop with MacOS version 10.15.7. Using a 2023 computer, the entire code should run within 2 hours.

Data Availability and Provenance Statements

Statement about Rights

- □ I certify that the author of the manuscript has legitimate access to and permission to use the data used in this manuscript.
- □ I certify that the author of the manuscript has documented permission to redistribute/publish the data contained within this replication package.

Summary of Availability

- □ Some data **cannot be made** publicly available.

Details on each Data Source

- NBS ASIF dataset

  The paper uses the Annual Survey of Above-scale Industrial Firms (ASIF), which is collected by the National Bureau of Statistics of China. These data are not available to the general public and can only be accessed by permission of China’s National Bureau of Statistics or their authorized data sellers. Recently, the NBS has also started to make the data, including more recent years, available through a few (regional) data centers. I refer to Brandt, Van Biesebroeck, Wang, and Zhang (AER, 2017) for a description of this dataset and how it is assembled.

- NBS Product-level quantity data

  Similarly to the ASIF, the product-level quantity data are not publicly available and can only be accessed by permission of China’s National Bureau of Statistics or their authorized data sellers. Researchers interested in access to the data may contact the National Bureau of Statistics of China (NBS) through info@stats.gov.cn or address: No. 57, Yuetan Nanjie, Sanlihe, Xicheng District, Beijing 100826.

- Chinese census of population for the year 2000

This dataset is publicly available and was obtained from Harvard Dataverse, doi:10.7910/DVN/VKGEBX. The dataset is merged to the ASIF data using the province and county codes mentioned further below.

- Brand characteristics

  Brand-level cigarette characteristics were collected from the paper "Cigarettes sold in China: design, emissions and metals" (O'Connor. et al., 2010), as cited in the paper. The full dataset and documentation can be downloaded from doi:10.1136/tc.2009.030163.

- Industry-specific deflators and codes

  Industry-level Input and output deflators are from Brandt et al. (AER, 2019).

- Weather data from CMA

  County-level weather data from the Chinese Meteorological Agency (CMA), and can be obtained from https://data.cma.cn/en. The required dataset can be downloaded by clicking on "Home"=>"Data and products" => "Surface data" => "Dataset 10: Annual Values of Climate Data From Chinese Surface Stations for Global Exchange, 1997-2008." This dataset was retrieved by the author on August 7, 2018, for every province in the data, resulting in data files climate1-climate29.txt (one per province). The variables retrieved are V01301-V13007, and are defined in the file 'variablenames_climate.csv', which is a direct translation of the metadata found on the CMA website under "Dataset 10".

- Agricultural price data from FAOSTAT.

  Agricultural price data are obtained from FAO's statistical agency, FAOSTAT through https://www.fao.org/faostat/en/#data, and then "Producer prices", "Download data" with the parameters 'Country = China', "item = unmanufactured tobacco", "Annual value", "Years = all", "Producer price (USD/tonne)". Downloading this results in the dataset FAOSTAT_prices.csv. Similarly, downloading under "Crops and livestock products" using the same parameters as described above, but now for the variable 'Yield', results in the dataset FAOSTAT_quantities.csv.

  The FAOSTAT database is subject to open data licensing, and is hence accessible for non-commercial usage, according to the licensing policy described in the following FAO document https://www.fao.org/3/ca7570en/ca7570en.pdf.

- English translations

  English translations of product names, firm names, and province names in the product-level ASIF data were translated using Google Translate. These files are not public because they are derived from ASIF.

- Retail prices

Chinese cigarette retail prices per year are obtained from the World Bank report 'Cigarette Affordability in China, 2001-2016', which is cited in the paper and can be accessed online through https://documents1.worldbank.org/curated/en/130301492424519317/pdf/114283-REVISED-PUBLIC-China-report-final-may-16-2017.pdf The annual prices are obtained from p.18 in this report, and are in the spreadsheet "retailprices.csv".

- Shapefiles and geographical data

  The shapefiles to make maps are obtained from the China Administrative Boundary Common Operational Database (COD-AB) through https://cod.unocha.org/ => "hds link". The shapefiles can be downloaded under 'chn_adm_ocha_2020_SHP.zip'. To obtain geographical coordinates for the counties in the dataset, the author proceeded as follows. First, each 4-digit zip code in the NBS ASIF dataset (described above) was manually linked to county names using wikipedia (https://en.wikipedia.org/wiki), with all pages used being documented in the file "county_coords.csv". Next, each county name was inserted in google maps (https://www.google.com/maps) in order to obtain a coordinate. Google maps reports the midpoint coordinate of each county. These coordinates were manually inserted in the spreadsheet "county_coords.csv". This dataset is made available by the OCHA Regional Office for Asia and the Pacific (ROAP) and is made publicly available under the Creative Commons Attribution for Intergovernmental Organisations.

