

Pricing from Observational Data

Dimitris Bertsimas

Massachusetts Institute of Technology, dbertsim@mit.edu, <http://web.mit.edu/dbertsim>

Nathan Kallus

Cornell Tech and Cornell University, kallus@cornell.edu, <http://www.nathankallus.com>

Given observational data on price and demand, the price optimization problem is sometimes addressed in the literature by a predictive approach: (a) fit a model to the data that best predicts demand given price and (b) substitute the predictive model into the overall profit and optimize for price. We show that, because historical demand at all prices but the observed one is missing, the price optimization problem is not well specified by the data, and in particular, the predictive approach fails to find the optimal price. We bound the suboptimality of the predictive approach, even when the optimal price cannot be identified from the data, by leveraging the special structure of the problem. Drawing from the causal inference literature, we provide sufficient conditions for the optimal price to be identifiable from the data. Given these conditions, we provide parametric and non-parametric algorithms for the price optimization problem. In the non-parametric case we prove consistency and asymptotic normality and establish rates of convergence. We develop a hypothesis test for asymptotic profit optimality of any algorithm for pricing from observational data. We use this test to demonstrate empirically in an auto loan dataset that both parametric and non-parametric predictive approaches lose significant profit relative to the optimum and that our prescriptive parametric framework leads to profit that cannot be distinguished from the optimal one, recovering 36-70% of profits lost by the predictive approaches.

Key words: Pricing, Revenue Management, Data-Driven Decision Making, Predictive Analytics, Prescriptive Analytics, Causal Inference

1. Introduction

Pricing is one of the most fundamental instruments for revenue management. Effective pricing in monopolistic settings hinges on the manager’s understanding of consumers’ response to price changes (Phillips 2005). In many pricing applications, this response, sometimes known as “demand elasticity,” is estimated from observations of past sale attempts. This can be done through repeated experiments (as in Bertsimas and Perakis (2006), Besbes and Zeevi (2009), Harrison et al. (2012)), but in most real-world applications this is done based on analytics of a corpus of observational data (examples include Besbes et al. (2010), Cohen et al. (2014), Ferreira et al. (2016)).

For the purpose of illustration, consider a simple example. Table 1 displays the unit demand D_i observed at the MIT Coop on each day $i = 1, \dots, n$ for the classic MIT hoodie and the price P_i at which it was offered. Only two prices, \$20 and \$28, have been observed and each was observed n_{20}

Table 1 Example Demand Data for the Classic MIT Hoodie

Day (i)	Price (P_i)	Observed Demand (D_i)	Unobserved Demand (D'_i)
1	20 (\$)	1 (units)	0
2	28	0	1
3	28	1	1
4	20	2	0
\vdots	\vdots	\vdots	\vdots
$n-1$	20	1	0
n	28	0	1

and n_{28} times, respectively. For each day, there are the demands $D_i(20)$ and $D_i(28)$ that would have been observed if the price were set to \$20 or \$28, respectively – these values represent the unseen *demand curve* associated with that day. We only observe $D_i = D_i(P)$. We do not observe $D'_i = D_i(48 - P)$. For example, on day 1, $D_1 = D_1(20) = 1$ and $D'_1 = D_1(28) = 0$. Using the observed data only, we can compute

$$\begin{aligned}\tilde{d}_n(20) &= \frac{1}{n_{20}} (D_1 + D_4 + \dots + D_{n-1}) = \text{Average}(\{D_i : i = 1, \dots, n, P_i = 20\}), \\ \tilde{d}_n(28) &= \frac{1}{n_{28}} (D_2 + D_3 + \dots + D_n) = \text{Average}(\{D_i : i = 1, \dots, n, P_i = 28\}).\end{aligned}$$

If the rows of Table 1 constitute independent and identically distributed (iid) data and $P, D, D(p)$ represent a generic random instance, then $\tilde{d}_n(20)$ and $\tilde{d}_n(28)$ are our best guesses for the value of demand D in a new random instance where $P = 20$ or $P = 28$, respectively. In particular, $\tilde{d}_n(p) \rightarrow \tilde{d}(p) := \mathbb{E}[D | P = p]$ almost surely as $n \rightarrow \infty$, and $\tilde{d}(p)$ is always the *best* predictor of demand given price (in squared error). This leads to a price optimization problem as follows

$$\tilde{p} \in \arg \max_{p \in \mathcal{P}} \left\{ \tilde{R}(p) := \mathbb{E}[r(p)D | P = p] = r(p)\tilde{d}(p) \right\}, \quad (1)$$

where \mathcal{P} is the feasible set of prices (here $\mathcal{P} = \{20, 28\}$) and $r(p) = p - c$ is the per-unit profit at price p with procurement cost $c \geq 0$. When we substitute any estimate $\tilde{d}_n(p)$ of $\tilde{d}(p)$ (or any estimate $\tilde{R}_n(p)$ of $\tilde{R}(p)$) into (1), we call the result a predictive approach to pricing from data because it hinges on fitting a predictive model of demand (or profit) given price to the data.

On average over a new random instance, $d(p) := \mathbb{E}[D(p)]$ is the expected demand if price were set to p – we call this the *price response function* (PRF). Therefore, the problem of selecting a price so to optimize expected profits for a new random instance is

$$p^* \in \arg \max_{p \in \mathcal{P}} \{R(p) := \mathbb{E}[r(p)D(p)] = r(p)d(p)\}. \quad (2)$$

In general, $d(p) \neq \tilde{d}(p)$. In particular, if the population distribution of the data is exactly the discrete distribution with weight $1/6$ on each of the six displayed rows of Table 1 ($n = 6$), then

$$\tilde{d}(20) = 4/3 \neq 7/6 = d(20), \quad \tilde{d}(28) = 1/3 \neq 1/6 = d(28).$$

This highlights that, in general, Problem (1) is different from Problem (2) for observational data.

Moreover, the correct estimate of $d(p)$ involves *unobserved* data:

$$d_n(p) = \frac{1}{n} (D_1(p) + D_2(p) + \dots + D_n(p)) = \text{Average}(\{D_i(p) : i = 1, \dots, n\}),$$

This highlights that, in general, Problem (2) is not necessarily well-specified by the observed data. In particular, if the data is distributed in the population like the six displayed rows, we can imagine filling in the unseen values in Table 1 with anything, changing $d(p)$, and correspondingly Problem (2), but keeping the observed data completely unchanged.

In practice, many real pricing datasets can exhibit this issue since historical prices are not set completely at random. Therefore, an important concern, which we have found is not fully and consciously addressed in data-driven pricing theory and applications, is the distinction between *prediction* and *prescription*, which we argue becomes relevant here due to issues of endogeneity and causality. Concretely, predictive approaches, which are common in the literature, may be leaving money on the table. In this paper, we explore this problem using a fundamental building block of pricing: the choice of a single price for a single product in a single sale instance, that is, Problem (2). Next we consider an example with continuous prices and demands to illustrate that as much as 100% of the profit could be potentially lost.

EXAMPLE 1 (CONTINUOUS EXAMPLE). Consider procurement cost $c = 0$, potential prices $\mathcal{P} = (0, \infty)$, and demand curve

$$D(p) = 27.75 - p^2 + 6Xp - 9X^2 + Y, \quad (3)$$

where $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(0, \sigma^2)$ are normal noise, and prices historically set as $P = 3X + Z$ where $Z \sim \mathcal{N}(0, \tau^2 = 15.1234)$.

The best predictor of D , given an observation of $P = p$, is

$$\mathbb{E}[D|P=p] = 27.75 - p^2 + 6p\mathbb{E}[X|3X=p-Z] - 9\mathbb{E}[X^2|3X=p-Z] + \mathbb{E}[Y] = 22.108 - 0.393p^2,$$

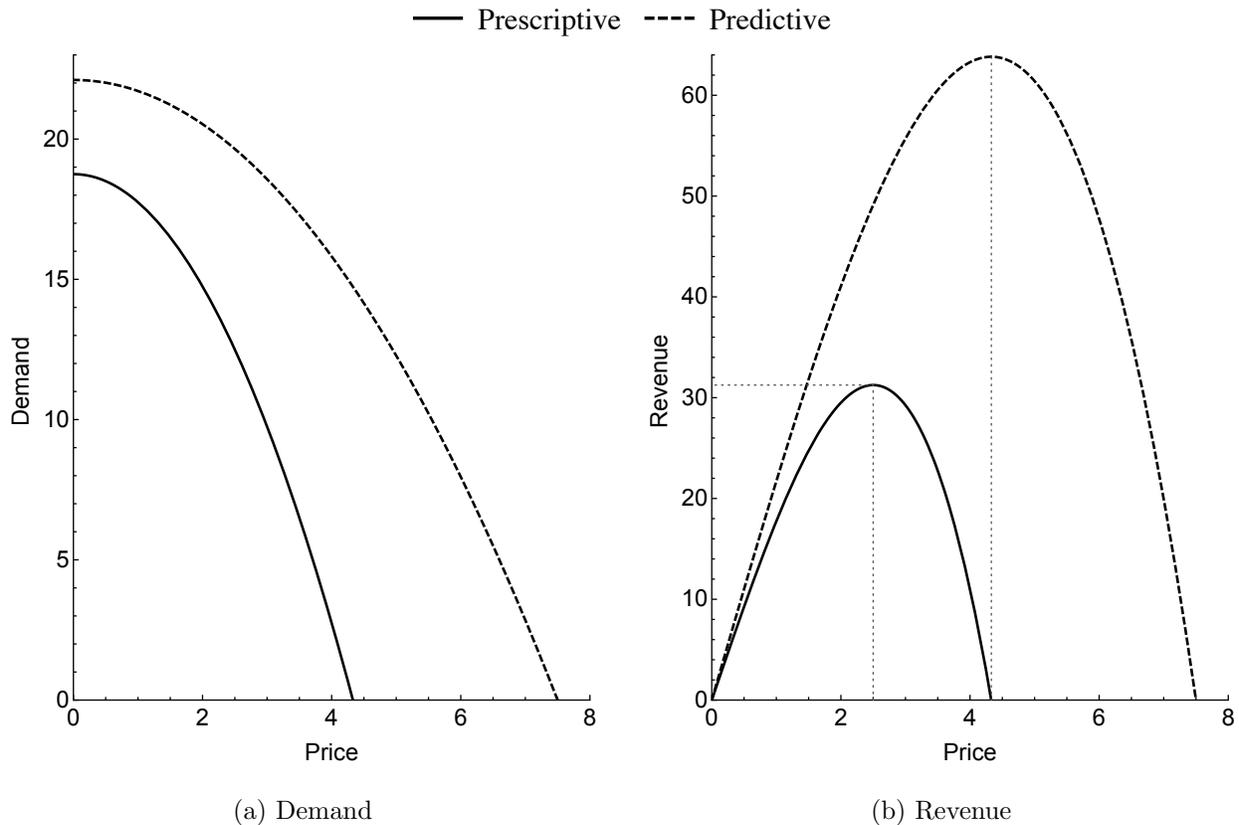
which we get by plugging in P into (3) and recognizing $(X | 3X = p - Z) \sim \mathcal{N}\left(\frac{3p}{9+\tau^2}, \frac{\tau^2}{9+\tau^2}\right)$. This is exactly the function we would arrive at if we used data to regress demand on price (e.g., by linear regression on P , P^2 or by non-parametric regression on P). However, the expected demand when the price is set to p is a different function,

$$\mathbb{E}[D(p)] = 27.75 - p^2 + \mathbb{E}[X] - 9\mathbb{E}[X^2] + \mathbb{E}[Y] = 18.75 - p^2,$$

which we get by taking the expectation of (3). We plot these two functions in Figure 1(a).

Now consider the price prescription problem. The true profit function, $R(p)$, is optimized at $p^* = 2.5$ with a value of 31.25. On the other hand, a predictive approach suggests we should optimize $\tilde{R}(p)$, leading to the price $\tilde{p} = 4.330$, which leads to exactly 0 profit under the true profit function. We plot R , \tilde{R} , p^* , and \tilde{p} in Figure 1(b).

Figure 1 Prediction vs Prescription in Example 1



Note: Solid lines show the true PRF and profit function and dashed lines show the conditional expectation of demand and the spurious profit function that would arise from it.

In the above example, the variable X confounds price P with the demand curve $D(p)$, i.e., it creates an association between those random variables. Because of this, the prediction and prescription problems are different. For prediction, the optimal solution is the conditional expectation $\tilde{d}(p) = \mathbb{E}[D|P=p]$, and, with data, we would solve this problem by fitting a linear, other parametric, or non-parametric regression. For example, the non-parametric Nadaraya-Watson kernel regression used in Besbes et al. (2010) as an estimator for $\tilde{R}(p)$ is a universally consistent estimator under mild conditions (Greblicki et al. 1984). Prediction gets at the hidden value of demand given an observation of price, but not at the causal effect on demand of setting the price. While prediction is a well-posed and valid question to raise, it may not in general have bearing on prescription in observational settings. Were the data a result of a randomized controlled trial (RCT) experiment, the association between price and demand curve would not be present. In observational data, it is generally present, giving rise to confounding. The distinction between association and causation is common (Spirtes 2010) as is addressing issues of endogeneity in econometric supply-demand-price

analyses (Phillips et al. 2012, Berry et al. 1995, Bijmolt et al. 2005). Here we focus on ramifications for the prescriptive problem of price optimization.

We summarize the paper below and refer the reader to the relevant sections.

1. We bound the suboptimality of any predictive approach with respect to the true optimum even when the optimum cannot be identified from the data (Section 2). Our bounds leverage the special structure of the pricing problem and common features of pricing data (e.g., discount promotions often coincide with advertising promotions).
2. We study the theoretical issue of identifiability of the optimal price (Section 3) – is it always possible to compute the optimum from data? – and demonstrate that observational data alone do not suffice to identify the optimal price in Problem (2) (Section 3.1).
3. We provide sufficient conditions, under which the optimal price can indeed be identified from observational data (Section 3.2).
4. Given these conditions and drawing on the literatures of nonparametric estimation and causal inference, we provide algorithms for solving Problem (2) (Section 4). We provide a nonparametric method for price optimization with guarantees of asymptotic optimality, but, since large amounts of data may be necessary before these asymptotics “kick in,” we also provide a parametric framework specially tailored to pricing from observational data.
5. In the non-parametric case, we prove consistency and asymptotic normality and establish rates of convergence (Section 4.1). That is, the algorithm solves Problem (2) based on observational data under only mild regularity conditions and without any model specification.
6. We develop a hypothesis test for revenue optimality of any pricing algorithm that is based on observational data (Section 5). The test, which extends the work of Besbes et al. (2010) to the case of Problem (2) and to observational data, allows us to determine whether profits generated by any one pricing algorithm, such as a predictive one, can be distinguished as suboptimal to a statistically significant degree based on purely observational data.
7. Using this test, we demonstrate empirically that predictive approaches generate profits that are suboptimal to a statistically significant degree and that our parametric algorithm cannot be distinguished from optimal and recovers 36-70% of profits lost by predictive approaches (Section 5.3). The first finding shows that the distinction between prediction and prescription is of real, practical relevance. The second finding expands the scope of recent work on the sufficiency of parametric models for pricing (Besbes et al. 2010, Besbes and Zeevi 2015) to pricing from observational data.
8. In Section 7, we discuss the ways that the ideas presented in the paper generalize to related problems such as customized pricing. Some proofs are given in Section 8.

