Causal Inference Models in Operations Management

Teck-Hua Ho*, Noah Lim**, Sadat Reza†, and Xiaoyu Xia‡

This version, June 21, 2017

Abstract

Operations management (OM) researchers have traditionally focused on developing normative mathematical models that prescribe what managers and firms should do. Recently, there has been increased interest in understanding what managers and firms actually do and the factors that drive these decisions. To advance this understanding, empirical investigation using causal inference models is critical. However, in many contexts, the ability to obtain causal inferences is fraught with the challenges of endogeneity and selection bias. This paper describes five empirical tools that have been widely used in economics to address these challenges, and how they can be adopted by OM researchers. We also present an example which illustrates how the various attributes of big data – variety, velocity and volume, can be useful in addressing the endogeneity bias.

Keywords: Causal Inference, Empirical Methods, Operations Management, Big Data

*Corresponding author, National University of Singapore; **University of Wisconsin-Madison and National University of Singapore; †Institute on Asian Consumer Insight, Nanyang Technological University; ‡Chinese University of Hong Kong. Email: Ho: dprhoth@nus.edu.sg; Lim: nlim@bus.wisc.edu; Reza: sreza@ntu.edu.sg; Xia: xiaoyuxia@baf.cuhk.edu.hk. The authors are listed in alphabetical order. The authors would like to thank Ingrid Koch and Ashish Sachdeva for their constructive comments and Rehan Ali for his excellent editorial help. We also thank the editor and two anonymous referees for their helpful comments and suggestions.

1 Introduction

Operations management (OM) is an applied field. Historically, OM researchers have primarily focused on developing normative mathematical models that prescribe what managers and firms *should* do, and placed relatively little emphasis on understanding what they *actually* do and why they do it. Recently however, there has been increased interest in the latter. The understanding of the causal impact of managerial actions is often obtained through empirical analysis using observational data. As is well known, such investigation is fraught with challenges.

Two such challenges that commonly arise in OM contexts are endogeneity and self-selection. It is difficult to find empirical contexts where neither of these challenges is present. Hence, researchers need to explicitly address the potential manifestations of these challenges in their research contexts, and apply appropriate tools that allow them to mitigate any associated problems.

In this paper, we discuss five empirical modeling tools that have been widely applied in economics to overcome the challenges posed by endogeneity and self-selection. These are (1) the instrumental variables estimator in cases where endogeneity is due to omitted variables; (2) the instrumental variables estimator with exclusion restrictions in cases where endogeneity is due to simultaneity; (3) the propensity score matching estimator in cases where the selection mechanism is explained by observables; (4) the regression discontinuity design in cases where the selection mechanism is not explained by observables; and (5) the difference-in-differences estimator in cases where the selection mechanism is not explained by observables but researchers have access to data that span multiple periods.

A review of the empirical literature in OM reveals two interesting observations. First, while the number of empirical papers published has increased over time, empirical OM researchers still remain a minority. Second, explicit discussion of the above challenges and applications of the above-mentioned tools is not widespread. Consider the three well-established journals in the field: *Management Science* (*MS*), *Manufacturing and Service Operations Management* (*MSOM*) and *Production and Operations Management* (*POM*). From 2010 to 2015, 1,015 articles were published on OM topics. Noticeably, only 17% of the articles contained empirical analysis. Among these articles, 81% were on causal inference using observational data, with the rest being on either predictive modeling or inference using experimental data. Only 37% of the papers using observational data explicitly addressed endogeneity and/or self-selection biases. These statistics, reported in Table 1, show that there is much scope for wider application of the above tools in empirical OM.

The availability of tools needs to be complemented with good quality data for robust empirical

investigation. One could argue that the historical lack of empirical research in OM could be attributed to the difficulty of obtaining data. Recently, wide-scale digitization and the availability of big data have rapidly changed the information landscape. However, big data analysis is not immune to the challenges of endogeneity or self-selection. Researchers still need to apply the appropriate econometric tools to draw robust causal inferences. Nonetheless, big data has three characteristics – variety, velocity and volume – that provide the following advantages: (1) variety can increase the sources of instrumental variables – a key ingredient in addressing the above challenges; (2) velocity, or the high frequency nature of big data, opens up new avenues for empirical research by allowing investigation of problems that would otherwise not be feasible; and (3) volume can enhance the quality of estimates by improving their precision.

The objective of this paper is to provide a concise introduction to the endogeneity and selection bias issues in empirical modeling for causal inference using non-experimental observational data.[1] In particular, we focus on linear models, and present several examples on how to address these issues in such models.[2] We hope that this paper will help OM researchers, particularly those who are less familiar with but interested in empirical modeling, gain a workable knowledge of a few useful tools for causal inference modeling.

We begin our discussion in section 2 by laying out the requirements for consistent estimation using a linear regression model. In section 3, we discuss the endogeneity bias and the tools to address it. In section 4, we elaborate on the selection bias and the tools to address it. In section 5, we present an example which illustrates how the various characteristics of big data can be useful in addressing the endogeneity bias. In section 6, we conclude with a discussion of the research implications for OM.

2 Linear regression – the workhorse model for causal inference

Causal inference models using observational data can be broadly categorized into two strands. The first strand of models imposes analytical structures on the data, and is known as structural

[1] We do not discuss empirical models for predictive purposes, such as time-series models. Researchers interested in forecasting can refer to Enders (2004) for a detailed introduction to such models. We also exclude empirical models for experimental data, either from the laboratory or field. Field experiments, particularly well-designed randomized controlled trials, are ideal for causal inference. However, such experiments are usually prohibitively expensive, and statistical analyses of experimental data are relatively simpler.

[2] We do not discuss non-linear models and panel data models in this review. The two empirical challenges that we highlight in this paper, endogeneity and selection biases, are also pertinent to such models. Interested readers can refer to Greene (2000) and Wooldridge (2008) for textbook-level discussions on such models.

econometric modeling. These analytical structures are typically based on the optimality conditions derived from profit or utility maximization models.[3] Our paper focuses on the second and more widely-used strand of causal inference models, known as reduced-form models. The reduced-form approach is useful when the analytical structure underlying the data is either not fully established, or cannot be taken directly to the available data. Instead, one imposes a statistical relationship between an outcome variable and a set of explanatory variables that may potentially determine the outcome variable. The most well-known tool for reduced-form estimation is the linear regression model. In this paper, we focus primarily on addressing endogeneity and selection bias issues for this model.

Applications using linear regression models are ubiquitous in the empirical OM literature. Suppose the researcher observes the data $(Y_i, \mathbf{X}_i)$ on $i = 1, ..., N$ individuals or firms, where Y_i denotes some outcome variable of interest, and $\mathbf{X}_i$ denotes an M-dimensional set of explanatory variables. The linear model assumes the following causal relationship:

$$Y_i = \boldsymbol{\beta}'\mathbf{X}_i + \epsilon_i \tag{1}$$

where ϵ_i is a zero-mean unobserved error term, and $\boldsymbol{\beta} = (\beta_1 \quad \beta_2 \ ... \ \beta_M)'$ is an M-dimensional parameter vector. Thus the conditional expectation $E[Y_i|\mathbf{X}_i]$ is $\boldsymbol{\beta}'\mathbf{X}_i$. The objective is to estimate each parameter $\beta_m \in R$, which captures the causal effect of the component X_m of $\mathbf{X}$ on Y, since $\beta_m = \partial E[Y_i|\mathbf{X}_i]/\partial X_{mi}$. Let $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ denote the N-dimensional vectors of the outcome variable and the unobserved variables, respectively, for the N individuals in the sample. The ordinary least squares (OLS) estimator is given by, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$. It can be shown that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{\mathbf{X}'\mathbf{X}}{N}\right)^{-1}\frac{\mathbf{X}'\boldsymbol{\epsilon}}{N}. \tag{2}$$

One of the conditions necessary for the OLS estimator to be consistent (*i.e.* $plim\ \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$) is $plim\ \frac{\mathbf{X}'\boldsymbol{\epsilon}}{N} = 0$. In other words, $\mathbf{X}$ should not be correlated with $\boldsymbol{\epsilon}$. When the above condition is not met, then causal inference based on the linear model is not valid. This is the fundamental concern in modeling causal relationships. In particular, we draw attention to endogeneity and self-selection problems, which are quite pervasive in OM contexts. If these problems are not properly addressed, then the validity of the empirical results would be questionable.

[3] A few examples of structural modeling papers in OM include those by Akşin et al. (2013), Caro et al. (2014), Golrezaei et al. (2014), Kim et al. (2014), Li et al. (2014) and Hyndman and Parmeter (2015).

3 Endogeneity bias

Endogeneity bias arises when the explanatory variables and the errors are correlated. Specifically, if any component X_m of $\mathbf{X}$ is correlated with the error term, then it follows from (2) that $E[X_{mi}\epsilon_i] \neq 0$. Consequently, $\hat{\beta}_m$ does not converge to the true parameter β_m, even if the number of observations $N \to \infty$. This leads to the OLS estimator being inconsistent, and thus the correct correlation between the dependent and explanatory variables cannot be established. Two important reasons for the endogeneity bias are: (1) omitted variables and (2) simultaneity.[4] Below we describe these issues, and also discuss the econometric tools that can generate consistent estimates in the presence of these problems.

3.1 Omitted variable bias

Omitted variable bias occurs when one ignores the possibility of some unobserved variable affecting both the outcome and key explanatory variables in the estimation models. We demonstrate this using a service management example where the research question is on how staffing levels affect store sales. Perdikaki et al. (2012) and Mani et al. (2015) investigated this question. Suppose the data we observe consists of average daily sales (in $'000) and average daily staffing levels (number of staff) for $i = 1, ..., N$ stores located in various markets for a particular week. Let Y_i and X_i denote store sales and staffing level, respectively, for store i.

A standard regression model to estimate the effect of the staffing levels on store sales would be as follows:

$$Y_i = \alpha_0 + \alpha_1 X_i + \mu_i \tag{3}$$

where μ_i is a random error term. However, if there exists some variable Z_i, such as store promotions conducted during the observation week, which is correlated with both store sales and the staffing levels, then a more accurate model for the data generating process is possibly given by $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i,\ Cov(X_i, Z_i) \neq 0$. In this situation, if we regress Y_i only on X_i, as in (3), then the estimated coefficient of X_i, $\hat{\alpha}_1$, will not be a consistent estimate of β_1, the true causal effect. Since Z_i is unobserved and omitted from the regression model, X_i and μ_i are correlated in (3), which makes X_i endogenous.

