Alternative Data and Hedonic Price Indexes in the U.S. Consumer Price Index: A Review of Recent Research

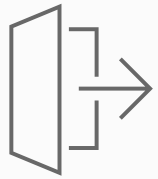
Craig Brown
Senior Economist | Bureau of Labor Statistics | Consumer Price Index
Jeremy Smucker
Economist | Bureau of Labor Statistics | Consumer Price Index

FCSM Conference
October 26, 2022

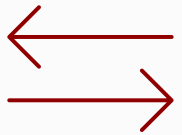
Hedonic Imputation

CPI's Matched-Model Methodology

Survey data is used to select a sample of goods that fall within an item definition



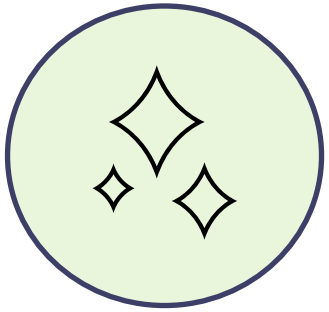
Once goods are selected, **price relatives** are calculated by comparing the previous period price to the current period



If a selected good is no longer available, it is replaced by a different good within the same item definition. Quality Adjusted Relatives are calculated between new and old goods

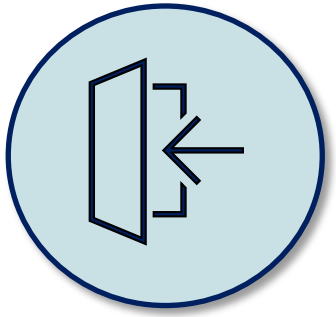
Price relatives from both continuing goods and substitutions are aggregated to form an index for a specific item

CPI Methodology - Problems



New Goods Problem

Occurs if the prices of new goods are not used in the construction of a price index and those prices are systematically lower/higher, on a quality-adjusted basis, than the prices of old items.



Entering & Exiting Goods

Goods in their first and last months of pricing show unique price movement. CPI's 6-month initiation cycle procedures mean we are not capturing the price behavior of entering goods

Product Life Cycle Effects

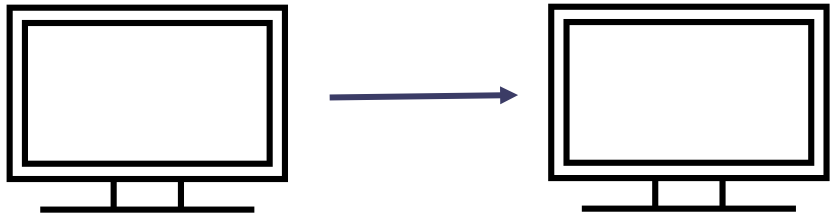
Many products decrease in price over their lifecycle. If a product ages, drops in price, and then rotates out of the sample without being replaced, the index may drop while average price in the sample increases

Hedonic Imputation

Hedonic Imputation is a price index methodology that compares **predicted** prices for products in the current and previous period.

Prices are predicted using hedonic regressions where price or log(price) is used as a dependent variable, and the product characteristics are the independent variables.

Matched Model

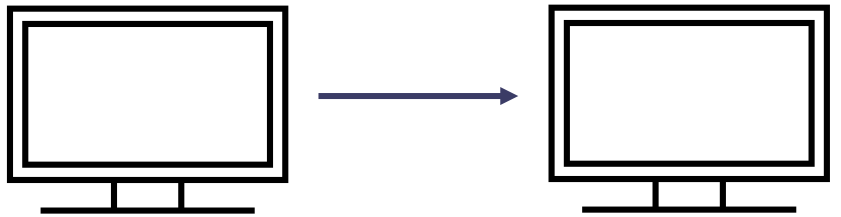


Collected 202201
price: 399.99

Collected 202202
price: 349.99

Observation Relative = 0.875

Hedonic Imputation



Predicted 202201
price: 381.88

Predicted 202202
price: 365.08

Observation Relative = 0.956

Problems Solved by Hedonic Imputation

- A quality adjusted index with no individual item replacement decisions
- Hedonic Imputation is not beholden to an item's product lifecycle like matched model
- Little analyst or data collection intervention, minimal respondent burden
- If we have a universe, we don't have to sample
- Hedonic Imputation imputes prices for entering or exiting goods, unlike current CPI procedures
- Hedonic Imputation innovations (Erickson & Pakes 2011) make it possible to control for unobserved characteristics

