Structural Estimation of the Newsvendor Model: An Application to Reserving Operating Room Time*

Marcelo Olivares · Christian Terwiesch · Lydia Cassorla
The Wharton School, University of Pennsylvania, Philadelphia PA, 19104
Department of Anesthesia and Perioperative Care, University of California, San Francisco CA, 94143
maolivar@wharton.upenn.edu · terwiesch@wharton.upenn.edu · cassorla@anesthesia.ucsf.edu

April 06, 2007

Abstract

The Newsvendor model captures the trade-off faced by a decision maker that needs to place a firm bet prior to the occurence of a random event. Previous research in Operations Management has mostly focused on deriving the decision that minimizes the expected mismatch costs. In contrast, we present two methods that estimate the unobservable cost parameters characterizing the mismatch cost function. We present a structural estimation framework that accounts for heterogeneity in the uncertainty faced by the newsvendor as well as in the cost parameters. We develop statistical methods that give consistent estimates of the model primitives, and derive their asymptotic distribution, which is useful to do hypothesis testing. We apply our econometric model to a hospital that balances the costs of reserving too much vs. too little operating room capacity to cardiac surgery cases. Our results reveal that the hospital places more emphasis on the tangible costs of having idle capacity than on the costs of schedule overrun and long working hours for the staff. We also extend our structural models to incorporate external information on forecasting biases and mismatch costs reported by the medical literature. Our analysis suggests that over-confidence and incentive conflicts are important drivers of the frequency of schedule overruns observed in our sample.

Keywords: Newsvendor model, empirical research, health care.

*We thank Gerard Cachon, Justin Ren, Elena Krasnokutskaya, Serguei Savin and three anonymous referees for their helpful ideas and comments.

1. Introduction

Many business decisions require that a decision maker takes a firm decision before the occurrence of a random event. Once the uncertainty is resolved, the decision maker observes that her decision was too "large" or too "small" and incurs costs reflecting the mismatch between her decision and the ex-post optimal decision. The Newsvendor model captures this trade-off. In Operations Management, the most frequently analyzed application of this type of decision deals with placing an inventory order in the presence of demand uncertainty.

While previous research related to the Newsvendor model has taken the mismatch cost parameters as given and has minimized the expected cost to obtain a cost minimizing decision, we take a different approach. Following the tradition of structural estimation models in Econometrics, we assume that the decision maker acts rationally and chooses the optimal decision for a cost function that is unobservable for us as researchers. Based on the observed decision making and a set of covariates, we use Maximum Likelihood estimation to obtain the cost parameters describing the latent cost function. For example, we can use this method to estimate how much value a retailer assigns to a stock-out.

In the Econometrics community, similar approaches have been taken by Rust (1987) and Berry et al. (1995). Rust (1987) combines a Markov decision process describing a maintenance problem with the empirically observed behavior of the person in charge of managing the maintenance to impute costs of regular cost maintenance and perceived costs of unexpected failures. Berry et al. (1995) use data from the automotive industry to estimate model markups based on a oligopoly model of price competition in a differentiated product market. We contribute to this stream of literature by providing estimable structural models of newsvendor type-decisions, which can be used in a broad variety of managerial applications.

In addition to deriving the estimation procedures for Newsvendor-like cost models as well as establishing their econometric properties, we apply our theory to a healthcare application. In the setting we studied, the hospital had to reserve a certain amount of

operating room (OR) time to specific cardiac procedures. Since the actual procedure time in the OR is random and will - in the best of all cases - vary around the expected procedure time, some procedures will exceed the forecasted durations while others will be completed ahead of schedule. If the hospital reserves too much time to a case, the OR is likely to incur excessive idle time. If, however, the hospital reserves too little time to a case, the hospital is likely to face schedule over-runs and decreased service quality. (See Strum et al. (1999) for an application of the Newsvendor model in OR management).

Our econometric models and its healthcare application enable us to make the following contributions. **First**, we extend the long line of Newsvendor research by developing a theory that allows for an estimate of the underlying cost function. We present two model specifications, both of which are sufficiently general to capture cost and demand heterogeneity and hence have the potential to be applied to various Operations Management decisions. **Second,** for each model, we derive a two-step estimation procedure and establish its key econometric properties, including the asymptotic distribution of the estimators and the associated standard errors required for hypothesis testing. **Third**, we apply our econometric framework to a healthcare setting. We analyze how a hospital balances the costs of reserving too much vs. too little OR capacity to individual cardiac surgery cases and demonstrate that our model has significant predictive power for this decision.

2. The Newsvendor Model and Structural Estimation

The newsvendor is a simple and intuitive model and is arguably one of the stepping stones for decision making in Operations Management. The model is defined as follows. Given a random variable D with distribution $F(\cdot)$, a decision maker (hereon the newsvendor) needs to make a decision Q, before the realization of the random variable D is known. The objective of the newsvendor is to minimize the expected mismatch cost between D and Q. This mismatch cost is assumed to be linear in the amount of the mismatch but typically is not symmetric. If the newsvendor's decision Q exceeds D, the incurred cost is equal to $C_o\left(Q-D\right)^+$, where $(x)^+ = \max\left\{x, 0\right\}$. If D exceeds Q, the incurred cost is equal to $C_u\left(D-Q\right)^+$. The model parameters C_o and C_u are referred to as the overage and underage cost, respectively, and are assumed to be strictly positive. The optimal decision

Q^* solves:

$$\min_Q E\left\{C_o\left(Q-D\right)^+ + C_u\left(D-Q\right)^+\right\} \tag{1}$$

We assume that the problem is unconstrained and therefore Q can take any value in the real domain. Since the objective function defined in (1) is convex in Q, the optimal solution can be characterized by the first order condition. If the random variable D is continuous, the optimal solution to (1) satisfies:

$$F(Q^*) = \frac{C_u}{C_u + C_o} = \frac{1}{1+\gamma} \tag{2}$$

where $\gamma = C_o/C_u$ is the ratio between overage and underage costs (see e.g. Porteus (2002)).

In this paper, we extend the long tradition of Newsvendor research by developing an econometric framework to impute the cost parameters of a newsvendor based on observed decisions. In our decision problem, a traditional Operations Research model takes the distribution function F and the cost parameters C_o and C_u as model input and then characterizes the optimal reservation decision Q^* (see Figure 1, left). Unlike the Operations Researcher, who is interested in providing a normative theory of how rational agents "should" behave, the Econometrician is interested in a descriptive theory how real-world decision makers actually do behave. There are two econometric approaches towards developing such a theory. In a method known as *reduced form estimation*, researchers collect data on a dependent variable of interest and use regression analysis (or other statistical methods) to explain its variation through a set of explanatory variables. In our decision problem, such models might take the amount of underage, $(D-Q)^+$, as a dependent variable and attempt to explain it through a set of explanatory variables. The outcome of this estimation would be a set of parameters characterizing the marginal impact of an explanatory variable on the amount of overage (see Figure 1, middle). This approach has been the dominant in empirical research in Operations Management (e.g. Lieberman and Demeester (1999), Brush and Karnani (1996)), in empirical work in healthcare operations management (e.g. Milne et al. (1989), Chorba (1976)) and is also widely used in the medical community (e.g. Pell et al. (2001), Urbach et al. (2003)).

In contrast to reduced form estimation, *structural estimation* first builds a decision model of the situation, similar to the ones used in Operations Research. When using structural estimation, the Econometrician assumes that decision makers already act rationally (and thereby optimally) and then uses observed decision making behavior (in our case the reservation decision Q) to impute the underlying parameters of the decision model for which this behavior is rational (see Figure 1, right). Structural estimation has actively been used in several fields of Economics, including Labour Economics and Industrial Organization[1]. With the exception of the work by Cohen et al. (2003), that estimates the cost parameters of a supplier in a semiconductor manufacturing context, structural estimation has had very few applications in the Operations Management literature.

The application that we present relates to managing capacity in a healthcare environment. An important stream of the Operations Research literature has created a number of tools that directly or indirectly relate to the management of healthcare capacity and its utilization (see Green (2004) for an overview). Given that patient demand for healthcare services is inherently uncertain, the newsvendor model has found interesting ground for its application. At the strategic level, decisions need to be made with respect to sizing the care capacity. This includes choosing occupancy rates (e.g., Smith-Daniels et al. (1988), Huang (1995), Green and Nguyen (2001)), making staffing decisions (e.g., Aiken et al. (2002), Kwak and Lee (1997), Green and Meissner (2002)) and choosing the right panel size for physicians (Green and Savin (2005)). At the tactical level, decisions need to be made with respect to allocating capacity to various demand types (e.g. Green et al. (2003)), such as the allocation of operating room time to services in a hospital (Strum et al. (1997)). Several of these decisions resemble the Newsvendor model and will be discussed more explicitly in Section 4.

3. Econometric Framework

The first order condition (2), which defines the optimal decision Q^*, is essential to our imputed cost framework. This equation, commonly referred to as the critical fractile solution, provides a direct relationship between the overage/underage cost ratio and

[1] See Reiss and Wolak (2006) for a review of structural estimation in Industrial Organization.

the probability of overestimating D at the optimum. Suppose we observe a sequence $(D_i, Q_i)_{i=1..n}$ of realizations of the random variable D and the observed decision Q made prior to each realization. We can then define the fraction of cases in which overage costs were incurred as $\bar{I} = \frac{1}{n}\sum_{i=1}^{n} 1\{D_i \leq Q_i\}$, where $1\{\cdot\}$ denotes the indicator function. $\bar{I}$ provides a crude estimate of the probability of overage of the newsvendor. Asuming that the newsvendor is behaving rationally and that the overage/underage ratio is constant among all observations $i = 1..n$, we can replace $\bar{I}$ for $F(Q^*)$ in (2) to obtain:

$$\frac{1}{1+\gamma} = \bar{I} \tag{3}$$

which gives $\gamma = \frac{1}{\bar{I}} - 1$, or equivalently, $C_o = \left(\frac{1}{\bar{I}} - 1\right) C_u$.

While the analysis outlined by equation (3) is useful as a preliminary data analysis, it suffers from three important problems. First, it does not allow for any statistical tests, which raises the question of whether the cost of overage is larger than the cost of underage (or vice-versa) with statistical significance. Second, this approach ignores the underlying heterogeneity of random component D and the ability of the newsvendor partially anticipate this heterogeneity. For example, there might be seasonal variation in demand, which changes the distribution D across observations. Since the newsvendor can anticipate this seasonal demand changes before choosing the quantity Q, she will adjust his quantity accordingly. Third, the cost ratio, $\gamma = C_o/C_u$, could vary across observations. In a retail example, the cost of a lost sale (which is related to the underage cost C_u) might vary with the time of the year or with the margin of the product. Moreover, some factors may affect both the distribution of the random variable and the cost ratio γ at the same time (e.g. price changes affect the demand distribution and the cost of underage faced by a retailer). Thus, a more elaborate model is needed. Below, we develop an estimation framework for the newsvendor problem which incorporates heterogeneity in both, the random component, D_i, and the overage/underage cost ratio, γ. Figure 2 illustrates this framework, which is sufficiently general to estimate different applications of newsvendor problems.

The upper right part of Figure 2 accounts for heterogeneity in the random component D_i. Specifically, we assume that the D_i's are given by independent random variables from

a common family of distributions $\{F(\cdot;\theta) : \theta \in \Theta\}$, where θ is a vector parameter from the parameter space Θ which characterizes each member of the class. The distribution of the duration of procedure i, denoted D_i, is given by $F(\cdot;\theta_i)$. We let this distribution depend on a vector of covariates X_i, which can include different kinds of variables depending on the context. Following common econometric practice (e.g. Bickel and Doksum (2001)), we assume the functional form:

$$\theta_i = h(X_i, \eta) \tag{4}$$

where η is a vector of parameters to be estimated. Throughout the paper, we will denote covariates (e.g. X_i) as *row* vectors and parameters (e.g. η) as *column* vectors. Thus, the distribution of the random component for observation i, $F(\cdot;\theta_i)$, is characterized by the functional form of the distribution, the function $h(\cdot,\cdot)$, the vector η and the vector of covariates X_i.

In addition to the ex-ante heterogeneity of the random component, the newsvendor might face different trade-offs between overage and underage costs across observations; i.e. the relative cost parameter, γ, might differ on a case to case basis. This is captured in the upper left part of Figure 2. Similar to (4) we let the cost trade-off, γ_i, vary across cases:

$$\gamma_i = g(Z_i, \alpha) \tag{5}$$

where Z_i is a vector of covariates, $g(\cdot)$ is a link function and α is a vector of parameters to be estimated. Note that the set of explanatory variables for the relative cost parameter underlying equation (5), Z, may have a non-empty intersection with the set of explanatory variables for the random component, X, as outlined in equation (4). However, the two sets are not necessarily identical.

Using equations (4), (5) and (2), we can express the optimal decision Q_i^* as:

$$F(Q_i^*; h(X_i, \eta)) = \frac{1}{1 + g(Z_i, \alpha)} \tag{6}$$

Equation (6) specifies the optimally reserved time, Q_i^*, for each observation $i = 1..n$ in the data and thereby introduces the Newsvendor solution into our estimation framework (see

Figure 2, lower part).

