Scanner Data, Product Churn and Quality Adjustment

June 9, 2023

Geneva, United Nations.

Erwin W. Diewert

(The University of British Columbia & UNSW)

and

Chihiro Shimizu

(Hitotsubashi University)

1. Introduction.

- **CPI Manual Chapter 8: Quality Adjustment.**

- An increasing number of business firms are willing to share **their price and quantity data on their sales of consumer goods and services** to a ***statistical office***.
- These data are often referred to as **scanner data**.
- Some scanner data involves ***high technology products*** which are characterized by ***product churn***; i.e., the rapid introduction of new models and products and the short time that these new products are sold on the marketplace.

- This study will look at possible methods that statistical offices could use for ***quality adjusting this type of data***.

2. Hedonic Regressions and Utility Theory: The Time Product Dummy Hedonic Regression Model.

- The problem of adjusting the prices of similar products due to ***changes in the quality*** **of the products should be related to the usefulness or utility of the products to purchasers**.

- ***A hedonic regression*** is typically based on regressing a product price (or a transformation of the product price) on the amounts of the various price determining characteristics of the product.

- **Disadvantages: Misspecification Problem.**

- An alternative hedonic regression model may be based on regressing the ***product prices on product dummy variables***.

- Each of these hedonic regression models can be related to ***specific functional forms for purchaser utility functions***.

- **"Hedonic Regressions and Utility Theory".**

3. The Time Dummy Hedonic Regression Model with Characteristics Information.

- It is assumed that there are N products that are available over a window of T periods.

- We assume that the quantity aggregator function for the N products is the linear function, $\mathbf{f(q) \equiv \alpha \cdot q = \sum_{n=1}^{N} \alpha_n q_n}$ where q_n is the quantity of product n purchased or sold in the period under consideration and α_n ***is the quality adjustment factor*** for product n.

- ***What is new is the assumption that the quality adjustment factors are functions of a vector of K characteristics of the products.***

- Thus it is assumed that product n has the vector of characteristics $\mathbf{z^n \equiv [z_{n1},z_{n2},...,z_{nK}]}$ for n = 1,...,N. We assume that this information on the characteristics of each product has been collected.
- The new assumption is that **the quality adjustment factors α_n are functions of the vector of characteristics z^n for each product and the same function, g(z) can be used to quality adjust each product;** i.e., we have the following assumptions:
- (29) $\alpha_n \equiv g(z^n) = g(z_{n1},z_{n2},...,z_{nK})$; n = 1,...,N.
- Thus each product n has its own unique mix of characteristics z^n but the ***same function g*** can be used to determine the ***relative utility to purchasers of the products***.

4. Laptop Data for Japan and Sample Wide Hedonic Regressions Using Characteristics.

4.1 The Laptop Data and Some Preliminary Price Indexes

- We obtained data from a private firm that collects price, **quantity and characteristic information on the *daily sales* of digital device across Japan, 2015-2022. *Scanner data and amazon*.**

- Laptop computer and the 24 months in the years 2020 and 2021.

- Thus the prices and quantities **are p_{tn} and q_{tn}** where **p_{tn} is the average monthly (unit value) price for product n in month t in Yen and q_{tn} is the number of product n units sold**.

- **CLOCK** is the clock speed of the laptop. The mean clock speed was 1.94 and the range of clock speeds was 1 to 3.4. The larger is the clock speed, the faster the computer can make computations.
- **MEM** is the memory capacity for the laptop. The mean memory size was 8188.9. There were only 4 clock speeds listed in our sample: 4096, 8192 and 16,384.
- **SIZE** is the screen size of the laptop. The mean screen size (in inches) was 14.49. There were 10 distinct screen sizes in our sample: 11.6, 12, 12.5, 13.3, 14, 15.4, 15.6,16, 16.1 and 17.3.

- **PIX** is the number of pixels imbedded in the screen of the laptop. The mean number of pixels was 24.82. There were only 10 distinct number of pixels in our sample: 10.49, 12.46, 12.96, 20.74, 33.18, 40.96, 51.84, 55.30, 58.98 and 82.94.

- **HDMI** is the presence (HDMI = 1) or absence (HDMI = 0) of a HDMI terminal in the laptop. If HDMI =1, then it is possible to display digitally recorded images without degradation.

- **BRAND** is the name of the manufacturer of the laptop. In the data file, BRAND takes on the values 1-12 but the second brand is not present in 2020-2021 so we have only 11 brands in our sample.
- BRAND is frequently used as an explanatory variable in a hedonic regression as a proxy for company wide product characteristics that may be missing from the list of explicit product characteristics that are included in the regression.

