README DATA/CODE DOCUMENTATION for "Market Structure and Cost Pass-Through in Retail"

Nicholas Li and Gee Hee Hong

August 25, 2015

1 Data

1.1 Single retailer data

The data are proprietary but are accessible to many academics. Access is managed through the SIEPR-Giannini Data Center through the following web-site: http://are.berkeley.edu/SGDC/. According to the web-site, the data are currently available to researchers at Stanford, the entire UC system, Cal State Sacramento, Northwestern, Princeton, and Washington State. For a partial list of publications using these data, see the web-site and references in the paper. For additional information contact: Professor Sofia Berto Villas-Boas at UC Berkeley (sberto@berkeley.edu). Note that there are several data sets available under the agreement, and we use the one for Canada and the US that covers approx. 200 categories over the 2004-mid 2007 period.

Description

"upc_us_all_2004.dta, upc_us_all_2005.dta, upc_us_all_2006.dta, upc_us_all_2007.dta"

These are the original weekly transaction files sent to us in text format, which we read into STATA and appended across weeks within each calendar year. We use the following variables:

- store_id: Unique store identifier

- upc_id: UPC code
- promo_week_id: year followed by week, e.g. 200451
- avt_cost: average wholesale cost
- gross_amt: gross revenue
- net_amt: net revenue including all member discounts, bulk discounts, rebates, etc.
- item_qty: number of physical units
- meas_qty: weighed quantity, non-zero for some goods (e.g. produce, bulk items), we drop these goods

"lu_upc.dta"

This is the original UPC description file that was sent to us in text format by the retailer, which we read into STATA format. There are many variables, the ones we use are:

- upc_id: UPC code
- upc_dsc: Text description of the product
- internet_item_dsc: Alternative text description of the product
- corporation_id: identifies US (=1) or Canadian (=20) stores (sometimes the same UPC has different descriptions in the two countries)
- group_id, category_id, subclass_id: these are the categorization variables we use, in descending order of aggregation
- last_update_ts and nobs: we only use these as an alternative (anonymized) unique identifier for products; these allow an authorized user of the data to link our private label classification to UPCs, and hence to the price and quantity data.

"pl_data.dta" (private label information)

As we are not allowed to disclose the retailer's identity, we cannot provide the code and web-site we use to identify retailer branded and manufactured private labels to interested

researchers not authorized to use the data.

However we include a file that allows for unique matching of private label product attributes based on two variables found in the lu_upc.dta data set ("last_update_ts" and "nobs"). The file contains

- pl: =1 for any of the retailer's private label brands, =0 otherwise
- last_update_ts and nobs: we only use these as an alternative (anonymized) unique identifier for products; these allow an authorized user of the data to link our private label classification to UPCs, and hence to the price and quantity data.

Interested researchers who have secured access to the data can contact Nicholas Li for the program we used to search text strings (and use first-five digits of UPC codes) to identify private labels, and details of the web-site we used to identify products manufactured by the retailer's manufacturing affiliate.

"lu_store.dta"

This is the original store description file that was sent to us in text format by the retailer, converted to STATA format. There are many variables, the ones we use are:

- store_id: Unique store identifier
- corporation_id: identifies US (=1) or Canadian (=20) stores
- store_state_id: the state in which the store operates – some of the stores with very different prices (Alaska, Hawaii) are technically part of a mainland operating area, we exclude these stores.
- op_area_id: Operating area (called divisions in the paper) for the retailer, the data are organized into 10 US and 3 Canadian operating areas of 25 stores each. The main sample we use corresponds to the sixth operating area we label within the US (Northern California).

1.2 Multi-retailer data

The multi-retailer data comes from the Symphony IRI marketing research company. The academic use data set is available for purchase by academics at reason-

able rates. See web-site and description at http://www.iriworldwide.com/solutions/Academic Data-Set and further discussion of the data set in Bronnenberg, Bart, Michael Krueger, and Carl Mela, "Database Paper: The IRI Marketing Data Set," Marketing Science, 2008, 27(4), 745748.

"category_groc_old.dta"

This is the weekly store-level data for all markets between 2001 and 2007, for each category (i.e. category=milk or spagsauc). We read the annual files into STATA and appended them to create this file. note that we will drop the year 2007 because of changes in the stub codes and recording of private labels in early 2007 (see IRI Symphony documentation).

We use the following variables:

- SY: UPC system code. Note that system code 88 corresponds to private labels (whose VEND codes do not identify manufacturers or retailers) which is how we identify them in the data.
- VEND: UPC vendor code (used to identify brands – first five digits)
- GE: UPC generation code (for UPCs with multiple versions)
- ITEM: UPC item code (last five digits)
- DOLLARS: total weekly sales in dollars
- UNITS: total weekly units sold
- IRI_KEY: This is the masked store number, allows for merging with the delivery stores files.
- WEEK: The IRI week classification.

"delivery_stores_yr11.xlsx"

This is an excel file provided by IRI Symphony that allows unique store identifiers to be linked to chain identifiers and markets.

- IRI_KEY: This is the masked store number and can be merged with the IRI_KEY variable in the weekly store scanner data.
- MskdName: this is the masked retail chain identifier provided by IRI

- Market_Name: this corresponds to the IRI classification of markets

"weekmonth.dta" This STATA file is a concordance from weeks to months that we created, based on assigning a value of 1 to the first week in the IRI data (week 1 2001) all the way up to the maximum weeks in the data (313, the last week in 2006). January 2001 corresponds to month 1 and December 2006 corresponds to month 72.

1.3 Commodity prices

We use monthly commodity price data from three sources:

1. Food and Agricultural Organization: We use the monthly average prices for Maize (US, No.2 Yellow) and Wheat(US, No.2 Hard Red Winter) from http://www.fao.org/economic/est/prices
2. Food and Agricultural Organization: We use the monthly dairy price index from http://www.fao.org/worldfoodsituation/foodpricesindex/en/
3. Department of Labor and BLS: Fresh tomatoes monthly producer price index 1947-2010, available at: http://www.economagic.com/em-cgi/data.exe/blswp/WPU01130217

We include the price series for each commodity in STATA format with the month codes matched to the ones in the scanner data sets.

- com_single.dta: this has the tomato, dairy, wheat and corn prices with months corresponding to the months in the single-retailer data (i.e. month=1 corresponds to January 2004). It also contains 12 monthly lags for each commodity price (to facilitate creating distributed lags of monthly differences and allow for use of data in months prior to the beginning of the retailer data).
- tomato_index.dta, dairy_index.dta, corn_index.dta: these are tomato, dairy and wheat prices with months corresponding to the months in the multi-retailer data (i.e. month=1 corresponds to January 2001).

2 Programs/Code

Note: all code was run using STATA MP Version 13.1.

"clean_data_final.do" This program converts the raw weekly-store level data (already in STATA format) into monthly, operating area average data. It also drops products with no fixed weights (non-zero values of "meas_qty").

"newprog_final.do" This program can be run after clean_data_final.do, and creates Table 1 Panel A, Figure 1, Table 2, Table 3 and Table 4.

"newprog_twostep_final.do" This program can be run after clean_data_final.do, and creates Table 5 (two-step estimation).

"newprog_conditional_final.do" This program can be run after clean_data_final.do, and creates Table 6 (conditional pass-through estimation).

"prep_and_one_step.do" This stand-alone program prepares the IRI Symphony data for our analysis and creates Figure 2, Web appendix Figure ', outputs the summary statistics used in Table 1 Panel B, and runs the regressions necessary for Table 7.