11

Price Index Estimation Using Price Imputation for Unsold Items

Ralph Bradley

11.1 Introduction

Although scanner price quotes and expenditures have great promise in improving the Consumer Price Index (CPI), their use has introduced new problems. One major problem confronting the Bureau of Labor Statistics (BLS) is the treatment of an item that experiences no purchases during a particular time period. Either a price can be explicitly imputed for this item or it can be implicitly imputed by ignoring it in the construction of the price index. This paper examines and compares different imputation methods. Indexes are then constructed using these various imputation methods using price and expenditure data from scanner sales of cereal in the New York area.

In scanner databases, when a particular item in a particular outlet does not sell in a certain time period, it is not possible to determine if this non-sale was the result of no inventories or of no consumer demand for the product, perhaps because the price was too high. Currently, under the BLS manual sampling system, if an item is still on the outlet shelf and experiences no purchase, the "list" price displayed on the shelf is sampled because the data collector can still observe the price on the shelf. In the scanner databases, when an item is not sold, it is not even possible to get a "list" price. Therefore, missing prices can be a result of at least one of two events in scanner data, but only one event in the current manual sampling system of the CPI.

Ralph Bradley is an economist at the Price and Index Number Research Division in the U.S. Bureau of Labor Statistics.

The author is grateful to Eva Sierminska for assistance and to David Richardson, Robert C. Feenstra, Eduardo Ley, and Matthew Shapiro for useful comments. The views expressed in this paper are solely those of the author and do not reflect either the policies or procedures of the U.S. Bureau of Labor Statistics.

Imputation of missing prices can be done implicitly or explicitly. Implicit imputation occurs when the missing item is ignored.[1] For example, similar items can be grouped together in such a way that the groups are large enough so that there is always at least one item sold in each group. One would calculate unit values of only the sold items in the groups and then compute a price index using the unit values. The reason that this is an implicit imputation method is that there is an *implied* imputation of the prices of the nonpurchased item with its group unit value. Explicit imputation involves the direct replacement of a missing price with an estimated price.

There is a rich literature on imputation of missing data. Little and Rubin (1987) provide a thorough method for imputing random variables that are "missing at random." Under these conditions, the probability that a random variable is missing is independent of the random variable itself, although one could use other exogenous variables to generate a replacement random variable whose distribution well approximates the distribution of the missing variable. Unfortunately, it is not possible to assume that the probability of a missing price is independent of its level. Therefore, an unbiased statistical imputation of prices needs to account for these selection effects.

Armknecht and Maitland-Smith (1999) and Feenstra and Diewert (2000) discuss alternative imputation methods for missing prices in the construction of price indexes. In their studies, missing prices are not necessarily the result of a nonsale. Seasonality, erratic reporting, and replacement with newer models are cited as possible causes. Feenstra and Diewert evaluate the alternatives in their study by their ability to both minimize the erratic movement in the price index and still incorporate all available information. This contrasts with the goal of Little and Rubin (1987), for whom the replacement variable should have a statistical distribution that closely approximates the distribution of the missing variable. The methods in this paper include the unit value approach, the carry forward approach, and the current BLS approach, as well as an economic approach that uses a combination of micro-theory and the methods of Little and Rubin. The reason that these methods are studied is that some are easy to implement but do not estimate the welfare effects of nonsales, whereas the economic approach is more difficult to implement and can be prone to specification and measurement error. Using the database that currently generates the BLS scanner cereal index for New York, I generate indexes for each of these methods and compare the results of the easier methods to the difficult ones. All of the methods except for the unit value approach produce indexes that are close in magnitude even though their pairwise differences are statistically significant.

1. Armknecht and Maitland-Smith (1999) discuss implicit and explicit imputation in great detail.

This study is organized as follows. Section 11.2 describes the various imputation methods that will be examined in this paper. Section 11.3 describes the cereal scanner data set, and finally section 11.4 describes the results of computing price indexes for cereal in New York using the alternative indexes imputation methods described in section 11.2.

11.2 Imputation Methods

11.2.1 Unit Values

Perhaps the easiest method from a computational standpoint is grouping items so that at least one item within each group is purchased and then calculating a unit value for each group. At a second-stage level, the "all-items" price index is computed from the group unit values. This grouping does not necessarily need to be across items within a time period, but can group across time periods. Hausman (1996) uses the unit value approach to get his elementary prices.

Suppose that there are G groups of goods. Then the unit value, UV_g, for the gth group with N_g items is

$$\mathrm{UV}_g = \frac{\sum_{i=1}^{N_g} p_i q_i}{\sum_{i=1}^{N_g} q_i},$$

where p_i and q_i are, respectively, the price and quantity sold of item i. Since this is a quantity-weighted average, UV_g is not a sufficient statistic if q_i is also a function of p_i.

Although computationally simple, using unit values is still controversial. Diewert (1995) recommends using unit values, yet he does not show that the unit value index will "closely" approximate a true price index. In response to Diewert's article, Balk (1998) investigates the sufficient conditions for an index using unit values to be an appropriate price index. One of three independent criteria must be satisfied: (a) there is no variance in price within the group; (b) all the products within the group are perfect substitutes; (c) the group has a Leontieff cost function. However, when none of these conditions is satisfied, it is not clear how closely the unit value approximates the true price index, because these conditions are sufficient but not necessary.

Since unit values implicitly impute the missing prices, if items within a group are not perfect substitutes or complements but are differentiated, then the imputed price is based on quality characteristics that are not necessarily embodied in the product.

11.2.2 Bureau of Labor Statistics Method

Currently, when BLS collects its monthly sample of prices and an item in the sample is missing, the agency does a combined implicit and explicit im-

putation of the missing price. In the first month that an item is missing, the price of the missing item is ignored and the resulting price index is calculated by ignoring the item. If the item continues to be missing after the first month, then the BLS staff selects an item that is similar to the missing one and is available for sale and uses its price to impute a price for the missing item.

To describe this imputation adjustment fully, I give a simple example. Suppose in period s, item h and item i are available for sale, and both items have similar characteristics. If item h disappears in period $s + 1$, then the month-to-month index is computed by dropping the price of the missing item. This is an implicit imputation, in which the imputed price of the missing item is merely the previous price times the month-to-month index.

If the item is still missing in $s + 2$, the imputed prices, $\hat{p}_h^{s+1}$ and $\hat{p}_h^{s+2}$ for the index from period $s + 1$ to $s + 2$, are

$$\hat{p}_h^{s+1} = \hat{p}_h^{s+1} \frac{p_h^s}{p_i^s},$$

$$\hat{p}_h^{s+2} = \hat{p}_h^{s+2} \frac{p_h^s}{p_i^s}.$$

This is an explicit imputation. However, it is based on the implicit assumption that the consumer will buy this replacement item and that there is no welfare loss from the disappearance of the item.

As mentioned previously, if an item is available for sale in an outlet, its list price is still used in the index. However, one cannot observe the list price of unsold items in the scanner data set.

11.2.3 Carry-Forward Imputation

Missing prices can be explicitly imputed by "carrying forward" the last recorded price. Like the unit value, this is computationally simple, and it does not require the grouping of items as in the unit value approach. As Armknecht and Maitland-Smith (1999) point out, this method can produce abrupt changes in the index when the item reappears. Additionally, if the list or market price does not equal the imputed carry-forward price, then the index could be biased.

However, the carry-forward approach has certain advantages over the unit value approach. The explicit imputation is done with the price of the same item and therefore with the same quality characteristic. However, if the time period, itself, is an important characteristic, then this imputation approach suffers the same disadvantage of the unit value approach since the time characteristic embodied in the price differs from the true time characteristic.

It should be self-evident that if nonsold items were offered at a price greater than the carry-forward price, then using the carry-forward price

could generate a bias in the price index. However, in these cases the carry-forward price could be greater than the price that makes the quantity demanded exactly equal to zero. For example, if there is a deeply discounted sale one week and consumers buy enough to supply themselves for over a week, in the second week there could be no sales at the deeply discounted price because consumers are saturated with a large inventory.[2]

11.2.4 An Economic Approach to Imputation

Unlike the methods of Little and Rubin (1987), the methods that are discussed above are not designed with the intent of generating a replacement random variable whose distribution closely approximates the distribution of the missing variable. The method in this section attempts to adapt the methods of Little and Rubin so that the imputed price method accounts not only for the expected value of the missing variable but also for the variance. If the imputed random variable has the same expectation as the missing random variable, but a different variance, then when these imputed values are used to compute regression parameters or are plugged into nonlinear functions, there can be resulting biases. Finally, when prices are missing one needs to account for the possibility that these prices are not missing at random but are missing because of their underlying value.

Fortunately, there is enough information in scanner databases to impute an estimate of the "reservation" or virtual price so that this imputed price estimate has a distribution that closely approximates the distribution of the true reservation price. This is the price that will make the quantity demanded equal to zero. To describe this method, I denote the **k** commodity vector as x with the associated price vector as **p**. The consumer problem is typically

$$\lim_{x} U(x) \quad s.t.\ \mathbf{p}x \leq y \quad x \geq 0 \tag{1}$$

where $U(x)$ is the direct utility function with the standard regularity conditions. Let λ and γ be the Lagrangian multiplier for the first and second constraint, respectively. If the "desired" quantity for the ith good is negative, the first-order condition at the bound of $x_i = 0$ is then

$$U_i(x)\big|_{x_i=0} - \lambda p_i + \gamma = 0, \tag{2}$$

where the subscript on U denotes the derivative of the ith item. The virtual price is then $U_i(x)\big|_{x_i} = 0/\lambda$. Although the "desired" $\{x_i^* = [x_i : U(x) - \lambda p_i = 0]\}$, is negative, we observe x_i is exactly zero. If the market price was $U_i(x)\big|_{x_i=0}/\lambda$ then the quantity demanded would be exactly zero. Letting π_i

2. Feenstra and Shapiro discuss this issue in chapter 5 in this volume.

denote the virtual or "reservation" price, the first-order conditions or tangency conditions are restated as

$$U_i(x)\big|_{x_i=0} - \lambda\pi_i = 0. \tag{3}$$

The role of the virtual price can be displayed graphically using a $k = 2$ example. In figure 11.1, the market price line is MM and the "desired" quantities are x_1^* and x_2^*. However, this solution violates the nonnegativity constraint, since $x_1^* < 0$. Therefore, there is a corner solution at x_2, and there are no sales of good 1. The indifference curve U represents the equilibrium utility and is lower than U', which could be reached if there is no nonnegativity constraint. Therefore, the shadow price of this nonnegativity constraint is $U' - U$. The slope of the price line MM is the ratio $-p_1/p_2$. The price line RR is tangent to U at the equilibrium quantities of $(0, x_2)$. The slope of the price line RR is $-\pi_1/p_2$. If the market price for good 1 had been π_1, the consumer would have reached the same utility that she does under the constrained problem. It is necessary that $\pi_1 \leq p_1$.

