2010

Testing the Validity of a Demand Model: An Operations Perspective

Omar Besbes
University of Pennsylvania

Robert Phillips

Recommended Citation
Besbes, O., & Phillips, R. (2010). Testing the Validity of a Demand Model: An Operations Perspective. *Manufacturing & Service Operations Management, 12* (1), 162-183. http://dx.doi.org/10.1287/msom.1090.0264

Testing the Validity of a Demand Model: An Operations Perspective

Abstract
The fields of statistics and econometrics have developed powerful methods for testing the validity (specification) of a model based on its fit to underlying data. Unlike statisticians, managers are typically more interested in the performance of a decision rather than the statistical validity of the underlying model. We propose a framework and a statistical test that incorporate decision performance into a measure of statistical validity. Under general conditions on the objective function, asymptotic behavior of our test admits a sharp and simple characterization. We develop our approach in a revenue management setting and apply the test to a data set used to optimize prices for consumer loans. We show that traditional *model-based* goodness-of-fit tests may consistently reject simple parametric models of consumer response (e.g., the ubiquitous logit model), while at the same time these models may "pass" the proposed *performance-based* test. Such situations arise when decisions derived from a postulated (and possibly incorrect) model generate results that cannot be distinguished statistically from the best achievable performance—i.e., when demand relationships are fully known.

Keywords
pricing, parametric and nonparametric estimation, model misspecification, hypothesis testing, goodness-of-fit test, asymptotic analysis, performance analysis

Disciplines
Other Business | Other Medicine and Health Sciences

Testing the Validity of a Demand Model: An Operations Perspective

Omar Besbes* University of Pennsylvania

Robert Phillips† Nomis Solutions

Assaf Zeevi‡ Columbia University

First submitted: 06/2008, Revised: 11/2008, 02/2009

To appear in *Manufacturing & Service Operations Management*

Abstract

The fields of statistics and econometrics have developed powerful methods for testing the validity (specification) of a model based on its fit to underlying data. Unlike statisticians, managers are typically more interested in the performance of a decision rather than the statistical validity of the underlying model. We propose a framework and a statistical test that incorporates decision performance into a measure of statistical validity. Under general conditions on the objective function, asymptotic behavior of our test admits a sharp and simple characterization. We develop our approach in a revenue management setting and apply the test to a data set used to optimize prices for consumer loans. We show that traditional *model-based* goodness-of-fit tests may consistently reject simple parametric models of consumer response (e.g., the ubiquitous logit model), while at the same time these models may "pass" the proposed *performance-based* test. Such situations arise when decisions derived from a postulated (and possibly incorrect) model, generate results that cannot be distinguished statistically from the best achievable performance – i.e., when demand relationships are fully known.

Keywords: pricing, parametric and non-parametric estimation, model misspecification, hypothesis testing, goodness of fit test, asymptotic analysis, performance

*The Wharton School, e-mail: obesbes@wharton.upenn.edu

†e-mail: robert.phillips@nomissolutions.com

‡Graduate School of Business, e-mail: assaf@gsb.columbia.edu

1. Introduction

1.1 Motivation and the main objective

Prescriptive solutions to applied problems in operations management invariably hinge on the specification of a model for the underlying system or phenomenon of interest. The key model primitives (e.g., demand distribution, service time distribution, functional relationships between decision variables and observed response, etc.) are typically estimated from historical observations, and subsequent to that a suitable model-based objective function is optimized to arrive at operating decisions.

An important diagnostic step in this roadmap involves "tweaking" the model to better fit the data or to increase its predictive power, relying on standard measures of statistical fit such as Mean Average Percentage Error (MAPE), concordance, Bayesian Information Criterion (BIC), etc. More generally, statisticians and econometricians have developed powerful tools to test the *validity* of a model based on its fit to underlying data. These tools allow an analyst to determine, by means of a suitable statistical test, whether the postulated model is well specified (or not) in a statistically significant sense. In contrast to the search for statistical significance, managers are likely to be more interested in the quality of *decisions* derived from a given model; this suggests the value of an operations-centric view as opposed to the traditional model-centric one.

Our goal in this paper is to propose a framework and a statistical test for evaluating models, that formalize an *operations perspective*. We focus on the problem of specifying a model for the relationship between price and realized demand, which is central to the areas of revenue management and economics. The main reason for this focal point, beyond the desire to be concrete, stems from the research having its origins in an empirical analysis of pricing data, the details of which are discussed in Section 8. The main question that surfaced there concerned the use of simple models of consumer choice (such as the logit or probit) in applied revenue management. In particular, in such settings the model is first calibrated to data (i.e., its parameters are estimated), and then placed within an optimization problem to support pricing and/or capacity allocation decisions; see Talluri and van Ryzin (2005) and Phillips (2005) for examples and further pointers to the literature.

The sheer parsimony of the logit/probit-type models, among other documented deficiencies, suggests that they are unlikely to pass a formal statistical model testing procedure. At the same time, as mentioned above, these are by far the most widely used models in practice. This naturally leads to the following question:

> What is the loss incurred by using decisions derived from a restricted class of models, relative to the best achievable profits had one known the true relationship between price and average demand?

If the above loss is suitably "small," the assumed demand model, whether well specified or not,

might be deemed adequate from a revenue manager's perspective. Much of this paper deals with putting these statements on rigorous ground, articulating a suitable formulation for testing them, and quantifying what "small" means in a precise statistical sense.

1.2 A simple illustration of the key idea

Consider the following simple setup which is characteristic of many pricing problems. Consumers inspect a product or service on offer by a company. The fraction who will purchase it is governed by a so-called *response function* or equivalently, a willingness-to-pay distribution. Denote by $\lambda(x)$ the probability that an arbitrary customer within this population will purchase at a price x. (This function is assumed not to change over the relevant time horizon over which decisions are made.) The revenue manager's objective is to maximize, say, the expected profit-per-customer $\pi(x) = (x - x_0)\lambda(x)$, by suitably setting the price x (where x_0 is the cost per unit sold). Since the true underlying response function is not known, a class of (demand) models is put in place based on which the profit maximization problem is solved.

In Figure 1(a), we depict two response functions (demand models) that are visibly quite distinct; think of one [solid line] as corresponding to the "true" underlying model $\lambda(\cdot)$, and the other [dashed line] as corresponding to the postulated one. In Figure 1(b) the profit function associated with each model is represented: the solid line depicts the true profit function; and the dashed line represents the profit function corresponding to the postulated model. The *optimal price* corresponding to the true response function [solid line] is \$5, while the price prescribed by the postulated model [dashed line] is \$6.

As discussed in Section 1.1, the traditional statistical perspective on model testing would ask whether the two curves in Figure 1(a) are suitably "close" or "far apart," and strive to conclude whether the postulated model is misspecified relative to the true one. (For this purpose, imagine that all models are constructed on the basis of a finite number of historical purchase decisions.) In contrast, the operations perspective advocated in this paper asks whether the expected profit rate generated by the model-based decision $\pi(6)$, is significantly different, in a statistical sense, from the *optimal profit rate* $\pi(5)$. The former describes the profits that are achieved using the model-based prescription, when consumer behavior is dictated by the *true* underlying response function. Figure 1(b) depicts this difference (Δ) graphically. If Δ is suitably small, so that based on a sample of past sales it cannot be distinguished from zero in a well defined statistical sense, then one *would not* reject the model-based estimate.

It is worth pointing out that the setting we are focusing on is static, in the sense that a single optimal operating point is sought. In dynamic settings, where the optimal price levels typically change over time due to capacity and product perishability considerations, the traditional model-testing approach may be more appropriate, as one needs to ensure a valid "global" representation

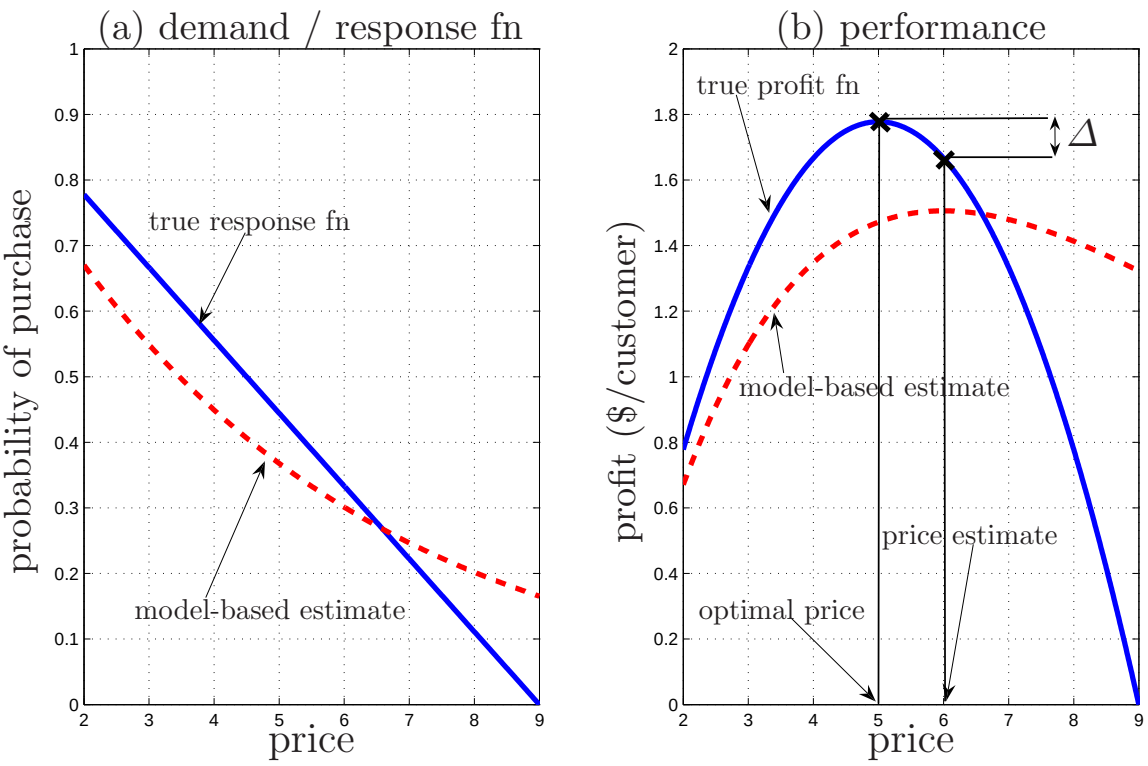

Figure 1: **Demand model misspecification and performance implications.** Δ = difference between optimal profit rate and profits achieved using model-based price-estimate. The magnitude of Δ quantifies the profit loss due to model misspecification.

of the demand model.

1.3 The main contributions and key qualitative insights

Main contributions. Our main objective is to provide a rigorous foundation for the operations-centric approach. This is done by first formalizing the aforementioned hypothesis test in Section 5. (Some preliminaries and background on the traditional model-based testing approach are given in sections 3 and 4.) The next step is fleshing out how Δ, the profit loss due to the use of a given model class, can be used as a formal test statistic to resolve the above mentioned test. This is done in sections 5 and 6.

To determine what values of Δ differ from zero in a statistically significant manner, we develop limit theory which establishes that a suitably normalized estimator of Δ converges in distribution to a simple limit as the sample size grows large (Theorems 2 and 3). This limit is given by a scaled Chi-squared distribution $\gamma\chi^2$, based on which p-values can be derived to determine whether one rejects the null or not (i.e., whether a postulated model should be rejected from this operations-based

perspective). The scalar γ is characterized explicitly and its magnitude is seen to be proportional to the level of "noise" associated with the observations, and inversely proportional to the "flatness" of the profit function in the neighborhood of the point of maximum. Thus, more noise and increased flatness contribute to a larger value of γ, which in turn makes it harder to reject the null; see Remark 1 following Theorem 2 for further discussion. Numerical experiments in Section 7 focus on synthetic examples similar to the one described in Figure 1, and illustrate some properties and efficacy of the proposed test. The proof of the pudding is in the eating, and to that end we discuss in Section 8 the application of our method to an empirical data set arising in financial services (see further comments below).

In terms of methodology, we rely on non-parametric estimation techniques to construct and study the properties of a consistent estimator of the true unknown response function; for accessible overviews of non-parametric estimation, see Härdle (1990) and Pagan and Ullah (1999). Large sample theory of maximum likelihood estimation in a misspecified environment also plays an important role in our derivations; see, e.g., White (1996) and the literature review below for general references on these topics.

To the best of our knowledge, the present paper is the first to propose a statistical test that strives to assess the validity of a model from purely operational considerations, i.e., focusing on decision making implications of model specification. In addition, the essence of the approach transcends the revenue management application which is the focal point of the paper. In Section 9, we indicate how to "cut and paste" the various modules in our proposed approach and apply them in the context of two prototypical inventory and queueing problems.