As the number of observations is much higher than the number of parameters that we wish to estimate, it is unlikely to find parameters α and η for which the observed decisions Q_i and the predicted optimal reservation time Q_i^* will exactly match. Thus, as in any econometric estimation, the model needs to account for some unexplained variation of Q_i. In the remainder of this section, we propose two models that account for this unexplained variation in different ways.

In the first model (Model N1), we assume that there are some unobservable (to the researcher) factors that are taken into account by the decision maker when determining the overage/underage ratio. Let ξ_i be an i.i.d. unobservable factor that affects the cost ratio for observation i, and assume that $E(\xi_i Z_i) = 0$. Given that γ_i is strictly positive and based on equation (5), we assume the following log-linear specification for the overage/underage cost ratio:

$$\log(\gamma_i) = Z_i\alpha + \xi_i \tag{7}$$

If we knew γ_i, we could use linear regression to estimate α. Of course, the problem is that we do not know the true γ_i. What we do know is that a rational decision maker will behave according to the critical ratio, which can be rewritten to:

$$\gamma_i = \frac{1}{F(Q_i;\theta_i)} - 1 \tag{8}$$

We propose the following two-step procedure to estimate α: Step 1: Using data from the realizations of D_i, estimate η through Maximum Likelihood. Use the estimate $\hat{\eta}$ to compute fitted values $\hat{\theta}_i = h(X_i, \hat{\eta})$. Step 2: Compute the fitted cost ratios $\hat{\gamma}_i = \frac{1}{F(Q_i;\hat{\theta}_i)} - 1$, and then estimate α in the linear model $\ln(\hat{\gamma}_i) = Z_i\alpha + \xi_i$ through Ordinary Least Squares (OLS). We refer to this procedure as TS-OLS. We show in the Appendix the consistency and asymptotic distribution of the estimator provided by TS-OLS.

The intuition behind TS-OLS is simple. In the first step, it estimates the distribution of D_i as seen by the newsvendor. Then, it computes the cost ratio $\hat{\gamma}_i$ that is consistent with the decision Q_i that was made by the newsvendor. Finally, it uses a regression to describe the variability of the cost ratios through the factors in Z_i. The asymptotic variance of the

estimator provided by the second step OLS regression is adjusted for the estimation error incurred in the first step.

Model N1 is not the only way to describe the unexplained variation in the observed decision. In Model N2, we assume that the decision maker behaves approximately rational with some random deviation from the optimal decision. Given that $F(Q|\theta_i)$ is monotone in Q, we can invert equation (6) to get:

$$Q^*\left(W_i, \alpha, \eta\right) = F^{-1}\left(\frac{1}{1+\exp\left(\alpha' Z_i\right)}; h\left(X_i, \eta\right)\right) \tag{9}$$

where $W_i = [X_i, Z_i]$. We assume that $E\left(Q_i|W_i\right) = Q^*\left(W_i, \alpha, \eta\right)$. Define the error term $\nu_i = Q_i - Q^*\left(W_i, \alpha, \eta\right)$. The model can be written as $Q_i = Q^*\left(W_i, \alpha, \eta\right) + \nu_i$, where ν_i is an i.i.d. random variable.

We proceed in a similar way as in the first model and estimate α through a two-step non-linear least squares method (TS-NLLS).[2] This method can be summarized as follows: Step 1: Using data from the realizations of D_i, estimate η through Maximum Likelihood. Step 2: Use Non-linear Least Squares to estimate the equation $Q_i = Q^*\left(W_i, \alpha, \hat{\eta}\right) + \nu_i$. We show in the appendix the consistency and asymptotic distribution of the estimator provided by TS-NLLS.

Model N1 and Model N2 differ in several aspects. The main difference is that they rely on different assumptions to account for the unexplained variability of Q_i. Model N1 assumes that the newsvendor behaves optimally but has private information regarding the cost ratio. Model N2 assumes that the Z vector describes all the factors that affect the cost ratio, but that the decision maker acts with a "trembling hand" around the optimal decision, i.e. the newsvendor acts optimally in expectation but the actual decision is adjusted by a zero-mean random variable. Which one of these assumptions is more appropriate

[2] The parameters α and η in equation (9) could be estimated directly through standard Non-linear Least Squares methods. We found that this approach may fail under some specifications due to identification problems. For example, suppose the D_i's are i.i.d. and that the cost ratio is constant, i.e. X_i and Z_i contain only a constant. This implies that $Q^*\left(W_i, \alpha, \beta\right)$ is constant: the optimal decision is one and the same for all the observations. But there are multiple pairs (α, η) that yield this decision, which makes the model un-identifiable.

depends on the context of the application. Both methods converge as unobserved factors that affect costs and trembling hand behavior are reduced to zero. The two models also differ in the complexity of the estimation method. Both two-step methods, TS-OLS and TS-NNLS are identical on the first step. On the second step, TS-OLS has a closed form solution, while TS-NNLS requires the inversion of the distribution, which might have to be done numerically.

4. Application to Operating Room Time Reservation

Operating Room (OR) management is a broad and complex problem which involves different levels of decision making. It includes strategic decisions such as deciding how much OR capacity to put in place, as well as operational decision such as the scheduling of cases. Given the high complexity of the OR management problem, it is reasonable to decompose the problem into multiple hierarchical decisions. Figure 3 illustrates a time-line with five important decisions related to OR management. Two of these decisions ((b) and (c)) involve trade-offs of reserving too much versus too little OR time, and therefore can be modeled as a Newsvendor problem.

On a yearly basis, the hospital management determines the OR capacity needed based on future case workload. Dexter et al. (2005a) provide a methodology for making this type of decisions using mixed integer programming. The second stage in Figure 3 determines how much OR time to allocate to a specific block (defined as a set of interchangeable operating suites and personnel). Since the workloads of any given day is uncertain, the optimally allocated time has to balance the costs of allocating too much time, which typically translates to idle time for the staff, with the costs of allocating too little time, which typically translates to overtime charges. Strum et al. (1997) develop a newsvendor model to find the optimal time to allocate to each block based on historical workloads. This decision can be revised annually to quarterly. This decision also sets the context in which subsequently case level time reservations are performed.

At a lower level in the hierarchy, each service needs to decide how much OR time to reserve for any given case. This decision is done for each patient individually and is

typically performed during pre-operative planning (see Figure 3 (c)). Reserving too much OR time to cases will very likely increase idle capacity. Reserving too little OR time will lead to more frequent schedule overruns and overtime hours for the hospital staff[3]. When deciding how much OR time to reserve to a case, the decision maker can use information of similar cases that were conducted in the past. This way, forecasts for the case durations can be constructed (Strum et al. (2000a) provide goodness of fit tests for various duration distributions). Given a forecast of the case duration, the decision maker decides how much OR time to reserve for a specific case. It is important to make the distinction between the "forecasted time" for the case duration and the "reserved time" for the case. The former is a purely statistical concept, while the latter takes into account the overage and underage costs. The decision of how much OR time to reserve for a specific case can be modeled as a newsvendor problem, where the random component (D_i) is the actual duration of cases and the decision variable (Q_i) is the amount of OR time to be reserved. (Weiss (1990) and Charnetski (1984) follow similar approaches to model OR time reservation of individual cases). We apply our structural estimation methods to this specific decision.

At some point after the OR time has been reserved to a case, the case needs to be scheduled to a specific day and time (see Figure 3 (d)). This operational decision will depend on the convenience for the surgeon and patient as well as on the urgency of the case, among other variables. Scheduling the case is a decision that is separate from the previously discussed time reservation decision. Dexter and Traub (2002) provide heuristics for these type of decisions. Even closer to the day of the surgery (Figure 3(e)), OR time is *released*, i.e. scheduled cases are cancelled or moved leaving idle OR time which can be used to schedule other cases, cases can be moved and emergencies are scheduled (see Dexter et al. (2003)).

[3]Predictable work hours are a key driver of employee satisfaction in the healthcare industry. For example, Shader et al. (2001) link schedule stability with work satisfaction of nurses as well as with work stress and employee turnover. Mueller and McCloskey (1990) identified eight dimensions of nursing job satisfaction, of which reliable scheduling is one. Similarly, Stachota et al. (2003) cited hours and schedules as one of the primary reasons for nurses terminating their employment and Thompson and Brown (2002) identified schedule conflicts as a major driver of nursing turnover.

The decision of how much time to reserve to a given case provides an excellent context to apply our structural estimation methods of the newsvendor model. First, there exists significant evidence in the medical and health care literature that the newsvendor model is used in practice to reserve OR time in order to balance under and over-utilization costs[4]. Furthermore, these models have been incorporated into decision support tools to assist OR management in hospitals and to optimize the staffing of these facilities (see http://www.mda-ltd.com and Dexter et al. (2001)). Second, the estimation of surgical procedure duration has been extensively analyzed in the medical literature (e.g. Strum et al. (2000a) and Strum et al. (2000b)). These statistical models can be incorporated in the first stage of our two step procedure described in Section 3 to fit the distribution of case durations in our dataset. Finally, the actual case duration, which is the random component of our model, is fully observed. This feature may not be present in those applications of the newsvendor model where demand is censored by the endogenous stocking quantity.

Model Specification

We apply our structural estimation method to the decision of how much OR time to reserve to a specific cardiac surgery case. The model input will be the observed reservation decision, Q and various case characteristics. In addition, we also observe the actual durations of each of the surgery cases (D). Our objective is to estimate the cost ratio γ.

The first stage of our two step procedure requires fitting the distribution of the duration for each case. Our unit of analysis, indexed by i, is an individual cardiac surgery case (e.g. a triple-bypass surgery for Mr. B conducted by surgeon W). The medical literature related to OR management suggests that the lognormal distribution provides a good statistical fit for the duration of surgery procedures (Strum et al. (2000a)). The parameters of this distribution are determined by several case characteristics (Strum et al. (2000b)), and can be anticipated by the decision maker. Let X_i denote the factors that describe the duration of case i. Recall that a random variable Y has log-normal distribution with parameters (μ, σ^2) if $\ln(Y)$ is normally distributed with parameters (μ, σ^2). Following the notation defined in Section 3, we assume that the distribution of the duration of case i is

[4]We thank two anonymous referees for suggesting references from the medical field.

characterized by the parameter vector $\theta_i = (\mu_i, \sigma_i^2)$. Given historical data of case durations, denoted D_i, and the characteristics of each case, denoted X_i, the actual duration of a case can be written as:

$$\log(D_i) = X_i\beta + \varepsilon_i \tag{10}$$

where the ε_i's are assumed to be i.i.d. normally distributed with mean zero and standard deviation σ.[5] Therefore, we have $\eta = (\beta, \sigma^2)$ and $h(X_i, \beta, \sigma^2) = (X_i\beta, \sigma^2)$. Estimating (β, σ^2)via Maximum Likelihood is equivalent to estimating β through OLS and σ based on the standard deviation of the regression residuals.

In fitting the log-normal distribution to cardiovascular procedures, we faced one statistical problem. Cardiovascular procedures are much longer than general surgery cases, and therefore are not well fitted by the log-normal distribution. May et al. (2000) suggest adding a third parameter, a location or "shift" parameter that we denote by δ, which fixes the lower bound of the support of D_i. This means that if D_i is the case duration, then $\tilde{D}_i \equiv D_i - \delta$ follows a log-normal distribution. We follow this approach and estimate the shift parameter as $\hat{\delta} = (d_{\max}d_{\min} - d_{\text{med}}^2) / (d_{\min} + d_{\max} - 2d_{\text{med}})$, where $d_{\min}, d_{\max}$ and d_{med} are the minimum, the maximum and the median case duration observed in our cardiac surgery dataset, respectively. For our sample, this estimate was 134.75 minutes. Note that for this particular application, the maximum likelihood estimates have closed form solutions and therefore numerical optimization routines are not required.

Now, we turn to the second step of the method. Using the log-normality of case duration, we have:

$$F\left(Q_i; \hat{\theta}_i\right) = \Pr(D_i \leq Q_i) = \Phi\left(\frac{\ln(Q_i - \hat{\delta}) - X_i\hat{\beta}}{\hat{\sigma}}\right) \tag{11}$$

where $\Phi(\cdot)$ denotes the standard normal distribution. Replacing $F(Q_i; \theta_i)$ in equation (8) with the fitted value (11), we obtain an estimate $\hat{\gamma}_i$ of the cost ratio. Model N1 estimates

[5] To our knowledge, all the previous work that have analyzed the distribution of case durations assume homoskedasticity of the error term, which for the log-normal case implies a constant coefficient of variation across cases.

α via OLS as described in Section 3.[6] Standard errors of the estimator for this specific application are described in the Appendix.

In order to estimate Model N2, we need to find $Q^*(W, \alpha, \eta)$. Combining equations (9) and (11) gives:

$$Q^*\left(W_i, \alpha, \hat{\beta}, \hat{\sigma}^2\right) - \hat{\delta} = \exp\left\{\hat{\beta}X_i + \hat{\sigma} \cdot \Phi^{-1}\left(\frac{1}{1+\exp(Z_i\alpha)}\right)\right\} \tag{12}$$

Defining $\tilde{Q}_i = Q_i - \delta$ and $\tilde{Q}^*\left(W_i, \alpha, \hat{\beta}, \hat{\sigma}^2\right) = Q^*\left(W_i, \alpha, \hat{\beta}, \hat{\sigma}^2\right) - \hat{\delta}$, the second step is equivalent to estimating the vector α in $\tilde{Q}_i = \tilde{Q}^*\left(W_i, \alpha, \hat{\beta}, \hat{\sigma}^2\right) + \nu_i$ via Non Linear Least Squares. Again, standard errors for the estimates in this application are given in the Appendix.