If we knew the virtual prices in the scanner data sets, then we could correctly account for the effects of a missing price quote that was either the result of no inventory or of a market price that was "too high." In either case, the virtual price would satisfy condition (c), and if we knew the functional

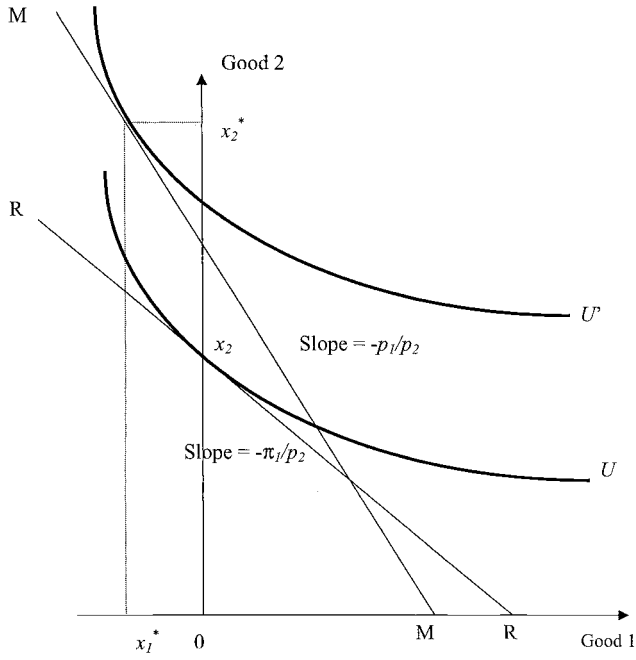

Fig. 11.1 A boundary condition

form of the consumer's utility function, we could use these virtual prices along with the prices of the purchased items to construct a true cost-of-living index. Additionally, we do not need to observe the "list" prices of the nonsold items because they would not be relevant in the construction of a cost of living index.

Unfortunately, one does not observe the virtual price. Instead, a demand system must be estimated, and, using the parameters of the estimated demand system and the prices of the purchased goods, one can impute an estimate of the virtual price. This method has been used in studies that attempt to determine the welfare effects in the introduction of new goods (see Hausman 1996 and Feenstra 1994, 1997). In this study, the indirect utility function is the "stochastic" translog:[3]

$$H(v;\alpha,\beta,\varepsilon) = -\sum_{i=1}^{k}\alpha_i v_i - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}\beta_{ij}v_i v_j + \sum_{i=1}^{k}\varepsilon_i v_i$$

where $v_i = \ln(p_i/y)$ and ε_i is a mean zero random variable. The following homogeneity constraints are imposed: $\sum_{i=1}^{k}\alpha_i = 1$, $\sum_{j=1}^{k}\beta_{ij} = 0$, $\forall i$. Using these constraints and Roy's Identity, and not imposing the nonnegativity constraint for the quantities, we get the "desired" share equations for the ith share:

$$w_i^* = \alpha_i + \sum_{j=1}^{k}\beta_{ij}v_j + \varepsilon_i \tag{4}$$

However, suppose that for the first good $w_1^* < 0$. Then the observed shares $\{w_i\}_{i=1}^{k}$ will not equal the "desired" shares $\{w_i^*\}_{i=1}^{k}$, $w_1 = 0$, and $\sum_{i=1}^{k}w_i = 1$. Suppose further that $k = 3$, $w_2 > 0$, and $w_3 > 0$. Then the system of equations becomes

$$\begin{aligned}
\pi_1 &= -\frac{1}{\beta_{11}}(\alpha_1 + \beta_{12}v_2 + \beta_{13}v_3 + \varepsilon_1)\\
w_2 &= \alpha_2 + \beta_{21}\pi_1 + \beta_{22}v_2 + \beta_{23}v_3 + \varepsilon_2\\
w_3 &= \alpha_3 + \beta_{31}\pi_1 + \beta_{32}v_2 + \beta_{23}v_3 + \varepsilon_3
\end{aligned}$$

In order to impute π_1, we need to estimate the parameters of this demand system. Since v_1 is not observable, the resulting model is truncated rather than censored. In this example, the structural equation for the second good is

$$w_2 = \alpha_2 - \frac{\beta_{21}}{\beta_{11}}\alpha_1 + \left(\beta_{22} - \frac{\beta_{12}\beta_{21}}{\beta_{11}}\right)v_2 + \left(\beta_{23} - \frac{\beta_{13}\beta_{21}}{\beta_{11}}\right)v_3 + \tilde{\varepsilon}_2,$$

where

3. The superlative BLS price will have a Törnqvist functional form, and I posit the translog aggregator since this is the aggregator that makes the Törnqvist exact.

$$\tilde{\varepsilon}_2 = \varepsilon_2 - \frac{\beta_{21}}{\beta_{11}}\varepsilon_1$$

The parameter estimation in this example requires the accounting of "selection effects" because the following event has occurred:

$$(5) \qquad \tilde{\varepsilon}_2 > -\left[\alpha_2 - \frac{\beta_{21}}{\beta_{11}}\alpha_1 + \left(\beta_{22} - \frac{\beta_{12}\beta_{21}}{\beta_{11}}\right)v_2 + \left(\beta_{23} - \frac{\beta_{13}\beta_{21}}{\beta_{11}}\right)v_3\right].$$

This event is denoted as A_2, and A_3 is the event that the third good is purchased. Therefore, the following holds:

$$E(w_2 \mid A_2, A_3, v_2, v_3) = \alpha_2 + -\frac{\beta_{21}}{\beta_{11}}\alpha_1 + \left(\beta_{22} - \frac{\beta_{12}\beta_{21}}{\beta_{11}}\right)v_2 + \left(\beta_{23} - \frac{\beta_{13}\beta_{21}}{\beta_{11}}\right)v_3 + E(\tilde{\varepsilon}_2 \mid A_2, A_3)$$

Since $E(\tilde{\varepsilon}_2 \mid A_2, A_3)$ is a function of the observed v_2, and v_3, and since the residuals across the equations are not independent (since each one will now contain ε_1), the regressors are now correlated with the residual. Therefore, the econometric estimation of the share equations cannot be solely done by nonlinear least squares estimation. It is these selection effects that make price imputation more difficult than the imputation in a "new goods" problem in which the time period of introduction is exogenous and therefore selection effects need not be incorporated. The appendix describes the estimation method used in this study in order to get the parameters of the demand system.

Once the parameters of the demand system have been estimated, it might be tempting to impute the virtual price $\hat{\pi}_1$ by

$$(6) \qquad \hat{\pi}_1 = -\frac{1}{\hat{\beta}_{11}}(\hat{\alpha}_1 + \hat{\beta}_{12}v_2 + \hat{\beta}_{13}v_3).$$

$\hat{\beta}_i$ denotes the parameter estimate of β_i. While $\text{plim}(\hat{\pi}_1) = E(\pi_1)$, the variance of the imputed virtual price, $\hat{\pi}_1$, will be smaller than variance of the actual virtual price, π_1. One needs to account for the variance that comes from both the residual and the parameter estimates. The appendix describes this imputation method in greater detail.

11.3 The Cereal Scanner Data Set

The data set used in this study is the source data set that the BLS has used to construct its real time New York Cereal Index, which is described in the Richardson article of this publication. It contains the price and quantity sold for the supermarket outlets for New York City and its surrounding counties. Most large grocery chains, price clubs, and drugstores use scan-

ner systems to monitor their inventory, to store prices, and to retrieve these stored prices when items pass through the checkout line. Each item has a twelve-digit bar code or Universal Product Code (UPC). It is the UPC that distinguishes the different items that are sold in an outlet. Different digits in the UPC are reserved to identify specific characteristics. For instance, there are five digits used to identify the manufacturer. These digits are assigned to the manufacturer by the Uniform Code Council. Another five digits are used by the manufacturer to identify each distinct product that is produced. Each new item or change in an existing item requires the issuing of a new UPC code.

When an item goes through the checkout line, the cash register scans the UPC code of the item, retrieves a price, and then records the sale on a computer tape or disk. From these records, one can find the weekly sales and prices for each bar-coded item in a grocery store. The outlet managers can use this information to monitor the turnover of the items on their shelves and make adjustments to improve their sales margins.

Even for a highly specified "item-area" such as cereal in New York, these data sets are extremely large. For instance, one month of data for New York cereal contains more observations than the entire data set that the BLS uses within a year. Because of this vast size, there is a need to establish a hierarchy. Figure 11.2 outlines the hierarchy for cereal. At the top level, the module identifies broad category type (ready-to-eat vs. hot cereal). At the next level is the brand name. Sometimes a brand name is the proprietary trademark of a firm (e.g., Cheerios) and other times it is not (e.g., Raisin Bran). At the lowest level is the UPC, each specific product having a unique UPC.

Table 11.1 lists the frequency of at least one unsold item for eight major brands over the 181-week period in the scanner data set. Except for Mini Wheats and Total, there is at least a 20 percent chance that for a given week that there will be at least one unsold UPC. It is evident that this probability increases with the number of UPCs within a brand. Although the probability of a nonsale is 20 percent for a particular brand, for every week in the data set used in this study, there are at least three nonsales.

In this study, when I compute indexes using the unit value approach, I group the UPCs by brand and then compute a unit value for each brand. I can do this since in this data set there is always at least one item within a brand that has strictly positive sales. I also investigate whether the items within a brand satisfy the Balk's criteria for a unit value index.

