High-dimensional high-frequency retail price dynamics with missing data

Alan Crawford Lars Nesheim

Universidad Carlos III de Madrid (UC3M) CeMMAP, UCL and IFS

November 2022

Acknowledgements

We gratefully acknowledge financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) grant number ES/I034021/1 and ESRC Large Research Grant ES/P008909/1. The data was supplied by Kantar UK Limited. The use of Kantar UK Ltd. data in this work does not imply the endorsement of Kantar UK Ltd. in relation to the interpretation or analysis of the data. All errors and omissions remain the responsibility of the authors.

Section 1

Introduction

Introduction: data

- Scanner data on retail prices
    - Records information on about 15,000 households over time.
    - Observe every shopping trip to any supermarket/grocery store for 1-3 years.
    - Stores include 4 major supermarket chains, other large competitors, and a large set of small stores.
    - Observe quantity, price and characteristics of every item purchased.
    - More than 300 categories of products (e.g. shampoo, laundry detergent, fresh fruit, beer, etc.)
    - Observe prices and quantities for tens of thousands of distinct products.
    - Each shopping trip each consumer purchases nonzero quantities of 1 - 50 items.
- High frequency sales and promotions.
- Price levels and dynamics vary across categories, brands, pack sizes, stores, time.

Introduction: statistical problem

- High dimensional data on prices and quantities.
- Missing data issues
  - Promotional status not observed.
  - No price observed if no purchases recorded in the sample.
  - Products with moderate to small market shares have large missing data problems.
- Missing data complicates inference on:
  - Price levels of missing items.
  - Price dynamics and promotional dynamics of missing items.
  - Estimates of aggregate inflation.
  - Estimates of consumer response to prices.

Introduction: proposed solution

- Develop a nonlinear factor model **(Chen et al., 2021)** also known as a “generalized low rank” model **(Udell et al., 2016)** to capture the dynamics.
  - Assume dynamics of high dimensional prices and consumer demand driven by a common set of low-dimensional factors.
  - Factors evolve according to a simple VAR(K) model.
- Model price process as a switching model with switching between a “regular price” process and a “sales price” process.
- Promotional status is “missing” or unobserved.
- Price is also missing when observed demand is zero.
- Correct for missing data problem with consumer demand based model of sample selection.

Section 2

Model

Model

- Observe data on J products for T time periods.
  - For shampoo, $J = 820$ and $T = 365$ or $T = 52$.
  - Product is defined by brand, pack size, and store.
- Model time series evolution of prices using a low dimensional factor structure.
- Account for switching between "sales-price" and "regular-price" with a hidden Markov state.
- Account for missing price data with a model of "selection" based on consumer demand.

Model

- Let $q_{jt} \geq 0$ be the aggregate quantity of product j purchased by consumers in the sample in period t.
- Let p^*_{jt} be the (potentially unobserved) price of product j in period t.
    - If $q_{jt} > 0$, observe $p_{jt} = p^*_{jt}$.
    - if $q_{jt} = 0$, p^*_{jt} is not observed.
- Let s^*_{jt} be the unobserved sales status of product j in period t.
    - s^*_{jt} is always missing.
    - if $s^*_{jt} = 1$, then p^*_{jt} is drawn from the "sales price" distribution.
    - if $s^*_{jt} = 0$, then p^*_{jt} is drawn from the "regular price" distribution.
- Dynamics of sales status s^*_{jt} and of demand q^*_{jt} are driven by an R dimensional set of factors f_t with R known and $R < J$.

Model

1. Sales status

$$s^*_{jt} = 1\left[\alpha_{sj} + \lambda_{sj}^T f_t + \varepsilon_{sjt} \leq 0\right]$$

2. Price

$$\begin{aligned} p^*_{jt} &= \left(1 - s^*_{jt}\right) p_{Hjt} + s^*_{jt} p_{Ljt} \\ p_{Hjt} &= p_{0j} + \varepsilon_{Hjt} \\ p_{Ljt} &= p_{0j} - \delta_j + \varepsilon_{Ljt} \end{aligned}$$

3. Consumer demand

$$q^*_{jt} = \alpha_{qj} + \sum_{i=1}^{J} \beta_{ji} p^*_{it} + \lambda_{qj}^T f_t + \varepsilon_{qjt}$$

Identification

- The factors f_t capture time-varying demand and cost shocks.
- Our assumptions impose functional form restrictions:
  - p_t^* is not a linear function of f_t.
- Functional form restrictions based on observation:
  - Firms do not adjust prices continuously while consumers can choose to purchase at any time.
- Ongoing work (with Crawford and Myśliwski):
  - Incorporate information on cost shocks.
  - Develop model of price setting competition with adjustment costs (Kydland, 1975; Judd, 1996).

