Published in final edited form as:

INFORMS J Comput. 2020 ; 32(4): 1143–1156. doi:10.1287/ijoc.2019.0947.

Predictive Analytics with Strategically Missing Data

Juheng Zhang^a, Xiaoping Liu^b, Xiao-Bai Li^a

^aDepartment of Operations and Information Systems, University of Massachusetts, Lowell, Massachusetts 01854;

^bD'Amore-McKim School of Business, Northeastern University, Boston, Massachusetts 02115

Abstract

We study strategically missing data problems in predictive analytics with regression. In many real-world situations, such as financial reporting, college admission, job application, and marketing advertisement, data providers often conceal certain information on purpose in order to gain a favorable outcome. It is important for the decision-maker to have a mechanism to deal with such strategic behaviors. We propose a novel approach to handle strategically missing data in regression prediction. The proposed method derives imputation values of strategically missing data based on the Support Vector Regression models. It provides incentives for the data providers to disclose their true information. We show that with the proposed method imputation errors for the missing values are minimized under some reasonable conditions. An experimental study on real-world data demonstrates the effectiveness of the proposed approach.

Keywords

business analytics; information disclosure; data manipulation; support vector regression; strategic learning

1. Introduction

Information concealment is a familiar practice in today's competitive business environment. In the context of e-commerce marketplace, online sellers present the information of their products in a way to attract more potential buyers and higher bids (Ba 2001). Certain information about the product is strategically concealed from buyers, which may potentially affect buyers' decision-making. In other contexts, job applicants may purposely hide certain information that is unfavorable for them to get the jobs (Goldberg 2010). College applicants disclose their records selectively in response to admission policies and rules (Braun et al. 2010). Similarly, health insurance applicants may hide their lifestyle habits or preexisting conditions in order to get approval and avoid high insurance premium (Insure.com 2010). For academic rankings of universities, a college or university may intentionally choose not to report certain unfavorable information to the ranking organization in hoping for a better position (Conlin et al. 2013, Morse and Flanigan 2013). Strategic information disclosure is

Contact: juheng_zhang@uml.edu (JZ); xia.liu@northeastern.edu (XL); xiaobai_li@uml.edu (X-BL).

Supplemental Material: Data and the online appendix are available at <https://doi.org/10.1287/ijoc.2019.0947>.

also common in financial markets. As evidenced in the Facebook IPO case in 2012 (Higgins 2012), the company's business in the second quarter of 2012 was forecasted weak, but this information was not disclosed to the public at its offering. As a result, many individual investors placed bids at much overvalued prices. As the stories about Facebook's limited information disclosure began to surface, its stock price plummeted in subsequent days and months after the IPO. Information manipulation is one of the serious concerns in financial industry (Cecchini et al. 2010).

This work studies decision-making problems with strategically *hidden* data but not strategically *distorted* data. In our problem setting, data providers may hide the information, but, by rule or by commonly accepted practice, they cannot lie by intentionally providing distorted information. In many real-world situations, it is acceptable if the data providers choose not to disclose certain information. Lying, however, is typically considered “across the line.” This nature of distinction between hiding and lying applies to those examples we mentioned above. For example, in college admission where an SAT subject test report is often optional, the applicants can decide whether to report their SAT subject test scores or not, but they cannot lie about their scores. In financial reporting, some items may be optional, but organizations cannot report false information about these items.

Predictive analytics has been increasingly used in today's data-driven businesses (Ryu et al. 2004, Cecchini et al. 2010, Yang et al. 2013). In predictive analytics, there are two stages directly related to predictive models. In the *model building* stage, predictive models are developed based on training data collected in the past. In the *prediction* stage, the developed models are used to predict or evaluate the outcome of new records. High-quality data are essential in performing predictive analytics. However, missing data are common in both model building and prediction stages. When building models on data with strategic missing values, the decision-maker and analyst should consider the impact of strategic behaviors on the models under construction. A common strategy is to adjust the models, making them adaptive to the strategic behaviors (Dalvi et al. 2004, Boylu et al. 2010, Zhang et al. 2014). When using existing models for prediction or evaluation of new records, however, it is not practical to change the models to deal with strategic behaviors, because the models are already in use and often known to the data providers. This scenario has not been investigated by existing studies on strategically manipulated data. It is important to have an effective approach for evaluating new records with strategically missing values.

The decision problem we consider in this study pertains to regression prediction, which is very common in practice. In finance, regression models have been used in IPO valuation and various investment analyses (Kim and Ritter 1999, Bartov et al. 2002, Pukthuanthong-Le and Varaiya 2007, Fabozz 2008). In marketing, researchers and analysts use regression analysis to study how consumer behaviors are affected by pricing and other marketing strategies (Degeratu et al. 2000). Regression models have also been used for ranking universities (Clarke 2002, Bernat and Grimson 2011). Regression analysis is an essential methodology for empirical and experimental studies and is widely used in management, social science, and health science research. Traditional linear regression methods assume that training data follows the same distribution as unseen data. Consequently, these methods build a regression model that best fits the training data. Such a model, however, could overfit

the training data and perform poorly when applied to unseen data. A machine learning based method, called Support Vector Regression (SVR), has shown great performance advantage over the traditional regression methods in predictive analysis (Smola and Schölkopf 2004, Ho and Lin 2012). The SVR approach is grounded on statistical learning theory and enabled to generalize well to unseen data. Our study involves using SVR to deal with strategically missing data.

In this work, we examine the problem of performing regression prediction on data with strategically missing values. We focus on the problem where, given the regression models that have been put into practice, the decision-maker evaluates new data while anticipating data providers' strategic behaviors. We propose a novel SVR-based approach to impute missing data. The proposed method derives imputed values based on the margins of the SVR models. We call it the Support Margin Imputation (SMI) approach. The main contributions of this work are summarized below.