Even within brands and among outlets, there is evidence of product differentiation. The graphs in figure 11.3 plot the range of each week's prices for the top-selling UPCs in New York at different stores.[4] It is evident that the outlet differentiates the product, since consumers do not purchase only

4. These are prices per ounce.

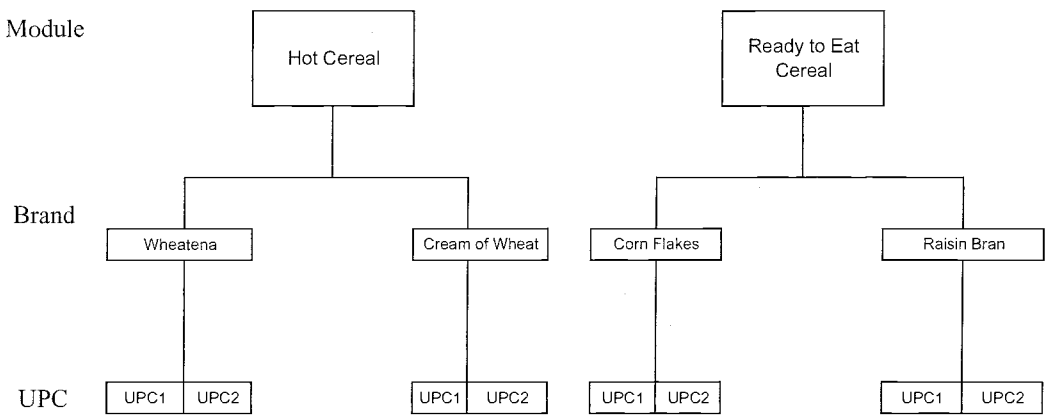

Fig. 11.2 Hierarchy for cereal

Table 11.1 The Frequency of Unsold Items by Brand

Brand	% of periods with Unsold Item	Number of Items (UPCs) within Brand
Cheerios	20.1	5
Corn Flakes	27.6	7
Raisin Bran	28.6	5
Rice Krispies	24.8	3
Honey Nut Cheerios	12.8	4
Frosted Flakes	33.1	7
Total	18.2	3
Mini Wheats	4.1	2

at the outlet offering the lowest price. Although Reinsdorf (1993) found evidence of outlet substitution bias, it seems clear that outlets are not perfect substitutes. Notice that the minimum price in general fluctuates more than the maximum price. The reason is that different stores put these items on sale at different times, and the percent reduction of the sale price varies across store. Most often the minimum price is a sale price.

When a brand has several UPCs assigned to it, it is usually the box size that distinguishes the two UPCs within the same brand. Conventional wisdom might conclude that box size is an "immaterial" characteristic. However, I find evidence to the contrary. I select three of the larger-selling brands and select two UPCs for each brand. The only characteristic that differentiates the two UPCs is box size. For each store, I subtract the price of the larger box from the price of the smaller box and then average these differences across stores. The results are listed in table 11.2, and the null hypothesis that the two prices are equal is always rejected. Therefore, we observe that outlets offer a choice of different box sizes for the same brand. These different box sizes have different per ounce prices. Both box sizes enjoy positive sales. Therefore, a 20-oz. box of Cheerios is not a perfect substitute for a 15-oz. box of Cheerios. Additionally, it is evident that a 20-oz. box of Cheerios in one store is not a perfect substitute for a 20-oz. box in another store. Therefore, both the outlet and the box size differentiate the product.

Based on this evidence, I conclude that the sufficient conditions for constructing a price index by taking unit values either across outlets or across items with a particular brand do not hold for the cereal market in New York. Obviously neither the items within a brand nor the outlets are complements. There is a variance of prices within a brand and among outlets. The items within a brand are not perfect substitutes. If they were perfect substitutes, then manufacturers and outlets would not be bearing the additional costs of offering different box sizes for the same item. However, it is still possible that a unit value approach might "closely" approximate a true price index.

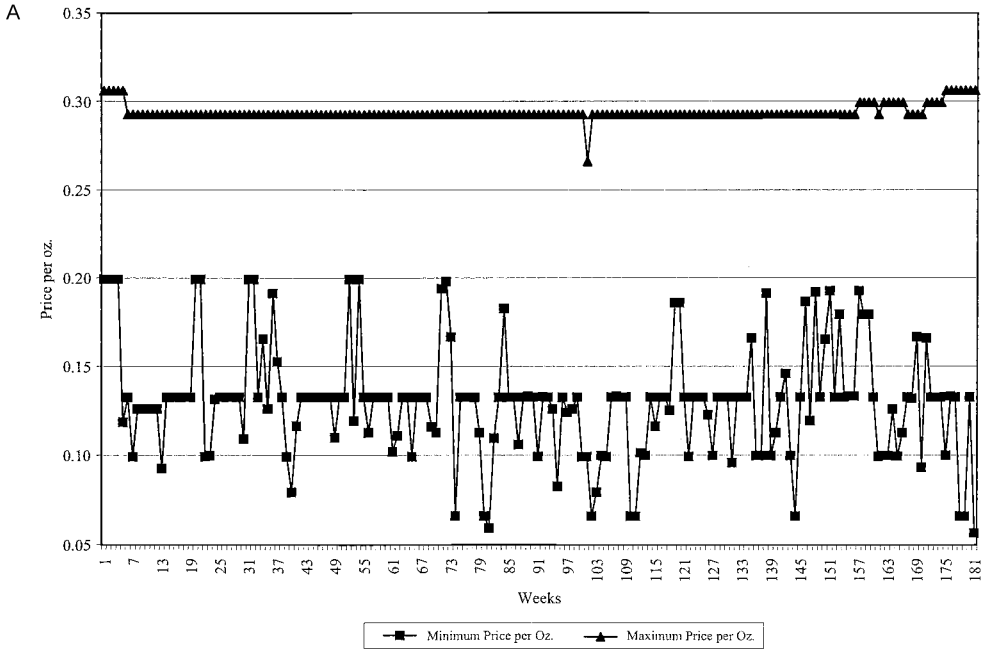

Fig. 11.3 *A,* **Price per ounce range across stores for 15-oz. box of Cheerios;** *B,* **price per ounce range across stores for 20-oz. box of Cheerios;** *C,* **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** *D,* **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** *E,* **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** *F,* **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

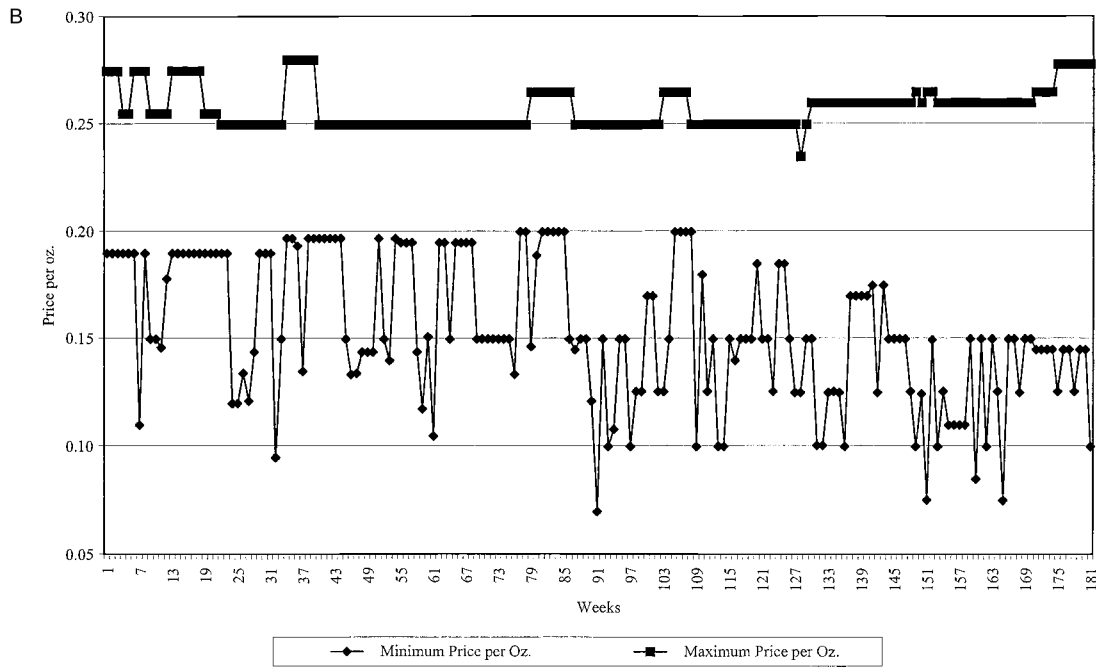

Fig. 11.3 (cont.) ***A,*** **Price per ounce range across stores for 15-oz. box of Cheerios;** ***B,*** **price per ounce range across stores for 20-oz. box of Cheerios;** ***C,*** **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** ***D,*** **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** ***E,*** **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** ***F,*** **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

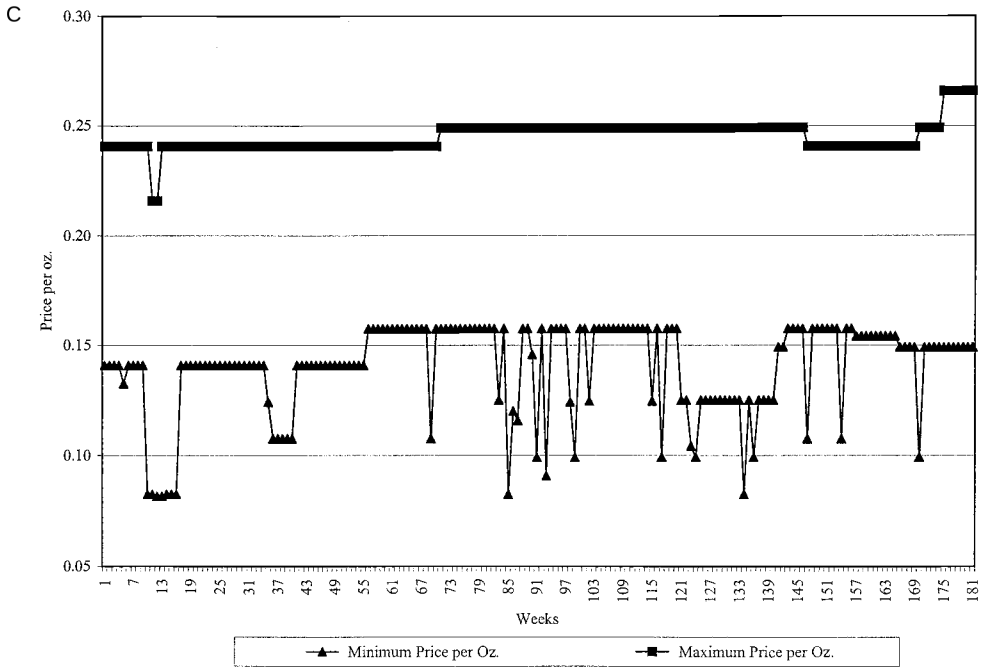

Fig. 11.3 (cont.) ***A,*** **Price per ounce range across stores for 15-oz. box of Cheerios;** ***B,*** **price per ounce range across stores for 20-oz. box of Cheerios;** ***C,*** **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** ***D,*** **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** ***E,*** **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** ***F,*** **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

