International Journal of Approximate Reasoning 60 (2015) 1–7

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Willingness-to-pay estimation using generalized maximum-entropy: A case study

Songsak Sriboonchitta, Jianxu Liu, Jirakom Sirisrisakulchai *

Faculty of Economics, Chiang Mai University, Chiang Mai 50200, Thailand

A R T I C L E I N F O

Article history:
Received 22 December 2014
Received in revised form 2 February 2015
Accepted 24 February 2015
Available online 28 February 2015

Keywords:
Maximum entropy
Econometrics
Ordered probit
Interval regression

A B S T R A C T

Estimation of potential customers' willingness-to-pay provides essential information for setting the price of new products. When no market data are available, one usually has to resort to customer surveys. To avoid biases encountered when directly asking respondent how much they would be willing to pay for some products, a useful strategy is to propose some tentative prices and ask the customers whether they would agree to buy the product at those prices. The resulting data can then be analyzed using latent variable models. However, it is often very difficult to specify the error distribution for such models. In this paper, we investigate the use of generalized maximum-entropy (GME) approach as a solution to this problem. Using simulations, this method is shown to be robust to misspecification of the error distribution. As an illustration, the approach is then applied to the determination of the entrance fee to the Royal Park Rajapruek in Chiang Mai, Thailand.

1. Introduction

Valid estimates of willingness-to-pay (WTP) are essential components to develop an optimal pricing strategy. For this reason, providing such estimates has been an important goal of marketing research in the last decades [1]. A first approach is to study customer preferences based on actual or simulated market data. However, this method is not feasible in the case of a totally new product. An alternative approach is to use survey data. In the so-called direct surveys, respondents (usually, selected customers) are asked how much they would be willing to pay for some product. However, this is a challenging task for respondents, and this kind of procedure may provide strongly biased estimates. For instance, a respondent may quote an artificially low price as a result of a "consumer collaboration effect" or, on the contrary, he or she may overestimate the price to avoid appearing stingy. To circumvent these difficulties, some indirect survey approaches have been proposed, in which prices are systematically varied and customers are asked to state whether they would be willing to purchase the good at that price. The difficulty is then to find a statistical model allowing us to draw reliable conclusions from such data.

An example of an indirect survey approach is given in [8], where the problem was to estimate the WTP for the entrance fee of Royal Park Rajapruek in Chiang Mai, Thailand. In this study, respondents were first proposed a randomly chosen initial bid and then, depending on their decision (accept or reject), a lower bid or an upper one [9]. This procedure thus provided interval-valued data consisting of a lower bound and/or an upper bound for each respondent's unobserved WTP. The problem was then to determine the influence on WTP of covariates such as the age, sex or income, using interval regression. The validity of such an analysis depends crucially on the adequacy of the model, characterized here by the

* Corresponding author.
E-mail address: sirisrisakulchai@hotmail.com (J. Sirisrisakulchai).

http://dx.doi.org/10.1016/j.ijar.2015.02.003

error distribution. If knowledge of the data generating process is available and the correct model is specified, the maximum likelihood (ML) estimator is consistent. However, such knowledge is rarely available, and model misspecification may result in inconsistent estimates. In the particular case of the study in [8], knowledge about the error distribution was not available, and none of the usual families of distribution had a good fit with the residuals of the regression. This example points to the need for distribution-free inference methods capable of extracting most of the relevant information in the data, with minimal assumptions about the data generating process. The objective of this paper is to study the application of one such method – the generalized maximum-entropy (GME) approach [3,2]. As will be shown using both Monte Carlo simulation and real data, this approach is more robust than the ML method and it performs well over a range of data distributions.

The rest of this paper is organized as follows. The model and the GME method will first be described in Section 2. Simulation results will then be presented in Section 3, and the method will be applied to real data from the Royal Park Rajapruek case study in Section 4. Finally, Section 5 will conclude the paper.

2. Generalized maximum entropy

In this section, we first define the interval regression model in Section 2.1. The reformulation of this model using the GME approach is then described in Section 2.2.

2.1. Interval regression model

Let y_i^* denote the willingness to pay for a potential customer i, $i = 1, \ldots, n$. We assume that each y_i^* can be explained by a vector of covariates $\boldsymbol{x}_i$ using an equation of the form

$$y_i^* = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is a vector of coefficients to be estimated and ε_i is an error term, assumed to be independent of the vector $\boldsymbol{x}_i$.

In the indirect survey approach considered here, the y_i^* are latent variables. Information about y_i^* is obtained as follows. First, we randomly generate an initial bid P_i. If this bid is accepted, we generate an upper bid $P_i^u > P_i$. Otherwise, we generate a lower bid $P_i^l < P_i$. This procedure yields, for each respondent, a lower bound μ_i^l and an upper bound μ_i^u for the unobserved valued y_i^*, with

- $\mu_i^l = -\infty$ and $\mu_i^u = P_i$ if the first and second bids are rejected;
- $\mu_i^l = P_i^l$ and $\mu_i^u = P_i$ if the first bid is rejected and the second is accepted;
- $\mu_i^l = P_i$ and $\mu_i^u = P_i^u$ if the first bid is accepted and the second is rejected;
- $\mu_i^l = P_i^u$ and $\mu_i^u = +\infty$ if the first and second bid are accepted.

With the assumption that the error terms ε_i are normal and identically distributed with zero mean and variance σ^2, the classical ordered probit model based on the maximum likelihood approach could be used to estimate the parameter vector $\boldsymbol{\beta}$ (see, e.g., [9]). In practice, however, no knowledge about the data generation process is usually available. As shown in Section 4, specifying a model with a good fit to the data may be a very difficult task. To avoid these complications, non-parametric methods such as the GME approach can be used. This approach, presented in the next section, is more flexible than maximum likelihood as it makes minimum assumptions about the distribution of the data.

2.2. Model reformulation

The GME method [5,2] is based on the maximum entropy (ME) principle [6]. The main idea of the method is to treat both the parameter vector $\boldsymbol{\beta}$ and the errors ε_i as discrete random variables with bounded support. The corresponding probability distributions are determined by maximizing the Shannon entropy under first-moment constraints.

More precisely, let $\{z_{k1}, \cdots, z_{kM}\}$ denote the support space of component β_k of $\boldsymbol{\beta}$. In the absence of prior knowledge about the possible values of β_k, a simple strategy is to define the support space to be centered on zero with wide range. The probability distribution of β_k is then denoted by $\boldsymbol{p}_k = (p_{k1}, \ldots, p_{kM})'$. Similarly, the error term ε_i is assumed to be a discrete random variable with support $\{v_1, \cdots, v_J\}$ and probability distribution $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iJ})$. In classical regression problems where the response variable y_i^* is observed, Golan et al. [4] recommend using the "three-sigma rule" to establish bounds on the error components: the lower bound is $v_1 = -3s_y$ and the upper bound is $v_J = +3s_y$, where s_y is the empirical standard deviation of the observations y_i^*. For example if $J = 5$, then the support space of the error can be set to $\{-3s_y, -1.5s_y, 0, 1.5s_y, 3s_y\}$. When the y_i^* are latent, some other strategy has to be employed (see Sections 3.1 and 4).

According to the ME principle, the probability distributions $\boldsymbol{p}_k$ and $\boldsymbol{w}_i$ can be chosen to maximize the following entropy function,

$$H(p, w) = -\sum_{k=1}^{K}\sum_{m=1}^{M} p_{km} \log p_{km} - \sum_{i=1}^{n}\sum_{j=1}^{J} w_{ij} \log w_{ij}, \tag{2}$$