Promoting the use of a publicly available scanner data set in price index research and for capacity building

Jens Mehrhoff*, Deutsche Bundesbank

This version: 25 September 2018
(Updated: 15 September 2019)

Abstract

The present paper demonstrates how a publicly available scanner data set can be used for price index research and capacity building. Discounting the incomparability of different data sets would add to transparency in this evolving strand of the literature. Likewise, training courses could be run even before such data become available to statistical offices.

The paper continues by estimating price index numbers using the weighted time-product dummy method which accounts for product churn and overcomes chain drift. It further derives an indicator of product heterogeneity from the results. Last, the December splice is introduced for calculating non-revisable indices in real time with a view to the European HICP.

Keywords: Product churn; Chain drift; Time-product dummy; December splice.

JEL: C23; C43; C82; E31.

*This paper represents the author's personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or its staff. Detailed results are available on request from the author. E-mail: jens.mehrhoff@bundesbank.de

Introduction

Scanner data from Dominick's Finer Foods – a now-defunct Chicago-area grocery store chain – is publicly available for download from the website of the James M. Kilts Center at the University of Chicago Booth School of Business: https://www.chicagobooth.edu/research/kilts/datasets/dominicks. The data is provided for academic research purposes only and all users should acknowledge this source in any working paper or publication that uses any portion of this data.

The data set covers 93 stores for 399 weeks from 14 September 1989 to 7 May 1997 and totals 98,884,285 observations (after cleansing) of 13,845 products (excluding re-launches) in 29 categories (from analgesics to toothpastes). Though the data set is historic in nature (and has some gaps), it shows all the specificities one finds in modern electronic transactions data, particularly a dynamic universe due to new and disappearing products (i.e. churn). Hence, this data set is most useful for both research and training purposes.

Currently, research results are largely derived from information that cannot be shared for data protection reasons. This, however, seriously hampers the academic discourse and rules out reproduction of the results. With the use of the Dominick's data set, the current discussions about multilateral methods, aggregation / classification, economic approaches and the like could be based on the very same figures, thus adding to transparency in this evolving strand of the literature.

Likewise, statistical offices could start running training courses on the basis of data that provides the same challenges they would face with actual data. A major obstacle is establishing the data flows from retailers and only a few countries globally have succeeded in this so far. Capacity building, on the other hand, could already start before such data becomes available. This appears to be particularly important in this field as scanner data is considerably more complex.

In what follows, we first elaborate on how to acquire and pre-process the data. We then turn to some peculiarities of the data set that need to be considered in the analysis. Finally, we present an estimation of price index numbers using the weighted time-product dummy method. We conclude by showing the impact of different splicing approaches, also introducing the *December splice* in the framework of the European Harmonised Index of Consumer Prices (HICP).

Acquiring the data

The Dominick's database contains category-specific files and general files (customer count and store-level demographics) in SAS format. For the exposition here, it is necessary to download all category-specific files, i.e. the UPC files and movement files (58 files in ZIP format for the 29 categories).

The UPC (Universal Product Code) files contain information about Dominick's commodity code, Dominick's item code, etc. of each UPC in a category. The movement files contain sales information at the store level on a weekly basis for each UPC in a category. The variables included in these files comprise: total dollar sales, number of units sold, sale code, etc.

Unfortunately, the UPC file for refrigerated juices ('RFJ') is not in SAS format. We provide a CSV file that contains the information in a SAS readable format. Furthermore, we provide two CSV files that prepare the information on the week variable and the stores included that was covered only in *Dominick's Data Manual*. All three files and subsequently introduced SAS codes are available for download from the article's website: https://github.com/eurostat/dff.

Pre-processing the data

The first part of SAS code ('upc.sas') reads in all UPC files and adds a category identifier. The SAS data set generated ('upc.sas7bdat') has 18,345 observations (including re-launches) and 4 variables, hierarchically organised as shown below.

