Extreme values in indexes

Long abstract, By Peter B Meyer, November 2019

Preliminary, incomplete, and please do not cite

Short abstract. We examine the behavior of index functions when its input data has outliers in levels or growth rates. A Tornqvist quantity index, for example, cannot process a quantity of zero, and an extreme growth rate in even a tiny subcategory can distort a larger index. We compare the behavior of a Tornqvist to a Fisher or geo-means index. We discuss advantages and disadvantages of addressing the issue by imputation, merging the item into a larger group, smoothing, or by hard-coding a replacement value. We report on findings from addressing outliers in a large sample of data used for productivity measurement.

The problem: Quantity microdata to be used in a quantity index calculation (notably a Tornqvist) has zeros, missing values, and/or negative or extreme values or extremely steep growth rates. What preprocessing steps are appropriate for this, and is there a library of standard code to do it?

The basic guidelines which we can explain and justify suggest to take these steps in order.

1) Impute replacements if possible based on context.

- Contextually relevant statistical or economic facts might include: awareness that a value is missing or zero because of a disclosure problem; or that expenditures were actually negative; or that it is a new line item in a later period; or that this line item has been discontinued -- these last cases tell us whether a zero is actually a good estimate. Similar rules for prices may apply but a price imputation can simply be smoothed when the quantities are small.
- Context may be drawn from other line items, e.g. on the assumption that trends would be the same, or that the products are good substitutes and are in competition.
- Note that imputations can change the totals per period of the original data, so it may be desirable to scale or RAS/raking afterward.

2) Patch over any remaining extreme outliers or missing values based on extreme thresholds of level of growth rates

- Note for user that this act changes totals from the original data

3) Merge line items if there are still zeros, or if the level or growth rates still exceed some inner threshold (e.g. 1:20 -- bounds of .05 and 20 are recommended to keep Tornqvist smooth)

- This does not change the totals and can incorporate price/weight information from the line items being merged; we can show an equation for the definition of the merged item
- A merge of line items need not be done for those time periods for which the values do not have a steepness problem, but the code to create and eliminate line items can be tricky

4) The best approaches having been exhausted, several alternatives remain:

- Exclude the problematic items
- Choose a replacement value based on smoothing outliers over time to a 20:1 ratio compared to the neighboring time periods, or an average of the neighboring values
- Reduce the weights on the remaining problematic observations
- Switch to another index function for this period
- Substitute topcoded or bottom-coded values for zeros or bad values only if still necessary.

Whatever the software code does in steps 2-4, the user/analyst should be alerted, since context may be relevant to improving the choice here.

TSP software built in some such rules.

We can show examples of how these work in practice, and which results seem appealing. We plan to apply these guidelines to our real product-line output data, and to report on the results for the Tornqvist and then compare to behavior of Fisher and perhaps geo-means indexes on the same data, and perhaps try out various implementations on them.

We have the ambition to provide open-source R and/or python code that addresses steps 2-4 in standard ways that appear best.