Base Pricing Analysis

37505 Data Science for Marketing Decision Making
Günter J. Hitsch
The University of Chicago Booth School of Business

2017

Overview

1. Introduction to scanner data
2. Marketing mix modeling: Base pricing analysis
3. IRI case study
4. Demand model building process
5. Panel data econometrics

Scanner data

- Timeline
  - First scan test at Kroger in Cincinnati in 1972
  - IRI's InfoScan introduced in 1987
- What do scanner data capture?
  - Sales at the UPC (universal product code) level
    - Honey Nut Cheerios 25.25 oz size
  - Brand aggregates
  - Prices and promotions
  - Aggregation levels
    - Market (Raleigh-Durham)
    - Chain/account (Kroger)
    - Store
  - Time
    - Weekly, monthly, ...

The two dimensions of scanner data

- Time series data of sales, prices, etc. in a store, chain (account), or market
- Prices, sales, etc. in a cross section of stores, chains (accounts), or markets
  - One time period in each cross sectional unit
- Data with a cross-sectional dimension and a time series dimension are called **panel data**

Week	Atlanta	Boston	Columbus	...	San Francisco
1	4.57	3.97	1.76		4.33
2	4.35	4.12	1.77		4.3
3	4.46	3.94	1.63		4.41
4	4.62	3.87	1.69		4.39
...					
52	4.91	4.04	1.98		4.45

cross section

time series

Time series vs. cross-sectional data

- If you could either use time series data or cross-sectional data for a demand analysis but not both, which would you prefer?
- Example: You attempt to estimate the own-price elasticity of demand from either
  - 104 weeks of price and sales data in one store
  - Price and sales data for 104 different stores in one week

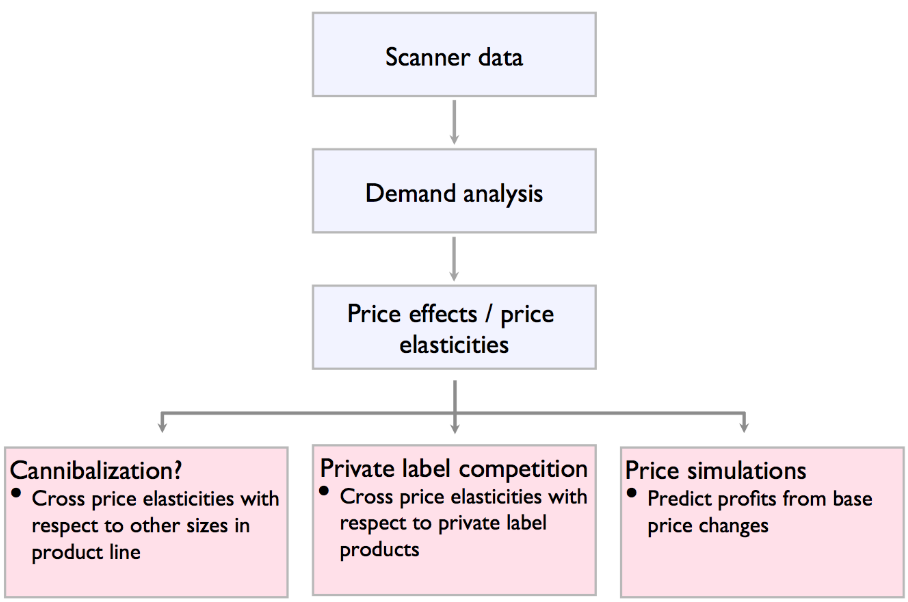
Base pricing analysis

- Key component of marketing mix modeling
- Base price = non-promoted price ("everyday" shelf price)
- Purpose of base pricing analysis
  - Understand competitive influence of prices on sales
  - Adjust / fine-tune base prices based on demand model prediction

Case study

- A CPG (consumer packaged goods) manufacturer approaches IRI
  - Company sells a national brand
  - Product line: 16 oz, 24 oz, and 32 oz bottle size (32 oz size has recently been added)
- Key problems faced by the brand manager
  - Price the different sizes separately or engage in product line pricing?
  - Is there cannibalization across product sizes (worry about new 32 oz size)?
  - Worry about private label (PL) competition — is it possible to assess the extent of the competitive threat?
  - Are the current base price points optimal, or should we change our prices?