Data and Variable Definition

Our empirical analysis is based on 258 cardiac surgery cases. Details of the data collection procedure and scope of the sample are defined in the Appendix. Our dataset includes the date, the time of entry into the OR (*TimeIn*) and the time of departure from the OR, the amount of OR time reserved when the case was booked (Q_i) as well as the actual OR time (D_i), patient characteristics and procedure characteristics. The actual and reserved time are measured in minutes, while *TimeIn* is measured in hours elapsed starting at 7 AM. Patient characteristics include a sex dummy (*SEX*=1 if male) and *AGE* (in years, normalized to have mean equal to one). Procedure characteristics include the type of the main procedure conducted (see Table 1, bottom), a dummy to indicate if the procedure was an emergency (*EMERG*), a dummy to indicate if more than one procedure was conducted during the operation (*MPROC*), and information about anesthesia classification (*ASA*). The conventional anesthesia risk assessment score has six levels: (I) No systemic disease; (II) Systemic disease, controlled; (III)Systemic disease, symptomatic or uncontrolled; (IV) Incapacitated; (V) In extremis, moribund; and (VI) Brain death pronounced - organ donor. For cardiovascular procedures, most of the cases fall into categories III or

[6]Note that our structural model implies that Q_i has to be greater than δ for every i, since it can never be optimal to reserve less than the minimum possible case duration. In our dataset, all of the observed reservation times were above the estimate of the location parameter δ.

IV (in our dataset, no cases were classified I or VI, only one case was classified II, and only two were classified V). Therefore, we defined *ASA* as a dummy variable that is equal to one when the anesthesia classification was equal to or above IV. We classified procedures into 5 categories which are defined, following Strum et al. (2000b), based on the CPT classification. Note that each case is classified based on the actual procedures that were conducted. Even though we do not have pre-operation data on the procedures that were planned to be conducted, we learned that there is not much variation between the planned and the actual procedures that are conducted during a surgery. All cases that included a coronary artery bypass procedure were included in the *CABG* category, regardless of additional procedures performed. For these cases, we also coded a measure capturing the number of arteries bypassed (*NBYP*). Aortic valve replacements that included repair of the ascending aorta were classified as *AVR*. Cases of repair of the ascending aorta without *AVR* were included in *OTHER* procedures. Each case was conducted by one of four surgeons, specified by three dummy variables (*S1* through *S3*). Table 1 includes some descriptive statistics for these variables. All dummy variables are coded as binary {0,1}.

Based on the work by Strum et al. (2000b) and conversations with the hospital management at our research site, we defined covariates for actual duration (X) and cost ratio parameter (Z) as follows. In X, we included procedure information (dummies for each type of procedure, *NBYP, MPROC*, *ASA* and *EMERG*), patient information (*AGE* and *SEX*) and dummies for surgeons. Even though we would expect post-operative measures to have significant explanatory power for actual duration, we did not include them as covariates in X since this information is not available to the hospital management when reserving the OR. The *NBYP, ASA* and *EMERG* measures are good proxies for case severity, providing valuable information to predict actual procedure duration. Surgeon dummies are included to account for different levels of experience of the surgeons, which can affect actual duration. Previous research in the medical literature (Strum et al. (2003)) has shown that predicting duration of multiple procedure cases can be difficult, mainly due to the lack of sufficient historical data for each procedure combination. Given the small size of our dataset, we opt to include a single dummy to indicate multiple procedures. This simple

approach could be improved if more data on multiple procedure cases were available. In Z, we included the same procedure characteristics (dummies for each type plus *MPROC*, *ASA* and *EMERG*), dummies for surgeons (*S1*,*S2* and *S3*) and *TimeIn*. We include procedure characteristics mainly because procedures may use different resources of the hospital, which would affect the overage and underage costs. Surgeon dummies are included to control for potential differences in overtime costs for the surgeons. The *TimeIn* covariate allows the cost ratio to vary during the day.

The service we analyzed was allocated one OR per day from 7:30 AM to 7 PM (11.5 hours). Hence, cardiac surgeons never conducted multiple surgeries in parallel by moving back and forth across ORs. Also, in our setting, a surgeon only very rarely conducted two consecutive surgeries on one day. This specific setting is somewhat unusual for a large hospital. But it provides a perfect empirical setting to apply our estimation framework as it supports the assumption of independence in the time allocation across any two cases. This unique setting clearly limits the generalizability of our cost estimates to other hospitals. In other words, while our estimation method applies to other settings (including other applications of the Newsvendor model), our cost estimates do not.

5. Results

Our estimation results are summarized in Table 2. First, consider the variables influencing actual times (variables in X, left part of Table 2)[7]. As we can see, the variation in case durations can partly be explained by patient characteristics such as *SEX* as well as by variables describing case severity (*ASA* and *NBYP*). Cases with multiple procedures (*MPROC*) tend to take longer. In other words, case durations are not identically distributed. The coefficient of determination (R^2) of the regression is equal to 0.39, which reflects that a significant fraction of the variation in D_i can be predicted through the factors in X_i. The point estimate of σ^2 is approximately 0.08, which reflects that there still is uncertainty in predicting actual duration after controlling for patient and procedure characteristics.

Second, consider the estimation of the cost ratio equation (variable Z, right part of

[7]Note that since the first step of the estimation method is the same for the two models, the coefficients for X in models N1 and N2 are identical..

Table 2). As we can see in Table 2, for cases with more than one procedure ($MPROC$), increased emphasis was placed on the costs of OR idle time. The same holds for the number of bypass arteries ($NBYP$) in cardiac bypass surgeries and the anesthesia risk factor (ASA). This reflects that complicated cases may use key hospital resources which have high utilization rates, increasing the overage cost for these procedures.

We also observe that emergency cases tend to have a lower cost ratio. Since emergency cases use dedicated resources which are not shared with regular scheduled operations (for example, at nights or during weekends), idle capacity has a lower impact on costs, lowering the cost ratio parameter. Moreover, as nearly all emergencies are performed within 24 hours of reserving the OR time, the decision maker can use last minute information such as cancellations of other procedures to schedule emergencies in "time windows" that would not otherwise be utilized, further decreasing overage costs.

After looking at the drivers of the cost parameter γ, we now estimate the numeric value of γ. Towards this task, we compute the predicted values, $\gamma_i = Z_i\hat{\alpha}$, for the observations in our sample. The resulting histogram is shown in Figure 4. The median for estimates of γ for Model N1 (Model N2) is 1.79 (1.56). This is evidence that the hospital indeed emphasizes the costs of OR idle time.

We used the estimate of the cost ratio γ and its associated standard errors to test if γ_i would be larger than one with statistical significance[8]. For Model N1 (Model N2) we could reject the hypothesis of a cost ratio less than one, $H_o : \gamma_i \leq 1$, at the 95% confidence level for 44% (46%) of the cases, in favor of the alternative hypothesis $H_1 : \gamma_i > 1$. Only 23% (21%) of the cases showed a cost ratio significantly less than one for Model N1 (Model N2).

Finally, Figure 4 also shows a substantial heterogeneity in γ. The centered R^2 of the second step regression in Model N1 is equal to 0.40, which reflects that a substantial part of this cost heterogeneity can be explained by the factors in Z. An F-test rejects the null of all Z coefficients being equal to zero (p-value less than 10^{-4}).

[8] Asymptotic standard errors for the cost ratio parameter were obtained via the delta method.

We conducted out of sample goodness of fit tests using Model N1 and N2. The results are described in the Appendix.

While our structural model exhibits good fit to the data, the magnitude and heterogeneity of our cost estimates seem to be at odds with some of the models developed in the OR management literature. Strum et al. (1999) suggest that the relevant costs in allocating OR time are staffing costs, which lead to a cost ratio γ less than one because over-time hours are more expensive than regular hours. In contrast, our results suggest an average cost ratio above one. In the next section, we develop two alternative models which extend the structural models developed in Section 3. These models provide results which are better aligned with those reported in the OR management literature.

6. A Newsvendor Model with Forecasting Bias

The structural models developed in Section 3 consistently estimate the overage/underage cost ratio under the assumption that the newsvendor decision Q is based on an unbiased estimate of the distribution of the random component D. However, Dexter et al. (2005b) report systematic biases in OR time forecasts provided by surgical services. There are two main reasons why we believe that the magnitude and variation of the cost estimates of Models N1 and N2 may be driven by systematic forecasting biases. First, we observe that more complicated cases are overrun more often, which is consistent with previous descriptive results related to "over-confidence" developed in other managerial settings (see Kahneman and Tversky (2000)). Consider the two sub-samples of cases with multiple procedures and single procedures, which correspond to 20% and 80% of the total cases, respectively. The average time required to complete multiple (single) procedure cases is around 430 minutes (350 minutes). Comparing the schedule overruns across these two types of cases, we find that multiple procedure cases go over the schedule 75% of the time while low complexity cases go over the scheduled time about 60% of the time. Hence, multiple procedure cases are 1.25 times more likely to go over the scheduled time, which suggests that the decision maker over-estimates his ability to perform complex procedures on time and might be able to improve her forecasting capability.

Second, forecasting biases may also arise due to incentive conflicts between the agents participating in the reservation process. As noted by Dexter et al. (2005b), systematic biases can be caused by surgeons "underestimating their case durations to get their cases to 'fit' into their allocated OR time". In our data, we find that emergencies are overrun 54% of the time versus 64% for non-emergency cases. For emergency cases, surgeons can use patient urgency as their "lever" to get the case on schedule soon, reducing the incentive to misreport their estimated duration for the case. This suggests that incentive conflicts may be a cause for the frequent schedule overruns in our sample.

We now extend our econometric framework to account for forecasting biases in the newsvendor's forecast of D. Continue to assume the same distribution for case duration (log normal with parameters $\mu_i = \beta X_i$ and σ). Instead of assuming a perfectly rational decision maker, we assume that the reservation time is based on a *biased* estimate of the mean distribution. Let D_i^b be the perceived duration of case i, so that $\log D_i^b$ is normal with mean $\mu_i^b = b_i + \mu_i$ and standard deviation σ. For this biased forecast, the reservation decision $\tilde{Q}_i$ is the solution to $\Phi\left(\frac{\ln(\tilde{Q}_i)-(b_i+\mu_i)}{\hat{\sigma}}\right) = \frac{1}{1+\gamma_i}$. Rearranging we obtain:

$$\ln\left(\tilde{Q}_i\right) - \mu_i = b_i + \sigma\Phi^{-1}\left(\frac{1}{1+\gamma_i}\right) \tag{13}$$

The bias term b_i may vary across cases. For example, surgeons may have less incentive to under-estimate emergency procedures relative to non-emergency procedures. We can model the bias as:

$$b_i = B + \varphi' W_i + \omega_i \tag{14}$$

where W_i is a vector of observable characteristics (without an intercept), B is a constant and ω_i denote unobservable factors that affect the bias. We assume that ω_i and ε_i are independent. We also transform W by subtracting its mean so that B represents the average forecast bias. Under these behavioral assumptions and a cost ratio of the form $\gamma_i = \exp(\alpha' Z_i)$, the actual reservation decision becomes:

$$\ln\left(\tilde{Q}_i\right) - \mu_i = B + \varphi' W_i + \sigma\Phi^{-1}\left(\frac{1}{1+\exp(\alpha' Z_i)}\right) + \omega_i \tag{15}$$

Model (15) can have poor identification of the parameters (B, φ, α). To see this, suppose that Z_i contains only a constant, so that the cost ratio is constant. Define the critical

value $t(\gamma) = \Phi^{-1}(1/(1+\gamma))$. Note that even if we know σ, we cannot identify $t(\gamma)$ and B separately. In other words, we cannot identify whether the observed reservation times are due to an under-estimation of procedure duration (low B) or due to a high cost ratio γ (low t) [9].

In order to obtain a more informative model, we impose restrictions on the parameters of (15) based on external information. One alternative is to restrict the model to have a constant cost ratio, leading to the following model:

$$\ln\left(\tilde{Q}_i\right) - \mu_i = c + \varphi' W_i + \omega_i \tag{16}$$

were $c = t(\gamma)\sigma + B$. In a study of OR block allocation conducted at two university hospitals, Abouleish et al. (2003) suggest that underage costs are 75% higher than overage costs, so we set $\gamma = 1/1.75 = 4/7$. We refer to (16) as Model B1. In this model, the variance of the adjusted reserved time $\ln\left(\tilde{Q}_i\right) - \mu_i$ is explained by heterogeneity in the forecasting bias across cases. Given an estimate of c and a value for γ, we can calculate the implicit average bias B and the coefficient vector φ.

Another alternative is to fix the forecasting bias. Letting $\varphi = 0$ and fixing the average bias at $B = B_o$, model (15) becomes:

$$\ln\left(\tilde{Q}_i\right) - \mu_i = B_o + \sigma\Phi^{-1}\left(\frac{1}{1+\exp(\alpha' Z_i)}\right) + \omega_i \tag{17}$$

We refer to (17) as Model B2. Following Dexter and Ledolter (2005), we define the proportionality bias as $\rho = E\left(D_i^b\right)/E(D_i)$. Based on the results reported by Strum et al. (1999) and Dexter et al. (2005b), we fix the value of B_o so that $\rho = 0.8$, i.e. the newsvendor forecast is 20% below the true forecast. Model B2 provides a new estimate of α, which we use to compute a bias-adjusted average cost ratio.