D

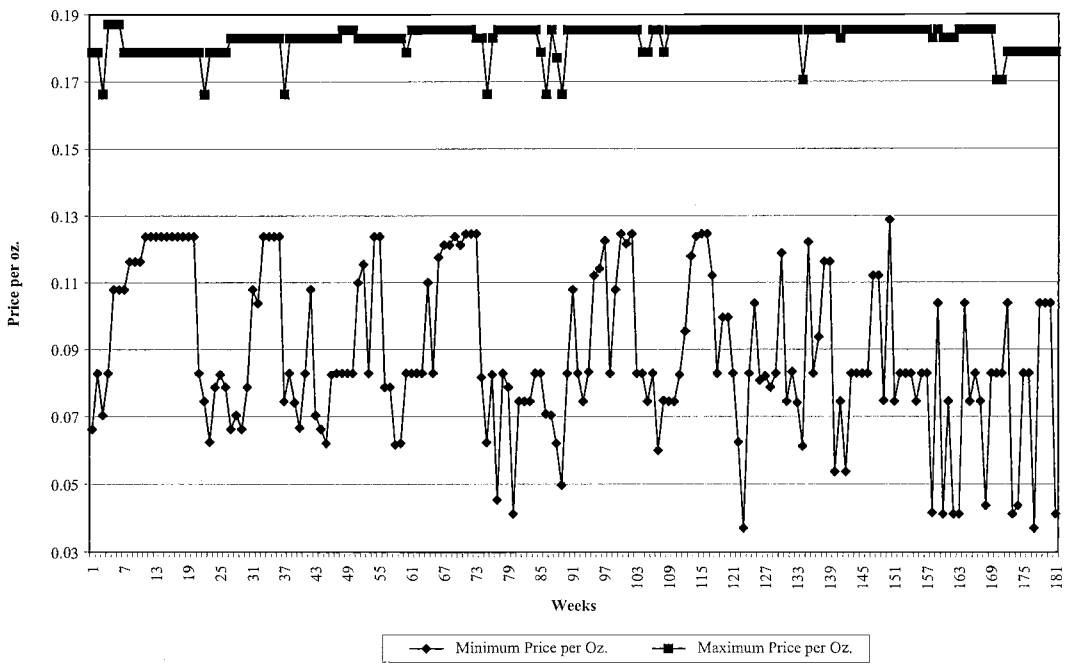

Fig. 11.3 (cont.) ***A,*** **Price per ounce range across stores for 15-oz. box of Cheerios;** ***B,*** **price per ounce range across stores for 20-oz. box of Cheerios;** ***C,*** **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** ***D,*** **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** ***E,*** **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** ***F,*** **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

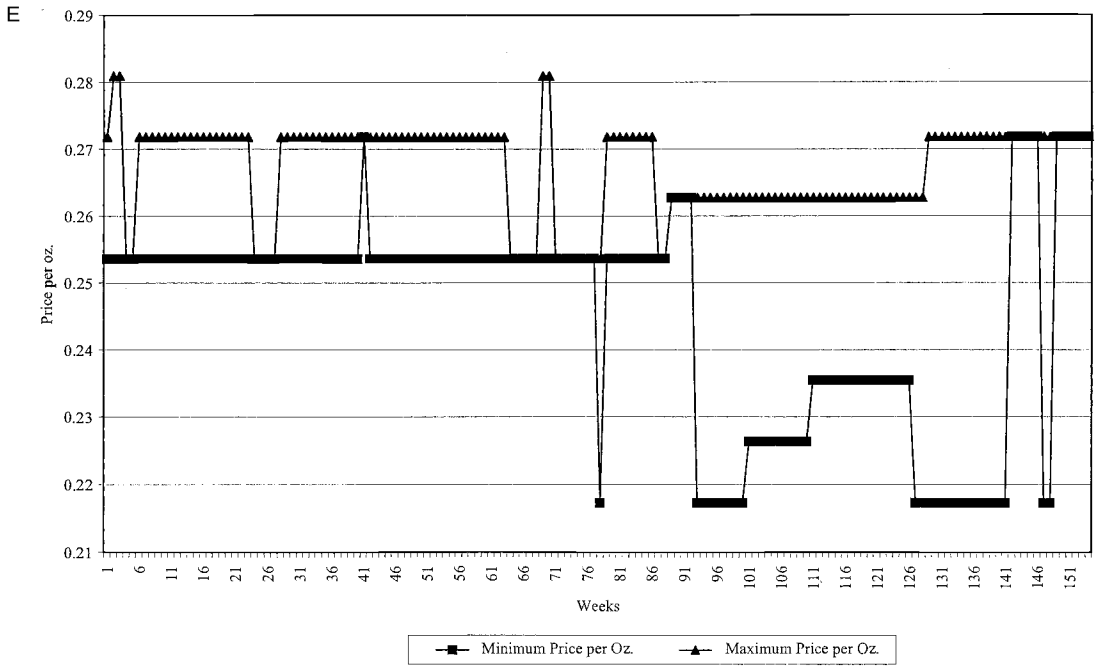

Fig. 11.3 (cont.) ***A,*** **Price per ounce range across stores for 15-oz. box of Cheerios;** ***B,*** **price per ounce range across stores for 20-oz. box of Cheerios;** ***C,*** **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** ***D,*** **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** ***E,*** **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** ***F,*** **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

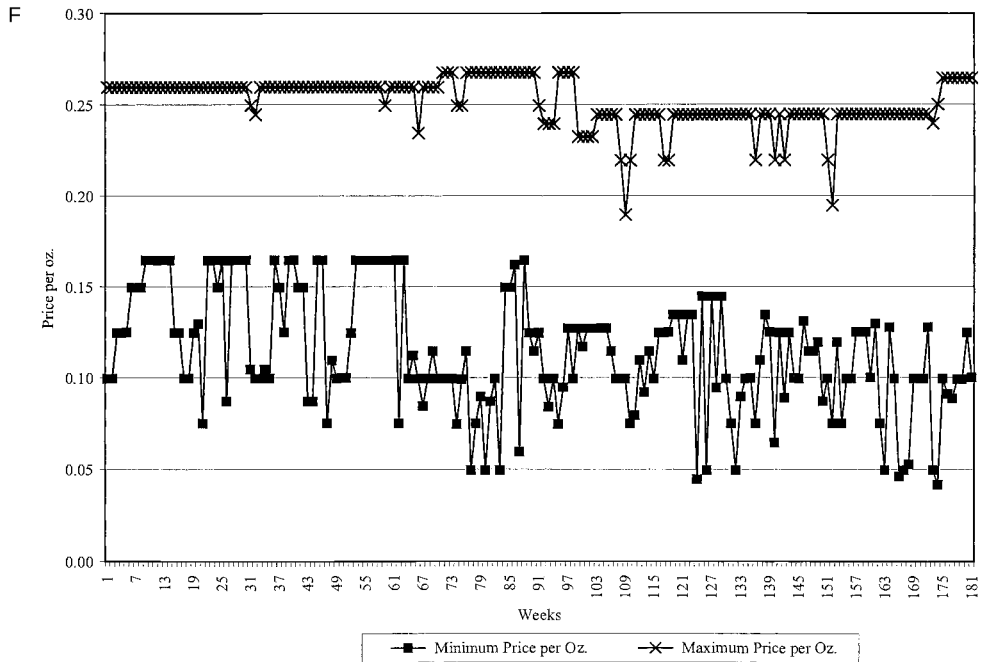

Fig. 11.3 (cont.) ***A,*** **Price per ounce range across stores for 15-oz. box of Cheerios;** ***B,*** **price per ounce range across stores for 20-oz. box of Cheerios;** ***C,*** **price per ounce range across stores for 12-oz. box of Kellogg's corn flakes;** ***D,*** **price per ounce range across stores for 24-oz. box of Kellogg's corn flakes;** ***E,*** **price per ounce range across stores for 11-oz. box of Kellogg's raisin bran;** ***F,*** **price per ounce range across stores for 20-oz. box of Kellogg's raisin bran**

Table 11.2 Within-Store Differences in Price per Ounce

Brand Name	Sizes	Average Difference in Price per Oz.	*t*-test Probability that Average Difference = 0	Maximum	Median	Minimum
General Mills Cheerios	15 oz, 20 oz	0.009024	0.0001	0.1665	0.017667	–0.16017
Kellogg's Corn Flakes	12 oz, 18 oz	0.043853	0.0001	0.1675	0.04125	–0.08375
Kellogg's Raisin Bran	11 oz, 22 oz	0.048706	0.0001	0.179136	0.045955	0.017773

To gain insight on the nonsale of items in this database, I define a sale as any price that is 95 percent or below the median price. Among the items that experience no sales, 28 percent of the nonsales follow its own sale *and* at the same time occur when at least one other item with its own brand experiences a sale. Twenty-one percent of the nonsales follow *only* its own sale, and 22 percent occur *only* during the sale of at least one other item within its own brand. The remaining 29 percent do not follow either its own sale or occur during the sale of at least one other sale in its own brand. These results should influence the results of the carry-forward imputation approach.