- 11 variables in vectors of dimension 2639: OBS (runs from 1 to 2639), **TD, JAN, CLOCK, MEM, SIZE, PIX, HDMI, BRAND**, Q and P.

- The information in the column vectors TD and JAN were used to generate **24 time dummy variables**, D1, D2, …, D24 and 366 product dummy variable vectors, DJ1, DJ2,…,DJ366.

4.2 A Hedonic Regression with *Clock Speed* as the Only Characteristic

- Of course, the price indexes P_A^t and P_{UV}^t make no adjustments for changes in the average quality of laptops over time. Thus we now consider hedonic regression models of the type defined by equations (38) in the previous section.

- We start our analysis by **regressing the price vector P on the time dummy variables D_1.,,,,D_{24}** ***and dummy variables for the clock speed*** of each laptop that was sold during the sample period.

- ***Our first hedonic regression*** **sets the dependent variable vector equal to the logarithms of the product price vector P (which we denote by lnP) and the vectors in the matrix of independent variables are the time dummy variable vectors D_2, D_3,…,D_{24} and the *new 7 clock speed dummy variable* vectors D_{C1}, D_{C2}, …, D_{C7}.**

- **(56)** $\ln P = \sum_{t=2}^{24} \rho_t D_t + \sum_{j=1}^{7} b_{Cj} D_{Cj} + e$

  where e is an error vector of dimension 2639.

4.3 A Hedonic Regression that Added *Memory Capacity* as an Additional Characteristic

- We add memory capacity as another price determining characteristic of a laptop. There were **only 3 sizes of memory capacity** (the variable MEM in the Data Appendix): 4096, 8192 and 16384. Construct dummy variable vectors of dimension 2639 for each value of MEM.
- Denote these vectors as D_{M1}, D_{M2} and D_{M3}. The new log price time dummy characteristic hedonic regression is the following counterpart to (58):
- **(61)** $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + e.$

4.4 A Hedonic Regression that Added *Screen Size* as an Additional Characteristic.

- There were 10 different screen sizes (in units of 10 inches) in our sample of laptop observations. This variable is listed as SIZE in the Data Appendix. The 10 screen sizes in our sample were: 1.16, 1.2, 1.25, 1.33, 1.4, 1.54, 1.56, 1.6, 1.61 and 1.73. The usual commands were used to generate **10 dummy variables** for this characteristic.
- **New Groups 1 to 7 aggregated old groups 1-3, 4-8, 8-9, 10-12, 13-15, 16-18 and 19-25** respectively. Thus the new dummy variable vector D_{C1} equals the sum of the old vectors $D_{C1} + D_{C2} + D_{C3}$, the new D_{C2} equals the sum of the old vectors $D_{C4} + D_{C5} + D_{C6} + D_{C7} + D_{C8}$ and so on.

- The new log price time dummy characteristic hedonic regression is the following counterpart to (61):

- **(62)** $\ln P = \sum_{t=2}^{24} \rho_t D_t + b_0 ONE + \sum_{j=2}^{7} b_{Cj} D_{Cj} + \sum_{j=2}^{3} b_{Mj} D_{Mj} + \sum_{j=2}^{7} b_{Sj} D_{Sj} + e.$

- The log of the likelihood function was – 202.270, a gain of 446.667 log likelihood points for adding 6 new screen size parameters.

4.5 A Hedonic Regression that Added *Pixels* as an Additional Characteristic.

- There were **10 different numbers of pixels** in our sample of laptop observations. A larger number of pixels per unit of screen size will lead to clearer images on the screen and this may be utility increasing for purchasers.
- The pixel variable is listed as PIX in the Data Appendix. There were 10 different PIX sizes in our sample.
- The 10 sizes (in transformed units of measurement) were: 1.049, 1.246, 1.296, 2.074, 3.318, 4.096, 5.184, 5.530, 5.898 and 8.294.

- We ended up with 5 pixel groups: the new group 1 combined groups 1, 2 and 3; old group 4 became the new group 2, old groups 5 and 6 were combined to give us the new group 3, old groups 7, 8 and 9 were combined to be the new group 4 and the old group 10 became the **new group 5**.
- Denote the new pixel dummy variable vectors as D_{P1}-D_{P5}. The number of observations in each of these new pixel cells was 330, 1769, 405, 96, 39.
- The new log price time dummy characteristic hedonic regression is the following counterpart to (62):
- **(63)** $\ln P = \sum_{t=2}^{24} \rho_t D_t + b_0 ONE + \sum_{j=2}^{7} b_{Cj} D_{Cj} + \sum_{j=2}^{3} b_{Mj} D_{Mj} + \sum_{j=2}^{7} b_{Sj} D_{Sj} + \sum_{j=2}^{5} b_{Pj} D_{Pj} + e.$