Base pricing approach

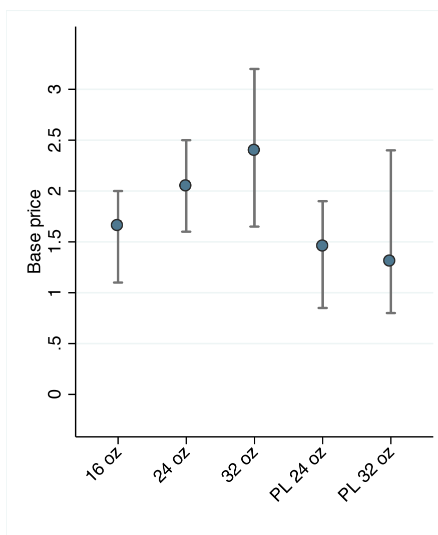


Data source in case study

- IRI's InfoScan database
  - About 1,000 stores across the U.S. used in this study (small subset of total data base)
  - 52 weeks
  - Prices and sales units of the three branded bottle sizes and the main private label products
- Focus on base prices and base unit sales
  - Collected in weeks without promotional activity
  - Promotions (details later):
    - Temporary price reductions (TPR's), display, and feature
- ACV: all commodity volume
  - Defined as the store revenue from all products sold in $ million (per annum)
  - Includes sales in all categories and departments (produce, milk, health and beauty, etc.), not just the products in the demand model
  - Proxy for store size

Data inspection: Variation in base prices

- Graph shows:
  - Average base price across all store-weeks
  - 95% range of base prices



Measurement detail: %ACV weighted price distributions

- Captures store-size weighted distribution of prices
- Example:
  - Market with 3 stores
  - Focus on price of one product

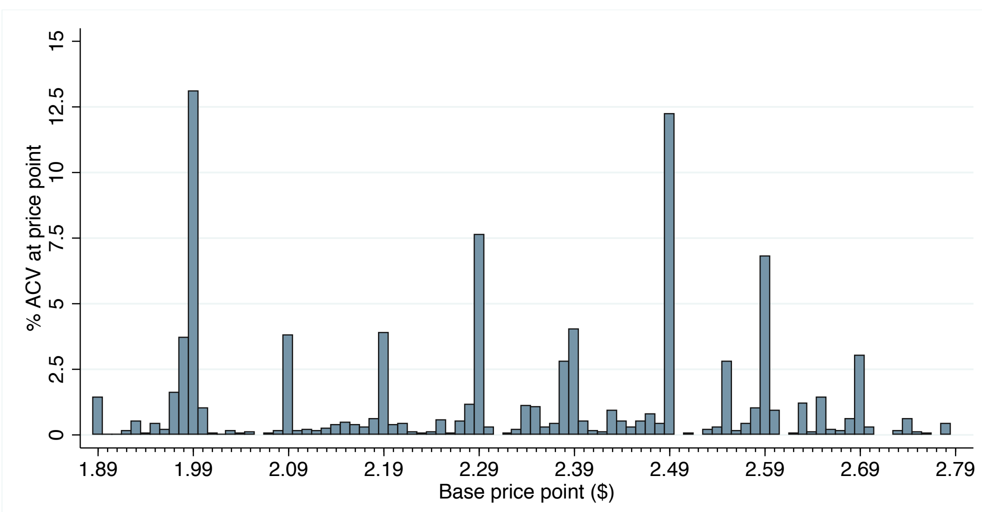
Store	ACV	Base price
A	100	$1.99
B	20	$2.69
C	80	$1.99

% of stores with price point $\$1.99 = \frac{2}{3} = 66.7\%$
%ACV at price point $\$1.99 = \frac{100+80}{100+20+80} = 90\%$

  - Takes into account that the store selling at $2.69 is small

- Roughly speaking, 90% of the stores sell the product at $1.99

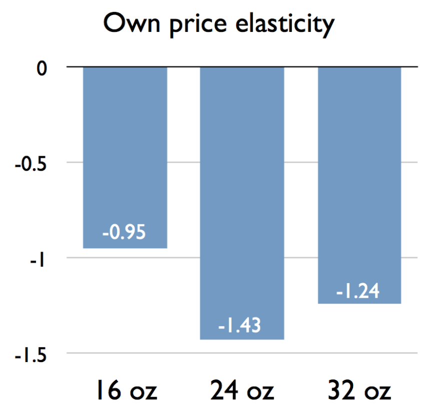
Variation in base prices across store-weeks