11.4 The Indexes Based on Five Methods

I estimate Törnqvist price indexes using one implicit imputation, the unit value, and four explicit methods. The first explicit method is the carry forward. The second is the BLS method. The last two are the imputations of the virtual price. I impute the virtual price two ways—first by using a simple direct approach using equation (6) and second by accounting for the selection effects and all the sources of variation. The imputation approach is described in the appendix.

My time series starts in August 1994 and ends in March 1998, and it has 181 weeks. The "cereal price war" occurs during this period as Kellogg's attempts to stop its falling market share. Price drops are most dramatic for those brands whose name is not proprietary, such as Raisin Bran and Corn Flakes.

I randomly select a store and start with the brands that have the highest expenditure share. I do this because the price indexes coming from these brands will be given a greater weight in the final index. It would almost be computationally impossible to do this for every brand, and therefore it is perhaps more important to correctly generate indexes for the brands that will get the most weight.

When I estimate the virtual prices, I first estimate the parameters of the

model that is depicted in equation (4) by nonlinear least squares without adjusting for the selection effects of product exit and entry. I use the model parameters to impute a virtual price as described in equation (6). Then I estimate the model again and account for the selection effect as shown in equation (5) by using a simulated moment method, described in the appendix. When I impute a virtual price this time, I account for the additional sources of variation.

The UPCs that fall under each brand include both the national trademark brand and the private-label counterpart that is intended to serve as a substitute for the national brand. Therefore, for a brand such as Cheerios, I include all the different boxes of General Mills Cheerios and the store's private-label cereal that is intended to be a substitute for Cheerios. I include the private labels because they are specifically manufactured to be a substitute for a national brand even though the characteristics of the cereal are not exactly the same. In the case of the "Raisin Bran" brand, I combine both Kellogg's and Post Raisin Bran along with the private label that is intended to be a substitute for the Post Raisin Bran.

Finally, I construct Törnqvist indexes for the fourteen top-selling brands using five different imputation methods.[5] The first index uses unit values, and the second index uses the imputed virtual prices that come from the parameter estimates that are done without adjusting for the selection effects. The third index is calculated using the full imputation procedures described in the previous section. The fourth index uses the carry-forward imputation, and the fifth method is the BLS method.

Table 11.3 lists the parameter estimates for the nonlinear least squares estimation that does not incorporate the selection effects. (Table 11A.1 lists the results that do incorporate the selection effects.) Because of space limitations, I do not report the results for each brand. Instead, I give the results for the top-selling brands. The first results are for the Cheerios brand. There are five UPCs that fall within this brand. The UPCs are a private-label 7-oz. box, a 35-oz. box of Cheerios, a 20-oz. box of Cheerios, a 10-oz. box of Cheerios, and a 15-oz. box of Cheerios. Generally, the cross effects among the 20-oz., 10-oz., and 15-oz. boxes increase after the selection effects are incorporated. This should be expected. The absolute value of the own price coefficients increases for those UPCs with a relatively large share of the Cheerios brand.

The next set of results is for the Corn Flakes brand. The first five UPCs are, respectively, the 45-oz. box, the 7-oz. box, the 12-oz. box, the 18-oz. box, and the 24-oz. box for Kellogg's Corn Flakes. The last are respectively the 12-oz. and 18-oz. boxes of the private label. The highest cross effect is between the 18-oz. box of Kellogg's and the 24-oz. box. In the appendix, I

5. Again, I focus on the Törnqvist, since the chained Törnqvist will be the functional form of the newly published BLS superlative index.

Table 11.3 **Parameter Estimates of Lower-Level Demand Systems by Brand without Selection Effects**

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
			General Mills Cheerios (R^2 = .808)				
Own and Cross Effect							
Item 1	0.42						
	(.023)						
Item 2	0.04	–0.23					
	(.019)	(.031)					
Item 3	0.11	0.12	–0.86				
	(.040)	(.047)	(.084)				
Item 4	0.13	0.24	0.40	–0.93			
	(.029)	(.037)	(.058)	(.066)			
Item 5	0.13	–0.17	0.22	0.15	–0.32		
	(.044)	(.042)	(.061)	(.049)	(.084)		
Constant	0.31	0.49	0.66	1.10	–1.56		
	(.073)	(.066)	(.142)	(.101)	(.201)		
			Kellogg's Corn Flakes (R^2 = .716)				
Own and Cross Effect							
Item 1	–0.23						
	(.003)						
Item 2	0.00	–0.26					
	(.004)	(.060)					
Item 3	0.00	0.25	–0.51				
	(.012)	(.065)	(.104)				
Item 4	0.07	0.02	0.12	–0.87			
	(.020)	(.032)	(.061)	(.094)			
Item 5	–0.03	–0.06	0.07	0.51	–0.59		
	(.017)	(.032)	(.058)	(.065)	(.076)		
Item 6	0.01	0.03	0.05	0.10	0.06	–0.26	
	(.005)	(.019)	(.048)	(.048)	(.041)	(.023)	
Item 7	–0.31	0.01	0.01	0.05	0.05	0.00	–0.09
	(.007)	(.010)	(.030)	(.044)	(.037)	(.013)	(.022)
Constant	–0.16	0.57	1.05	0.26	–0.66	–0.09	0.04
	(.033)	(.053)	(.129)	(.209)	(.170)	(.068)	(.078)
			Kellogg's Raisin Bran (R^2 = .663)				
Own and Cross Effect							
Item 1	–0.06						
	(0.004)						
Item 2	0.07	–0.86					
	(0.038)	(0.097)					
Item 3	–0.05	0.47	–0.73				
	(0.032)	(0.088)	(0.135)				
Item 4	0.00	0.00	0.026	–0.27			
	(0.006)	(0.037)	(0.048)	(0.014)			
Item 5	0.03	0.32	0.05	0.00	–0.41		
	(0.027)	(0.078)	(0.099)	(0.039)	(0.116)		
Constant	0.21	0.98	–1.04	0.20	0.66		
	(.022)	(.169)	(.193)	(.042)	(.159)		

Table 11.3 (continued)

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
	Kellogg's Rice Krispies (R^2 = .781)						
Own and Cross Effect							
Item 1	–0.70						
	(.001)						
Item 2	0.21	–0.57					
	(.025)	(.046)					
Item 3	0.49	0.36	–0.85				
	(.032)	(.046)	(.055)				
Constant	–0.15	1.17	–0.02				
	(.141)	(.095)	(.123)				
	General Mills Honey Nut Cheerios (R^2 = .835)						
Own and Cross Effect							
Item 1	–0.32						
	(.034)						
Item 2	0.01	–0.12					
	(.019)	(.012)					
Item 3	0.18	0.01	–0.65				
	(.034)	(.026)	(.057)				
Item 4	0.14	0.10	0.47	–0.70			
	(.035)	(.025)	(.048)	(.054)			
Constant	–0.01	0.45	–0.58	1.13			
	(0.104)	(0.058)	(0.154)	(0.158)			

Brand/Item	Item Description
Cheerios	
Item 1	General Mills 7 oz.
Item 2	General Mills 35 oz.
Item 3	General Mills 20 oz
Item 4	General Mills 10 oz.
Corn Flakes	
Item 1	Kellogg's Corn Flakes 45 oz.
Item 2	Kellogg's Corn Flakes 7 oz.
Item 3	Kellogg's Corn Flakes 12 oz.
Item 4	Kellogg's Corn Flakes 18 oz.
Item 5	Kellogg's Corn Flakes 24 oz.
Item 6	Private Label 12 oz.
Item 7	Private Label 10 oz.
Raisin Bran	
Item 1	Kellogg's Raisin Bran 23.5 oz.
Item 2	Kellogg's Raisin Bran 20 oz.
Item 3	Post Raisin Bran 20 oz.
Item 4	Private Label 51 oz.
Item 5	Kellogg's 25.5 oz.
Rice Krispies	
Item 1	Kellogg's Rice Krispies 15 oz.
Item 2	Kellogg's Rice Krispies 10 oz.
Item 3	Kellogg's Rice Krispies 19 oz.
Honey Nut Cheerios	
Item 1	General Mills 27 oz.
Item 2	General Mills 48 oz.
Item 3	General Mills 14 oz.
Item 4	General Mills 20 oz.

Note: Standard errors in parentheses.

show that the cross effect is increased when the parameters are estimated with the selection effect. In both estimations, there are negative cross effects between the 45-oz. box and the private labels. This is a disturbing result.

When estimating these parameters, there is one factor that perhaps can create bias. Oftentimes, an outlet will place an item on sale and temporarily run out of the item during the sale period. In these situations, the desired quantity purchased does not equal the actual quantity purchase. Unfortunately, the event of an "item run-out" is not recorded on the scanner data sets. This measurement error problem should include a downward bias in the absolute value of both the own and cross effects since the magnitude of the quantity change coming from a price change is underreported. For the time being, this problem will persist, and it will also affect the values of any superlative index that is calculated from this data.

Additional results from the top-selling brands are also listed in table 11.2 but will not be discussed in this paper.

Table 11.4 gives summary statistics for the Törnqvist indexes generated in this study, and figures 11.4A and 11.4B plot the value of the indexes. I compute both a chained and a direct Törnqvist index for each method over the 181-week period. The column head "simple imputation" refers to the simple economic imputation that is done without incorporating the selection effects and that uses a price using the form in equation (6). Additionally, it does not account for all the sources of variation. The column head "full imputation" is the economic imputation that does account for selection effects and all source of variation. Besides giving summary statistics, this table also gives the last period value for each one of the methods.