Model

4. Measurement equations

$$\begin{aligned} q_{jt} &= 1\left[q_{jt}^* > 0\right] q_t^* \\ p_{jt} &= 1\left[q_{jt}^* > 0\right] p_{jt}^* \end{aligned}$$

5. Factor dynamics

$$f_t = A_0 + \sum_{k=1}^{K} A_k f_{t-k} + \eta_t$$

Reduce dimension of β: exploit observable product characteristics

- Number of parameters increases with J and T.
- In particular, when J is large, β consists of J^2 parameters.
- Assume consumer substitution patterns depend on a lower dimensional vector of product characteristics (e.g. store, brand, "childrens' shampoo", or for alcohol, lager, ale, alcohol content, etc.)

$$\beta_{ji} = \sum_k \widetilde{\beta}_k z_{kji}$$

Model

- Joint distribution of the unobservables. For all (j, t)

$$\varepsilon_{jt} = \begin{bmatrix} \varepsilon_{sjt} \\ \varepsilon_{Hjt} \\ \varepsilon_{Ljt} \\ \varepsilon_{qjt} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sigma^2_{Hj} & 0 & \sigma_{Hyj} \\ 0 & 0 & \sigma^2_{Lj} & \sigma_{Lyj} \\ 0 & \sigma_{Hyj} & \sigma_{Lyj} & \sigma^2_{yj} \end{bmatrix} \right) \tag{1}$$

- Heteroscedasticity across products.
- Assume ε_{jt} independent over products and time periods.
- Write the above covariance matrix as $\Sigma_j = C_j C_j^T$.
- Also, assume $\eta_t \sim N\left(0, \Sigma_\eta\right)$.

Parameterisation of covariance matrix

- Assume

$$C_j = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_{Hj} & 0 & 0 \\ 0 & 0 & c_{Lj} & 0 \\ 0 & c_{Hyj} & c_{Lyj} & c_{yj} \end{bmatrix}.$$

  with

$$\begin{aligned} c_{Hj} &= c_{H0} + \Delta c_{Hj} \\ c_{Lj} &= c_{L0} + \Delta c_{Lj} \\ c_{yj} &= c_{y0} + \Delta c_{yj} \\ c_{Hyj} &= c_{Hy0} + \Delta c_{Hyj} \\ c_{Lyj} &= c_{Ly0} + \Delta c_{Lyj}. \end{aligned}$$

- Add L1 penalty to “regularise” the Δ terms.

Asymptotic distribution

- Theorem's 1 and 2 in Chen et al. (2014) apply in the homoscedastic case.
- Model parameters $(\alpha_s, \alpha_y, \beta_{ji})$ and average partial effects (e.g. demand elasticities) can be consistently estimated.
- Due to incidental parameters problem, asymptotic bias exists.
- Remove bias either with analytic formula or with split-sample bias estimator.

Estimation: maximum likelihood

- Parameters are $\theta = \left\{ f, \alpha_s, \lambda_s, \alpha_y, p_0, \delta, \beta, C_j, A, C_\eta \right\}$.
- Likelihood estimation computationally costly due to cost of integration across 2^J sales states. Likelihood function is

$$
\begin{aligned}
L\left(p, q, f\right) &= T^{-1} \sum_{t=1}^{T} \log L_t\left(p_t, q_t \,|f_t\right) \\
&+ \left(T - K\right)^{-1} \sum_{t=1+K}^{T} \log \phi \left(f_t - A_0 - \sum_{k=1}^{K} A_k f_{t-k}, C_\eta \right).
\end{aligned}
$$

where

$$
\begin{aligned}
L_t\left(p_t, q_t \,|f_t\right) &= \sum_{s^* \in \mathcal{S}_J} \text{Ł}_{st}\left(p_t, q_t \,|f_t, s^*\right) \Pr\left(s_t^* | \, f_t\right) \\
\Pr\left(s_t^* | \, f_t\right) &= \prod_j \Phi\left(-a_{sjt}\right)^{s_j^*} \left[1 - \Phi\left(-a_{sjt}\right)\right]^{1-s_j^*} \\
a_{sjt} &= \alpha_{sj} + \lambda_{sj}^T f_t
\end{aligned}
$$

Estimation: EM algorithm (1)

- Use EM algorithm combined with numerical approximation:
  1. Guess θ_{w-1}.
  2. For M draws of $\{\varepsilon_{sjtm}\}$, compute $\pi_w(t,m) = \Pr\left(s_{tm} \middle| P, Q, F, \theta_{w-1}\right)$.
  3. Choose θ_w to maximize

$$L^{EM}(p,q,f) = T^{-1}\sum_{t=1}^{T} L_t^{EM}(p_t, q_t \,|f_t) \\ + (T-K)^{-1}\sum_{t=K+1}^{T} \log\phi\left(f_t - A_0 - \sum_{k=1}^{K} A_{t-k} f_{t-k}, \Sigma_\eta\right)$$

where

$$L_t^{EM}(p_t, q_t \,|f_t) = \sum_{m=1}^{M} \pi_w(t,m) \log L_{st}(p_t, q_t \,|f_t, s_{tm}). \\ + \sum_{m=1}^{M} \pi_w(t,m)\left(\sum_j \log\Pr\left(s_{jtm} \middle| f_t\right)\right)$$

