Quality Adjustment and Hedonic Regressions

Mick Silver

UN ECE-ILO Group of Experts Meeting on Consumer Price Indices, May 26, 2014, Geneva

Four problems:

- A temporary missing item problem
- A seasonal item problem
- A permanently missing item
  - Comparable replacement
  - Non-comparable replacement – imputation/linking or quality adjustment
  - Rapid turnover of models and characteristics, e.g. consumer electronics– new models are introduced and old ones disappear
- A new products problem

Price statisticians **do** take account of quality changes

- They use the matched models method:
  - sampled models (varieties) selected using detailed product descriptions on "initiation"
  - prices are recorded in initial month, and monitored in subsequent months
  - *like* is compared with *like*.

Importance of price collector

Price collectors need characteristic specifications, characteristics of replacements when permanently missing, and codes for:

- temporarily missing
- seasonal
- permanently missing: replacement variety's price is:
  - comparable replacement
  - non-comparable replacement
    - previous period's price (overlap)
    - explicit adjustment (specification)

Imputation techniques for **temporary** missing prices

- Calculate change in average price from ***matched*** observations of items from the same class using short-term (geometric mean) price changes.
- Estimate the price for the missing observations and mark them as "imputed."

Carry forward

- Induces undue stability into the index, especially for high inflation countries.
- Not for use unless assured prices do not change

Permanently missing varieties

- Comparable replacement
- Simply compare price of replacement with previous price
- Requires well trained price collectors and tightly-defined specifications
- There is an incentive to assume replacements are comparable.

Permanently missing: no comparable replacement

- Alternatives: Direct or Indirect Quality Adjustments
  - Indirect Quality Adjustments with Imputations
    - overlap price available – store manager;
    - if not need to impute overlap price (not carry forward)
  - Direct Quality Adjustment
    - data collector or analyst knowledge,
    - information from producers,
    - differences in production/option costs
    - hedonic regression models

Differentiated items with high rates of model turnover:

	jan	feb	mar	apr	may	jun
A	x	x	x	x	x	x
B	x	x	x			
C			X	x	x	x
D	x	x	x	x		
E					x	x
F	x	x	x	x	x	

Quantity as a quality dimension

- Say a 300 tablet box of a specified store brand of aspirins, sold in January at $12.00 is replaced by a 365 tablet box, sold for $15.00 in February – a non-comparable replacement. Assume linearity in the price-size relationship.
- The rescaling of the January price to the February size is 365/300*12=14.6. The constant-quantity size price change is 15/14.6 =1.0274, i.e. 2.74 percent. The assumption here is that every extra pill cost 12/300=$0.04.
- Phrased another way: $\hat{p}_{365}^{Jan} = p_{300}^{Jan} + \beta \Delta size$
- 12 + 0.04 (365-300) = 14.6