We illustrate the extent of the omitted variable bias using simulated data generated based on

[4] A third reason for the endogeneity bias is due to measurement error in the explanatory variables. Because the solution to address this issue is similar to the case with omitted variables, *i.e.*, we need to find an instrumental variable, we do not discuss the measurement error problem separately.

the above example with $\beta_1 = 10$.[5] In Figure 1(A), we draw a scatter plot of $\mathbf{Y}$ and $\mathbf{X}$ for a few observations. We do not distinguish between promotional and non-promotional sales, since this data is unobserved. As shown in the graph, the estimated coefficient of the staffing level based on regression model (3) is 13.7 (s.e. 1.2), which is significantly larger than the true parameter value. However, if we also observe which stores offered promotions (Z_i), we can group the observations on the basis of whether the store had offered a promotion, and re-estimate regression model (3) for these two groups. We draw the predicted lines based on the two regressions in Figure 1(B). The estimates of the coefficient of X_i for both the regressions are not significantly different from 10, the true parameter value.

– Figure 1 here – , – Figure 2 here –

Solution: One way to correct for the omitted variable bias is to introduce an instrumental variable (IV). Intuitively, the IV filters out the influence of the omitted variable in a regression model. Let V_i be such an IV. Then two conditions must be satisfied: (1) V_i must be correlated with X_i, *i.e.*, $Cov(V_i, X_i) \neq 0$; and, (2) V_i must not be correlated with the omitted variable Z_i, *i.e.*, $Cov(V_i, Z_i) = 0$. These conditions allow us to generate an estimate of X_i, $\hat{X}_i$, which contains the same information as in X_i but is free from the influence of Z_i.

The correction for omitted variable bias in a linear regression framework is usually conducted using a two-stage least squares (2SLS) approach.[6] In the first step, we regress the endogenous explanatory variable X_i on the instrumental variable (see Figure 2(A)), V_i, and other exogenous variables in the model. Denote the predicted values by $\hat{X}_i$. In the second step, we regress the outcome variable Y_i on $\hat{X}_i$ and the other exogenous variables.

In our example above, the staffing level is endogenous since we do not observe which store offered the promotion. Finding an appropriate set of IVs is usually a challenge. Both Perdikaki et al. (2012) and Mani et al. (2015) used the lagged staffing level as an instrument. This is a plausible IV, as stores are likely to maintain some continuity in the staffing levels over consecutive periods. In our example, we do not have other exogenous variables, therefore it is sufficient to regress only on the lagged staffing level. Notice that the effect of lagged staffing level on current

[5] The complete data generating process is as follows: $Y_i = 500 + 10X_i + 100Z_i + \epsilon_i$ and $X_i = 20 + 0.5V_i + 6Z_i + \nu_i$. We randomly assign half the observations into the groups $Z_i = 1$ and $Z_i = 0$. V_i is randomly generated from $N(50, 5)$, and ϵ_i and ν_i are randomly generated from $N(0, 5)$, where N denotes the normal distribution.

[6] Alternatives to the 2SLS approach include 3SLS method, limited information maximum likelihood estimation, control function approach, and generalized methods of moment. Readers can refer to graduate level econometrics textbook such as Wooldridge (2008) for discussions of these methods.

staffing level is significant with a coefficient estimate of 0.5 (s.e. 0.1). In Figure 2(B), we show the regression line for Step 2 of the 2SLS approach, where store sales is regressed on the predicted staffing level from Step 1. The estimated slope coefficient from Step 2 turns out to be 10.7 (s.e. 3.3), which is statistically not different from the true value of β_1, which equals 10. In addition to correcting for the omitted variable bias, researchers may be interested in the direction of the bias. It follows from (2) that in a single equation regression model the direction of bias will depend on the covariance between the endogenous variable and the omitted variables. If the covariance is positive then the OLS regression will overestimate the effect, and vice versa. In our simulated example, the explanatory variable (staffing level) is positively correlated with the omitted variable (store promotions). As a result, the OLS estimate turns out to be upward biased.

It is important to note that the IV should not be correlated with the omitted variable. While it is testable whether the IV is correlated with the endogenous explanatory variable, there are no statistical tests that can demonstrate that it is uncorrelated with the omitted variable. This is due to the fact that the omitted variables are not observed. However, there are statistical tests (*e.g.* Anderson's LM test, Cragg-Donald F-test) for checking whether the IV is sufficiently correlated with the explanatory variables. Standard software packages report these test statistics routinely with IV-regression outputs.

3.2 Simultaneity bias

The second cause of endogeneity bias is simultaneity. This occurs when there are multiple dependent variables each influencing at least one other dependent variable. An example of the simultaneity problem in empirical OM research is found in Kesavan et al. (2010), in which the authors argue that cost of goods sold, gross margin and inventory levels are simultaneously determined.

To demonstrate the problems associated with simultaneity, we use an example of the causal relationship between air ticket sales and the price of air tickets. The price of air tickets is an important determinant of the demand for air travel. However, airline companies adjust ticket prices based on expected demand. Therefore, air ticket sales are a function of price from the demand perspective, but the price of air tickets is also a function of expected air ticket sales from the supply perspective. Suppose that we have data on the number of tickets sold for a particular route originating in a particular city for various quarters, along with the average prices of those tickets

sold. We can use the following system of equations to represent the demand and supply functions:

$$\begin{aligned} Demand: Q_t &= \beta_{10} + \beta_{11} P_t + \beta_{12} X_{1t} + \epsilon_{1t} \\ Supply: P_t &= \beta_{20} + \beta_{21} Q_t + \beta_{22} X_{2t} + \epsilon_{2t} \end{aligned} \tag{4}$$

where Q_t and P_t are the log transformed values of quantity of tickets sold and average price per ticket for periods $t = 1, ..., T$. X_{1t} and X_{2t} are factors that affect customer demand for tickets and airline companies' pricing decisions, respectively, and ϵ_{1t} and ϵ_{2t} are random shocks. Consider the demand function. Regressing Q_t on P_t and X_{1t} will generate a biased estimate of the demand elasticity β_{11}. This is due to the fact that P_t is also determined by Q_t in the supply function. Consequently, we cannot claim that $Cov(P_t, \epsilon_{1t}) = 0$. In other words, P_t is endogenous in the demand equation, and the estimate of β_{11} using such a regression model would be subject to the simultaneity bias.

We illustrate this problem using simulated data in Figure 3 using $\beta_{10} = 10,\ \beta_{11} = -0.6,\ \beta_{12} = 0.2$ and $\beta_{20} = 1,\ \beta_{21} = 0.5,\ \beta_{22} = 0.3$. In equilibrium, the demand for air travel equates supply. We generated the data using this equilibrium condition and the above parameter values. X_1, X_2 and the errors were drawn from independent standard normal distributions. A direct regression of quantity on price yields a slope estimate of -0.1 (s.e. 0.2), which is not significant and reflects neither the elasticity of demand ($dQ_t/dP_t = -0.6$) nor the elasticity of supply ($dQ_t/dP_t = 2$).

– Figure 3 here –

Solution: To correct for the simultaneity bias, we need one IV for every endogenous variable. Since there are two endogenous variables, price and quantity, in our example, we need two IVs. Importantly, for each endogenous variable there must be at least one IV that does not directly affect any other endogenous variable. This is known as the exclusion restriction. In our case, we assume that per capita income affects only the quantity demanded, and fuel costs affects only the pricing strategy of airlines. Therefore, per capita income and fuel costs satisfy the exclusion restrictions, and can be used as IVs. We can use these IVs to estimate both the demand and supply function parameters, using two separate sets of regressions, via the 2SLS method. Consider estimating the demand function. In the first step, the endogenous variable (ticket prices) is regressed on all the instrumental variables (crude oil prices and per capita income) to generate the predicted ticket prices.[7] In the second step, we regress the quantity of tickets sold on these predicted ticket

[7] If there were any exogenous variables in any of the two equations, then those variables should be in the regression as well.

prices. The estimated coefficient is -0.6 (s.e. 0.2), which is statistically no different from the true parameter value of -0.6.

In the discussion on omitted the variable bias correction, we had mentioned how the direction of bias is dependent on the covariance between the explanatory variable and the omitted variable. However, in the presence of simultaneity, the direction of bias is not as easy to establish. In such cases, the direction of the bias will depend on both the signs and magnitudes of the actual parameters, which are *a priori* unknown. Thus in our example, the direction of bias will depend on the signs and magnitudes of β_{11} and β_{21}.

4 Selection bias

Many OM questions involve investigating the effects of adopting a program or policy on one or more dependent variables. Researchers have investigated the effects of adopting programs such as total quality management (TQM) or ISO certification on productivity and other aspects of firm performance (Levine and Toffel 2010, Gray et al. 2015). When participants self-select into different programs, versus when they are randomly assigned, standard regression models do not adequately estimate the effect of the program.

Let us consider the example of studying the effects of firms obtaining ISO 9000 certification. In this case, researchers may be inclined to introduce an indicator variable W_i, which takes the value of 1 if firm i is exposed to a treatment (*i.e.*, adopted ISO 9000 standards in this context) and 0 otherwise, in the regression model $Y_i = \beta_1 W_i + \boldsymbol{\beta}_2' \mathbf{X}_i + \epsilon_i$. However, such straightforward inclusion of the treatment indicator in the regression model will generally not lead to a consistent estimate of the actual impact of ensuring ISO 9000 standards on firm productivity, due to self-selection by firms into the programs.

Why is self-selection an issue for generating a consistent estimate of the causal effect? For any firm i, we observe the outcome conditional on the adoption decision. Letting $Y_i(W_i)$ denote the outcome conditional on the treatment $W = \{0, 1\}$, then we observe $Y_i = Y_i(1)\mathbf{1}[W_i = 1] + Y_i(0)\mathbf{1}[W_i = 0]$. In the program evaluation literature, this equation is known as the Rubin causal model. If a firm adopts the treatment, then the actual observed outcome is $Y_i(1)$, but we do not observe the potential outcome in the non-adoption condition, $Y_i(0)$. Similarly, if a firm does not adopt the treatment, then the actual observed outcome is $Y_i(0)$, but we do not observe the potential outcome in the treatment condition, $Y_i(1)$. Since $Y_i(1)$ and $Y_i(0)$ are never observed simultaneously, the causal effect of the treatment $E[Y_i(1) - Y_i(0)]$ cannot be directly estimated using the observed

data. If a firm self-selects into the treatment group based on another factor that can also influence the outcome, such as the quality of its management, then we cannot be sure whether the estimated effect is due to the treatment itself or due to the firm's management quality.

We illustrate the self-selection bias using an example of the decision to obtain ISO 9000 certification. Suppose we are interested in estimating the causal effect of obtaining the certification on firm output. If it is possible to estimate the outputs of firms having obtained and not obtained the certificate, *i.e.*, if we were to observe both the actual outcomes and potential outcomes, then it is fairly straightforward to estimate the causal effect. In Table 2, we provide an example with two firms, where we observe the actual and potential outcomes for both firms. Firm B is ISO certified and its actual output is 800 and the potential output without certification is estimated to be 700. Therefore, the treatment effect for Firm B is 100. Firm A is non-certified and its actual output is 400 and the potential output with certification is estimated to be 500. Therefore, the treatment effect for Firm A is also 100. However, the potential outcomes are not observable. What we observe are only the actual outcomes, also shown in Table 2. In this example, if we assume that the two firms are similar, then we would incorrectly infer that the treatment effect is $800 - 400 = 400$, which is the difference between the outputs of Firms A and B. Without accounting for selection, a regression model would generate biased estimates.