Estimation: EM algorithm (2)

- Weights in L_t^{EM} are given by

$$\pi_w(t,m) = \Pr\left(s_{tm}\,|p_t, q_t, f_t, \theta_{w-1}\right)$$

$$\Pr\left(s_{tm}\,|p_t, q_t, f_t, \theta_{w-1}\right) = \frac{\Pr\left(s_{tm}, p_t, q_t\,|f_t, \theta_{w-1}\right)}{\sum_m \Pr\left(s_{tm}, p_t, q_t\,|f_t, \theta_{w-1}\right)}$$

$$\Pr\left(s_{tm}, p_t, q_t\,|f_t, \theta_{w-1}\right) = L_{st}\left(p_t, q_t\,|f_t, s_{tm}, \theta_{w-1}\right)\prod_j \Pr\left(s_{jtm}|\, f_t, \theta_{w-1}\right)$$

Penalisation

1. Assume most off-diagonal elements of matrix β equal zero. Add $\text{L1}\,(\beta)$ penalty to impose sparsity.
2. Need to normalise factors and loadings.
    1. Impose constraints $\frac{1}{T} ff^T = I$ and $\lambda^T \lambda =$ diagonal.
3. Impose sparsity on covariance matrix by adding penalties: $\text{L1}(\Delta C_{Hj})$, $\text{L1}(\Delta C_{Lj})$, $\text{L1}(\Delta C_{Yj})$, $\text{L1}(\Delta C_{HYj})$, and $\text{L1}(\Delta C_{LYj})$.

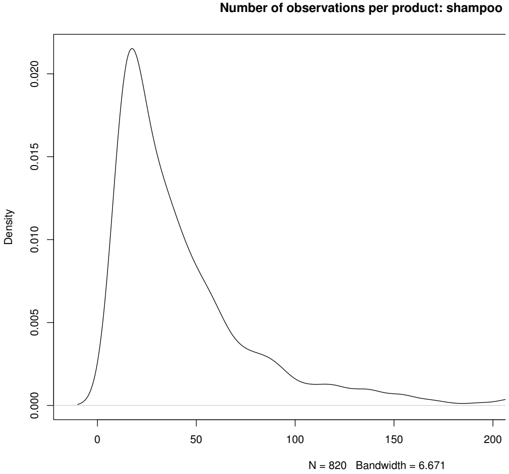
Section 3

Data

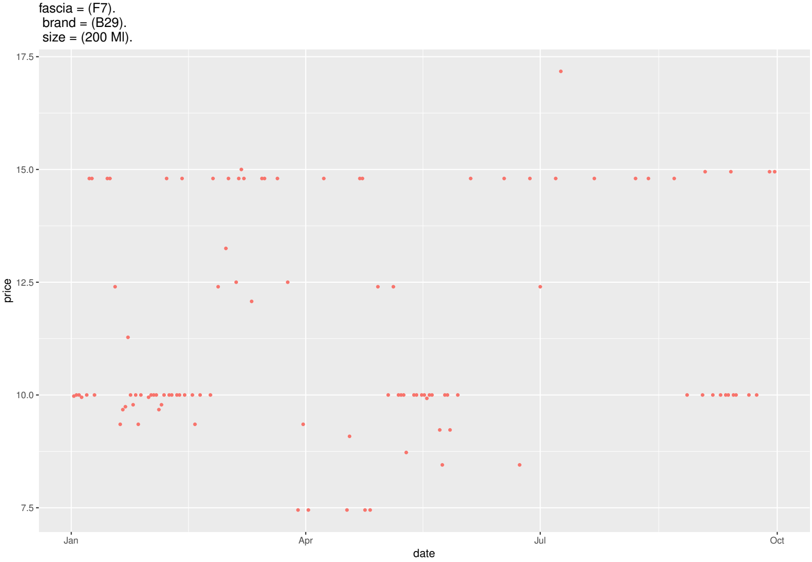
Shampoo data

- 13 fascia: Aldi, Asda, Co-Op, Costco, Iceland, Lidl, Morrisons, Ocado, other, Sainsbury's, Tesco, Tesco Metro, and Waitrose.
- 220 brands. For example, Alberto Balsam, Aldi Shampopo, Aussie Aussome Vol Shamp, Head & Shoulders, Tesco Standard Shampoo.
- Multiple bottle sizes. E.g. 150ml, 400 ml, 1000 ml.
- 820 "products" defined by fascia, brand and pack size.