- *The originality and significance of the problem.* Existing missing data imputation methods assume that data are missing randomly; thus they are not appropriate for dealing with strategically hidden data. Current work on strategic learning focuses on model building aspect in the classification domain instead of data imputation in a regression prediction context. Given the significant managerial implications of the data concealment problem, it is important to develop an approach to address the issue.
- *The novelty of the approach.* The idea of using a “biased” margin value for hidden data imputation is new. It departs significantly from traditional statistical principal of unbiasedness for handling missing data. Because of strategic interactions, however, the SMI approach produces imputation outcomes that are much closer to the true values (i.e., having smaller bias) than the traditional unbiased methods.
- *Theoretical justification of the methodology.* We show that the SMI method minimizes the imputation errors for the hidden values under certain reasonable conditions. Our analytical results are derived with very little assumption about the knowledge between the decision-maker and the data providers. With the proposed SMI method the regression model remains unchanged and unaffected by the strategic behaviors of data providers (while existing strategic learning methods typically alter the decision models).

This paper is organized as follows. We present an overview of related literature in the next section. Following that, we formulate data providers' strategies and the decision-maker's problem and explain the idea of the proposed approach with an illustrative example. We then derive the optimal solutions to the decision-maker's problem and discuss the analytical properties of the solutions. We further include a simplified approach that is computationally more efficient. We then present the results of an experimental study and discuss managerial implications of our work. The final section concludes our paper with a discussion of limitations and future research.

2. Literature Review

Missing data analysis literature in statistics typically assumes that data are missing randomly. Little (1992) reviews different assumptions of randomly missing data and conventional statistical imputation methods in regression analysis. Conventional statistical methods dealing with missing data in regression analysis include complete case analysis, available case method, Least squares on imputed data, maximum likelihood estimation, Bayesian methods, and multiple imputation (Little 1992). Both complete case analysis and available case analysis are deletion methods, which discard the records with missing values completely or partially. Complete case analysis, also called listwise deletion method, deletes all records having missing data on any variable and performs analysis on the remaining data. Available case method, also called pairwise deletion method, deletes observations having missing data on the pair of two variables being used for calculating covariance. Deletion methods are easy to implement but the new data set after deletion can be very small or even empty, resulting in significant information loss.

Another group of methods dealing with missing data are imputation methods, which include mean substitution, least squares (LS), maximum likelihood (ML), Bayesian methods, and multiple imputation (MI). For any attribute X with some missing values, the mean substitution method replaces the missing values with the sample mean of X calculated from nonmissing values of X . LS fills in the missing values with the values estimated using least squares regression. ML chooses estimates to maximize the likelihood of present values. Bayesian method adds a prior to the likelihood and makes inference on the posterior distribution. MI draws multiple imputed values from the distribution of the data.

The conventional methods for handling missing data are based on the assumption that data are missing randomly. One assumption, missing completely at random (MCAR), refers to the case where missing values of an attribute are independent of any attribute. Another scenario, missing at random (MAR), assumes that missing values do not depend on the attribute itself if the other attributes are controlled. In addition, the assumption of missing not at random considers that the missing values of an attribute depend on the attribute itself (Rubin 1976, Schafer and Graham 2002). Other methods such as the worst-value substitution or best-value substitution are used for sensitivity analysis in experimental treatments (e.g., clinical trials), when there are only a few missing values of binary outcomes (Sterne et al. 2009).

Statistical imputation methods are based on the assumption of data missing randomly, which does not consider the strategic behaviors of the data providers. There are some studies in the strategic learning field that examines data manipulation issues. Dalvi et al. (2004) consider the case where one player, the adversary, strategically changes attribute values to mislead the other player, the decision-maker, to assign the adversary to a favorable group. The adversary's altered values would beat the decision-maker's initial classification rule, but the decision-maker gets better results with an updated classification rule when anticipating such strategic behaviors of the adversary. Following Dalvi et al. (2004), Lowd and Meek (2005) and Barreno et al. (2006) investigate the same strategic learning problem but take a different perspective. They propose a method to test the vulnerability of classification rules to

adversarial attacks. For instance, in the email spam context, Dalvi et al. (2004) study how the system can improve its email filtering algorithm to counter email spammers, while Lowd and Meek (2005) examine an attacking technique for email spammers to learn the filtering algorithm.

The problem of missing data and data manipulation has also been studied in the information systems area. Zhang et al. (2014) investigate the strategically manipulated data problem and propose a method to thwart the strategic behaviors of data providers in the classification contexts. They assume data providers are either positive or negative and the negative data providers attempt to obtain a positive group membership by manipulating and providing their data strategically. The authors found an optimal separating hyperplane for decision-makers to thwart negative data providers from obtaining the undeserved classification, and this new separating hyperplane is tilted and shifted from the original hyperplane. The authors assume that there is a known cost for manipulating data, including changing and hiding data. Boylu et al. (2010) use the similar assumption and investigate strategic behaviors of data providers in a classification problem, focusing on distorted data. With their proposed mechanism, the decision model is altered in response to how data providers can distort data.

Our study differs from existing studies of strategic data manipulation in the literature in several ways. First, we focus on data concealment problem rather than data distortion problem as in many of existing works (we have pointed out earlier the critical difference between the two problems in practice). Second, the existing studies for dealing with strategically manipulated data focus on building adaptive decision models. The main idea of these studies is to alter the decision model to cope with strategic behaviors. Our study concerns evaluating new records with strategically missing data when applying the decision model that is already in use, which precludes altering the model. Third, we examine the data concealment issue in a regression context, whereas existing studies in strategic learning focus on classification domain. Fourth, our approach handles strategic behaviors of data providers by imputing incentive-based values without affecting the decision rules, whereas prior studies attempt to adjust the decision rules to deal with strategic players. Last but not least, we do not make assumptions typically associated with the work on strategic learning and data manipulation. Specifically, our approach does not assume a cost structure for data concealment or manipulation, which is hard to justify in practice.