Bearing on practice. Both theory and practice of revenue management, certainly within the context of the example discussed in Section 1.2, stress the use of simple models of demand and consumer choice; the logit and probit are two of the most ubiquitous examples. This is due by and large to a combination of convenience and mathematical tractability: simple models are easy to understand, calibrate and optimize. On the other hand, it is difficult to expect such models to fully capture the true underlying behavior, and hence also unlikely that they survive rigorous statistical tests for model specification; see the literature review in Section 2 for references. This has led to the introduction of increasingly complex models that are variants of the basic ones mentioned above (e.g., nested logit and hierarchical models), with most of this work being pursued within the marketing community; see, e.g., Leeflang et al. (2000). On the other hand, if one inspects various application areas of revenue management one still finds prevalent use of simple models such as logit. The approach and tools developed in this paper can help in providing further theoretical support for the use of such models.

With that point in mind, we refer the reader to Section 8 in the paper that contains an empirical study based on data from a firm that offers automobile loans. Our analysis shows that at least

in several instances within that data set, a simple logit model is rejected based on a standard model-based test, but there is not enough statistical evidence to reject it from the perspective of our operations-centric test. This offers a potential explanation for the prevalent use of the logit model in practice. Roughly speaking, the intuition behind the above observations lies in the fact that traditional model-based tests are predicated on *global* criteria of functional fit, while our test is *local* in nature, focusing only on the relevant region of interest (the neighborhood of the maximum profits).

Given that the test we propose relies on nonparametric estimates, it requires the availability of sufficient data. One could argue that in such settings it is possible to resort to non-parametric modeling of the response function. However, practitioners would most likely be reticent to depart from the world of simple parametric models in favor of more complicated and opaque non-parametric ones, even in data-rich environments. Our proposed test, while relying on non-parametric estimation techniques for diagnostic purposes, is actually more likely to lend support to simple parametric models due to its less stringent nature (relative to model-based tests).

2. Literature review

Our focus in this paper is on a revenue management application which centers on consumer choice behavior. There is a large stream of literature focusing on such models; Ben-Akiva and Lerman (1985) and Train (2002) provide an overview on the topic. As discussed earlier, various measures of fit are used in practice, based on a significant body of work dating back to the pioneering paper of Akaike (1974) that deals with model *selection*; the typical approach there is to formulate an optimization problem that penalizes the complexity of a model (e.g., log-likelihood with penalty for the number of parameters in the model). It is important to note that all models considered in these comparisons may still be misspecified with respect to the true mechanism that generates the data. Amemiya (1981, Section II.C) reviews some criteria typically used for model selection, and gives a general econometrics perspective on this and related issues; Leeflang et al. (2000) provide a general overview from a marketing perspective.

Distinct from that line of research is the model *testing* paradigm in which the specification of a model is tested against the true underlying structure of interest; this approach is covered in almost any graduate level textbook on statistical theory; see, e.g., Borovkov (1998). Broadly speaking, our paper falls into this category. For the purpose of our revenue management application, the response function is a conditional probability and hence one can draw on general results that have been developed in the literature for testing the validity of regression (conditional expectation) type models. Perhaps the two most notable examples are the conditional moment tests of Bierens (1990), and the conditional Kolmogorov test of Andrews (1997); see also references therein for further pointers to this literature. Both these tests enjoy certain optimality properties in terms of

their power against local alternatives. Unlike these types of tests that do not directly estimate the regression function, there are various others that use intermediary non-parametric approximations to the regression function; see, e.g., Härdle and Mammen (1994).

Various instances of the model specification tests mentioned above have been applied to empirical data sets in order to assess the validity of widely used models such as the logit or probit. Horowitz (1993) analyzes the binary response model of choice between automobile and public transit, and Bartels et al. (1999) apply a non-parametric test to scanner panel data and reject the multinomial logit model; see also references therein. While not focusing on testing per se, Abe (1995) discusses benefits and drawbacks of non-parametric models relative to simple parametric ones (logit) in the context of marketing research.

Our work is also related to a stream of work in operations management which highlights operational objectives when estimating a model for an underlying system. Cachon and Kok (2007) analyze the perils associated with salvage value estimation in the context of the newsvendor problem and Liyanage and Shantikumar (2005) study the interaction between optimization and estimation of the demand distribution in a newsvendor setting; see also Ernst and Cohen (1990) for an earlier study of similar flavor. Cooper et al. (2006) study the interaction between demand distribution/parameter estimation and operational objectives, but in a capacity control revenue management problem, highlighting potential consequences of model misspecification. These studies do not focus on model testing per se, nor on statistical analysis, and therefore intersect with our work more in terms of philosophy rather than actual focus and methods.

Our study also shares a common theme with the field of statistical decision theory (see, e.g., the overview in Berger (1980)), as an important feature of the approach we propose is that it takes into account decision making. Similarly, there has long been a recognition within the decision analysis literature that the value of quantitative modeling should be judged primarily by the quality of the decisions they support (see, for example, Nickerson and Boyd (1980)). However, there has been a lack of methodologies for evaluating the adequacy of a particular model from this vantage point.

3. Problem Formulation

A single product can be sold for a price $x \in \mathcal{X} := [\underline{x}, \overline{x}]$, with $0 < \underline{x} < \overline{x} < \infty$. At the prevailing price, x, a consumer will purchase the product with probability $\mathbf{P}\{Y = 1|x\}$. Here $Y \in \{0, 1\}$ is a random variable such that $Y = 1$ corresponds to a purchase decision, and $Y = 0$ corresponds to a situation where the consumer declines to purchase the product. We refer to $\lambda(x) := \mathbf{P}\{Y = 1|x\}$ as the consumer *response function*.

Let $r(x)$ denote a function that describes the revenue/*profit* resulting from a given sale. The decision-maker's objective is to set a price that maximizes the expected profit per customer. That

is, for the expected *profit function*

$$\pi(x) := r(x)\lambda(x), \tag{1}$$

the objective is to seek $x^* \in \arg\max\{\pi(x) : x \in \mathcal{X}\}$. Under mild conditions on $\pi(\cdot)$, e.g., continuity, such a point of maximum exists. The optimal profits are then $\pi^* := \pi(x^*)$. This is an example of the so-called *customized pricing problem* as described in Phillips (2005).

The decision-maker does not know the true response function $\lambda(\cdot)$ characterizing the market. S/he only has access to data in the form of n past observations $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$; each pair (X_i, Y_i) describes the sales outcome $Y_i \in \{0, 1\}$ when offering the product at price $X_i \in \mathcal{X}$. Using this data, a model for the response function is fitted from a parametric family $\mathcal{L}(\Theta) = \{\ell(\cdot;\theta) : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is a compact set. (As discussed in the introduction, this parametric estimation procedure is the typical approach used in practice.) In what follows, we use $\mathbf{P}_\theta(y|x)$ to denote the conditional probability of observing y when x is chosen under the assumed parametric model $\ell(\cdot;\theta)$; in particular, note that this probability might differ from the actual probability of observing y given a choice of x. For any $\theta \in \Theta$, we let $p(x;\theta) := r(x)\ell(x;\theta)$ denote the profit function under the parametric assumptions describing the *postulated* model. Put $x^*(\theta)$ to be a maximizer of $p(x;\theta)$, where $x \in \mathcal{X}$. Hence for a fixed choice of the parameter $\theta \in \Theta$, $x^*(\theta)$ represents the optimal *model-based* decision (price). In practice, the value of the parameter θ is estimated from the data $\mathcal{D}_n$ and we let $\hat{\theta}$ denote such an estimate, i.e., a mapping from $\mathcal{D}_n$ to Θ. While many such estimation procedures exist, we will focus here on the method of maximum likelihood, which is by far the one most prevalently used in practical settings.

Our main objective is to assess the validity of the class of models $\mathcal{L}(\Theta)$ selected by the decision-maker.

4. General Background and Model-Based Testing

We first describe briefly the classical *model-based* approach which is the one found in the statistics and econometrics literature. In passing, we also introduce some necessary background on hypothesis testing and the key concepts that will be used throughout the paper. Subsequent to that, we review a specific model-based test developed by Andrews (1997), that will later serve as a basis for comparison against our proposed performance-based test which is described in Section 6.

4.1 The traditional model-based approach and some general background on hypothesis testing

The traditional statistical approach strives to determine, based on observations $\mathcal{D}_n$, whether the true unobservable conditional probability $\lambda(\cdot)$ can be distinguished from the "best approximation"

within the model class $\mathcal{L}(\Theta)$. Formally, one can formulate the hypothesis test as follows:

$$H_0: \quad \lambda(\cdot) = \ell(\cdot;\theta^*) \quad \text{for some } \theta^* \in \Theta \tag{2}$$

$$H_1: \quad \lambda(\cdot) \neq \ell(\cdot;\theta) \quad \text{for all } \theta \in \Theta, \tag{3}$$

where the '$\neq$' in the alternative hypothesis means that for any $\theta \in \Theta$ there exists some $x \in [\underline{x}, \overline{x}]$ such that $\lambda(x) \neq \ell(x;\theta)$. It is worth emphasizing that what one is testing via this traditional statistical formulation is a hypothesis about the model that generates the data.

The decision rule that is used to resolve the test typically hinges on a suitably chosen *test statistic* $T_n : \mathcal{D}_n \to \mathbb{R}_+$. The idea is that when T_n (properly scaled) exceeds, say, a suitably chosen threshold τ, the null hypothesis H_0 is rejected. Since the distribution of "good" test statistics is often difficult to compute, one resorts to an asymptotic analysis. (Note also that the hypotheses in (2)-(3) are not simple hypotheses, as the distribution of the response Y conditional on the covariate X is not fully specified under H_0.)

Suppose that there is a scaling sequence of positive real numbers $\{a_n\}$ such that $a_n T_n$ converges in distribution to a limit random variable Z; we denote this as $a_n T_n \Rightarrow Z$ as $n \to \infty$[1]. The threshold τ is then chosen so that

$$\mathbf{P}\{a_n T_n > \tau \mid H_0\} \to \alpha \qquad \text{as } n \to \infty, \tag{4}$$

where $\alpha \in (0,1)$ is called the *significance level* of the test. In other words, the choice of τ ensures that the Type 1 probability of error, i.e., the likelihood of rejecting the null when it is true, is asymptotically equal to α. A decision rule or test (we use the two interchangeably in what follows) is said to be *consistent* if

$$\mathbf{P}\{a_n T_n \leq \tau \mid H_1\} \to 0 \qquad \text{as } n \to \infty. \tag{5}$$

This restriction ensures that the Type 2 error, namely, not rejecting the null when it is false, has vanishingly small probability as the sample size increases. Finally, the *p-value* associated with the test is the minimum level of significance at which one rejects the null hypothesis (the smaller the p-value, the more evidence there is to reject the null). In particular, if the p-value falls below the significance level α, the null is rejected.

As a side comment, we note that it is not always possible to directly use the limiting distribution (of Z) to compute p-values, as the latter will in general depend on characteristics of the true demand model which are unknown (for example, this distribution may depend on θ^*). As a result, an additional step is typically needed to arrive at a fully implementable test.

Testing model specification as in (2)-(3) has received significant attention in the fields of economics (econometrics) and statistics as discussed in Section 2. A specific test that falls into this category and has certain desired properties is described in the next section.

[1]Note that identifying an appropriate scaling sequence a_n is often is often a significant step in the analysis, and that such convergence results will in general hinge on structural assumptions characterizing the observations $\mathcal{D}_n$.

4.2 Example of a model-based test

Below, we briefly describe a model-based test developed by Andrews (1997) which resembles the classical Kolmogorov-Smirnov test used to assess if two samples are drawn from the same distribution (see, e.g., Borovkov (1998)). Based on historical data $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$, the test uses the conditional Kolmogorov test statistic which is defined as follows

$$CK_n = \sqrt{n} \max_{j \leq n} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \leq Y_j\} \, \mathbf{1}\{X_i \leq X_j\} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{P}_{\hat{\theta}}(Y \leq Y_j | X = X_i) \, \mathbf{1}\{X_i \leq X_j\} \right|, \qquad (6)$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ, and in our setup $\mathbf{P}_\theta(Y_i \leq Y_j | X_i) = 1$ if $Y_j = 1$ and $\mathbf{P}_\theta(Y_i \leq Y_j | X_i) = 1 - \ell(X_i; \theta)$ if $Y_j = 0$. The conditional Kolmogorov test statistic compares two terms. The first is an empirical version of the cumulative distribution function of the vector (X, Y). As the sample size grows large, this term converges to the true cumulative distribution function (cdf) of the vector (X, Y) by the law of large numbers. The second term is a semi-parametric, semi-empirical version of the cdf of (X, Y), and under the null hypothesis, it also converges to the same limit as above. However, when the null is false, the two limits will differ, and the statistic, which is scaled by $\sqrt{n}$, will diverge to infinity. In that regard, we make the following assumption

Assumption 1 For $i \geq 1$, (X_i, Y_i) are independent and identically distributed (iid) with response function given by $\mathbf{P}(Y = 1 | X = x) = \lambda(x)$ and marginal distribution G with density function $g(\cdot)$ which is positive and continuous everywhere on its support $[\underline{x}, \overline{x}]$.