Model Selection

Classical Statistical Approach vs. Machine Learning Approach

Aspect	Statistical	Machine
Modeling	Rigid – A particular model is specified after thorough data analysis	Flexible – Few, if any assumptions made on underlying distribution
Inference	Explaining and understanding is the focus; Considers confidence intervals and parameter estimates	Prediction or classification is usually the focus
Variable Selection/ Model Fitting	Relies on a battery of model misspecification and goodness of fit tests	Variable selection and overfitting models can be problematic

Model Selection Process

- **Stepwise selection procedure** that uses the *F* statistic to gauge improvement in model fit

- **K-fold Cross Validation** predicted residual sum of squares statistic terminates the process
  - Results in more parsimonious models
  - Prevents overfitting the models
  - Reduces prediction error

Estimate Model on each Fold

1. Partition the data into K subsets (folds)
2. Estimate model for each K iteration with training data
3. Compute predicted RSS on validation folds for each K iteration
4. Sum of five predicted RSS is estimate of prediction error

Estimate Model on Full Data Set

5. Estimate model on full data set using selection criteria
6. Compute RSS for model with each “best candidate” variable added/removed
7. If model RSS score is greater than the “trained” model RSS score, selection process terminates

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS	F Value	Pr > F
0	Intercept		1	23.0463	0.00	1.0000
1	Num_Lines		2	8.2428	262.65	<.0001
2	log_hotdata		3	5.4414	76.17	<.0001
3	PREPAID_CONT		4	3.9586	61.66	<.0001
4	P_TMB		5	3.6136	17.11	<.0001
5	CallCanMex		6	3.5346	6.37	0.0127
6	P_BOO		7	3.4804*	3.93	0.0495

* Optimal Value of Criterion

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details

Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Entry	log_hsd	3.6002	>	3.4804
Removal	P_BOO	3.5346	>	3.4804