We should emphasize that our study focuses on the problem where the decision models have already been built and the decision-maker attempts to use the models to evaluate new data that have strategically missing values. We consider the models that have been put into practice. How to build a model from data with strategically missing values is a different problem, which has been studied to some extent, as reviewed in this section. The models in use may be built based on historical data without missing values or by using an existing method discussed earlier when the data have missing values. In our problem, it is expected (from both the decision-maker and data providers) that the same models are to be used for the set of new records within a given timeframe (e.g., the same evaluation models/criteria for accepting students for a given admission cycle).

3. The Support Margin Imputation Approach

Support Vector Machines (SVM), introduced by Vapnik (1998), are supervised learning techniques that analyze data and recognize patterns. SVM is widely used for solving classification problems. The technique can also be used for regression, where it is referred to as Support Vector Regression (SVR). Comparing to the ordinary least square (OLS) regression model, the SVR models take model complexity into account to limit the overfitting problem that the traditional OLS models do not address.

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ be a training set of N records, where $\mathbf{x}_i = (x_{i1}, \dots, x_{iM})$ denotes the values of M input attributes (also known as predictor or feature variables), $x_1, \dots, x_M$ for record i , and y_i is the corresponding value of the outcome attribute (also known as target variable), Y . In a standard SVR setting, all attribute values are of numeric type (categorical attributes are converted into 0–1 binary attributes). The linear SVR model parameters ($\mathbf{w}$, b) are obtained by solving the following optimization problem (Smola and Schölkopf 2004):

$$\min_{\mathbf{w}, b, s^+, s^-} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N (s_i^+ + s_i^-) \quad (1a)$$

$$\text{s.t.} \quad \mathbf{w} \cdot \mathbf{x}_i + b - y_i \leq \varepsilon + s_i^-, \quad i = 1, \dots, N \quad (1b)$$

$$y_i - \mathbf{w} \cdot \mathbf{x}_i - b \leq \varepsilon + s_i^+, \quad i = 1, \dots, N \quad (1c)$$

$$s_i^+, s_i^- \geq 0, \quad i = 1, \dots, N. \quad (1d)$$

The parameter ε determines a lower margin in (1b) and an upper margin in (1c). A deviation from the central regression plane is not considered an error if it is within the margins. When the deviation is larger than ε , it is captured by the slack variables s_i^+ or s_i^- . So, the second term in the objective (1a) measures the fitness of the model. The first term in (1a) represents the complexity of the model; intuitively, the smaller the $\mathbf{w}$ values, the flatter the regression plane (i.e., the less complex the model). Therefore, SVR attempts to minimize both the prediction error and model complexity (to limit the overfitting problem). Parameter $C > 0$ can be viewed as a “weight” to balance the trade-off between the two components of the objective. This idea, called structural risk minimization, is essential to all SVM/SVR models and related techniques (Vapnik 1998, Ryu et al. Jacob 2004, Smola and Schölkopf 2004, Honghai et al. 2005, Aydilek and Arslan 2013).

In the formulation above, we assume a linear relationship between the outcome attribute and the input attributes. It is natural to start with a linear relationship because linear regression is arguably the most commonly used technique for numeric prediction. For nonlinear cases, we can replace the linear kernel in the SVR model with a nonlinear kernel. Detailed

development for the nonlinear case is beyond the scope of this study, and we leave it for future study.

To discuss SVR models with missing data, denote the number of attributes that may have missing values by m ($< M$). For convenience, we label the attributes with complete data first as $X_1, \dots, X_{M-m}$, followed by the m attributes that may have missing values. With the m attributes in mind, we define two types of SVR models below.

Definition 1. The SVR model with all M input attributes is called a *full model*, written as

$$Y = w_1X_1 + \dots + w_MX_M + b. \quad (2)$$

An SVR model with some (say, k) attributes excluded is called a *reduced model*, written as

$$Y_r = w_{r1}X_1 + \dots + w_{r,M-m}X_{M-m} + w_{r,M-m+1}X_{M-m+1} + \dots + w_{r,M-k}X_{M-k} + b_r. \quad (3)$$

A reduced model can be trained based on incomplete data with missing values on the excluded attributes. For example, suppose there are five attributes: X_1 and X_2 have complete data, and X_3 , X_4 , and X_5 may have missing values. The full model is written as

$$Y = w_1X_1 + w_2X_2 + w_3X_3 + w_4X_4 + w_5X_5 + b.$$

There are a total of 7 ($= 2^3 - 1$) reduced models, which are listed below:

$$Y_1 = w_{11}X_1 + w_{12}X_2 + w_{13}X_3 + w_{14}X_4 + b_1$$

$$Y_2 = w_{21}X_1 + w_{22}X_2 + w_{23}X_3 + w_{25}X_5 + b_2$$

$$Y_3 = w_{31}X_1 + w_{32}X_2 + w_{34}X_4 + w_{35}X_5 + b_3$$

$$Y_4 = w_{41}X_1 + w_{42}X_2 + w_{43}X_3 + b_4$$

$$Y_5 = w_{51}X_1 + w_{52}X_2 + w_{54}X_4 + b_5$$

$$Y_6 = w_{61}X_1 + w_{62}X_2 + w_{65}X_5 + b_6$$

$$Y_7 = w_{71}X_1 + w_{72}X_2 + b_7.$$

As mentioned earlier in the Introduction section, this study focuses on the problem scenario where the decision-maker has developed the regression models, and the decision-maker's main concern is how to use the models to appropriately evaluate new data supplied by data providers who have an incentive to disclose favorable information while concealing unfavorable information.