Table 11.4 **Index Results Summary Statistics**

Method	Last Period Value	Average	Standard Deviation	Median	Minimum	Maximum
Unit value						
Direct	0.71174988	0.8690262	0.073431105	0.868337	0.575281	1.0302505
Chained	0.77702322	0.8805631	0.058646995	0.879293	0.66946	1.02783
Carry forward						
Direct	0.89017	0.9513015	0.031027043	0.95083	0.81719	1.00501
Chained	0.7946	0.9118548	0.051323889	0.89991	0.75456	1.00848
BLS method						
Direct	0.8584	0.9376107	0.034104332	0.93665	0.8141	1.00246
Chained	0.88219	0.9771714	0.035775544	0.97913	0.8512	1.05026
Simple imputation						
Direct	0.8312895	0.9277948	0.040486383	0.935425	0.751042	1.0178959
Chained	0.87883525	0.9484294	0.034320196	0.957048	0.841315	1.0136196
Full imputation						
Direct	0.833335324	0.9293684	0.039747923	0.936333	0.751782	1.0170632
Chained	0.88074915	0.9500644	0.0033277902	0.95858	0.845506	1.0123342

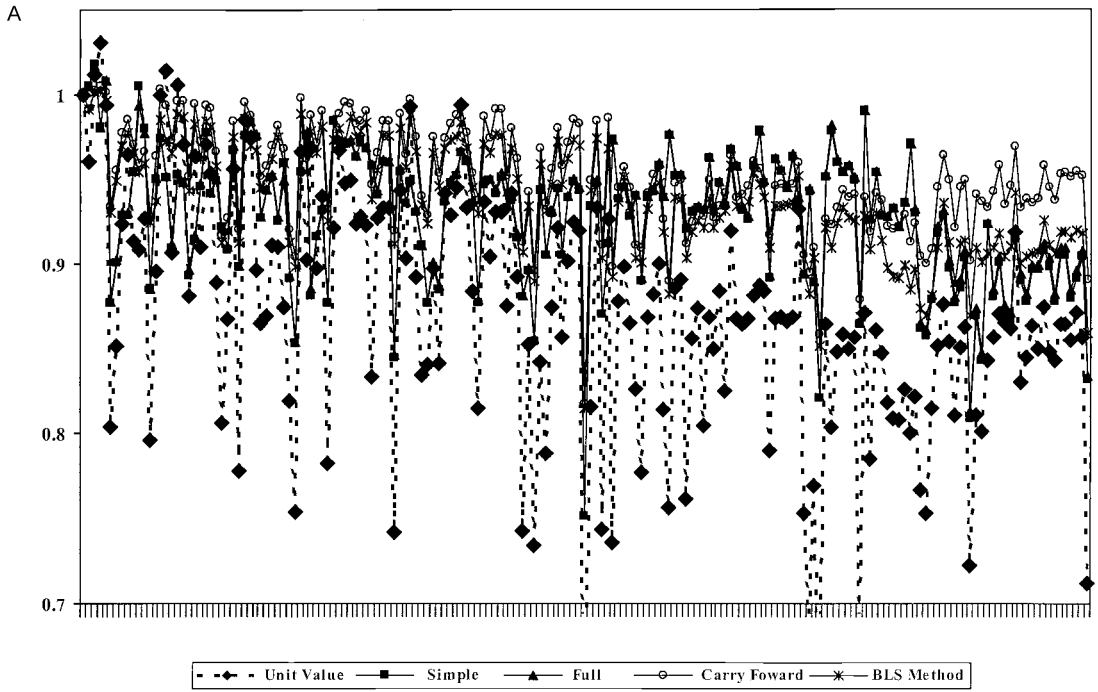

Fig. 11.4 *A*, A comparison of direct indexes; *B*, a comparison of chained indexes

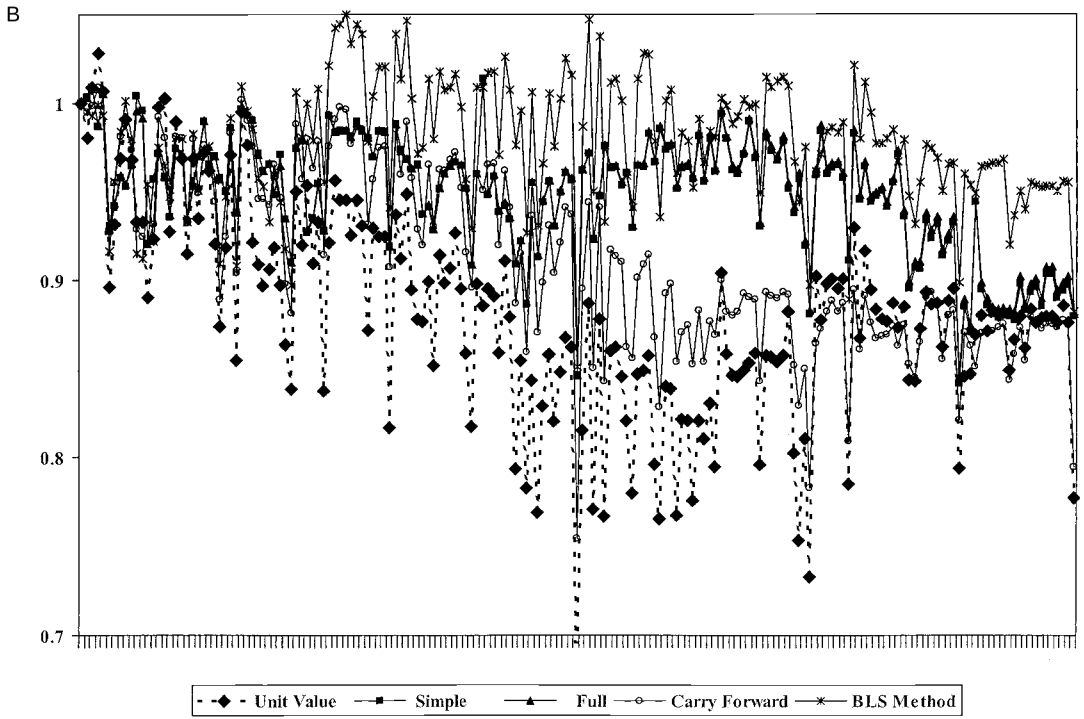

Fig. 11.4 (cont.) *A,* **A comparison of direct indexes;** *B,* **a comparison of chained indexes**

Figures 11.4A and 11.4B illustrate an important result. The lines that depict the indexes that are not based on unit values are indistinguishable. The line for the unit values is distinguishable and lies below the other indexes. While the other methods generate significantly different results when compared to each other, the magnitude of these differences is not as large as their difference with the indexes that use unit values. It is important to note that in the unit value method all the prices within each group are being replaced by their unit value. In the other methods, the prices of the sold items are always the same; therefore, the only difference is solely the result of the imputation method itself. For cereal in New York, the Balk criteria for using unit values do not hold, and the indexes using unit values produce dramatically different results.

What is surprising is that under the direct method there is the small magnitude of difference between the BLS method and the full and simple economic imputations. Clearly, the current BLS method is far simpler than the economic approach, and it seems that the BLS's current price replacement methods at least somewhat incorporate the welfare effects of a nonsale of an item.

Sections A and B of table 11.5 list the mean differences between indexes and give a standard errors for these differences. The index for each time period using the method listed in the row down the page is subtracted from the

Table 11.5 Difference across Methods in Estimates of Price Change

	Unit Value	Carry Forward	BLS Method	Simple Imputation	Full Imputation
		A. Direct			
Unit value	–0.082	–0.069	–0.059	–0.060	
		(0.003)	(0.003)	(0.004)	(0.004)
Carry forward	0.082		0.014	0.024	0.022
	(0.003)		(0.001)	(0.002)	(0.002)
BLS method	0.069	–0.014		0.010	0.008
	(00.003)	(0.001)		(0.002)	(0.002)
Simple imputation	0.059	–0.024	–0.010		–0.002
	(0.004)	(0.002)	(0.002)		(0.000)
Full imputation	0.060	–0.022	–0.008	0.002	
	(0.004	(0.002)	(0.002)	(0.000)	
		B. Chain			
Unit value		–0.0313	–0.0966	–0.0679	–0.0695
		(0.002)	(0.004)	(0.004)	(0.004)
Carry forward	0.0313		–0.0653	–0.0366	–0.0382
	(0.002)		(0.003)	(0.003)	(0.003)
BLS method	0.0966	0.0653		0.0287	0.0271
	(0.004)	(0.003)		(0.002)	(0.002)
Simple imputation	0.0679	0.0366	–0.0287		–0.0016
	(0.004)	(0.003)	(0.002)		(0.000)
Full imputation	0.0695	0.0382	–0.0271	0.0016	
	(0.004)	(0.003)	(0.002)	(0.000)	

Note: Standard errors in parentheses.

index method listed in the column and then averaged. All of the differences are significant at the 5 percent level. There is even a significant difference between the full and the simple economic imputation approach. Clearly, the regression results between the model that incorporated the selection effects and the one that did not were not large in magnitude, and thus the resulting differences in the price indexes are not large.

There is an interesting result for the chained index based on the carry-forward method. The chained index drifts below its direct value. Usually, the reverse happens. A major pitfall of the Törnqvist is that it is not reversible, so that if prices return to their base-period value, a chained Törnqvist will not necessarily equal one. It is very important to note from table 11.4 that the chained Törnqvist using the carry-forward method has a higher variance than the other explicit methods. This confirms the conclusions of Armknecht and Maitland-Smith (1999), and, as mentioned earlier, the carry-forward method carries forward a sales price for over 49 percent of the imputations, so that when the item is purchased again there is a "bounce" in the index. However, this "bounce" after reappearance does not completely offset the downward drift that occurs from deeply discounted sales. Perhaps these are important reasons that the carry-forward method might be avoided.

11.5 Conclusions

Among the alternative imputation methods that are reviewed in this study, it seems that the unit value methods generate the largest difference when applied to the scanner cereal database for New York. Clearly, the cereal market does not have the characteristics that are sufficient for a price index that uses unit values to be a true price index. However, there could easily be other product areas where at least one of Balk's criteria is met. For example, in the tuna market, Feenstra and Shapiro's study (chap. 5) in this volume indicates that the conditions for taking unit values across time might provide a true price index.

The economic approach in this study did not produce an index whose difference with the index using the BLS approach was large in magnitude even though the difference was statistically significant. At least in the cereal market, it seems that the BLS imputation method produces indexes with relatively smaller variances and whose results are close in magnitude to the indexes based on the economic approach.

Appendix

Here, I describe the full economic imputation method in detail. I posit the translog indirect utility function that generates the "desired" share equations depicted in equation (6).