4.6 A Hedonic Regression that Added *HDMI* as an Additional Characteristic.

- The dummy variable that indicates the presence of HDMI in the laptop has already been generated and is listed in the Data Appendix as the column vector HDMI. Denote this column vector as D_{H2} in the following hedonic regression which adds D_{H2} to the other regressor columns in (63):

- **(64)** $\ln P = \sum_{t=2}^{24} \rho_t D_t + b_0 ONE + \sum_{j=2}^{7} b_{Cj} D_{Cj} + \sum_{j=2}^{3} b_{Mj} D_{Mj} + \sum_{j=2}^{7} b_{Sj} D_{Sj} + \sum_{j=2}^{5} b_{Pj} D_{Pj} + D_{H2} + e.$

4.7 A Hedonic Regression that Added *Brand* as an Additional Characteristic.

- BRAND takes on values from 1 to 12 but there are no brands that correspond to the number 2 in our sample for the 24 months in the years 2020 and 2021.
- Here are the numbers of observations in each of the 12 BRAND categories: 4, 0, 3,101, 6, 235, 107, 389, 489, 439, 327, 479.
- We calculated the sample wide average price for each brand and re-ordered the brands according to their average prices with the lowest average price brands listed first and the highest average brand listed last.

- Add the column vectors D_{B2}-D_{B11} to the other regressor columns in (64) in order to obtain the following hedonic regression model:

- **(65)** $\ln P = \Sigma_{t=2}^{24} \rho_t D_t + b_0 ONE + \Sigma_{j=2}^{7} b_{Cj} D_{Cj} + \Sigma_{j=2}^{3} b_{Mj} D_{Mj} + \Sigma_{j=2}^{7} b_{Sj} D_{Sj} + \Sigma_{j=2}^{5} b_{Pj} D_{Pj} + D_{H2} + \Sigma_{j=2}^{11} b_{Bj} D_{Pj} + e.$

4.8 A *Weighted Time Dummy Characteristics* Hedonic Regression Model.

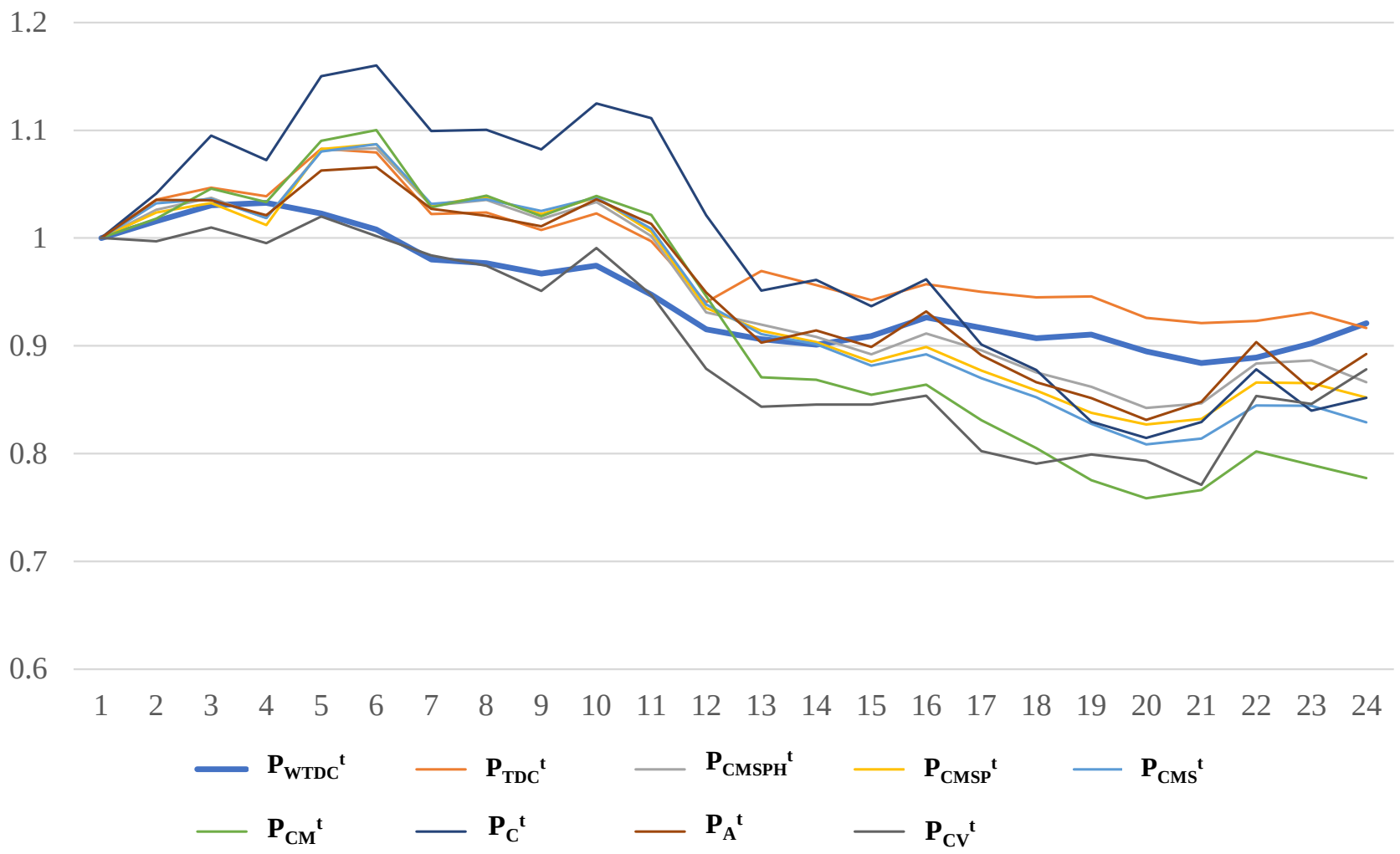
- Recall that the **expenditure share** that corresponds to purchased product n in month t is defined as $s_{tn} = p_{tn}q_{tn}/\Sigma_{j \in S(t)} p_{tj}q_{tj}$ for $t = 1,\ldots,24$ and $n \in S(t)$.