Corp14

Corp14 Data

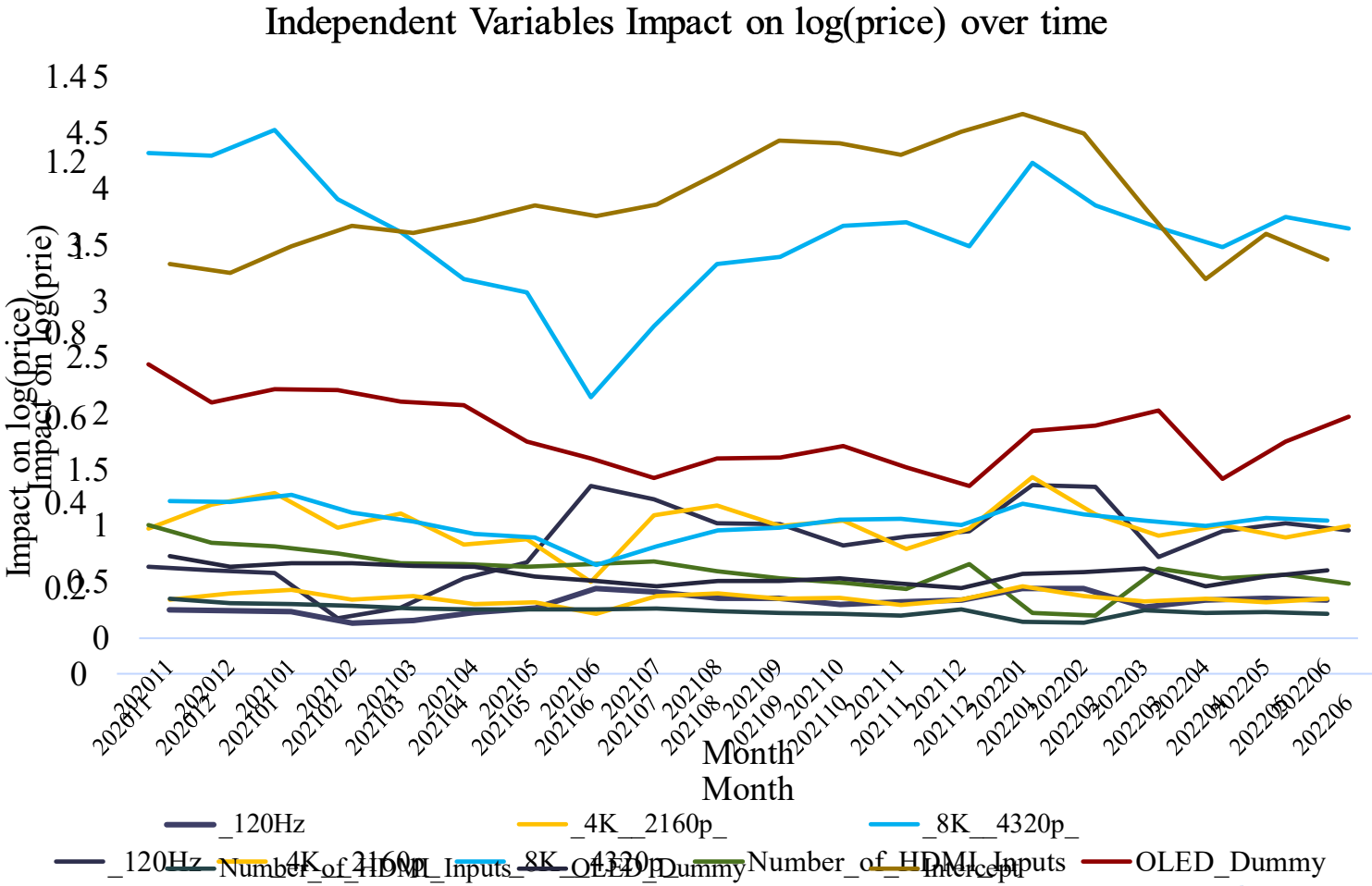
- API collected data from a consumer electronics retail store
- Prices collected 3 times per month
- 94 Unique Specification Variables
- 3rd party household survey data used to weight observations