Suppose in time period t goods that are indexed from 1 to m are unsold and the remaining $k - m$ goods experience positive sales. The system of equations for this time period is then

$$w^t_{m+1} = \alpha_{m+1} + \sum_{j=1}^{m} \beta_{m+1,j} \overline{\pi}^t_j + \sum_{i=m+1}^{k} \beta_{m+1,j} v^t_j + \tilde{\varepsilon}^t_{m+1} \tag{7}$$

$$\cdots$$

$$w^t_k = \alpha_k + \sum_{i=1}^{m} \beta_{k,i} \overline{\pi}^t_i + \sum_{i=m+1}^{k} \beta_{k,i} v^t_i + \tilde{\varepsilon}^t_k,$$

where

$$\begin{pmatrix} \overline{\pi}^t_1 \\ \cdot \\ \cdot \\ \overline{\pi}^t_m \end{pmatrix} = -\begin{pmatrix} \beta_{11} \cdots & \beta_{1m} \\ \cdots & \cdot \\ \cdots & \cdot \\ \beta_{m1} \cdots & \beta_{mm} \end{pmatrix}^{-1} \left[\begin{pmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \alpha_m \end{pmatrix} + \begin{pmatrix} \beta_{1,m+1} \cdots & \beta_{1,k} \\ \cdot & \cdot \\ \cdot & \cdot \\ \beta_{m,m+1} \cdots & \beta_{m,k} \end{pmatrix} \begin{pmatrix} v^t_{m+1} \\ \cdot \\ \cdot \\ v^t_k \end{pmatrix} \right] \tag{8}$$

$$\begin{pmatrix} \tilde{\varepsilon}^t_{m+1} \\ \cdot \\ \cdot \\ \tilde{\varepsilon}^t_k \end{pmatrix} = -\begin{pmatrix} \beta_{m+1,1} \cdots & \beta_{m+1,m} \\ \cdots & \cdot \\ \cdots & \cdot \\ \beta_{k,1} \cdots & \beta_{k,m} \end{pmatrix} \begin{pmatrix} \beta_{11} \cdots & \beta_{1m} \\ \cdots & \cdot \\ \cdots & \cdot \\ \beta_{m1} \cdots & \beta_{mm} \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon^t_1 \\ \cdot \\ \cdot \\ \varepsilon^t_m \end{pmatrix} + \begin{pmatrix} \varepsilon^t_{m+1} \\ \cdot \\ \cdot \\ \varepsilon^t_k \end{pmatrix} \tag{9}$$

Note that $\overline{\pi}^t_i$ is that part of π^t_i that contains the effect of the prices and not the residuals. Likewise, the effect of the residuals in π_i are incorporated in $\tilde{\varepsilon}^t_{m+1}, \ldots \tilde{\varepsilon}^t_k$. As mentioned in the study, the conditional expectation $E(\tilde{\varepsilon}^t_i \mid w_{m+1} > 0, \ldots, w_k > 0)$ for $i > m$ can be correlated with the right-hand-side regressors. To account for this residual correlation, I posit that $[\varepsilon_1, \ldots, \varepsilon_k]' \sim N(0_k, \Sigma)$, where $\mathbf{0_k}$ is a k vector of zeros and Σ is a positive semidefinite $k \times k$ matrix. Then $[\tilde{\varepsilon}_{m+1}, \ldots, \tilde{\varepsilon}_k] \sim N(0, \Gamma)$ where

$$\Gamma = V \sum \Sigma V'$$

and

$$V = \left(-\begin{pmatrix} \beta_{m+1,1} \cdots & \beta_{m+1,m} \\ \cdots & \cdot \\ \cdots & \cdot \\ \beta_{k,1} \cdots & \beta_{k,m} \end{pmatrix} \begin{pmatrix} \beta_{11} \cdots & \beta_{1m} \\ \cdots & \cdot \\ \cdots & \cdot \\ \beta_{m1} \cdots & \beta_{mm} \end{pmatrix}^{-1}, I_{k,k} \right)$$

The event $w_i > 0$ is the event

$$\tilde{\varepsilon}^t_i > -(\alpha_k + \sum_{j=1}^{m} \beta_{i,j} \overline{\pi}^t_j + \sum_{j=m+1}^{k} \beta_{i,j} v^t_j) = B_i$$

The pdf for the $k - m$ vector $\tilde{\varepsilon}^t = \{\tilde{\varepsilon}^t_{m+1}, \ldots, \tilde{\varepsilon}^t_k\}$ is denoted as $\phi[\tilde{\varepsilon}^t; 0_{m-k'}, \Gamma]$, which is a $k - m$ variate normal probability density function with mean $0_{k-k'}$ and variance covariance matrix equal to Γ. Denote the $k - m$ vector of $\{v^t_{m+1}, \ldots, v^t_k\}$ as v^t. Then, for $i > m$,

$$(10)\quad E\left(\tilde{\varepsilon}^t \mid w^t_{m+1} > 0, \ldots, w^t_k > 0, v^t; \beta, \sum\right) = \frac{\int_{B_{m+1}}^{\infty} \ldots \int_{B_k}^{\infty} z\phi(z; 0_{m-k}, \Gamma)dz}{\int_{B_{m+1}}^{\infty} \ldots \int_{B_k}^{\infty} \phi(z; 0_{m-k}, \Gamma)dz}$$

The matrix of $\beta_{i,j}$ is β. The reason that β and v^t are conditioning values in equation (10) is that the integration limits B_i have these values in their domain. There is no analytical solution to equation (10). However, using a simulation technique that is essentially a variant of the Geweke-Hajivassiliou-Keane (GHK) simulator that is described in Gourieroux and Monfort (1996), I can generate a $k-m$ random vector whose expected mean is $E(\tilde{\varepsilon}^t \mid w_{m+1} > 0, \ldots, w_k > 0, v^t; \beta, \Sigma)$. In this study, I generate 200 simulations of this random variable and then average them. I denote this average as $h(v^t; \beta, \Sigma)$. Since this is an unbiased estimate of equation (10) of $O_p(1/\sqrt{200})$, the variance of $h(v^t; \beta, \Sigma)$ should be small. Letting

$$\hat{w}^t_i\left(v^t; \beta, \sum\right) = \alpha_i + \sum_{j=1}^{m} \beta_{i,j}\overline{\pi}^t_j + \sum_{j=m+1}^{k} \beta_{i,j} v^t_j + h\left(v^t; \beta, \sum\right)$$

and letting $\hat{w}^t(v^t; \beta, \Sigma)$ be the $k-m$ vector of $\hat{w}^t_i(v^t; \beta, \Sigma)$, I solve

$$\min_{\alpha,\beta,\Sigma} S^2 = \sum_{t=1}^{T}\left(w_t - \hat{w}_t\left(v^t; \beta, \sum\right)\right)' A_t\left(w_t - \hat{w}_t\left(v^t; \beta, \sum\right)\right)$$

A_t is a weighting matrix that accounts for the "within time period" variance covariance $E(\tilde{\varepsilon}^t - h(v^t; \beta, \Sigma))(\tilde{\varepsilon}^t - h(v^t; \beta, \Sigma))'$.[6] This model assumes independence across time for the residual.[7] The regression results of this estimation are listed in table 11A.1.

Once the parameters are estimated, I can impute the virtual prices. Here, I rely on the method of Little and Rubin (1987). The true virtual price vector is

$$(11)\quad \begin{pmatrix} \pi^t_1 \\ \cdot \\ \cdot \\ \pi^t_m \end{pmatrix} = -\begin{pmatrix} \beta_{11} \ldots & \beta_{1m} \\ \ldots\ldots & \cdot \\ \ldots & \cdot \\ \beta_{m1} \ldots & \beta_{mm} \end{pmatrix}^{-1} \left[\begin{pmatrix} \alpha_1 \\ \cdot \\ \cdot \\ \alpha_m \end{pmatrix} + \begin{pmatrix} \beta_{1,m+1} \ldots & \beta_{1,k} \\ \cdot & \cdot \\ \cdot & \cdot \\ \beta_{m,m+1} \ldots & \beta_{m,k} \end{pmatrix}\begin{pmatrix} v^t_{m+1} \\ \cdot \\ \cdot \\ v^t_k \end{pmatrix} + \begin{pmatrix} \varepsilon^t_1 \\ \cdot \\ \cdot \\ \varepsilon^t_m \end{pmatrix}\right]$$

However, I only have parameter estimates in place of the true parameter, and I cannot observe the residuals. Therefore, to impute the virtual price I take

6. In this study, I estimate A_t by taking the 200 draws that are used to compute h and calculate a variance matrix around h.

7. Future study should focus on time dependence of the residual and on the possibility that historical prices influence demand.

Table 11A.1 **Parameter Estimates of Lower-Level Demand Systems by Brand with Selection Effects**

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
			General Mills Cheerios (R^2 = .840)				
Own and Cross Effect							
Item 1	–0.29						
	(0.019)						
Item 2	0.04	–0.28					
	(0.016)	(0.040)					
Item 3	0.10	0.14	–1.02				
	(0.037)	(0.051)	(0.100)				
Item 4	0.16	0.35	0.50	–1.25			
	(0.024)	(0.046)	(0.060)	(0.074)			
Item 5	–0.02	–0.24	0.28	0.24	–0.25		
	(0.036)	(0.045)	(0.069)	(0.040)	(0.089)		
Constant	0.25	0.83	0.91	1.66	–2.65		
	(0.080)	(0.095)	(0.168)	(0.114)	(0.232)		
			Kellogg's Corn Flakes (R^2 = .723)				
Own and Cross Effect							
Item 1	0.00						
	(0.001)						
Item 2	0.00	–0.25					
	(0.001)	(0.060)					
Item 3	0.00	0.27	–0.51				
	(0.004)	(0.065)	(0.106)				
Item 4	0.03	0.05	0.12	–0.91			
	(0.007)	(0.037)	(0.068)	(0.103)			
Item 5	–0.01	–0.11	0.05	0.57	–0.63		
	(0.006)	(0.037)	(0.058)	(0.071)	(0.081)		
Item 6	0.00	0.04	0.06	0.12	0.07	–0.31	
	(0.002)	(0.019)	(0.048)	(0.051)	(0.041)	(0.029)	
Item 7	–0.02	0.01	0.01	0.02	0.05	0.01	–0.09
	(0.003)	(0.010)	(0.030)	(0.050)	(0.045)	(0.011)	(0.023)
Constant	–0.08	0.71	1.05	0.35	–0.68	–0.03	–0.32
	(0.024)	(0.066)	(0.158)	(0.224)	(0.177)	(0.062)	(0.114)
			Kellogg's Raisin Bran (R^2 = .667)				
Own and Cross Effect							
Item 1	–0.07						
	(0.006)						
Item 2	0.07	–0.84					
	(0.033)	(0.108)					
Item 3	–0.06	0.48	–0.74				
	(0.039)	(0.093)	(0.138)				
Item 4	0.00	–0.01	0.030	–0.31			
	(0.007)	(0.037)	(0.052)	(0.019)			
Item 5	0.06	0.29	0.02	0.02	–0.39		
	(0.032)	(0.084)	(0.109)	(0.040)	(0.116)		
Constant	0.19	0.93	–1.10	0.33	0.65		
	(0.027)	(0.186)	(0.199)	(0.052)	(0.179)		