Assumption 1 is adopted for convenience as it facilitates the mathematical analysis in what follows (namely, the derivation of large sample properties of certain test statistics). From that perspective, it can be significantly weakened to allow for some dependency between (X_i, Y_i) pairs; this is well documented in the statistical and econometrics literature (see, for example, White (1996)). Essentially, the iid assumption together with the postulated density condition are put in place to ensure sufficient "dispersion" in the decision variable X. In the absence of such dispersion, it would be difficiult to reconstruct the response function consistently. In the practice of revenue management, such dispersion can sometimes arise as a consequence of price experiments that are aimed at inferring the nature of consumer preferences and purchasing behavior. The empirical study presented in Section 8 indicates the presence of such price dispersion in Figure 4. (An important problem, which is beyond the scope of the current paper, is the formal quantification of price dispersion present in datasets as well as the design of appropriate price experimentation schemes to achieve adequate dispersion.)

We also make the following technical assumption, which is quite standard in the context of parameter estimation problems.

Assumption 2 *Each member of the family of response function models $\ell(x; \theta) \in \mathcal{L}(\Theta)$ is continuously differentiable on $[\underline{x}, \overline{x}] \times \Theta$.*

Theorem 1 (Andrews (1997)) *Let Assumptions 1 and 2 hold. Under the null hypothesis,*

$$CK_n \Rightarrow \mathcal{K} \qquad \text{as } n \to \infty.$$

The precise characterization of the limit $\mathcal{K}$ is given in Andrews (1997). The limit distribution depends on several nuisance parameters including the true parameter vector θ^* (under the null) and the marginal distribution $G(\cdot)$. In addition, the form of the limit distribution is fairly complicated. To implement the test, Andrews (1997) suggests a bootstrapping procedure. We come back to this test in sections 7 and 8, where we compare its conclusions to those of the approach proposed in the present paper.

5. The Proposed Approach: Key Ideas

Motivation. The model-based approach, and hence the test described in (2)-(3), focuses on the specification of the response model. Roughly speaking, for a given estimator $\hat{\theta}$, the null hypothesis will be rejected when the estimated response function $\ell(\cdot;\hat{\theta})$ differs in a statistically significant manner from the true response model $\lambda(\cdot)$. Yet even under such circumstances it is possible that the estimated response function would still give rise to a "good" pricing prescription, i.e., a price that performs well under the true underlying response function. As illustrated in Figure 1 in the introduction, this is possible even if the postulated model class $\mathcal{L}(\Theta)$ is misspecified in a significant manner relative to the underlying response model $\lambda(\cdot)$.

We now describe a new test that focuses directly on the *performance* of decisions derived from the parametric model class $\mathcal{L}(\Theta)$. In particular, the question that will be answered by this test is whether decisions that are derived from the "best" parametric model in $\mathcal{L}(\Theta)$ lead to actual performance (profits) that differ significantly from (i.e., are inferior to) the *best achievable performance*. The latter corresponds to profits generated by the optimal decision derived from the *true* underlying response model $\lambda(\cdot)$.

Background on parametric inference under model misspecification. To describe our proposed test, we first need to articulate what is meant by the "best" parametric model in $\mathcal{L}(\Theta)$, as that class need not include the true response function $\lambda(\cdot)$; a good reference on the topic is the book by White (1996).

Let

$$\hat{\theta} \in \arg\max \left\{ \frac{1}{n} \sum_{i=1}^{n} \log\big(\mathbf{P}_\theta(Y_i|X_i)\big) : \theta \in \Theta \right\}$$

be the maximum likelihood estimator based on the sample $\mathcal{D}_n$, where $\mathbf{P}_\theta$ denotes the conditional probability distribution of Y given X for a parameter vector $\theta \in \Theta$. In our context, $\mathbf{P}_\theta(y = 1|x) = 1 - \mathbf{P}_\theta(y = 0|x) = \ell(x;\theta)$. The right-hand-side above, which is being maximized, is a sample-based approximation to the expected log-likelihood $\mathbf{E}\Big[\log\big(\mathbf{P}_\theta(Y|X)\big)\Big]$, where the expectation is

with respect to the *true distribution* $\mathbf{P}$ of (X, Y), which may be distinct from any distribution $\{\mathbf{P}_\theta : \theta \in \Theta\}$. It is easily seen that the expected log-likelihood is maximized for a value of $\theta = \theta^*$ which minimizes

$$\mathbf{E}\Big[\log\big(\mathbf{P}(Y|X)/\mathbf{P}_\theta(Y|X)\big)\Big], \tag{7}$$

over the parameter space Θ.

The expression in (7) is called the Kullback-Leibler divergence (KL) and can be shown to be non-negative and equal to zero if and only if θ is such that $\lambda(\cdot) = \ell(\cdot; \theta)$ for almost all x in its support $\mathcal{X}$. Hence one can think of KL as a measure of "distance" between the true underlying response function, and members of the parametric class $\mathcal{L}(\Theta)$ (although KL is not a metric since it does not satisfy the triangle inequality). As the sample size n grows large, the empirical log-likelihood converges to the expected log-likelihood, and hence one expects that $\hat{\theta}$ will converge to the point θ^* which minimizes the KL (7); This can be shown to hold under some mild technical conditions (see White (1996)). Thus, when using ML estimation, one is effectively using a finite sample estimate of the parameter θ^* that minimizes the KL distance between the parametric class $\mathcal{L}(\Theta)$ and the true underlying response function which generates the data.

Formulation of the performance-based test. The pertinent hypothesis test can be formulated as follows

$$H_0 : \quad \pi^* = \pi(x^*(\theta^*)) \tag{8}$$

$$H_1 : \quad \pi^* > \pi(x^*(\theta^*)), \tag{9}$$

where: π^* represent the *optimal profit rate*; $x^* \in \arg\max\{\pi(x) : x \in \mathcal{X}\}$ is the *optimal price* relative to the true underlying profit function; $x^*(\theta)$ is the maximizer of the profit rate corresponding to the parametric model class $p(x; \theta) = r(x)\ell(x; \theta)$ for $\theta \in \Theta$; and $\pi(x^*(\theta^*))$ is the true profit rate achieved by the latter decision when $\theta = \theta^*$. Thus, what is being compared above is the *best achievable performance* π^*, and the performance achieved by a decision (price) that optimizes the parametric model that "best fits" the data.

This test is quite different in flavor from the traditional statistical one given in (2)-(3). The latter would reject a given parametric model class unless it provides a "good" *global* fit to the true underlying response function. The test described above is *local* in nature: it will reject the null only if the resulting price prescriptions do not fall within the region where the true profit function achieves its maximum. In this manner, we substantially relax the statement of the null H_0 in comparison with the model-based test (2). In particular, whenever the null hypothesis is not rejected in the latter, it will also not be rejected in the performance-based test. However, the new notion of a null hypothesis can hold under a much broader set of scenarios; the example depicted in Figure 1 in Section 1.2 provides such an illustration.

The reader would have obviously noted that the best achievable performance π^*, as well as the best performance achieved by pricing based on the parametric model class $\pi(x^*(\theta^*))$ are not directly computable, as $\lambda(\cdot)$ is not known to the decision-maker. The remaining challenge is therefore to prescribe a procedure for executing the performance-based test so as to meet a required significance level as well as the requirement of consistency. This is spelled out in the next section.

6. The Proposed Approach: The Test Statistic and its Properties

6.1 A non-parametric estimate of the profit function

The first step towards operationalizing the test (8)-(9) is to define a consistent estimator of the true profit function $\pi(x)$; the postulated model class $\mathcal{L}(\Theta)$ need not contain the response function $\lambda(\cdot)$. Letting $Z_i = r(X_i)Y_i$ for $i = 1, ..., n$, the available data can be viewed as noisy observations of the profit rate at n discrete points given by the X_i's. Indeed,

$$Z_i = r(X_i)Y_i = \pi(X_i) + \varepsilon_i, \tag{10}$$

where $\varepsilon_i = r(X_i)Y_i - \pi(X_i)$, and given X_i, ε_i is a random variable with zero mean and variance $\mathbf{E}[\epsilon_i^2|X_i] = (r(X_i))^2\lambda(X_i)(1-\lambda(X_i))$. We let Z denote a generic random variable with the same distribution as $r(X)Y$ and $\sigma^2(\cdot)$ denote the function $x \mapsto (r(x))^2\lambda(x)(1-\lambda(x))$. One of the most straightforward nonparametric estimator is the Nadaraya-Watson estimator:

$$\widehat{\pi}_n(x) \quad := \quad \frac{\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{x-X_i}{h}\Big)Z_i}{\frac{1}{nh}\sum_{j=1}^n K\Big(\frac{x-X_j}{h}\Big)}, \tag{11}$$

where h, the so-called bandwidth, is a positive tuning parameter and the kernel, $K : \mathbb{R} \to \mathbb{R}_+$ is such that $\int K(u)du = 1$ and $K(u) = K(-u)$. The approximation $\widehat{\pi}(x)$ of $\pi(x)$ can be seen to take the form of a weighted sum of the observations Z_i, where the weight of observation i depends on the proximity of X_i to x. One also notes that the denominator in (11)

$$\widehat{g}_n(x) := \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{x-X_i}{h}\Big) \tag{12}$$

approximates the density of X at the point x, $g(x)$.

For concreteness, we will assume throughout that the kernel is Gaussian, i.e., $K(x) = (2\pi)^{-1/2}\exp\{-x^2/2\}$ for $x \in \mathbb{R}$, noting that other choices are possible; cf. Härdle (1990).

Remarks i.) Various nonparametric techniques can be used to approximate the profit function; The Nadaraya-Watson kernel estimator has been widely used and studied in the literature, and hence is a natural candidate. The specific results we present can be derived in a similar manner for many other nonparametric techniques; cf. Härdle (1990) for an overview of alternatives. ii.)

The bandwidth parameter h plays a crucial role in the approximation and its specification requires care. There exist simple prescriptions for specifying the magnitude of the bandwidth based on the sample size. Additional considerations and techniques are discussed in Härdle (1990, Chapter 5). iii.) Nonparametric methods typically become less practical to apply in higher dimensions due to significant data requirements. This fact limits the applicability of our test to higher dimensional problems.

6.2 The test statistic

Note that $\widehat{\pi}_n(\cdot)$ is continuous and let $\hat{x}_n \in \arg\max\{\widehat{\pi}_n(x) : x \in [\underline{x}, \overline{x}]\}$ where $\widehat{\pi}_n(\cdot)$ is defined in (11). Recall that $x^*(\theta) \in \arg\max\{p(x;\theta) : x \in [\underline{x}, \overline{x}]\}$ and that $\hat{\theta}$ denotes the maximum likelihood estimator of the parameter vector θ based on the same observations used to estimate $\widehat{\pi}_n(\cdot)$, $\{(X_i, Y_i) : i = 1, ..., n\}$. The performance-based test statistic Δ_n is then defined as

$$\Delta_n \quad = \quad \widehat{\pi}_n(\hat{x}_n) - \widehat{\pi}_n(x^*(\hat{\theta})). \tag{13}$$

The motivation behind this construction is as follows. As the sample size grows large, the approximation $\widehat{\pi}_n(\cdot)$ should eventually provide a good approximation for the true profit function $\pi(\cdot)$, and hence we anticipate that $\widehat{\pi}_n(\hat{x}_n) \approx \pi(x^*)$. Similarly, $\hat{\theta}$ should be close to θ^* and this should imply that $\widehat{\pi}_n(x^*(\hat{\theta})) \approx \pi(x^*(\theta^*))$. As a result Δ_n can be viewed as a noisy version of the difference between the two terms in the null hypothesis (8), i.e., $\Delta_n \approx \pi(x^*) - \pi(x^*(\theta^*))$. Based on this, we anticipate that Δ_n, properly scaled, would converge in distribution to some random variable under H_0, while it would diverge under the alternative hypothesis. The intuition above and the conditions under which it is valid will be formalized in Theorem 2. We first impose the following technical assumptions.