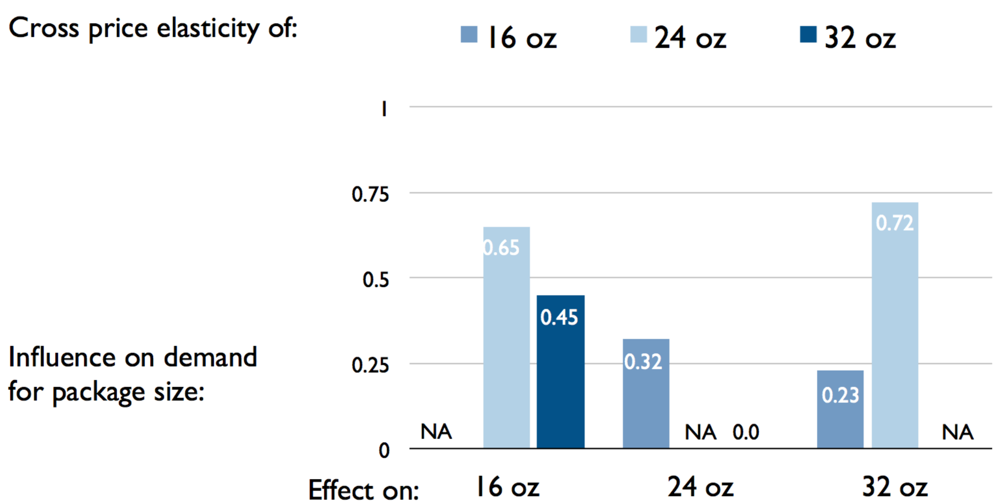


- Branded 32 oz bottle size
- % ACV weighted price distribution

Results of base pricing analysis: Elasticities

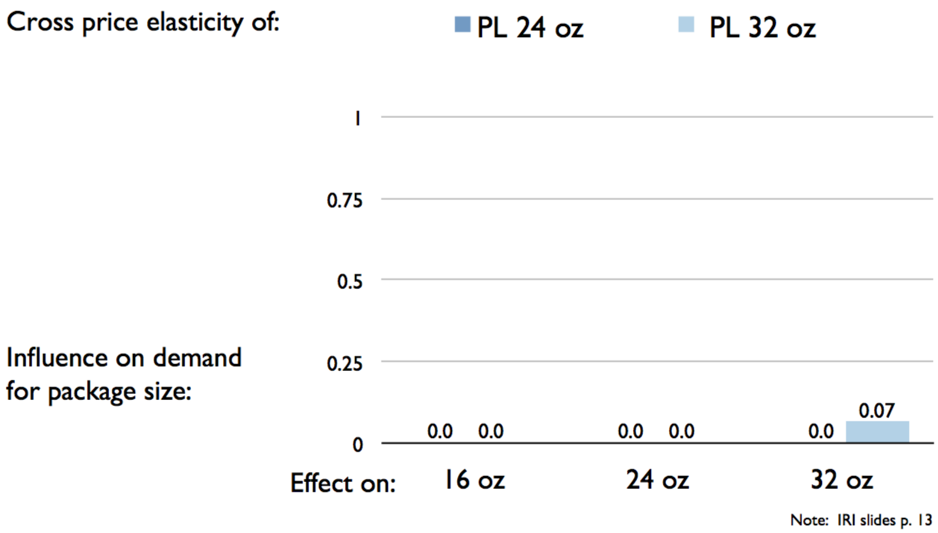
- How responsive is sales volume to own price changes?
- Note
  - Here, IRI does not report standard errors of estimates
  - Always question marketing consultants about statistical precision of results





Note: IRI slides p. 13

- Is there evidence of cannibalization?
- Consequence?



- How severe is the competitive effect of the private label products?