We estimated Model B1 and B2 using a two-step method similar to TS-NLLS. The properties of the estimators and the calculation of the standard errors are detailed in

[9] However, if case duration is heteroskedastic so that $Var(\varepsilon_i) = \sigma_i^2$ is not constant across cases, then it is possible to identify t and B separately, and so γ can be identified. In our results, post-regression statistical tests do not provide significant evidence of heteroskedasticity. This can be due to the relatively homogenous cases in our sample.

the Appendix. The estimate for the φ coefficients of Model B1 are reported in Table 3. The estimate for the intercept c implies a proportionality bias equal to $\rho = 85\%$ (approximately) [10]. We also have added the estimates of the corresponding X coefficients in Table 3 for comparison. The results show that longer procedures tend to exhibit a significantly larger bias. Furthermore, the magnitude of the estimates suggest that, while the decision maker seems to be adjusting its forecast for longer procedures, this adjustment is insufficient. To explore this issue further, we looked at the average reserved and actual times of cases with low and high ASA risk factor for each of the four surgeons in our sample, which are illustrated in Figure 5. The figure shows how the adjustment made in the reservation time, while in the right direction, is insufficient to account for all the heterogeneity among surgeons. Similar results were found by Cachon and Schweitzer (2000), who study newsvendor decisions made in a controlled experiment. They find that subjects make an insufficient adjustment in their newsvendor decisions and tend to be biased towards the mean. In contrast to the substantial forecasting bias found for an average case, the positive sign and magnitude of the φ coefficient for *EMERG* suggests that emergency cases exhibit almost no forecasting bias[11]. This suggests that surgeons are intentionally underestimating the duration of non-emergency cases to get them sooner in schedule.

The estimates of Model B2[12] imply an average bias-adjusted cost ratio of approximately 0.5. This adjusted cost measure is in line with the overage/underage cost ratio reported by Strum et al. (1999).

7. Discussion and Conclusion

As in any single-site research study, one should be cautious to generalize our estima-

[10] This can be calculated by noting that $\rho = \exp(B + \varphi W_i) \cdot E(\exp(\omega_i))$. Assuming ω_i is normally distributed with mean zero, we obtain $E(\exp(\omega_i)) \approx \exp(\hat{s}/2)$ where $\hat{s}$ is the standard deviation of the residuals ω_i in equation (16).

[11] Recall that the W variables are de-meaned, so the average bias for emergency cases is $B + \varphi_{EMERG} \cdot (1 - \overline{EMERG_i}) \approx -0.02$

[12] The coefficient estimates of Model B2 are not reported but can be obtained from the authors upon request.

tion results to other hospitals; the methods that we developed are generalizable to other settings, but the reported estimation results are specific to our study. Specifically, the trade-offs present in our application depend on how much OR time was allocated to cardiac services at the aggregate level (see Figure 3 (a)). The estimated costs are valid to the cardiac service we analyze under the fixed allocated time of one OR, 11.5 hours per day.

We believe that our results are of substantial interest, both from an academic and a managerial perspective. From an academic perspective, the main contribution of this paper is to provide a general structural model to impute the overage and underage costs in newsvendor-type decisions. Our models are sufficiently general to allow for arbitrary parametric distributions of the random variable, and can accommodate observed heterogeneity in this distribution. The model also allows for observed and unobserved heterogeneity in the overage/underage cost ratio and therefore can be used to compute different costs estimates for each observation in the sample. We develop methods that give consistent estimates of the parameters of each of the two models, and derive the asymptotic distribution of the estimators which can be used to compute standard errors of the estimates. Therefore, our methodology can be used to conduct hypothesis testing and is useful for empirical research.

We applied our structural estimation methods to the decision of how much OR time to reserve to a specific surgical case, using real data from cardiac surgery. From the perspective of healthcare management, our analysis reveals that the hospital underlying this study is apparently placing much greater emphasis on OR idle time compared to delays and running over the scheduled time. Specifically, we showed that the costs of OR idle time were perceived, in average, as approximately 60% higher than the cost of schedule overrun. It should be emphasized that using such cost parameters is not right or wrong per se: it simply reflects how the hospital balances partly conflicting objectives. It is the role of the hospital administration to evaluate the alignment of these cost estimates with the overall strategic objectives of the hospital.

The structural models we developed are relevant in practice and can be used in different ways. Consider Models N1 and N2, which assume a fully rational decision maker with no

forecasting biases. We found the cost ratio to be approximately 1.6 in this case. This is a helpful piece of information for the hospital management to diagnose if the current system is in line with the expectations and objectives of the hospital. As an example, consider a hospital manager who views the costs of over-time as 75% higher than idle time costs. Further, assume that using Model N1 or N2 she finds that the hospital appears to value over-time costs 50% *less* than the costs of idle capacity, as we find in our research. This provides statistically significant evidence that there exists a mismatch between the management objectives and the actual behavior of the system. Our analysis is useful to identify potential biases introduced by surgeons while controlling for the procedure mix performed by each surgeon. A more detailed analysis of the reservation process is in order.

This is where Model B1 can be used. Model B1 helps hospital management to diagnose whether there exists a bias in the forecasting process, and if so, to identify the variables that lead to larger biases. This forecasting bias may arise because of insufficient forecasting capability or due to incentive conflicts. For example, our hospital manager might find that the hospital systematically under-estimates the duration of complex cases, as we did in our research. The system would benefit from an enhanced forecasting system, potentially based on an implementation of the methods developed by Dexter and Ledolter (2005).

Finally, model B2 provides a refined estimate for the costs of too much versus too little capacity reservation. Unlike Models N1 and N2, Model B2 corrects for the forecasting bias and hence leads to more realistic estimates of the underlying cost parameters. Based on the adjustment for the forecasting bias, we find the average cost ratio to be 0.5, which is in line with the earlier study by Strum et al. (1999).

One advantage of using structural estimation is that it provides a better understanding of the mechanism by which the different factors affect decisions. In the context of the newsvendor, we can disentangle whether a specific factor affects the observed decision Q_i through the distribution of the random variable D_i or through the overage/underage cost ratio γ_i. This can be helpful for a prescriptive analysis of the system. For example, often it might be easier to adjust factors that affect the cost ratio than changing factors that affect D_i. In addition, disentangling these effects can provide a more robust tool to do

prospective analysis when major changes in the system are introduced. For example, we could use our model to measure the economic impact of subsidies of overage and underage costs. In a decentralized supply chain, the imputed parameters could be used to design contracts to coordinate the supply chain and increase efficiency. Most of the contracts suggested in the literature that coordinate newsvendor decisions depend on the overage and underage costs (see Cachon (2003)). However, these costs are usually private information of each of the agents negotiating the contract, who might not want to reveal them during the bargaining process. Our structural model can be used to impute these costs parameters from historical data, which can facilitate the specification of such contracts. All of these contributions easily carry over to the broad range of existing Newsvendor applications.

We found the application domain of hospital capacity planning to be particularly well suited for structural estimation methods, as hospital operations have been researched both from a analytical and an empirical perspective. We believe that future research could apply our estimation methods to other hospital decisions, such as inventory decisions at blood banks, service level decisions of trauma surgeons, or resource allocations for elective and emergency procedures. Given the broad range of Newsvendor applications in Operations Management, however, the potential usage of our econometric framework extends to Supply Chain Management, capacity planning, and project management.

References

Abouleish, A., F. Dexter, R. Epstein, D. Lubarsky, C. Whitten, D. Prough. 2003. Labor costs incurred by anesthesiology groups because of operating rooms not being allocated and cases not being scheduled to maximize operating room efficiency. *Anesthesia and Analgesia* **96** 1109–13.

Aiken, L. H., S. P. Clarke, D. M. Sloane, J. Sochalski, J. H. Silber. 2002. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association* **288** 1987–1993.

Berry, S., J. Levinson, A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica* **963(4)** 841–890.

Bickel, P.J., K.A. Doksum. 2001. *Mathematical Statistics*. 2nd ed. Prentice Hall, Upper Saddle River, New Jersey 07458.

Brush, T., A. Karnani. 1996. Impact of plant size and focus on productivity: An empirical study. *Management Science* **42**(7) 1065–1081.

Cachon, G. 2003. Supply chain coordination with contracts. Handbooks in Operations Research and Management Science: Supply Chain Management, S. Graves and T. de Kok. 11:229-339.

Cachon, G., M. Schweitzer. 2000. Decision bias in the newsvendor problem with known demand distribution: Experimental evidence. *Management Science* **46(3)** 404–420.

Charnetski, J.R. 1984. Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management* **5** 91–102.

Chorba, R.W. 1976. Potential avoidability: A statistic for controlling in-patient utilization in acute care hospitals. *Management Science* **22**(6) 694–700.

Cohen, M.A., T.H. Ho, J.Z. Ren, C. Terwiesch. 2003. Measuring imputed cost in the semiconductor equipment supply chain. *Management Science* **49**(12) 1653 – 1670.

Dexter, F., R. Epstein, H.M. Marsh. 2001. A statistical analysis of weekday operating room anesthesia group staffing costs at nine independently managed surgical suites. *Anesthesia and Analgesia* **92** 1493–8.

Dexter, F., J. Ledolter. 2005. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historical data. *Anesthesiology* **103** 1259–1267.

Dexter, F., J. Ledolter, R. Wachtel. 2005a. Tactical decision making for selective expansion of operating room resources incorporating financial criteria and uncertainty in subspecialties' future workloads. *Anesthesia and Analgesia* **100** 1425–32.

Dexter, F., A. Macario, R. Epstein, J. Ledolter. 2005b. Validity and usefulness of a method to monitor surgical services' average bias in scheduled case durations. *Canadian Journal of Anesthesia* **52** 935–939.

Dexter, F., R. Traub. 2002. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia and Analgesia* **94** 933–42.

Dexter, F., R. Traub, A. Macario. 2003. How to release allocated operating room time to increase efficiency: predicting which surgical service will have the most underutilized operating room time. *Anesthesia and Analgesia* **96** 507–12.

Green, L. V. 2004. Capacity planning in hospitals. Handbook of Operations Research/Management Science Applications in Health Care, Kluwer Academic Publishers.

Green, L. V., J. Meissner. 2002. Developing insights for nurse staffing. Columbia Business School, working paper.

Green, L. V., V. Nguyen. 2001. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36** 421–442.

Green, L. V., S. Savin, B. Wang. 2003. Managing competing demands in a medical diagnostic facility. Columbia Business School, working paper.

Green, L.V., S. Savin. 2005. Advanced access: What is the right panel size? Conference abstract.

Huang, X. 1995. A planning model for requirement of emergency beds. *Journal of Mathematics Applied in Medicine Biology* **12** 345–353.

Kahneman, D., A. Tversky. 2000. *Choices, Values and Frames*. Cambridge University.

Kwak, N.K., C. Lee. 1997. A linear programming model for human resource allocation in a health-care organization. *Journal of Medical Systems,* **21** 129–140.

Lieberman, M. B., L. Demeester. 1999. Inventory reduction and productivity growth: Linkages in the Japanese automotive industry. *Management Science* **45**(4) 466–485.

May, J., D. Strum, L. Vargas. 2000. Fitting the lognormal distribution to surgical procedure times. *Decision Sciences* **31**(1) 129–148.

Milne, R.G., A. Abebe, B. Torsney. 1989. The impact of teaching on hospital costs: A budgetary approach to non-market institutions. *The Journal of Operational Research Society* **40**(12) 1089–1098.

Mueller, C.W., J. C. McCloskey. 1990. Nurses' job satisfaction: A proposed measure. *Nursing Research* **39(2)** 113–117.

Pell, J., J. Sirel, A. Marsden, I. Ford, S. Cobbe. 2001. Effect of reducing ambulance response times on deaths from out of hospital cardiac arrest: Cohort study. *BMJ* **322** 1385–8.

Porteus, E.L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford Business Books.

Reiss, P., F. Wolak. 2006. Structural econometric modeling: Rationales and examples from Industrial Organization. Working paper, http://www.stanford.edu/ preiss/makeit.pdf.

Rust, J. 1987. Optimal replacement of GMC bus engines: An empirical model of Harold Zucher. *Econometrica* **55(5)** 999–1033.

Shader, K., M. E. Broome, C. D. Broome, M. E. West, M. Nash. 2001. Factors influencing satisfaction and anticipated turnover for nurses in an academic medical center. *Journal of Nursing Administration* **31(4)** 210–6.

Smith-Daniels, V.A., S. B. Schweikhart, D. E. Smith-Daniels. 1988. Capacity management in health care services: Review and future research directions. *Decision Sciences* **19** 889–919.

Stachota, E., P. Normandin, N. O'Brien, M. Clary, B. Krukow. 2003. Reasons registered nurses leave or change employment status. *Journal of Nursing Administration.* **33(2)** 111–117.

Strum, D., J. May, A. Sampson, L. Vargas. 2003. Estimating times of surgeries with two component procedures. *Anesthesiology* **98**(1) 232–40.

Strum, D., J. May, L. Vargas. 2000a. Modeling the uncertainty of surgical procedure times. *Anesthesiology* **92**(4).