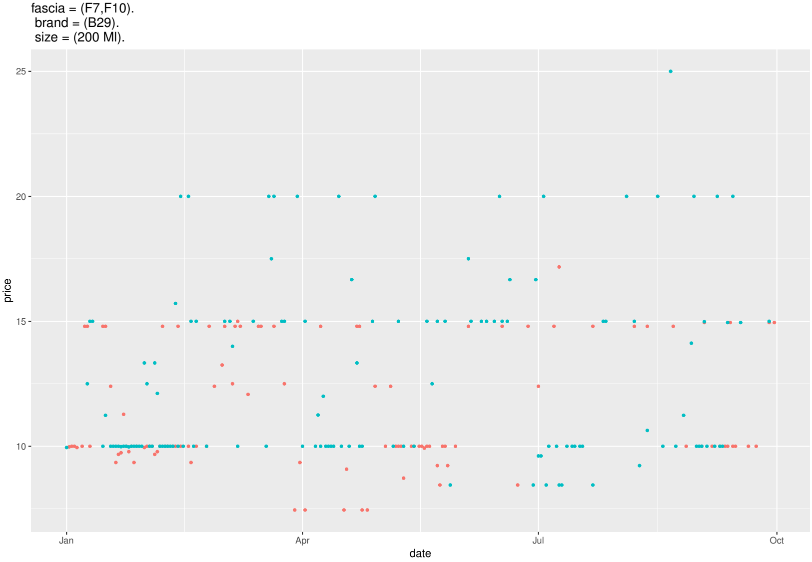
Number of observations per product: shampoo 2016