To identify the causal impact of a treatment, the estimation depends crucially on the assumptions behind the assignment mechanism. If assignment in the treatment group can be assumed to be dependent only on observable covariates, we can use models applicable for selection-on-observables. If there exists some unobservable covariate that influences the outcome and treatment assignment rule, then we can use models that are able to address selection-on-unobservables. Below we discuss a few popular models for these two types of selection problems.

4.1 Selection on observables

Consider the same question of estimating the causal effect of obtaining ISO 9000 certification on firm output. Suppose that we have data on both certified and non-certified firms. Firms which have obtained the certification are in the treatment group, and firms which have not obtained the certification are in the control group. ISO certification is granted on the basis of a firm's performance on eight criteria: (1) customer focus, (2) leadership, (3) people involvement, (4) quality management process, (5) management system, (6) continual improvement, (7) approach to decision making and (8) supplier relationships. If we observe data on all of the above eight criteria, then in the estimation model we can control for these characteristics that determine the assignment rule (to the treatment

group). Once these characteristics have been controlled for, the firms' decisions to obtain ISO certification can be considered to have random differences, and in this case selection is no longer a confounding factor. We can then compare the outputs of the firms belonging to the treatment and control groups.

Solution: Matching Model. To estimate the treatment effect given the above scenario, we can use a matching estimator. For every treated unit, the goal of the matching estimator is to find a comparison unit among the controls that has similar values of observable characteristics X_i. This comparison unit need not be a single unit – rather, it can be a composite (*i.e.*, a weighted average) of several different control units that have similar values of X_i. After computing the average difference between the treated units and the control units, we can use a function of this distance as weights to construct the composite comparison units. More detailed discussion on different matching methods can be found in Todd (2010). Note however that when the observable characteristics X_i has many dimensions, it would be difficult to find a comparison unit that is comparable to the treated unit in every dimension, an issue known as the "curse of dimensionality".

Among the matching models, the propensity score matching (PSM) model has gained significant popularity, as it is intuitive and does not suffer from the "curse of dimensionality". The essential idea behind the PSM model is that if one can estimate (1) the propensity (or probability) of selecting the treatment conditional on the observed covariates, and (2) the expected outcomes with and without treatment for units with similar propensities, then the treatment effect can be estimated by comparing the average outcomes of the treatment group and control group firms with similar propensities. In order to estimate the treatment effect, we need to group the firms according to their similarity in characteristics. The primary advantage of PSM is that the propensity score is a scalar, thus it is easy to group the firms on the basis of their propensities to participate in the treatment.

– Table 3 here –

Continuing with the example of estimating the impact of obtaining ISO 9000 certification on firm output, suppose that in addition to certification status and output, we also observe data on all of the eight certification criteria. Data on these characteristics will allow us to calculate each firm's propensity to obtain ISO 9000 certification. This can be done through a simple probit or logit model. Once these propensities are estimated, we can group firms according to their similarity in propensities. In Table 3, we present the sample data for our illustration. As shown, for each group of firms based on high or low propensities, we can find the average outputs for firms that have and

have not obtained certification. Using this procedure, we can see that firms with low propensities can expect a treatment effect of an output increase of 75, while firms with high propensities can expect a treatment effect of 150. This highlights an additional advantage of the PSM method. Not only are we able to match firms using scalar propensities, which can be derived from multiple characteristics, but we are also able to see if the treatment effects differ across propensities.

Denote the propensity of firm i self-selecting into a treatment by $P_i(\mathbf{X}_i)$, where $\mathbf{X}_i$ is the vector of characteristics that determine assignment to the treatment group. Assume that there are firms with similar values of $\mathbf{X}$ in both the treatment group (ISO certified) and the control group (non-certified). Thus, we should be able to estimate $E[Y_i(1)|P_i(\mathbf{X}_i) = p]$ and $E[Y_i(0)|P_i(\mathbf{X}_i) = p]$, where p is the value of propensity score. One can then proceed to estimate the treatment effect, which is given by, $\beta_{PSM} = E[Y_i(1)|P_i(\mathbf{X}_i) = p] - E[Y_i(0)|P_i(\mathbf{X}_i) = p]$.

In the first step of the PSM method, researchers need to estimate the propensities. Typically a probit or a logit model is used to generate these estimates. In the next step, once the propensities are estimated, we need to estimate the average outcomes of the treatment and control group firms around the various propensity measures. For this purpose, we need to find groups of firms matched according to their propensities. We can use a specific matching algorithm, such as nearest-neighbor matching, caliper and radius matching, classification and interval matching, kernel weighting function and regression weighting. To estimate the treatment effect, one can then compare the average difference between the treated units and the composite comparison units that are matched by the chosen algorithm. Alternatively, one can use the classification approach suggested by Rosenbaum and Rubin (1983). This approach first divides the estimated propensities into J groups, and then it estimates the average outcome for treated and untreated firms within each group. Following the estimation of the treatment effect for each group, we can then find the average treatment effect in the sample.

4.2 Selection on unobservables

If the assignment to a treatment condition depends on factors that are not observed and if such factors also affect the outcome, then there is selection on unobservables. In such situations, a matching estimator such as PSM is not applicable, since assignment to the treatment group is not solely determined by observable factors. There is no general approach for causal inference or estimation of the treatment effect in the presence of selection-on-unobservables (Imbens and Wooldridge 2009). Researchers have to rely on quasi-experimental designs to identify the causal effect of treatment. Two popular methods for estimating treatment effects are the regression discontinuity design (RDD)

and the difference-in-differences (DID) method . For RDD, the researcher observes a threshold or cutoff value of a continuous variable that separates the treatment and control group units. The units around this cutoff could be considered to be similar. Comparison of the outcomes of these units allows estimation of the treatment effect. To apply the DID method, the researcher observes data for at least two periods, one before exposure to the treatment and one after exposure to the treatment, for both the treatment and control group units. Under the assumption of a common time trend, comparison of pre- and post-treatment outcomes for the two groups yields the treatment effect. We discuss the two methods below.

Solution 1: Regression Discontinuity Design (RDD). The RDD allows for an estimation of the causal effect of a treatment, when the treatment is assigned only above (or below) some cutoff or threshold of an observable variable X_f, known as the forcing variable. We estimate the treatment effect by comparing observations lying closely on either side of the threshold. Such forcing variables exist in many contexts, often based on the application of some administrative criterion.

In the previous example of examining the effect of obtaining ISO 9000 certification on firm output, suppose now that we do not observe all of the eight characteristics. In this case, PSM may not be a useful approach since the unobserved variables may also have an impact on firm output. However, suppose that we are able to obtain data on another variable that is continuous in nature, such that above a certain value of this variable, all firms are in the treatment group. For example, if the certification granting agency generates a score for each firm, on the basis of which the certification would be awarded, then surely there will be a cutoff value above which firms would be in the treatment group. If the researcher has access to these scores, then the score can be used as a forcing variable. Suppose that this score is generated on a scale of 1-10, and that there is a cutoff value of 5, above which the firms are able to receive certification. It is plausible to assume that firms scoring just above 5 and just below 5 are similar in many ways. Therefore, we can compare the average output of firms with scores just above 5 (certified) and those just below 5 (non-certified) to estimate the effect of ISO certification on output. We illustrate this in Figure 4.[8]

– Figure 4 here –

It is recommended that one start with a visual presentation of the data in order to evaluate the possibility of implementing RDD. If the plot of the being-treated ratio against any continuous-valued covariate suggests that there is a clear threshold at a particular value of the covariate, then one can

[8] Another example of having a discontinuity in the data is Bennett et al. (2013). The authors exploit changes in the testing protocols for car emissions and find a discontinuity in the passing rate based on the cars' model year.

consider that covariate to be the forcing variable. It is then advisable to plot the forcing variable and other observable covariates. If there is any indication of discontinuity around the threshold for any other covariate, then the treatment effect is considered to be not identified. One routine procedure for RDD is to plot every covariate against the forcing variable and test the existence of discontinuity.

The advantage of RDD is that it allows for selection on unobservables. Firms may self-select to be in a treatment regime if the realized value of a forcing variable is above a threshold c, *i.e.*, the selection equation is given by, $W_i = 1\Big[X_{fi} \geq c\Big]$, where X_{fi} is the forcing or assignment variable. While various unobservable factors may determine the value X_f, we can assume that around $X_f = c$, firms are similar. Firms with realizations of the forcing variable just above c are in the treatment group and firms with realizations just below c are in the control group. Since we assume that firms belonging to these two groups are similar, we can estimate the potential outcome in the non-treatment regime using the average outcomes of the firms just below c. This is the essential idea behind RDD. Firms just below the cutoff are considered to be the quasi-control group. The treatment effect is given by, $\beta_{RDD} = \lim_{\epsilon \downarrow 0} E[Y_i | X_{fi} = c + \epsilon] - \lim_{\epsilon \uparrow 0} E[Y_i | X_{fi} = c + \epsilon]$.

Once the sample near the cutoff is selected, one can then use the following regression to estimate the treatment effect that is given by β in the model $Y_i = \alpha + \beta W_i + \gamma_1 (X_{fi} - c) + \gamma_2 W_i \cdot (X_{fi} - c) + \epsilon_i$. It is not necessary to include other covariates in the regression, even if those covariates are important in the selection criterion. However, including available covariates can help reduce any small-sample bias (see Imbens and Lemieux 2008). Finally, one would also need to consider how to determine the optimal intervals around the cut-off point. We refer readers to Imbens and Kalyanaraman (2012) for more detailed discussion.

Solution 2: Difference-in-differences (DID) method . We discussed the usefulness of the PSM model and RDD for causal inference on the treatment effect using cross-sectional data. If there are observations on pre- and post-treatment periods for both the treatment and control group firms, then the DID method is a very useful technique for estimating the causal effect of the treatment. This method allows for the calculation of the treatment effect by comparing the difference in the change in average outcomes over the two periods for the treatment and control group firms.

We continue with the same example of estimating the treatment effect of obtaining ISO 9000 certification. One needs to address the potential self-selection of high-performing firms into the treatment group (ISO certified) and low-performing firms into the control (non-certified) group. Suppose that we observe data on these two groups of firms before and after ISO certification was

obtained. Let $G_i = \{0, 1\}$ denote the group that firm i belongs to, where $G = 0$ is for the control group and $G = 1$ is for the treatment group. Let $T_i = \{0, 1\}$ denote the observation period for firm i, where $T = 0$ is the pre-certification period and $T = 1$ is the post-certification period. In this case, we can define $W_{iT_i} \equiv G_i \cdot T_i$, which takes a value of 1 if firm i is in the treatment group and the observation is for the post-treatment period, and 0 otherwise. Given this setting, we can obtain the DID estimate of the treatment effect on outcome Y_i through the coefficient of W_{iT_i} in the regression, $Y_{iT_i} = \alpha + \beta_{DID} W_{iT_i} + \gamma G_i + \delta T_i + \epsilon_{iT_i}$. It follows from this equation that, $\beta_{DID} = \Big(E[Y|G = 1, T = 1] - E[Y|G = 1, T = 0]\Big) - \Big(E[Y|G = 0, T = 1] - E[Y|G = 0, T = 0]\Big)$.