Strum, D., A. Sampson, J. May, L. Vargas. 2000b. Surgeon and type of anesthesia predict variability in surgical procedure times. *Anesthesiology* **92** 1454–66.

Strum, D., L. Vargas, J. May. 1999. Surgical subspecialty block utilization and capacity planning: A minimal cost analysis model. *Anesthesiology* **90**(4).

Strum, D., L. Vargas, J. May, G. Bashein. 1997. Surgical suite utilization and capacity planning: A minimal cost analysis model. *Journal of Medical Systems* **21**(5).

Thompson, T. P., H.N. Brown. 2002. Turnover of licensed nurses in skilled nursing facilities. *Nursing Economics* **20(2)** 66–69.

Urbach, D.R., C.M. Bell, P.C. Austin. 2003. Differences in operative mortality between high- and low-volume hospitals in ontario for 5 major surgical procedures: estimating the number of lives potentially saved through regionalization. *Canadian Medical Association Research Journal* **168**(11) 1409–14.

Weiss, E. N. 1990. Models for determining estimated start times and case ordering in hospital operating rooms. *IIE Transactions* **22** 143–150.

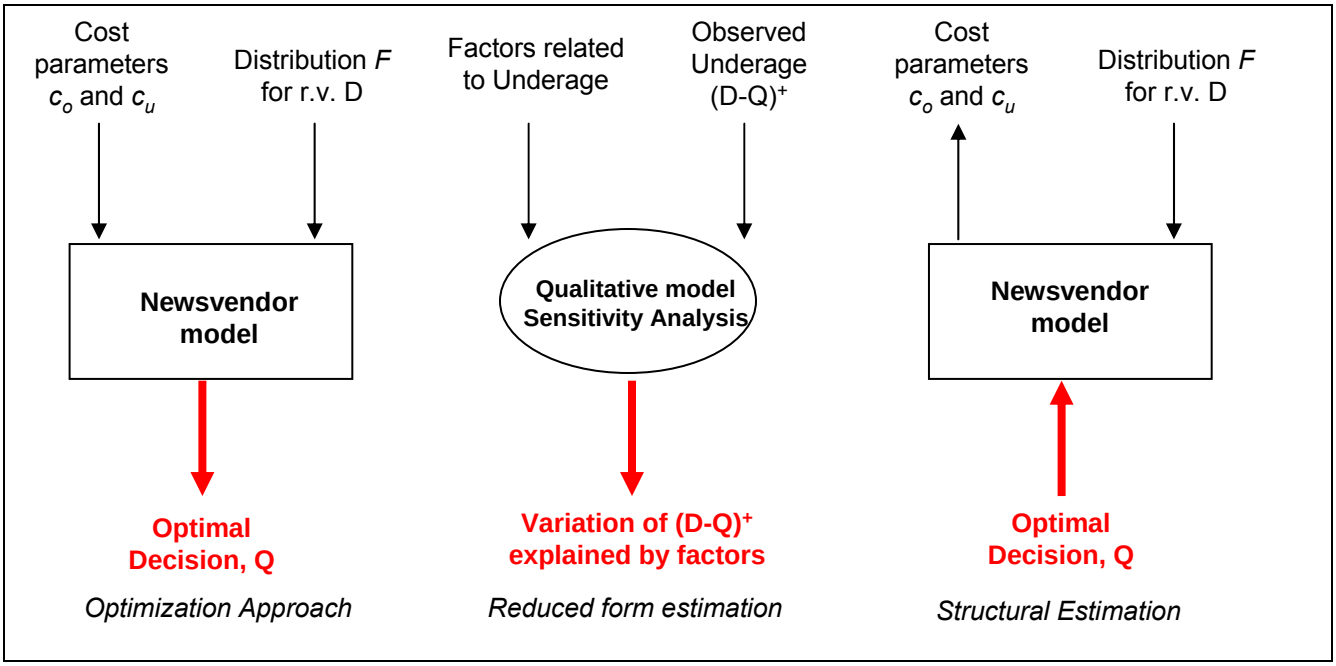


Figure 1: Comparison of different approaches to the Newsvendor Problem.

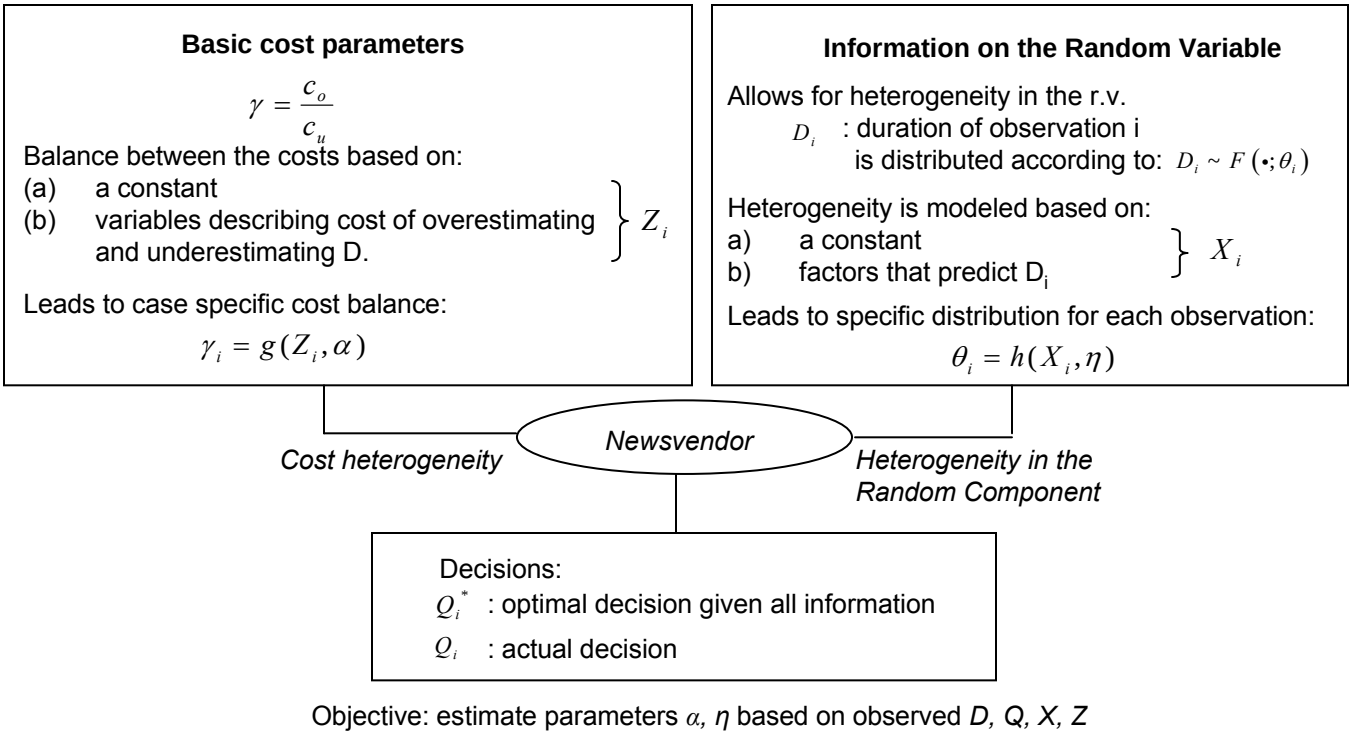


Figure 2: General Econometric Framework.

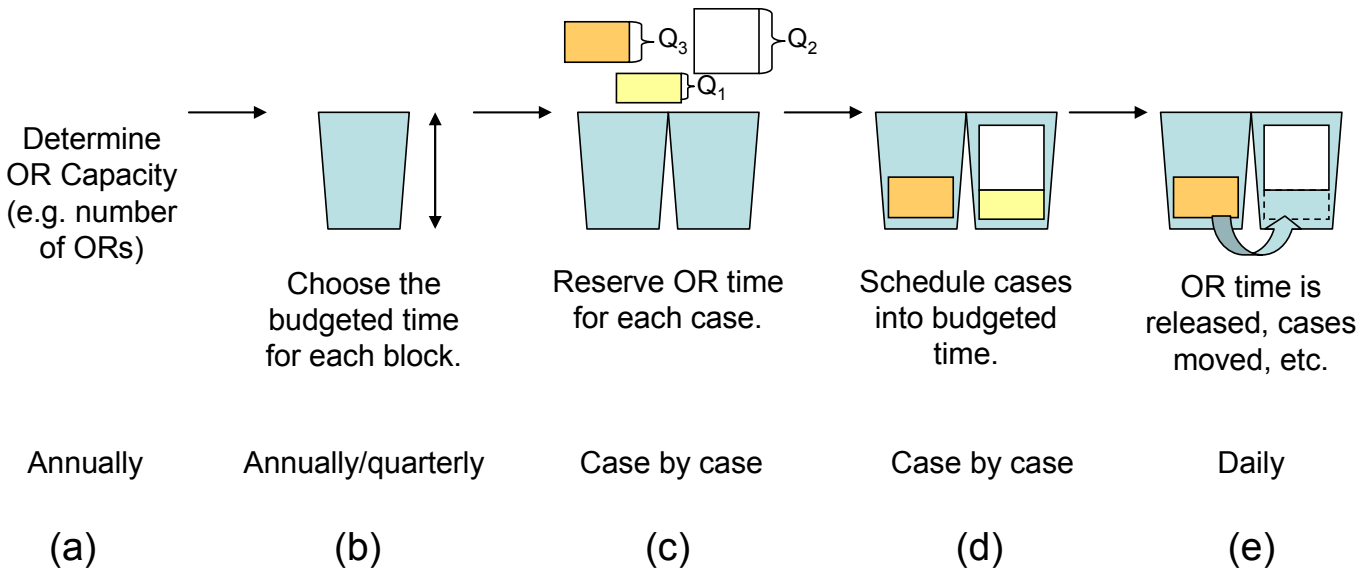


Figure 3. OR Management decision time-line.

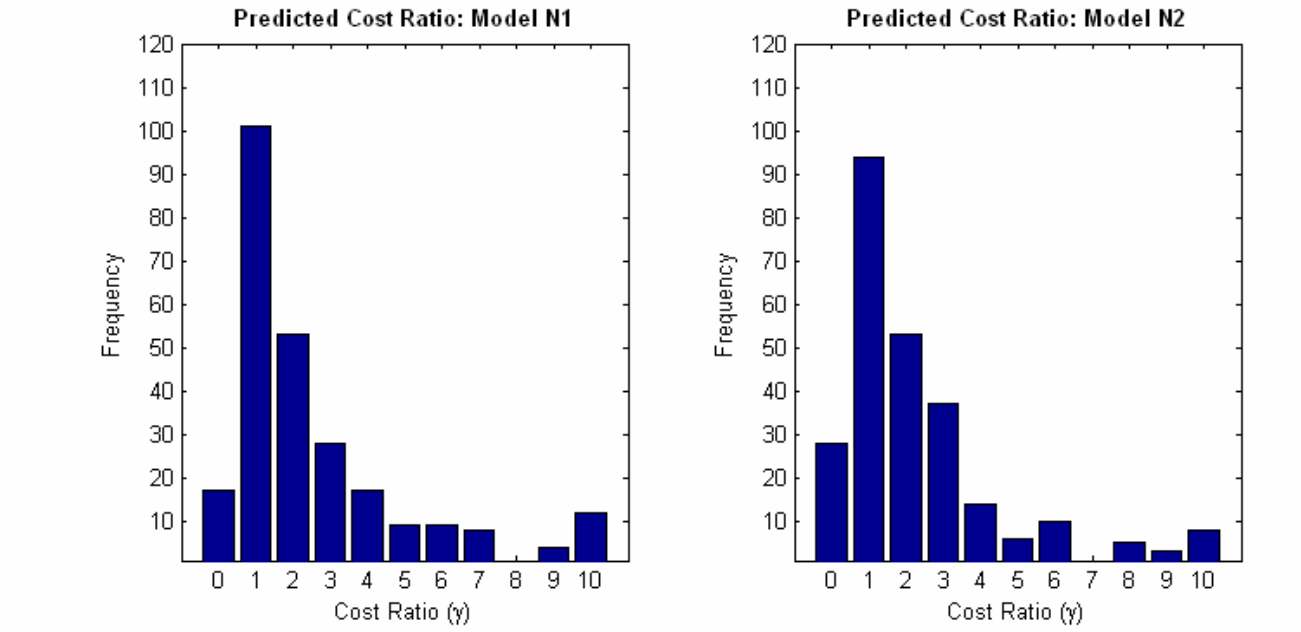
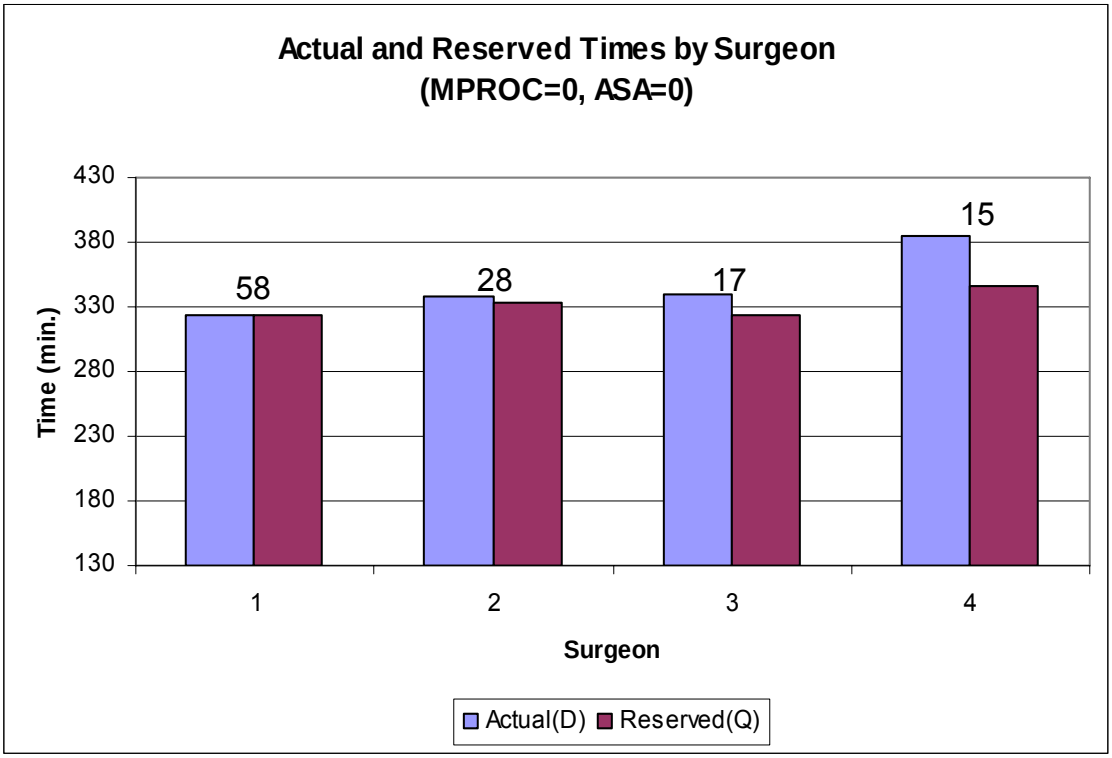