Variable	Description
CAT	Category
COM_CODE	Commodity code (Dominick's version of categories)
NITEM	Item code (attempt at tracking products across multiple UPCs)
UPC	Universal Product Code

The second part of SAS code ('move.sas') reads in all movement files, adds a category identifier, and calculates total dollar sales (SALES) from the retail price (PRICE), the number of units sold (MOVE) and the number of items bundled together (QTY). The Dominick's database includes a flag set by the James M. Kilts Center to indicate data that is suspect. We do not use flagged data in our analysis nor non-positive prices. The SAS data set generated ('move0.sas7bdat') has 98,884,285 observations (after dropping suspect data) and 7 variables.

Variable	Description
CAT	Category
WEEK	Week variable
STORE	Store number
UPC	Universal Product Code
MOVE	Number of units sold
SALES	Total dollar sales (SALES = PRICE $\times$ MOVE $\div$ QTY)
SALE	Sale code (indicates whether the product was sold on a promotion)

The third and last part of SAS code ('weeks_stores.sas') reads in the week and store files and merges them with the movement and UPC files. The SAS data set generated ('move.sas7bdat') still has 98,884,285 observations but has now 5 additional variables over and above of the 9 variables in total before.

Variable	Description
MONTH	Month in which the week ends (weeks run from Thu. to Wed.)
MONTH1	MONTH, if the week starts and ends in the same month
WEEK1	Consecutive week number within each MONTH1
PRICE_TIER	Price tiers (Cub-Fighter, low, medium, and high)
ZONE	Dominick's pricing zones (16) with uniform regular prices

Peculiarities of the data

For some observations in the movement files, the corresponding observation in the store or UPC files is missing – in this case missing (or incorrect) values were replaced with dummies. The variables affected by this are PRICE_TIER (dummy: 'N/A') and ZONE ('0') from the store file, and NITEM (UPC) and COM_CODE ('999') from the UPC files. In particular for refrigerated juices ('RFJ') there was no information on NITEMs (and COM_CODEs), which in effect leads to every UPC being treated as an NITEM (i.e. re-launches cannot be identified).

Although a single commodity code (COM_CODE) cannot be present in more than one UPC file, due to the above imputations, an individual product should always be defined as the unique combination of CAT $\times$ COM_CODE $\times$ NITEM (or UPC).

It should be noted that the item code (NITEM) scheme is not fool proof. There are many instances where two UPCs have different NITEM codes although they are the same products. The same is true for the sale code (SALE), which we replaced with a dummy; this variable is not set by Dominick's on a consistent basis. If the variable is set it indicates a promotion, if it is not set, there might still be a promotion that week.

Last, there is no (trustworthy) data for week 219 for any of the categories. But there are more gaps for the different categories, with most of them at the beginning, although there are some gaps over the course of time in several categories. The table below shows for each category the number of weeks covered and the first week of data (all categories cover data until week 399); the average and minimum number of products (NITEMs, excluding missing weeks); and the within-category average duration of products in weeks and as a percentage of non-missing weeks (the overall average duration in the sample is 138 weeks). It is noteworthy that a mere quarter of a per cent of products are available in 'all' 398 weeks, while almost one-and-a-half per cent are available in a single week only.