2. Bounding the Suboptimality of a Predictive Approach to Pricing

In this section, we provide two bounds on how suboptimal a predictive approach may be in general. The intention is to leverage the special structure of the pricing problem to address error in the metric of interest, which is true profit $R(p)$, and express the error in relative terms involving few parameters. Both bounds are expressed relative to the *size of the market*

$$d_0 = \sup_{p \in \mathcal{P}} d(p).$$

As exemplified in the introduction, predictive approaches to pricing from observational data fail because there is confounding (i.e., association or dependence) between the price P and the demand curve $D(p)$. We can quantify confounding as the difference between predicted demand and the effect of a price prescription, $E(p) := \tilde{d}(p) - d(p)$. Let $\epsilon(p) = D(p) - d(p)$ be the deviation of the demand curve from the population mean. Following our convention for $D = D(P)$, let $\epsilon = \epsilon(P)$. Then, $E(p) = \mathbb{E}[\epsilon | P = p]$. We call $E = E(P) = \mathbb{E}[\epsilon | P]$ the confounding error. Thus, the discrepancy E is directly related to the association of price P and the idiosyncrasies ϵ of historical sale events. Independence of the two – i.e., that prices were chosen without regard to the particular sale event as in an RCT – would imply that the confounding error is zero, but generally it is not. Note, in particular, the distinction between ϵ and the regression errors (residuals) $\xi = D - \mathbb{E}[D | P = p]$. Regression errors, by their very definition, will always have $\mathbb{E}[\xi | P] = 0$ even in general observational data, but this is not generally true of ϵ .

The first bound is based on the magnitude of confounding error relative to the size of the market.

THEOREM 1. *Let $\mathcal{P} = (c, \infty)$. If $d(p)$ is decreasing and linear and $|E|/d_0 \leq \gamma$ then*

$$1 - 4\gamma - 4\gamma^{3/2} - \gamma^2 \leq \frac{R(\tilde{p})}{R(p^*)} \leq 1,$$

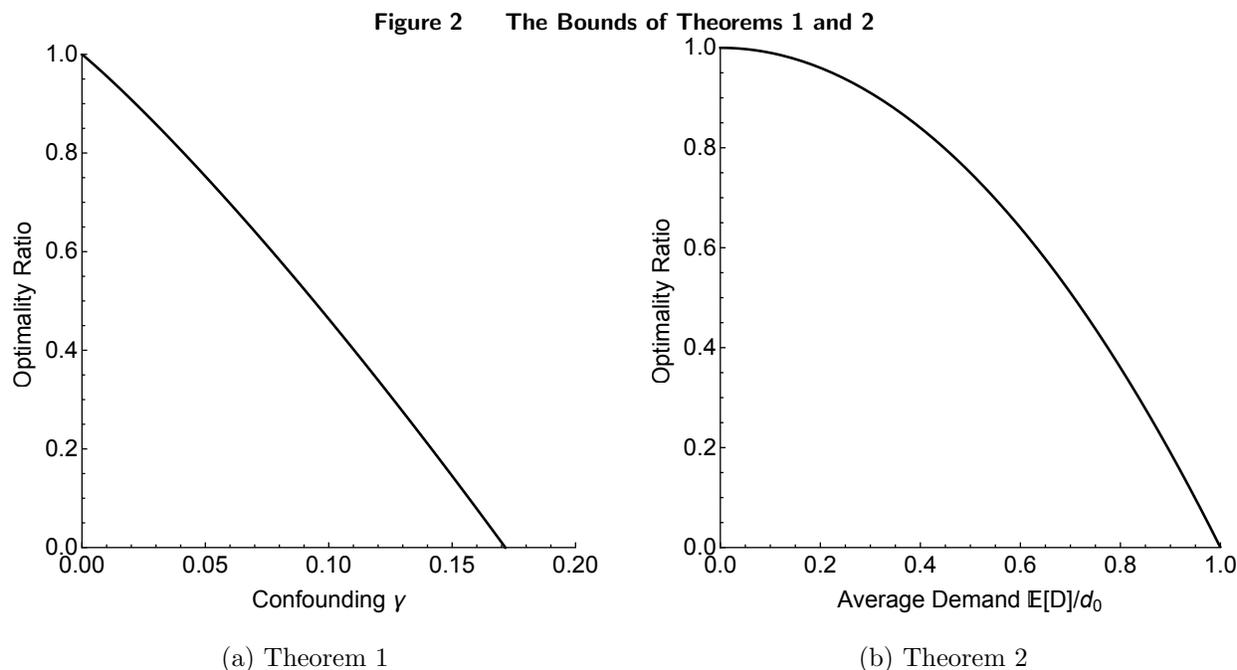
where \tilde{p} and p^* are the optimal solutions to Problem (1) and (2), respectively.

The proof is given in the appendix. The bound is non-negative for values of γ up to $3 - \sqrt{8} \approx 0.17$. We plot the bound in Figure 2(a).

In general, the magnitude of the confounding error is neither known nor estimable from data. The second bound seeks to leverage particular structure that is symptomatic of pricing data to express suboptimality in terms of average demand, which can be estimated from data.

THEOREM 2. *Let $\mathcal{P} = (c, \infty)$. If $d(p)$ is linear and decreasing and ϵ and P are non-positively correlated and jointly normal then*

$$1 - \left(\frac{\mathbb{E}[D]}{d_0} \right)^2 \leq \frac{R(\tilde{p})}{R(p^*)} \leq 1.$$



The proof is given in the appendix.

Note that $\mathbb{E}[D]$ can be unbiasedly and consistently estimated by the sample average of observed demands. The assumption of non-positive correlation between ϵ and P corresponds to a common feature of pricing datasets. For example, promoting a product via advertising, which would increase its potential demand at any given price, would often coincide with promotion via price discounts and this would lead to such non-positive correlation. This assumption, which can be reasoned about in such a way, leads to a stronger bound that is independent of the size of confounding error. We plot the bound in Figure 2(b).

3. Identifiability

In the last section, we saw that the prescription Problem (2) is distinct from Problem (1), while in some cases (1) can still be used to approximate (2). Next we address solving (2) directly and ask whether it can be solved based on data alone. The next example shows that, without further assumptions, the optimal price as prescribed by (2) is not identifiable based on data (we define this precisely below).

EXAMPLE 2 (CONSULTING FOR THE MIT COOP). Alice and Bob are hired by the MIT Coop to help determine an optimal sale price for the classic MIT hoodie, which the MIT Coop procures at a unit price of $c = \$19$ ($r(p) = p - 19$). The MIT Coop is debating between a retail price of \$20 and a retail price of \$28. In any given day in the past, the MIT Coop has offered the hoodie at either of the two prices and observed either no units sold, one unit sold, or two units sold. The Coop has a great deal of observational data.

Table 2 Data for Example 2

P	D	\mathbb{P}
20	0	0
20	1	8/18
20	2	1/18
28	0	8/28
28	1	1/18
28	2	0

P	$D(20)$	$D(28)$	\mathbb{P}
20	1	0	32/81
20	1	1	4/81
20	2	0	4/81
20	2	1	1/162
28	1	0	32/81
28	1	1	4/81
28	2	0	4/81
28	2	1	1/162

P	$D(20)$	$D(28)$	\mathbb{P}
20	1	0	40/99
20	1	1	4/99
20	2	0	0
20	2	1	1/18
28	1	0	4/9
28	1	1	2/45
28	2	0	0
28	2	1	1/90

(a) Joint distribution of historical price and demand

(b) Alice's demand model

(c) Bob's demand model

Alice and Bob collate the data into a table that shows the frequency of each price-demand combination over history shown in Table 2(a). Due to the abundance of data, Alice and Bob are confident that this is a faithful representation of the joint distribution of (P, D) . Naturally, the data only has the demand that was in fact observed and the demand that would have been observed under any other price is missing. A full demand model models the distribution of the demand curve $D(\cdot)$ for a new sale event.

Alice regresses demand on price by computing a weighted average in each of the columns of Table 2(a) and finds that

$$\mathbb{E}[D|P=p] = \begin{cases} 10/9, & p=20, \\ 1/9, & p=28. \end{cases} \quad (4)$$

She constructs a demand model wherein the PRF is equal to the conditional expectation and arrives at the one shown in Table 2(b). Alice verifies that her model fully agrees with the observed data (via the transformation $D = D(P)$) and computes

$$R(p) = r(p)\mathbb{E}[D(p)] = \begin{cases} (20-19) \times (\frac{72}{81} \times 1 + \frac{9}{81} \times 2) = 10/9, & p=20, \\ (28-19) \times (\frac{72}{81} \times 0 + \frac{9}{81} \times 1) = 1, & p=28, \end{cases} \quad (5)$$

concluding that $p^* = 20$ is the optimal price.

Bob, working from home that day and unaware of Alice's progress, has independently come up with another model, shown in Table 2(c), in order to explain the observed pricing data. Bob, too, verifies that his model completely agrees with the observed data and calculates

$$R(p) = r(p)\mathbb{E}[D(p)] = \begin{cases} (20-19) \times (\frac{14}{13} \times 1 + \frac{1}{15} \times 2) = 16/15, & p=20, \\ (28-19) \times (\frac{28}{33} \times 0 + \frac{9}{33} \times 1) = 15/11, & p=28, \end{cases} \quad (6)$$

concluding, differently from Alice, that $p^* = 28$ is in fact the optimal price.

Alice and Bob had both come up with demand models that fully concur with the observed data but recommended different prices as optimal. Both models support the data fully. Both models

give rise to the same conditional expectation (regression) function as in (4), which, in particular, agrees with the PRF $\mathbb{E}[D(p)]$ under Alice's model, but not under Bob model. Both models, as well as the data, fully agree with a homoscedastic linear model,

$$D = \frac{65}{18} - \frac{1}{8}P + \xi, \quad \xi = \begin{cases} -1/9, & \text{with prob. } 8/9, \\ 8/9, & \text{with prob. } 1/9, \end{cases} \quad \xi \perp P,$$

where regressor P is independent of error ξ . Since the two models agree on this form but recommended different prices, this highlights that this is not an issue of misspecifying a functional form for the demand model.

3.1. Non-Identifiability

The issue we encountered above is one of identifiability and shows that the optimal price is non-identifiable in general.

DEFINITION 1. Let $\Pi = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a model for the distribution of the observed data. We say that $\phi : \Theta \rightarrow \Phi$ is *identifiable* if for any $\mathbb{P}_{\theta_1}, \mathbb{P}_{\theta_2} \in \Pi$ such that $\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2}$, we have $\phi(\theta_1) = \phi(\theta_2)$.

In the above definition, Θ and Φ may be arbitrary sets, that is, the model need not be parametric. Note that if any ϕ is not identifiable then any finer quantity, such as θ itself, is not identifiable.

To connect the above definition with our pricing setting, we let θ denote the joint distribution of $(P, D(\cdot))$, we let \mathbb{P}_θ be the corresponding distribution of the data (P, D) (begotten via the transformation $D = D(P)$), and we let ϕ map θ to the optimal price (or, set thereof) as described by Problem (2). Then, Example 2 above provides a proof by example of the following result:

COROLLARY 1. *The optimal price p^* is not identifiable on the basis of observations of (P, D) .*

In fact, we have shown a stronger result:

THEOREM 3. *The optimal price p^* is not identifiable on the basis of observations of (P, D) even under the Gauss-Markov assumptions:*

- a. *Linearity: there is a random variable ξ such that $D = \beta_0 + \beta_1 P + \xi$.*
- b. *Exogeneity of independent variables: $\mathbb{E}[\xi|P] = 0$.*
- c. *Homoscedasticity: $\text{Var}(\xi|P) = \text{Var}(\xi)$ is constant.*
- d. *No collinearity: P is not constant.*

In Example 2, exogeneity and homoskedasticity are a consequence of $\mathbb{E}[\xi] = 0$ and $\xi \perp P$. Exogeneity implies $\text{Cov}(\xi, P) = 0$. Note that whenever the optimal price is not identifiable, the PRF $d(p)$, a finer quantity, is not identifiable either.

3.2. Conditions for Identifiability

To identify the optimal price we need to factor out the association between P and the demand curve $D(\cdot)$. For this purpose, let us now also consider observing auxiliary covariates X associated with historical sale events, e.g., characteristics of the customer for whom price was customized, whether a product was featured in a promotional flyer, external signals about demand used for pricing, etc. Even with this data, for the moment, we still restrict ourselves to the problem of choosing a single price for the whole population of sale events; we consider the extension where prices can be customized on the basis of covariates in Section 7.1.

In Section 1, we saw that one of the important distinctions between prediction and prescription is that, in prediction, price P is a random variable, whereas, in prescription, price p is a control. Sometimes covariates X can help us disassociate the random variable P and a particular sale event and its demand curve $D(p)$, which is of sole relevance in the prescription problem. One such sufficient condition for X to account for this association is the following continuous generalization of the assumption made by Rosenbaum and Rubin (1983).

ASSUMPTION 1. *For every $p \in \mathcal{P}$, we have that, conditioned on X , $D(p)$ is independent of P . That is,*

$$\forall p, \quad D(p) \perp\!\!\!\perp P \mid X.$$

Under this condition, which we discuss further below, we have identifiability.