Base price simulations

- Goal: Predict profit for each package size k in the product line

$$\text{profit}_k = Q_k \cdot [P_k(1 - \text{retail margin}) - VC_k]$$

  - P...retail shelf price
  - VC...variable cost

- Total profit from product line (if we drop the 24 oz pack size):

$$\text{total profit} = \text{profit}(16 \text{ oz}) + \text{profit}(32 \text{ oz})$$

- What we need to conduct base price simulations:
  - Data
    - Current price levels
    - Retail margin
    - Variable cost (per unit or case)
  - Prediction of unit sales conditional on own and private label prices
    - Demand model

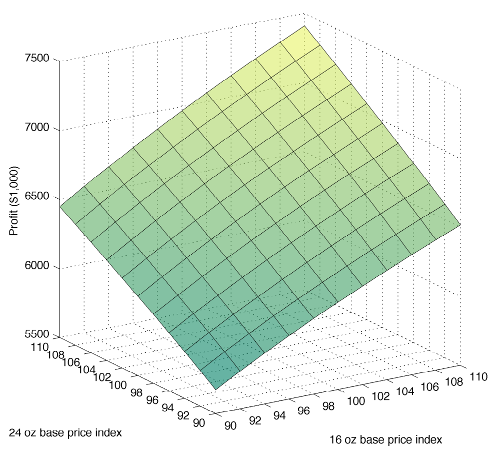
Profits for different 16 oz and 24 oz price combinations

16 oz base price index

24 oz base price index	90	92	94	96	98	100	102	104	106	108	110
90	5,672	5,771	5,866	5,957	6,046	6,131	6,213	6,292	6,369	6,443	6,515
92	5,756	5,856	5,952	6,045	6,135	6,221	6,304	6,385	6,462	6,538	6,610
94	5,838	5,939	6,037	6,131	6,222	6,310	6,394	6,476	6,555	6,631	6,705
96	5,918	6,021	6,121	6,216	6,308	6,397	6,482	6,565	6,645	6,722	6,797
98	5,998	6,102	6,203	6,299	6,393	6,482	6,569	6,653	6,734	6,812	6,888
100	6,076	6,181	6,283	6,381	6,476	6,567	6,655	6,740	6,822	6,901	6,978
102	6,152	6,260	6,363	6,462	6,558	6,650	6,739	6,825	6,908	6,988	7,066
104	6,228	6,336	6,441	6,541	6,638	6,732	6,822	6,909	6,993	7,074	7,153
106	6,302	6,412	6,518	6,619	6,718	6,812	6,903	6,992	7,077	7,159	7,239
108	6,376	6,487	6,594	6,697	6,796	6,892	6,984	7,073	7,160	7,243	7,323
110	6,448	6,561	6,668	6,773	6,873	6,970	7,064	7,154	7,241	7,325	7,407

Profits in $1,000

Note: IRI slides p. 23



- Profits highest if both prices are increased by 10%
- Why not increase prices even further?
  - Price constraints — acknowledges that statistical reliability of model decreases when proposed prices are very different from prices in the data

IRI base pricing analysis: Take-aways

- IRI's client had very limited access to sales and price data and only a limited understanding of the key pricing issues
  - Competition (private label)
  - Cannibalization
  - Optimality of base prices
- Insights to the client
  - Private label competition poses only a very limited threat to the brand
  - There is cannibalization within the product line, but the main offender is not the new 32 oz size but mainly the 24 oz size
    - Consider eliminating 24 oz size to save on costs (packaging, distribution, . . .)
  - Base price points are sub-optimally low

Demand estimation for base pricing: Details

Assignment: Estimate log-linear demand models for three brands

- Tide
- Tropicana
- ReaLemon

Data

- Nielsen RMS scanner data, 15,000+ stores, weekly data 2010-2013
- Estimate and examine cross-price elasticities
- Evaluate current pricing tactics

Model building process

- Output
  - $\log(Q)$
  - Problem with store data in particular: $Q = 0$ in some weeks
  - Alternative: $\log(1 + Q)$
- Key input
  - $\log(P)$

Model building process: Add all necessary controls

- Control for store size
  - Old school: Normalize sales by variable that is proportional to store size
  - Example: (log) *sales velocity* as output

$$\text{sales velocity} = \frac{Q}{ACV}$$

  - Modern approach: Include store dummy variables (fixed effects)
- Time effects
  - Capture shifts in demand that occur over time
    - Changes in consumer preference, availability of substitutes, market structure, ...
  - Controls
    - Time trend (linear, polynomial, ...)
    - Time dummy variables (fixed effects)

- Control for competition — what could go wrong if we did not include such controls?
  - (log) prices of key competing products
  - Which and how many competing products to include?
    - Requires intuition and experimentation
    - Alternative: Variable selection using machine learning methods—only at experimental stage as of now