No.	Category ('CAT')	Weeks (First)	Products (Min)	Duration (%)
1	Analgesics ('ANA')	392 (1)	256 (212)	199 (51)
2	Bath soap ('BAT')	265 (128)	137 (53)	76 (29)
3	Beer ('BER')	302 (91)	257 (135)	118 (39)
4	Bottled juice ('BJC')	392 (1)	172 (112)	153 (39)
5	Cereals ('CER')	366 (1)	194 (175)	178 (49)
6	Cheese ('CHE')	392 (1)	261 (208)	178 (45)
7	Cigarettes ('CIG')	398 (1)	18 (11)	104 (26)
8	Cookies ('COO')	388 (1)	354 (299)	140 (36)
9	Crackers ('CRA')	380 (1)	118 (50)	149 (39)
10	Canned soup ('CSO')	378 (1)	224 (93)	204 (54)
11	Dish detergent ('DID')	392 (1)	84 (68)	174 (44)
12	Front-end candies ('FEC')	396 (1)	175 (141)	167 (42)
13	Frozen dinners ('FRD')	255 (142)	102 (25)	110 (43)
14	Frozen entrees ('FRE')	396 (1)	297 (187)	143 (36)
15	Frozen juices ('FRJ')	396 (1)	83 (65)	203 (51)
16	Fabric softener ('FSF')	396 (1)	89 (62)	167 (42)
17	Grooming products ('GRO')	271 (128)	377 (284)	114 (42)
18	Laundry detergents ('LND')	396 (1)	132 (95)	145 (37)
19	Oatmeal ('OAT')	305 (91)	52 (47)	169 (55)
20	Paper towels ('PTW')	388 (1)	34 (16)	145 (37)
21	Refrigerated juices ('RFJ')	396 (1)	85 (37)	150 (38)
22	Soft drinks ('SDR')	390 (1)	496 (283)	141 (36)
23	Shampoos ('SHA')	215 (128)	891 (516)	85 (40)
24	Snack crackers ('SNA')	384 (1)	149 (127)	149 (39)
25	Soap ('SOA')	275 (122)	113 (76)	126 (46)
26	Toothbrushes ('TBR')	388 (1)	138 (97)	140 (36)
27	Canned tuna ('TNA')	374 (1)	106 (50)	188 (50)
28	Toothpastes ('TPA')	398 (1)	163 (131)	173 (43)
29	Toilet papers ('TTI')	384 (1)	42 (29)	158 (41)

Estimation of price index numbers

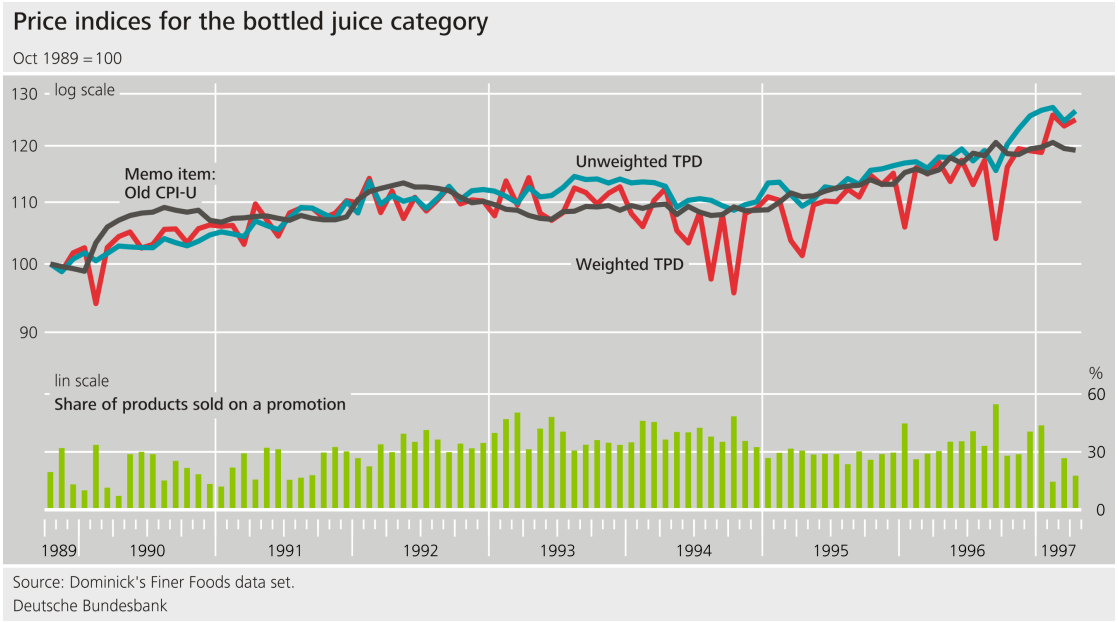
For the sake of exposition the example SAS code ('wtpd.sas') aggregates the weekly store-level UPC data to chain-wide item codes at monthly frequency using the month in which the week ends, covering the period from October 1989 to April 1997 (91 months; 467,605 observations of 13,818 products). It calculates further unit prices and their natural logs (LOGPRICE), monthly expenditure shares per category (SHARE), and creates an index for months (MONTH2) as well as products (NITEM2) to be used in the weighted time-product dummy regression.

