Missing in action

testing alternative imputation methods in price statistics

Patrick Kelly and Matlhatsi Mogalanyane
Statistics South Africa
8 May 2019
Aim

1. Why prices go missing and why do we care?

2. What others have said about dealing with missing prices.

3. Test the most effective ways to deal with missing prices.
Why do prices go missing?
And why do we care?
Why do prices go missing?

- Temporarily missing
- Out of stock
- Seasonal
- Permanently missing
- Discontinued
- Temp unavailable > 2 months

Have you seen me?
Aim of imputing

• Missing price creates a gap in sampled items – could create bias

• Objectives of imputing
  • “First do no harm” – Minimise bias
  • Complete matrix of prices for immediate use of next price when available
Imputation methods

**Overall mean**
Price relatives of all matched varieties in elementary index

**Targeted mean**
Price relatives of a subset of matched varieties in elementary index

**Carry forward**
Price of the variety from the previous month

**Omit**
Do nothing
What about regression?

- Time product dummy commonly used for quality adjustment when substituting for permanently missing prices.

- Uses longer time series than traditional imputation techniques.

- Test it as a method for imputing temporarily missing prices (will it help with seasonal products?)
Literature
What have others said?
Literature

• Very little quantitative analytical work

• Exception is Swiss study – test performance of overall and class mean for quality adjustment when substituting permanently missing varieties of clothing

• Class mean performed better than overall mean but needs bigger sample
Methodology
Methodology

- 25 month dataset for 7 products using South African CPI
- Imputed all missing prices to get complete matrix of data
- Randomly deleted 10% of prices in each product group

Description of data

<table>
<thead>
<tr>
<th>COICOP group</th>
<th>Product</th>
<th>Pricing behaviour</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread and cereals</td>
<td>Bread</td>
<td>Stable</td>
<td>758</td>
</tr>
<tr>
<td>Fruit</td>
<td>Peaches</td>
<td>Strong seasonal</td>
<td>10</td>
</tr>
<tr>
<td>Vegetables</td>
<td>Broccoli</td>
<td>Weak seasonal</td>
<td>57</td>
</tr>
<tr>
<td>Milk, cheese and eggs</td>
<td>Eggs</td>
<td>Stable</td>
<td>414</td>
</tr>
<tr>
<td>Clothing</td>
<td>Men’s shorts</td>
<td>Strong seasonal</td>
<td>197</td>
</tr>
<tr>
<td>Clothing</td>
<td>Men’s jeans</td>
<td>Stable</td>
<td>180</td>
</tr>
<tr>
<td>Furniture</td>
<td>Bedroom suites</td>
<td>Sticky</td>
<td>135</td>
</tr>
</tbody>
</table>
Methodology

Imputed missing prices using 4 methods

• Overall mean - all available price relatives for product

• Targeted mean - price relatives of a subset of varieties for that product (specific geographic area)

• Carry forward - use price in t-1 as the price in t.

• TPD - all data from the product group for the current and previous 12 months.
Methodology

Compare imputed price and index to:

• Original data
• Data with missing prices
Results
Measure deviation of price and index from the original data set
### Price level: Difference between actual and imputed price

<table>
<thead>
<tr>
<th>Item</th>
<th>Targeted mean</th>
<th>Overall mean</th>
<th>TPD</th>
<th>Carry forward</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bread</strong></td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Eggs</strong></td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>Peaches</strong></td>
<td>0.03</td>
<td>0.02</td>
<td>0.16</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>Broccoli</strong></td>
<td>0.04</td>
<td>0.03</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Men's shorts</strong></td>
<td>0.03</td>
<td>0.03</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Men's jeans</strong></td>
<td>0.05</td>
<td>0.04</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Bedroom suites</strong></td>
<td>0.05</td>
<td>0.04</td>
<td>0.10</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Index level: Deviation of index by imputation type

Do nothing
Index level: Deviation of index by imputation type – Carry forward
Index level: Deviation of index by imputation type – TPD

<table>
<thead>
<tr>
<th>Bread</th>
<th>Eggs</th>
<th>Broccoli</th>
<th>Peaches</th>
<th>Shorts</th>
<th>Jeans</th>
<th>Bedroom</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing deviation by imputation type with categories: Bread, Eggs, Broccoli, Peaches, Shorts, Jeans, Bedroom, showing points for 'Do nothing', 'Carry forward', and 'TPD'.]
Index level: Deviation of index by imputation type – Overall

<table>
<thead>
<tr>
<th>Bread</th>
<th>Eggs</th>
<th>Broccoli</th>
<th>Peaches</th>
<th>Shorts</th>
<th>Jeans</th>
<th>Bedroom</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

-2.0  -1.0  0.0  1.0  2.0  3.0  4.0  5.0

-2.0  -1.0  0.0  1.0

-2.0  -1.0  0.0  1.0  2.0

Deviation

- Do nothing
- Carry forward
- TPD
- Overall

The graph shows the deviation of index by imputation type for various items. The x-axis represents different categories such as Bread, Eggs, Broccoli, Peaches, Shorts, Jeans, and Bedroom. The y-axis represents deviation values ranging from -2.0 to 5.0. The graph includes data points for each category, indicating different deviation levels under three imputation methods: Do nothing, Carry forward, and TPD, with an overall deviation.
Index level: Deviation of index by imputation type – Targeted

<table>
<thead>
<tr>
<th>Category</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Eggs</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Broccoli</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Peaches</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Shorts</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Jeans</td>
<td>Do nothing</td>
</tr>
<tr>
<td>Bedroom</td>
<td>Do nothing</td>
</tr>
</tbody>
</table>
Index level: Deviation of index by imputation type

- Bread
- Eggs
- Broccoli
- Peaches
- Shorts
- Jeans
- Bedroom

Deviation

- Do nothing
- Carry forward
- TPD
- Overall
- Targeted
Conclusions
Conclusions

- Confirm overall and targeted mean as most reliable methods
- Overall mean performed better with seasonal items
- Overall mean has larger sample – ensure targeted mean has adequate observations
- Performance of TPD was mixed – needs further investigation
- Carry forward performed well for sticky prices – but will miss change when it does happen
- Do nothing creates biggest bias – Not meeting criteria of ‘do no harm’
The end

Thank you