- Remember our goal: Estimate base price elasticities
  - Focus is on the effect of the "everyday" price, not on the effect of price promotions
  - Price promotions are typically associated with sales spikes
  - Old school: Eliminate periods with promotions from the sample
  - Modern approach: Capture promotions using promotion indicators (dummies) or interaction of promotion dummy with (log) price

- Store fixed effects
  - To control for store size
  - Above and beyond controlling for size differences across stores: See discussion in panel data econometrics section

Model building in practice

Building a model step-by-step is typically highly illuminating

- Illustrates the importance of all the controls
- Illustrates how estimates change after the addition of controls

Pricing simulations

- Goal: Predict the impact on total (product line) profits if we change the current base prices of one or more of the products in the product line
- For each product j,
$$\text{profit}_j = Q_j \cdot [P_j(1 - \text{retail margin}_j) - C_j]$$
    - C_j is the unit (variable) cost of production
    - Fixed costs only influence the overall profit level, not profit differences associated with price differences
- The prediction of Q_j is based on an estimated demand model
    - To be precise, Q_j is a function of all inputs (prices, promotions, store effects, ...)

Mechanics of quantity prediction—formula approach

For each product j evaluate percentage price changes of the form

$$\Delta \log(P_k) = \log((1+\gamma_k) \cdot P_k) - \log(P_k) = \log(1+\gamma_k)$$

Quantity prediction based on log-linear demand model:

$$\log(Q_j) = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} \log(P_k) + \dots$$

- Take difference, after vs. before price change:

$$\begin{aligned}\Delta \log(Q_j) &= \sum_{k=1}^{K} \beta_{jk} \Delta \log(P_{jk}) \\ &= \sum_{k=1}^{K} \beta_{jk} \log(1+\gamma_{jk})\end{aligned}$$

- Take $\exp$ on both sides of equation:

$$\frac{Q'_j}{Q_j} = (1+\gamma_1)^{\beta_{j1}} \cdots (1+\gamma_K)^{\beta_{jK}}$$

- Instead of using a formula can simply predict quantity using R (or any other statistical software)
- If there are multiple products in the product line, predict sum over all product-level profits

Panel data econometrics

Many data sets have the form $(Y_{it}, X_{1it}, \ldots, X_{Kit})$

- Two indices, i and t
- K inputs

Cross-sectional dimension: $i = 1, \ldots, N$

- Each i is a *unit*
- Examples of units: Firms, stores, markets, households, or consumers

Time-series dimension: $t = 1, \ldots, T$

- *Balanced panel*: Same time-series length T for all units
- *Unbalanced panel*: Time-series length differs across units. Each unit has T_i observations

Use of panel data

Many data sets are panel data

- Nielsen Homescan
- Nielsen RMS scanner data
- Any CRM data base, ...

Advantages

1. Data size: Large number of observations
   - *Pooling* of data: We use the sheer size of the data and treat each observation indexed by i, t as we would in a standard regression model
2. The structure of panel data allows us to account for heterogeneity across the units, and potentially avoid bad estimates (e.g. omitted variables bias)

Linear panel data model

There are different (and more general) models, but one of the most prominent specifications is

$$Y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kit} + \phi(Z_{1i}, \ldots, Z_{Li}) + \epsilon_{it}$$

We assume conditional mean-independence of the error term:

$$\mathbb{E}(\epsilon_{it} | X_{1it}, \ldots, X_{Kit}) = 0$$

Two types of variables:

- The X_{kit} variables are in our data set
- The Z_{li} variables are not in our data set—because they cannot be observed directly or are hard to measure

- $\phi(Z_{1i}, \ldots, Z_{Li})$ represents *heterogeneity across the units*—different outcomes, Y_{it}, depending on the values of $Z_{1i}, \ldots, Z_{Li}$. Effect can but need not be linear in Z_{li}
- Note that the Z_{li} variables do not vary over time, unlike X_{kit}
- Examples:
  - Consumer preferences
  - Competition and demographics of customers that a store faces

Goal of estimation

$$Y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kit} + \phi(Z_{1i}, \ldots, Z_{Li}) + \epsilon_{it}$$

We work off the assumption that this statistical model is a causal model

Key objective

- Estimate the true, causal parameters β_k
- Corresponds to effect of *manipulation* of X_{kit} while holding $\phi(Z_{1i}, \ldots, Z_{Li})$ and ϵ_{it} fixed

How to estimate this model?