Price indices are derived by means of the GLM (general linear models) procedure in SAS since this is more convenient than simple linear regression with a lot of dummy variables. The log-linear model

$$\ln p_i^t = \alpha + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$$

is estimated for each category by weighted least squares, where the expenditure shares serve as weights. The time dummies δ^t and the product dummies γ_i are extracted from the output.

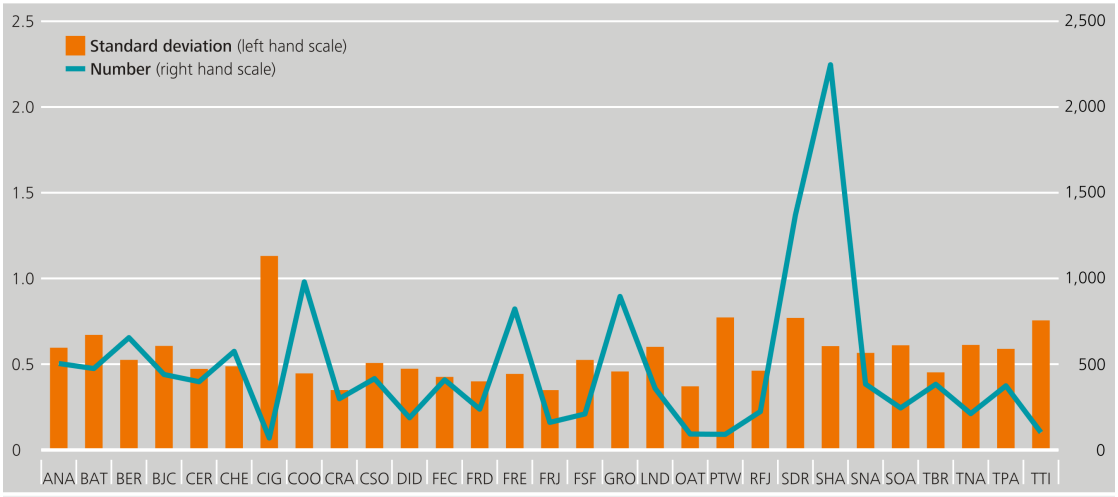
The price index is derived as the exponentiated time dummy, i.e. $P^t = \exp \delta^t$ and $P^{t=0} = 1$. The standard deviation of the estimated product dummies serves as an indicator of product heterogeneity. Below we present the results for bottled juice ('BJC') as an example and contrast the heterogeneity measure across all 29 categories. Further, we exemplify the product churn in the bottled juice category by showing the availability of item codes over time.



As it can be seen from a comparison of the weighted and the unweighted time-product dummy (TPD) index[1], where the latter is less affected by quantity increases due to price decreases – very much like web-scraped data –, the troughs are highly correlated to sale periods.

[1] According to information from the U.S. Bureau of Labor Statistics, the old series of the CPI-All Urban Consumers (CPI-U) for fresh, canned or bottled fruit juices is a good match to the bottled juice category in the Dominick's data set for dates prior to 1998.