Consider a new record that has unknown y value and strategically missing data. Our idea is to impute the missing values with somewhat "biased" values that are likely to cause a worse evaluation outcome than if the true values were disclosed. As a result, some data providers will opt to report the true values. Even if data providers choose not to disclose the information, the decision-maker's decision models should not be affected by the strategic behaviors of the data providers when the imputed "biased" values are used to replace the missing values. This idea of keeping decision models unchanged even with biased imputed values can be implemented in an SVR model because it allows margins for errors. The idea cannot be implemented in a traditional OLS model.

3.1. Problem Formulation

Let $\mathbf{x}_a^0$ represent the true values of a new record associated with a strategic data provider. The decision-maker's problem is to evaluate the outcome value y_a given $\mathbf{x}_a^0$ (we will use y_a instead of $\hat{y}_a$ throughout this paper because the true Y value for the new record does not exist anyway - It is the subject for evaluation). Without loss of generality, suppose a larger Y value indicates a better outcome.

The data provider of $\mathbf{x}_a^0$ considers hiding data on an attribute, X_h , $h \in \{M - m + 1, \dots, M\}$, whose true value is x_{ah}^0 . Clearly, the data provider is motivated to hide x_{ah}^0 if he believes that hiding will lead to a favorable outcome. Let $y(\mathbf{x}_a^0)$ be the value of y_a when the values of all attributes are provided and $y(\mathbf{x}_a^0 | x_{ah}^0)$ be the value of y_a when all the values are provided except for x_{ah}^0 . Then, the data provider's strategy is to hide x_{ah}^0 if the data provider estimates that $y(\mathbf{x}_a^0) < y(\mathbf{x}_a^0 | x_{ah}^0)$ and to disclose x_{ah}^0 if $y(\mathbf{x}_a^0) \geq y(\mathbf{x}_a^0 | x_{ah}^0)$.

The decision-maker's primary objective is to make accurate prediction for the Y value when facing strategic data providers. This is best achieved when the data provider truthfully reports the X values: It is obvious if the data provider hides unfavorable information, then the prediction made will be inaccurate. So, the decision-maker should discourage such a strategic hiding behavior. On the other hand, the decision-maker should also consider that some data providers indeed do not have the information to disclose. Therefore, the imputed value should be such that it disincentivizes strategic hiding as much as possible while still being fair to those with truly missing data. This idea can be well implemented by using the most unfavorable value within the SVR margin for imputation. The SVR margin value is

reasonable and justifiable to those with truly missing data in that the value fits the current decision model. It also discourages strategic hiding behaviors by using the most unfavorable value afforded by the SVR model for imputation.

Consider a new record that has hidden values in the m attributes ($m < M$, i.e., the new record has at least one attribute with complete value). Because reduced models are useful when the new record has missing values, it is important to keep both the full model and reduced models unaffected by the strategic behaviors of the data provider. To apply the full and reduced models, y_a cannot be outside of any of the SVR margins. That is, it must satisfy constraints that are similar to (1b) and (1c), with respect to the full and all reduced models. Because a larger Y value indicates a better outcome, a y_a value unfavorable to the data provider should be close to the lower margins. Therefore, the upper margin constraints corresponding to (1c) will not be binding and thus can be removed. With the lower margin boundary constraints, the imputed values should cause the evaluation outcome y_a for the data provider as small as possible in order to motivate the data provider to disclose the true values. Let $\mathbf{x}_a = (x_{a,1}, \dots, x_{a,M-m}, x_{a,M-m+1}^*, \dots, x_{a,M}^*)$ be a candidate record for replacing $\mathbf{x}_a^0$. Then, the decision-maker's problem, in terms of finding the best $\mathbf{x}_a$ and y_a , can be formulated as follows:

$$\min y_a \quad (4a)$$

$$\text{s.t.} \quad \mathbf{w} \cdot \mathbf{x}_a + b - y_a \leq \varepsilon \quad (4b)$$

$$\mathbf{w}_r \cdot \mathbf{x}_a + b_r - y_a \leq \varepsilon, r = 1, \dots, (2^m - 1). \quad (4c)$$

The number of constraints in (4c) is $2^m - 1$, which corresponds to the number of all possible reduced models formed by excluding 1, ..., m attributes from the full model. For a case where a smaller Y value indicates a better outcome, objective (4a) should be changed from minimizing y_a to maximizing y_a , and (4b) and (4c) should be replaced by upper margin constraints similar to (1c). Unlike in Problem (1) where the decision variables are $(\mathbf{w}, b)$ while $(\mathbf{x}_i, y_i)$ are known data values, the decision variables in Problem (4) are y_a and $x_{a,M-m+1}, \dots, x_{a,M}$ (corresponding to the m missing elements in $\mathbf{x}_a$), while $(\mathbf{w}, b)$ and $(\mathbf{w}_r, b_r)$ are known. We call (4) the Support Margin Imputation (SMI) model. Note that a larger ε means a larger margin, which induces more truthful disclosures from the data providers. However, ε value cannot be chosen freely. It should be the same as the one used to build the SVR models in order for the imputed values to fit the SVR models in use.

In some applications, the decision -maker can build a feedback mechanism to provide calculated value for imputation to a data provider based on the provider's other available inputs. In this case, the data provider's strategy is very simple. Let x_{ah}^* be the calculated imputation value for X_h . If Y increases in X_h , then the data provider's strategy is to hide x_{ah}^0 if $x_{ah}^0 < x_{ah}^*$, and to disclose x_{ah}^0 if $x_{ah}^0 \geq x_{ah}^*$. If Y decreases in X_h , the strategy will be the

opposite. In this case, the proposed SMI method does not need to assume that the data provider has the knowledge about the decision model and imputation method, as typically assumed by existing approaches. The only assumption is that the data provider knows whether Y is positively or negatively related to X_h . This is very reasonable because if the data provider does not know a larger or smaller X_h value is beneficial, there will be no incentive for the data provider to hide the X_h value to get a better Y value.