(*continued*)

Table 11A.1 (continued)

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
	Kellogg's Rice Krispies (R^2 =. 79)						
Own and Cross Effect							
Item 1	–0.71						
	(0.038)						
Item 2	0.18	–0.53					
	(7.427)	(13.14)					
Item 3	0.52	0.35	–0.88				
	(0.033)	(0.042)	(0.054)				
Constant	–0.12	1.49	–0.36				
	(0.145)	(0.108)	(0.141)				
	General Mills Honey Nut Cheerios (R^2 = .836)						
Own and Cross Effect							
Item 1	–0.37						
	(0.042)						
Item 2	0.00	–0.11					
	(0.019)	(0.012)					
Item 3	0.20	0.01	–0.68				
	(0.038)	(0.025)	(0.060)				
Item 4	0.16	0.10	0.48	–0.74			
	(0.036)	(0.026)	(0.056)	(0.065)			
Constant	0.03	0.42	–0.54	1.09			
	(0.124)	(0.128)	(0.156)	(0.199)			

Note: Standard errors in parentheses.

a random draw from the distribution of the parameter to account for the variance in the parameter estimates, and to account for the residual I take a random draw from $N(0_k, \hat{\Sigma})$ and then substitute these into equation (11).

References

Armknecht, P. A., and F. Maitland-Smith. 1999. Price imputation and other techniques for dealing with missing observations, seasonality, and quality change in price indices. IMF Working Paper no. WP/99/78. Washington, D.C.: International Monetary Fund.

Balk, B. M. 1998. On the use of the unit value indices as consumer price indices. In *Proceedings of the fourth meeting of the International Working Group on Price Indexes,* ed. Walter Lane, 112–20. Washington, D.C.: Bureau of Labor Statistics.

Diewert, W. E. 1995. Axiomatic and economic approaches to elementary price indexes. Discussion Paper no. 95-01. University of British Columbia, Department of Economics.

Feenstra, R. C. 1994. New product varieties and measurement of international prices. *American Economic Review* 84 (1): 157–77.

———. 1997. Generics and new goods in pharmaceutical price indexes: Comment. *American Economic Review* 87 (4): 760–67.
Feenstra, R. C., and E. W. Diewert. 2000. Imputation and price indexes: Theory and evidence from the international price program. Working Paper no. 00-12. University of California, Davis.
Gourieroux, C., and A. Monfort. 1996. *Simulation-based econometric methods.* London: Oxford University Press.
Hausman, J. 1996. Valuation of new goods under perfect and imperfect competition. In *The economics of new goods,* ed. T. Breshnahan and R. Gordon, 209–48. NBER Studies in Income and Wealth, vol. 58. Chicago: University of Chicago Press.
Little, R. J. A., and D. Rubin. 1987. *Statistical analysis with missing data.* New York: Wiley.
Reinsdorf, M. 1993. The effect of outlet price differentials in the U.S. Consumer Price Index. In *Price measurements and their uses,* ed. M. F. Foss, M. E. Manser, and A. H. Young, 227–54. NBER Studies in Income and Wealth, vol. 57. Chicago: University of Chicago Press.

Comment

Eduardo Ley

Congratulations to the author for a most interesting and informative paper. The paper deals with the issue of estimating price indexes when some price observations are missing because the quantity observed is zero—as is often the case with high-frequency highly disagreggated data. The approach followed in the paper is to estimate demand systems that can then be used to attribute virtual (shadow) prices, π_is, to the zero-consumption goods (Lee and Pitt 1986). The paper develops an innovative two-stage estimation method for estimating the virtual prices. I will not comment on the econometric methodology but will, rather, focus on aggregation and data issues.

Aggregation Issues

Equation (1) in the paper displays what looks like a standard consumer-choice problem:

$$(1) \qquad \max_{\mathbf{x}} U(\mathbf{x}) \quad s.t. \ \frac{\mathbf{px}}{y} = 1.$$

However, the problem treated in the paper is not a standard problem for two reasons (a) there is aggregation (separation) over goods—that is, y only represents the expenditure on cereal—and (b) there is aggregation over consumers—that is, y refers to the aggregated expenditure at a particular es-

Eduardo Ley is a senior economist at the International Monetary Fund Institute.

tablishment during one week. I would have liked to see these two aggregation issues addressed in the paper. (Furthermore, note that the relationship being estimated is not a proper consumer demand function but rather an "establishment sales function." Only after making further assumptions—for example, fixing the distribution of consumers across establishments—is it permissible to jump to demand functions.)

Although arguably the aggregation across goods could be easily handled—that is, through functional separability—the aggregation across consumers requires, in my opinion, further reflection. What is being assumed at the household level to give rise to this translog cost function at the retail-establishment level?

As an example, if we assumed translog preferences at the household level, the demand equation for good i by household h becomes

$$x_i^h = \left[\alpha_i^h + \sum_j \beta_{ij}^h \ln(p_i)\right]\frac{y^h}{p_i},$$

and—without making assumptions on the distribution of income—the aggregation condition requires linear Engel curves with identical slopes—that is, for all h: $\alpha^h = \alpha$ and $\beta_{ij}^h = \beta_{ij}$. The resulting household demands would generally result in positive shares for all cereal products—a highly unrealistic scenario. Because of the highly disaggregated data, I conjecture that most households only consume a small number of brands (typically one or two per household member), and the possibilities of substitution among brands are probably rather small, whereas the substitution among different-sized packages of the same product is large. I believe that aggregation over heterogeneous households would be a more realistic approach in this case.

In that vein, we could, for instance, get aggregate Cobb-Douglas consumption functions (a particular case of the translog) from individual Leontief-type preferences. In an extreme case, if households of type i buy 1 unit of x_i regardless of prices, the aggregate demands will be Cobb-Douglas:

$$\sum_h x_i^h = \left(\frac{H_i}{\sum_i H_i}\right)\frac{\sum_h y^h}{p_i},$$

where H_i is the number of type-i households. A better specification is probably given by linear preferences over x_is, which have the same or similar content but differ in package size and Leontief type over essentially different cereals. In any event, I would have liked to find in the paper some arguments providing some justification for what is ultimately done.

Generalized Axiom of Revealed Preference

There is, of course, a more basic question. Do the data satisfy the Generalized Axiom of Revealed Preference (GARP)? If the observed $(\mathbf{p}_t, \mathbf{x}_t)$ were

generated by a utility-maximizing aggregate consumer, the data must satisfy GARP:

$$\mathbf{x}_t R \mathbf{x}_s \Rightarrow \mathbf{p}_s \mathbf{x}_s \leq \mathbf{p}_s \mathbf{x}_t,$$

where R is the transitive closure of the directly revealed preferred relation, R^D,

$$\mathbf{p}_t \mathbf{x}_t \geq \mathbf{p}_t \mathbf{x} \Leftrightarrow \mathbf{x}_t R^D \mathbf{x}.$$

If there are (large) violations to GARP it does not make much sense to worry about Slutsky symmetry implicit in the translog cost function—GARP is a necessary and sufficient condition for utility maximization. Thus, every maximizing consumer's behavior must satisfy GARP, and if the data satisfy GARP they can be interpreted as being generated by a utility-maximizing entity—see, for example, Samuelson (1948), and also Varian (1983) for tests on GARP. Jerison and Jerison (1999) relate violations of the Slutsky conditions to sizes of revealed preference conditions—the inconsistencies measured by the highest possible minimum rate of real income growth along revealed preference cycles in a particular region.

Therefore, the data should be checked for GARP before one attempts to estimate any demand functions. If the data violate the restrictions implied by the consumer optimization model, there is little justification in using that model to describe them.

Missing Data Are Observable

The motivation of the paper is that some weekly data pairs,

$$(x_{it}, p_{it})$$

are completely missing whenever $x_{it} = 0$; that is, p_{it} is not observed whenever $x_{it} = 0$. However, the reporting retail establishment does have available a much richer data set.

The establishment knows p_{it} regardless of the recorded sales whenever the stock of the product at the end-period, s_{it}, is not zero. When $s_{it} = 0$, provided that some sales were made during that week, the desired price data would also be available.

It follows that the problems raised in the paper can be easily circumvented if the reported data become

$$(x_{it}, p_{it}, s_{it})$$

whenever $s_{it} > 0$ or $x_{it} > 0$. Still, the problem of which price to use would arise when $s_{it} = x_{it} = 0$. Nevertheless, this case would be equally problematic for the method developed in the paper.

References

Jerison, David, and Michael Jerison. 1999. Measuring consumer inconsistency: Real income, revealed preference, and the Slutsky matrix. Discussion Paper no. 99-01. State University of New York at Albany, Department of Economics.

Lee, Lung-Fei, and Mark M. Pitt. 1986. Microeconometric demand systems with binding nonnegativity constraints: The dual approach. *Econometrica* 54 (5): 1237–42.

Samuelson, Paul. 1948. Consumption theory in terms of revealed preference. *Economica* 15:243–53.

Varian, Hal R. 1983. Non-parametric tests of consumer behaviour. *Review of Economic Studies* 50 (1): 99–110.