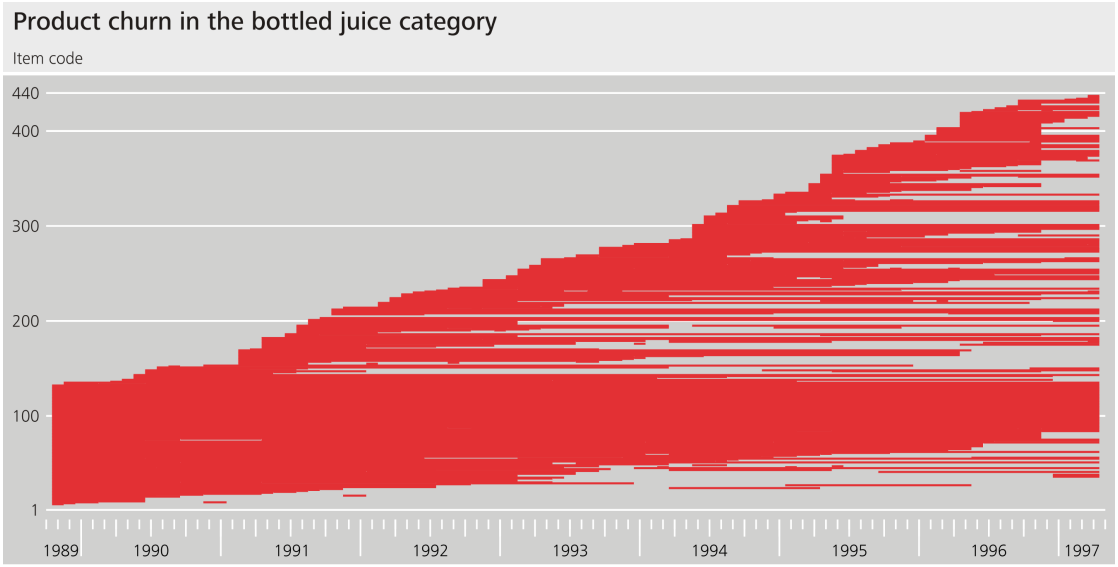
Standard deviation of product dummies and number of products by category



Source: Dominick's Finer Foods data set.
Deutsche Bundesbank

As to the standard deviation (SD) of product dummies, cigarettes ('CIG') stand out. This is based on the broad heterogeneity within this rather small category as well as considerable churn and is not the result of a few individual outliers.

Product churn in the bottled juice category



Source: Dominick's Finer Foods data set.
Deutsche Bundesbank

The within-category average duration of bottled juice products in the sample is 37 months (average over all categories: 34 months). Notably, more than 10% (5%) of products are available throughout the 91 months studied. On the other hand, 3½% (2½%) of products are available in just one month. Compared to the first month, half of the initial products are still sold after 82 months.

Impact of different splicing approaches

As one of the many possible examples how this data set can be exploited, we demonstrate the impact of different splicing approaches to calculate non-revisable indices in real time. We consider a rolling estimation window of 14 months length in order to allow for moving seasonality. When the estimation window is shifted one month forward, revisions occur to previous estimates. To avoid this problem, the new index is spliced onto a previous value (dependent on the particular approach) and revised indices are ignored. This ensures the full information contained in the estimation window is sufficiently exploited.

The three approaches covered are:

- Movement splice: $CP^{m,t} = CP^{m-1,t} \times P^{m-1,t \to m,t}$, i.e. the chained index in month m of year t is the previous month's chained index times the index link from the previous to the current month.
- Window splice: $CP^{m,t} = CP^{m,t-1} \times P^{m,t-1 \to m,t}$, i.e. the current chained index is the previous year's chained index of the same month times the index link from the previous to the current year (of the same month). Strictly speaking, this would be true only if the estimation window would have been 13 months long but this specification makes the window splice approach comparable to the *over-the-year technique*.
- December splice (Mehrhoff, 2017): $CP^{m,t} = CP^{12,t-1} \times P^{12,t-1 \to m,t}$, i.e. the current chained index is the previous December's chained index times the index link from the previous December to the current month. This approach is fully in line with the annual chain-linking principle of the European HICP and consistent with the *one-quarter (month) overlap technique*.