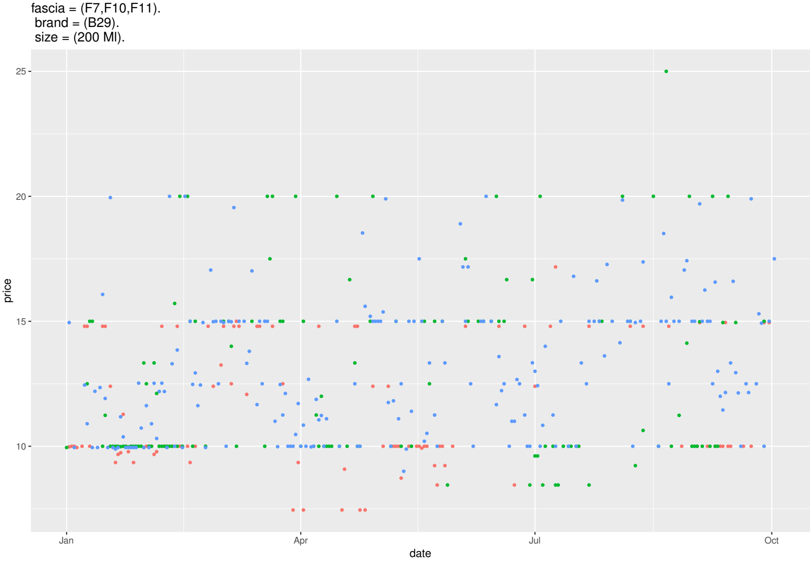
Shampoo: data



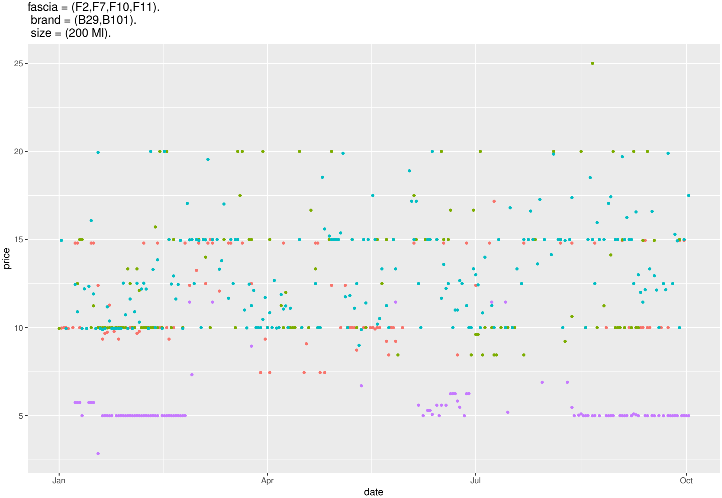
Shampoo: data



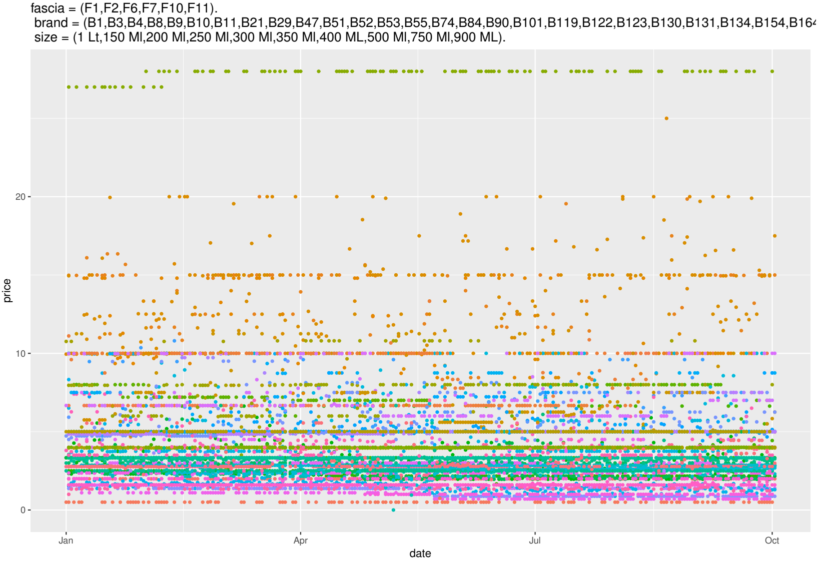
Shampoo: data



Shampoo: data



Shampoo: data



Shampoo data summary

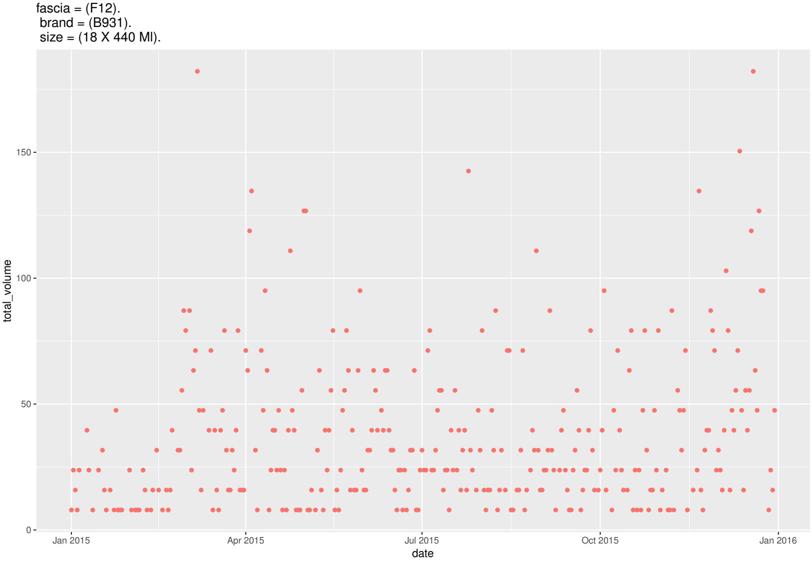
- Large volume of missing prices.
- Current model does not allow for "sticky prices" nor for prices that never move.
- Work in progress: "sticky price" model.
  - Prices only move when sales state changes.
  - Otherwise, price equals price from previous period.

$$\begin{aligned} p^*_{jt} &= \left|\Delta s^*_{jt}\right| \left(s^*_{jt} p_{Hjt} + \left(1 - s^*_{jt}\right) p_{Ljt}\right) \\ &\quad + \left(1 - \left|\Delta s^*_{jt}\right|\right) p^*_{j,t-1} \end{aligned}$$

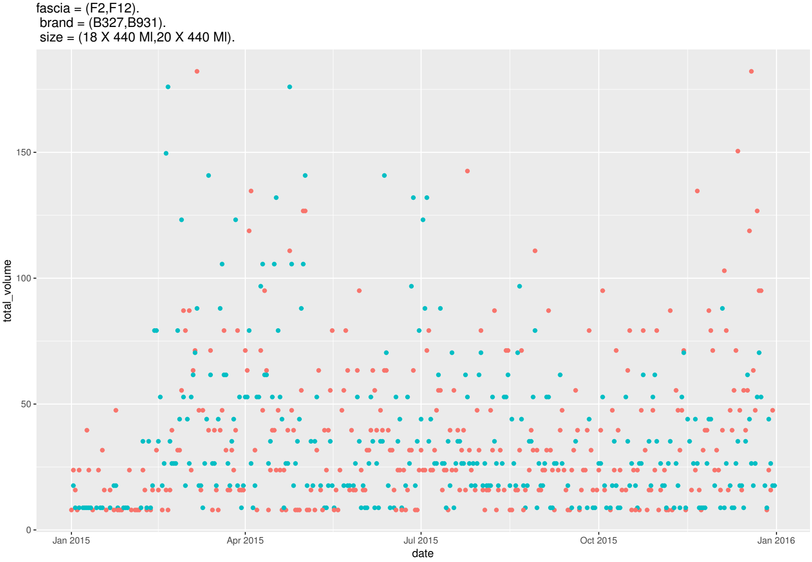
Beer data

- More than 1,000 brands including Stella Artois, Adnams Broadside Ale, Tesco lager, etc.
- Multiple sizes and pack types: can vs. bottle, single item vs multi-pack, 250ml, 500ml, etc.
- Submarkets for lager, ale, cider, bitter, etc.
- Missing data and price plots look similar to shampoo.