Assumption 3 (Interior maximum) i.) $\pi(\cdot)$ is twice continuously differentiable $[\underline{x}, \overline{x}]$ with unique maximizer $x^* \in (\underline{x}, \overline{x})$ such that $\pi''(x^*) < 0$.

ii.) $p(\cdot;\cdot)$ is twice continuously differentiable on $[\underline{x}, \overline{x}] \times \Theta$. For all $\theta \in \Theta$, $p(\cdot;\theta)$ has unique maximizer $x^*(\theta) \in (\underline{x}, \overline{x})$ and $\partial^2 p(x^*(\theta);\theta)/\partial x^2 < 0$.

Assumption 4 (Maximum Likelihood) i.) $\mathbf{E}[|\log \mathbf{P}(Y|X)|] < \infty$ and $|\log \mathbf{P}_\theta(y|x)| \leq f_0(x)$ for all $\theta \in \Theta$, where $f_0(\cdot)$ is bounded on $\mathcal{X}$.

ii.) The KL $\mathbf{E}\Big[\log\big(\mathbf{P}(Y|X)/\mathbf{P}_\theta(Y|X)\big)\Big]$ admits a unique minimum $\theta^* \in \Theta$, which is interior.

iii.) $\log \mathbf{P}_\theta(y|x)$ is twice continuously differentiable with respect to θ on Θ.

iv.) There exist functions $f_1(\cdot)$ and $f_2(\cdot)$ bounded on $\mathcal{X}$ such that for all $\theta \in \Theta$, for all $i, j = 1, ..., d$, and for all $(x, y) \in \mathcal{X} \times \{0, 1\}$,

$$\Big|\frac{\partial \log \mathbf{P}_\theta(y|x)}{\partial \theta_i}\Big| \leq f_1(x), \qquad \Big|\frac{\partial^2 \log \mathbf{P}_\theta(y|x)}{\partial \theta_i \partial \theta_j}\Big| \leq f_2(x).$$

v.) For each $\theta \in \Theta$, let A_θ and B_θ be the matrices whose elements are defined as:

$$\begin{aligned} A_\theta(i,j) &:= \mathbf{E}_\theta\Big[\frac{\partial^2 \log \mathbf{P}_\theta(Y|x)}{\partial\theta_i\theta_j}\Big] \qquad i,j = 1,...,d. \\ B_\theta(i,j) &:= \mathbf{E}_\theta\Big[\frac{\partial \log \mathbf{P}_\theta(Y|x)}{\partial\theta_i}\frac{\partial \log \mathbf{P}_\theta(Y|x))}{\partial\theta_j}\Big] \qquad i,j = 1,...,d. \end{aligned}$$

Assume that B_{θ^*} is nonsingular and A_θ has constant rank in some open neighborhood of θ^*.

Assumption 4 is standard in the context of asymptotic analysis of maximum likelihood estimators (cf. White (1982)). Assumption 3 ensures that the optimal decision is an interior point. (The analysis of boundary maximizers constitutes a straightforward extension of the results in section 6.3, and is omitted for space considerations.)

6.3 Large sample theory

Theorem 2 (Consistency and Asymptotic Distribution) *Let Assumptions 1, 3 and 4 hold. Put*

$$\gamma := -\frac{1}{2\pi''(x^*)}\frac{\sigma^2(x^*)}{g(x^*)}\int_{-\infty}^{+\infty}(K'(\psi))^2 d\psi, \tag{14}$$

and let $h_n \downarrow 0$ be a sequence of positive real numbers such that $nh_n^6 \to \infty$ and $nh_n^7 \to 0$, then

$$\begin{aligned} &\textit{i.) Under } H_0\textit{:} \qquad nh_n^3\Delta_n \;\Rightarrow\; \gamma\,\chi^2 \qquad \textit{as } n \to \infty, \\ &\textit{ii.) Under } H_1\textit{:} \qquad nh_n^3\Delta_n \;\Rightarrow\; \infty \qquad \textit{as } n \to \infty, \end{aligned}$$

where χ^2 is a Chi-squared random variable with one degree of freedom.

Remark 1 (value of γ). The value of the constant γ given in (14) plays a crucial role in the proposed test: the higher this value is, the "harder" it will be to reject the model. Consistent with basic intuition, γ is increasing in the variance of the noise associated with observations at the optimal operating point $\sigma^2(x^*)$; higher variance in the noise induces larger confidence bands around any non-parametric estimator of the profit function, and as a result makes it harder to reject any given model. The constant γ is also inversely proportional to $\pi''(x^*)$. To this end, note that the "flatter" the profit function is in the region of the optimum (i.e., the smaller the value of $|\pi''(x^*)|$), the harder it should be to reject a model. This stems from the fact that operating away from the optimum will have lesser ill effects on performance.

Discussion. Given that $nh_n^3\Delta_n$ converges under the null hypothesis to a scaled χ^2 random variable, and diverges to infinity under the alternative, a consistent decision rule can be constructed straightforwardly on the basis of the limiting distribution under H_0. At a significance level α, the decision would be to reject H_0 if $nh_n^3\Delta_n > \tau_\alpha$ where τ_α is the $(1-\alpha)$-quantile of $\gamma\,\chi^2$. Note that the

test is also consistent, as the probability of not rejecting H_0 when H_1 is true $\mathbf{P}(nh_n^3\Delta_n \leq \tau_\alpha|H_1)$ converges to zero as $n \to \infty$ by Theorem 2 ii.). At a more qualitative level, the above test procedure can only distinguish the best performance of the model-based decision from the best achievable performance up to fluctuations of order $(nh_n^3)^{-1}$. In other words, if $\Delta = \pi(x^*) - \pi(x^*(\theta^*)) > 0$ is of this order for a given sample, the test would conclude that the model-based decision gives rise to performance that is statistically indistinguishable from the best achievable performance.

An implementable test. The value of the constant γ depends on characteristics of the *true* profit function, which is not known. Hence to compute p-values based on the asymptotic result in Theorem 2 one would need to approximate the value of γ. Indeed, $g(x^*)$ can be approximated by $\widehat{g}_n(\hat{x}_n)$, where $\widehat{g}_n(x)$ is given in (12). Then $\pi''(x^*)$ can be approximated by $\widehat{\pi}''_n(\hat{x}_n)$, and $\sigma^2(x^*)$ can be approximated by using a kernel approximation to compute $\mathbf{E}[(Z - E[Z|X = \hat{x}_n])^2|X = \hat{x}_n]$ as follows

$$\widehat{\sigma}^2(\hat{x}_n) = \frac{\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{\hat{x}_n - X_i}{h}\Big)\big(Z_i - \widehat{\pi}(\hat{x}_n)\big)^2}{\frac{1}{nh}\sum_{i=1}^n K\Big(\frac{\hat{x}_n - X_i}{h}\Big)}. \tag{15}$$

The "plug-in" estimator

$$\widehat{\gamma}_n := -\frac{1}{\min\{2\widehat{\pi}''_n(\hat{x}_n)\widehat{g}_n(\hat{x}_n); -1/n\}}\widehat{\sigma}^2(\hat{x}_n)\int_{-\infty}^{+\infty}(K'(\psi))^2 d\psi + \frac{1}{n} \tag{16}$$

is then expected to converge to γ in probability (this is made rigorous in the proof of Theorem 3); the $1/n$ correction factors are introduced to ensure that the denominator does not equal zero and that $\widehat{\gamma}_n > 0$. The following corollary provides an implementable version of the performance-based test using this idea

Theorem 3 (Implementable test) *Let Assumptions 1, 3 and 4 hold. Set* $h_n = c_h n^{-1/7}/(\log n)^{1/7}$ *where* c_h *is a positive constant, then*

$$\begin{aligned}
&\textit{i.) Under } H_0\textit{:} \qquad c_h^3 n^{4/7}(\log n)^{-3/7}\widehat{\gamma}_n^{-1}\Delta_n \;\Rightarrow\; \chi^2 \qquad \textit{as } n \to \infty,\\
&\textit{ii.) Under } H_1\textit{:} \qquad c_h^3 n^{4/7}(\log n)^{-3/7}\widehat{\gamma}_n^{-1}\Delta_n \;\Rightarrow\; \infty \qquad \textit{as } n \to \infty.
\end{aligned}$$

It is worth stressing that failure to reject the null in the performance-based framework does not imply that we have identified the "right" model, nor that we have used the "right" decision. Rather, it asserts that based on the available data one cannot *distinguish* (statistically) the performance induced by the best decision based on the parametric model, and the best achievable performance had we known the true underlying model.

An alternative procedure to estimate γ. For finite samples, the approximation resulting from the weak convergence result in Theorem 3 i.) depends critically on the quality of the approximation of γ by $\widehat{\gamma}_n$. Given that this involves estimating the second derivative of the profit function,

we expect the estimate to be fairly noisy. To improve the finite sample performance of the test, we propose a bootstrapping procedure for estimating the constant γ.

Algorithm 1: Bootstrapping estimate of γ

1) For a fixed positive integer b and $j = 1, ..., b$, randomly draw n vectors with replacements from $\{(X_i, Y_i) : 1 \leq i \leq n\}$. Let $\mathcal{D}_n^{(j)} = \{(X_i^{(j)}, Y_i^{(j)}) : 1 \leq i \leq n\}$ denote the resulting draw, and let $\widehat{\pi}_n^{(j)}(\cdot)$ denote the kernel based estimator of the profit function based on the dataset $\mathcal{D}_n^{(j)}$.

2) Let $\hat{x}_n^{(j)} \in \arg\max\{\widehat{\pi}_n^{(j)}(x) : x \in [\underline{x}, \overline{x}]\}$ and

$$\Delta_n^{(j)} = \widehat{\pi}_n^{(j)}(\hat{x}_n^{(j)}) - \widehat{\pi}_n^{(j)}(\hat{x}_n). \tag{17}$$

3) Let

$$\widehat{\gamma}_n^b = \frac{1}{b} \sum_{j=1}^{b} n h_n^3 \Delta_n^{(j)}. \tag{18}$$

4) For the significance level α, let τ_α to be the $(1-\alpha)$ quantile of $\widehat{\gamma}_n^b \chi^2$. Then one rejects the null if and only if $n h_n^3 \Delta_n > \tau_\alpha$.

Based on classical results for bootstrapping (see, e.g., Efron and Tibshirani (1993) and Giné and Zinn (1990)) we expect that, under the conditions of Theorem 3, $\widehat{\gamma}_n^b$ converges to γ under H_0 as n and b grow to ∞ in a suitable sense. While spelling out this limit theory is beyond the scope of this paper, we do compare numerically results obtained using the bootstrapping procedure with those using the estimation procedure (16) in the next section and in Appendix C.

7. Properties of the Proposed Test: Illustrative Numerical Examples

We present below numerical results that illustrate properties of the performance-based test, and contrast them with the model-based test discussed in Section 4.2. For illustrative purposes, we consider two response function models: a logit structure $\ell(x;\theta) = \exp\{\theta_1 - \theta_2 x\}\big(1+\exp\{\theta_1 - \theta_2 x\}\big)^{-1}$ with parameters (θ_1, θ_2); and an exponential structure $\ell(x;\theta) = \theta_1 \exp\{-\theta_2 x\}$ with parameter (θ_1, θ_2). In all cases, the per-customer revenue function is given by $r(x) = (x-1)$, and the X_i's are drawn from a uniform distribution on $[1, 9]$. The bandwidth is taken to be $h_n = c_h n^{-1/7}/(\log n)^{1/7}$, with $c_h > 0$ a tuning constant whose effects are examined below. The procedure we follow is to approximate the distribution of the scaled test statistic on the basis of Theorem 3, as well as based on the bootstrapping procedure that was discussed thereafter. Given the limit distribution and a

significance level α, we can define the rejection region for the null. In particular, for an estimated value of the scaling constant $\widehat{\gamma}$, we take the $(1-\alpha)$-quantile τ_α such that $\mathbf{P}\{\widehat{\gamma}\chi^2 > \tau_\alpha\} = \alpha$. In our experiments, we focus on a standard choice of $\alpha = 0.05$.

In Table 1, we consider a scenario where the sample size is $n = 500$, and is generated according to a logit with parameters $(\theta_1 = 3, \theta_2 = -.9)$. The assumed logit structure is well-specified and hence the null hypothesis will be true in both the model-based formulation (2) as well as the performance-based one (8). We study the impact of the constant c_h affecting the bandwidth and compare the accuracy of the test using the asymptotic distribution in Theorem 3, versus the bootstrap procedure in Algorithm 1, with $b = 250$. In particular, we depict the number of times one rejects the null at the $\alpha = 0.05$ level based on 500 independent replications of the experiment.

c_h	1.5	2	3	4
performance-based test	18.2%	8.4%	4.0%	0.8%
performance-based test (bootstrap)	9.8%	8.2%	4.4%	1.2%

Table 1: **Efficacy of the performance-based test.** Fraction of time H_0 is rejected at the $\alpha = 0.05$ level (based on 500 replications), as a function of the bandwidth parameter. The data-generating model is a logit and the assumed structure is a logit (well-specified case).