Essentially, β_{DID} captures the difference between two components. The first component is the difference in population average outcomes between the pre- and post-treatment periods for the firms in the treatment group, while the second component is the difference in population average outcomes between the pre- and post-treatment periods for the firms in the control group. We illustrate this in Figure 5. Suppose the average output of firms in group $G = 0$ increased by 100, from 300 in $T = 0$ to 400 in $T = 1$. On the other hand, the average output of firms in group $G = 1$ increased by 200, from 600 in $T = 0$ to 800 in $T = 1$. Inferring that all of this output increase was due to obtaining the certification would clearly be misleading. Instead, we can assume that the output increase in group $G = 1$ would have followed the same time trend as firms in group $G = 0$, had the former not obtained the certification. Specifically, if firms in $G = 1$ had not obtained the certification, they would have potentially produced 700 in $T = 1$. The DID estimate of the treatment effect of certification is thus calculated as the difference between the actual and potential outcome for firms in $G = 1$, which in this case is 100. The direct DID estimate of the treatment effect uses the sample averages before and after treatment according to the formula given above. In practice, researchers usually adopt the regression approach using the regression model above, as it allows for the incorporation of control variables.

– Figure 5 here –

While the DID regression model has been widely used, it also faces two common critiques. First, researchers are often asked to justify the validity of the DID model. Specifically, the most important assumption of the DID model is that of the parallel trend between the control and treatment groups.[9] To examine the DID model's validity, if the data span multiple periods, one can visually examine

[9]This assumption is more plausible when the observable heterogeneity among the firms is explicitly taken into account in the estimation model. Angrist and Pischke (2009) provide a discussion on the types of variables that are acceptable as control variables in a DID framework.

the time trends in the treatment and control groups. In addition, one can conduct a placebo test by re-estimating the model with randomly chosen treated groups. A valid DID design would generate zero causal effect of any "placebo treatment". The second critique to the DID model is about the estimated standard errors of the treatment effect, and these need to be adjusted for potential serial correlation. Bertrand et al. (2004) discuss how to estimate the correct standard errors using block bootstrap or arbitrary variance-covariance matrix corrections.

5 Does big data alleviate causal inference concerns?

Wide-scale digitization of information is rapidly changing the data landscape. Digital data is now collected and stored at unprecedented levels of variety, velocity and volume (3Vs), and we are now unquestionably in the 'big data' era. While big data has become the buzz word in empirical research, the challenges to causal inference, such as endogeneity and selection-bias, still exist no matter how big the 3Vs of the data are. However, the richness of the data can help researchers better address these challenges, and we believe that the combination of big data and causal inference tools represents the future of empirical OM research.

To illustrate our point better, we define the 3Vs of big data in the OM context. A relevant dataset for OM usually consists of various information on a set of firms or individuals, possibly over several time periods. Let the number of firms be N, the number of time periods be T, and the set of information be M. We will characterize a dataset to be big data if N, T, M are large, which includes that (1) the dimension of information M is large for each firm or individual (variety), (2) the sampling rate $(1/T)$ is granular with observations available for every instance of time within the sample period (velocity), and (3) the sample size N constitutes a large proportion of the population (volume). We use an example to demonstrate how the 3Vs may help causal inference models.

Example: IV estimation to address the endogeneity bias Our example relates to services management and investigates how a customer's choice to join an on-call taxi queue is determined by the number of empty taxis. In the taxi service context, there is a pool of taxi drivers who serve a common group of customers. The customers can choose to be in either of two queues: a physical queue and an on-call queue. In the physical queue, the waiting time is uncertain, whereas in the on-call queue, the waiting time is more certain once the customer receives a booking confirmation. Customers have to pay an additional fee in order to join the on-call queue, so that taxi drivers also earn more revenue per trip from an on-call booking. Our research question carries policy

implications on optimal price setting and taxi allocation for the taxi operator. For instance, if the number of empty taxis is found to have a negative impact on on-call sales, then the taxi operator should keep the number of empty taxis low to boost on-call revenues.

Our dataset contains electronic records of taxi usage recorded by 3,184 taxi drivers over the span of three consecutive months, in a city with 5.5 million people spread across 29 geographical districts. The 3,184 taxi drivers represent 12% of all taxi drivers in this city. The taxi usage records were collected from information stored in a console installed in each taxi. The console updates the information every 15 seconds. The recorded information includes: taxi location (via GPS) and vehicle status (*i.e.*, on break, empty, on-call or passenger on board). Our dataset consists of 588,764 GPS location points and 18,747,792 observations on vehicle status. Given the nature of the dataset, we consider this to be an example of big data with the 3V characteristics. For our study, we can construct the following variables for each district and each hour in a day:

- On-call sales: This is the dependent variable. We calculate this as the number of fulfilled requests for taxis through the on-call service, for each district and hour of the day.

- Number of empty taxis: This is the key explanatory variable. We calculate this in two steps. In the first step, we calculate the number of sessions a taxi recorded the status as "free" in each district for each hour of the day. Then we compute the total number of "free" sessions for all the taxis in each district for each hour.

- Average speed (km/hour) of the taxis: This is a control variable that captures congestion. For each taxi we calculate the average speed while located in a district during each hour of the day. We then compute the average speed of all taxis in each district for each hour of the day.

- Average duration (in minutes) of on-call trips and non-on-call trips: We use these as instrumental variables. For each trip originating from a district in a given hour, we calculate the duration of the trip. Then we compute the average duration of the on-call and non-on-call trips.

In our study we examine the effect of the number of empty taxis on on-call sales. However, our dependent variable is potentially censored. This is because we do not observe unfulfilled demand for on-call taxi services. To address this, we use data only from the off-peak hours (all hours except for 8:00am-10:00am and 5:00pm-7:00pm on weekdays), because the demand for on-call service during

these hours is almost always met (*i.e.*, the potential censoring problem is more likely to occur during the peak hours).

We assume that customers who wish to take a taxi can correctly estimate the number of empty taxis in their district during a given hour. We wish to test whether this estimate has any influence on their decision to join the on-call queue. Other observable factors that possibly influence their decision include their location, the time of day, the day of the week and the level of congestion. Let Y_{ldt}, X_{ldt} and Z_{ldt} denote the on-call sales, the number of empty taxis and the average speed of taxis (proxy for congestion), respectively, at location l, day of the week d and hour t. Let $\mathbf{L}$ be the vector of dummy variables for the districts, $\mathbf{D}$ be the vector of dummy variables for the days in the week and $\mathbf{H}$ be the vector of dummy variables for the hours in a day. Finally, let $\epsilon_{Y,ldt}$ denote the random error term. A linear regression model for causal inference would be as follows:

$$Y_{ldt} = \beta_0 + \beta_1 X_{ldt} + \beta_2 Z_{ldt} + \boldsymbol{\delta}_1' \mathbf{L} + \boldsymbol{\delta}_2' \mathbf{D} + \boldsymbol{\delta}_3' \mathbf{H} + \epsilon_{Y,ldt}. \tag{5}$$

In the model above, the parameter β_1 captures the causal effect of the number of empty taxis on on-call sales. This is our parameter of interest.

However, an important question to ask is whether the key explanatory variable (X_{ldt}) is affected by the dependent variable (Y_{ldt}). In our case, it is possible that the number of empty taxis is itself determined by on-call sales. In general, the taxi drivers would incorporate their knowledge of the demand side for their supply choices. In other words, we cannot rule out simultaneity in the above regression model. If there is simultaneity between these two variables, a direct estimation of this model using OLS regression will not yield a consistent estimate of β_1.

In order to address the simultaneity bias problem in the estimation of β_1, we need at least one instrumental variable (IV) that is correlated with X_{ldt} but not with $\epsilon_{Y,ldt}$. Therefore, the IV should not affect on-call demand, but should affect taxi availability in the vicinity of a waiting customer. Once a suitable IV is found, we can then proceed to estimate the regression model (5) using the 2SLS approach. Below, we demonstrate how the three attributes of big data helped us in implementing this procedure.

Variety: The most crucial component in addressing the issue of potential endogeneity is the availability of IVs. A greater variety of data allows us to find suitable IVs. In our context, we need IVs that affect the taxi drivers' decisions, but not those of the customers. Specifically, a taxi driver's decision can be seen as having to choose among three mutually exclusive options: accept an on-call request, accept a passenger from the physical queue or roadside, or remain empty for the time

being. We assume that the taxi drivers form an expectation of the benefits and costs for these options before making a choice. A key factor in whether to accept an on-call request is the expected revenue from the trip, which in turn depends on the expected duration of the on-call trip. Note that the destinations of most on-call requests are revealed to the taxi drivers before they decide whether to accept the request. Similarly, the expected revenue of a non-on-call pick-up is dependent on the expected duration of such a trip. Therefore, we can assume that the observed number of empty taxis is related to the expected durations of on-call and non-on-call trips originating in each location, day and hour. In fact, we hypothesize that trips with potentially longer durations are more attractive for the taxi drivers during the off-peak hours, since there are fewer customers during these hours. Consequently, the correlation between the expected durations (for both on-call and non-on-call trips) and the number of empty taxis should be negative.

To operationalize this conceptual connection between the taxi driver's decision to remain empty and the expected durations of the trips in the estimation model, we assume that the actual mean durations of the trips can be a reasonable proxy for the drivers' expectations. That is, we assume that on average, the taxi drivers make correct estimates about the trip durations. Due to the variety in data, we are able to calculate the amount of time taken until completion for all the trips originating in any location, district and hour. Then we calculate the average durations of on-call and non-on-call trips for each location, day and hour combination. Importantly, these variables (*i.e.*, the average durations) are not observable by the passengers, and thus can only affect the supply of empty taxis, and not the demand for on-call taxis. Hence, we can use these variables as IVs.

Before applying the IV-2SLS method to estimate the causal effect of the number of empty taxis on on-call sales, we estimate the model in equation (5) using OLS. We log transform the variables Y_{ldt}, X_{ldt} and Z_{ldt} in our estimation, so that our coefficient of interest β_1 captures the elasticity of on-call sales. In Table 4, column 1 displays the OLS estimates, which are potentially biased, and column 2 displays the estimates using IV-2SLS, which corrects for endogeneity bias. Our results indicate that on-call sales decrease in locations where there are more empty taxis. Specifically, the bias-corrected coefficient is -2.739 (s.e. 0.134), which suggests that during off-peak hours, if the number of empty taxis increases by 1%, then on-call sales decrease by 2.739%. In contrast, the estimated coefficient using OLS is -0.132 (s.e. 0.008), which is upward biased.