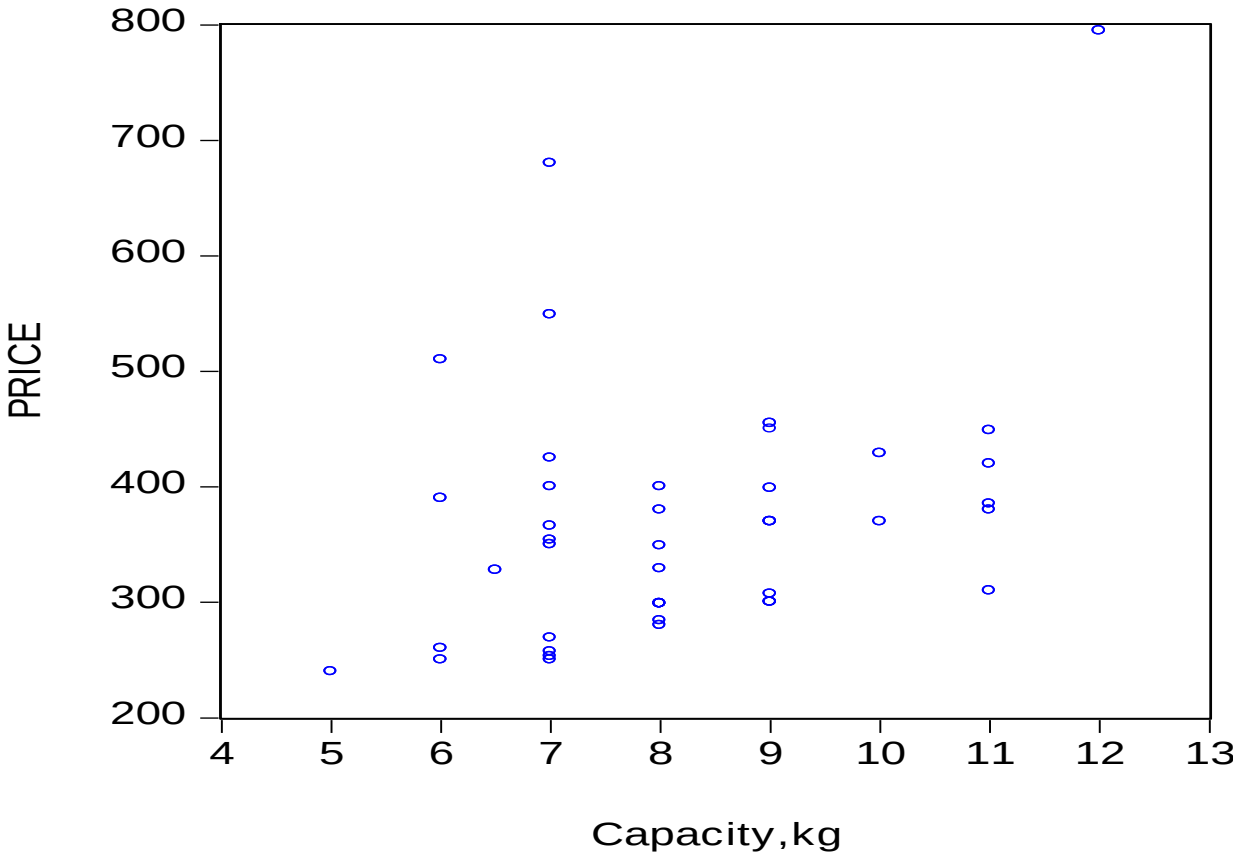
Hedonic regressions

A set of (z_k = 1,….,*K*) characteristics of the models are identified and data over *i*=1,…,*N* models are collected. A hedonic regression of the (log) price of model *i*, p_i , on its set of quality characteristics z_{ki} is given by:

$$1\mathrm{n}\ \boldsymbol{p_i} = \boldsymbol{\beta}_0 + \sum_{\boldsymbol{k}=1}^{\boldsymbol{K}} \boldsymbol{\beta_k z_{ki}} + \boldsymbol{\varepsilon_i}$$

The β_k are estimates of the marginal valuations the data ascribes to each characteristic.

A simple case of one explanatory variable: price of washing machines and their load capacity in kg.