Dataset list

- `1998.dta`- `2007.dta`: These are the annual versions of the NBS ASIF dataset from Brandt et al. (2017).
- `master.dta`: This is the merged NBS ASIF dataset from Brandt et al. (2017), across all years.
- `HS-CIC.dta` This is the concordance file between HS product codes and CIC industrial classification codes, from Brandt et al. (2019). A version without duplicates for the CIC codes is generated in the data do-file, and is stored as a temp file `HS-CIC_nodup.dta`.
- `qycp0006.txt - qycp0612.txt`: These files contain monthly output in physical quantities at the firm-product level, in between June 2000 and December 2006. The dataset records current quantity, quantity in the same month in the previous year, a product code and description, and a unit code and description.
- `J11A0101.tab-J65L0814.tab`: census data files from the 2000 China population census.

- `provincecodes.csv`: province codes from the 2000 China population census.
- `countycodes.csv`: county codes from the 2000 China population census.
- `data_characteristics.csv`: brand-level cigarette characteristics from O'Connor et al. (2010). The following files are used to match these data to the firm-level ASIF:
- `data_brandfirm.csv`: lists the firm for every brand in O'Connor et al. (2010).
- `data_concord_firm.csv`: concordance between firm names of brands in O'Connor et al. (2010) and firm names in the ASIF dataset. Not public because contains ASIF information.
- `simple_correction_input_deflators.dta`: Industry-level Input and output deflators from Brandt et al. (AER, 2019).
- `climate1.txt-climate29.txt`: county-level weather data from the Chinese Meteorological Agency (CMA).
- `station_names.csv`: names of weather stations.
- `FAOSTAT_prices.csv`: producer prices of tobacco leaf in China, from FAOSTAT.
- `FAOSTAT_quantities.csv`: agricultural yields for tobacco leaf in China, from FAOSTAT.
- `prodnames.xlsx`: English translations of product names in the product-level ASIF data. Translated using Google Translate. Not public because derived from ASIF.
- `mergers.csv`: English translations of firm names in the firm-level ASIF data. Translated using Google Translate. Not public because derived from ASIF.
- `provincenames_EN.csv`: English translations of province names in the firm-level ASIF data. Translated using Google Translate. Not public because derived from ASIF.
- `retailprices.csv`: annual retail prices for cigarettes.
- `chn_adm_ocha_2020_SHP.zip`: shapefiles for Chinese administrative boundaries, from OCHA.

Computational requirements

Software Requirements

- The programs were run using STATA 17, but should equally work with older versions of STATA. The code was last run on a 4-core Intel-based

laptop with MacOS version 10.15.7. Using a 2023 computer, the entire code should run within 2 hours.
  - The package `estout` should be installed.

Controlled Randomness

- ☐ Random seeds are set at:
  - line 53 of program `china_tobacco_ap_acf_bs.do`
  - line 45 of program `china_tobacco_ap_nestedlogit_bis_bs.do`
  - line 47 of program `china_tobacco_ap_nestedlogit_bs.do`
  - line 68 of program `china_tobacco_ap_robchecks_bs.do`
  - line 51 of program `china_tobacco_ap_substit_bs.do`
  - line 82 of program `china_tobacco_baseline_bs.do`

Memory and Runtime Requirements

Summary Approximate time needed to reproduce the analyses on a standard 2023 desktop machine:

- ☐ 1-2 hours

Details The code was last run on a 4-core Intel-based laptop with MacOS version 10.15.7.

Description of programs/code

- Programs in `main/` generate all tables, figures and results in the main text. The file `main/china_tobacco_master.do` runs all other files.
- Programs in `appendix/` generate all tables, figures and results in the appendix.
- Output files are called appropriate names (`table1.tex`, `Figure4.pdf`) and should be easy to correlate with the manuscript.
- Programs in `bootstrap/` run the code to estimate the bootstrapped standard errors.
- The folder `data/` contains all the data files to run the code
- The folder `tempfiles/` generate all temporary data files that are internally generated by the code.