For simplicity, we restrict the analysis to those 17 categories that have full coverage in all 91 months of the data set. The first month in which splicing was applied is December 1990 (i.e. the first 14 month are identical for all three approaches; in fact, since our specification of the window splice coincides with the December splice in December of each year, for these two approaches the first 15 months are identical).

For each of the three approaches we calculate two measures of chain drift vis-à-vis the index estimated from the full data set: root mean square percentage error (RMSPE) and maximum absolute percentage error (maxAPE). The first month is excluded from this calculation since all indices are set equal to 100 in October 1989, i.e. the error would be zero for all three approaches.

		Movement splice		Window splice		December splice	
No.	CAT	RMSPE	maxAPE	RMSPE	maxAPE	RMSPE	maxAPE
1	ANA	1.01	2.05	0.47	1.82	0.44	1.41
4	BJC	1.04	3.89	1.53	4.41	0.95	3.34
6	CHE	1.15	9.68	1.31	8.02	1.81	6.94
7	CIG	0.73	1.52	0.53	1.08	0.27	0.83
10	CSO	1.53	2.65	1.61	3.83	1.76	2.77
11	DID	1.13	2.82	1.48	4.17	1.27	3.37
12	FEC	0.67	1.98	1.02	3.05	0.90	2.25
14	FRE	2.37	5.49	2.20	5.51	3.29	6.83
15	FRJ	2.16	4.13	3.72	8.62	3.16	5.57
16	FSF	5.51	10.36	4.44	8.89	4.77	9.43
18	LND	3.53	7.70	3.12	7.07	2.96	6.81
20	PTW	1.03	3.79	2.21	7.62	2.03	7.66
21	RFJ	2.49	4.89	1.31	3.77	1.45	3.48
26	TBR	6.64	11.78	5.22	8.01	5.14	8.68
27	TNA	3.13	16.86	2.92	9.44	3.43	13.22
28	TPA	1.79	4.00	2.65	6.28	1.58	4.04
29	TTI	3.40	8.19	3.60	13.02	3.30	9.74
	Average	2.31	5.99	2.31	6.15	2.26	5.67

There is some evidence of chain drift in all of the categories. The extent to which splicing re-introduces chain drift in an otherwise drift-free index obviously greatly depends on the particular category.

Although the December splice has been suggested on theoretical grounds – in the same way the overall HICP is calculated for reasons of consistency and explainability – rather than as yet another approach which comes closer to the benchmark index with one or the other data set, it performs remarkably well compared to both the movement and window splice.

For the HICP it is conceptually unquestionable that the splicing of multilateral methods should be performed using December of the previous year as link in every month. This anchoring mitigates path dependency of the index, while not resulting in revisions to published figures. Hence for other consumer price indices (CPIs), too, it might be better to address this question by looking at the way the overall index is calculated.

Discussion and outlook

The Dominick's data set is unique for the breadth of its coverage and for the information available on retail prices. Although it is rather old and has some issues that remain after dropping suspect data (this is true for more modern scanner data, too!), it can prove very useful in price index research and for capacity building.

Yet, its range of applications is not confined to theoretical questions. Quite the contrary, with this data set it is possible to share and compare scanner data methods globally, including institutions that do not have access to such data sets. This would eventually allow for a more effective monitoring of production methods even before their introduction, discounting the incomparability of different data sets.

In the same way, web-scraping methods can be harmonised. Web-scraped data can be simulated by dropping the information on quantities. In fact, since this information is available in the data set, approximate weighting schemes can be explored rather than relying on unweighted indices at the lowest level of aggregation for this data source.

The methodological scope of the Dominick's data set is also not limited to consumer prices alone. If, for example, Dominick's pricing zones are considered as countries, this data set is applicable to purchasing power comparisons as well.