Velocity: High velocity data permits the investigation of questions that are not feasible with low velocity data. Suppose that instead of having data at the hourly level, we only had access to lower

velocity data at the daily level. In this case, we would lose important variation in both the dependent and independent variables, and would also lose the hourly dummies as control variables. In Table 4, column 3 reports the estimates generated by the low-velocity data. Note that the estimated effect of empty taxis on on-call sales is markedly different with the low velocity data, with the sign of the coefficient reversed. A major driver of this is that an important set of control variables, the hourly dummies, is absent in the model. In the model estimated on the high velocity data, we find that many of the hourly dummies are statistically significant. We also note that the Sargan statistic for the over-identification test indicates that the null hypothesis of the instruments being jointly valid is rejected. In other words, the IVs are not admissible using the low velocity data.

Volume: The primary advantage of having a high volume dataset is that it helps in generating more precise estimates. To see this, in column 4 of Table 4 we report the model estimates obtained by using a randomly generated sub-sample of our data, set at 10% of the size of the original data. Although the estimated effect of empty taxis on on-call sales is similar to that for the full sample, the standard errors with the smaller sample are more than twice as large. In some cases, the larger standard errors due to lower-volume data may result in low power for testing the effect of an explanatory variable on the dependent variable.

6 Research implications for OM and Conclusion

An important question is which areas within OM will benefit from more causal empirical research. To examine this, we review the empirical literature published in three top OM journals, MS, MSOM and POM, from 2000 to 2015.[10] We include only the papers which focus on establishing causal relationships among variables of interest using observational data; thus, papers which are experiment-based, simulation-based or forecasting-oriented, and those employing SEM are excluded. For papers published in 2000-2009, we include only those with Google Scholar citation counts of 50 or more (as of May 2017). For articles published during 2010-2015, we include only those with 30 or more Google Scholar citation counts. We categorize these papers into the following broad topics: supply chain management, quality management, services management and retailing, pricing and revenue management, workforce management, and miscellaneous topics. We list these papers in Tables A1 to A6 in the Appendix.

Supply chain management (Table A1). The major research questions in inventory manage-

[10]For MS, we include empirical papers accepted in the OM department, based on the names of the department editors. Since the department editor information is available only from 2004, we exclude articles published during 2000-2003.

ment have centered on (a) which factors are the critical determinants of inventory, and (b) how various aspects of inventory management impact firm performance.[11] Gaur et al. (2005) show that inventory is simultaneously determined with other firm operational outcomes, which suggests that simultaneous equations models would be appropriate empirical tools. Kesavan et al. (2010) apply simultaneity bias correction in their estimation model to study the relationship among cost of goods sold, gross margin and inventory levels.

Other papers on inventory management employ single equation models, which include Ton and Raman (2010), Randall et al. (2006), Rumyantsev and Netessine (2007), and Jain et al. (2013). Industry-specific papers include Olivares and Cachon (2009) and Cachon and Olivares (2010), which identify the key drivers that explain the variation in the finished goods inventory within the automotive distribution system. There is also work which investigates the factors that affect inventory record accuracy (DeHoratius and Raman (2008)) and when managers decide to deviate from the inventory recommendations of an automated ordering system (van Donselaar et al. (2010)).

Since simultaneous equations models can be more useful in modeling inventory decisions, it is important to use appropriate IVs and exclusion restrictions for the identification and estimation of the parameters associated with the endogenous variables. Consequently, access to a greater variety of data will be particularly useful. In addition, access to a greater velocity of data will allow for the construction of more precise variables. For instance, van Donselaar et al. (2010) explain how access to a more accurate inventory data would have improved the construction of their key dependent variable.

Other well-cited empirical work in the supply chain management area include topics on purchase operations and e-procurement (Boyer and Olson (2002), Mithas and Jones (2007)), the bullwhip effect (Cachon et al. (2007), Bray and Mendelson (2012)), the effect of supply disruptions on stock market and firm performance (Hendricks and Singhal (2005a), Hendricks and Singhal (2005b)), buyer-supplier relationships and communications (Terwiesch et al. (2005), Jira and Toffel (2013)), and the effect of adopting a buy-online-pickup-in-store distribution strategy (Gallino and Moreno (2014)). The questions that study both the buyers' and the suppliers' decisions are potentially subject to endogeneity issues, as one side's decisions can affect those of the other side. Moreover, research that measures the impact of adopting particular procurement or distribution strategies

[11] The early empirical papers were primarily concerned with examining inventory performance over time (*e.g.* Rajagopalan and Malhotra (2001), Chen et al. (2005), Chen et al. (2007)). More recently, researchers have shifted towards establishing causal relationships between inventory and other operational variables.

must account for potential selection bias (Gallino and Moreno (2014)). Finally, we note that while there are many analytical models on inventory management, there are only a handful of empirical studies that directly test the predictions of these models (*e.g.*, Olivares et al. (2008)).

Quality management (Table A2). The empirical papers published in the quality management area can be divided into two groups. The first group focuses on examining the impact of quality standards on firm outcomes. Corbett et al. (2005) use the event study method to investigate the impact of ISO 9000 certification on financial performance. Levine and Toffel (2010) use a matching model to examine how the adoption of ISO 9001 quality management standards affects employee outcomes such as employee earnings, turnover and safety. Thirumalai and Sinha (2011) examine what happens to firms when quality drops, and applies it to the context of product recalls of medical devices. They apply an event study methodology and find that the product recalls do not have a significant impact on aggregate stock returns. Guajardo et al. (2015) examine the moderating effect of quality on the effects of service attributes on demand in the auto sector. They use the random coefficient model following Berry et al. (1995), which has been widely applied in economics and marketing for demand analysis. An advantage of the approach of Berry et al. (1995) is that it suggests various instruments that can be generated from the data.

The second group of papers examine quality outcomes in the healthcare context. The papers that have attempted to address the omitted variables problem include Kc and Terwiesch (2011)), Kc and Terwiesch (2012) and Kim et al. (2015). They all use IV estimation to investigate the causal effects of hospital focus, discharge strategies and admission policies. Other studies include Theokary and Ren (2011) who investigate whether hospital volume and teaching status affect service quality, and Chandrasekaran et al. (2012) who use a random-effects regression to examine how process management affects the quality of hospital stays.

Services management and retailing (Table A3). In the area of services management, the main topics that have been empirically examined are: (a) the causal relationship between customer satisfaction and various operational metrics, and (b) the effects of introducing self-service techonology and e-services. Two notable empirical studies on the first topic are by Lapré (2011) and Gu and Ye (2014). The notable papers on the second topic include those by Tsikriktsis et al. (2004), Xue et al. (2007), Buell et al. (2010) and Campbell and Frei (2010). In particular, Campbell and Frei (2010) examine how self-service (specifically, online banking) alters services consumption, cost to serve, and customer profitability. Because customers self-select whether to adopt online banking, the authors use the propensity score matching method to control for selection.

In the area of retailing, Heim and Sinha (2001) use data from 52 electronic food retailers to explore how website navigation, timeliness of delivery, and ease of return affect customer loyalty. Chong et al. (2001) use shopping trips and purchase records to estimate a category assortment model which can be used by managers to assess the revenue implications of alternative category assortments. Perdikaki et al. (2012) use a dynamic panel data model to study the relationship between store traffic, labor, and sales performance. They find that while the effect of store traffic on sales exhibits diminishing returns, having more in-store labor can alleviate this effect.

Pricing and revenue management (Table A4). Notable works include Anderson and Xie (2012), Subramanian and Subramanyam (2012), Li et al. (2014) and Phillips et al. (2015). In general, the problems examined in this literature relate to the factors that affect pricing decisions and how different pricing schemes affect revenues. For example, Phillips et al. (2015) investigate the effect of centralized versus decentralized pricing strategies on firm performance. They noted the potential endogeneity of price and consumer response, and address this using a control function approach in their model. Notice that pricing mechanisms are typically self-selected by firms. Therefore, the tools discussed in section 4 – propensity score matching, difference-in-differences and regression discontinuity design, would be particularly germane.

Workforce management (Table A5). There has also been recent empirical work examining how firms can better manage their workforce to achieve better outcomes (see Siemsen et al. (2009), Bendoly (2014), Narayanan et al. (2011) and Staats (2012)). Kc and Terwiesch (2009), Powell et al. (2012) and Kc (2014) examine the effect of workload on various operational outcomes in the healthcare sector. Kc and Staats (2012) and Staats and Gino (2012) examine the impact of specialization on performance. Another line of research examines how the structure of teams, specifically the team's diversity, affects firm performance. Huckman and Staats (2011) answer this question in the context of fluid teams (*i.e.*, teams which are formed and dissembled quickly) in the software services industry. Green et al. (2013) examine how the workload of the team affects the participation of team members. Although the endogeneity issue is likely to be absent given the specific contexts of their studies, future research should account for the fact that workers self-select into teams and that there are potentially unobserved factors that guide how managers construct teams.

Miscellaneous topics (Table A6). Other OM topics that have been empirically investigated, but with a relatively fewer number of papers, include innovation management, information technology management, queue management and operations strategy. In innovation management, two notable

studies are those by Bajaj et al. (2004) and Boudreau et al. (2011). Bajaj et al. (2004) assess the impact of management levers such as oversight, design specialization, and customer interaction on cost savings and scheduling during the two phases of new product development – the design phase and the manufacturing phase. Since the design phase outcomes are explanatory variables in the manufacturing phase regressions, the authors employ a two-stage regression approach. Boudreau et al. (2011) use data from software contests to examine the optimal number of competitors in innovation contests. They rely on their quasi-experimental setting to directly estimate the effect of the number of competitors on innovation outcomes.

The well-cited empirical papers in IT management include Ahmad and Schroeder (2001) and McAfee (2002). McAfee (2002) exploits a natural experiment conducted at a U.S. high-tech manufacturer and documents the longitudinal effect of IT adoption on operational performance. The event-study approach used in the paper follows a DID design, where the causal effect of IT adoption is captured by the difference between the control and treatment groups. The causal link between queuing time and firm revenues is an important topic in queue management research. Recent well-cited studies on this topic include Allon et al. (2011), Lu et al. (2013), Akşin et al. (2013), Batt and Terwiesch (2015) and Song et al. (2015). Song et al. (2015) use a DID model to show that a dedicated queuing system reduces waiting time relative to a pooled queuing system. On issues related to operations strategy, empirical researchers have examined how various strategies affect operational performance (see Stratman (2007), Lapré and Scudder (2004), Tsikriktsis (2007), Rawley and Simcoe (2010) and Kroes et al. (2012)). In particular, Rawley and Simcoe (2010) investigate whether firm diversification leads to increased outsourcing. They employ both the PSM model and IV regression to account for the fact that firms self-select into diversifying their operations.