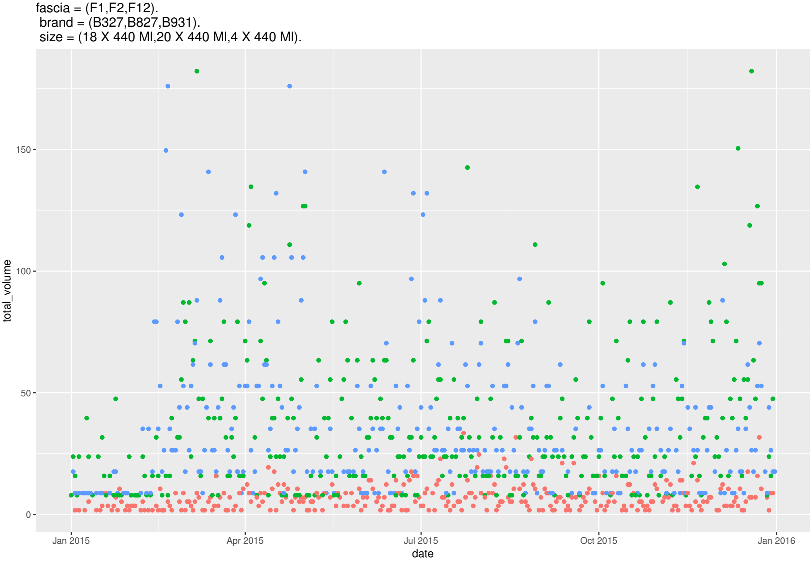
Data: beer quantities



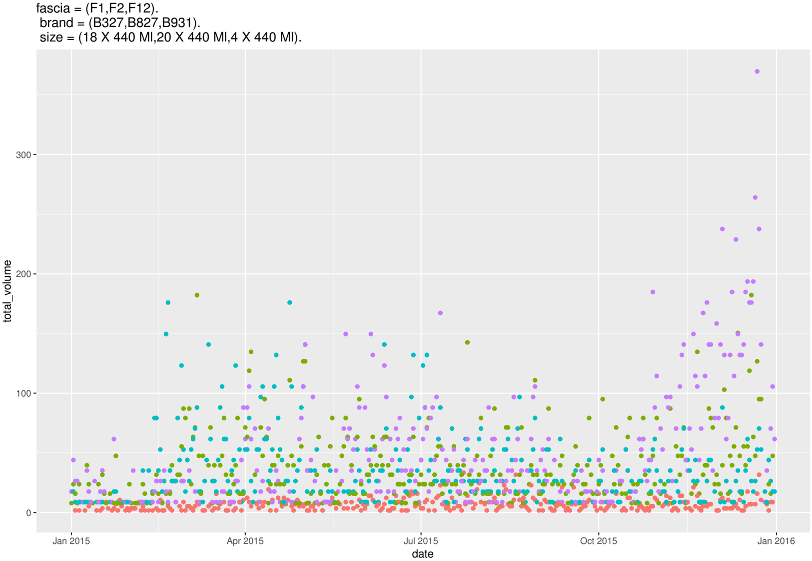
Data: beer quantities



Data: beer quantities

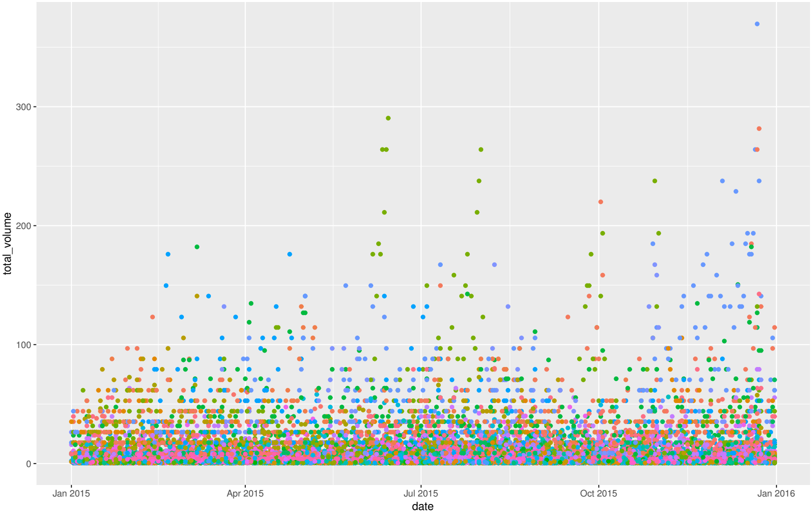


Data: beer quantities



Data: beer quantities

fascia = (F1,F2,F4,F6,F9,F10,F12).
brand = (B26,B28,B62,B76,B84,B93,B109,B125,B129,B160,B173,B211,B219,B220,B318,B327,B357,B358,B415,B416,B418,B435,B
size = (10 X 250 Ml,12 X 440 Ml,15 X 440 Ml,18 X 440 Ml,20 X 440 Ml,4 X 440 Ml,4 X 500 Ml,4 X 568 Ml,500 Ml,6 X 330 Ml,660 ML,8



Section 4

Simulation results

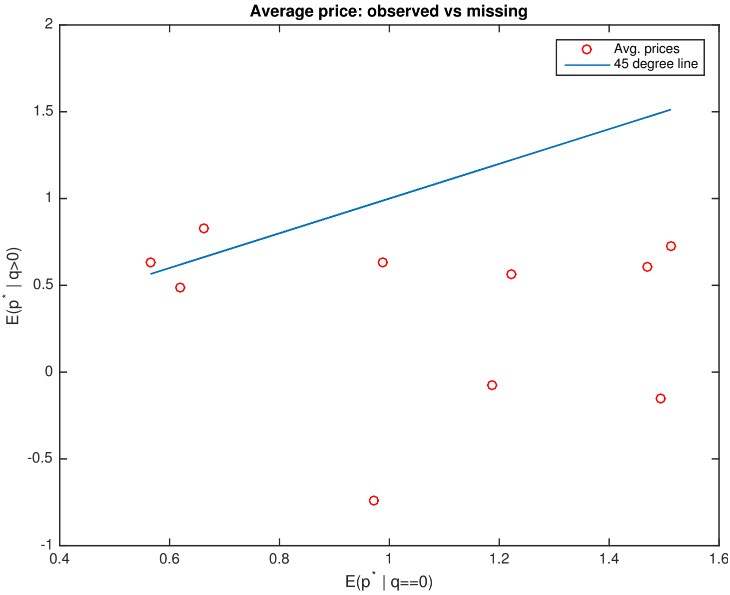
Preliminary simulation results

- Simulate data for an economy with J=10, T=100, R = 2.
- Estimated parameters using EM algorithm and penalised likelihood.