THEOREM 4. *Under Assumption 1, the optimal price p^* is identifiable on the basis of observations of (P, X, D) .*

Proof We have

$$\mathbb{E}[D(p)] = \mathbb{E}[\mathbb{E}[D(p)|X]] = \mathbb{E}[\mathbb{E}[D(p)|P=p, X]] = \mathbb{E}[\mathbb{E}[D(P)|P=p, X]] = \mathbb{E}[\mathbb{E}[D|P=p, X]],$$

where the first equality is by iterated expectations and the second is by Assumption 1. The last expectation is expressed solely in terms of the joint distribution of (P, X, D) , which gives the identifiability of the PRF and hence the profit function $R(p)$. The optimal price is given by optimizing $R(p)$, completing the proof. \square

Note the last expectation is not conditioned on $P=p$ and cannot be marginalized via iterated expectations. In words, it says to take the conditional expectation of D given $X=x, P=p$ and to average it over all x using the marginal distribution of X (and *not* the conditional distribution given $P=p$).

In words, Assumption 1 says that, historically, X accounts for all the sale-event-specific features that influenced managerial price-setting. Managers usually set prices strategically rather than at

random. Thus, if prices are set in (partial) anticipation of demand, then prices and demand are confounded. If X accounts for the information based upon which the price was selected then Assumption 1 is nonetheless satisfied. Because of this, this sort of condition is sometimes termed selection on observables (this is, however, imprecise because Assumption 1 does not imply that price is a *function* of observables). Note that the conditional independence in Assumption 1 is between the historical price and the *potential* demand at some price p , not historically observed demand. The independence need only hold separately for each price p .¹

If the manager chooses prices without regard to any specific sale event, then Assumption 1 holds with a null X variable (formally, $\sigma(X) = \{\Omega, \emptyset\}$). In particular, this is the case in dynamic demand learning and pricing as in Bertsimas and Perakis (2006), Besbes and Zeevi (2009), Harrison et al. (2012) because each sale event is assumed independent and nothing about a present sale event is considered when setting the price. This is the experimental setting. Unfortunately, it rarely holds in practice for observational data.

If there is not sufficient recorded information in X to merit Assumption 1, it is said that there is residual endogeneity. In this case, Theorem 4 fails, but there may be other conditions that enable identification such as the availability of instrumental variables (see e.g. Bijmolt et al. (2005)). Here we focus on data that arises from historical pricing by managers whose behavior is well understood or even documented and hence what information influenced pricing decisions is known and observable.

Let us consider Assumption 1 and its ramifications in some specific examples.

EXAMPLE 3 (CONSULTING FOR THE MIT COOP). Consider again the hypothetical case of Example 2. Recall, Alice and Bob both came up with models for demand that completely agreed with the data but gave rise to different optimal prices. Thus, we concluded that the data observed could not possibly identify the right optimal price.

Suppose Assumption 1 holds with X being a null variable, i.e., without any extra information. This condition eliminates Bob’s model – it no longer agrees with both the data and this condition. On the other hand, Alice’s model remains valid – in fact it turns out to be the unique model that agrees with both the data and this condition. Hence, under this condition, $p^* = 20$ is the correct optimal price. But for Assumption 1 to hold with X being a null variable we would have needed experimental data, where prices are set at random for the sake of experiment.

Suppose instead that we recorded additional information about each sale event: whether there was a major home game that day ($X = 1$) or not ($X = 0$). On average, there is a game 2 days of each month ($\mathbb{P}(X = 1) = 2/30$). Suppose tallying the historical observations led to the summary of the data shown in Table 3. If prices were chosen independently of whether there was a game, the

¹ Also note that conditional mean-independence is sufficient to prove Theorem 4. However, we will need the stochastic independence later on.

Table 3 Data for Example 3

	$X = 0$		$X = 1$	
	$P = 20$	$P = 28$	$P = 20$	$P = 28$
D=0	0	4/9	0	0
D=1	4/9	2/45	0	1/90
D=2	0	0	1/18	0

previous scenario still holds and Alice’s model is the uniquely correct one. If, however, to capitalize on the promotion offered by a major home game, prices were more often cut to \$20 when there was a game, then we are no longer in the experimental setting. In fact, if we assume Assumption 1 holds with X being whether there is a game, then it turns out that Alice’s model is ruled out and Bob’s model is the unique model that accommodates both this condition and the data observed, in which case $p^* = 28$ is the correct optimal price. In fact, Bob’s model can be written as follows. If there is no game then $D(20) = 1$, $D(28) = 0$ with probability $10/11$ and otherwise $D(20) = D(28) = 0$, and $P = 20$ with probability $10/21$ and otherwise $P = 28$, each independently. If there is a game then $D(20) = 2$ and $D(28) = 1$ with probability 1 and $P = 20$ with probability $5/6$ and otherwise $P = 28$, each independently. In this hypothetical example, we are seeking a universal price, to be set a priori without regard to whether there is a game, but the price can also be customized (as was done historically); we consider customization in Section 7.1.

EXAMPLE 4 (AUTO LOAN RATE OPTIMIZATION). In Besbes et al. (2010), the authors study the problem of prescribing interest rates for automobile loans based on historical, observational data. The on-line auto lending data, provided by Columbia University Center for Pricing and Revenue Management (2012), consists of past sale events where a customer fills out a loan application, if approved an interest rate is quoted (price), and the customer either accepts or rejects the loan (binary demand). The authors study the differences – and how to test for them – between pricing strategies generated by optimizing based on either non-parametric (Nadaraya-Watson kernel) regression or parametric (logistic) regression. Both approaches address the prediction problem and start by estimating $\tilde{d}(p)$ or $\tilde{R}(p)$ within each of several predefined subpopulations.

The dataset description says that approval and rate is based on “credit information and other criteria.” Such criteria would almost certainly also be associated with the potential likelihood of the consumer to accept a loan offer at any one particular rate (the demand curve). Even if rates are not chosen strategically in response to demand, they could be chosen based on default risk or expected loss, which may be in turn associated with the demand curve. Therefore, Assumption 1 does not hold with null X . The data, however, contains much information about each loan applicant and the associated sale event, including the FICO credit score of the applicant, the length of the term over which the loan is to be repaid, the dollar amount of the loan, whether the car to be purchased is

new, used, or refinanced, competitors’ rate, prime rate, and who referred the applicant. From here on, we let X denote these covariates. If these encompass the aforementioned “credit information and other criteria,” then Assumption 1 would hold. Note that FICO score only accounts for credit information and not the “other criteria” that influence rates.

In Besbes et al. (2010), the authors consider setting a single price for each of eight customer segments prescribed by predefined ranges of FICO scores, term lengths, and season (see their paper or Example 7 for additional detail). Let $Y = 1, \dots, 8$ denote membership in each of the segments. Note that $Y = s(X)$ is strictly coarser than X . Paraphrased, their approach to pricing is to estimate $\mathbb{E}[D|P = p, Y = i]$ (which is the same as $\mathbb{P}(D = 1|P = p, Y = i)$ because demand is binary) within each segment either parametrically or non-parametrically, prescribing the segment-wide price that maximizes the estimated conditional expectation times per-unit profit $r(p)$. The authors assume that data points (D_i, P_i) are iid (Assumption 1 therein) but this does not imply the independence within each data point prescribed by Assumption 1.

Since $Y = s(X)$ is coarser than X , Assumption 1 with respect to X does not generally imply the same with respect to Y . In particular, since, besides binning, Y also removes price-driving dimensions such as loan amount, Assumption 1 with respect to Y is not a reasonable assumption, suggesting that even the non-parametric model the authors consider need not converge to the true PRF.²

It is important, nonetheless, to keep in mind that this distinction is completely moot if, at the end of the day, profits generated by such a pricing scheme are indistinguishable from optimal. It is for this reason that, inspired by Besbes et al. (2010), we seek to develop a statistical test for profit optimality. Using the statistical test we develop in Section 5, we test whether it is the case that profits generated by these predictive approaches are indistinguishable from optimal to a statistically significant degree – or whether a finer analysis leads to greater profit – when we next return to this example.

4. Solutions to the Prescriptive Problem

In the last section, we saw that Assumption 1 enables identification, i.e., the data-driven pricing problem is hypothetically solvable. Now we turn to specific solutions, assuming Assumption 1.

4.1. A Non-Parametric Solution

We begin with a non-parametric solution that is model-independent in that it will converge to the optimal price regardless of the true underlying distribution, given sufficient data and some assumptions. Henceforth, we assume that X is a vector of covariates taking values in \mathbb{R}^k .

² Conditional expectations given X can be estimated based on partitioning into segments – as in regression trees and k -nearest neighbors – but these must be data-driven and shrinking with sample size, not fixed and removing certain dimensions a priori.

The proof of Theorem 4 says that, under Assumption 1, the profit function can be written as $R(p) = \mathbb{E} [\mathbb{E} [r(P)D | P=p, X]]$. Thus, to estimate the profit function, one approach may be to estimate the regression function $\mathbb{E} [r(P)D | P=p, X=x]$ and then average the estimated function over an estimate for the marginal distribution of X . Then, the optimizer of this estimate can be used as an estimator for the optimal price.

First, one non-parametric estimate of the marginal distribution of X is the empirical distribution, which places unit mass at each of the observations X_i . Second, to estimate the regression function $\mathbb{E} [r(P)D | P=p, X=x]$ non-parametrically, we can use Nadaraya-Watson kernel regression (Nadaraya 1964, Watson 1964). The estimate, based on a choice of kernel $K : \mathbb{R}^{1+k} \rightarrow \mathbb{R}_+$ and bandwidth $h_n > 0$, is

$$\bar{R}_n(p, x) = \frac{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right) r(P_i)D_i}{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right)}, \quad (7)$$

where $K\left(\frac{p-P_i}{h_n}, \frac{x-X_i}{h_n}\right) = K\left(\frac{p-P_i}{h_n}, \frac{x_1-X_{i1}}{h_n}, \dots, \frac{x_k-X_{ik}}{h_n}\right)$. A kernel mimics a continuous distribution centered at the data points, the width of which is determined by the bandwidth. This regression estimator arises as the conditional expectation with respect to the Parzen window density estimator for the joint distribution of $(P, X, r(P)D)$ and that of (P, X) (Parzen 1962). There are a variety of kernels used in practice (Härdle 1990). Our requirements for a kernel function and bandwidth are as summarized as follows:

ASSUMPTION 2 (Kernel Conditions).

- a. $0 < \int_{\mathbb{R}^{1+k}} K(u) du < \infty$.
- b. K is zero outside a bounded set.
- c. K is twice Lipschitz-continuously differentiable.
- d. K has order at least $s \in \mathbb{N}$, that is, $\int K(u)u^\alpha du = 0 \quad \forall \alpha \in \mathbb{N}^{1+k} : |\alpha| < s$.
- e. $h_n \rightarrow 0$ and $nh_n^{2s+3} \rightarrow 0$.
- f. $nh_n^{k+5} / \log(n) \rightarrow \infty$ and $nh_n^{2k+1} / \log(n)^2 \rightarrow \infty$.

Combining the two estimators as detailed above, we arrive at the following estimate for the profit function

$$\bar{R}_n(p) = \frac{1}{n} \sum_{i=1}^n \bar{R}_n(p, X_i) = \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^n K\left(\frac{p-P_j}{h_n}, \frac{X_i-X_j}{h_n}\right) r(P_j)D_j}{\sum_{i=1}^n K\left(\frac{p-P_i}{h_n}, \frac{X_i-X_j}{h_n}\right)}. \quad (8)$$

Optimizing the above estimate over price yields a non-parametric data-driven price prescription

$$\bar{p}_n \in \arg \max_{p \in \mathcal{P}} \bar{R}_n(p). \quad (9)$$

One question that arises is how do these prices behave asymptotically. In particular, does this pricing strategy lead to prices and profits that converge to the optimal price and profit. Since the estimates are non-parametric, the hope is that this can occur under model-free assumptions. Next we show that this is indeed the case under the following regularity conditions.

ASSUMPTION 3 (Optimal Price Conditions).

- a. \mathcal{P} is compact.
- b. p^* uniquely maximizes $R(p)$ on \mathcal{P} .
- c. p^* lies in the interior of \mathcal{P} .
- d. $R(p)$ is twice continuously differentiable and $R''(p^*) < 0$.

ASSUMPTION 4 (Distributional Conditions).

- a. X and P are continuously distributed on a compact support where the joint density, $f_{P,X}(p, x)$, is bounded away from zero.
- b. The marginal density of X , $f_X(x)$, is bounded and continuously differentiable.
- c. $\mathbb{E}[D^4] < \infty$ and $\mathbb{E}[D^4|P=p, X=x]$ is bounded.
- d. $\mathbb{E}[D^2|P=p, X=x]$ is continuously differentiable.
- e. $\mathbb{E}[D|P=p, X=x]$ and $f_{P,X}(p, x)$ are $s+1$ times continuously, boundedly differentiable.

Under these conditions, we can show the following asymptotic optimality and rates.

THEOREM 5. *Under Assumptions 1, 2, 3, and 4, we have that*

$$\begin{aligned} \sqrt{nh_n}(R(p) - \bar{R}_n(p)) &\xrightarrow{d} \mathcal{N}(0, \eta_p \kappa) \quad \forall p \in \mathcal{P}, \\ \sqrt{nh_n^3}(p^* - \bar{p}_n) &\xrightarrow{d} \mathcal{N}\left(0, \frac{\eta_{p^*} \kappa'}{R''(p^*)^2}\right) \\ (nh_n^3)(R(p^*) - R(\bar{p}_n)) &\xrightarrow{d} \frac{-\eta_{p^*} \kappa'}{2R''(p^*)} \chi_1^2, \end{aligned}$$

and, if also $nh_n^{2s+1} \rightarrow 0$, then

$$\sqrt{nh_n}(R(p^*) - \bar{R}_n(\bar{p}_n)) \xrightarrow{d} \mathcal{N}(0, \eta_{p^*} \kappa),$$

where $\mathcal{N}(0, \sigma^2)$ denotes a centered normal distribution with variance σ^2 , χ_1^2 denotes a chi-squared distribution with one degree of freedom, and η_p , κ , κ' are constants defined as follows

$$\eta_p = r(p)^2 \mathbb{E} \left[\frac{\text{Var}(D|P=p, X)}{f_{P|X}(p|X)} \right], \quad \kappa = \int \tilde{K}(p)^2 dp, \quad \kappa' = \int \tilde{K}'(p)^2 dp,$$

where $\tilde{K}(p) = \int K(p, x) dx$ and $f_{P|X}(p|x) = f_{P,X}(p, x)/f_X(x)$ is the conditional density of P .

The proof is given in the appendix.