CORP14 - Hedonic Model

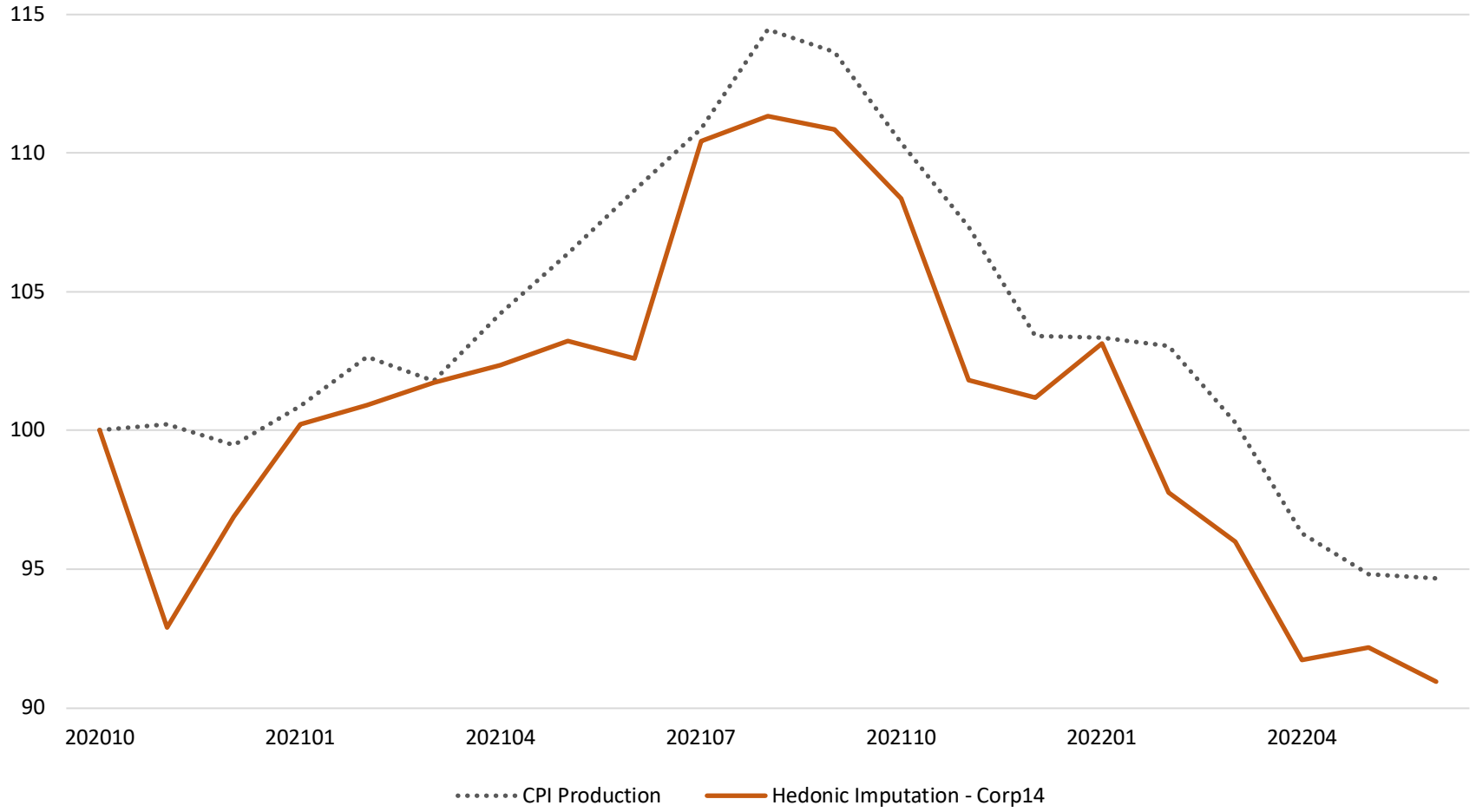
- Selected independent variables in the televisions model remained very consistent during the time period that we have CORP14 data (Oct 2020 – Jun 2022).
- Technology change for the item is minimal within the short time period we have data for.
- While relevant independent variables were constant across time periods, the impact of each independent variable on log(price) varied over time

CORP14 - Televisions Model

- Newer technology (8k, OLED, HDMI 2.1 ports) become cheaper to produce over the time period
- 4k resolution (now by far the highest selling resolution) and refresh rate were more stable
- But the effects of each individual coefficient don't tell the whole story…



