A Framework for Large Scale Use of Scanner Data in the Dutch CPI

Jan de Haan

Statistics Netherlands and Delft University of Technology

Ottawa Group, 20-22 May 2015

The basic idea

“Ideally, to make the production process as efficient as possible, a limited number of fully or semi-automated methods would be used. The purpose of this paper is to propose a framework supporting these plans. Our basic aggregation formula is what we refer to as a ‘quality-adjusted unit value index’, which is equal to the value index divided by a quantity index that is defined as the ratio of quality-adjusted or standardized quantities. Time dummy regression models play an important role in the estimation of the quality-adjustment (standardization) factors. There are two extreme cases. If information on all relevant item characteristics is available, then the use of time dummy hedonic models is preferred. When characteristics information is lacking, the use of time-product dummy (fixed-effects) models is proposed.”

(De Haan, 2015)

Outline

- Expanding the use of transactions (and online) data
- The framework
  - Starting point: retail chains
  - Unit values, unit value indexes and quality-adjusted unit value indexes
  - Estimating the quality-adjustment factors
    - weighted time dummy hedonic regressions
    - weighted time-product dummy (fixed effect) regressions
  - Accounting for revisions: two splicing methods
- An example using NZ consumer electronics scanner data
- Fixed effects and lack of matching
- Online data
- Conclusions

Expanding the use of transactions (and online) data

- Budget cuts - reduction of field price collection
- Expanding the use of ‘secondary data’ - new project just started
  - scanner data provided by retail chains (supermarkets, drugstores, DIY stores, department stores, etc.)
  - online data obtained through web-scraping
- New methods required
  - Limited number of different methods
  - Lower levels: trends more important than short-term changes
  - Weighted (superlative-type) indexes
  - Hedonic quality adjustment
- Focus on multilateral approaches; maximum use of matches in the data while being (approximately) free of chain drift

Retail chains and elementary aggregates

- Elementary aggregates: cross-classification of store types and product categories
- Each retail chain (and field collection) treated as a separate store type
- Product categories: COICOP classification at publication level
- Below lowest publication level: chain-specific elementary aggregates
- Typically a single price index method used for a particular chain (not required – type of product more important than type of chain)
- Most elementary price indexes currently unweighted
- Aggregation within and across chains: scanner data where possible

Unit value index

Time periods $t = 0,...,T$, where 0 is the base period or starting period of the time series to be constructed

Prices $p_i^0 .... p_i^T$, quantities $q_i^0 .... q_i^T$, expenditure shares $s_i^0 .... s_i^T$

(Dynamic) sample of items purchased/sold $S^0 .... S^T$

For a homogeneous item, the average transaction price or unit value is the appropriate measure of price and the unit value index

$$P_{UV}^{0t} = \frac{\sum_{i \in S^t} p_i^t q_i^t \Big/ \sum_{i \in S^1} q_i^t}{\sum_{i \in S^0} p_i^0 q_i^0 \Big/ \sum_{i \in S^0} q_i^0} = \left[\frac{\sum_{i \in S^t} s_i^t (p_i^t)^{-1}}{\sum_{i \in S^0} s_i^0 (p_i^0)^{-1}} \right]^{-1}$$

is the appropriate measure of price change between periods 0 and *t* $(t = 1,...,T)$

Quality-adjusted unit value index

Unit value index is not appropriate for heterogeneous products

Standardization: quantity of each item *i* expressed in units of an arbitrary base item *b* using standardization factors or quality-adjustment factors

Equivalent to adjusting the price of each item for difference in quality with the base item: $\tilde{p}_i^t = p_i^t / \lambda_{i/b}$

Quality-adjustment factors $\lambda_{i/b}$ assumed constant across time

Quality-adjusted unit value index

$$P_{QAUV}^{0t} = \frac{\left[\sum_{i \in S^t} s_i^t (\tilde{p}_i^t)^{-1} \right]^{-1}}{\left[\sum_{i \in S^0} s_i^0 (\tilde{p}_i^0)^{-1} \right]^{-1}}$$

Quality-adjusted unit value index and quantity index

Implicit quantity index

$$Q^{0t} = \frac{\sum_{i \in S^t} p_i^t q_i^t}{\sum_{i \in S^0} p_i^0 q_i^0} \frac{1}{P_{QAUV}^{0t}} = \frac{\sum_{i \in S^t} \lambda_{i/b} q_i^t}{\sum_{i \in S^0} \lambda_{i/b} q_i^0}$$

Simple ratio of quantities expressed in constant-quality units; easy to interpret – change in number of quality-adjusted sales

Quantity index simplifies to the ratio of number of (unadjusted) sales when all items or of the 'same quality'