- To obtain the weighted counterpart to the hedonic regression, we just form a share vector of dimension 2639 that corresponds to the $\ln p_{tn}$ and then form a new vector of dimension 2639 that consists of the positive square roots of each s_{tn}.

Price Indexes

- **(1)P_{WTDC}^{t} :** ***Weighted Time Dummy Characteristics Price Index.***
- **(2) P_{TDC}^{t} :** Unweighted (or equally weighted) Time Dummy Characteristics Price Index.
- **(3) P_{C}^{t}:**A Hedonic Regression with Clock Speed as the Only Characteristic
- **(4)P_{CM}^{t}:** Clock Speed + ***Memory Capacity***.
- **(5)P_{CMS}^{t}:**Clock Speed+ Memory Capacity***+ Screen Size.***
- **(6)P_{CMSP}^{t}:**Clock Speed+ Memory Capacity+ Screen Size+ ***Pixels.***
- **(7)P_{CMSPH}^{t}:**Clock Speed+ Memory Capacity+ Screen Size+ Pixels
- + ***HDMI***

Chart 1: Weighted and Unweighted Time Product Dummy Price Indexes



5. The adjacent period time dummy Characteristics.

- There are ***two problems*** **with our "best" directly defined weighted hedonic price index using characteristics,** P_{WTDC}^{t}, which was defined in the previous section:
- **It is not a real time index; i.e., it is a *retrospective* index that is calculated using the data covering two years;**
- **It does not allow for gradual taste change on the part of purchasers.**
- These difficulties can be avoided if we restrict the number of months **T to be equal to 2. →Adjacent period Hedonic Regression.**
  - P_{ATDC}^{t} : ***Adjacent period equally weighted characteristics index***.
  - P_{WATDC}^{t} : ***Weighted Adjacent period characteristics index***.

- **Adjacent period Hedonic Regression.**

- ***Time dummy characteristic hedonic regression model***:

