Physica A 451 (2016) 266–276

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Estimating sales and sales market share from sales rank data for consumer appliances

Samir Touzani *, Robert Van Buskirk

Lawrence Berkeley National Laboratory, 1, cyclotron Rd., Berkeley, CA 94720, USA

HIGHLIGHTS

- The log-normal distribution is well suited to fit the appliances sales distributions.
- The sales ranks are translated into the sales volume using the log-normal distribution.
- Using log-normal sales proxies produces realistic estimates of different market averages.

ARTICLE INFO

Article history:
Received 24 September 2015
Received in revised form 5 January 2016
Available online 28 January 2016

Keywords:
Sales rank
Log-normal distribution
Market average quantities
Price indices
Statistical model

ABSTRACT

Our motivation in this work is to find an adequate probability distribution to fit sales volumes of different appliances. This distribution allows for the translation of sales rank into sales volume. This paper shows that the log-normal distribution and specifically the truncated version are well suited for this purpose. We demonstrate that using sales proxies derived from a calibrated truncated log-normal distribution function can be used to produce realistic estimates of market average product prices, and product attributes. We show that the market averages calculated with the sales proxies derived from the calibrated, truncated log-normal distribution provide better market average estimates than sales proxies estimated with simpler distribution functions.

1. Introduction

Internet-based data collection is beginning to revolutionize market and economics research. One area of activity is called "Big Data" analysis and consists of collecting very large amounts of data from the Internet and exploring a large number of inferential relationships that may exist in the data [1]. Another related, but more specialized form of Internet-based data collection is called "Scraped Data". Scraped data are available from on-line websites and databases and are collected by means of specialized computer programs that either "crawl" the websites and parse the web page text [2] or interface more directly to back-end databases through the utilization of an application programmer interface (API) made available by the website providers [3].

One very important category of scraped data is product prices and attributes, which are available from on-line retailers. These prices, which are available in real-time, are beginning to find application to real-time price monitoring, especially for the calculation of price indices [4]. With the increasing availability and use of such product market and sales data collected over the Internet, it becomes important to estimate the sales volume corresponding to different product models and product offers observed on Internet websites. Sales volumes estimates are necessary for assigning appropriate weights to different sales offers when product market average quantities are calculated.

* Corresponding author.
E-mail address: stouzani@lbl.gov (S. Touzani).

http://dx.doi.org/10.1016/j.physa.2016.01.030

The price index calculations by Cavallo [2] based on scraped data use weights when aggregating price indices from different product categories, but do not use quantity weights for individual products within a product category. However, several Internet sources provide sales ranks for individual products relative to other products in their category. Sales rank data can be included in the scraped data acquired from the Internet when prices and product attributes are collected. Conceptually these sales rank data can be used to develop quantity proxies that can be incorporated into price index calculations. When sales rank can be used to estimate sales volume this can provide a means of estimating weights for the different model-specific price changes which enable the calculation of price indices that utilize these weights. This approach has been used to measure book prices [5], but has not yet been generalized to other products.

In this study, we investigate the details of how sales quantity estimates can be constructed from sales rank data for household appliances. This study aims to improve the derivation of quantity proxies for market data acquired over the Internet by refining the modeling of the relationship between sales rank and sales volume. Specifically, we use detailed point-of-sales (POS) data for refrigerators, freezers, and clothes washers to estimate improved distribution functions for model-by-model sales.

A review of the literature shows that both power law functions and log-normal distribution functions have been used to approximate economic distributions. Chevalier and Goolsbee [5] used a power law function to describe online books sales, while Hisano and Mizuno [6] observed that the sales distribution of consumer electronics follows both power law and log-normal distributions. Stanley et al. [7] used a log-normal distribution to fit the size distribution of firms. Okuyama et al. [8] showed that a power law holds for the distribution of Japanese companies income. Mizuno et al. [9] observed that the density of the expenditure of a person in a single shopping trip follows a power law. Antoniou et al. [10] applied the log-normal distribution to fit some stock market data. More examples of occurrence of power law and log-normal distribution in empirical data can be found in Stanley et al. [11], Amaral et al. [12], D'hulst et al. [13], Newman [14], Podobnik et al. [15,16], Clauset et al. [17], Ghosh et al. [18], Pinto et al. [19] and Wang et al. [20].

In this work we empirically show that the log-normal distribution (specifically the truncated version) produces an accurate approximation of sales using sales rank data in the context of appliances. We then demonstrate the application of this improved distribution function to providing quantity proxies for the estimation of market average product prices and product attributes (specifically appliance capacity). We also examined power law distributions, but their performance was consistently inferior to the simple log normal distribution, and we do not report the details here.

When there are no limitations to data access, the data-weighting technique commonly used is to weight the data for each model in proportion to the actual sales of that model. If there are no data for estimating quantity proxies, then the simplest alternative data-weighting method is to calculate market averages weighting the values for each product model equally. We show that the method of giving each model an equal weight performs poorly in estimating market average prices or market average appliance attributes or prices compared to estimates of sales-weighted market averages.

Finally we show that the parameters of the improved distribution function are sufficiently stable, that a distribution function calibrated with historical POS data can be used to forecast sales weights from future sales rank data. Specifically, we show that, if one calibrates the parameters of the truncated log normal distribution function with historical POS data, then it is possible to accurately estimate market average quantities (such as market average price or market average appliance capacity) with future sales rank data only. The market average estimates are made by using the truncated log normal distribution function to estimate the sales quantity proxy from the sales rank data. This method of estimating market averages from sales rank can provide an accurate approximation of actual sales-weighted market average quantities if sales ranks are known with accuracy and if the distribution function parameters can be estimated.

The paper is organized into the following sections. Section 2 introduces our method to compute sales proxies using the sales rank. Section 3 provides an overview of the data used in this study. Section 4 presents the empirical results of fitting the distribution function to POS data. Section 5 compares the performance of a log normal distribution with the performance of a truncated log-normal distribution for forecasting sales from sales rank data and shows that a truncated log-normal distribution provides improved performance. In Section 6 we demonstrate the application of the distribution function to the estimation of market average quantities from sales rank data, and demonstrate the superior performance of the truncated log normal distribution function. In Section 7 we demonstrate the application of our method for estimating sales proxies to the calculation of a weighted version of an online price index.

2. From sales ranks to sales volume

If we consider sales quantity as a random variable S and let $(s_1, \ldots, s_N)$ be a set of N observations, then if the data are observed in rank order, i.e. the largest is first and the smallest is last, then the function that provides the observed rank to sales can be written in the following form:

$$F_S(s_i) = 1 - \frac{r_i}{N} \tag{2.1}$$

where s_i is the quantity of sales of the ith model, r_i the corresponding sales rank, N the number of considered appliance models, and $F_S(s_i)$ corresponds to the cumulative distribution function defined as

$$F_S(s) = P(S \leq s) = \int_{-\infty}^{t=s} f(t)\mathrm{d}t \tag{2.2}$$