It is evident that the choice of bandwidth parameter c_h impacts the behavior of the test statistic and a constant c_h in the range $[2, 3]$ seems appropriate; "rules-of-thumb" for selecting the bandwidth are discussed in Härdle (1990, Chapter 5). We note that the bootstrapping procedure provides more consistent results across bandwidths, and improves the finite sample performance of the test.

In Table 2, we focus again on the fraction of time one rejects the null at the 5% level based on 500 replications. We compare the results provided by the performance-based test to those of the model-based test. In the first (case 1), the true model is a logit with parameters $(\theta_1 = 3, \theta_2 = -.9)$ and the assumed structure is also logit. In other words, the true response function belongs to the postulated family, and H_0 is true in this case from both model-based and performance-based perspectives. The second (case 2) considers again a true model which is a logit with parameters as above but the assumed structure is exponential $\ell(x;\theta) = \theta_1 \exp\{-\theta_2 x\}$. In the last setting (case 3), the true model is a logit with parameters $(\theta_1 = 4.5, \theta_2 = -.9)$ and the assumed structure is exponential. Thus both cases 2 and 3 correspond to misspecified settings, and one anticipates that at least the model-based test would reject the null. The bandwidth for the performance-based test in all cases is taken to be $h_n = 2n^{-1/7}/(\log n)^{1/7}$ and the number of bootsraps used is $b = 250$.

We observe that when the model is well specified (case 1), and hence H_0 is correct for both model- and performance-based tests, the latter rejects H_0 about 8% of the times at the $\alpha = 0.05$ level. Turning to case 2, where the assumed structure is incorrect, we observe that the model-based test rejects the exponential model more than 98% of the times. This is in sharp contrast with the

	case 1		case 2		case 3	
true model	logit: $(3, -0.9)$		logit: $(3, -0.9)$		logit: $(4.5, -0.9)$	
assumed structure	logit: (θ_1, θ_2)		exp: (θ_1, θ_2)		exp: (θ_1, θ_2)	
Data size (n)	5×10^2	10^3	5×10^2	10^3	5×10^2	10^3
model-based test	3.0%	6.2%	98.6%	100%	100%	100%
performance-based test (bootstrap)	8.2%	8.4%	12.8%	14.4%	91.4%	98.8%

Table 2: **Comparison of the model- and performance-based tests.** Fraction of time one rejects H_0 at the $\alpha = 0.05$ level (based on 500 replications). In case 1, the demand model is well-specified while in both cases 2 and 3, it is misspecified.

performance-based test that only rejects the exponential model about 13% of the times. In case 3, where again the assumed structure is misspecified relative to the true response function, both tests reject the null more than 90% of the times.

To better understand the phenomena at play in cases 2 and 3, we present in Figures 2 and 3 the true demand model (logit) and the "best" exponential fit (see discussion in Section 5). We observe that the discrepancy between the two response curves is quite noticeable, exceeding in places 10% in absolute value (Figures 2(a), 3(a)). Analyzing the profit curves corresponding to case 2 in Figure 2(b), we observe that the difference between the optimal performance under the true model, and the performance of the decision dictated by the best exponential fit, differ by an amount Δ which is quite small. Thus the performance-based test indicates that it is difficult to distinguish the difference in performance with dataset sizes of 500 or 1000. In case 3, focusing on the profit curves in Figure 3(b), we see that the difference Δ becomes significant when compared to case 2. This is why the performance-based test now rejects the null more than 90% of the time. Note also that the amount of times one rejects the null increases with the size of the data, illustarting that the difference Δ becomes more significant for larger sample sizes, as one would expect based on the theory contained in Theorems 2 and 3.

Further experiments investigating the approximation of the test statistic by a χ^2 distribution and the impact of the estimation technique used to approximate γ (plug-in versus bootstrapping) are reported in Appendix C.

8. Empirical Example

This section illustrates an application of the proposed performance-based test to sales data obtained from an auto-lender operating in an online direct-to-consumer sales channel. We provide a description of the dataset in Section 8.1, then discuss the setup and profit maximization objective of the auto-lender in Section 8.2 and present our results in Section 8.3.

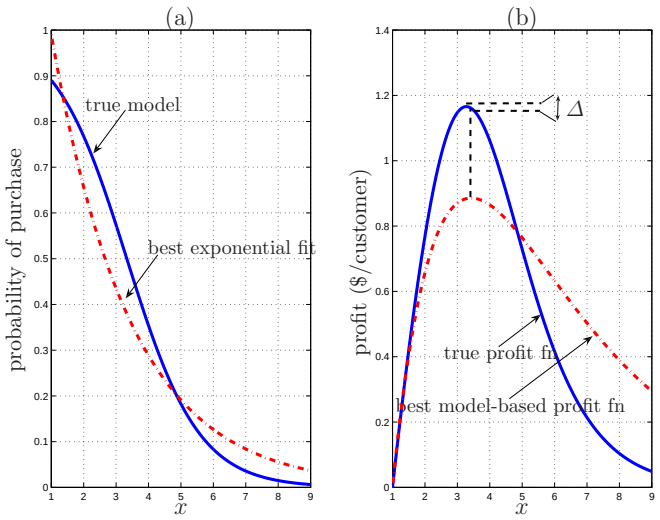

Figure 2: **Model misspecification that is not rejected by the performance-based test.** The true model is a logit with parameters $(\theta_1 = 3, \theta_2 = -0.9)$. Panel (a) gives the true demand model and the best exponential fit; Panel (b) depicts the true profit function and that based on the best exponential fit. Δ indicates the difference between the optimal profits, and those achieved on the basis of the optimal decision of the best exponential fit.

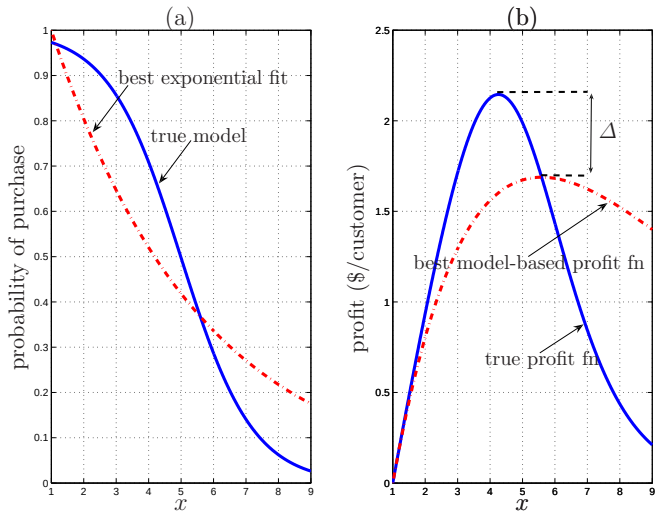

Figure 3: **Model misspecification that is rejected by the performance-based test.** The true model is a logit with parameters $(\theta_1 = 4.5, \theta_2 = -0.9)$. Panel (a) gives the true demand model and the best exponential fit; Panel (b) depicts the true profit function and that based on the best exponential fit. Δ indicates the difference between the optimal profits, and those achieved on the basis of the optimal decision of the best exponential fit.

8.1 Data description

The data is obtained from the following loan process whose key steps are summarized as follows:

1. A customer fills out an online loan request consisting of the amount and term of loan desired, the type of transaction (purchase a new car, purchase a used car, or refinance) as well as a questionnaire with personal data required for credit evaluation.

2. The lender then analyzes the loan application and either directly rejects the application or quotes a rate to the applicant.

3. The customer, upon receiving the offer, has 30 days to decide whether or not to accept the offer.

There were many different alternative sources of automobile loans for customers with good credit scores. Based on discussions with the lender, it is likely that the majority of customers who decline an offered loan will end up borrowing from another lender. However, rates from competitors, as well as the final (lender) choice of customers, in case there is such, are not, in general, publicly available[2] preventing the lender from being able to model the "full" choice of customers (among all lenders). As a result, the approach we take models this as a choice between the lender's own offer and the "outside" world, i.e., we consider a response function with respect to the lender's offer. Note that this accounts indirectly for the competitive environment, and is appropriate as long as the latter is reasonably stable. When prices of competition are observable and not stable, then they could be directly included as explanatory variables in the estimation of price-response. If competitive rates were likely to vary in response to changes in rates offered by this lender, then a game-theoretic analysis might be required. However, the auto-lending market in the United States is very dispersed, with no single lender enjoying a market share greater than 16%. Furthermore, the lender examined in this study had a market share of less than 2% of the market and did not feel that other lenders responded directly to changes in its rates.

The data set contains all instances of incoming customers who were offered a loan during a period ranging from December 2003 to December 2004. For each such customer i, the following information is available:

1) **characteristics of the loan requested:**

   a) date of the request

   b) amount requested (in dollars), denoted $W_{1,i}$.

[2]The only on-line competitive rates available are those advertised by competitors, which are typically the rates offered to the best credit quality customers (roughly, these are customers with FICO scores higher than 740). In this study, we do not consider this customer segment.

c) term requested (in months), denoted $W_{2,i}$. This variable indicates the length of time over which the loan is repaid.

d) loan type, denoted $W_{3,i}$. This variable indicates whether the loan was used for a used car, a new car or to refinance a car.

2) **annual percentage rate**: quoted to the customer, denoted X_i (this is a decision variable determined by the firm).

3) **decision of the customer:**

a) accept/reject decision, denoted $Y_i \in \{0, 1\}$, indicating whether the customer accepted ($Y_i = 1$) or rejected ($Y_i = 0$) the offer.

b) date of acceptance of the offer for customers who accepted the quoted rate.

4) **customer characteristics:**

a) the FICO score, denoted $W_{4,i}$. FICO score is a widely used measure of credit quality. The FICO score for a particular customer is computed through a proprietary algorithm using the customer's credit history and other factors to estimate the probability of default. The score ranges from 300 to 850 with higher scores denoting lower default probabilities.

b) the state in which the customer lives, denoted $W_{5,i}$.

8.2 The profit maximization problem

Let W be the vector summarizing the loan and customer characteristics. The firm would ideally like to have a handle on the response function, $\lambda_W(x)$, i.e., the probability of acceptance as a function of the quoted rate x for a given loan/customer profile. If x_0 denotes the cost of funds for the auto-lender, the profit maximization problem can be approximated by

$$\max_{x \geq x_0} \quad (x - x_0 - \text{risk factor})\lambda_W(x), \tag{19}$$

where the risk factor might also depend on other characteristics of the loan and the customer. The risk factor term was not available and we will not consider it for this illustration. However, it should be apparent that a proper description of the risk factor, if available, can be easily incorporated. Similarly, the exact value of the cost of funds x_0 was not available and in what follows, we will take $x_0 = 2\%$ for illustrative purposes[3].

Rather than solving (19) for every profile W, we will segment the space of profiles W along various dimensions and and maximize the profits over each segment. Note that for each segment, we are in the setup considered in the problem formulation. For the purposes of this study, we will

[3]Experiments with other values yileded similar results.

only focus on loans for used cars and customers with FICO scores in the range 690 to 740; see the discussion in §8.1 and footnote 1 for further explanation about why we do not focus the highest FICO scores. Within that group of customers/loans, we segment customers with two possible FICO score ranges ($(690-715]$ and $(715-740]$) and four possible requested term values which will be referred to as: term 1, term 2, term 3, term 4.

For each segment, the parametric family of models for the acceptance probability is assumed to be the logit class[4]

$$\ell(x;\theta_1,\theta_2) = \frac{\exp\{\theta_1+\theta_2 x\}}{1+\exp\{\theta_1+\theta_2 x\}}, \tag{20}$$

where θ_1 and θ_2 are parameters to be estimated from data.

8.3 Results

Price dispersion. An important prerequisite for the asymptotic theory outlined in previous sections is encoded in Assumption 1 which requires some dispersion in the decision variable X as seen in historical data. While we have not made an attempt to spell out a formal test for the presence of sufficient dispersion, we illustrate below in Figure 4 that in the empirical application under consideration one indeed observes significant variation in the decision variable (offered rate).