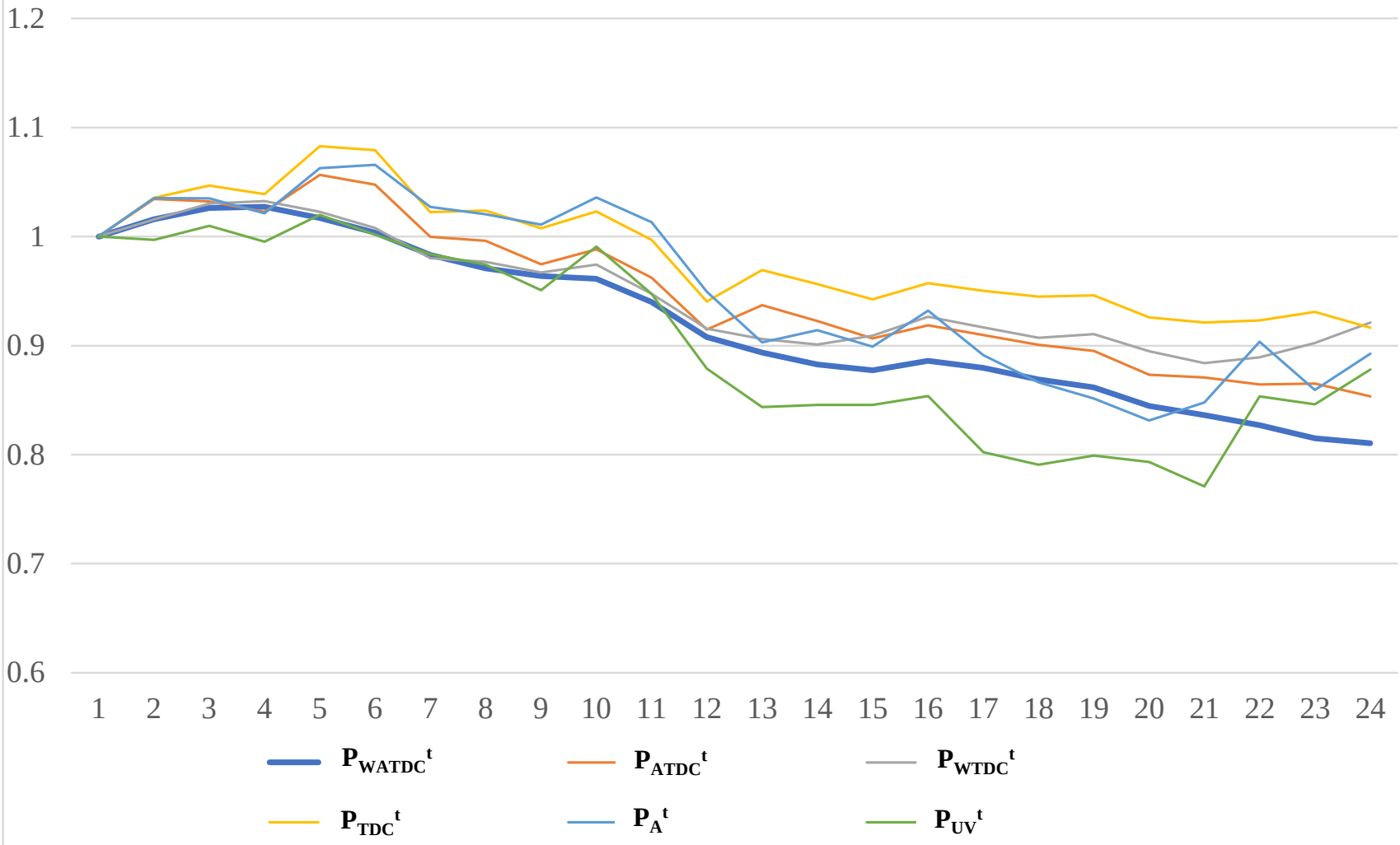
- **(75)** $\ln P = \rho_2 D_2 + b_0 ONE + \sum_{j=2}^{7} b_{Cj} D_{Cj} + \sum_{j=2}^{3} b_{Mj} D_{Mj} + \sum_{j=2}^{7} b_{Sj} D_{Sj} + \sum_{j=2}^{5} b_{Pj} D_{Pj} + b_{H2} D_{H2} + \sum_{j=2}^{11} b_{Bj} D_{Bj} + e$

- where lnP is now the vector of log prices for the products which were **sold only in months 1 and 2**.

Price Indexes

- **(1)P_{WATDC}^{t} :** ***Weighted Adjacent period characteristics index*****.**
- **(2)P_{ATDC}^{t} :** ***Adjacent period equally weighted characteristics index*****.**
- **(3)P_{WTDC}^{t} :** *Weighted Time Dummy Characteristics Price Index.*
- **(4)P_{TDC}^{t} :** Unweighted (or equally weighted) Time Dummy Characteristics Price Index.
- **(5)P_{A}^{t} : Average Price.**
- **(6)P_{UV}^{t} : Unit Value Price.**

Chart 3: Sample Wide and Adjacent Period Weighted and Unweighted Characteristics Price Indexes



- Here are some of the **advantages and disadvantages of the Weighted Adjacent Period Time Dummy Characteristics indexes P_{WATDC}^{t} over the Weighted Time Dummy Characteristics indexes P_{WTPC}^{t}:**
  - The adjacent period indexes **fit the data much better since each bilateral regression estimates a new set of quality adjustment parameters** whereas the panel regression approach fixes the quality adjustment parameters over the entire window of observations.
  - If the number of characteristics is large relative to the number of observations in a bilateral regression, **the estimates for the quality adjustment parameters could be unreliable which could lead to unreliable estimates for the price levels**.
  - The adjacent period methodology that allows the quality adjustment parameters to change every month means that **purchasers may not have stable consistent preferences over time** and some economists may object to this fact.