Figure 4: Histogram of cost ratio $\gamma_i = Z_i \alpha$ for Model N1 and Model N2.



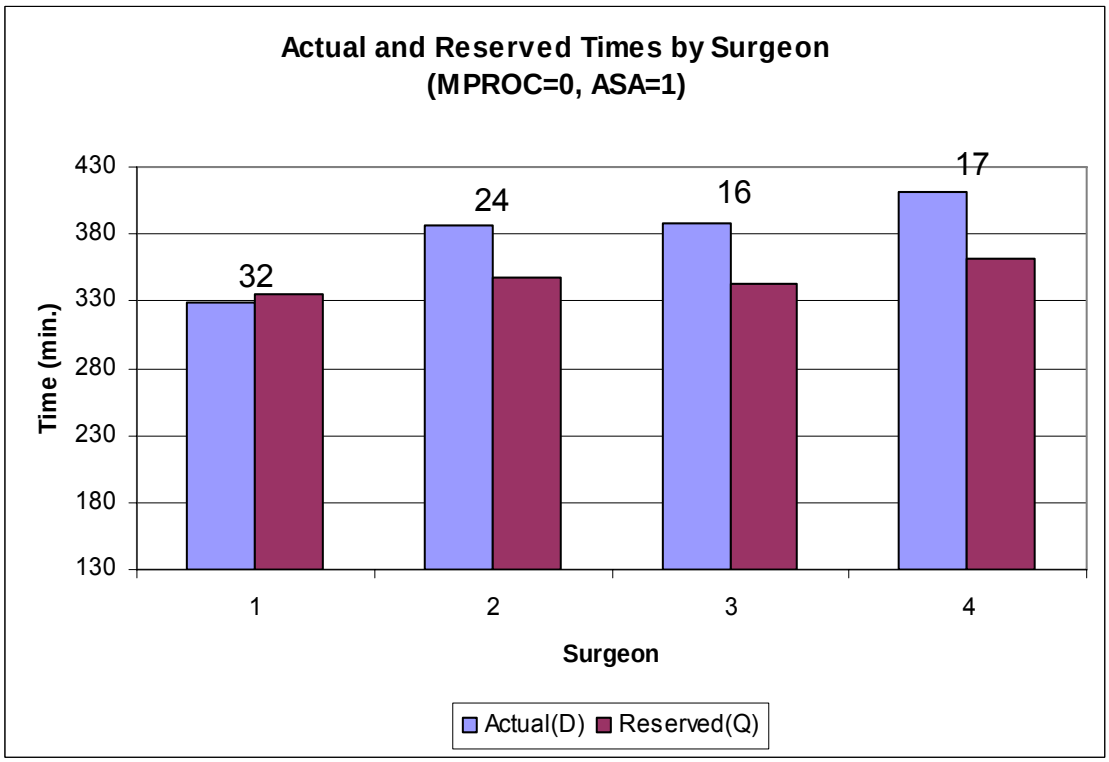


Figure 5 – Average actual case duration (D) and reserved time (Q) by surgeon, for single procedure cases with low (top) and high (bottom) ASA risk factor. Numbers above bars indicates the number of observations on each group.

Variable	Mean	St.Dev.	Min	Max
NBYP	1.74	1.59	0.00	6.00
Age	1.00	0.21	0.32	1.42
Time in	2.76	3.03	0.13	15.58
Q	338.19	42.13	210.00	467.00
D	366.69	84.35	215.00	742.00

Variable	Mean
MPROC	0.20
SEXdum	0.67
EMERG	0.11
ASA	0.45
S1	0.47
S2	0.23
S3	0.15

Procedure	Abbreviation	Mean
Aortic valve replacement	AVR	0.17
Coronary artery bypass graft(s)	CABG	0.62
Mitral valve repair	MV	0.06
Mitral valve replacement	MVR	0.09

Table 1: Summary Statistics

Coefficient	Models N1 and N2	Coefficient	Model N1	Model N2
X covariates		**Z covariates**		
intercept	5.1757 (0.1169)***	*intercept*	0.0065 (0.5475)	- 0.4157 (0.4473)
AVR	- 0.0796 (0.0854)	*TimeIn*	- 0.0083 (0.0256)	- 0.0028 (0.0230)
CABG	- 0.1295 (0.1041)	*AVR*	0.0642 (0.5082)	0.5458 (0.4196)
MV	0.0237 (0.1034)	*CABG*	- 1.0991 (0.5784)*	- 0.7279 (0.4977)
MVR	- 0.0471 (0.0925)	*MV*	0.2734 (0.4969)	0.6697 (0.4033)*
NBYP	0.1019 (0.0229)***	*MVR*	- 0.2039 (0.5853)	0.0834 (0.4476)
MPROC	0.3862 (0.0468)***	*NBYP*	0.6633 (0.0963)***	0.6532 (0.1052)***
SEX	0.0920 (0.0382)**	*MPROC*	1.7790 (0.1940)***	1.6981 (0.2306)***
Age	0.0527 (0.0929)	*S1*	- 0.6768 (0.2303)***	- 0.6488 (0.2263)***
EMERG	0.0947 (0.0598)	*S2*	- 0.1081 (0.2679)	- 0.1031 (0.2561)
ASA	0.0973 (0.0379)**	*S3*	- 0.0054 (0.2805)	0.0166 (0.2608)
S1	- 0.1824 (0.0563)***	*EMERG*	- 0.7071 (0.2799)**	- 0.7002 (0.2411)***
S2	- 0.0859 (0.0588)	*ASA*	0.3876 (0.1605)**	0.2958 (0.1610)*
S3	- 0.0765 (0.0644)			
Sigma^2	0.0784 (0.0001)***			

Table 2: Estimation results for Model N1 and N2. Left (right) of the table shows the estimates for the first (second) step. Standard errors are shown in parenthesis. ***,**,* denote significance at the 1%, 5% and 10% confidence level.

Coef.	Model B1	X Coefficients
c	-0.0906 (0.0116)***	
MPROC	-0.2799 (0.0286)***	0.3862 (0.0468)***
ASA	-0.0632 (0.0250)**	0.0973 (0.0379)**
NBYP	-0.1065 (0.0152)***	0.1019 (0.0229)***
CABG	0.1751 (0.0522)***	-0.1295 (0.1041)
S1	0.1098 (0.0354)***	-0.1824 (0.0563)***
S2	0.0238 (0.0411)	-0.0859 (0.0588)
S3	0.0022 (0.0431)	-0.0765 (0.0644)
EMERG	0.1126 (0.0429)***	0.0947 (0.0598)
TimeIn	0.0018 (0.0040)	
R-sq	0.4099	

Table 3 – Estimates for Model B1 (with cost ratio equal to 4/7). The two right columns are the coefficient estimates of the actual duration estimation, shown for comparison. Standard errors are shown in parenthesis. ***,**,* denote significance at the 1%, 5% and 10% confidence level.

Model	Assumptions			Input	Output
	Cost Ratio	Forecast Bias	Unobservables		
N1	Varies among cases.	None	Cost heterogeneity	*Q, D, X, Z*	•Average cost ratio=1.7 •Factors affecting cost ratio.
N2	Varies among cases.	None	Deviations from optimal reservation (additive)	*Q, D, X, Z*	•Average cost ratio=1.6 •Factors affecting cost ratio.
B1	Constant and equal to 4/7(*).	Varies among cases	Bias heterogeneity	*Q, D, X, W,γ*	•Average forecasting bias = 15%. •Variation of bias among cases.
B2	Varies among cases.	80% of true forecast(**)	Cost heterogeneity	*Q, D, X, Z,B*	•Average cost ratio = 0.5. •Factors affecting cost ratio.

(*) As reported by Abouleish et al. 2003 [4]
(**) As reported by Dexter et al. 2005 [1]

Table 4 – Comparison of Structural Models

APPENDIX

Notation

Throughout this appendix, we use $D_x g = \left[\frac{\partial g}{\partial x_1} ... \frac{\partial g}{\partial x_n}\right]$ to denote the gradient of the real valued function g with respect to x in *row* vector format and $J_x[G]$ to denote the Jacobian of the vector valued function G with respect to x. $F(t; h(X_i, \eta))$ and $f(t; h(X_i, \eta))$ denote the distribution and the density function of D_i, respectively. We use X' to denote the transpose of the vector/matrix X.

Asymptotic Distribution of the TS OLS estimator:

Proposition 1 *Assume: (i) $E(Z_i \xi_i) = 0$; (ii) $E(Z_i' Z_i)$ is finite and of full rank; (iii) the MLE $\hat{\eta}$ is a consistent estimator of η; and (iv) $F(D; h(X_i, \eta))$ is continuous in η. Then, the TS-OLS Method provides a consistent estimator of α and is asympotically normal.*

A consistent estimator of the asymptotic variance of the estimator is given by:

$$\widehat{Avar}(\hat{\alpha}) = \hat{A}_o^{-1} \hat{D}_o \hat{A}_o^{-1} / n$$

where:

$$\begin{aligned}
\hat{A}_o &= n^{-1} \sum_{i=1}^{n} Z_i' Z_i \\
\hat{D}_o &= n^{-1} \sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \\
\hat{g}_i &= s_i(\hat{\alpha}, \hat{\eta}) + \hat{G}_o \cdot \hat{r}_i \\
s_i(\alpha, \eta) &= Z_i' (\ln(\gamma_i(\eta)) - Z_i \alpha) \\
\hat{G}_o &= n^{-1} \sum_{i=1}^{n} -\frac{(1+\hat{\gamma}_i)^2}{\hat{\gamma}_i} Z_i' \cdot D_\eta F(Q_i; h(X_i, \eta)) \\
\hat{r}_i &= I(\hat{\eta})^{-1} \cdot D_\eta l_i(\hat{\eta})'
\end{aligned}$$

and where $I(\hat{\eta})$, $l_i(\hat{\eta})$and $D_\eta l_i(\hat{\eta})$ are the the information matrix, the log-likelihood and the score of the first step maximum likelihood evaluated at the estimate $\hat{\eta}$, respectively.