- Analysis of Monte Carlo results in progress.
- Some limited experiments with J = 50 and J=100. Requires parallel computation to speed up estimation. Work in progress.
- Below, I show some results on missing data bias from small scale experiment.

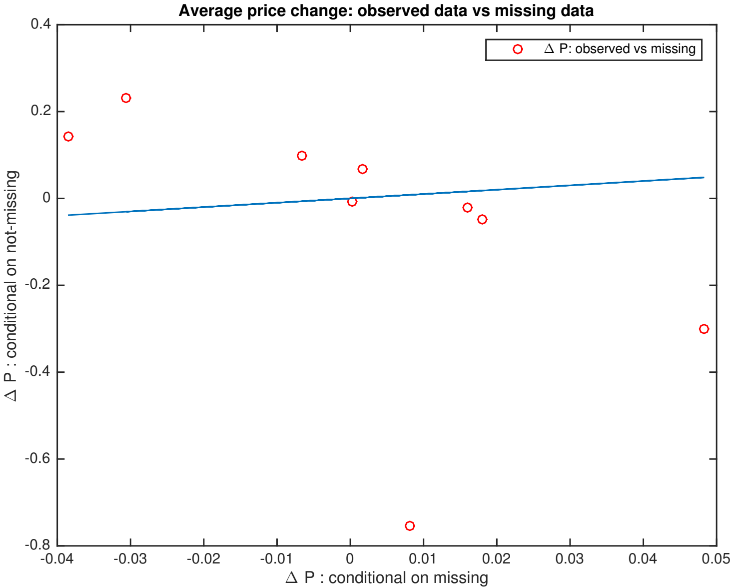
Bias due to missing data

- Plot $E\left(p_{jt}^{*}\middle|\, q_{jt} > 0\right)$ vs. $E\left(p_{jt}^{*}\middle|\, q_{jt} = 0\right)$.
- Plot $E\left(s_{jt}^{*}\middle|\, q_{jt} > 0\right)$ vs. $E\left(s_{jt}^{*}\middle|\, q_{jt} = 0\right)$.
- Plot $E\left(\Delta p_{jt}^{*}\middle|\, q_{jt} > 0 \text{ and } q_{j,t-1} > 0\right)$ vs. $E\left(\Delta p_{jt}^{*}\middle|\, q_{jt} = 0 \text{ or } q_{j,t-1} = 0\right)$.
- Plot true β_{jj} vs estimate from OLS regression of q_{jt} on (p_{jt}, f_t).