3.2. An Illustrative Example

To illustrate the idea of the SMI approach, consider a salary offering example. As shown in Table 1, the data set consists of 12 training records with two input attributes, education (X_1) and experience (X_2), and an outcome attribute, salary (Y). Let $\epsilon = 5$. The full SVR model on the training data is

$$Y_F = 2.1923X_1 + 1.7694X_2 - 1.926.$$

Suppose a candidate, A, reported only the education value of 16. The reduced SVR model on education is

$$Y_R = 3.6667X_1 - 1.3333,$$

which is the solid line plotted in Figure 1. The corresponding lower and upper margins are shown as the dashed lines in Figure 1.

Substituting $x_{A1} = 16$ into the reduced model Y_R , we have the predicted y value of 57.3 (which is shown in Figure 1 as the intersection between the center regression line and the vertical line representing $x_1 = 16$). Because $\epsilon = 5$, the lower margin y value is 52.3 (the intersection between the lower SVR margin line and $x_1 = 16$). Substituting x_{A1} and the lower margin y values into the full model Y_F and solving for x_{A2} , we have the calculated experience value $x_{A2}^* = 10.8$. This value (or a slightly larger value) will be used to replace the missing value, and with this imputed value the full and reduced model will not change because the resulting y value is within the margin. Suppose candidate A's true experience value is $x_{A2}^0 = 12$. The candidate will be incented to report the true value x_{A2}^0 because the resulting y value with x_{A2}^* would be lower than that with x_{A2}^0 based on the full model Y_F . Note that the mean of X_2 is $\bar{x}_2 = 13.2$. So, the candidate will not report the true x_{A2} value if the mean substitution method is used.

Next, consider a new candidate B, who reported only the education value of 12. Following the same SMI procedure above, the lower margin y value is 37.7 (the intersection between the lower SVR margin line and the vertical line $x_1 = 12$ in Figure 1), and the calculated experience value is $x_{B2}^* = 7.5$. Suppose candidate B's true experience value is $x_{B2}^0 = 7$. Then, the candidate may still not disclose the information, but the imputed value calculated with the SMI method is much closer to the true value and thus has much smaller imputation error than the mean substitution (which replaces the missing value with $\bar{x}_2 = 13.2$).

3.3. Decision-Maker's Optimal Solutions

Problem (4) is a linear programming problem, which can be solved using a traditional iterative method such as the Simplex method. In linear programming, it is typically assumed that decision variables are nonnegative. Any free variable, $x \in (-\infty, +\infty)$, is generally handled by introducing two nonnegative variables, x' and x'' , and replacing x with $x' - x''$. In Problem (4), a free variable is handled in the same way.

Problem (4) considers the full model and all reduced models. In general, we can expect that all attributes in the full model are necessary (otherwise, if a model with fewer attributes turned out to be more effective, then it should serve as the full model). A reduced model has one or more relevant attributes omitted from the full model. It has been well documented in the literature that omitted variables can cause some biases for estimated regression coefficients. When the bias is large, a coefficient of a reduced model may have an opposite sign to its counterpart in the full model (Greene 2011). In the five-attribute example following Definition 1, for instance, it is possible for w_{13} , the coefficient of X_3 in the first reduced model, to have an opposite sign to w_3 , the coefficient of X_3 in the full model. Reduced models with such inconsistent signs, caused by omitting relevant attributes, are hard to interpret and often require some remedy (Greene 2011). In practice, it is reasonable to simply ignore these models and keep the reduced models that have coefficient signs consistent with those of the full model. Next, we show that, with these “consistent reduced models,” the linear programming problem can be solved analytically, without using the Simplex method.

Let $R(2^m - 1)$ be the number of consistent reduced models in Problem (4). To simplify notation, denote the parameters of the full model as $(\mathbf{w}_0, b_0)$. Furthermore, relabel the consistent reduced models with indices, $r = 1, \dots, R$, such that the number of attributes decreases or does not increase as r increases. So, the R th model refers to the *fully reduced model* where none of the m attributes is included. Then, Problem (4) with consistent reduced models can be written as

$$\min y_a \quad (5a)$$

$$\text{s.t.} \quad \mathbf{w}_r \cdot \mathbf{x}_a + b_r - y_a \leq \varepsilon, \quad r = 0, 1, \dots, R. \quad (5b)$$

Recall that we label the attributes with complete data first from 1 to $M - m$, followed by the m attributes that may have missing values. For convenience, we call these m attributes *optional* attributes. Problem (5) can be written in a scalar format as below.

$$\min y_a \quad (6a)$$

$$\text{s.t.} \quad \left(\sum_{j=1}^{M-m} w_{rj} x_{aj} + b_r \right) + \sum_{j=M-m+1}^M w_{rj} x_{aj} - y_a \leq \varepsilon, \quad r = 0, 1, \dots, R. \quad (6b)$$

Let $K_r = -\varepsilon + \left(\sum_{j=1}^{M-m} w_{rj}x_{aj} + b_r\right)$, where $x_{aj}, j = 1, \dots, M-m$, are the values of the attributes of the new record $\mathbf{x}_a$ already provided by the data provider. Thus, K_r is a fixed known value for each constraint r . Substituting K_r into (6b), we can write Problem (5) or (6) as

$$\min y_a \quad (7a)$$

$$\text{s.t.} \quad K_r + \sum_{j=M-m+1}^M w_{rj}x_{aj} - y_a \leq 0, \quad r = 0, 1, \dots, R. \quad (7b)$$

Intuitively, if a coefficient w_{rj} of any constraint in (7b) is negative, then the left-hand side can be arbitrarily small by allowing an arbitrarily large positive value for x_{aj} . Therefore, the optimal value of y_a will generally not depend on any constraint having a negative coefficient on an optional attribute. Let W^+ be the set of indices of the reduced models whose coefficients of the optional attributes are nonnegative; that is, $W^+ = \{r \mid w_{rj} \geq 0, j = M-m+1, \dots, M\}$. We have the following analytical solution for Problem (7).