Proof of Proposition 1

Define $y_i = \ln(\gamma_i)$ and $\hat{y}_i = \ln(\hat{\gamma}_i)$. Because $F(Q_i; h(X_i, \eta))$ is continuos in η, $\ln\left(\frac{1}{F(Q_i; h(X_i, \eta))} - 1\right)$ is continuos in η. Therefore, the consistency of η implies that plim $\hat{y}_i = y_i$. Define

$\zeta_i = \hat{y}_i - y_i$ and $\hat{y}$, y, ζ and ξ as the stack vectors of $\hat{y}_i$, y_i, ζ_i and ξ_i respectively. The estimator $\hat{\alpha}$ is given by:

$$\begin{aligned}
\hat{\alpha} &= (Z'Z)^{-1} Z'\hat{y} \\
&= (Z'Z)^{-1} Z' (y + \zeta) \\
&= (Z'Z)^{-1} Z' (Z\alpha + \xi + \zeta) \\
&= \alpha + n \cdot (Z'Z)^{-1} \left(n^{-1} Z'\xi + n^{-1} Z'\zeta\right)
\end{aligned}$$

The law of large numbers implies that plim $n^{-1}Z'\xi = E(Z_i\xi_i) = 0$ and plim $n \cdot (Z'Z)^{-1} = [E(Z_i Z_i')]^{-1} < \infty$ (by assumption), which together with plim$\zeta = 0$ implies that the second term on the right hand side of the equality converges in probability to zero. Therefore, plim $\hat{\alpha} = \alpha$, which proves consistency. We can also write:

$$\sqrt{n}(\hat{\alpha} - \alpha) = n \cdot (Z'Z)^{-1} \left(n^{-1/2} Z'\xi + n^{-1/2} Z'\zeta\right)$$

By the central limit theorem (CLT), $n^{-1/2}Z'\xi$ is asymptotically normal with mean $E(Z_i\xi_i) = 0$ and variance $E\left(\xi_i^2 Z_i' Z_i\right)$. Since y_i is a continuous function of η, we can use first order Taylor approximation to get:

$$n^{-1/2} \sum_i Z_i' (y_i - \hat{y}_i) = -\left[n^{-1} \sum_i Z_i' \frac{(1+\gamma_i)}{\gamma_i} D_\eta F(Q_i; h(X_i, \eta))\right] \cdot \sqrt{n}(\hat{\eta} - \eta) + o(1)$$

Standard results from maximum likelihood imply:

$$\sqrt{n}(\hat{\eta} - \eta) = I_0(\eta)\, n^{-1/2} \sum_i D_\eta l_i(\eta)' + o(1)$$

where $D_\eta l_i(\eta)$ is the gradient of the log-likelihood of observation i and $I_o(\eta) = E\left(D_\eta l_i(\eta)^t\right)$ is the information matrix for the MLE of step 1. Defining $A_o = E(Z_i' Z_i)$, $s_i(\alpha, \beta) = Z_i\xi_i$, $G_o = E\left\{Z_i' \frac{(1+\gamma_i)}{\gamma_i} D_\eta F(Q_i; h(X_i, \eta))\right\}$, $r_i = I_0(\eta) \cdot D_\eta l_i(\eta)^t$ and $g_i = s_i(\alpha, \beta) + G_o \cdot r_i$ we get that $\sqrt{n}(\hat{\alpha} - \alpha)$ converges in distribution to $A_o^{-1}\left(n^{-1/2}\sum_i g_i\right)$. By the CLT, $n^{-1/2}\sum_i g_i$ is asymptotically normal, which implies the asymptotic normality of $\sqrt{n}(\hat{\alpha} - \alpha)$. The asymptotic variance is given by:

$$Avar\left(\sqrt{n}(\hat{\alpha} - \alpha)\right) = A_o^{-1} D_o A_o^{-1}$$

where $D_o = E\left(g_i' g_i\right)$, and so $Avar\left(\hat{\alpha}\right) = A_o^{-1} D_o A_o^{-1}/n$. Given that $\hat{A}_o$ and $\hat{D}_o$ are consistent estimates of A_o and D_o, plim $\widehat{Avar}\left(\hat{\alpha}\right) = Avar\left(\hat{\alpha}\right)$. ■

Asymptotic Distribution of the TS NLLS estimator.

Proposition 2 *Assume: (i) $E\left(Z_i' Z_i\right)$ is finite and of full rank; (ii) $Q^*\left(W_i, \alpha, \eta\right)$ is continuous in α and η; and (iii) the MLE of α obtained in the first step is consistent. Then, the TS NLLS method provides a consistent estimator of α and is asymptotically normal.*

A consistent estimator of the asymptotic variance of the estimator is given by:

$$\widehat{Avar}\left(\hat{\alpha}\right) = \hat{A}_o^{-1} \hat{D}_o \hat{A}_o^{-1}/n$$

where:

$$\begin{aligned}
\hat{A}_o &= n^{-1} \sum_{i=1}^{n} D_\alpha Q^{*\prime} \cdot D_\alpha Q^* \\
D_\alpha Q_i^* &= -\frac{1}{f\left(Q_i^*; h\left(X_i, \hat{\eta}\right)\right)} CR_i \left(1 - CR_i\right) Z_i \\
CR_i &= \frac{1}{1 + \hat{\gamma}_i} \\
s_i\left(\alpha, \eta\right) &= D_\alpha Q_i^{*\prime} \cdot \nu_i\left(\alpha, \eta\right) \\
\nu_i\left(\alpha, \eta\right) &= Q_i^* - Q_i \\
Q_i^* &= Q^*\left(W_i, \hat{\alpha}, \hat{\eta}\right) \\
J_\eta\left[D_\alpha Q_i^{*\prime}\right] &= \frac{CR_i\left(1 - CR_i\right)}{\left[f\left(Q_i^*; h\left(X_i, \hat{\eta}\right)\right)\right]^2} \cdot Z_i' \cdot \left[\left[D_\eta f\left(t; h\left(X_i, \hat{\eta}\right)\right)\right]_{t=Q_i^*} + \left[\frac{\partial f\left(t; h\left(X_i, \hat{\eta}\right)\right)}{\partial t}\right]_{t=Q_i^*} D_\eta Q^*\left(W_i, \hat{\alpha}, \hat{\eta}\right)\right] \\
\hat{G}_o &= n^{-1} \sum_{i=1}^{n} J_\eta\left[D_\alpha Q_i^{*\prime}\right] \cdot \nu_i\left(\hat{\alpha}, \hat{\eta}\right) + D_\alpha Q^*\left(W_i, \hat{\alpha}, \hat{\eta}\right)' \cdot D_\eta Q^*\left(W_i, \hat{\alpha}, \hat{\eta}\right)
\end{aligned}$$

and $\hat{D}_o$, $\hat{g}_i$ and $\hat{r}_i$ are defined as in Proposition 1.

Proof of Proposition 2

The estimator is a special case of two stage linear least squares (NLLS) (see Wooldridge (2002), pg. 353). The assumptions in the proposition provide the general conditions for which this class of estimators are consistent. Since $F\left(\cdot; \theta_i\right)$ is monotone increasing, $F^{-1}\left(\frac{1}{1+\gamma_i}; h\left(X_i, \eta_o\right)\right)$ is one to one in γ_i and so assumption (i) implies that $\sum_i [Q^*\left(W_i, \alpha, \eta_o\right) - Q_i]^2$ is uniquely minimized at α_o, where (α_o, η_o) denote the true parameters. Therefore,

given η_o, the model is identified for α. Assumption (ii) and (iii) are standard to provide the consistency of the two step M-estimator.

The asymptotic variance of the two-stage NLLS is given by:

$$Avar\left(\sqrt{n}\left(\hat{\alpha}-\alpha\right)\right)=A_o^{-1}D_oA_o^{-1}$$

where

$$\begin{aligned} D_o &= E\left(g_i'g_i\right) \\ g_i &= s_i\left(\alpha,\eta\right)+G_o\cdot r_i \\ s_i\left(\alpha,\eta\right) &= D_\alpha Q^*\left(W_i,\alpha,\hat{\eta}\right)\left(Q^*\left(W_i,\alpha,\hat{\eta}\right)-Q_i\right) \\ G_o &= E\left\{J_\eta s_i\left(\alpha,\eta\right)\right\} \end{aligned}$$

Because $\hat{\eta}$ is identical to the first step estimate for Model N1, r_i has the same form as in Proposition 1. The Jacobian of $s_i\left(\alpha,\eta\right)$ is given by:

$$J_\eta s_i\left(\alpha,\eta\right)=J_\eta\left[D_\alpha Q_i^{*\prime}\right]\cdot\nu_i\left(\hat{\alpha},\hat{\eta}\right)+D_\alpha Q^*\left(W_i,\hat{\alpha},\hat{\eta}\right)'\cdot D_\eta Q^*\left(W_i,\hat{\alpha},\hat{\eta}\right)$$

To compute the gradient with respect to α of $Q^*\left(W_i,\alpha,\eta\right)=F^{-1}\left(\frac{1}{1+\exp(\alpha'Z_i)};h\left(X_i,\eta\right)\right)$ we use implicit differentiation:

$$D_\alpha Q_i^*=-\frac{1}{f\left(Q_i^*;h\left(X_i,\eta\right)\right)}CR_i\left(1-CR_i\right)Z_i.$$

Taking derivatives with respect to η gives the expression for the Jacobian $J_\eta\left[D_\alpha Q_i^{*\prime}\right]$. Replacing the gradient and the Jacobian on the general equations for the asymptotic variance of the two-stage NLLS gives the asymptotic variance of the TS-NLLS specified in Proposition 2. ■

Standard Errors for the OR Reservation Application

To calculate the Asymptotic variance of the estimators, we need to specify $\hat{G}_o$, $\hat{r}_i$ and $s_i\left(\alpha,\eta\right)$ for each model. Since the first step method used to compute $\hat{\eta}$ is the same for both models, we start by specifying $\hat{r}_i$. The MLE of model (**??**) gives the following expressions

for the information matrix $I(\beta,\sigma^2)$ and the score function $D_{(\beta,\sigma^2)}l_i(\beta,\sigma^2)$:

$$\begin{aligned} I(\eta) &= \begin{pmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \\ D_\beta l_i &= \frac{1}{\sigma^2}X_i\left(\ln(D)-X_i\beta\right) \\ \frac{\partial l_i}{\partial\sigma^2} &= \frac{-n}{2\sigma^2}+\frac{1}{2\sigma^4}\left(\ln(D)-X_i\beta\right)'\left(\ln(D)-X_i\beta\right) \end{aligned}$$

These defines r_i and its sample counterpart $\hat{r}_i$ for both methods.

For model 1, using equation (12) we have:

$$\begin{aligned} D_\beta F_i &= -\frac{X_i}{\sigma}\phi\left(\frac{\ln\tilde{Q}_i-X_i\beta}{\sigma}\right) \\ \frac{\partial F_i}{\partial\sigma^2} &= -\frac{\left(\ln\tilde{Q}_i-X_i\beta\right)}{2\sigma^3}\phi\left(\frac{\ln\tilde{Q}_i-X_i\beta}{\sigma}\right) \end{aligned}$$

which can be replaced in the expression for $\hat{G}_o$.

For model 2, note that

$$\begin{aligned} \tilde{Q}^*\left(W_i,\alpha,\hat{\beta},\hat{\sigma}^2\right) &= \exp\left\{\hat{\beta}X_i+\hat{\sigma}\cdot\Phi^{-1}\left(\frac{1}{1+\exp(Z_i\alpha)}\right)\right\} \\ D_\alpha\tilde{Q}_i^* &= \tilde{Q}^*\left(W_i,\alpha,\hat{\beta},\hat{\sigma}^2\right)\cdot\hat{\sigma}\cdot\frac{1}{\phi\left(\Phi^{-1}(CR_i)\right)}\cdot(-CR_i)(1-CR_i)Z_i \\ D_\beta Q_i^* &= X_i\tilde{Q}^*\left(W_i,\alpha,\hat{\beta},\hat{\sigma}^2\right) \\ D_{\sigma^2}Q_i^* &= \frac{1}{2\sigma}\Phi^{-1}(CR_i)\,\tilde{Q}^*\left(W_i,\alpha,\hat{\beta},\hat{\sigma}^2\right) \\ f\left(t;\mu,\sigma^2\right) &= \frac{1}{\sigma\sqrt{2\pi}}\frac{1}{t}\exp\left\{-\frac{(\ln t-\mu)^2}{2\sigma^2}\right\} \\ \frac{\partial f\left(t;\mu,\sigma^2\right)}{\partial t} &= -f\left(t;\mu,\sigma^2\right)\frac{1}{t}\left(1+\frac{\ln t-\mu}{\sigma^2}\right) \\ \left[f\left(t;X_i\beta,\sigma^2\right)\right]_{t=Q_i^*} &= \frac{1}{\sigma Q_i^*}\phi\left(\Phi^{-1}(CR_i)\right) \\ \left[D_{(\beta,\sigma^2)}f\left(t;X_i\beta,\sigma\right)\right]_{t=Q_i^*} &= f\left(Q_i^*;X_i\beta,\sigma\right)\frac{\Phi^{-1}(CR_i)}{\sigma}\left[X_i,\frac{\Phi^{-1}(CR_i)}{2\sigma}-\frac{1}{2\sigma\Phi^{-1}(CR_i)}\right] \end{aligned}$$

The score function is calculated by replacing the expression for $D_\alpha\tilde{Q}_i^*$ on $s_i(\alpha,\eta)=D_\alpha\tilde{Q}_i^*\left(Q_i^*-Q_i\right)$. Replacing the expressions above in the equation for $\hat{G}_o$ and simplifying

terms gives the following expression:

$$\hat{G}_o = n^{-1} \sum_i -\frac{CR_i\left(1-CR_i\right)}{\phi\left(\Phi^{-1}\left(CR_i\right)\right)} \sigma Q_i Q_i^* \cdot Z_i' \cdot \left[\begin{array}{c} X_i \\ \frac{1}{2\tilde{\sigma}} \Phi^{-1}\left(CR_i\right) + \frac{\nu_i(\alpha,\eta)}{2\sigma^2 Q_i} \end{array} \right]'.$$

Estimation Methods for Models B1 and B2

We estimated Models B1 and B2 by running a two-stage non-linear least square regression of equations (17) and (18). As in TS-OLS and TS-NLLS, the first stage estimates (β, σ) using data on actual duration (D). In the second stage, we replace the fitted values $\hat{\mu} = \beta X_i$ and $\hat{\sigma}$ in the corresponding model and use standard non-linear least squares. Because the models are special cases of two step M-estimators, the estimators are consistent (the proof follows the same lines as Proposition 1 and 2).