6. Time Product Dummy Variable Regression Models.

- This section also defined the **366 product dummy variable vectors** of dimension 2639, D_{J1}, …, D_{J366}. Define the vector of the logarithms of observed laptop prices as lnP as was done in previous sections.
- Then ***the unweighted Time Product Dummy regression model*** can be expressed as the following estimating equation for the log price levels ρ_2, ρ_3, …, ρ_{24} and the 366 product log quality adjustment factors β_1, β_2, …, β_{366}:

- **(77)** $\mathbf{lnP = \Sigma_{t=2}^{24} \rho_t D_t + \Sigma_{k=1}^{366} \beta_k D_{Jk} + e^t.}$

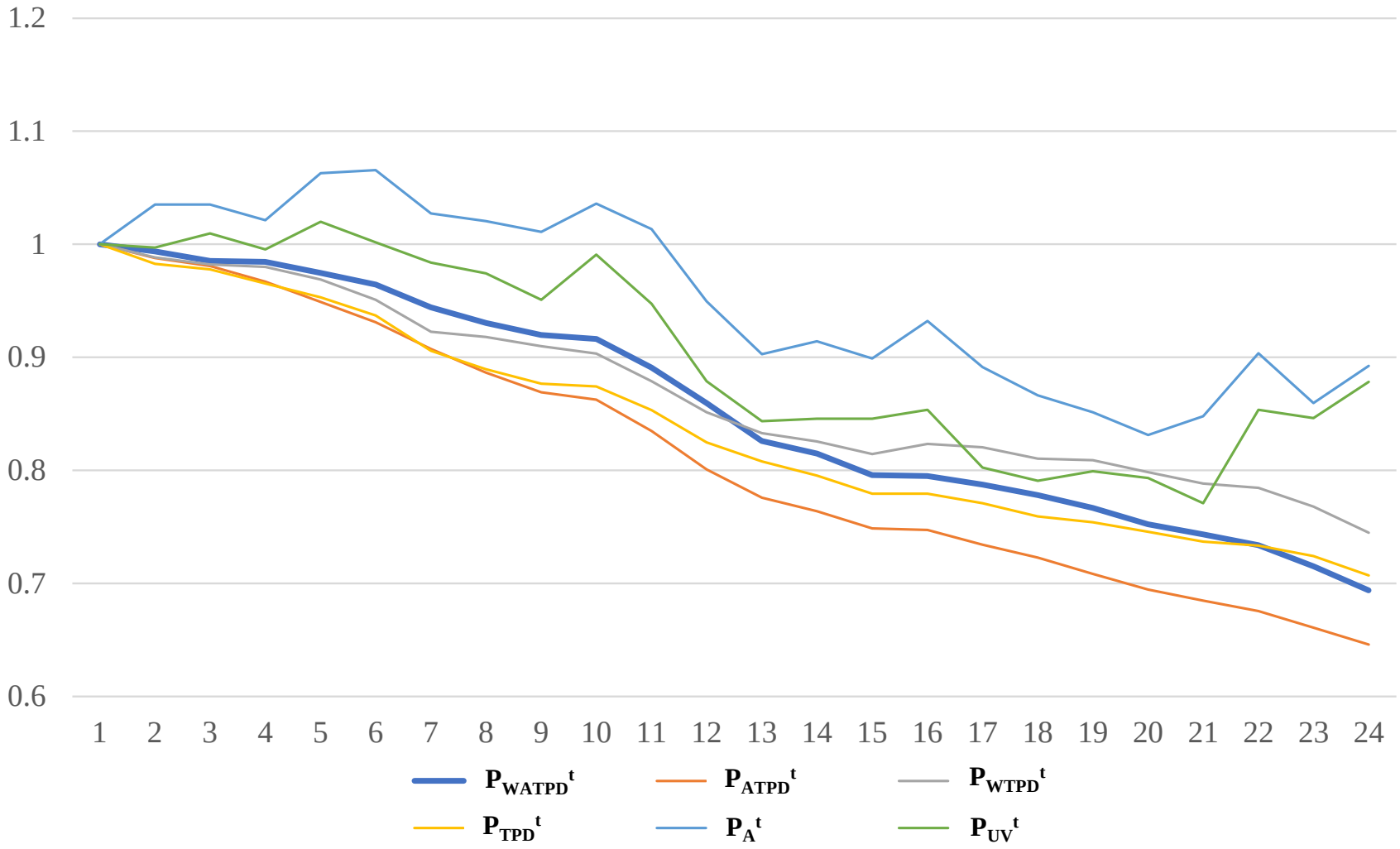
- ***Adjacent Period Time Product Dummy Price Indexes*** P_{ATPD}^{t} for t = 2, 3, …, 24.