Theorem 1. *With nonnegative $\mathbf{x}$, the optimal objective value of Problem (7) is*

$$y_a^* = \max\{K_r \mid r \in W^+\}, \text{ where } K_r = -\varepsilon + \left(\sum_{j=1}^{M-m} w_{rj}x_{aj} + b_r\right).$$

The proofs of Theorem 1 and all other mathematical results are provided in the online appendix. Theorem 1 states that, when $\mathbf{x}$ are nonnegative, the optimal objective value y_a^* is determined by the constraint that has the largest K_r value among those whose coefficients of all optional attributes are nonnegative. For the special case where every model except the fully reduced model has some negative coefficients on the optional attributes, we have Corollary 1 below.

Corollary 1. *If every model except the fully reduced model, that is, the R th model, in (7b) has some negative coefficients on the optional attributes, then $y_a^* = K_R$.*

In other words, when every model except the fully reduced model has some negative coefficients on the optional attributes, the optimal value y_a^* is the K_R value in the constraint corresponding to the fully reduced model, which does not include any optional attribute.

It follows from Theorem 1 that the optimal objective value y_a^* is bound by only one constraint out of the R constraints in (7b). The other constraints generally have some “slack” values. This suggests that the optimal solution $\mathbf{x}_a^*$ to Problem (7) (and also (5)) generally has multiple solutions. We state this property in Corollary 2 below.

Corollary 2. *For nonnegative $\mathbf{x}$, the optimal solutions $\mathbf{x}_a^*$ of Problem (7) is not unique unless $\mathbf{x}_a^*$ is determined by the full model.*

Corollary 2 indicates that, while the optimal objective value y_a^* is unique, there are typically multiple optimal solutions for $\mathbf{x}_a^*$. As a result, our goal after solving Problem (7) is to further formulate the problem to find the unique optimal solution that best induces the data providers to report truthfully while still satisfying all the model constraints. To simplify discussion, we have assumed in this section that a larger outcome attribute Y value means a better result. Therefore, when $w_j > 0$ (i.e., X_j is positively related to Y), the imputed X_j value should be as small as possible in order to incent the data provider to disclose; when $w_j < 0$, the X_j value should be as large as possible. Together, this will cause the Y value to be as small as possible. However, the Y value calculated from any full or reduced model based on the solution $\mathbf{x}_a^*$ cannot be smaller than y_a^* ; that is, $\mathbf{w}_r \mathbf{x}_a^* + b_r$ should be lower-bound by y_a^* for each r . Based on these observations, we formulate Problem (8) below, where (8a) minimizes the total benefit to the data provider if the true values are replaced by imputed values (in order to maximize the incentive for the data provider to truthfully report), (8b) are the model-margin constraints taken directly from (7b), and (8c) are the y_a^* -related lower-bound constraints.

$$\min \left\{ \begin{array}{l} \sum_{w_j > 0, j \in \{M-m+1, \dots, M\}} x_{aj} - \sum_{w_j < 0, j \in \{M-m+1, \dots, M\}} x_{aj} \end{array} \right\} \quad (8a)$$

$$\text{s. t.} \quad \sum_{j=M-m+1}^M w_{rj} x_{aj} \leq y_a^* - K_r, r = 0, 1, \dots, R \quad (8b)$$

$$\sum_{j=1}^M w_{rj} x_{aj} + b_r \geq y_a^*, r = 0, 1, \dots, R. \quad (8c)$$

The optimal solution to the above problem $(\mathbf{x}_a^*, y_a^*)$ is located on the lower margins of some reduced models. So, the optimal solution fits all the existing models.

For convenience and without loss of clarity, we now denote $\mathbf{x}_a^0$ to be the true values of the m optional attributes. That is, $\mathbf{x}_a^0$ and $\mathbf{x}_a^*$ pertain to the same attributes. The optimal solution $\mathbf{x}_a^*$ of Problem (8) may be insufficient for a data provider to disclose $\mathbf{x}_a^0$ if $\mathbf{x}_a^0$ is worse than $\mathbf{x}_a^*$. We define $\mathbf{x}_a^0$ is *inferior* to $\mathbf{x}_a^*$ if $y(\mathbf{x}_a^0) < y(\mathbf{x}_a^*)$, i.e., if the y value calculated based on $\mathbf{x}_a^0$ is smaller than that based on $\mathbf{x}_a^*$. In this case, the data provider benefits by hiding $\mathbf{x}_a^0$ and letting decision maker to use $\mathbf{x}_a^*$ for imputation. This unfair gain associated with imputation error should be kept as small as possible. Theorem 2 below states that the optimal solution $\mathbf{x}_a^*$ minimizes such imputation error (and unfair gain of strategic data providers by taking advantage of the imputed values) while still preserving the decision-maker's decision models.

Theorem 2. When $\mathbf{x}_a^0$ is inferior to $\mathbf{x}_a^*$, the optimal solution $\mathbf{x}_a^*$ minimizes the sum of absolute imputation errors while preserving the decision-maker's decision models. That is, for any data point $\mathbf{x}'_a$ that satisfies the decision models, we have

$$\sum_{j=M-m+1}^M |x_{aj}^* - x_{aj}^0| \leq \sum_{j=M-m+1}^M |x'_{aj} - x_{aj}^0|.$$

We have discussed the optimal solution in terms of one new record. It is straightforward to apply the SMI approach for multiple new records by considering one record at a time repeatedly. All the properties of the optimal solution discussed in this section remain the same for the case with multiple new records.