Model B1 can be rewritten as:

$$y_i\left(\beta\right) = \bar{W}_i'\varphi + \omega_i$$

where $y_i\left(\beta\right) = \log Q_i - X_i'\beta$ and the vector $\bar{W}_i$ includes the the covariates W_i and the intercept c. Using the same previous notation, the asymptotic variance of the estimator of φ for Model B1 is given by:

$$Avar\left(\hat{\varphi}\right) = A_o^{-1} D_o A_o^{-1} \cdot n^{-1}$$

where

$$\begin{aligned} A_o &= n^{-1} \cdot \bar{W}'\bar{W} \\ D_o &= n^{-1} \sum g_i g_i' \\ g_i &= s\left(\varphi, \hat{\beta}\right) + G_o r_i \\ G_o &= E\left[J_\beta s_i\left(\varphi, \hat{\beta}\right)\right] \\ s\left(\varphi, \beta\right) &= \bar{W}_i\left(y\left(\beta\right) - \bar{W}_i'\varphi\right) \end{aligned}$$

The Jacobian is given by: $J_\beta s_i\left(\varphi, \beta\right) = \bar{W}_i'\left(-X_i\right)$, which gives:

$$G_o = -n^{-1}\bar{W}'X$$

For Model B2, we write equation (18) as:

$$y_i(\beta) - B = \sigma t(\alpha; Z_i) + e_i$$

where $t(\alpha; Z_i) = \Phi^{-1}\left([1 + \exp(Z_i'\alpha)]^{-1}\right)$. We estimate α by replacing the equation above with the fitted values $\hat{\beta}$ and $\hat{\sigma}$. Define $m_i(\alpha, \sigma) = \sigma t(\alpha, Z_i)$ and $m(\alpha, \sigma) = [m_1(\alpha, \sigma) ... m_n(\alpha, \sigma)]'$. Using the same notation as above for the asymptotic variance, we calculate each of the terms as follows:

$$\begin{aligned}
A_o &= n^{-1} \cdot D_\alpha m(\alpha; \hat{\sigma})' D_\alpha m(\alpha; \hat{\sigma}) \\
s_i(\alpha; \beta, \sigma) &= D_\alpha m_i(\alpha, \sigma)(y_i(\beta) - B - m_i(\alpha, \sigma)) \\
&= D_\alpha t(\alpha, Z_i) \cdot \sigma(y_i(\beta) - B - m_i(\alpha, \sigma)) \\
J_{(\beta,\sigma)} s_i &= D_\alpha t(\alpha, Z_i)' \cdot \begin{bmatrix} -\sigma X_i \\ \hat{e}_i - m_i(\alpha, \sigma) \end{bmatrix} \\
G_o &= n^{-1} \sum_i J_{(\beta,\sigma)} s_i
\end{aligned}$$

The gradient $D_\alpha t(\alpha, Z_i)$ is calculated as:

$$D_\alpha t(\alpha, Z_i) = Z_i \cdot \left\{ -\frac{1}{\phi(\Phi^{-1}(CR_i))} CR_i(1 - CR_i) \right\}$$

Details on data collection

Our analysis is based on a data set that was collected in a large US teaching hospital. After obtaining approval from our Committee on Human Research, we conducted a retrospective study using data from patients who underwent cardiac surgery. The study period was January 1, 2003 to December 31, 2003. Throughout the study period, the cardiac service was allocated one OR from 7:30 AM to 7:00 PM on each day.[1] Patients were included in the sample if they were over 18 years of age and if they underwent one or more of the following procedures: coronary artery bypass surgery, cardiac valve surgery, excision of a cardiac mass, or repair of the ascending aorta. Other cardiac surgery procedures were included if cardiopulmonary bypass support was required. Emergency cases were included, however we excluded cases of repair of complex congenital heart disease

[1] In the afternoon of of each day, the remaining time on the following day not reserved for cardiac surgery is opened up to book surgeries of other services.

and all heart or lung transplants.. Congenital heart disease is conducted by a completely different group of surgeons, and in the context we study, can be considered as a different service. Transplants, all of which are classified as emergencies, are a small fraction of the total cardiac surgeries conducted by the service we analyze (about 5%). We excluded these cases because the variability on their actual duration is much larger than for the rest of cardiac procedures, mainly due to waiting times for organ donations. Therefore, the OR reservation process used to book these cases can be quite different. Finally, we only considered those cases that were conducted by surgeons who conducted two or more procedures in our study period.

We also excluded extremely severe cases in which the patient died during the same hospitalization in which they had the surgery. We repeated our analysis over a larger sample which included these 15 cases and found that the coefficient estimates did not change substantially. However, cases in which the patient died, sometimes on the day of the surgery, were persistent outliers in all the specifications analyzed and reduced the goodness of fit significantly. Estimating the distribution of actual duration for these extremely severe cases seems to be particularly difficult and therefore we decided to exclude them from the sample.

The data were collected from multiple internal sources including OR scheduling, hospital billing, heart-lung bypass records, and anesthesia records. This accomplished our objective to obtain high quality data via triangulation, i.e. by verifying data through multiple, distinct sources, and provided the additional benefit of compensating for imperfections in record keeping practices of individual units within the hospital. Like most hospitals, the hospital at which this study was undertaken did not incorporate planning-related information such as the forecasted procedure time in the same systems as patient-related clinical information.

Out of sample goodness-of-fit test

To do the out of sample test, we randomly picked five points and excluded them from the estimation. Using the rest of the sample we estimated the parameters of the model

and used these to predict the excluded points. We repeated this process 100 times, giving us 500 out-of-sample predictions. We focused on predicting the adjusted reserved time, defined as the difference between $\log \tilde{Q}_i$and the conditional median time ($E[\log \tilde{D}_i|X_i]$). Based on equations (11) and (13), our structural model implies:

$$\begin{aligned} \log \tilde{Q}_i - E\left[\log \tilde{D}_i|X_i\right] &= \log \tilde{Q}_i - \beta X_i \\ &= \hat{\sigma} \cdot \Phi^{-1}\left(\frac{1}{1+\exp\left(Z_i\alpha\right)}\right) \end{aligned}$$

where $\tilde{D}_i = D_i - \delta$ is the random part of actual case duration. We chose the adjusted reserved time as the predicted variable in order to compare our structural model to a model in which the decision maker follows a simple rule of reserving a fixed percentage of the median time. According to this simple rule, $\log \tilde{Q}_i - E\left[\log \tilde{D}_i|X_i\right]$ should be a constant. Figure 1 shows a scatter of plot of the actual versus the fitted values of the adjusted reserved time, for Model N1 and Model N2. Regressing the actual values against the fitted values plus an intercept gives a centered R^2 of 0.37 for both models. Thus, our structural newsvendor model was able to explain around 40% of the variation in adjusted reserved time, which suggests that the newsvendor model has significant predictive power of the actual behavior observed in the data.

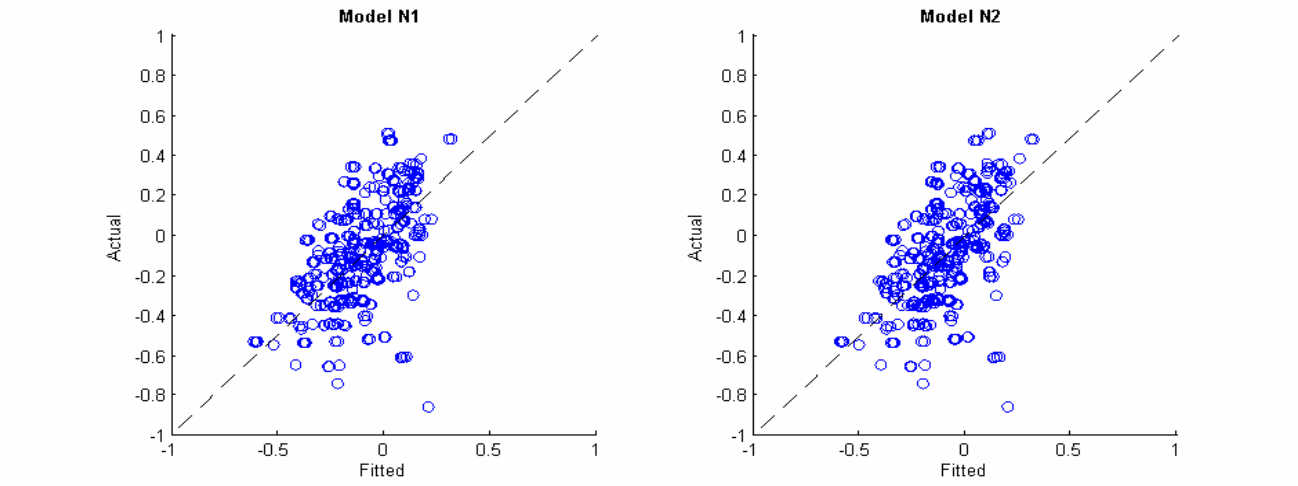


Figure 1: Out of sample goodness of fit tests for Model N1 and Model N2 to predict adjusted reserved time.

Validation and limitations of the empirical results

Our cost ratio estimates are based on a selected group of surgeries. Cardiac surgeries are significantly longer than general cases (average duration of cardiac cases is about 6 hours). The underage and overage costs inherent to this type of surgeries are likely to be different from general cases, and so our cost estimates might not apply to a more general selection of cases.

In order to ensure the robustness of our findings, we validated the econometric assumptions underlying our work. First, by applying a Newsvendor model instead of a scheduling model, we implicitly assume that each reservation decision is made in isolation, and given γ_i, the cases are independent of each other. This assumption reflects the fact that the time reservation to the procedure is usually made well in advance and before observing any procedure durations during that day. The lack of statistical significance in the *TimeIn* variable on the cost ratio provides some evidence in support for this assumption. We also analyzed whether conducting other cardiac cases in the same day had any effect on the overage/underage cost ratio. For this, we considered specifications that included indicators on whether the case is: followed or preceded by another cardiac surgery (*POST* and *PREV*, respectively); followed or preceded by a surgery conducted by the same surgeon (*POSTSS* and *PREVSS*, respectively). We found that *POSTSS* was positive and statistically significant, which implies that cases followed by a case conducted by the same surgeon tend to be overrun more often (all the other indicators were not significant). Only 5% of the cases in our sample fall into the *POSTSS* category. The coefficients on the other covariates are similar in sign, magnitude and statistical significance to those reported in Table 2. Our main results seem to be robust to the effect of other cardiac surgeries been conducted in the same day. To provide further support, we looked at days which had more than one scheduled cardiac case and plotted the estimated times of consecutive procedures. In absence of any systemic pattern in the scatter plot, we believe the independence assumption is reasonable.

Second, we assume that conditional on X_i and Z_i, Q_i^* and D_i are statistically independent, i.e. doctors don't "rush" to finish on-time[2]. Given the highly difficult nature

[2]Observe that this assumption is sometimes violated in the traditional Newsvendor application to

of cardiac surgery and the associated risks of death for the patient and liability for the surgeon, we believe this is a reasonable assumption. To explore this assumption formally, we looked for an abnormal concentration of the actual duration just before the end of the reserved time: if doctors rush to finish on time, we should see over-proportionally many observations just around the reservation time. However, we found this distribution as well as the distribution for the overrun time, $D_i - Q_i$, to be smooth and well-behaved.

We also checked for any weekday effects, using two approaches: (i) comparing the A/F ratios for each weekday; and (ii) introducing weekday dummies in the covariates. The average A/F ratios are not significantly different between days, and none of the weekday dummies introduced in either X and Z were significant.

Note that the endogeneity of the *TimeIn* variable may be introducing a bias in our estimation. Since the scheduling of the case is endogenous, this variable could be correlated with unobservable factors (from the researchers perspective) that affect procedure duration (this will happen, for example, if more complicated cases are scheduled early in the morning). If this is the case, then *TimeIn* should predict some of the variation in actual case duration. When adding this covariate to X, its coefficient was negative and not significant (t-value equal to 1.33). This suggests that this bias is unlikely to be important.

To validate the assumptions about the shifted log-normal distribution of case durations, we analyzed the residuals generated on the first step of our estimation method. A Jarque-Bera test cannot reject the null hypothesis of normally distributed residuals (p-value .16). Similar results were found using the Shapiro-Wilks test of normality (p-value of 0.11). We also analyzed quantile-quantile plots of the residuals and found further support for the log-normality of case duration.

We evaluated the robustness of our results to the type of econometric model used by comparing the estimates of Model N1 and N2. Assuming the estimation of α is independent across the two models, none of the coefficients shown in Table 2 are different with statistical significance (95% confidence level). Overall, we found that the estimation of α and the

retailing: demand is frequently correlated with the number of items on the retailer's shelf, i.e. there is an endogenous effect of stocks and product variety in demand.

cost ratios are robust across the two specifications.

In Section 6 we develop two alternative models which account for forecasting bias in the newsvendor decision. To validate the fit of these models, we compared models B1 and B2 using Cox's test for non-nested models (Pesaran and Deaton (1978)). None of the models could be rejected against each other (lowest p-value equal to 0.3), which suggests that both models provide reasonably good fit to the data.

References

Pesaran, M., A. Deaton. 1978. Testing non-nested nonlinear regression models. *Econometrica* **46**(3) 677–694.

Wooldridge, Jeffrey. 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, Massachusetts.