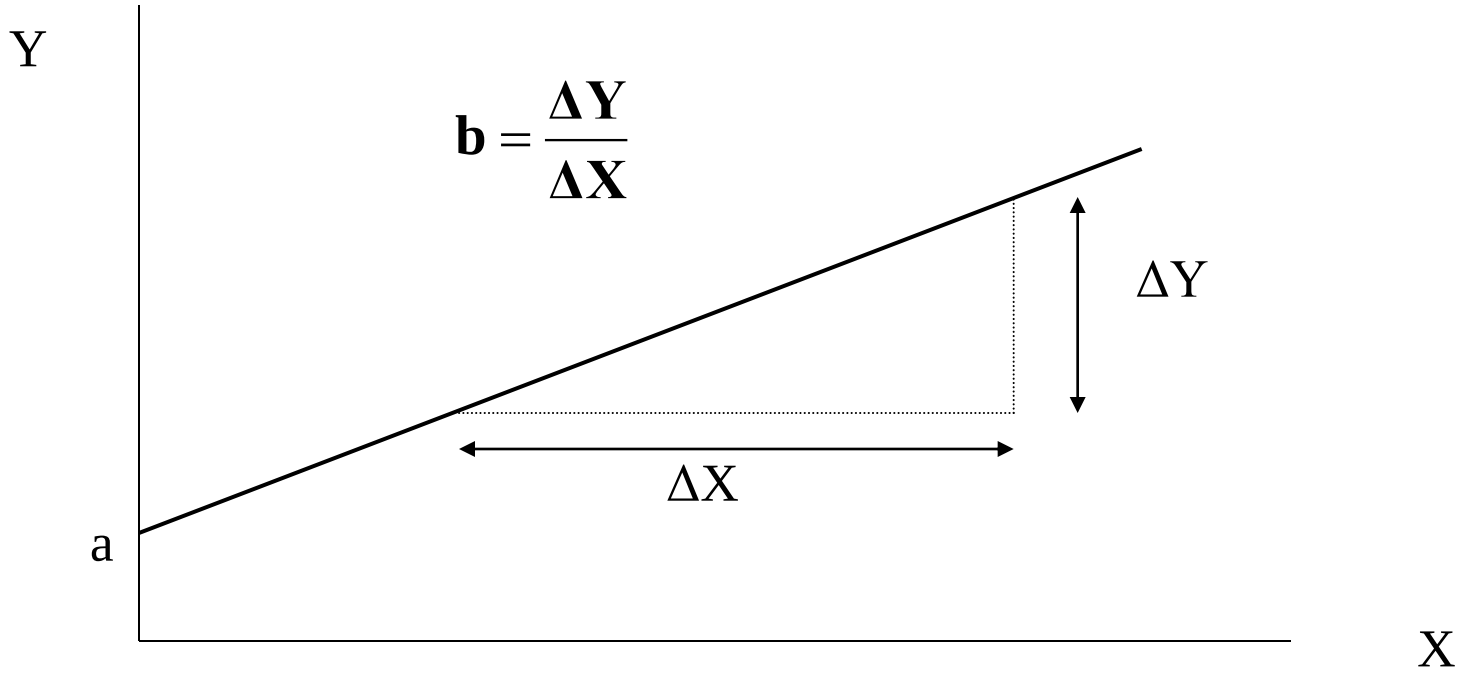


The slope coefficient

Used to estimate the equation of the line that best fits the data.

$Y = a + b\,X$

a is the intercept – the value of Y when X is zero
b is the slope – the change in Y arising from a unit change in X



A line of best fit

- Least squares criterion: minimises the sum of squared vertical differences between the points and the line.
- A one kg. increase in capacity increases price by 5.88 percent.

$$\text{1n price} = 5.3961 + 0.058808 capacity, kg$$

$$\bar{R}^2 = 0.139$$

- R-bar squared is the proportion of variation in (the log of) P that is explained by the right-hand-side variables.
- A *t*-statistic to test the null hypothesis of the coefficient being zero was 2.543: the null hypothesis of capacity having a zero effect on log price was rejected at a 5 percent level.

Hedonic regressions: an illustrative example of washing machines, continued

- Data from the UK Which website on advertised characteristics and prices
  - http://www.which.co.uk/home-and-garden/laundry-and-cleaning/reviews/washing-machines/product-finder/
- All models of 29 brands compared in source data: 3 brands taken for illustration April 2014: LG, Hotpoint and AEG comprising 42 models in all.
- Log of price (pounds sterling) on:

Explanatory characteristics

- Capacity k.
- Spin speed rpm
- Stainless steel
- Non white
- Energy label (5 categories)
- Energy costs
- Brand LG
- Brand AEG
- Warranty years
- Time for standard load, mins
- Easy iron
- Hand wash
- Extra rinse
- Variable spin
- Delayed start
- Time remaining display