- ***Weighted Adjacent Period Time Product Dummy Price Indexes,*** P_{WATPD}^{t}.
  for t = 2, 3, …, 24.

Price Indexes

- **(1)P_{WATPD}^{t} :** ***Adjacent Period Weighted Time Product Dummy Indexes .***
- **(2)P_{ATPD}^{t} :** ***Adjacent Period Time Product Dummy Price Indexes*** .
- **(3)P_{WTDC}^{t} :** *Weighted Time Dummy Characteristics Price Index.*
- **(4)P_{TPD}^{t} :** Unweighted (or equally weighted) Time Dummy Characteristics Price Index.
- **(5)P_{A}^{t} : Average Price.**
- **(6)P_{UV}^{t} : Unit Value Price.**

Chart 4: Sample Wide and Adjacent Period Weighted and Unweighted Time Product Dummy Price Indexes



- We ***prefer the Adjacent Period Weighted Time Product Dummy Indexes*** P_{WATPD}^t over their single regression counterpart indexes, the Weighted Time Product Dummy Indexes $\mathbf{P_{WTPF}^t}$ for two reasons:
  - (i) the regressions which generate the P_{WATPD}^t **fit the data much better than the single regression** which generated the P_{WTPD}^t and
  - (ii) the P_{WATPD}^t appear to be ***smoother*** than the P_{WTPD}^t. **Thus** $\mathbf{P_{WATPD}^t}$ **is our preferred index thus far.**

- Our preferred index, the ***Adjacent Period Weighted Time Product Dummy Index*** P_{WATPD}^{t}***, is a chained index*** and thus, it is subject to possible ***chain drift***.

- In order to reduce or eliminate possible ***chain drift***, we will calculate ***Predicted Share Price Similarity linked indexes***.

7. Similarity Linked Price Indexes for Laptops.

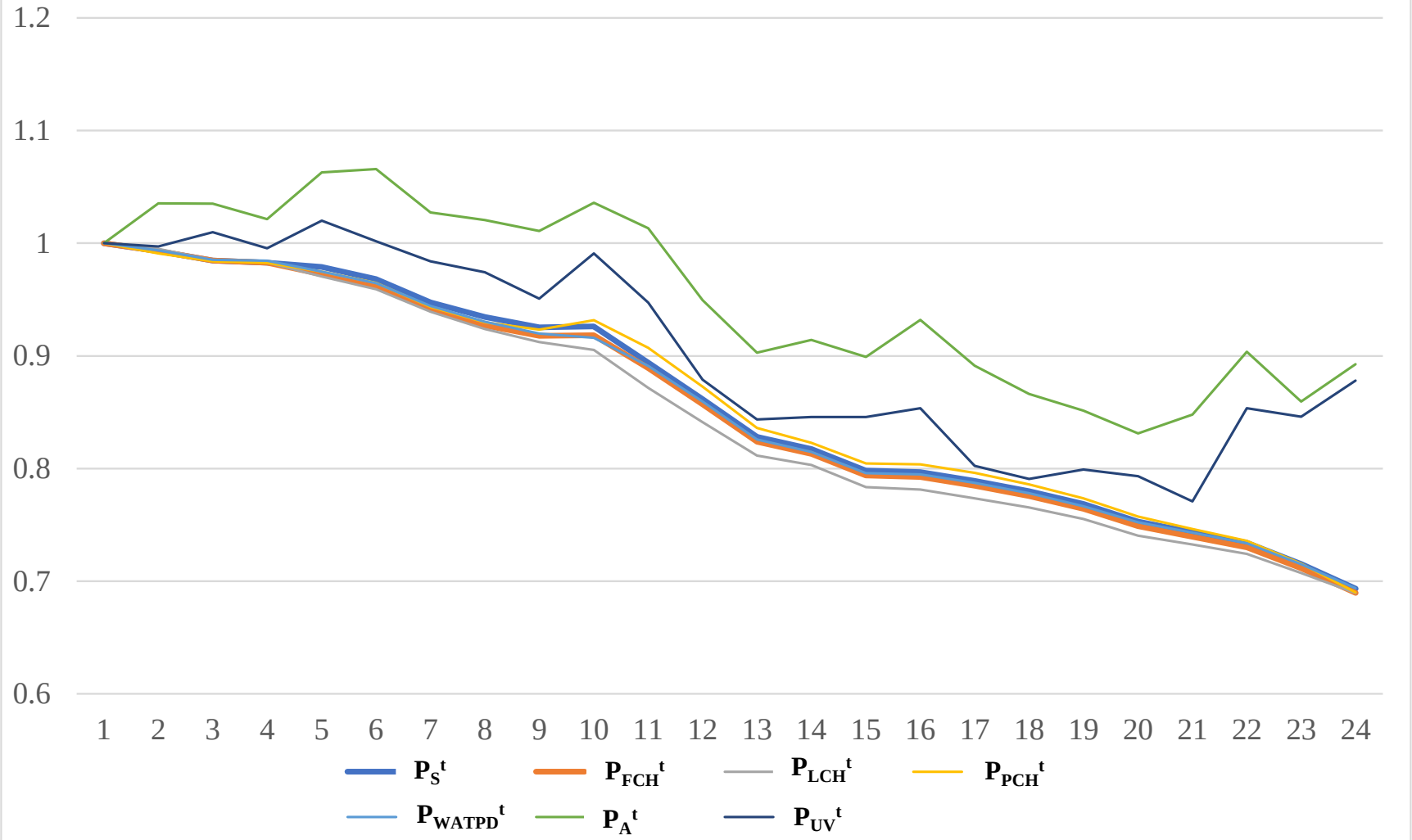
- **The *Predicted Share* method of linking months with the most similar structure of relative prices will be explained under the assumption that it is necessary to construct a price Index P^t in real time.**