3.4. A Simplified Support Margin Imputation Model

The linear programming problem (4) involves a large number of constraints (on the order of 2^m), which can be computationally expensive to solve. In practice, it is likely that only a few reduced models are really important. Therefore, in addition to the complete model (4), we propose a light version of the SMI model, called SMI-Lite (SMILE), which attempts to keep only m important reduced models. The idea is that if the number of optional attributes is limited to k ($k = 1, \dots, m-1$), there is one and only one reduced model with k optional attributes. Clearly, SMILE is a heuristic approach and does not guarantee optimality for the complete model (4).

We use a *backward elimination* procedure to select the m reduced models. The importance of an optional attribute X_j is determined by the average sum of squared error (SSE) in Y computed from the reduced model that removes X_j . First, out of m optional attributes, the attribute whose deletion causes the smallest SSE is removed, resulting in a reduced model of $m-1$ optional attributes. Next, from the remaining $m-1$ attributes, a reduced model of $m-2$ optional attributes is found similarly. Continuing this process will result in a total of m reduced models. Without loss of generality, we relabel the optional attributes such that the r th reduced model has $M-r$ input attributes. As a result, Problem (4) is simplified to

$$\min \quad y_a \tag{9a}$$

$$\text{s. t.} \quad \mathbf{w} \cdot \mathbf{x}_a + b - y_a \leq \varepsilon \tag{9b}$$

$$\mathbf{w}_r \cdot \mathbf{x}_a + b_r - y_a \leq \varepsilon, r = 1, \dots, m. \tag{9c}$$

Following Corollary 2, the optimal solution $\mathbf{x}_a^*$ of Problem (9) is generally not unique. We can again set up an optimization problem similar to Problem (8) to find an optimal solution of (9). However, given the special structure of the constraints in (9) determined by the backward elimination procedure, there is a very simple algorithm for solving the entire problem efficiently. We describe this SMILE algorithm below.

Algorithm 1 (SMI-Lite (SMILE) Algorithm)

1. Given a full SVR model with M attributes

$$Y_0 = w_{01}X_1 + w_{02}X_2 + \dots + w_{0M}X_M + b_0,$$

consider a new record $\mathbf{x}_a$ with missing values in $m(< M)$ attributes, which appears as the last m attributes(after) relabeling.

2. Use the backward elimination procedure to determine m reduced models:

$$Y_1 = w_{11}X_1 + \dots + w_{1,M-m}X_{M-m} + w_{1,M-m+1}X_{M-m+1} + \dots + w_{1,M-1}X_{M-1} + b_1$$

$$Y_{m-1} = w_{m-1,1}X_1 + \dots + w_{m-1,M-m}X_{M-m} + w_{m-1,M-m+1}X_{M-m+1} + \dots + w_{m-1,M-1}X_{M-1} + b_{m-1}$$

$$Y_m = w_{m,1}X_1 + \dots + w_{m,M-m}X_{M-m} + b_m$$

3. Find (MI) by substituting $\mathbf{x}_a$'s known values $(x_{a,1}, \dots, x_{a,M-m})$ into equation Y_m in step 2.

4. For $j = 1$ to m , find $x_{a,M-m+j}^*$ by substituting $y_a^* - \varepsilon$ and

$$(x_{a,1}, \dots, x_{a,M-m}, x_{a,M-m+j}^*, \dots, x_{a,M-m+j-1}^*) \text{ into equation } Y_{m-j}.$$

It is straightforward to apply the SMILE approach for multiple new records. Steps 1 and 2 of the SMILE algorithm are run only once for multiple new records. Steps 3 and 4 are run multiple times, one for each new record. The computational complexity of Steps 1 and 2, which is related to SVR's, is somewhere between linear and quadratic in the training set size (Platt 1998). The time complexity of Steps 3 and 4 for n new records is clearly linear in n . In short, SMILE is computationally very efficient and should be useful when the number of new records to be evaluated is large.

Similar to the SMI solution, the SMILE solution also has the property of minimizing the sum of imputation errors, as stated below:

Theorem 3. When x_a^0 is inferior to x_a^* , the optimal SMILE solution x_a^* minimizes the sum of absolute imputation errors while preserving the decision maker's decision models.

Note that while the statement in Theorem 3 appears the same as that in Theorem 2, the number of decision models in SMILE is fewer than that in the complete SMI formulation.

4. Experimental Evaluation

To evaluate the effectiveness of the proposed approach, we conducted an experimental study on three real-life data sets, related to housing price, university rankings, and small business earnings. All three data sets have realistic data concealment contexts. In housing markets,

some sellers may purposely hide housing information to increase the selling price of their houses. The hidden information can be costly for buyers and mortgage companies. In terms of university rankings, Morse and Flanigan (2013) and Conlin et al. (2013) report that some universities may intentionally choose not to disclose certain student information in order to increase their rankings. Lipsett (2008) points out that some universities hide poor research information related to university rankings. Furthermore, small business earnings data are likely to be used in a credit evaluation context, where a small business may selectively disclose some nonmandatory information.

The first data set, Housing, has 506 records with 14 attributes, such as housing price (outcome attribute), age, number of rooms, tax, crime rate (of the town), teacher/student ratio (of the town), and so on (Harrison and Rubinfeld 1978). The second data set, University Rankings, consists of 381 university records with 12 attributes, including ranking score (outcome attribute), which is a real number between 0 (worst) and 100 (best), university size, age, type, faculty/student ratio, citations per faculty, and several other measures on research, teaching, and employability (Quacquarelli Symonds 2012). The third data set, Earnings, extracted from the U.S. Census Bureau's data website (Brand et al. 2002), includes earnings related data for 1,080 individuals who run small business. There were 13 numeric attributes originally, including various categories of income, interests, taxes, and deductions. We found that for each record, the value of one attribute, "total person earnings," exactly equals the sum of two other attribute values, causing the perfect collinearity problem. So, this attribute was removed from the data set, resulting in a total of 12 attributes, including net earnings, salary, personal income, other income, interest income, adjusted gross income, state income tax liability, taxable income amount, federal income tax liability, etc. The attribute net earnings is the outcome attribute.