Quantity index and quality-adjusted unit value index are both transitive

Quantity index satisfies identity test in matched-item context but quality-adjusted unit value index doesn't

Estimating the quality-adjustment factors

Multilateral regression-based approach

Time dummy regressions based on pooled data for periods $t = 0,\ldots,T$

Weighted least squares: expenditure shares serve as weights to reflect items' economic importance

Model 1) Time dummy hedonic model

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{k=1}^{K} \beta_k z_{ik} + \varepsilon_i^t$$

with item characteristics z_{ik} and time dummy variables D_i^t

Quality-adjustment factors estimated by

$$\hat{\lambda}_{i/b} = \hat{p}_i^0 / \hat{p}_p^0 = \hat{p}_i^t / \hat{p}_b^t = \exp\left[\sum_{k=1}^{K} \hat{\beta}_k (z_{ik} - z_{bk})\right]$$

Time dummy hedonic index

Using $\hat{\tilde{p}}_i^0 = p_i^0 / \hat{\lambda}_{i/b}$ and $\hat{\tilde{p}}_i^t = p_i^t / \hat{\lambda}_{i/b}$, the time dummy index can be written as

$$\hat{P}_{TD}^{0t} = \frac{\prod_{i \in S^t} (\hat{\tilde{p}}_i^t)^{s_i^t}}{\prod_{i \in S^0} (\hat{\tilde{p}}_i^0)^{s_i^0}}$$

Time dummy index: ratio of expenditure-share weighted geometric means of quality-adjusted prices

Quality-adjusted unit value index: ratio of expenditures-share weighted harmonic means of (the same) quality-adjusted prices

The two indexes are transitive, hence free of chain drift, and independent of choice of base item

Time dummy index and quality-adjusted unit value index

Relation between the two indexes

$$\hat{P}_{QAUV}^{0t} = \hat{P}_{TD}^{0t} \left[\frac{\sum_{i \in S^0} s_i^0 \exp(u_i^0)}{\sum_{i \in S^t} s_i^t \exp(u_i^t)} \right]$$

$u_i^t = \ln(\hat{p}_i^t / p_i^t)$ are regression residuals

Bracketed factor

- changes the ‘geometric quality-adjusted unit value index’ (i.e., the time dummy index) into the desired arithmetic quality-adjusted unit value index

- depends on variance of residuals and is expected to be very small

Time-product dummy (fixed effects) index

Many scanner data sets: not enough characteristics information available for hedonic regressions

Model 2) Time-product dummy or fixed effects model

$$\ln p_i^t = \delta^0 + \sum_{t=1}^{T} \delta^t D_i^t + \sum_{i=1}^{N-1} \gamma_i D_i + \varepsilon_i^t$$

with dummy variables (indicators) D_i for the various items

The item-specific fixed effects γ_i can be viewed as approximations of the unknown hedonic price effects $\sum_{k=1}^{K} \beta_k z_{ik}$

Time-product dummy (fixed effects) index

The time-product dummy or fixed effects (FE) model is a special case of the time dummy hedonic model, so

- same choice of regression weights (expenditure shares)

- similar relation with quality-adjusted unit value index

- TPD/FE index is transitive/free from chain drift

TPD/FE index

- might be called a quality-adjusted price index because it uses a form of ’overlap pricing’ for new and disappearing items

- needs at least two observations for an item to be non-trivially included; for example, new items in the last period (*T*) are not included!

Decomposition of time dummy hedonic and FE indexes

Because the weighted time dummy indexes are transitive, they can be written as chained period-on-period indexes

A single index movement can be decomposed as follows:

adjacent-period Tornqvist index

X

effect of disappearing items

X

effect of new items

X

‘residual factor’ (to make the chained index transitive)

Accounting for revisions

All multilateral indexes, including time hedonic, FE, and quality-adjusted unit value indexes, suffer from revisions

- when the sample period is extended and new data is added, the results for all periods will change

Rolling window approach

estimation window (fixed length) is moved forward and indexes (re-) estimated

Two issues:

- how should the estimates from the most recent window be linked to the existing time series, i.e. what is the preferred splicing method?

- what is the optimal window length?