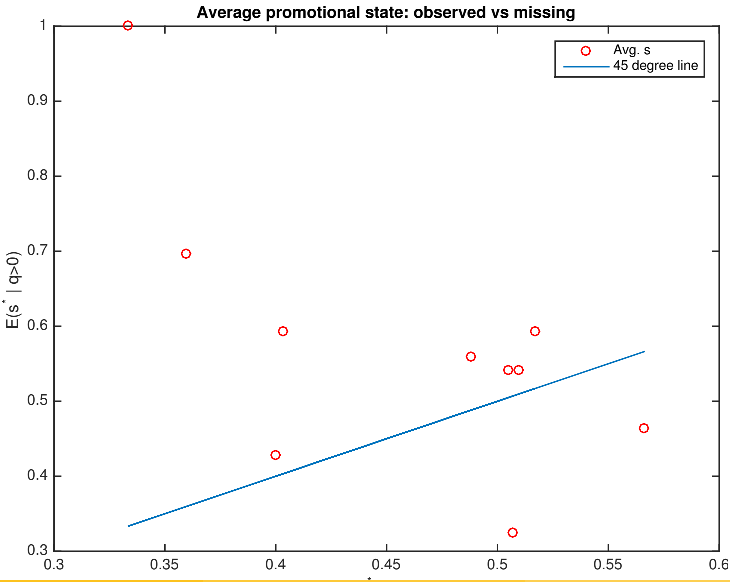
Bias in price levels: simulation



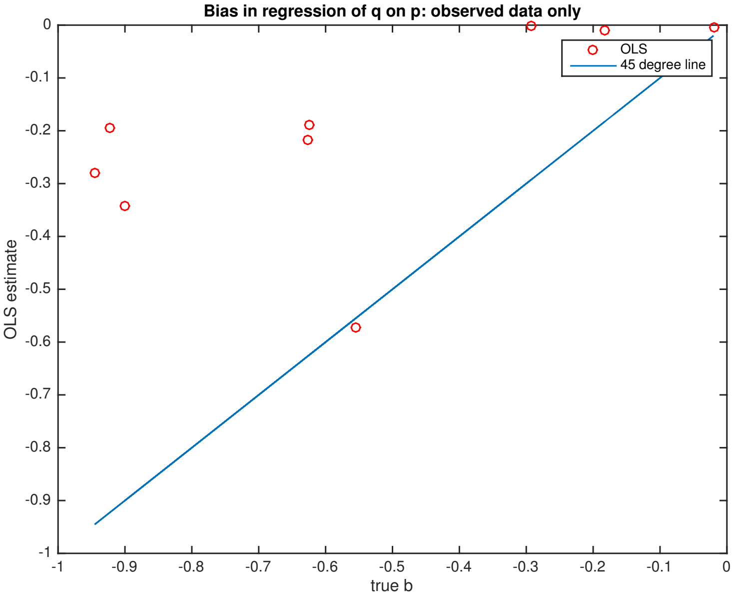
Bias in price changes: simulation



Average sales state: conditional on missing status: simulation



Bias in slope of demand function: simulation



Section 5

Conclusions and extensions

Price index

- From 2023, ONS will integrate scanner data with web-scraped and traditional data sources.
- Our method allows to construct multilateral price index (Diewert and Fox, 2018) that accounts for missing data.
- For new product introductions, our method can be used to estimate the virtual price at which demand equals zero.

Ongoing work

- Estimate model for a few categories of food (shampoo, beer, laundry detergent)
- Empirically analyse how results affect price indices.
- Complete work on sticky price model.
- Applications to competition policy, optimal taxation.
- Developing model of firms' high frequency price setting competition.

Mingli Chen, Iván Fernández-Val, and Martin Weidner. Nonlinear panel models with interactive effects. *arXiv preprint arXiv:1412.5647*, 2014.

Mingli Chen, Iván Fernández-Val, and Martin Weidner. Nonlinear factor models for network and panel data. *Journal of Econometrics*, 220(2): 296–324, 2021.

W Erwin Diewert and Kevin J Fox. Substitution bias in multilateral methods for cpi construction using scanner data. *UNSW Business School Research Paper*, (2018-13), 2018.

Kenneth L Judd. Cournot versus bertrand: A dynamic resolution. *Hoover Institution*, 1996.

Finn Kydland. Noncooperative and dominant player solutions in discrete dynamic games. *International economic review*, pages 321–335, 1975.

Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.