The performance of the logit. We present in Table 3 the results obtained when applying the model- and performance-based tests to four of the segments described above over the period of six months.[5] Each segment is analyzed over the first and the second half of the year where data was available. The four other segments either did not have an interior maximum in the region of rates that were experimented with by the lender, or there was insufficient data. For the experiments, 250 bootstrap samples were used for both tests and the bandwidth h for the performance-based test was set to $h_n = 2n^{-1/7}/(\log n)^{1/7}$.

We observe that at the $\alpha = 0.05$ level, one rejects the logit model in four out of eight instances. In contrast, the performance-based test rejects the logit model only in one of these cases. We also observe that the model-based test rejects the logit model in all term 3 segments but not in the term 1 segments. This result is probably driven by the limited number of data points available in the latter segments and does not imply that the logit is an appropriate global fit for those. Interestingly, even in the presence of a high number of data points (term 3 segments), the performance-based test rejects the logit in only one of those four instances.

To summarize, the simple logit model (20) appears to perform well enough for practical purposes of revenue management in seven out of eight instances studied here; this was the case despite the

[4]Experiments were also conducted using the probit class and lead to similar results.

[5]While we did not have access to the re-estimation points that the lender was using, a period of six months between parameter updates appears to be appropriate given the volume of incoming customers in the present case, and is quite in line with industry practices.

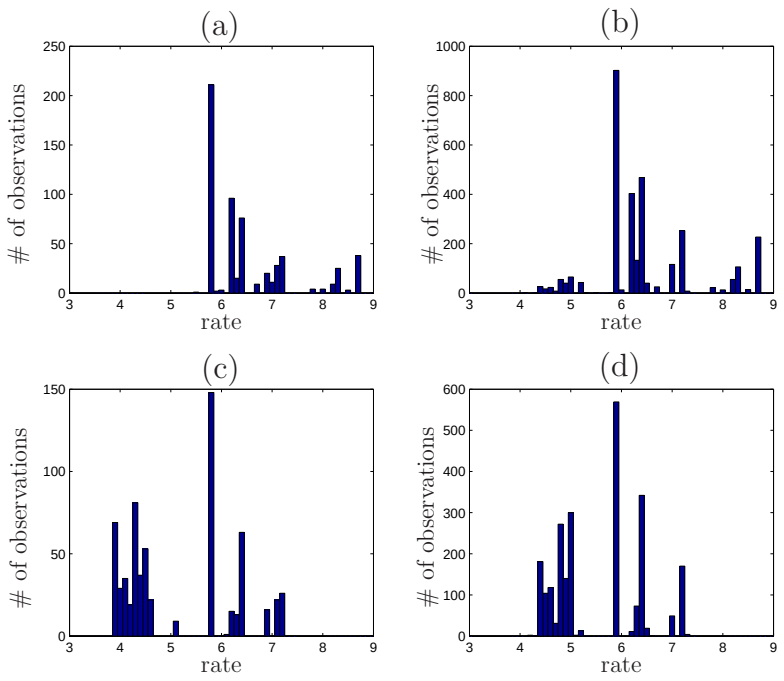

Figure 4: **Price dispersion.** Histograms of rates offered over the course of one year in four different segments. (a) Segment with FICO scores in $(690-715]$ and term value 1; (b) Segment with FICO scores in $(690-715]$ and term value 3; (c) Segment with FICO scores in $(715-740]$ and term value 1; (d) Segment with FICO scores in $(715-740]$ and term value 3.

	time period	term 1 (1st half)	term 1 (2nd half)	term 3 (1st half)	term 3 (2nd half)
FICO range 1 (690-715]					
sample size	n	323	269	1528	1552
parameter estimates	$(\hat{\theta}_1, \hat{\theta}_2)$	$(6.91, -1.35)$	$(2.79, -0.56)$	$(1.92, -0.50)$	$(0.64, -0.30)$
model-based test	p-value	10.8%	47.2%	0.4%	0.0%
performance-based test	p-value	100%	93.5%	1.0%	45.9%
FICO range 2 (715-740]					
sample size	n	352	306	1141	1261
parameter estimates	$(\hat{\theta}_1, \hat{\theta}_2)$	$(0.89, -0.30)$	$(2.71, -0.58)$	$(1.86, -0.51)$	$(1.52, -0.46)$
model-based test	p-value	22.8%	54.4%	0.0%	3.2%
performance-based test	p-value	16.8%	55.7%	66.9%	59.7%

Table 3: **Empirical example.** Comparison of the p-values for the model- and performance-based tests for the Logit model (250 bootstrap samples). A p-value below 5% indicates that the Logit model is rejected.

fact that it does not provide a good global model fit, as evident in the model-based test results.

9. Additional Applications

A performance-based perspective on model testing is pertinent in a wide range of operations management problems. In this section we briefly outline how the approach developed in this paper can be used to assess the validity of models in two additional settings: an inventory problem involving a newsvendor model; and a capacity planning problem involving setting staffing levels in a telephone call-center.

9.1 The newsvendor problem

Consider a "classical" repeated newsvendor problem where the decision maker orders a non-negative quantity X_i in period i and faces demand D_i, where the D_i's are independent and identically distributed with cumulative distribution function $F(\cdot)$. The profit in period i is given by

$$\pi(X_i) = p\mathbf{E}[\min\{D_i, X_i\}] + s\mathbf{E}[(X_i - D_i)^+] - cX_i.$$

Here, $c > 0$ is the unit cost, p is the selling price and s is the salvage value of unsold items, and it is assumed that $p > c > s \geq 0$. Given $F(\cdot)$, it is straightforward to compute the optimal ordering quantity, namely, the familiar fractile solution

$$x^* = F^{-1}\Big(\frac{p-c}{p-s}\Big).$$

Let $\pi^* := \pi(x^*)$ denote the optimal performance with knowledge of the true distribution function.

In practice the distribution $F(\cdot)$ is not known, and one only has access to historical order quantities and associated profits. Let $Y_i = \min\{X_i, D_i\}$, $i \geq 1$. Historical data in this setting would take the form $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$, or equivalently $\mathcal{D}'_n = \{(X_i, Z_i) : 1 \leq i \leq n\}$, where $Z_i = p\min\{D_i, X_i\} + s(X_i - D_i)^+ - cX_i$.

When one postulates a family of models $\mathcal{L}(\Theta) = \{\Phi(\cdot;\theta) : \theta \in \Theta\}$ (where Θ is the space of parameters) for the distribution $F(\cdot)$ and then estimates the corresponding parameters, the central question is to assess the validity of the family $\mathcal{L}(\Theta)$.

In this setting, the model-based approach would attempt to assess the validity of the conditional distribution of Y given an ordering level x. In contrast, the approach developed in Section 5 would focus on the performance of the decision induced by the postulated parametric family. Given the historical data, it is possible, just as in the pricing setting, to construct a non-parametric approximation $\widehat{\pi}(\cdot)$ to the true performance function $\pi(\cdot)$. If $\hat{x}_n$ denotes the maximizer of $\widehat{\pi}(\cdot)$, the test statistic would again be $\Delta_n = \widehat{\pi}_n(\hat{x}_n) - \widehat{\pi}_n(x^*(\hat{\theta}))$, where $\hat{\theta}$ denotes the ML estimator of θ, and $x^*(\hat{\theta})$ denotes the optimal order quantity when facing demand with cumulative distribution

function $\Phi(\cdot;\hat{\theta})$. A similar path to that taken in Section 6 can be followed to analyze the asymptotic behavior of Δ_n, and devise a decision rule to distinguish between the two hypotheses (8) and (9) in the performance-based test.

9.2 A call center staffing problem

Consider a call center with a single class of customers, and one pool of agents (servers). Calls are handled on a first-come first-serve basis and those who are not handled upon arrival are placed in queue. Customers are impatient and hence may abandon from the queue while waiting. The objective of the system manager is to determine the staffing level N to minimize the sum of staffing, delay and abandonment costs in steady-state:

$$\pi(N) \quad = \quad wN + c_d\mathbf{E}[D] + c_a\mathbf{P}(\mathit{Abandon}), \tag{21}$$

where w is the wage rate, N is the number of agents, c_d is a delay penalty and c_a is an abandonment penalty. Here, D denotes the steady-state delay. Let N^* denote the optimal decision and π^* the optimal cost corresponding to an *oracle* that knows the distributions of the inter-arrival times, the service times as well as the times to abandonment.

In practice, the system manager does not have access to these primitive distributions and would attempt to infer them based on historical data $\mathcal{D}_n = \{(t_i, S_i, N_i, \tilde{D}_i, A_i\}) : 1 \leq i \leq n\}$. Here t_i is the arrival time of customer i, S_i the service time, N_i is the number of servers when customer i arrives, $\tilde{D}_i$ is the actual time spent in the system by customer i before being served or abandoning, and A_i is an indicator variable registering if customer i abandoned or not.

In order to devise a staffing plan, an approach often adopted in practice is to postulate an Erlang A model ($M/M/N+M$), and estimate the arrival rate $\hat{\lambda}$, average service time $\hat{\mu}^{-1}$, and average patience $\hat{\beta}^{-1}$, based on the available data. Then, based on the Markovian assumptions underlying the Erlang A model together with these estimates, it is possible to compute $\mathbf{E}[D]$ and $\mathbf{P}(\mathit{Abandon})$ and optimize the objective (21). The question in this context is whether the postulated $M/M/N+M$ model is valid.

A natural model-based approach would be to test if the inter-arrival times, service times and abandonment times are exponential (e.g., using some variant of a Kolmogorov-Smirnov test is the spirit of the one discussed in Section 4). Roughly speaking, the model will then be declared valid if all three distributions cannot be distinguished from exponential. It is apparent that this approach is likely to be overly stringent. In fact, a recent empirical study by Brown et al. (2005) based on call center data, found that service times have a distribution which is "closer" to a lognormal, not in line with the prevalent exponential assumption. Interestingly, Mandelbaum and Zeltyn (2005) observe that the performance predicted by the Erlang A formulae are fairly close to *observed* performance despite the various sources of misspecification. This is consistent with the performance-based approach advocated here.

Referring back to Section 5, a similar test to the one proposed there can be developed to address the problem at hand. First, one would construct a non-parametric estimate of the profit function $\widehat{\pi}_n(\cdot)$ using the historical observations and derive the optimal strategy staffing $\hat{N}_n$. This can be done by estimating the fraction of abandoning customers $\mathbf{P}(abandon)$ and average delay as a function of the staffing level. Second one would obtain the model-based estimate $N^*(\hat{\theta})$ based on the estimated parameters $\hat{\theta} = (\hat{\lambda}, \hat{\mu}, \hat{\beta})$ needed to calibrate the Erlang A model to the data $\mathcal{D}_n$. Then one should compare the *performance* of the decisions induced by the former and the latter. In other words, compute again $\Delta_n = \widehat{\pi}_n(\hat{N}_n) - \widehat{\pi}_n(N^*(\hat{\theta}))$. The final task is to characterize the limiting distribution of Δ_n, properly centered and scaled, following a similar template to the one followed in Theorems 2 and 3.

References

Abe, M. (1995), 'A nonparametric density estimation method for brand choice using scanner data', *Marketing Science* **14**, 300–325.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE Transactions on Automatic Control* **19**, 716–723.

Amemiya, T. (1981), 'Qualitative response models: A survey', *Journal of Economic Literature* **XIX**, 1483–1536.

Andrews, D. W. (1997), 'A conditional Kolmogorov test', *Econometrica* **65**, 1097–1128.

Bartels, K., Boztug, Y. and Uller, M. M. (1999), 'Testing the multinomial logit model', *working paper, University Potsdam, Germany* .

Ben-Akiva, M. and Lerman, S. R. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press.

Berger, J. O. (1980), *Statistical Decision Theory: Foundations, Concepts, and Methods*, Springer-Verlag.

Bierens, H. J. (1990), 'A consistent conditional moment test of functional form', *Econometrica* **58**, 1443–1458.

Borovkov, A. (1998), *Mathematical Statistics*, Gordon and Breach Science Publishers.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Zeltyn, S., Zhao, L. and Shen, H. (2005), 'Statistical analysis of a telephone call center: A queueing-science perspective', *Journal of the American Statistical Association* **100**, 36–50.

Cachon, G. and Kok, A. (2007), 'How to (and how not to) estimate the salvage value in the newsvendor model', *Manufacturing & Service Operations Management* **9**, 276–290.

Cooper, W. L., Homem-de-Mello, T. and Kleywegt, A. J. (2006), 'Models of the spiral-down effect in revenue management', *Operations Research* **54**, 968–987.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction To The Bootstrap*, Chapman & Hall.

Ernst, R. and Cohen, M. A. (1990), 'Operations related groups (orgs): A clustering procedure for production/inventory systems', *Journal of Operations Management* **9**, 574–598.