Two splicing methods

Every splicing method impairs the transitivity property of multilateral price indexes, so chain drift in the linked time series cannot be completely ruled out

1) movement splice

after moving forward the window one month and re-estimating the model, the most recently estimated month-on-month movement of the index is spliced on to the existing time series

2) (Frances Krsinich’s) window splice

splices the entire newly estimated 13-month series on to the index level pertaining to 12 months ago

(by construction no chain drift in annual changes)

Two splicing methods

Movement splice is typically used for time dummy hedonic indexes (and GEKS indexes, an alternative multilateral approach with no characteristics information)

Window splice is probably more useful for FE indexes

Choice of estimation window length

- at least 5 quarters (or 13 months) to include strongly seasonal items

- not too long; assumption of fixed (underlying) characteristics parameters

With window splicing, the estimation window may differ from the splicing window

Quality-adjusted unit value indexes: an example

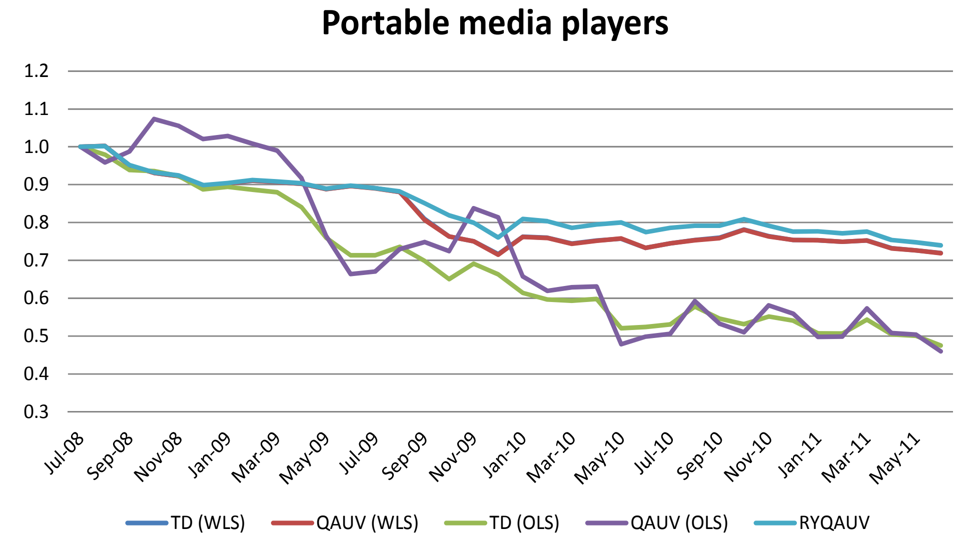
- New Zealand scanner data from market research company GfK
- Seven consumer electronics products
- Monthly data from mid-2008 to mid-2011
- Close to full coverage of New Zealand consumer market
- Items defined as unique combination of brand, model and available set of physical characteristics
- Data aggregated across outlet types
- High degree of churn (new and disappearing items)

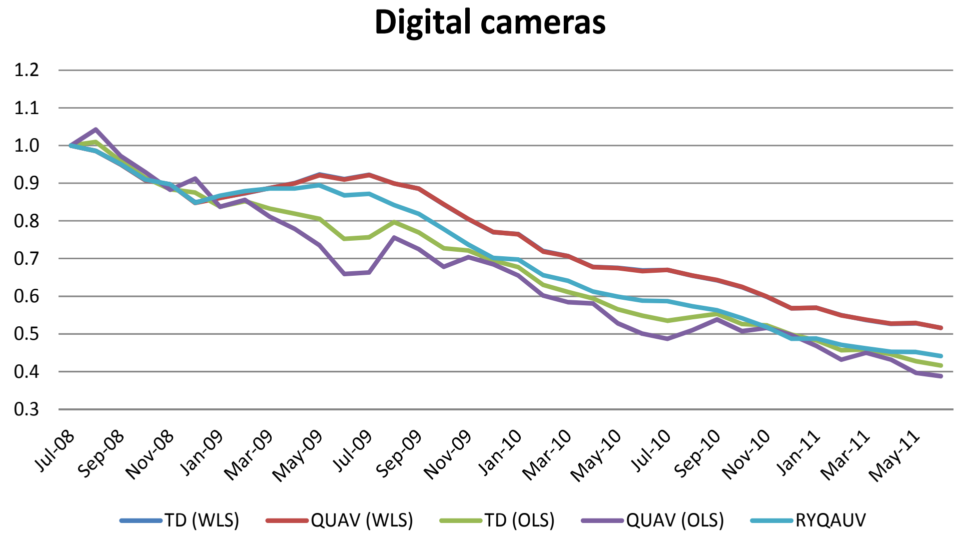
Quality-adjusted unit value indexes: an example

- (Adjusted) R square values for WLS time dummy hedonic regressions range from 0.964 (DVD players) to 0.989 (portable media players)
- High R squares are partly due to aggregation over goods (barcodes) with identical characteristics