- Define
$$\tilde{\epsilon}_{it} = \phi(Z_{1i}, \ldots, Z_{Li}) + \epsilon_{it}$$
- Allows us to write the model in simplified form:
$$Y_{it} = \beta_0 + \sum_{k=1}^{K} \beta_k X_{kit} + \tilde{\epsilon}_{it}$$
- *Pooled regression*: We *pool* over all i, t observations and estimate a regular linear regression
- When will this regression yield consistent estimates of the true parameters, β_k?

Consistent estimation

Classic linear regression model assumption

$$\mathbb{E}(\tilde{\epsilon}_{it}|X_{1it},\dots,X_{Kit}) = 0$$

- Yields consistent and—in the linear model case—unbiased estimates

In our specific model with unit-level heterogeneity, we already assumed that $\mathbb{E}(\epsilon_{it}|X_{1it},\dots,X_{Kit}) = 0$
Hence, the assumption requires that also

$$\mathbb{E}(\phi(Z_{1i},\dots,Z_{Li})|X_{1it},\dots,X_{Kit}) = 0$$

- $\phi(Z_{1i},\dots,Z_{Li})$ is uncorrelated with the X_{kit} variables

Think of examples: Are prices necessarily uncorrelated with store or market level heterogeneity?

Example: Goal is to estimate a demand model using data on sales and prices from stores in different geographies or markets

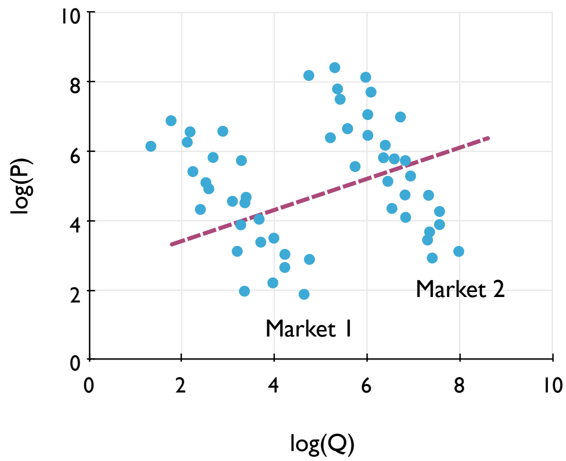
- Using data from different markets can be useful — price variation across markets may be higher than within markets

But markets often differ across other dimensions:

- Consumer tastes
- Demographics
- Competition
- Distribution

Problem: Prices might systematically differ across markets according to differences in consumer tastes, demographics, etc.

Illustration: Prices systematically higher in high-demand markets



This is an example of **omitted variable bias**

- Variable of interest (price) is correlated with omitted variable(s) (store or market level heterogeneity)
- Estimated price coefficient is biased and inconsistent—no matter how large the sample size, the estimate will never converge to the true value

How to fix this problem?

Review: Omitted variable bias

Consider the following regression:

$$Y = \alpha + \beta X + \gamma Z + \epsilon$$

Suppose you estimate the model

$$Y = \alpha + \beta X + \tilde{\epsilon}$$

because you do not have data on Z (or simply forget to include Z) in the regression.

Suppose the following conditions are met:

1. Z affects Y, i.e. $\gamma \neq 0$
2. Z and X are correlated

Then the estimate of β will be biased and inconsistent, i.e. the estimate will never converge to the true value irrespective of sample size

Note the assumption that the unmeasured variables, $Z_{1i}, \ldots, Z_{Li}$, do not vary across time, but only across units
Hence the effect of these variables is given by one value for each unit,

$$\alpha_i = \phi(Z_{1i}, \ldots, Z_{Li})$$

- The α_i's are called *fixed effects*

Define dummy variables for each unit:

$$D_n = \begin{cases} 0 & \text{if } i \neq n \text{ in observation } i, t \\ 1 & \text{if } i = n \text{ in observation } i, t \end{cases}$$

Then re-write the model:

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{kit} + \sum_{n=1}^{N} \alpha_n D_n + \epsilon_{it}$$

Note the intercept had to be removed
Or, even simpler:

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{kit} + \alpha_i + \epsilon_{it}$$