Conclusion. We conclude this paper with the following remarks. First, the five econometric tools discussed in this paper are non-exhaustive. We chose to focus on these because they are quite intuitive and have been applied to many empirical contexts. For a comprehensive discussion on causal inference models, readers can consult Angrist and Pischke (2009) and Imbens and Rubin (2015). Second, when applying causal inference models to analyzing big data, there are high-dimensional econometric and machine learning techniques such as LASSO (least absolute shrinkage and selection operator), post double selection method, random-forest and bagging (bootstrap aggregating), that researchers can use to handle large datasets. Interested readers can refer to Tibshirani (1996), Belloni et al. (2013), Belloni et al. (2014) and Varian (2014) for discussions on such methods. These methods have yet to see widespread applications. However, with data sizes getting increasingly

larger, we foresee greater adoption of these methods in empirical OM research over the next few years.

References

Ahmad, S. and Schroeder, R. G. (2001). The impact of electronic data interchange on delivery performance. *Production and Operations Management*, 10(1):16–30.

Akşin, Z., Ata, B., Emadi, S. M., and Su, C. L. (2013). Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746.

Allon, G., Federgruen, A., and Pierson, M. (2011). How much is a reduction of your customers' wait worth? An empirical study of the fast-food drive-thru industry based on structural estimation methods. *Manufacturing & Service Operations Management*, 13(4):489–507.

Anderson, C. K. and Xie, X. (2012). A choice-based dynamic programming approach for setting opaque prices. *Production and Operations Management*, 21(3):590–605.

Angrist, J. D. and Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Bajaj, A., Kekre, S., and Srinivasan, K. (2004). Managing NPD: Cost and schedule performance in design and manufacturing. *Management Science*, 50(4):527–536.

Batt, R. J. and Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59.

Belloni, A., Chernozhukov, V., and Hansen, C. (2013). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bendoly, E. (2014). System dynamics understanding in projects: Information sharing, psychological safety, and performance effects. *Production and Operations Management*, 23(8):1352–1369.

Bennett, V. M., Pierce, L., Snyder, J. A., and Toffel, M. W. (2013). Customer-driven misconduct: How competition corrupts business practices. *Management Science*, 59(8):1725–1742.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, pages 841–890.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, 119(1):249–275.

Boudreau, K. J., Lacetera, N., and Lakhani, K. R. (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, 57(5):843–863.

Boyer, K. K. and Olson, J. R. (2002). Drivers of internet purchasing success. *Production and Operations Management*, 11(4):480–498.

Bray, R. L. and Mendelson, H. (2012). Information transmission and the bullwhip effect: An empirical investigation. *Management Science*, 58(5):860–875.

Buell, R. W., Campbell, D., and Frei, F. X. (2010). Are self service customers satisfied or stuck? *Production and Operations Management*, 19(6):679–697.

Cachon, G. P. and Olivares, M. (2010). Drivers of finished-goods inventory in the US automobile industry. *Management Science*, 56(1):202–216.

Cachon, G. P., Randall, T., and Schmidt, G. M. (2007). In search of the bullwhip effect. *Manufacturing & Service Operations Management*, 9(4):457–479.

Campbell, D. and Frei, F. (2010). Cost structure, customer profitability, and retention implications of self-service distribution channels: Evidence from customer behavior in an online banking channel. *Management Science*, 56(1):4–24.

Caro, F., Martínez-de Albéniz, V., and Rusmevichientong, P. (2014). The assortment packing problem: Multiperiod assortment planning for short-lived products. *Management Science*, 60(11):2701–2721.

Chandrasekaran, A., Senot, C., and Boyer, K. K. (2012). Process management impact on clinical and experiential quality: Managing tensions between safe and patient-centered healthcare. *Manufacturing & Service Operations Management*, 14(4):548–566.

Chen, H., Frank, M. Z., and Wu, O. Q. (2005). What actually happened to the inventories of American companies between 1981 and 2000? *Management Science*, 51(7):1015–1031.

Chen, H., Frank, M. Z., and Wu, O. Q. (2007). US retail and wholesale inventory performance from 1981 to 2004. *Manufacturing & Service Operations Management*, 9(4):430–456.

Chong, J.-K., Ho, T.-H., and Tang, C. S. (2001). A modeling framework for category assortment planning. *Manufacturing & Service Operations Management*, 3(3):191–210.

Corbett, C. J., Montes-Sancho, M. J., and Kirsch, D. A. (2005). The financial impact of ISO 9000 certification in the United States: An empirical analysis. *Management Science*, 51(7):1046–1059.

DeHoratius, N. and Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science*, 54(4):627–641.

Enders, W. (2004). Applied time series econometrics. *Hoboken: John Wiley and Sons. ISBN X*, 52183919.

Gallino, S. and Moreno, A. (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science*, 60(6):1434–1451.

Gaur, V., Fisher, M. L., and Raman, A. (2005). An econometric analysis of inventory turnover performance in retail services. *Management Science*, 51(2):181–194.

Golrezaei, N., Nazerzadeh, H., and Rusmevichientong, P. (2014). Real-time optimization of personalized assortments. *Management Science*, 60(6):1532–1551.

Gray, J. V., Anand, G., and Roth, A. V. (2015). The influence of ISO 9000 certification on process compliance. *Production and Operations Management*, 24(3):369–382.

Green, L. V., Savin, S., and Savva, N. (2013). "Nursevendor problem": Personnel staffing in the presence of endogenous absenteeism. *Management Science*, 59(10):2237–2256.

Greene, W. H. (2000). *Econometric Analysis*. Prentice Hall.

Gu, B. and Ye, Q. (2014). First step in social media: Measuring the influence of online management responses on customer satisfaction. *Production and Operations Management*, 23(4):570–582.

Guajardo, J. A., Cohen, M. A., and Netessine, S. (2015). Service competition and product quality in the US automobile industry. *Management Science*, 62(7):1860–1877.

Heim, G. R. and Sinha, K. K. (2001). Operational drivers of customer loyalty in electronic retailing: An empirical analysis of electronic food retailers. *Manufacturing & Service Operations Management*, 3(3):264–271.

Hendricks, K. B. and Singhal, V. R. (2005a). Association between supply chain glitches and operating performance. *Management Science*, 51(5):695–711.

Hendricks, K. B. and Singhal, V. R. (2005b). An empirical analysis of the effect of supply chain disruptions on long-run stock price performance and equity risk of the firm. *Production and Operations Management*, 14(1):35–52.

Huckman, R. S. and Staats, B. R. (2011). Fluid tasks and fluid teams: The impact of diversity in experience and team familiarity on team performance. *Manufacturing & Service Operations Management*, 13(3):310–328.

Hyndman, K. and Parmeter, C. F. (2015). Efficiency or competition? A structural econometric analysis of Canada's AWS auction and the set-aside provision. *Production and Operations Management*, 24(5):821–839.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.

Imbens, G. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

Imbens, G. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Jain, N., Girotra, K., and Netessine, S. (2013). Managing global sourcing: Inventory performance. *Management Science*, 60(5):1202–1222.

Jira, C. and Toffel, M. W. (2013). Engaging supply chains in climate change. *Manufacturing & Service Operations Management*, 15(4):559–577.

Kc, D. S. (2014). Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management*, 16(2):168–183.

Kc, D. S. and Staats, B. R. (2012). Accumulating a portfolio of experience: The effect of focal and related experience on surgeon performance. *Manufacturing & Service Operations Management*, 14(4):618–633.

Kc, D. S. and Terwiesch, C. (2009). Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science*, 55(9):1486–1498.

Kc, D. S. and Terwiesch, C. (2011). The effects of focus on performance: Evidence from California hospitals. *Management Science*, 57(11):1897–1912.

Kc, D. S. and Terwiesch, C. (2012). An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management*, 14(1):50–65.

Kesavan, S., Gaur, V., and Raman, A. (2010). Do inventory and gross margin data improve sales forecasts for US public retailers? *Management Science*, 56(9):1519–1533.

Kim, S. H., Chan, C. W., Olivares, M., and Escobar, G. (2015). ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science*, 61(1):19–38.

Kim, S. W., Olivares, M., and Weintraub, G. Y. (2014). Measuring the performance of large-scale combinatorial auctions: A structural estimation approach. *Management Science*, 60(5):1180–1201.

Kroes, J., Subramanian, R., and Subramanyam, R. (2012). Operational compliance levers, environmental performance, and firm performance under cap and trade regulation. *Manufacturing & Service Operations Management*, 14(2):186–201.

Lapré, M. A. (2011). Reducing customer dissatisfaction: How important is learning to reduce service failure? *Production and Operations Management*, 20(4):491–507.

Lapré, M. A. and Scudder, G. D. (2004). Performance improvement paths in the US airline industry: Linking trade-offs to asset frontiers. *Production and Operations Management*, 13(2):123–134.

Levine, D. I. and Toffel, M. W. (2010). Quality management and job quality: How the ISO 9001 standard for quality management systems affects employees and employers. *Management Science*, 56(6):978–996.

Li, J., Granados, N., and Netessine, S. (2014). Are consumers strategic? Structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137.

Lu, Y., Musalem, A., Olivares, M., and Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science*, 59(8):1743–1763.

Mani, V., Kesavan, S., and Swaminathan, J. M. (2015). Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management*, 24(2):201–218.

McAfee, A. (2002). The impact of enterprise information technology adoption on operational performance: An empirical investigation. *Production and operations management*, 11(1):33.

Mithas, S. and Jones, J. L. (2007). Do auction parameters affect buyer surplus in E-auctions for procurement? *Production and Operations Management*, 16(4):455–470.

Narayanan, S., Balasubramanian, S., and Swaminathan, J. M. (2011). Managing outsourced software projects: An analysis of project performance and customer satisfaction. *Production and Operations Management*, 20(4):508–521.

Olivares, M. and Cachon, G. P. (2009). Competing retailers and inventory: An empirical investigation of General Motors' dealerships in isolated US markets. *Management Science*, 55(9):1586–1604.

Olivares, M., Terwiesch, C., and Cassorla, L. (2008). Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Science*, 54(1):41–55.

Perdikaki, O., Kesavan, S., and Swaminathan, J. M. (2012). Effect of traffic on sales and conversion rates of retail stores. *Manufacturing & Service Operations Management*, 14(1):145–162.

Phillips, R., Şimşek, A. S., and Van Ryzin, G. (2015). The effectiveness of field price discretion: Empirical evidence from auto lending. *Management Science*, 61(8):1741–1759.

Powell, A., Savin, S., and Savva, N. (2012). Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management*, 14(4):512–528.

Rajagopalan, S. and Malhotra, A. (2001). Have US manufacturing inventories really decreased? An empirical study. *Manufacturing & Service Operations Management*, 3(1):14–24.