Giné, E. and Zinn, J. (1990), 'Bootstrapping general empirical measures', *Annals of Applied Probability* **18**, 851–869.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press.

Härdle, W. and Mammen, E. (1994), 'Comparing nonparametric versus parametric regression fits', *Annals of Statistics* **21**, 1926–1947.

Horowitz, J. L. (1993), 'Semiparametric estimation of a work-trip mode choice model', *Journal of Econometrics* **58**, 49–70.

Leeflang, P. S., Wittink, D. R., Wedel, M. and Naert, P. A. (2000), *Building Models for Marketing Decisions*, Kluwer Academic Publishers.

Liyanage, L. H. and Shantikumar, J. G. (2005), 'A practical inventory control policy using operational statistics', *Operations Research Letters* **33**, 341–348.

Mandelbaum, A. and Zeltyn, S. (2005), 'Service engineering in action: The palm/erlang-a queue, with applications to call centers', *Technion Technical Report* .

Nickerson, R. C. and Boyd, D. W. (1980), 'The use and value of models in decision analysis', *Operations Research* **28**, 139–155.

Pagan, A. and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press.

Phillips, R. (2005), *Pricing and Revenue Optimization*, Stanford University Press.

Talluri, K. T. and van Ryzin, G. J. (2005), *Theory and Practice of Revenue Management*, Springer-Verlag.

Train, K. (2002), *Discrete Choice Methods with Simulation*, Cambridge University Press.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica* **50**, 1–25.

White, H. (1996), *Estimation, Inference and Specification Analysis*, Cambridge University Press.

Ziegler, K. (2002), 'On nonparametric kernel estimation of the mode of the regression function in the random design model', *Nonparametric Statistics* **14**, 749–774.

A. Algorithmic Description of the Proposed Testing Approach

Consider a family of demand models $\{\ell(\cdot;\theta), \theta \in \Theta\}$, a price domain $[\underline{x}, \overline{x}]$ and a set of observations $\{X_i, Y_i : i = 1, ..., n\}$.

Assessing the validity of a demand model

Step 1. Initialization

- Set the significance level that will be used for the test: $\alpha \in (0, 1)$. Let τ_α to be the $(1-\alpha)$ quantile of χ^2 (Chi-squared distribution with one degree of freedom).
- Set the bandwidth: $h = c_h n^{-1/7}/(\log n)^{1/7}$. (Some preliminary analysis may be needed to select c_h.[6])
- Set the number of bootstrap samples b to be used for the estimation of γ.
- Define a uniform grid of N points $v_1, v_2, ..., v_N$ on the price domain $[\underline{x}, \overline{x}]$.

Step 2. Parameter Estimation and Optimization

Compute the Maximum Likelihood estimate $\hat{\theta}$ based on the observations:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \big[Y_i \log \ell(X_i;\theta) + (1 - Y_i) \log(1 - \ell(X_i;\theta)) \big]$$

Optimize the profit rate based on the assumed model:

$$x^*(\hat{\theta}) = \arg\max_{x \in [\underline{x}, \overline{x}]} r(x) \ell(x; \hat{\theta})$$

Step 3. Computing and Optimizing a non-parametric approximation to the profit function

---

For $j = 1, ..., N$, compute

$$\widehat{\pi}_n(v_j) \quad := \quad \frac{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{v_j - X_i}{h}\right) Z_i}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{v_j - X_i}{h}\right)},$$

End

Set

$$\hat{x}_n = \arg\max\{\widehat{\pi}_n(x) : x = v_1, ..., v_N\}$$

[6] see, e.g., Härdle (1990, Chapter 5)

Step 4. Computing the test statistic

$$\Delta_n \quad = \quad \widehat{\pi}_n(\hat{x}_n) - \widehat{\pi}_n(x^*(\hat{\theta})).$$

Step 5. Estimating γ:

---

1) For $j = 1, ..., b$, randomly draw n vectors with replacements from $\{(X_i, Y_i) : 1 \leq i \leq n\}$. Let $\mathcal{D}_n^{(j)} = \{(X_i^{(j)}, Y_i^{(j)}) : 1 \leq i \leq n\}$ denote the resulting draw, and let $\widehat{\pi}_n^{(j)}(\cdot)$ denote the kernel based estimator of the profit function based on the dataset $\mathcal{D}_n^{(j)}$.
2) Let $\hat{x}_n^{(j)} \in \arg\max\{\widehat{\pi}_n^{(j)}(x) : x = v_1, ..., v_N\}$ and

$$\Delta_n^{(j)} = \widehat{\pi}_n^{(j)}(\hat{x}_n^{(j)}) - \widehat{\pi}_n^{(j)}(\hat{x}_n).$$

3) Compute the bootstrap-based estimate:

$$\widehat{\gamma}_n^b = \frac{1}{b} \sum_{j=1}^{b} n h_n^3 \Delta_n^{(j)}.$$

Step 6. Assessing the validity of the demand model

If $nh_n^3 \Delta_n / \widehat{\gamma}_n^b > \tau_\alpha$

Reject the null hypothesis H_0
[the postulated model is not valid]

Else

Do not reject the null hypothesis H_0
[the postulated model cannot be rejected]

End

B. Proof of the Main Result

Proof of Theorem 2. The proof is organized as follows. We first decompose Δ_n into two terms, A_n and B_n, separating the misspecification source of error. We then analyze the asymptotic behavior of each of those terms under both H_0 and H_1 in Lemmas 1 and 3, respectively.

We decompose Δ_n as follows

$$\Delta_n \quad = \quad A_n + B_n,$$

where

$$\begin{aligned} A_n &= \widehat{\pi}_n(\hat{x}_n) - \widehat{\pi}_n(x^*) \\ B_n &= \widehat{\pi}_n(x^*) - \widehat{\pi}_n(x^*(\hat{\theta})) \end{aligned}$$

Next, we analyze each term A_n and B_n separately under the null and the alternative hypotheses.

Lemma 1 *Let Assumptions 1, 3 and 4 hold. Suppose that $nh_n^6 \to \infty$ and $nh_n^7 \to 0$, then*

$$nh_n^3 A_n \Rightarrow \gamma\chi^2,$$

where γ was defined in (14) and χ^2 is a Chi-squared random variable with one degree of freedom.

Proof of Lemma 1. Noting that $\widehat{\pi}(\cdot)$ is differentiable, a Taylor expansion gives that for some $x_{1,n} \in [\min\{\hat{x}_n, x^*\}, \max\{\hat{x}_n, x^*\}]$

$$\begin{aligned} A_n &= -\widehat{\pi}_n'(\hat{x}_n)(x^* - \hat{x}_n) - \frac{1}{2}\widehat{\pi}_n''(x_{1,n})(x^* - \hat{x}_n)^2 \\ &= -\frac{1}{2}\widehat{\pi}_n''(x_{1,n})(x^* - \hat{x}_n)^2 \end{aligned} \tag{B-1}$$

Now, by Ziegler (2002, Theorem 3.1), we have under the assumption that $nh_n^6 \to +\infty$ and that $nh_n^7 \to 0$

$$\sqrt{nh_n^3}(x^* - \hat{x}_n) \Rightarrow \mathcal{N}\Big(0, \frac{\sigma^2(x^*)}{(\pi''(x^*))^2 g(x^*)} \int (K'(\psi))^2 d\psi\Big), \tag{B-2}$$

where $\mathcal{N}(0, s^2)$ denotes a centered normal distribution with variance s^2.

Let $\{x_n\}$ be any sequence of reals in $\mathcal{X}$. We next establish that $\widehat{\pi}_n''(x_n)$ converges to $\pi''(x)$ whenever $|x_n - x|$ converges to zero in probability.

$$\begin{aligned} |\widehat{\pi}_n''(x_n) - \pi''(x)| &\leq |\widehat{\pi}_n''(x_n) - \pi''(x_n)| + |\pi''(x_n) - \pi''(x)| \\ &\leq \sup_{\underline{x} \leq r \leq \overline{x}} |\widehat{\pi}_n''(r) - \pi''(r)| + |\pi''(x_n) - \pi''(x)| \end{aligned}$$

The second term on the right-hand-side converges to zero in probability by continuity of $\pi''(\cdot)$. The first term on the right-hand-side also converges to zero under the current assumption that $nh_n^6 \to \infty$ (see Ziegler (2002, Theorem 1.5)).

Now note that the sequence $x_{1,n}$ converges in probability to x^* (since $\hat{x}_n$ converges in probability to x^* by (B-2) and $|x_{1,n} - x^*| \leq |\hat{x}_n - x^*|$) . Applying the result above, we have that $\widehat{\pi}_n''(x_{1,n})$ converges to $\pi''(x^*)$ in probability. Noting that (B-1) implies that

$$nh_n^3 A_n = -\frac{1}{2}\widehat{\pi}_n''(x_{1,n})\Big[\sqrt{nh_n^3}(\hat{x}_n - x^*)\Big]^2,$$

Slutsky's theorem and the continuous mapping theorem, in conjunction with (B-2), imply that

$$nh_n^3 A_n \Rightarrow \Big(-\frac{1}{2\pi''(x^*)}\frac{\sigma^2(x^*)}{g(x^*)} \int (K'(\psi))^2 d\psi\Big)\chi_1^2$$

■

We now turn to analyze the second contribution to Δ_n, B_n. We start with a result on the estimate $\hat{\theta}$.

Lemma 2 *Let Assumptions 1, 3 and 4 hold. Then*

$$\sqrt{n}(\hat{\theta} - \theta^*) \Rightarrow \mathcal{N}(0, \Sigma^2_{\theta^*}) \qquad as\ n \to \infty, \tag{B-3}$$

for some positive definite matrix $\Sigma^2_{\theta^*}$.

Proof of Lemma 2. Note that the conditions spelled out in Assumption 4 ensure asymptotic normality of the ML estimator by White (1982, Theorem 3.2). ■

We now characterize the asymptotic behavior of B_n under the null and alternative hypotheses.

Lemma 3 *Let Assumptions 1, 3 and 4 hold. Suppose that* $nh_n^6 \to \infty$ *and* $nh_n^7 \to 0$, *then*

$$\begin{aligned} &i.)\ Under\ H_0 \qquad nh_n^3 B_n \ \Rightarrow\ 0 \qquad as\ n \to \infty. \\ &ii.)\ Under\ H_1 \qquad nh_n^3 B_n \ \Rightarrow\ \infty \qquad as\ n \to \infty. \end{aligned}$$

Proof of Lemma 3. To prove the result, we decompose B_n into two terms. We have $B_n = C_n + D_n$, where

$$\begin{aligned} C_n &= \widehat{\pi}_n(x^*) - \widehat{\pi}_n(x^*(\theta^*)) \\ D_n &= \widehat{\pi}_n(x^*(\theta^*)) - \widehat{\pi}_n(x^*(\hat{\theta})) \end{aligned}$$

Analysis under H_0*:* By a Taylor expansion, we have

$$-D_n = \widehat{\pi}_n'(x^*(\theta^*))\big(x^*(\hat{\theta}) - x^*(\theta^*)\big) + \frac{1}{2}\widehat{\pi}_n''(x_{1,n})\big(x^*(\theta^*) - x^*(\hat{\theta})\big)^2, \tag{B-4}$$

for some $x_{1,n} \in [\min\{x^*(\theta^*), x^*(\hat{\theta})\}, \max\{x^*(\theta^*), x^*(\hat{\theta})\}]$.

We next establish that $x^*(\cdot)$ is Lipschitz continuous. Let $C_2 = \max\{|\partial^2 p(x^*(\theta);\theta)/\partial\theta_i \partial x| : i = 1, .., d,\ \theta \in \Theta\}$ and $C_3 = \min\{|\partial p(x^*(\theta);\theta)/\partial^2 x| : i = 1, .., d,\ \theta \in \Theta\}$. Note that C_2 and C_3 are well defined as the maximum and minimum of continuous functions over compact sets. In addition, note that by Assumption 3 ii.) that $C_3 > 0$. The fact that $x^*(\theta)$ is an interior maximizer implies that it satisfies

$$\frac{\partial p(x^*(\theta);\theta)}{\partial x} = 0.$$

In addition, the fact that $\frac{\partial^2 p(x^*(\theta);\theta)}{\partial^2 x} < 0$ (Assumption 3 ii.)) implies that the equation above has a unique solution in the neighborhood of $x^*(\theta)$. Applying the implicit function theorem yields that $x^*(\theta)$ is differentiable and

$$\frac{\partial x^*(\theta)}{\partial \theta_i} = -\frac{\partial^2 p(x^*(\theta);\theta)/\partial\theta_i \partial x}{\partial^2 p(x^*(\theta);\theta)/\partial^2 x}.$$

We then have that $\left|\partial x^*(\theta)/\partial\theta_i\right| \leq C_2/C_3$ and $x^*(\theta)$ is Lipschitz continuous with constant C_2/C_3.