The main implication of Theorem 5 is that, under regularity conditions but without model specification, the non-parametric pricing strategy has profits that converge to optimal with rate of convergence $1/n$. Note that Assumption 2 implies that $s \geq k$ when $k \geq 3$ and $s \geq k + 1$ when $k \leq 2$. This means that a kernel of order strictly greater than two, also known as a “bias-reducing” kernel (Hansen 2009), is necessary when $k \geq 2$.

4.2. A Parametric Solution

In the preceding section we developed a non-parametric pricing strategy that converged to optimal without requiring any model to be specified. Non-parametric approaches, however, can sometimes be unwieldy because their shapelessness makes them uninterpretable and they may be slow to converge. In fact, there is a growing body of work (Besbes et al. 2010, Besbes and Zeevi 2015) arguing that parametric models are often sufficient for pricing problems, as the model may need only fit well near the optimum. In particular, what matters is not model fit but objective performance. In this section, we develop a particular parametric pricing strategy using a generalization of the propensity score.

The propensity score is a common matching metric used in the comparison of binary treatments in observational data (Rosenbaum and Rubin 1983). The (conventional) propensity score of a study subject is equal to the conditional probability of receiving a treatment (rather than control) given the subject’s covariates X . If treatments are continuous, the generalized propensity score of a unit is defined as the conditional density of the unit receiving whatever treatment it did receive given the subject’s covariates (Robins et al. 2000, Hirano and Imbens 2004, Imai and Van Dyk 2004).

In our problem, the generalized propensity score is $Q = f_{P|X}(P, X)$, that is, one takes the conditional density $f_{P|X}(p|x)$, which is non-random, and plugs in as values the random variables P and X . The key property of the generalized propensity score is that it is sufficient as a control for identifying the PRF (Hirano and Imbens 2004).

THEOREM 6. *Suppose Assumption 1 holds. Then the PRF satisfies*

$$\mathbb{E}[D(p)] = \mathbb{E}[d(p, f_{P|X}(p, X))], \text{ where } d(p, q) = \mathbb{E}[D|P = p, Q = q].$$

The proof follows Hirano and Imbens (2004) and is given in the appendix. The implication of Theorem 6 is that it is sufficient to control for the univariate generalized propensity score rather than all of X .

EXAMPLE 5 (CONTINUOUS EXAMPLE). Recall the setup from Example 1. First, note that Assumption 1 holds with respect to X . Moreover, since $P = 3X$, we have that

$$f_{P|X}(p, x) = \frac{\sqrt{9 + \tau^2}}{\sqrt{2\pi} \times \tau} e^{-\frac{(3p - (9 + \tau^2)x)^2}{2 \times \tau^2 \times (9 + \tau^2)}}.$$

Hence the generalized propensity score is $Q = f_{P|X}(P, X)$ and, moreover,

$$-2 \times \tau^2 (\log(Q) + \log(2\pi)/2 + \log \tau) = (p - 3x)^2.$$

This means that we can write $D = D(P)$ as $D = 96.624 + 30.247 \log(Q)$ and hence $d(p, q) = \mathbb{E}[D|P = p, Q = q] = 96.624 + 30.247 \log(q)$. Since

$$\mathbb{E}[\log(f_{P|X}(p, X))] = -\frac{p^2 + 9}{2 \times \tau^2} - \log(\sqrt{2\pi} \times \tau),$$

and $\tau = 3.889$, we conclude that $\mathbb{E}[d(p, f_{P|X}(p, X))] = 18.75 - p^2$, which agrees with the PRF $\mathbb{E}[D(p)]$ as calculated in Example 1.

Theorem 6 motivates the following pricing strategy: estimate a probability model $\hat{f}_{P|X}(p, x)$ for $f_{P|X}(p, x)$, impute generalized propensity scores $\hat{Q}_i = \hat{f}_{P|X}(P_i, X_i)$, regress demand on price and imputed scores to produce an estimate $\hat{d}(p, q)$ of $d(p, q)$, for each price p plug in the estimated $\hat{f}_{P|X}(p, X)$, average the result over the empirical distribution of X , and prescribe the price p that maximizes per-unit profit times this estimate. Specifically, the following parametric approach can be followed:

1. Regress P on X by fitting a generalized linear model (GLM) in order to estimate $f_{P|X}(p, x)$. I.e., choose $\hat{\beta}_n, \hat{\tau}_n$ by maximum likelihood estimation, given the parametric model

$$f_{P|X}(p|x; \beta, \tau) = h(p, \tau) \exp\left(\frac{b(\beta_0 + \beta^T x)T(p) - A(\beta_0 + \beta^T x)}{d(\tau)}\right).$$

See McCullagh et al. (1989) for choices of b, T, A, d, h . For example, the choices $b(\mu) = \mu$, $T(p) = p$, $A(\mu) = \mu^2/2$, $d(\tau) = \tau^2$, and $h(p, \tau) = \frac{1}{\sqrt{2\pi\tau}} e^{-\frac{p^2}{2\tau^2}}$ lead to ordinary least squares (OLS). Other examples of GLMs include logistic regression, Poisson regression, Gamma regression, and loglinear regression.

2. Use the fitted GLM to impute generalized propensity scores, setting $\hat{Q}_i = f_{P|X}(P_i|X_i; \hat{\beta}_n, \hat{\tau}_n)$.
3. Regress D on P and \hat{Q} based on the imputed data $\{(P_i, D_i, \hat{Q}_i) : i = 1, \dots, n\}$ using parametric regression (e.g. linear regression for general demand or logistic regression for binary demand) to produce an estimate $\hat{d}_n(p, q)$ of $d(p, q)$. For example, we could fit $D = b^{-1}(\alpha_0 + \alpha_1 p + \alpha_2 q + \alpha_3 q^2 + \epsilon)$ via some link function b (e.g., if a log-log model is appropriate, we regress $\log(D)$ on $\log(P)$ and \hat{Q}).
4. Use these to estimate the PRF and prescribe the price that optimizes estimated profits,

$$\hat{p}_n \in \arg \max_{p \in \mathcal{P}} \left\{ r(p) \times \frac{1}{n} \sum_{i=1}^n \hat{d}_n(p, \hat{f}_{P|X}(p|X_i; \hat{\beta}_n, \hat{\tau}_n)) \right\}.$$

The above procedure provides a flexible parametric framework for data-driven pricing with observational data. When we apply it to examples with both real and synthetic data in Section 5.3 we find that it performs well and produces profit that is statistically indistinguishable from optimal.

5. A Test for Revenue Optimality

In the previous sections we considered various data-driven pricing strategies for observational data. All of these proceeded by estimating the PRF and optimizing resulting estimated profit. Similarly, in their own context, each of Besbes et al. (2010), Cohen et al. (2014), Ferreira et al. (2016) first estimated demand then optimized price. It may be argued that it is important that estimated profits faithfully represent true profits, but in fact this point is moot insofar as actual profits generated by the resulting pricing strategy are satisfactory. This is the key point made by Besbes et al. (2010), where the authors develop a hypothesis test to inspect profit optimality instead of predictive model fit. In the perspective presented herein, this test does not apply to observational data because it relies on a kernel estimate of conditional expectation of profits that may in general bear no relationship to the true profit function. We build on their work in developing an analogous test for the observational setting under Assumption 1.

Suppose we wish to test the profit optimality of a data-driven pricing strategy that prescribes the price \hat{p}_n based on n data points. Let \hat{p} be the corresponding full-information pricing strategy, i.e., the hypothetical price this strategy would pick with infinite data. For added generality, we leave this somewhat vague and only require the following condition in defining what \hat{p} means.

ASSUMPTION 5 (Convergent Pricing Strategy). $\hat{p}_n - \hat{p} = O_p(1/\sqrt{n})$ for some fixed $\hat{p} \in \mathcal{P}$.

The notation $Y_n = O_p(a_n)$ means that for any $\epsilon > 0$ there is $M > 0$ such that $\mathbb{P}(|Y_n/a_n| > M) < \epsilon$ eventually. In particular, if Y_n/a_n converges in distribution (to anything) then $Y_n = O_p(a_n)$.

For example, the strategy \hat{p}_n that fits a kernel regression to estimate conditional expectation of profits given price and optimizes this estimate will (under some regularity) eventually arrive at the price $\hat{p} = \bar{p}$ of the hypothetical Problem (1), and, in particular, satisfies $\hat{p}_n - \hat{p} = O_p(1/\sqrt{n})$ because the difference is asymptotically normal (Ziegler 2002). A similar condition, with a potentially different \hat{p} , is true of the strategy that uses a parametric maximum-likelihood regression (Besbes et al. 2010, cf. Lemma 2 and Lipschitz condition in the proof of Lemma 3). Similarly, Theorem 5 shows that (under some regularity) our non-parametric pricing strategy \bar{p}_n from (9) satisfies $\bar{p}_n - \hat{p} = O_p(1/\sqrt{n})$ with $\hat{p} = p^*$.

We would like to test the following null hypothesis H_0 against the alternative H_1 :

$$H_0 : R(p^*) = R(\hat{p}), \quad H_1 : R(p^*) > R(\hat{p}).$$

That is, we would like to test whether the nominal price that our pricing strategy would be prescribing is generating optimal profits.³

A test for the hypothesis H_0 can be interpreted as rejecting a pricing strategy if it *generates profits that are distinguishable from optimal to a statistically significant degree based on the data*.

³ Note that our null hypothesis differs from the one considered by Besbes et al. (2010) in the definition of $R(p)$, that is, our $R(p) = \mathbb{E}[r(p)D(p)]$ vs. $\tilde{R}(p) = \mathbb{E}[r(p)D|P=p]$ in Besbes et al. (2010).

5.1. Test Statistic and Large Sample Theory

The impediment to verifying our hypothesis is that $R(p)$, p^* , and \hat{p} are all unknown; were they known, we would compute $\rho = R(p^*) - R(\hat{p})$ and compare it to 0. Therefore, we must come up with an observable test statistic as a proxy to ρ . We do this by replacing the unknowns by our consistent estimates for them. We replace $R(p)$ and p^* by our non-parametric estimates $\bar{R}(p)$ as in (8) and \bar{p} as in (9) and we replace \hat{p} by \hat{p}_n . The resulting test statistic is

$$\rho_n = \bar{R}_n(\bar{p}_n) - \bar{R}_n(\hat{p}_n).$$

If ρ_n is small, we have reason to believe that $\rho = 0$, whereas if ρ_n is large, we would believe that $\rho > 0$. The question is where to draw the line.

THEOREM 7. *Suppose Assumptions 1, 2, 3, 4, and 5 hold. Let $\Gamma = \frac{-\eta_{p^*} \kappa'}{2R''(p^*)}$. Then,*

- i. under H_0 , $(nh_n^3) \rho_n \xrightarrow{d} \Gamma \chi_1^2$, and
- ii. under H_1 , $(nh_n^3) \rho_n \xrightarrow{d} \infty$.

The proof is given in the appendix.

Theorem 7 says that if we only reject H_0 when $\rho_n > n^{-1} h_n^{-3} \Gamma F_{\chi_1^2}^{-1}(1 - \alpha)$ (where $F_{\chi_1^2}^{-1}$ is the chi-squared quantile function), then when H_0 is true we would only falsely reject H_0 at most α fraction of the time (asymptotically). On the other hand, if H_0 is false, then we would eventually reject it using such a procedure (a property known as *consistency* of a hypothesis test). The problem is that Γ is unknown meaning that this exact procedure cannot be implemented in practice.

5.2. A Hypothesis Test

One way to implement a hypothesis is to estimate Γ and replace the estimate into the results of Theorem 7. In particular, given any estimate $\hat{\Gamma}_n$ that converges in probability to Γ , we would have as an immediate consequence of Theorem 7 that $(nh_n^3) \hat{\Gamma}_n^{-1} \rho_n$ converges in distribution to χ_1^2 under H_0 and to ∞ under H_1 . This would give an implementable test. Non-parametric estimators for Γ , however, would tend to be convoluted and unwieldy, involving partial means of estimators of conditional variance and density as well as fragile estimates of second derivatives of partial means.

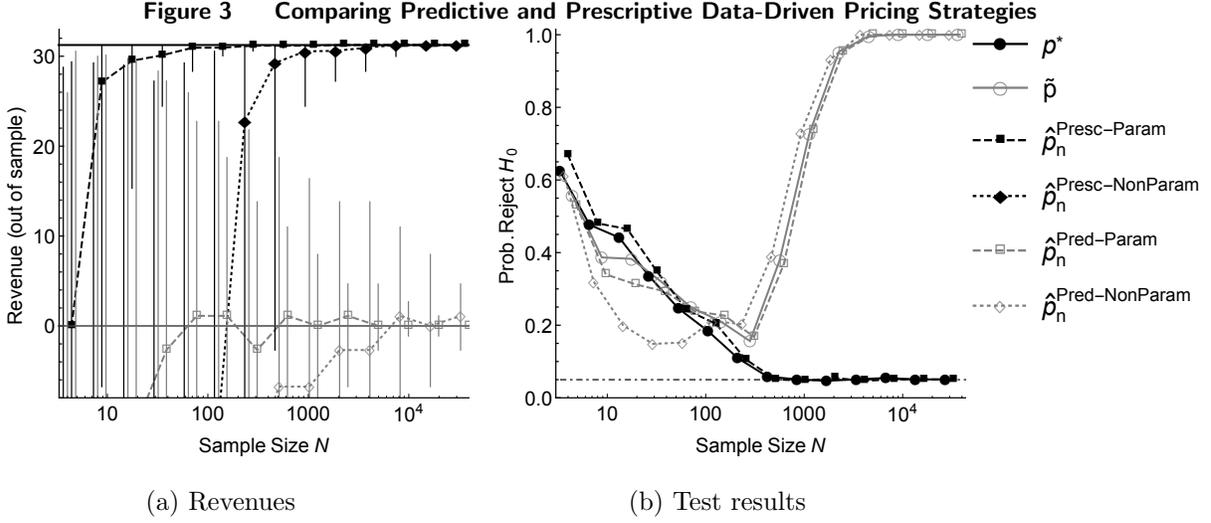
Instead, we use the bootstrap (Efron and Tibshirani 1993) and the following observation.

THEOREM 8. *Suppose Assumptions 1, 2, 3, and 4 hold. Let $A_n = \bar{R}_n(\bar{p}_n) - \bar{R}_n(p^*)$. Then, $(nh_n^3) A_n \xrightarrow{d} \Gamma \chi_1^2$. Consequently, $(nh_n^3) \mathbb{E}[A_n] \rightarrow \Gamma$.*

The proof is given in the appendix.