Finally, the UPC files can be used as a labelled data set in order to train and test supervised machine learning techniques for automatic product classification. In addition to information on categories, product codes (UPCs), descriptions (i.e. text) and other metadata (e.g. size) are available. Also, re-launches can be identified (record linkage).

References

Australian Bureau of Statistics (2016), *Making Greater Use of Transactions Data to Compile the Consumer Price Index, Australia*, Information Paper 6401.0.60.003.

de Haan, J., Willenborg, L., Chessa A.G. (2016), *An Overview of Price Index Methods for Scanner Data*, Thirteenth Session of the Group of Experts on Consumer Price Indices, Geneva: United Nations Economic Commission for Europe.

Kilts Center for Marketing (2013, updated 2018), *Dominick's Data Manual*, Chicago.

Mehrhoff, J. (2017), *Multilateral Methods*, Luxembourg (mimeo).

Appendix

This appendix explains the various data sets generated in the process flow by the SAS codes introduced in the paper. All codes are kept deliberately simple by abstaining from using more efficient facilities such as macros or loops.

All codes assume the SAS and CSV files are stored in C:\Data (Windows operating environment) whenever the 'DFF' library is (de)assigned and CSV files are read. Hence, this would need to be adapted should the location be different. Below we indicate how often this would be the case for each of the SAS codes.

Last, in order to clear the 'WORK' library, it is recommendable to restart SAS after generating the data sets 'upc.sas7bdat' and 'move0.sas7bdat', i.e. before trying to generate the data set 'move.sas7bdat' and running the example code 'wtpd.sas'.

upc.sas

The 'DFF' library is assigned twice and one CSV file is read for refrigatered juices ('RFJ'); see Acquiring the data. Further note that the UPC file for paper towels ('PTW') includes an undocumented variable WSTART which has been dropped.

Data set	Description
WORK.upc	Master data set for all 29 categories
WORK.upc1	Auxiliary data set to be appended (overwritten)
DFF.upc	Exported data set 'upc.sas7bdat'

move.sas

The 'DFF' library is assigned twice.

Data set	Description
WORK.move	Master data set for all 29 categories
WORK.move1	Auxiliary data set to be appended (overwritten)
WORK.move0	Dropped suspect data
DFF.move0	Exported data set 'move0.sas7bdat'

weeks_stores.sas

The 'DFF' library is assigned once and two CSV files are read for weeks and stores; see Acquiring the data.

Data set	Description
WORK.weeks	Read 'weeks.csv' (added MONTH, MONTH1, WEEK1)
WORK.stores	Read 'stores.csv'
WORK.upc	Read 'upc.sas7bdat'
WORK.move0	Read 'move0.sas7bdat'
WORK.move_weeks	Merged WORK.move0 with WORK.weeks
..._weeks_stores	Merged WORK.move_weeks with WORK.stores
..._weeks_stores_upc	Merged WORK.move_weeks_stores with WORK.upc
DFF.MOVE	Replaced missing values with dummies, and exported data set 'move.sas7bdat'

wtpd.sas

The 'DFF' library is assigned once and one CSV file is written if the export procedure is uncommented. The data set could then be used to analyse it with another statistical software package.

Data set	Description
WORK.move_monthly0	Auxiliary data set (aggregated DFF.move)
WORK.move_monthly	Dropped incomplete months (added LOGPRICE, SHARE, MONTH2, NITEM2)
WORK.move_monthly1	Auxiliary data set to be merged (SHARE)
WORK.wtpd	Weighted time-product dummy parameter estimates
WORK.wtpd1	Exponentiated time dummies by category, where $P^{t=91} = 1$
WORK.wtpd0	Auxiliary data set (product dummies by item code, where $\gamma_{i=N} = 0$)
WORK.wtpd2	Standard deviation (unbiased) of product dummies and number of products by category