Instructions to Replicators

- Access and download the confidential data files referenced above. These should be stored in the folders `data/`, in .dta format for the NBS ASIF data and .txt format for the NBS product-level quantity data.
- Edit `/main/master.do` to adjust the default path
- Run `/main/master.do`

Details

- `china_tobacco_master.do`: This is the master do-file that calls all the other programs.
- `china_tobacco_data.do`: Combines the different datasets and contains the data cleaning. This do-file calls the auxiliary program china_tobacco_weatherstation_matching.do, which matches weather stations to counties, in order to merge the weather data into the main dataset.
- `china_tobacco_maps.do`: Generates the maps in Figure 1.
- `china_tobacco_reducedform.do`: Contains the reduced-form analysis. Generates Figure 2, Figure A1, Table 1, Table A3, Table A12.
- `china_tobacco_baseline.do`: Contains the estimation of the model in the main text. Generates Tables 1, 2, and 3, Figure 3, 4, Table A11, Figure A2.
- `china_tobacco_ap_nestedlogit`: Estimates the nested logit model. Generates Table A1.
- `china_tobacco_ap_nestedlogit_bis`: Estimates the nested logit model, but without including the TFP instrument. Generates Table A2.
- `china_tobacco_ap_acf`: Estimates the model using a control function approach. Generates Table A4. This do-file calls four auxiliary do-files to estimate the production function:
- `china_tobacco_acf_endexit_stage1.do`: The first stage of the two-step procedure to estimate ACF, for the endogenous exit specification.
- `china_tobacco_acf_endexit.do`: The second stage of the two-step procedure to estimate ACF, for the endogenous exit specification.
- `china_tobacco_acf_exgexit_stage1.do`: The first stage of the two-step procedure to estimate ACF, for the exogenous exit specification.
- `china_tobacco_acf_exgexit.do`: The second stage of the two-step procedure to estimate ACF, for the exogenous exit specification.
- `china_tobacco_ap_substit`: Estimates the substitutable leaf model. Generates Table A6.
- `china_tobacco_ap_robchecks.do`: All other robustness checks. Generates Table A5, Table A7, Table A8, Table A9, and Table A10.
- The standard errors are bootstrapped in separate do-files, that carry the same name as the file they are providing the bootstrapped standard errors for, but with '_bs' added. They are in ./bootstrap.
- The number of bootstrap iterations is defined as B=200 in `china_tobacco_master.do` on line 10. In each of the bootstrapping do-files, the seeds are set equal to 'b' for each iteration 1<b<B, with 'b' being the number of the bootstrap iteration.

List of tables and programs

With access to the confidential datasets, the provided code reproduces:

☐ All numbers provided in text in the paper

□ All tables and figures in the paper, as summarized below:

Figure/Table		Line		Conf.
#	Program	Number	Output file	data?
Table 1	china_tobacco_baseline.do	381	table1.tex	X
Table 2	china_tobacco_baseline.do	409	table2.tex	X
Table 3	china_tobacco_baseline.do	436	table3.tex	X
Table A1	china_tobacco_ap_nestedlogit.do	471	tableA1.tex	X
Table A2	china_tobacco_ap_nestedlogit_bis.do	492	tableA2.tex	X
Table A3	china_tobacco_reducedform.do	248	tableA3.tex	X
Table A4	china_tobacco_ap_acf.do	263	tableA4.tex	X
Table A5	china_tobacco_ap_robchecks.do	482	tableA5.tex	X
Table A6	china_tobacco_ap_substit.do	179	tableA6.tex	X
Table A7	china_tobacco_ap_robchecks.do	506	tableA7.tex	X
Table A8	china_tobacco_ap_robchecks.do	550	tableA8.tex	X
Table A9	china_tobacco_ap_robchecks.do	586	tableA9.tex	X
Table A10	china_tobacco_ap_robchecks.do	598	tableA10.tex	X
Table A11	china_tobacco_baseline.do	474	tableA11.tex	X
Table A12	china_tobacco_reducedform.do	262	tableA12.tex	X
Figure 1a	not generated using do-files	N.A.	Figure1a.pdf	X
Figure 1b	china_tobacco_maps.do	42	Figure1b.pdf	X
Figure 1c	china_tobacco_maps.do	46	Figure1c.pdf	X
Figure 2a	china_tobacco_reducedform.do	138	Figure2a.pdf	X
Figure 2b	china_tobacco_reducedform.do	114	Figure2b.pdf	X
Figure 3a	china_tobacco_baseline.do	89	Figure3a.pdf	X
Figure 3b	china_tobacco_baseline.do	92	Figure3b.pdf	X
Figure 4	china_tobacco_baseline.do	568	Figure4.pdf	X
Figure A1	china_tobacco_reducedform.do	304	FigureA1.pdf	X
Fig. A2a	china_tobacco_baseline.do	285	FigureA2a.pdf	X
Fig. A2b	china_tobacco_baseline.do	293	FigureA2b.pdf	X