Dependent Variable: LOGPRICE					
Variable	**Coefficient**	**Std. Error**	**t-Statistic**	**Prob.**	
CAPACITYKG	0.046096	0.035383	1.302791	0.2041	
ENERGYCOSTS	0.02356	0.007973	2.955101	0.0066	
ENERGYLABLE	0.022671	0.047903	0.473263	0.64	
SPINSPEEDRPM	-0.000154	0.000396	-0.390167	0.6996	
TIMELOADMINS	-0.003533	0.001622	-2.178362	0.0386	
WARRANTYYEARS	2.500903	0.299634	8.346537	0	
DUMMYAEG	0.323514	0.134174	2.411148	0.0233	
DELAYEDSTART	0.049834	0.204742	0.243397	0.8096	
EASYIRON	0.14711	0.097878	1.502997	0.1449	
EXTRARINSE	0.045771	0.096656	0.473544	0.6398	
TIMEREMAININGDISPLAY	0.164096	0.21402	0.76673	0.4501	
HANDWASHPROGRAM	2.623436	0.322079	8.145317	0	
VARIABLESPIN	0.125619	0.233856	0.537163	0.5957	
NONWHITE	0.006836	0.089611	0.076284	0.9398	
STAINLESSSTEEL	0.157542	0.155097	1.015763	0.3191	
R-squared	0.722656	Mean dependent var		5.876148	
Adjusted R-squared	0.573316	S.D. dependent var		0.265973	
S.E. of regression	0.173736	Akaike info criterion		-0.38632	
Sum squared resid	0.784792	Schwarz criterion		0.240593	
Log likelihood	22.91964	Hannan-Quinn criter.		-0.15804	

Dependent Variable: LOGPRICE				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
CAPACITYKG	0.052047	0.018858	2.760026	0.0094
ENERGYCOSTS	0.018567	0.006003	3.092679	0.004
TIMELOADMINS	-0.002898	0.001151	-2.517834	0.0168
WARRANTYYEARS	2.552257	0.147547	17.29798	0
DUMMYAEG	0.356404	0.101557	3.509394	0.0013
EASYIRON	0.15951	0.078837	2.023292	0.0512
HANDWASHPROGRAM	2.592142	0.104254	24.86368	0
TIMEREMAININGDISPLAY	0.192852	0.069758	2.764597	0.0093
R-squared	0.699144	Mean dependent va		5.876148
Adjusted R-squared	0.635326	S.D. dependent var		0.265973
S.E. of regression	0.160616	Akaike info criterion		-0.64642
Sum squared resid	0.851322	Schwarz criterion		-0.31206
Log likelihood	21.25153	Hannan-Quinn criter		-0.52466
Durbin-Watson stat	1.312802			

Results

- Selection of brands in appropriate market segment: e.g. up-market brands.
- Three outliers (AEG).
- Bundling of features: a few can represent many; multicollinearity.
- Additional kg. of capacity increases price by 5.2 percent.
- Scatter diagram shows a lot of price variation for higher-capacity models.
- AEG valued at 100*(exp(0.356404-1))= 42.82 percent more than Hotpoint. Limited sample of high-priced AEG.
- More observations needed.

How to use hedonics

- Use individual coefficients as estimates of the value of changes in the quality of individual features – not advised
- Include a time dummy:

$$\ln p_i = \beta_0 + \sum_{k=1}^{K} \beta_k z_{ki} + \gamma TimeDum_i^t + \varepsilon_i$$

- Use predicted prices: single vs double imputation:

$$\frac{\sum_i \hat{p}_{hed}^t q_{char}^t}{\sum_i \hat{p}_{hed}^0 q_{char}^t} \qquad \frac{\sum_i \hat{p}_{hed}^t q_{char}^0}{\sum_i \hat{p}_{hed}^0 q_{char}^0}$$

- Use the web to see difference in pricing of a similar model for different specifications.
- Use some else's results.

Statistical issues in hedonic regressions

- Data often readily available on websites; just use mark-ups for characteristics?
- Software makes estimation easy – even spreadsheets have regression.
- Need for diagnostic tests and expertise to validate model: heteroskedasticity, normality of residuals, multicollinearity, non-linear estimators – available in statistical/econometric software.
- Use of predicted values rather than coefficients for quality adjustment. Use of dummy time variable
- Functional form
- Inclusion of which variables