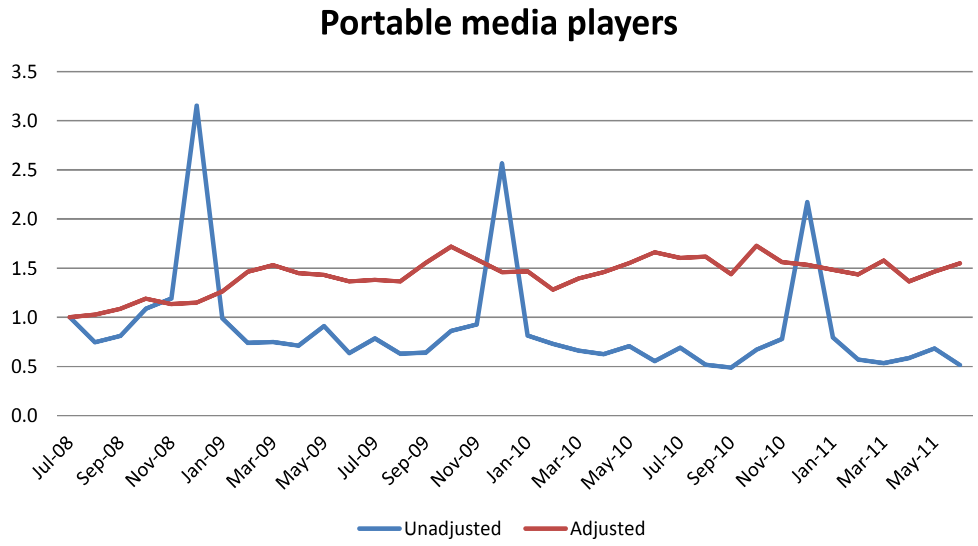
- We also estimated time dummy models using OLS, the resulting time dummy indexes and corresponding quality-adjusted unit value indexes
- (Adjusted) R squares for OLS regressions range from 0.859 to 0.913 (digital cameras)

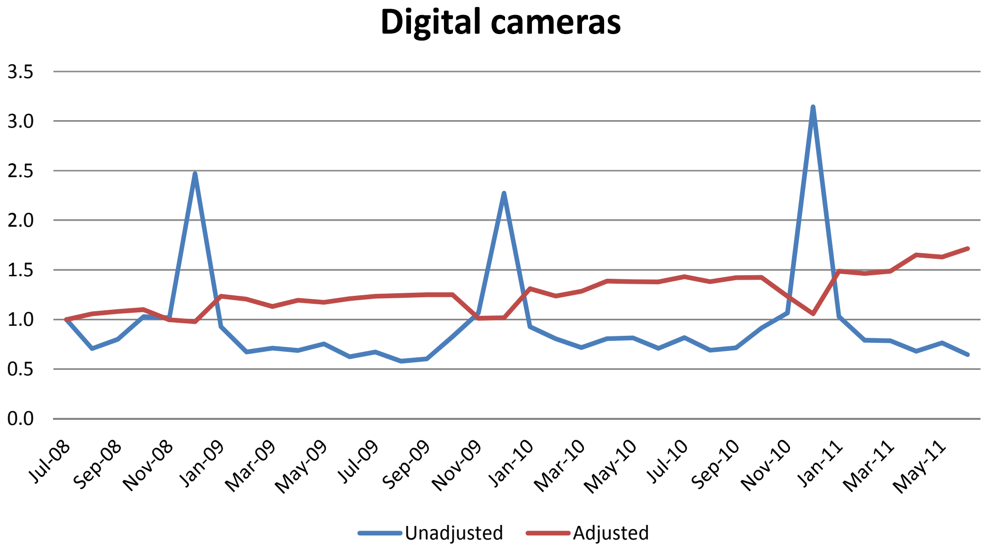
Time dummy hedonic and quality-adjusted unit value indexes





Indexes of number of sales and quality-adjusted sales





Fixed effects and lack of matching

EAN (GTIN) can be too detailed to be useful as item identifier

- different EANs may relate to the ‘same’ item

- lack of matching over time

- rate of item churn overestimated

- hidden price changes will be missed in matched-model indexes, including FE indexes

Aggregating across EANs pertaining to the ‘same’ item required – in Australia: SKU

Not a big problem for time dummy hedonic indexes (although results will be increasingly model-based)

Online data

Prices collected from retailers’ websites via web scraping

Many issues involved

- coverage of items sold (websites versus physical stores)
- price differences between websites and stores?
- online versus offline purchases
- characteristics information?
- changes in websites

Quantities/expenditures are unobservable

- no unit values across the month or quarter, just ‘average prices’
- quality-adjusted unit values and weighted indexes not possible

Conclusions

Statistics Netherlands: increasing use of scanner data and online data

For efficiency reasons: limited number of fully or semi-automated methods

Regression-based methods most promising

- time dummy hedonic models (scanner data)
- time-product dummy / FE effects models (scanner data, online data)

Proposal: quality-adjusted unit value indexes rather than initial indexes (scanner data)

New project started – research and implementation; international collaboration?