So, to estimate Γ , we use a scaled estimate of the mean of A_n . In the spirit of Besbes et al. (2010), we use the bootstrap to achieve this. The fact that A_n is asymptotically pivotal suggests that a bootstrap procedure could be particularly powerful (Horowitz 2001).

The procedure is as follows. Compute \bar{p}_n as in (9) based on data \mathcal{S}_n . Fix B large. For $b = 1, \dots, B$,



1. Draw n samples with replacement from \mathcal{S}_n to form the resampled dataset $\mathcal{S}_n^{(b)}$.
2. Compute $\bar{R}_n^{(b)}$ and $\bar{p}_n^{(b)}$ as in (8)-(9) based on the data $\mathcal{S}_n^{(b)}$.
3. Set $A_n^{(b)} = \bar{R}_n^{(b)}(\bar{p}_n^{(b)}) - \bar{R}_n^{(b)}(\bar{p}_n)$.

Let $\hat{\Gamma}_n = \frac{nh_n^3}{B} \sum_{i=1}^n A_n^{(b)}$. Reject H_0 if $\rho_n > n^{-1}h_n^{-3}\hat{\Gamma}_n F_{\chi_1^2}^{-1}(1 - \alpha)$.

This bootstrap procedure is more attractive than convoluted kernel estimates of Γ because it is less dependent on parameters and it deals more directly with the finite-sample distribution of ρ_n . We use this bootstrap test in our numerical experiments in the next section.

5.3. Examples

In this section, we use our test to study the distinction between prediction and prescription in both synthetic and real examples and discern whether it has any consequence in revenue management. We consistently find that ignoring the distinction in cases with observational data can significantly hurt profits. On the other hand, we find that a parametric approach is sufficient for good performance as long as it takes into account this distinction.

EXAMPLE 6 (CONTINUOUS EXAMPLE). Consider again the setup from Example 1 and consider recording n observations from (P, X, D) . We compare four different data-driven pricing strategies: a prescriptive non-parametric approach $\hat{p}_n^{\text{Presc-NonParam}}$, a prescriptive parametric approach $\hat{p}_n^{\text{Presc-Param}}$, a predictive non-parametric approach $\hat{p}_n^{\text{Pred-NonParam}}$, and a predictive parametric approach $\hat{p}_n^{\text{Pred-Param}}$. We also consider the (non-data-driven) true optimal price p^* of (2) and full-information predictive pricing strategy \tilde{p} of (1). The prescriptive non-parametric strategy is as in (9) using a second order Gaussian kernel $K(u) = e^{-\frac{\|u\|_2^2}{2h_n^2}}$ and $h_n = 0.1 \times (n \log(n))^{-1/7}$, which satisfy Assumption 2 with $s = 2, k = 1$. For the prescriptive parametric strategy, we follow our procedure from Section 4.2 using OLS linear regression of P on X for the GLM in step 1

and an OLS linear regression of D on P and $\log(\widehat{Q})$ in step 3. For the predictive non-parametric strategy, we use kernel regression to regress $r(P)D$ on P (using the same kernel and bandwidth $h_n = 2.5 \times (n \log(n))^{-1/7}$) and optimize the estimated regression function. Finally, for the predictive parametric strategy, we perform OLS linear regression of D on P and optimize the estimated regression function times $r(p)$.

First, we consider the profit performance of each of these strategies. We plot the corresponding out-of-sample profits, $R(\hat{p}_n)$, along with optimal profit $R(p^*)$, in Figure 3(a). The plot displays the median profit (center lines) and the 10th and 90th percentiles (vertical lines) over 256 replicate runs of each sample size. The example shows that a predictive approach, whether parametric or not, can potentially leave much on the table in terms of profits. In contrast to the predictive approach, the prescriptive non-parametric approach converges to optimum, in agreement with Theorem 5. On the other hand, the prescriptive parametric approach offers significantly better out-of-sample performance for small samples.

Next, we apply our hypothesis test for profit optimality. We plot the frequencies of rejecting a pricing strategy as significantly suboptimal at a significance of 0.05 in Figure 3(b). The plot displays the fraction of times the null hypothesis is rejected out of 256 replicate runs of each sample size. We see that with sufficient data, the test can distinguish those pricing strategies that generate suboptimal profits (i.e. solely predictive strategies) from those that cannot be distinguished from optimal for all prescriptive intents and purposes. In particular, it takes about a hundred data points before the test has the desired significance of 0.05 (i.e., p^* is rejected no more than 5% of the time).

EXAMPLE 7 (AUTO LOAN RATE OPTIMIZATION). Consider again the case of Example 4. In Besbes et al. (2010), the authors consider whether a parametric model suffices for the problem of fixed pricing within various customer segments of loan applicants, defined in terms of three factors:

1. FICO score: (690, 715] (range 1) or (715, 740] (range 2),
2. Loan term in months: ≤ 36 (class 1), (36, 48] (class 2), (48, 60] (class 3), or > 60 (class 4).
3. Season: first half of data (half 1) or second half (half 2).

Customers with FICO scores outside of (690, 740] are not considered (see Besbes et al. (2010) reasoning). Term classes 2 and 4 are not considered either, but we consider these here. The authors use a per-unit profit function $r(p) = r - 2\%$. Within each segment, the authors' approach is to estimate (either parametrically or non-parametrically) the conditional expectation of demand given price and to optimize per-unit profit times this conditional expectation. Using a test that compares the parametric and non-parametric approaches, they conclude that a parametric model suffices.

We consider the same problem again here, paying closer attention to the observational nature of the data. In Example 4 we argued that even within each segment, the data cannot be treated as experimental (i.e., satisfying Assumption 1 with respect to segment alone) and therefore that

Table 4 Testing Revenue Optimality in the Auto Loan Rate Optimization Example

		FICO range 1 (690, 715]		FICO range 2 (715, 740]		
		Half 1	Half 2	Half 1	Half 2	
n		1359	732	1386	781	
ρ_n	Prescriptive, Param	0.37 (0.15)	0.21 (0.11)	0.24 (0.49)	0.23 (0.018*)	Term cl. 1
(p -val)	Predictive, Param	0.86 (0.030*)	0.50 (0.013*)	0.25 (0.48)	0.70 (< 0.001***)	
	Predictive, Non-Param	1.91 (0.0012**)	1.51 (< 0.001***)	1.6 (0.07)	1.35 (< 0.001***)	
n		1394	832	1327	690	
ρ_n	Prescriptive, Param	0.23 (0.21)	0.18 (0.073)	0.28 (0.053)	0.074 (0.67)	Term cl. 2
(p -val)	Predictive, Param	0.87 (0.015*)	0.24 (0.039*)	0.35 (0.033*)	0.051 (0.73)	
	Predictive, Non-Param	1.6 (0.0011**)	1.59 (< 0.001***)	1.19 (< 0.001***)	1.76 (0.040*)	
n		4495	3147	3803	2865	
ρ_n	Prescriptive, Param	0.55 (0.32)	0.26 (0.33)	0.088 (0.061)	1.4 (0.066)	Term cl. 3
(p -val)	Predictive, Param	1.19 (0.14)	0.22 (0.37)	0.28 (< 0.001***)	1.89 (0.034*)	
	Predictive, Non-Param	1.19 (0.14)	1.1 (0.046*)	0.86 (< 0.001***)	2.49 (0.015*)	
n		2347	1506	1834	1206	
ρ_n	Prescriptive, Param	0.40 (0.0071**)	0.0059 (0.63)	1.86 (0.30)	0.14 (0.46)	Term cl. 4
(p -val)	Predictive, Param	0.27 (0.026*)	0.045 (0.19)	2.19 (0.26)	0.31 (0.28)	
	Predictive, Non-Param	1.5 (< 0.001***)	1.54 (< 0.001***)	2.92 (0.19)	1.7 (0.012*)	

* denotes reject H_0 at significance 0.05, ** at 0.01, and *** at 0.001. Gray denotes p -value ≥ 0.05 .

Note: The data clearly distinguishes the profits generated by predictive approaches as suboptimal, rejecting the null hypothesis in all but 3 segments (non-parametric) or 7 segments (parametric) out of 16. A prescriptive approach, albeit parametric, generates profits that cannot be statistically distinguished from optimal in all but 2 segments (in which the other approaches also fail the test).

purely predictive approaches may not be estimating the true PRF. We now use our hypothesis test to determine whether this distinction is moot from a profit-generated point of view. We also test whether our parametric prescriptive approach from Section 4.2 is successful. For the predictive approaches, we reproduce those in Besbes et al. (2010): kernel regression with the Gaussian kernel (non-parametric) and logistic regression (parametric). For our parametric prescriptive approach we fit a log-normal model for price via linear regression on X , i.e., $(\log(P)|X = x) \sim \mathcal{N}(\beta_0 + \beta_1^T x, \sigma^2)$, and we fit a logistic regression for demand that is linear in price and quadratic in generalized propensity score, i.e., $\hat{d}(p, q) = \left(1 + e^{-\alpha_0 - \alpha_1 p - \alpha_2 q - \alpha_3 q^2}\right)^{-1}$.

We let X consist of FICO score, the loan amount, the loan term, whether the car is new or used, whether the loan is refinancing, and if so what was the previous rate (otherwise 0). In our experience, each of these covariates has direct impact on the interest rate quoted to applicants – and each can arguably impact the decision of the applicant to accept any one rate. At the same time, this summarizes all relevant data provided and thus encapsulates all customer-specific information that could have gone into a rate quote decision. Therefore, we reason that Assumption 1 holds with respect to X , while it is likely to fail with respect to any subset of X .

We run the test within each of the 16 customer segments. In Table 4, we report the estimated suboptimality ρ_n and its corresponding p -value according to our bootstrap procedure with $B = 100$

draws. The results overwhelmingly support the case that a predictive approach is insufficient and leaves profit on the table – the data clearly distinguishes the profits generated by these approaches from optimal in all but 3 segments (non-parametric) or 7 segments (parametric) out of 16. Our prescriptive parametric approach, on the other hand, passes the test in all but 2 segments (in each of which, the predictive approaches also failed the test). Moreover, we see that the estimated suboptimality of our prescriptive parametric approach is smaller than that of the predictive approaches in all segments (non-parametric) or all but 3 segments (parametric). Averaging the estimated suboptimalities of each approach (weighted by n) and comparing, we see that our prescriptive parametric approach recoups 70% (non-parametric) or 36% (parametric) of the total profits lost by using the suboptimal predictive approaches.

We note that in the same segments in which the analysis in Besbes et al. (2010) suggested that logistic regression on P is sufficient to estimate the PRF for optimal pricing purposes, our findings show that it is in fact insufficient. On the other hand, a parametric approach that addresses the observational nature of the data and the prescriptive nature of the problem seems to suffice for pricing in most cases, generating profits that the data cannot outright distinguish from optimal. In practice, it is known that parametric models, even if misspecified, can be helpful in extracting useful conclusions from smaller datasets. Our findings confirm this and while, refuting the evidence provided by, agree with the final conclusion of Besbes et al. (2010) that parametric approaches work well for pricing.

6. Conclusions

We studied the distinction between prediction and prescription in the context of data-driven pricing. This distinction was relevant in the case of pricing based on analytics of a corpus of observational data – a very practically relevant case. We bounded the suboptimality of predictive approaches, but when it came to solving the prescriptive problem directly we saw that the solution could not necessarily be gleaned from data. Under sufficient conditions for identifiability, we developed specific data-driven pricing schemes and a hypothesis test for profit optimality. Applying this test to data from a loan provider to study the practical relevance of the distinction between prediction and prescription, we found that predictive approaches are practically insufficient, while parametric approaches to pricing often suffice, but only when they take into full account the problem’s prescriptive nature.

References

- Berry, Steven, James Levinsohn, Ariel Pakes. 1995. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society* 841–890.
- Bertsimas, Dimitris, Vishal Gupta, Nathan Kallus. 2014. Robust SAA. *arXiv preprint arXiv:1408.4445* .

- Bertsimas, Dimitris, Nathan Kallus. 2015. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481* .
- Bertsimas, Dimitris, Georgia Perakis. 2006. Dynamic pricing: A learning approach. Siriphong Lawphongpanich, Donald W Hearn, Michael J Smith, eds., *Mathematical and Computational Models for Congestion Charging, Applied Optimization*, vol. 101. Springer, 45–79.
- Besbes, Omar, Robert Phillips, Assaf Zeevi. 2010. Testing the validity of a demand model: An operations perspective. *Manufacturing & Service Operations Management* **12**(1) 162–183.
- Besbes, Omar, Assaf Zeevi. 2009. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research* **57**(6) 1407–1420.
- Besbes, Omar, Assaf Zeevi. 2015. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science* **61**(4) 723–739.
- Bijmolt, Tammo HA, Harald J van Heerde, Rik GM Pieters. 2005. New empirical generalizations on the determinants of price elasticity. *Journal of marketing research* **42**(2) 141–156.
- Cohen, Maxime C, Ngai-Hang Z Leung, Kiran Panchangam, Georgia Perakis, Anthony Smith. 2014. The impact of linear optimization on promotion planning. *Available at SSRN 2382251* .
- Columbia University Center for Pricing and Revenue Management. 2012. Dataset cprm-12-001: On-line auto lending.
- Efron, Bradley, Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall.
- Elmaghraby, Wedad, Pinar Keskinocak. 2003. Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science* **49**(10) 1287–1309.
- Ferreira, Kris Johnson, Bin Hong Alex Lee, David Simchi-Levi. 2016. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management* **18**(1) 69–88.
- Flores, Carlos Arturo. 2005. Estimation of dose-response functions and optimal treatment doses with a continuous treatment. Ph.D. thesis, University of California, Berkeley.
- Gallego, Guillermo, Garrett Van Ryzin. 1994. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management science* **40**(8) 999–1020.
- Goldberger, Arthur S. 1972. Structural equation methods in the social sciences. *Econometrica* 979–1001.
- Greblicki, Wlodzimierz, Adam Krzyzak, Mirosław Pawlak. 1984. Distribution-free pointwise consistency of kernel regression estimate. *The Annals of Statistics* 1570–1575.
- Hansen, Bruce E. 2009. Lecture notes on nonparametrics.
- Härdle, Wolfgang. 1990. *Applied nonparametric regression*. Cambridge Univ Press.
- Harrison, J Michael, N Bora Keskin, Assaf Zeevi. 2012. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* **58**(3) 570–586.