Randall, T., Netessine, S., and Rudi, N. (2006). An empirical examination of the decision to invest in fulfillment capabilities: A study of internet retailers. *Management Science*, 52(4):567–580.

Rawley, E. and Simcoe, T. S. (2010). Diversification, diseconomies of scope, and vertical contracting: Evidence from the taxicab industry. *Management Science*, 56(9):1534–1550.

Rosenbaum, P. and Rubin, D. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B (Methodological)*, 45(2):212–218.

Rumyantsev, S. and Netessine, S. (2007). What can be learned from classical inventory models? A cross-industry exploratory investigation. *Manufacturing & Service Operations Management*, 9(4):409–429.

Siemsen, E., Roth, A. V., Balasubramanian, S., and Anand, G. (2009). The influence of psychological safety and confidence in knowledge on employee knowledge sharing. *Manufacturing & Service Operations Management*, 11(3):429–447.

Song, H., Tucker, A. L., and Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053.

Staats, B. R. (2012). Unpacking team familiarity: The effects of geographic location and hierarchical role. *Production and Operations Management*, 21(3):619–635.

Staats, B. R. and Gino, F. (2012). Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science*, 58(6):1141–1159.

Stratman, J. K. (2007). Realizing benefits from enterprise resource planning: Does strategic focus matter? *Production and Operations Management*, 16(2):203–216.

Subramanian, R. and Subramanyam, R. (2012). Key factors in the market for remanufactured products. *Manufacturing & Service Operations Management*, 14(2):315–326.

Terwiesch, C., Ren, Z. J., Ho, T. H., and Cohen, M. A. (2005). An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Management Science*, 51(2):208–220.

Theokary, C. and Ren, Z. (2011). An empirical study of the relations between hospital volume, teaching status, and service quality. *Production and Operations Management*, 20(3):303–318.

Thirumalai, S. and Sinha, K. K. (2011). Product recalls in the medical device industry: An empirical exploration of the sources and financial consequences. *Management Science*, 57(2):376–392.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B (Methodological)*, 58(1):267–288.

Todd, P. E. (2010). Matching estimators. In *Microeconometrics*, pages 108–121. Springer.

Ton, Z. and Raman, A. (2010). The effect of product variety and inventory levels on retail store sales: A longitudinal study. *Production and Operations Management*, 19(5):546–560.

Tsikriktsis, N. (2007). The effect of operational performance and focus on profitability: A longitudinal study of the us airline industry. *Manufacturing & Service Operations Management*, 9(4):506–517.

Tsikriktsis, N., Lanzolla, G., and Frohlich, M. (2004). Adoption of e-processes by service firms: An empirical study of antecedents. *Production and Operations Management*, 13(3):216.

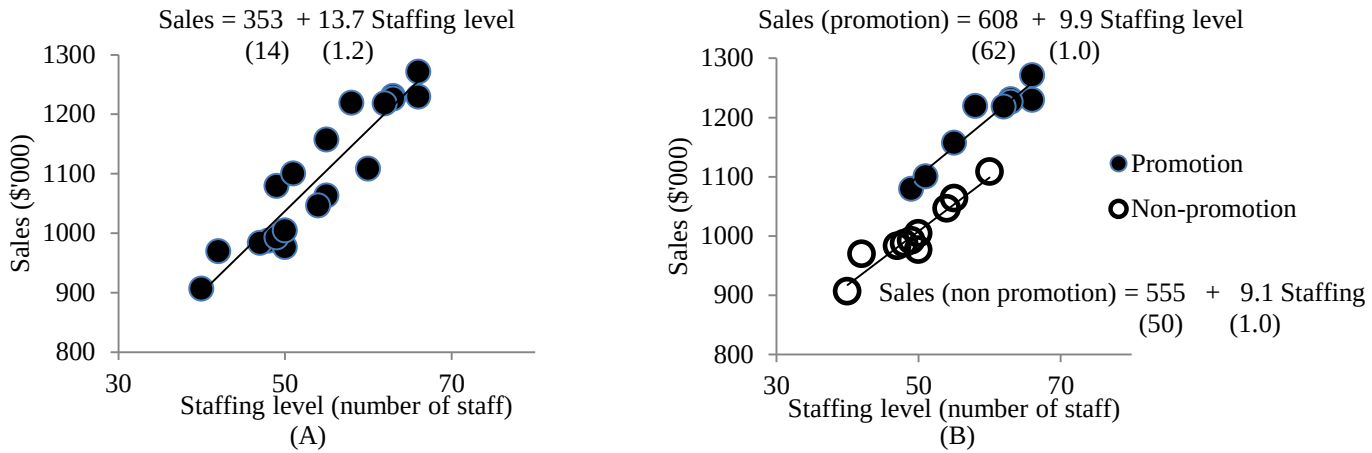
van Donselaar, K. H., Gaur, V., van Woensel, T., Broekmeulen, R. A., and Fransoo, J. C. (2010). Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56(5):766–784.

Varian, H. (2014). Big data: new tricks for econometrics. *Journal of Economic Perspective*, 28(2):3–28.

Wooldridge, J. (2008). *Econometric analysis of cross section and panel data*. MIT Press.

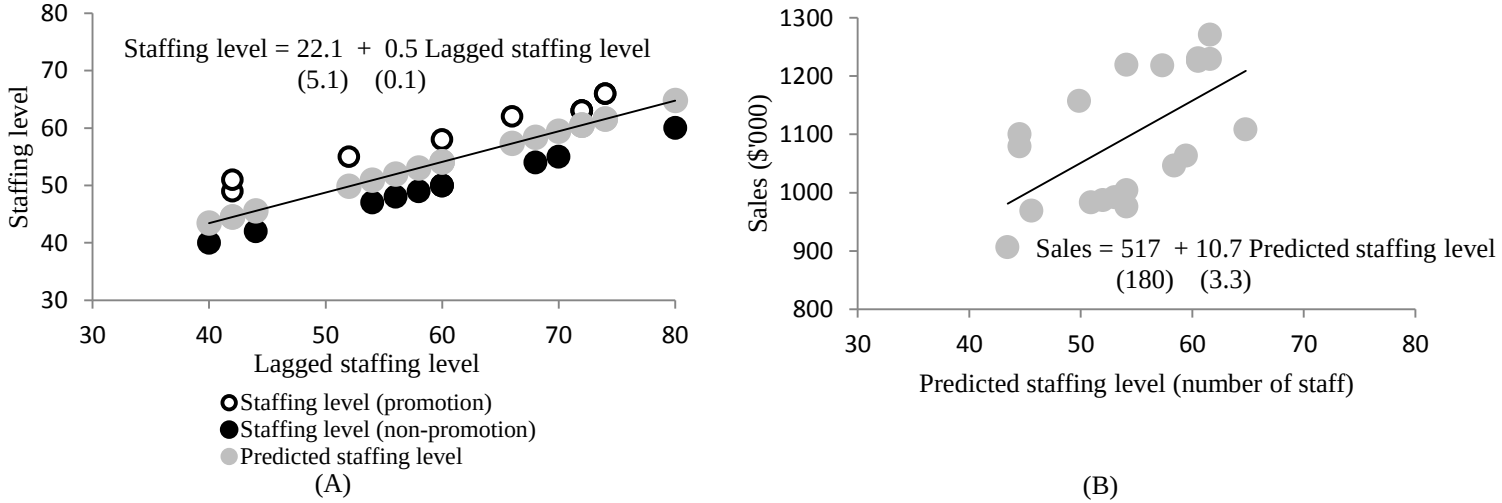
Xue, M., Hitt, L. M., and Harker, P. T. (2007). Customer efficiency, channel usage, and firm performance in retail banking. *Manufacturing & Service Operations Management*, 9(4):535–558.

Figure 1. Illustrating the omitted variable bias



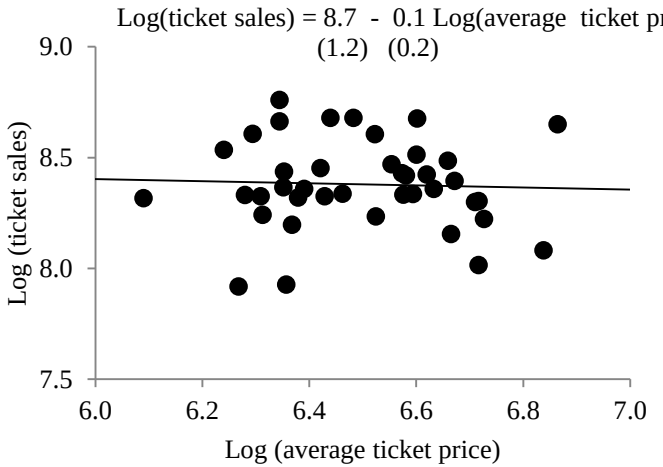
Notes: (A) Regression when the promotion variable is not available. (B) Regression when the promotion variable is available. Standard errors are shown in parentheses.

Figure 2. IV-2SLS estimation



Notes: (A) Step 1 of the IV-2SLS estimation: regression of the endogenous variable on instrumental variable. (B) Step 2 of the IV-2SLS estimation: regression of the dependent variable on the predicted values generated in step 1. Standard errors are shown in parentheses.

Figure 3. Illustration of the simultaneity bias



Notes: Based on simulated data. Standard errors are shown in parentheses.

Figure 4. Regression discontinuity design

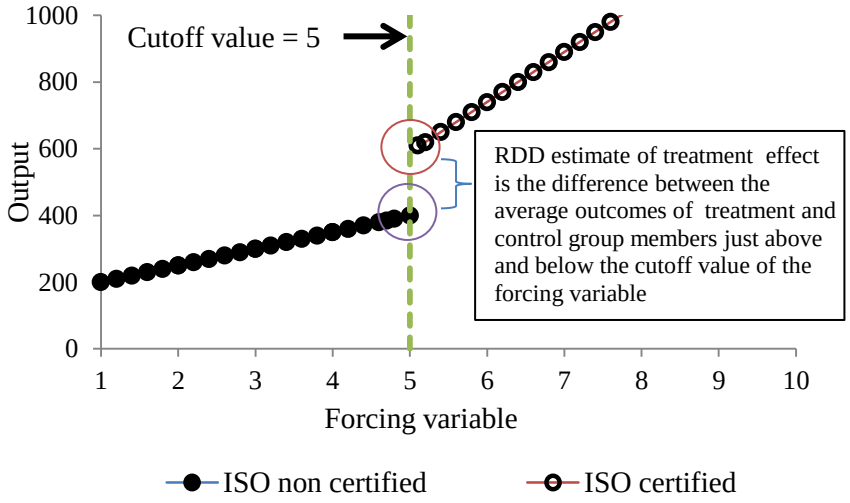


Figure 5. Difference-in-differences estimation

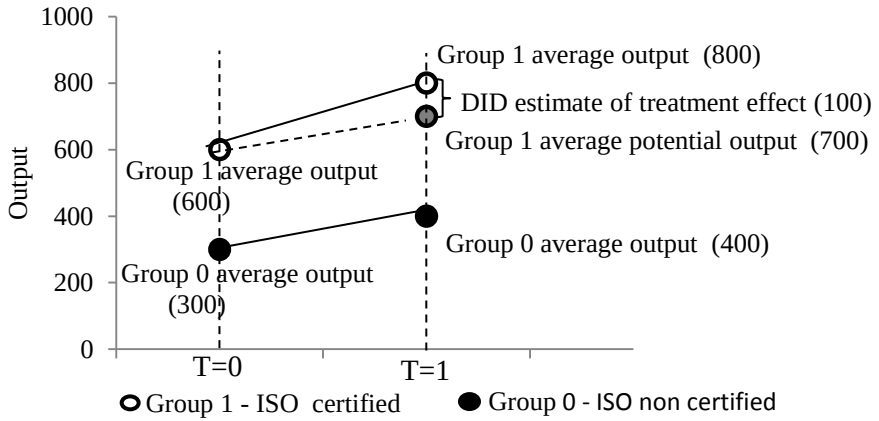


Table 1. Empirical papers in top OM journals: 2010-2015

Journal	Number of OM articles	# with empirical focus	# using observational data for causal inference	# addressing endogeneity and selection bias
MS	214	45	38	20
MSOM	245	54	34	13
POM	556	76	69	19