Under H_0, $x^*(\theta^*) = x^*$ and hence $\pi'(x^*(\theta^*)) = 0$. Coming back to (B-4), it follows that

$$\begin{aligned} nh_n^3 D_n &= -nh_n^3 \widehat{\pi}_n'(x^*(\theta^*))\big(x^*(\hat{\theta}) - x^*(\theta^*)\big) - nh_n^3 \frac{1}{2}\widehat{\pi}_n''(x_{1,n})\big(x^*(\theta^*) - x^*(\hat{\theta})\big)^2 \\ &= -h_n^{1/2}\sqrt{nh_n^3}\Big(\widehat{\pi}_n'(x^*(\theta^*)) - \pi'(x^*(\theta^*))\Big)\, h_n\sqrt{n}\big(x^*(\hat{\theta}) - x^*(\theta^*)\big) \\ &\quad -\frac{1}{2}\widehat{\pi}_n''(x_{1,n}) nh_n^3\big(x^*(\theta^*) - x^*(\hat{\theta})\big)^2. \end{aligned} \tag{B-5}$$

Focusing on the second term on the right-hand-side above, one can establish as in Lemma 1 that $\widehat{\pi}_n''(x_{1,n})$ converges to $\widehat{\pi}_n''(x^*(\theta^*))$. We also have that $nh_n^3\big(x^*(\theta^*) - x^*(\hat{\theta})\big)^2 \leq h_n^3(C_2/C_3)^2[\sqrt{n}(\hat{\theta} - \theta^*)]^2$ and the right hand side converges to zero in probability by (B-3) and the fact that $h_n \to 0$.

Turning to the first term on the right-and-side in (B-5), we have that $h_n^{1/2}\sqrt{nh_n^3}(\widehat{\pi}_n'(x^*(\theta^*)) - \pi'(x^*(\theta^*)))$ converges to zero under the assumption that $nh_n^7 \to 0$ (see Pagan and Ullah (1999, Theorem 4.3)). On another hand, $h_n\sqrt{n}\big(x^*(\theta^*) - x^*(\hat{\theta})\big) \leq h_n(C_2/C_3)\sqrt{n}(\hat{\theta} - \theta^*)$ and again the right hand side converges to zero in probability by (B-3) and the fact that $h_n \to 0$. We deduce that

$$nh_n^3 D_n \Rightarrow 0 \tag{B-6}$$

Under H_0, $x^*(\theta^*) = x^*$, which implies that $C_n = 0$ and hence i.) follows.

Analysis under H_1: We now analyze D_n and C_n under H_1. Under H_1, it is clear that D_n converges to zero in probability from (B-4), the continuity of $x^*(\theta)$ and (B-3). On another hand, we have

$$C_n = \widehat{\pi}_n(x^*) - \widehat{\pi}_n(x^*(\theta^*)) \quad = \quad \widehat{\pi}_n(x^*) - \pi(x^*) + \pi(x^*) - \pi(x^*(\theta^*)) + \pi(x^*(\theta^*)) - \widehat{\pi}_n(x^*(\theta^*)),$$

Both terms $\widehat{\pi}_n(x^*) - \pi(x^*)$ and $\pi(x^*(\theta^*)) + \pi(x^*(\theta^*)) - \widehat{\pi}_n(x^*(\theta^*))$ converge to zero by the consistency of the non-parametric estimator. We deduce that C_n converges to $\pi(x^*) - \pi(x^*(\theta^*)) > 0$ in probability. Hence B_n converges to $\pi(x^*) - \pi(x^*(\theta^*))$ and $nh_n^3 B_n \Rightarrow \infty$ since $nh_n^3 \to \infty$. ii) is now established and the proof of Lemma 3 is complete. ■

Combining the results of Lemmas 1 and 3, the result of the theorem follows by an application of Slutsky's theorem. This completes the proof. ■

Proof of Theorem 3. If one establishes that $\widehat{\gamma}_n$ converges to γ in probability, then result will follow from Theorem 2 and an application of Slutsky's theorem. Next, we show that $\widehat{\gamma}_n$ converges to γ in probability by analyzing separately every term in the definition of $\widehat{\gamma}_n$ provided in (16).

Recalling (B-2) and the argument that followed in the proof of Lemma 1, it has already been established that $\hat{x}_n$ converges in probability to x^* and that $\widehat{\pi}_n''(\hat{x}_n)$ converges in probability to $\pi''(x^*)$.

We now proceed with a similar argument to establish that $\widehat{g}_n(\hat{x}_n)$ converges in probability to

$g(x^*)$.

$$\begin{aligned} |\widehat{g}_n(\hat{x}_n) - g(x^*)| &\leq |\widehat{g}_n(\hat{x}_n) - g(\hat{x}_n)| + |g(\hat{x}_n) - g(x^*)| \\ &\leq \sup_{\underline{x} \leq r \leq \overline{x}} |\widehat{g}_n(r) - g(r)| + |g(\hat{x}_n) - g(x^*)| \end{aligned}$$

The second term on the right-hand-side above converges to zero in probability by continuity of $g(\cdot)$. The first term on the right-hand-side above also converges to zero under the current assumption that $nh_n^2 \to \infty$ (see Pagan and Ullah (1999, Theorem 2.8)). We conclude that $\widehat{g}_n(\hat{x}_n)$ converges in probability to $g(x^*)$.

We are left with the analysis of $\widehat{\sigma}^2(\hat{x}_n)$, which was defined in (15). Note that $\sigma(\cdot)$ is continuous on $[\underline{x}, \overline{x}]$. A similar argument as the one just developed for $g_n(\hat{x}_n)$ in conjunction with Pagan and Ullah (1999, Theorem 3.4) yields that $\widehat{\sigma}^2(\hat{x}_n)$ converges in probability to $\sigma^2(x^*)$.

Now, an application of Slutsky's theorem yields that

$$\widehat{\gamma}_n \to \gamma \qquad \text{in probability as } n \to \infty.$$

This concludes the proof. ■

C. Additional Numerical Results

This appendix complements section 7. In particular, we investigate the quality of the approximation of $(nh_n^3/\widehat{\gamma})\Delta_n$ by a χ^2 distribution under the null hypothesis, as well as the impact of the estimation technique used to approximate the constant γ. The setting considered is following that of Section 7 and we take throughout $h_n = 2n^{-1/7}/(\log n)^{1/7}$. In the experiment, we compute the empirical distributions of $(nh_n^3/\widehat{\gamma})\Delta_n$ and of the p-value associated with the test by replicating the sampling and estimation procedures 500 times.

In Figures 5 and 6, we analyze the quantiles of the empirical distribution $(nh_n^3/\widehat{\gamma})\Delta_n$ against the quantiles of a χ^2 distribution. Figure 5 corresponds to a logit model with $(\theta_1 = 3, \theta_2 = -.9)$ while Figure 6 corresponds to a logit model with $(\theta_1 = 4.5, \theta_2 = -.9)$. The top plot in each figure is associated with the basic "plug-in" estimation procedure provided in (16) while in the bottom plot, γ is estimated using the bootstrapping procedure given in (18).

We observe that when one estimates γ through (16) (Figure 5(a) and Figure 6(a)), the Q-Q plots are not fully aligned with the 45-degree line, which might be due to potential slow convergence associated with the plug-in estimates used in the estimation of γ. We note that the Q-Q plots are much closer to the 45-degree line when using the bootstrapping procedure for estimating γ (16) (Figure 5(b) and Figure 6(b)). This illustrates that the bootstrapping procedure is in general more reliable, in line with what was observed in Table 1.

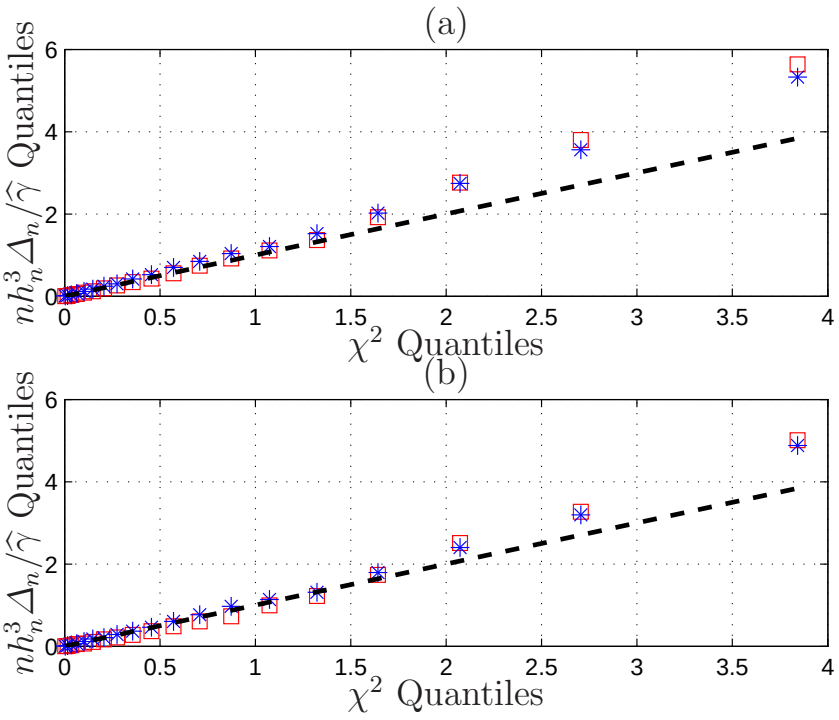

Figure 5: **Q-Q plots and quality of the approximation.** 20-Quantiles of the empricial distribution of $nh_n^3\Delta_n/\widehat{\gamma}$ against 20-Quantiles of the limiting distribution (χ^2). Squares ($\square$) correspond to a data size $n = 500$ and stars ($*$) correspond to a data size $n = 10^3$. Data is generated according to a logit demand model with parameters $(3, -0.9)$ and the fitted model is logit. In (a) the constant γ is estimated directly through (16); and in (b) the constant γ is estimated via bootstrapping through (18) (with 250 bootstraps).

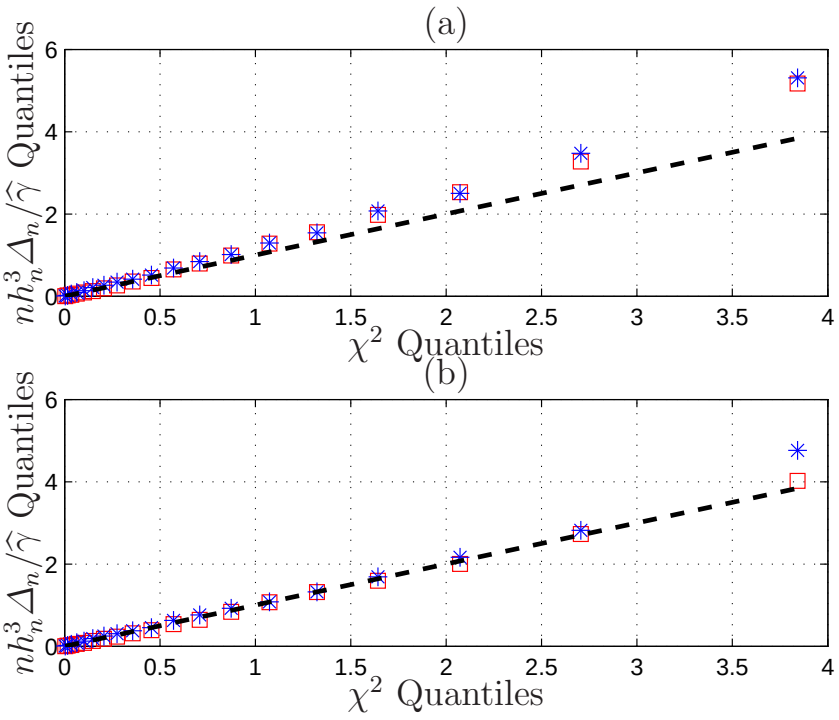

Figure 6: **Q-Q plots and quality of the approximation.** 20-Quantiles of the empricial distribution of $nh_n^3\Delta_n/\widehat{\gamma}$ against 20-Quantiles of the limiting distribution (χ^2). Squares ($\square$) correspond to a data size $n = 500$ and stars ($*$) correspond to a data size $n = 10^3$. Data is generated according to a logit demand model with parameters $(4.5, -0.9)$ and the fitted model is logit. In (a) the constant γ is estimated directly through (16); and in (b) the constant γ is estimated via bootstrapping through (18) (with 250 bootstraps).