- Hirano, Keisuke, Guido W Imbens. 2004. The propensity score with continuous treatments. Andrew Gelman, Xiao-Li Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, chap. 7. Wiley, New York, 73–84.
- Horowitz, Joel L. 2001. The bootstrap. James J Heckman, Edward Leamer, eds., *Handbook of Econometrics*, vol. 5, chap. 52. Elsevier, Amsterdam, 3159–3228.
- Imai, Kosuke, David A Van Dyk. 2004. Causal inference with general treatment regimes. *Journal of the American Statistical Association* **99**(467) 854–866.
- Kleywegt, Anton, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* **12**(2) 479–502.
- Lee, Soonhui, Tito Homem-de Mello, Anton J Kleywegt. 2012. Newsvendor-type models with decision-dependent uncertainty. *Mathematical Methods of Operations Research* **76**(2) 189–221.
- McCullagh, Peter, John A Nelder, P McCullagh. 1989. *Generalized linear models*. 2nd ed. Chapman and Hall, London.
- Nadaraya, Elizbar. 1964. On estimating regression. *Theory Probab. Appl.* **9**(1) 141–142.
- Newey, Whitney K. 1994. Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10**(02) 1–21.
- Oren, Shmuel, Stephen Smith, Robert Wilson. 1984. Pricing a product line. *Journal of Business* S73–S99.
- Parzen, Emanuel. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 1065–1076.
- Pearl, Judea. 2000. *Causality: models, reasoning and inference*. Cambridge University Press.
- Pearl, Judea. 2009. Remarks on the method of propensity score. *Statistics in Medicine* **28**(9) 1415–1416.
- Phillips, Robert. 2005. *Pricing and revenue optimization*. Stanford University Press.
- Phillips, Robert, A Serdar Simsek, G VanRyzin. 2012. Endogeneity and price sensitivity in customized pricing. *Columbia University Center for Pricing and Revenue Management Working Paper* **4**.
- Robins, James M, Miguel Angel Hernan, Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5) 550–560.
- Rosenbaum, Paul R, Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**(1) 41–55.
- Rubin, Donald B. 2009. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine* **28**(9) 1420–1423.
- Shrier, Ian. 2009. Propensity scores. *Statistics in Medicine* **28**(8) 1317–1318.
- Spirtes, Peter. 2010. Introduction to causal inference. *The Journal of Machine Learning Research* **11** 1643–1662.

Watson, Geoffrey. 1964. Smooth regression analysis. *Sankhyā A* 359–372.

Ziegler, Klaus. 2002. On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric Statistics* **14**(6) 749–774.

APPENDICES

7. Extensions

We explored the predictive-prescriptive dichotomy through a particular, simple pricing problem and presented one treatment of the issue. The ideas, however, extend and relate to a wider range of operational problems and statistical techniques. We discuss these extensions in this section.

7.1. Customized Pricing

Up to now, we have considered the problem of assigning a single price based on data, but in this data prices were potentially set based partially on observed covariates. The sole customization considered was of the form of discrete binning and splitting of the data. The single-pricing problem offered the clearest parallel to existing work and, as a building block of price optimization, provided a framework in which to study the distinction between prediction and prescription.

In this section we briefly expand our scope to the problem of customized pricing, where each price can be dependent on customer characteristics. In the full-information case, the problem is the same as the single-price problem and simply involves adjusting one's definition of the relevant population. In the data-driven case, however, the difference is that data from heterogeneous customers must be used to estimate the PRF for a particular customer either because customer characteristics are defined using continuous quantities or because there are many segments. The standing assumption will still be, as before, Assumption 1 with respect to X .

Let us consider the fully customized pricing problem where price should be customized on the basis of the full set of covariates X . That is, we are interested in the problem of choosing a unit price $p(x) \in \mathcal{P} \subset \mathbb{R}_+$ for each customer characteristic x . The hypothetical price optimization problem we would then like to solve can be expressed as follows:

$$p^*(x) \in \arg \max_{p \in \mathcal{P}} \{R(p, x) := r(p) \mathbb{E}[D(p)|X = x]\}. \quad (10)$$

Assuming customers arrive from some stationary distribution, our expected profit generated from a measurable pricing strategy $p(x)$ is

$$R(p(\cdot)) = \mathbb{E}[r(p(X))D(p(X))].$$

Note that we have

$$R(p^*(\cdot)) = \mathbb{E} \left[\max_{p \in \mathcal{P}} r(p) \mathbb{E}[D(p)|X] \right].$$

One approach to customized pricing is to estimate the customized PRF, i.e. $\mathbb{E}[D(p)|X = x]$, and optimize customized pricing with respect to it. To estimate the customized PRF, we can rely on Assumption 1. The proof of Theorem 4 argued that under Assumption 1, $\mathbb{E}[D(p)|X = x] =$

$\mathbb{E}[D|P=p, X=x]$ so that the customized PRF is given by regressing D on P and X . Since we customize the price based on X we do not average over it as we have before. (If customization were done on the basis of a subset $Y = s(X)$ of the features, we would need to average over the conditional distribution of X given Y , which would require additional estimation.)

Non-parametric approach. As before, we can use Nadaraya-Watson kernel regression to come up with a consistent non-parametric estimate for $\mathbb{E}[r(P)D|P=p, X=x]$. This is exactly what we did in deriving $\bar{R}_n(p, x)$ in eq. (7) in Section 4.1, which leads to the following non-parametric data-driven customized price prescription

$$\bar{p}_n(x) \in \arg \max_{p \in \mathcal{P}} \bar{R}_n(p, x). \quad (11)$$

Estimating the marginal distribution of X by the empirical distribution, a corresponding non-parametric estimate of the expected profit generated from a pricing strategy $p(\cdot)$ is

$$\bar{R}_n(p(\cdot)) = \frac{1}{n} \sum_{i=1}^n \bar{R}_n(p(X_i), X_i). \quad (12)$$

A hypothesis test. As before, it can be argued that for pricing purposes, the fit of a customized demand model is irrelevant insofar as the model leads to profits that cannot be discerned from optimal. We can develop a hypothesis test akin to that of Section 5, which assesses whether this is the case in the customized pricing case.

Let us consider any customized pricing scheme $\hat{p}_n(\cdot)$ and let $\hat{p}(\cdot)$ be its large-sample equivalent. That is, let us assume the following.

ASSUMPTION 6 (Convergent Customized Pricing Strategy). *For some fixed $\hat{p}(\cdot)$,*

$$\hat{p}_n(X) - \hat{p}(X) = O_p(1/\sqrt{n}).$$

We are interested in testing the hypotheses

$$H_0 : R(p^*(\cdot)) = R(\hat{p}(\cdot)),$$

$$H_1 : R(p^*(\cdot)) > R(\hat{p}(\cdot)).$$

Again, the impediment to testing this is that $R(p(\cdot))$, $p^*(\cdot)$, and $\hat{p}(\cdot)$ are all unknown; were they known, we could compute $\kappa = R(p^*(\cdot)) - R(\hat{p}(\cdot))$ and compare it to 0. A test statistic that proxies κ that uses our non-parametric estimates from the last section is

$$\kappa_n = \bar{R}_n(\bar{p}_n(\cdot)) - \bar{R}_n(\hat{p}_n(\cdot)) = \frac{1}{n} \sum_{i=1}^n (\bar{R}_n(\bar{p}_n(X_i), X_i) - \bar{R}_n(\hat{p}_n(X_i), X_i)).$$

As before, it can be shown that under appropriate conditions, our statistic diverges under H_1 and converges in distribution under H_0 , with

$$\tilde{A}_n = \frac{1}{n} \sum_{i=1}^n (\bar{R}_n(\bar{p}_n(X_i), X_i) - \bar{R}_n(p^*(X_i), X_i))$$

being the dominating term. Therefore, an approximate rejection threshold for κ_n can be gleaned from the bootstrap estimates

$$\tilde{A}_n^{(b)} = \frac{1}{n} \sum_{i=1}^n (\bar{R}_n^{(b)}(\bar{p}_n^{(b)}(X_i), X_i) - \bar{R}_n^{(b)}(\bar{p}_n(X_i), X_i)).$$

The details are beyond the scope of this paper.

7.2. Related Problems

The distinction between prediction and prescription extends to other data-driven prescriptive problems that leverage observational data. Within pricing, this includes data-driven multi-product or inventory-constrained pricing (Oren et al. 1984, Gallego and Van Ryzin 1994, Elmaghraby and Keskinocak 2003). More generally, the distinction is relevant whenever the effect of the decision being optimized on the objective is unknown and needs to be estimated from experimental data, including, e.g., newsvendor models where on-hand inventory affect demand (Lee et al. 2012). The same concepts, such as the central role of Assumption 1, extend to these problems.

Problems where the effect of the decision on the objective is known a priori are unaffected by this distinction. Consider, for example, the classic stochastic optimization problem,

$$\min_{z \in \mathcal{Z}} \mathbb{E}[c(z; Y)],$$

with decision z and random disturbance Y . Here, knowledge of the cost structure $c(z; y)$ encapsulates the decision's effect on the objective and, in data-driven settings, all that needs to be estimated from the data is the distribution of Y – e.g. via the sample average approximation (Kleywegt et al. 2002) or robust sample average approximation (Bertsimas et al. 2014). The same is true of the more intricate conditional stochastic optimization problem studied in Bertsimas and Kallus (2015),

$$\min_{z \in \mathcal{Z}} \mathbb{E}[c(z; Y) | X = x],$$

where X represents predictive observations. The cost function gives the prescriptive effect of the decision z and is assumed known. The effect of the predictive features X are to be estimated, but because they are not being optimized but only observed, their causal effect is not of interest.

7.3. Related Statistical Methods

Using $D(p)$ to denote the counterfactual result one would see if a particular intervention were used is known as the Neyman-Rubin potential outcome notation. The Neyman-Rubin framework is generally applied to binary interventions (control vs treatment), but here the interventions are prices that are potentially continuous. The Neyman-Rubin potential outcome framework is not the only framework used to describe causal relationships, although it is largely the most popular in statistics. We find that potential outcome notation fits well with the problem we explore here and also matches with familiar notation already used in other work in OR/MS, such as Lee et al. (2012), where the notation $G(q, \cdot)$ is used for the distribution of demand when the initial on-hand inventory is set to q .

Other notable frameworks for causality include structural equation models (SEM; cf. Goldberger (1972)), popular in econometrics, and Pearl’s framework of causal Bayesian networks and do-calculus (cf. Pearl (2000)), popular in epidemiology. The SEM framework may well be applied to the problem but we choose not to use it because of its need for a priori models, the common restriction to linear relationships, incompatible notation, and the less clear question of model-free identifiability, which drives our pricing solution and the nonparametric test for profit optimality.

Pearl’s framework in some senses encompasses both potential outcomes and SEM. Its dependence on directed acyclic graph (DAG) models to describe a priori causal relationships, however, makes it potentially too unwieldy for application to the problem herein and its notation and extensive nomenclature too complex for a succinct presentation. In effect, a causal DAG, correctly specified, can specify the correct subset of the covariates X that should be included in order to achieve Assumption 1. The standard practice in applications of the Neyman-Rubin framework is generally to condition on all observed covariates X that are potentially relevant (cf. Rubin (2009)), but one can come up with contrived scenarios where the inclusion of a covariate in such conditioning can (asymptotically) bias causal estimates (cf. Shrier (2009), Pearl (2009)). Because these scenarios are usually restricted to self-selection via hidden factors, rather than selection by a manager based on available data, the relevance of such concerns to the problems explored herein is limited.

8. Omitted Proofs

Proof of Theorem 1 That $d(p)$ is linear and decreasing implies that $d(p) = d_0 - \lambda(p - c)$ with $\lambda > 0$. Hence, $R(p) = d_0(p - c) - \lambda(p - c)^2$, which is unimodal and uniquely maximized at $p^* = c + d_0/(2\lambda)$ with value $R(p^*) = d_0^2/(4\lambda)$. Let $\delta(p) = \mathbb{E}[\epsilon|P = p]$, $\eta = d_0\gamma$. Then $|\delta(p)| \leq \eta$. Note that

$$\mathbb{E}[D|P = p] = \mathbb{E}[D(p)|P = p] = \mathbb{E}[d(p) + \epsilon(p)|P = p] = d(p) + \mathbb{E}[\epsilon|P = p] = d_0 - \lambda(p - c) + \delta(p).$$

Hence, the theorem is trivial if $\eta = 0$ so let us assume $\eta > 0$.

Next we ask the question, what is the largest and smallest that the maximizer \tilde{p} of $\tilde{R}(p)$ can be. By assumption, $|\delta(p)| \leq \eta$ for all $p \in \mathcal{P}$. So, defining $\tilde{R}_{\delta_0}(p) := (p - c)(d_0 - \lambda(p - c) + \delta_0(p))$, we are interested in

$$\tilde{p}_{\max} = \sup \left\{ \sup \left(\arg \max_{p \in \mathcal{P}} \tilde{R}_{\delta_0}(p) \right) : |\delta_0(p)| \leq \eta \right\}, \quad (13)$$

$$\tilde{p}_{\min} = \inf \left\{ \inf \left(\arg \max_{p \in \mathcal{P}} \tilde{R}_{\delta_0}(p) \right) : |\delta_0(p)| \leq \eta \right\}, \quad (14)$$

where we define $\sup(\emptyset) = -\infty$ and $\inf(\emptyset) = \infty$ without loss of generality because we assumed an optimizer \tilde{p} exists for $\tilde{R}(p)$ so we are only interested in those functions $\delta(p)$ that induce a nonempty argmax. In what follows, define $\tilde{R}_+(p) = (p - c)(d_0 - \lambda(p - c) + \eta)$ and $\tilde{R}_-(p) = (p - c)(d_0 - \lambda(p - c) - \eta)$, which are both unimodal and uniquely maximized at $\tilde{p}_+ = c + (d_0 + \eta)/(2\lambda)$ and $\tilde{p}_- = c + (d_0 - \eta)/(2\lambda)$ respectively ($\tilde{p}_- < \tilde{p}_+$ because $\eta > 0$). Notice that $\tilde{R}_-(p) \leq \tilde{R}_{\delta_0}(p) \leq \tilde{R}_+(p)$ whenever $|\delta_0(p)| \leq \eta$ with equality when $\delta_0(p) = \pm\eta$ is extremal.