- The *matched model Laspeyres and Paasche indexes*, $P_L(r,t)$ and $P_P(r,t)$, that relate the prices of month t to month r are defined as follows:
- (79) $P_L(r,t) \equiv \Sigma_{k \in S(r,t)} p_k^t q_k^r / \Sigma_{k \in S(r,t)} p_k^r q_k^r$; $1 \leq r, t \leq 24$;
- (80) $P_P(r,t) \equiv \Sigma_{k \in S(r,t)} p_k^t q_k^t / \Sigma_{k \in S(r,t)} p_k^r q_k^t$; $1 \leq r, t \leq 24$.

Price Indexes

- **(1)P_S^t:Predicted Share Similarity Linked indexes .**
- **(2)P_{FCH}^t :** Chained maximum overlap ***Fisher*** indexes.
- **(3)P_{LCH}^t :** Chained maximum overlap ***Laspeyres*** indexes.
- **(4)P_{PCH}^t :** Chained maximum overlap ***Paasche*** indexes.
- **(5)P_{WATPD}^t :**Weighted Adjacent Period Time Product Dummy Index.
- **(6)P_A^t : Average Price.**
- **(7)P_{UV}^t : Unit Value Price.**

Chart 5: The Predicted Share Similarity Linked Price Index and Other Comparison Price Indexes



- It can be seen that ***the similarity linked indexes P_S^t, the Chained Fisher maximum overlap indexes P_{FCH}^t*** **and the *Adjacent Period Weighted Time Product Dummy price indexes P_{WATPD}^t*** **are all extremely close to each other.**

- **These three indexes seem to be "*best*" for our particular application.**

- It can also be seen that the chained Laspeyres and Paasche indexes, P_{LCH}^t and P_{PCH}^t, are very close to our "***best***" indexes.

6.Conclusions.

- If quantity or expenditure weights are available in addition to price information, **then it is important to use these weights in the calculation of a weighted by economic importance price index**.
- Hedonic regressions that use amounts of product characteristics as independent variables in **the regressions are not recommended for two reasons:**
  - **(i) it is expensive to collect information on characteristics and**
  - **(ii) it is likely that some important price determining characteristics are not included in the list of characteristics**.

- ***The Adjacent Period Weighted Time Product Dummy index* is a preferred index provided** that:
  - (i) prices and quantities ***do not fluctuate violently*** from period to period due to product sales or strong seasonality and
  - (ii) the products in scope are thought to be ***close substitutes***.
- The ***Predicted Share Similarity Linked index*** is also a preferred index that should be satisfactory even if there are product sales or strong seasonality or if the products in scope are not close substitutes.
- **The disadvantages of this method are the complexity of the computations and the difficulty of explaining the method to the public**.

- In our particular application, our two preferred indexes were virtually identical.
  - **The chained maximum overlap Fisher indexes were also extremely close to our two preferred indexes and the chained maximum overlap Laspeyres and Paasche indexes were very close to our preferred indexes**.
  - We do not expect these close approximations to occur in other applications.
- Future works: New product effect.

W. Erwin Diewert,

University of British Columbia and University of New South Wales;
Email: erwin.diewert@ubc.ca

Chihiro Shimizu,

Hitotsubashi University;

Email: c.shimizu@r.hit-u.ac.jp