Index Results - Televisions



Wireless Telephone Service

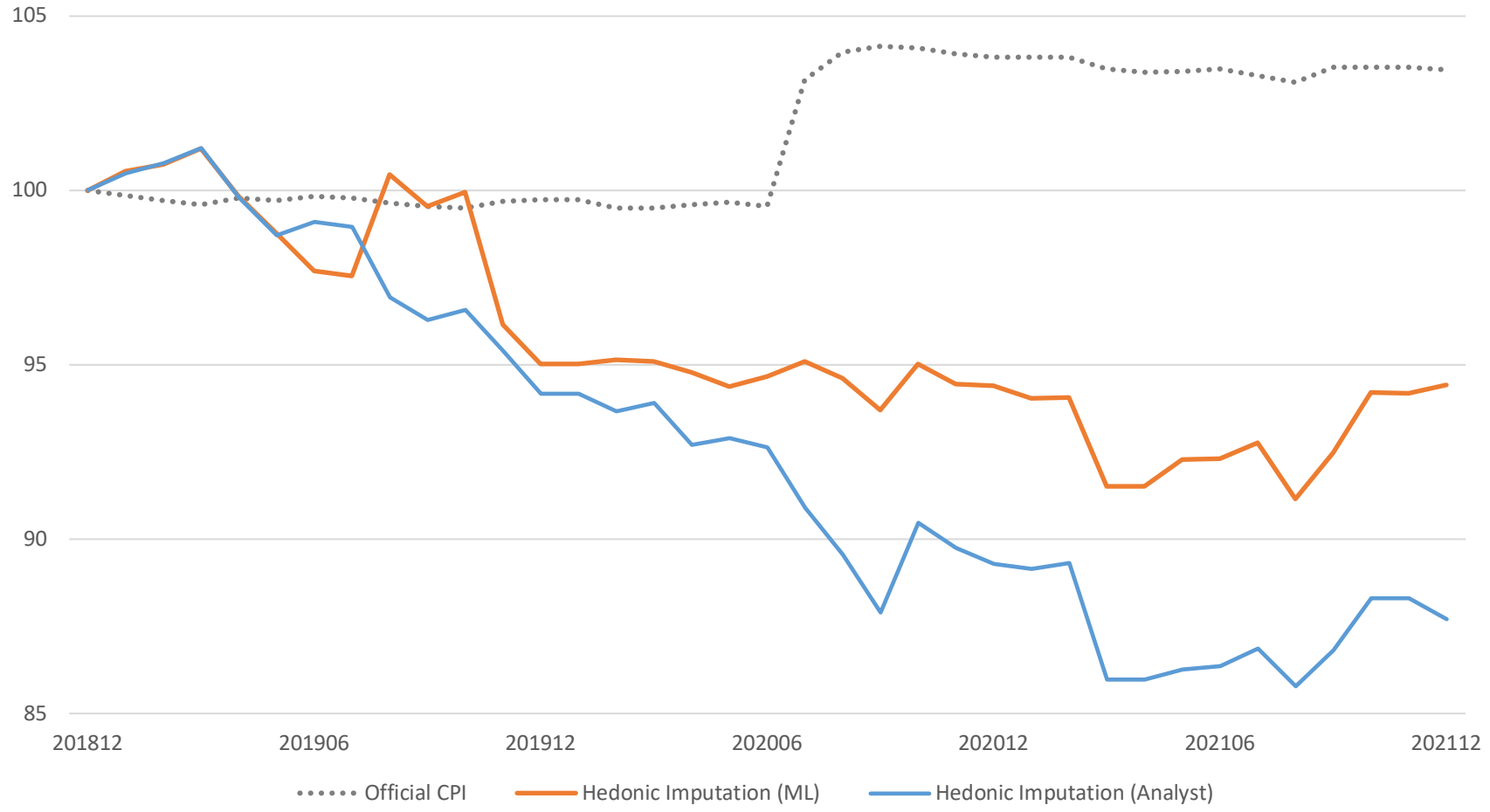
Wireless Telephone Services Data

- Prices and Characteristics
  - Data purchased from a 3rd party vendor
  - Collected via web-scraping and non-automated methods
  - Detailed information about offer prices and service plan features
- Weights
  - Data purchased from a 3rd party vendor
  - Calculate expenditure shares for plans to weight the regression and aggregate relatives
  - Adjust the plan shares by carrier national market shares

Results Summary

- Service Plans by Availability and Month
  - Average 171 plans per month
  - 6.5% entering, 6.5% exiting, 87% continuing
- Weighted Regression Models by Month
  - Dependent variable is the log(plan price)
  - Average of 8 independent variables per model
  - Average R-squared value of 0.88

Index Results – Wireless Telephone Services



Summary

- Research hedonic imputation indexes deviate from official CPI in important ways
- Model selection methodology may lead to high variances if model specifications change frequently
- Looking to extend methodology to residential telecom services and other electronic devices

Thank you

brown.craig@bls.gov

smucker.jeremiah@bls.gov