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Service Operations Management; *POM* – Production and Operations Management

Table 2. Estimating treatment effect of obtaining ISO certification

	Output without certification $Y_i(0)$	Output with certification $Y_i(1)$	Treatment effect $E[Y_i(1) - Y_i(0)]$
Both actual and potential outcomes are known			
Firm A (non-certified)	400	500	100
Firm B (certified)	700	800	100
Only actual outcomes are known			
Firm A (non-certified)	400	-	?
Firm B (certified)	-	800	?

Table 3. Illustration of propensity score matching model for estimation of the treatment effect of obtaining ISO certification

	Obtained certification	Output without certification	Output with certification	Propensity score	Group	Sub-group	Average output within sub-group	PSM estimate of treatment effect by sub-group
Firm 1	Yes	-	800	85%	High propensity (propensity score>50%)	Certified	700	150
Firm 2	Yes	-	700	75%				
Firm 3	Yes	-	600	70%				
Firm 4	Yes	-	700	85%				
Firm 5	No	600	-	80%		Not certified	550	
Firm 6	No	500	-	70%				
Firm 7	Yes	-	500	40%	Low propensity (propensity score<=50%)	Certified	450	75
Firm 8	Yes	-	400	30%				
Firm 9	No	450	-	50%		Not certified	375	
Firm10	No	400	-	40%				
Firm 11	No	350	-	35%				
Firm 12	No	300	-	30%				

Note: The above are based on simulated data for illustration purposes.

Table 4. Estimates of the various specifications in the example in section 5

DV: log(on-call sales)	Variety	Low	High	High	High
	Velocity	High	High	Low	High
	Volume	High	High	High	Low
		(1) OLS	(2) IV2SLS	(3) IV2SLS	(4) IV2SLS
Log(number of empty taxis)		-0.132***	-2.739***	1.002***	-2.073***
		(0.008)	(0.134)	(0.161)	(0.271)
Log(average speed)		-0.685***	-0.845***	-1.049***	0.760***
		(0.022)	(0.042)	(0.082)	(0.108)
District fixed effect (28 district dummy variables)		Yes	Yes	Yes	Yes
Day of week fixed effect (6 day dummy variables)		Yes	Yes	Yes	Yes
Hour of the day fixed effect (23 hour dummy variables)		Yes	Yes	**Not feasible**	Yes
Observations		42,116	42,116	2,592	4,211
Anderson's LM test		NA	523	103	87.410
P-value of Anderson's LM test		NA	<0.001	<0.001	<0.001
Cragg-Donald weak instrument test statistic		NA	264.4	52.9	43.98
Sargan over-identification test statistic		NA	0.016	22.42	0.088
P-value of Sargan statistic		NA	0.901	<0.001	0.767

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Log(number of empty taxis) instrumented by log(duration of on-call trips) and log(duration of non-on-call trips), where duration is the average duration of trips of on-call sales and non-on-call sales generated at each location and hour combination. Anderson's LM test is the under-identification test.

Appendix

Table A1. Empirical research using observational data for causal inference in inventory management

Journal	Year	Volume	Issue	Authors	Topic
MS	2005	51	2	Terwiesch et al.	Forecast sharing in supply chain
MS	2005	51	2	Gaur et al.	Inventory turnover performance in retail services
MS	2005	51	5	Hendricks and Singhal	Supply chain glitches and operating performance
MS	2005	51	7	Chen et al.	Inventories of American companies during 1981 - 2000
MS	2006	52	4	Randall et al.	Decision to invest in fulfillment capabilities
MS	2008	54	1	Olivares et al.	Structural estimation of the newsvendor model
MS	2008	54	4	DeHoratius and Raman	Inventory record inaccuracy
MS	2009	55	9	Olivares and Cachon	Competing retailers and inventory
MS	2010	56	1	Cachon and Olivares	Drivers of finished goods inventory
MS	2010	56	5	van Donselaar et al.	Ordering behavior and automated replenishment
MS	2010	56	9	Kesavan et al.	Inventory, gross margin and sales forecast
MS	2012	58	5	Bray and Mendelson	Information transmission and the bullwhip effect
MS	2013	60	5	Jain et al.	Managing global sourcing
MS	2014	60	6	Gallino and Moreno	Integrating online and offline channels
MSOM	2001	3	1	Rajagopalan and Malhotra	US manufacturing inventories
MSOM	2007	9	4	Cachon et al.	Bullwhip effect
MSOM	2007	9	4	Rumyantsev and Netessine	Classical inventory models
MSOM	2007	9	4	Chen et al.	US retail and wholesale inventory performance
MSOM	2013	15	4	Jira and Toffel	Climate change vulnerability information sharing
POM	2002	11	4	Boyer and Olson	Drivers of internet purchasing success
POM	2005	14	1	Hendricks and Singhal	Supply chain disruptions, stock price performance, equity risk
POM	2007	16	4	Mithas and Jones	Auction parameters, buyer surplus in e-auctions
POM	2010	19	5	Ton and Raman	Product variety and inventory level on store sales

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.

Table A2. Empirical research using observational data for causal inference in quality management

Journal	Year	Volume	Issue	Authors	Topic
MS	2005	51	7	Corbett et al.	Financial impact of ISO 9000 certification
MS	2010	56	6	Levine and Toffel	Effects of ISO 9001 on employees and employers
MS	2011	57	2	Thirumalai and Sinha	Product recalls and health risks in medical device industry
MS	2011	57	11	Kc and Terwiesch	Effect of focus on performance
MS	2015	61	1	Kim et al.	ICU admission control
MS	2015	62	7	Guajardo et al.	Service competition and product quality
MSOM	2012	14	1	Kc and Terwiesch	Patient flow in the cardiac ICU
MSOM	2012	14	4	Chandrasekaran et al.	Effect of process management on quality
POM	2011	30	3	Theokary and Ren	Patient volume, teaching quality and quality of services

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.

Table A3. Empirical research using observational data for causal inference in services management and retailing

Journal	Year	Volume	Issue	Authors	Topic
MS	2010	56	1	Campbell and Frei	Self service distribution, operational performance
MSOM	2001	3	3	Heim and Sinha	Operational drivers of customer loyalty
MSOM	2001	3	3	Chong et al.	Category assortment planning
MSOM	2007	9	4	Xue et al.	Customer efficiency, channel usage, and firm performance
MSOM	2012	14	1	Perdikaki et al.	Effect of traffic on store sales
POM	2004	13	3	Tsikriktsis et al.	Adoption of e-processes by service firms
POM	2010	19	6	Buell et al.	Self service and customer retention and satisfaction
POM	2011	20	4	Lapré	Importance of learning to reduce service failure
POM	2014	23	4	Gu and Ye	Online management response and customer satisfaction

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.

Table A4. Empirical research using observational data for causal inference in pricing and revenue management

Journal	Year	Volume	Issue	Authors	Topic
MS	2014	60	9	Li et al.	Are consumers strategic
MS	2015	61	8	Phillips et al.	Effectiveness of field price discretion
MSOM	2012	14	2	Subramanian and Subramanyam	Factors in the market for remanufactured products
POM	2012	21	3	Anderson and Xie	Dynamic programming and opaque prices

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.

Table A5. Empirical research using observational data for causal inference in workforce management

Journal	Year	Volume	Issue	Authors	Topic
MS	2009	55	9	Kc and Terwiesch	Impact of workload on service time and patient safety
MS	2012	58	6	Staats and Gino	Specialization and variety in repetitive tasks
MS	2013	59	10	Green et al.	Factors affecting nurse absenteeism
MSOM	2009	11	3	Siemsen et al.	Psychological safety, confidence and knowledge sharing
MSOM	2011	13	3	Huckman and Staats	Experience, diversity and team familiarity on performance
MSOM	2012	14	4	Kc and Staats	Effect of experience on surgeon performance
MSOM	2012	14	4	Powell et al.	Overworked physicians and hospital revenue
MSOM	2014	16	2	Kc	Multitasking and performance
POM	2011	20	4	Narayanan et al.	Drivers of project performance and customer satisfaction
POM	2012	21	3	Staats	Team familiarity and performance
POM	2014	23	8	Bendoly	Systems dynamics understanding

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.

Table A6. Empirical research using observational data for causal inference in other OM topics

Journal	Year	Volume	Issue	Authors	Topic
Innovation management					
MS	2004	50	4	Bajaj et al.	Schedule and cost in design and manufacturing
MS	2011	57	5	Boudreau et al.	Optimal number of competitors in innovation contests
Information technology management					
POM	2001	10	1	Ahmad and Schroeder	Impact of electronic data interchange
POM	2002	11	1	McAfee	Impact of enterprise information technology
Queuing					
MS	2013	59	8	Lu et al.	Effect of queues on customer purchases
MS	2013	59	12	Akşin et al.	Callers' delay sensitivity in call centers
MS	2015	61	1	Batt and Terweisch	Queue abandonment in an emergency department
MS	2015	61	12	Song et al.	Diseconomies in queue pooling
MSOM	2011	13	4	Allon et al.	Waiting time, firm valuation and pricing
Operations strategy					
MS	2010	56	9	Rawley and Simcoe	Effect of diversification on outsourcing
MSOM	2007	9	4	Tsikriktsis	Effect of operational performance and focus on profitability
MSOM	2012	14	2	Kroes et al.	Effect of emission reduction strategies on performance
POM	2003	13	2	Lapré and Scudder	US airline industry performance
POM	2007	16	2	Stratman	Enterprise resource planning: role of strategic focus

Notes: *MS* – Management Science; *MSOM* – Manufacturing and Services Operations Management; *POM* – Production and Operations Management.