First we argue that the bounds (13)-(14) are finite. For any $p \geq p' = c + (d_0 + \eta + 2\sqrt{\lambda + d_0\eta})/(2\lambda)$ and $|\delta_0(p)| \leq \eta$, since $\tilde{R}_+(p)$ is decreasing past \tilde{p}_+ and $p' \geq \tilde{p}_+$, we have that

$$\tilde{R}_{\delta_0}(p) \leq \tilde{R}_+(p) \leq \tilde{R}_+(p') = (d_0 - \eta)^2/(4\lambda) - 1 < (d_0 - \eta)^2/(4\lambda) = \tilde{R}_-(\tilde{p}_-) \leq \tilde{R}_{\delta_0}(\tilde{p}_-).$$

Since $\tilde{p}_- \leq p'$ we conclude that $\tilde{p}_{\max} \leq p' < \infty$. Finally, since $\delta_0(p) = 0$ is feasible in (13)-(14), we have $c \leq \tilde{p}_{\min} \leq p^* \leq \tilde{p}_{\max} \leq p'$.

Next we argue that in (13) it is sufficient to consider functions $\delta_0(p)$ taking values in $\{-\eta, +\eta\}$ that are monotonic increasing, i.e. constant or step functions. Let $\delta_0(p)$ be feasible in (13) and let $\tilde{p}_0 = \sup \left\{ \arg \max_{p \in \mathcal{P}} \tilde{R}_{\delta_0}(p) \right\}$. If at any $p_1 \geq \tilde{p}_0$ we have $\delta_0(p_1) < \eta$, then increasing the value of $\delta_0(p_1)$ to η can only increase the value of $\tilde{R}_{\delta_0}(p_1)$, which in turn may only increase the largest maximizer since $p_1 \geq \tilde{p}_0$. Moreover, if at any $p_1 < \tilde{p}_0$ we have $\delta_0(p_1) > -\eta$, then decreasing the value of $\delta_0(p_1)$ to $-\eta$ can only decrease the value of $\tilde{R}_{\delta_0}(p_1)$, which must already be at or below the maximal value and hence must leave the largest maximizer unchanged. The argument is unchanged even if \tilde{p}_0 is $\pm\infty$. A symmetric argument shows that in (14) it is sufficient to consider functions $\delta(p)$ taking values in $\pm\eta$ that are monotonic decreasing.

Next we evaluate \tilde{p}_{\max} . Fix $\tilde{p}' = c + (\sqrt{d_0} + \sqrt{\eta})^2/(2\lambda)$ and let us consider the step function $\delta_{\max}(p) = \eta\mathbb{I}[p \geq \tilde{p}'] - \eta\mathbb{I}[p < \tilde{p}']$. Since $\tilde{p}' > p_-$, $\tilde{R}_{\delta_{\max}}(p)$ is uniquely maximized on (c, \tilde{p}') at p_- , with value $\tilde{R}_{\delta_{\max}}(p_-) = \tilde{R}_-(p_-) = (d_0 - \eta)^2/(4\lambda)$. Since $\tilde{p}' > p_+$, $R_{\delta_{\max}}(p)$ is uniquely maximized on $[\tilde{p}', \infty)$ at \tilde{p}' , with value $\tilde{R}_{\delta_{\max}}(\tilde{p}') = \tilde{R}_+(\tilde{p}') = (d_0 - \eta)^2/(4\lambda)$. Hence, $\arg \max_{p \in \mathcal{P}} \tilde{R}_{\delta_{\max}}(p) = \{p_-, \tilde{p}'\}$ and $\sup\{p_-, \tilde{p}'\} = \tilde{p}'$. Now we show that it is impossible to achieve a higher maximizer with $|\delta(p)| \leq \eta$, which would lead to $\tilde{p}_{\max} = \tilde{p}'$. By our previous argument we need only consider

functions $\delta(p)$ taking values in $\pm\eta$ that are monotonic increasing. The constant functions taking values in $\pm\eta$ induce the maxima \tilde{p}_- and \tilde{p}_+ , both of which are smaller than \tilde{p}' . Next, consider any step function $\delta_0(p) = \eta\mathbb{I}[p \geq \tilde{p}_0] - \eta\mathbb{I}[p < \tilde{p}_0]$ with $\tilde{p}_0 \neq \tilde{p}'$. If $\tilde{p}_0 \leq \tilde{p}_+$ then, for any $p \neq p_+$, we have that $\tilde{R}_{\delta_0}(\tilde{p}_+) = \tilde{R}_+(\tilde{p}_+) > \tilde{R}_+(p) \geq \tilde{R}_{\delta_0}(p)$ since \tilde{p}_+ is the unique maximizer of $\tilde{R}_+(p)$; hence $\tilde{p}_+ < \tilde{p}'$ is the unique maximum of $\tilde{R}_{\delta_0}(p)$. Consider $\tilde{p}_0 > \tilde{p}_+$. Then, since $\tilde{p}_0 > p_+ > p_-$, $\tilde{R}_{\delta_{\max}}(p)$ is uniquely maximized on (c, \tilde{p}_0) at p_- , with value $\tilde{R}_{\delta_0}(p_-) = \tilde{R}_-(p_-) = (d_0 - \eta)^2 / (4\lambda)$. Since $\tilde{p}_0 > p_+$, $\tilde{R}_{\delta_{\max}}(p)$ is uniquely maximized on $[\tilde{p}_0, \infty)$ at \tilde{p}_0 , with value $\tilde{R}_{\delta_0}(\tilde{p}_0) = \tilde{R}_+(\tilde{p}_0)$. If $\tilde{p}_0 < \tilde{p}'$, then either of these potential maximizers are smaller than \tilde{p}' . If $\tilde{p}_0 > \tilde{p}'$ then, since $\tilde{R}_+(p)$ is strictly decreasing past p_+ and $\tilde{p}' \geq p_+$, we have $\tilde{R}_{\delta_0}(\tilde{p}_0) = \tilde{R}_+(\tilde{p}_0) < \tilde{R}_+(\tilde{p}') = (d_0 - \eta)^2 / (4\lambda) = \tilde{R}_-(p_-) = \tilde{R}_{\delta_{\max}}(p_-)$. Hence $\tilde{p}_- < \tilde{p}'$ is the unique maximum of $\tilde{R}_{\delta_0}(p)$. When $\eta < d_0$, a symmetric argument applied to (14) shows that $\tilde{p}_{\min} = c + (\sqrt{d_0} - \sqrt{\eta})^2 / (2\lambda)$. If $\eta \geq d_0$, the lower bound $\tilde{p}_{\min} = c$ is achieved by $\delta_-(p)$. Hence, $\tilde{p}_{\min} = c + \max\{0, \sqrt{d_0} - \sqrt{\eta}\}^2 / (2\lambda)$.

To summarize, we conclude that since $|\mathbb{E}[\epsilon|P]| \leq \eta$, we must have

$$\tilde{p} \in [\tilde{p}_{\min}, \tilde{p}_{\max}] \quad \text{where} \quad \tilde{p}_{\min} = c + \frac{\max\{0, \sqrt{d_0} - \sqrt{\eta}\}^2}{2\lambda}, \quad \tilde{p}_{\max} = c + \frac{(\sqrt{d_0} + \sqrt{\eta})^2}{2\lambda}.$$

Plugging these bounds into $R(p)$ we have

$$R(\tilde{p}_{\max}) = \frac{d_0^2 - 4d_0\eta - \eta^2 - 4\eta\sqrt{d_0\eta}}{4\lambda}, \quad R(\tilde{p}_{\min}) = \begin{cases} \frac{d_0^2 - 4d_0\eta - \eta^2 + 4\eta\sqrt{d_0\eta}}{4\lambda} & \eta < d_0 \\ 0 & \eta \geq d_0 \end{cases}$$

Notice that if $\eta < d_0$ then $R(\tilde{p}_{\max}) = R(\tilde{p}_{\min}) - 2\eta\sqrt{d_0\eta}/\lambda \leq R(\tilde{p}_{\min})$ and if $\eta \geq d_0$ then $R(\tilde{p}_{\max}) \leq 0 = R(\tilde{p}_{\min})$. Therefore, $\min\{R(\tilde{p}_{\max}), R(\tilde{p}_{\min})\} = R(\tilde{p}_{\max})$. Since $R(p)$ is unimodal and $\tilde{p} \in [\tilde{p}_{\min}, \tilde{p}_{\max}]$, we have

$$R(\tilde{p}) \geq \min\{R(\tilde{p}_{\max}), R(\tilde{p}_{\min})\} = R(\tilde{p}_{\max}) = \frac{d_0^2 - 4d_0\eta - \eta^2 - 4\eta\sqrt{d_0\eta}}{4\lambda}.$$

Finally, using $R(p^*) = d_0^2 / (4\lambda)$,

$$\frac{R(\tilde{p})}{R(p^*)} \geq 1 - 4 \left(\frac{\eta}{d_0} \right) - 4 \left(\frac{\eta}{d_0} \right)^{3/2} - \left(\frac{\eta}{d_0} \right)^2,$$

which is a univariate polynomial in $\sqrt{\eta/d_0}$. □

Proof of Theorem 2 Recall that $\mathbb{E}[\epsilon] = 0$. That ϵ and P are jointly normal implies that

$$\mathbb{E}[\epsilon|P] = \zeta(P - \mu), \quad \text{where } \mu = \mathbb{E}[P], \quad \zeta = \frac{\text{Cov}(\epsilon, P)}{\text{Var}(P)}.$$

By assumption of non-positive correlation, $\zeta \leq 0$.

That $d(p)$ is linear and decreasing implies that $d(p) = d_0 - \lambda(p - c)$ with $\lambda > 0$. Hence, $R(p) = d_0(p - c) - \lambda(p - c)^2$, which is unimodal and uniquely maximized at $p^* = c + d_0 / (2\lambda)$ with value

$R(p^*) = d_0^2/(4\lambda)$. Also, recall from the proof of Theorem 1 that $\mathbb{E}[D|P=p] = \mathbb{E}[D(p)|P=p] = \mathbb{E}[d(p) + \epsilon(p)|P=p] = d(p) + \mathbb{E}[\epsilon|P=p]$ and hence $\tilde{R}(p) = d_0(p-c) - \lambda(p-c)^2 + \zeta(p-c)(p-\mu)$, which is unimodal and uniquely maximized at its critical point $\tilde{p} = (2\lambda c + d_0 - \zeta(c + \mu))/(2\lambda - 2\zeta)$ because it is feasible since $\lambda > 0 \geq \zeta$ and $(2\lambda - 2\zeta)(\tilde{p} - c) = d_0 - \zeta(\mu - c) \geq d_0 \geq 0$.

Plugging \tilde{p} into $R(p)$ we get

$$R(\tilde{p}) = \frac{(d_0 - \zeta(\mu - c))(d_0(\lambda - 2\zeta) + \lambda\zeta(\mu - c))}{4(\lambda - \zeta)^2}.$$

Rearranging and using $R(p^*) = d_0^2/(4\lambda)$, we have

$$\frac{R(\tilde{p})}{R(p^*)} = 1 - \left(\frac{\zeta}{\lambda - \zeta}\right)^2 \left(\frac{d_0 + \lambda(c - \mu)}{d_0}\right)^2 = 1 - \left(\frac{\zeta}{\lambda - \zeta}\right)^2 \left(\frac{\mathbb{E}[D]}{d_0}\right)^2,$$

where we plugged in $\mathbb{E}[D] = \mathbb{E}[D(p)] = \mathbb{E}[d_0 - \lambda(P - c)] = d_0 - \lambda(\mu - c)$. Moreover,

$$\sup_{\zeta \leq 0} \left(\frac{\zeta}{\lambda - \zeta}\right)^2 = \lim_{\zeta \rightarrow -\infty} \left(\frac{\zeta}{\lambda - \zeta}\right)^2 = 1.$$

Hence, we have the result in the statement of the theorem. \square

Proof of Theorem 5 Assumption 1 gives $R(p) = \mathbb{E}[\mathbb{E}[r(P)D|P=p, X]]$, i.e. profit is given by taking a partial mean with $P = p$ fixed of the regression of $r(P)D$ on P and X .

By Assumption 4 part i, there exists $\delta > 0$ such that $[p^* - \delta, p^* + \delta]$ is contained inside the support of P . By Assumption 4 part i, $f_{P,X}(p, x)$ is bounded away from 0 on its support, and, by Assumption 4 part v, $\frac{\partial f_{P,X}(p, x)}{\partial x}$ is bounded. Hence,

$$\left| \frac{\partial}{\partial x} \log(f_{P,X}(p, x)) \right| = \frac{\left| \frac{\partial f_{P,X}(p, x)}{\partial x} \right|}{f_{P,X}(p, x)} \leq L < \infty$$

on the support of (P, X) . Therefore, we have that, for any x and $|\psi| \leq \delta$,

$$\log(f_{P,X}(p^* + \psi, x)) \leq \log(f_{P,X}(p^*, x)) + L\delta,$$

and consequently,

$$\int \sup_{|\psi| \leq \delta} f_{P,X}(p^* + \psi, x) dx \leq \int e^{L\delta} f_{P,X}(p^*, x) dx < \infty.$$

By Assumption 4 part iii, there exists M such that $\mathbb{E}[D^4|P=p, X=x] \leq M$ for all p, x . Combined, this yields

$$\begin{aligned} \int \sup_{|\psi| \leq \delta} (1 + \mathbb{E}[(r(P)D)^4|P=p^* + \psi, X=x]) f_{P,X}(p^* + \psi, x) dx \\ \leq (1 + r(p^* + \delta)^4 M) \int \sup_{|\psi| \leq \delta} f_{P,X}(p^* + \psi, x) dx < \infty. \end{aligned} \quad (15)$$

To study our profit function estimator (8) for each fixed p , we employ Theorem 4.1 of Newey (1994) (henceforth, N in this proof), which provides convergence results for two-step kernel m -estimators, a specific case of which is our profit function estimator. We let the “trimming function” of N be $\tau(x) = \mathbb{I}[f_X(x) > 0]$, an indicator for the compact support of X (where its density is assumed bounded away from zero). Since our estimator (8) has K both in the numerator and denominator, it is unchanged if we rescale K by a positive constant. Similarly, the conditions of Assumption 2 remain unchanged. Hence, by Assumption 2 part i, without loss of generality we may assume $\int_{\mathbb{R}^{1+k}} K = 1$. Then, Assumption K of N is satisfied with $\Delta = 2$ by Assumption 2 parts i-iv. Assumption H of N is satisfied with $d = s + 1$ by Assumption 4 part v. These constitute condition (ii) of N’s Theorem 4.1. By Assumption 4 part i, there exists $c < p_{\max} < \infty$ such that $P \leq p_{\max}$ almost surely. Let $r_{\max} = r(p_{\max}) < \infty$. Then, by Assumption 4 part iii, $\mathbb{E}[(r(P)D)^4] \leq r_{\max} \mathbb{E}[D^4] < \infty$ and $\mathbb{E}[(r(P)D)^4 | P = p, X = x] = r(p) \mathbb{E}[D^4 | P = p, X = x]$ are bounded. Combined with Assumption 4 part ii, we satisfy condition (i) of N’s Theorem 4.1. Condition (iii) of N’s Theorem 4.1 is satisfied by our choice of $\tau(\cdot)$ and by Assumption 4 part i. The first clause of condition (iv) of N’s Theorem 4.1 is satisfied by our choice of $\tau(\cdot)$ and by Assumption 4 part ii. The second clause is satisfied by Assumption 4 parts iv-v combined with the fact that for any $m > 0$, $\mathbb{E}[(r(P)D)^m | P = p, X = x] = r(p)^m \mathbb{E}[D^m | P = p, X = x]$ and $r(p)^m$ is continuous. The third clause is satisfied by (15). Since $X \in \mathbb{R}^k$ and $P \in \mathbb{R}$, condition (v) of N’s Theorem 4.1 is satisfied by Assumption 2 parts v-vi. Applying N’s Theorem 4.1 for each fixed $p \in \mathcal{P}$, we get

$$\sqrt{nh_n}(R(p) - \bar{R}_n(p)) \xrightarrow{d} \mathcal{N}(0, \eta_p \kappa) \quad \forall p \in \mathcal{P},$$

where $\eta_p \kappa$ is a simplification of the asymptotic variance in eq. (14) in N.