The within estimator/fixed effects estimator

For each unit i take the difference between the variables and the mean of the variables over all time periods,

$$Y_{it} - \bar{Y}_i = Y_{it} - \frac{1}{T} \sum_{t-1}^{T} Y_{it}$$

Can then write the model as

$$\begin{aligned} Y_{it} - \bar{Y}_i &= \sum_{k=1}^{K} \beta_k \left(X_{kit} - \bar{X}_{ki} \right) + (\alpha_i - \bar{\alpha}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \\ &= \sum_{k=1}^{K} \beta_k \left(X_{kit} - \bar{X}_{ki} \right) + \nu_{it} \end{aligned}$$

Here the error term is

$$\nu_{it} = \epsilon_{it} - \bar{\epsilon}_i$$

The original assumption, $\mathbb{E}(\epsilon_{it}|X_{1it}, \ldots, X_{Kit}) = 0$, implies that

$$\mathbb{E}(\nu_{it}|X_{1it} - \bar{X}_{1i}, \ldots, X_{Kit} - \bar{X}_{Ki}) = 0$$

Hence we are assured consistent (and unbiased, because this is a linear regression model) estimates of β_k

Within estimator:

- Idea is that estimation of β_k parameters is based on within-unit variation of the X_{kit} variables over time
- Of course this requires that X_{kit} varies within unit i — otherwise $X_{kit} - \bar{X}_{ki} \equiv 0$ and there is no information on the effect from the differenced data

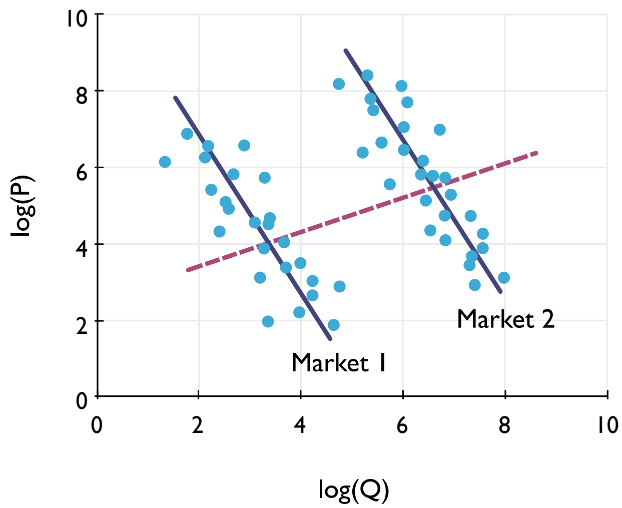
Dummy variable implementation

Estimate the regression

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{kit} + \sum_{n=1}^{N} \alpha_n D_n + \epsilon_{it}$$

- Will yield estimates of β_k that are identical to the estimates from the within estimator
- Consistent estimation of the fixed effects, α_i, is not possible unless T becomes large
- Computationally, estimating the regression in this form will be difficult or impossible for many units N

Illustration:



Advantage of panel data

- Variation in the inputs, X_k, that is correlated with unobserved heterogeneity across units is a common problem
  - Omitted variables problem: Estimated coefficients are not causal effects
  - Occurs in many marketing applications and more generally in most social science settings
- Panel data help to solve this problem using techniques such as the within (fixed effects) estimator
  - Use cross-sectional dimension to isolate the effect of heterogeneity across units on the output
  - Use time-series dimension to estimate the causal effects

Common extension: Time fixed effects

Model with time fixed effects:

$$Y_{it} = \sum_{k=1}^{K} \beta_k X_{kit} + \alpha_i + \gamma_t + \epsilon_{it}$$

The time fixed effects γ_t account for trends in the data or unmeasured variables that systematically affect the output at time t for all units

- Inclusion may increase precision of estimates or solve an omitted variables problem

Time fixed effects can also be formulated for different time-period definitions

- Example: Year/month fixed effects instead of week fixed effects

Computational implementation

Several excellent packages are available in R and other languages to estimate panel data regressions with fixed effects

Recommended: `lfe`

- Uses differencing techniques (or repeated differencing in case of multiple fixed effects) to allow estimation with tens of thousands or more fixed effects

Summary

- Base pricing analysis: Important component of marketing mix modeling
  - Understand how the prices of the products in the category influence demand for the products that we sell
  - Do the prices of competing products influence our own sales — competition?
  - Do the prices of other products in our product line influence our own sales — cannibalization?
- Data-driven approach to base pricing
  - Use demand model estimates to evaluate and predict effect of price changes
  - Profit simulations
  - Key steps in model building
- Panel data econometrics
  - Within or fixed effects estimator
  - Solves omitted variables problem