To study the optimizer of our profit function estimator, we employ Flores (2005) (henceforth, F in this proof). Conditions (ii-vi) of F’s Theorem 3 are satisfied in a similar way to the case of N’s Theorem 4.1. Condition (i) of F’s Theorem 3 is satisfied by Assumption 3 parts i and iii, condition (vii) by Assumption 4 part v, condition (viii) by Assumption 3 part iv, condition (ix) by Assumption 3 parts ii and iv, and finally condition (x) by Assumption 2 parts v-vi. Applying F’s Theorem 3, we get

$$\sqrt{nh_n^3}(p^* - \bar{p}_n) \xrightarrow{d} \mathcal{N}\left(0, \frac{\eta_{p^*} \kappa'}{R''(p^*)^2}\right), \quad (16)$$

simplifying the asymptotic variance.

By Assumption 3 part iv and using Taylor’s theorem to expand $R(p)$ around $p = p^*$, there exists $p_n \in [\min(p^*, \bar{p}_n), \max(p^*, \bar{p}_n)]$ such that

$$R(\bar{p}_n) = R(p^*) + R'(p^*)(\bar{p}_n - p^*) + \frac{1}{2}R''(p_n)(\bar{p}_n - p^*)^2.$$

By first order optimality conditions, $R'(p^*) = 0$. Hence, rearranging, we have

$$R(p^*) - R(\bar{p}_n) = -\frac{1}{2}R''(p_n)(\bar{p}_n - p^*)^2. \quad (17)$$

By continuous transformation of eq. (16), we have

$$(nh_n^3)(\bar{p}_n - p^*)^2 \xrightarrow{d} \frac{\eta_{p^*}\kappa'}{R''(p^*)^2}\chi_1^2. \quad (18)$$

Eq. (16) also implies $\bar{p}_n \xrightarrow{\mathbb{P}} p^*$, which also implies $p_n \xrightarrow{\mathbb{P}} p^*$ since p_n is sandwiched between \bar{p}_n and p^* . Since $R''(p)$ is continuous, we also get by continuous transformation that

$$R''(p_n) \xrightarrow{\mathbb{P}} R''(p^*). \quad (19)$$

Combining eqs. (17)-(19), we get the desired result,

$$(nh_n^3)(R(p^*) - R(\bar{p}_n)) \xrightarrow{d} \frac{-\eta_{p^*}\kappa'}{2R''(p^*)}\chi_1^2.$$

If $nh_n^{2s+1} \rightarrow 0$, then we also satisfy the conditions of F's Theorem 4 with equal bandwidths. Applying F's Theorem 4, we get

$$\sqrt{nh_n}(R(p^*) - \bar{R}_n(\bar{p}_n)) \xrightarrow{d} \mathcal{N}(0, \eta_{p^*}\kappa),$$

simplifying the asymptotic variance. □

Proof of Theorem 6 We begin by showing that $D(p) \perp P|f_{P|X}(p, X)$. On the one hand we have

$$\begin{aligned} f_{P|f_{P|X}(p, X)}(p|q) &= \mathbb{E}[\delta(P - p)|f_{P|X}(p, X) = q] \\ &= \mathbb{E}[\mathbb{E}[\delta(P - p)|f_{P|X}(p, X) = q, X]|f_{P|X}(p, X) = q] \\ &= \mathbb{E}[\mathbb{E}[\delta(P - p)|X]|f_{P|X}(p, X) = q] \\ &= \mathbb{E}[f_{P|X}(p, X)|f_{P|X}(p, X) = q] \\ &= q. \end{aligned}$$

On the other hand, using Assumption 1, we have

$$\begin{aligned} f_{P|f_{P|X}(p, X), D(p)}(p|q, d) &= \mathbb{E}[\delta(P - p)|f_{P|X}(p, X) = q, D(p) = d] \\ &= \mathbb{E}[\mathbb{E}[\delta(P - p)|f_{P|X}(p, X) = q, D(p) = d, X]|f_{P|X}(p, X) = q, D(p) = d] \\ &= \mathbb{E}[\mathbb{E}[\delta(P - p)|D(p) = d, X]|f_{P|X}(p, X) = q, D(p) = d] \\ &= \mathbb{E}[\mathbb{E}[\delta(P - p)|X]|f_{P|X}(p, X) = q, D(p) = d] \\ &= \mathbb{E}[f_{P|X}(p, X)|f_{P|X}(p, X) = q, D(p) = d] \\ &= q. \end{aligned}$$

Equality between the two conditional probabilities implies the desired independence.

Using this independence and then plugging in $P = p$, we have

$$\mathbb{E} [D(p) | f_{P|X}(p, X) = q] = \mathbb{E} [D(p) | P = p, f_{P|X}(p, X)] = \mathbb{E} [D | P = p, Q = q] = d(p, q).$$

By iterated expectations, we get

$$\mathbb{E} [D(p)] = \mathbb{E} [\mathbb{E} [D(p) | f_{P|X}(p, X)]] = \mathbb{E} [d(p, f_{P|X}(p, X))]$$

as desired. \square

Proof of Theorem 7 The proof borrows the outline of the proof of Theorem 2 of Besbes et al. (2010), but applied to our new testing case and causal estimators.

Decompose the test statistic ρ_n into three terms:

$$\rho_n = \bar{R}_n(\bar{p}_n) - \bar{R}_n(\hat{p}_n) = A_n + B_n + C_n,$$

where

$$\begin{aligned} A_n &= \bar{R}_n(\bar{p}_n) - \bar{R}_n(p^*), \\ B_n &= \bar{R}_n(p^*) - \bar{R}_n(\hat{p}), \\ C_n &= \bar{R}_n(\hat{p}) - \bar{R}_n(\hat{p}_n). \end{aligned}$$

We begin by showing that $(nh_n^3) A_n \xrightarrow{d} \Gamma \chi_1^2$. By Assumption 2 part iii, we have that $\bar{R}_n(p)$ is twice continuously differentiable. Thus, using Taylor's theorem to expand $\bar{R}_n(p)$ around $p = \bar{p}_n$, we get that there exists $p_n \in [\min(p^*, \bar{p}_n), \max(p^*, \bar{p}_n)]$ such that

$$\bar{R}_n(p^*) = \bar{R}_n(\bar{p}_n) + \bar{R}'_n(\bar{p}_n)(p^* - \bar{p}_n) + \frac{1}{2} \bar{R}''_n(p_n)(p^* - \bar{p}_n)^2.$$

By first order optimality conditions, $\bar{R}'_n(\bar{p}_n) = 0$. Hence, rearranging, we have

$$A_n = -\frac{1}{2} \bar{R}''_n(p_n)(p^* - \bar{p}_n)^2. \quad (20)$$

Next we show that $\bar{R}''_n(p_n) \xrightarrow{\mathbb{P}} R''(p^*)$. Note that

$$\left| \bar{R}''_n(p_n) - R''(p^*) \right| \leq \left| \bar{R}''_n(p_n) - R''(p_n) \right| + |R''(p_n) - R''(p^*)|. \quad (21)$$

As in the proof of Theorem 5, Assumptions 2, 3, and 4 imply the assumptions of Lemma 5.1 of Newey (1994) applied to $R''(p)$, which in turn yields the uniform convergence in probability of $\bar{R}''_n(p)$ over \mathcal{P} since \mathcal{P} is compact by Assumption 3 part i. Hence,

$$\left| \bar{R}''_n(p_n) - R''(p_n) \right| \leq \sup_{p \in \mathcal{P}} \left| \bar{R}''_n(p) - R''(p) \right| \xrightarrow{\mathbb{P}} 0. \quad (22)$$

By Theorem 5, $\bar{p}_n \xrightarrow{\mathbb{P}} p^*$. Because p_n is sandwiched between \bar{p}_n and p^* , we also get $p_n \xrightarrow{\mathbb{P}} p^*$. Since $R''(p)$ is continuous by Assumption 3 part iv, we have

$$|R''(p_n) - R''(p^*)| \xrightarrow{\mathbb{P}} 0 \quad (23)$$

by continuous transformation of the former. Combining eqs. (21)-(23), we get

$$\bar{R}_n''(p_n) \xrightarrow{\mathbb{P}} R''(p^*). \quad (24)$$

By continuous transformation of the result of Theorem 5 (eq. (16)), we have

$$(nh_n^3) (\bar{p}_n - p^*)^2 \xrightarrow{d} \frac{\eta_{p^*} \kappa'}{R''(p^*)^2} \chi_1^2. \quad (25)$$

Combining eqs. (20)-(25), we get

$$(nh_n^3) A_n \xrightarrow{d} \frac{-\eta_{p^*} \kappa'}{2R''(p^*)} \chi_1^2 = \Gamma \chi_1^2. \quad (26)$$

Next, we show that $(nh_n^3) C_n \xrightarrow{\mathbb{P}} 0$. By Assumption 2 part iii, we have that $\bar{R}_n(p)$ is twice continuously differentiable. Thus, using Taylor's theorem to expand $\bar{R}_n(p)$ around $p = \hat{p}$, we get that there exists $p'_n \in [\min(\hat{p}, \hat{p}_n), \max(\hat{p}, \hat{p}_n)]$ such that

$$\bar{R}_n(\hat{p}_n) = \bar{R}_n(\hat{p}) + \bar{R}_n'(\hat{p})(\hat{p}_n - \hat{p}) + \frac{1}{2} \bar{R}_n''(p'_n)(\hat{p}_n - \hat{p})^2.$$

Rearranging, we have

$$(nh_n^3) C_n = -h_n^{3/2} \left(\sqrt{nh_n^3} \bar{R}_n'(\hat{p}) \right) (\sqrt{n}(\hat{p}_n - \hat{p})) - \frac{1}{2} h_n^3 \bar{R}_n''(p'_n) (\sqrt{n}(\hat{p}_n - \hat{p}))^2. \quad (27)$$

By Assumption 5, we have that

$$\sqrt{n}(\hat{p}_n - \hat{p}) = O_p(1), \text{ and hence also } (\sqrt{n}(\hat{p}_n - \hat{p}))^2 = O_p(1). \quad (28)$$

Applying Theorem 4 of Newey (1994) we get the convergence in distribution of $\sqrt{nh_n^3} (\bar{R}_n'(p) - R'(p))$ for any fixed p , including \hat{p} and hence we have

$$\sqrt{nh_n^3} \bar{R}_n'(\hat{p}) = O_p(1). \quad (29)$$

Next we show that $\bar{R}_n''(p'_n) = O_p(1)$. Note that

$$\left| \bar{R}_n''(p'_n) - R''(\hat{p}) \right| \leq \left| \bar{R}_n''(p'_n) - R''(p'_n) \right| + |R''(p'_n) - R''(\hat{p})|. \quad (30)$$

As before, $\bar{R}_n''(p)$ converges uniformly to $R''(p)$ in probability over \mathcal{P} and so

$$\left| \bar{R}_n''(p'_n) - R''(p'_n) \right| \leq \sup_{p \in \mathcal{P}} \left| \bar{R}_n''(p) - R''(p) \right| \xrightarrow{\mathbb{P}} 0. \quad (31)$$

By Assumption 5, $\hat{p}_n \xrightarrow{\mathbb{P}} \hat{p}$. Because p'_n is sandwiched between \hat{p}_n and \hat{p} , we also get $p'_n \xrightarrow{\mathbb{P}} \hat{p}$. Since $R''(p)$ is continuous by Assumption 3 part iv, we have

$$|R''(p'_n) - R''(\hat{p})| \xrightarrow{\mathbb{P}} 0 \quad (32)$$

by continuous transformation of the former. Combining eqs. (30)-(32), we get

$$\overline{R}_n''(p'_n) \xrightarrow{\mathbb{P}} R''(\hat{p}). \quad (33)$$

Combining eqs. (27)-(33) gives $(nh_n^3)C_n = -h_n^{3/2}O_p(1) - h_n^3O_p(1)$. Hence, because $h_n \rightarrow 0$, we get $(nh_n^3)C_n \xrightarrow{\mathbb{P}} 0$.

Finally, we treat B_n . Under H_0 , $B_n = 0$ because Assumption 3 part ii (unique optimizer) and H_0 ($R(p^*) = R(\hat{p})$) imply that $p^* = \hat{p}$. Next, we show that under H_1 , $(nh_n^3)B_n \xrightarrow{\mathbb{P}} \infty$. By applying the first results of Theorem 5 to each term, we have that $B_n \xrightarrow{\mathbb{P}} R(p^*) - R(\hat{p})$. Since $k \geq 0$, Assumption 2 part vi implies $nh_n^5/\log(n) \rightarrow \infty$, which, since we also assume $h_n \rightarrow 0$, implies $nh_n^3 \rightarrow \infty$. Hence, since $R(p^*) - R(\hat{p}) > 0$ under H_1 , we have that $(nh_n^3)B_n \xrightarrow{\mathbb{P}} \infty$. \square

Proof of Theorem 8 Proven above. See eq. (26). \square