4.1. Experiment Setup and Performance Measures

The SMI method based on Models (7) and (8) is compared with several existing methods, including mean substitution (Mean), least squares imputation (LS), and maximum likelihood (ML). Mean substitution is chosen because it is one of the most widely used methods in practice. LS imputation is selected because, similar to our method, it also uses a regression approach. ML is considered one of the methods with the best performance (Little 1992, Schafer and Graham 2002). For comparison, we also included a worst-value (WV) imputation method, which panelizes strategic hiding behaviors by replacing any missing attribute value with the worst value of the attribute in the data set.

There are no missing values originally in the three data sets, which is desirable as we need the original values to evaluate imputation and prediction errors. In the experiment, we considered some optional attributes for each data set, based on the context of the attributes. In general, attributes that are more likely to have missing values were selected to be optional. In the Housing test set, we selected two attributes, crime rate and teacher/student ratio, as optional attributes. In the University Rankings test set, we selected faculty/student ratio and citations per faculty to be optional. In the Earnings test set, eight attributes were selected to be optional, including salary, personal income, other income, interest income, adjusted gross income, state income tax liability, taxable income amount, and federal income

tax liability. Our experiments tested different settings, involving either a subset or the entire set of the optional attributes. We used 10-fold cross-validation in evaluating the methods. A data set was randomly divided into 10 equally sized blocks for 10 test runs. Each time one block was used as the test set, and the remaining 9 blocks together were the training set. All the values of the optional attributes in the test sets were subject to imputation. The five imputation methods were used to compute imputing values in the test sets, based on the information obtained and the models developed from the training data.

We used the Weka implementation of SVR (Witten et al. 2011) to build SVR models in our experiments, with the default parameters except ε . To examine the impact of this parameter, we set five values for ε : 0.05, 0.10, 0.15, 0.20, and 0.25. This implies that the margins are 5%, 10%, 15%, 20%, and 25% of the ranges of the outcome attributes, respectively, because all of the original values are normalized to the range [0, 1] in Weka. Note that the selection of five ε values is to examine the effect of different ε values. In practice, only one ε value (the one that is used in building the SVR models) should be used for imputation and prediction.

The effectiveness of the proposed approach is evaluated based on its performance on two measures: (1) imputation error and (2) prediction error with imputed data. The imputation error is measured by the imputation mean absolute deviation (*IMAD*), defined for an attribute X_j as

$$IMAD_j = (1/n) \sum_{i=1}^n |\hat{x}_{ij} - x_{ij}|, \quad (10)$$

where n is the number of records in the test set, x_{ij} is the true value of the j th attribute of the i th record in the test set, and $\hat{x}_{ij}$ is the corresponding imputed value when the true value is unfavorable relative to $\hat{x}_{ij}$. This imputed value is calculated using each of the imputation methods in the experiment, respectively. For multiple attributes, the overall *IMAD* value is calculated as the average of the individual *IMAD_j* values.

The prediction error is measured by the regression mean absolute deviation (*RMAD*):

$$RMAD = (1/n) \sum_{i=1}^n |\hat{y}_i - y_i|, \quad (11)$$

where $\hat{y}_i$ is the predicted value of y_i when the missing values of the i th record in the test set are imputed. For SMI, $\hat{y}_i$ is calculated using SVR model built on the training set; for the other methods, corresponding estimation methods are used respectively. All attribute values in (10) and (11) are normalized.

4.2. Experimental Results and Discussion

We performed two sets of experiments for comparing imputation and prediction errors. The first set of our experiments compares the performances of the Mean, LS, ML, WV, and SMI methods in terms of imputation and prediction errors. We consider four different scenarios

where 0%, 25%, 50%, and 75% of the values of the optional attributes are respectively considered actually missing. These actually missing data will be imputed with estimated values without strategic reactions (i.e., no matter the true value is better than the estimated value or not). The second set of experiments compares the SMI and the SMILE methods in terms of imputation and prediction errors, as well as computation time.

The experimental results of imputation and prediction errors are reported in Table 2, which is calculated based on the part of the data with unfavorable true values being replaced by imputed values. As shown in Table 2, the *IMAD* and *RMAD* values of the SMI method are much smaller than those of the Mean, LS, ML, or WV method when there are both strategically and actually missing data (i.e., 25%, 50%, and 75% actually missing). The WV method has smaller *IMAD* and *RMAD* values than the SMI method when there is no (0%) actually missing data. The imputation error is zero in this case because all data providers would disclose the true values. However, it is very unlikely in real life that all missing data are due to strategic behavior without any actually missing value. The results with the WV method become worse when the proportion of actually missing values increases. The SMI's advantage is obviously due to its ability to minimize imputation error on strategically missing data, as specified in Theorem 2. Therefore, when imputed values are used, the value calculated with SMI is closer to the true value than those with Mean, LS, ML, and WV. We also observe that for most cases the *IMAD* and *RMAD* values of the SMI method have the smallest values when $\varepsilon = 0.15$ (compared with those when $\varepsilon = 0.05, 0.10, 0.20$, and 0.25). This is understandable because when ε is smaller, there will be more values falling outside the margins and thus strategically missing, causing larger errors. When ε is too big, the errors on the records with actually missing values will become dominant, also causing larger errors. Nevertheless, we should point out that ε is a parameter that was determined when building the SVR model not a parameter that can be chosen freely when applying the SMI method for prediction.

To evaluate the performance attributed to both truthfully disclosure and imputation effectiveness, we report in Table 3 the results based on all data, where an attribute value is either the true value or its imputed value, whichever is more favorable to data providers. It is very clear from Table 3 that SMI outperforms the other methods in the *IMAD* measure substantially on the three datasets when there are both strategically and actually missing data. The relative performances of the different methods in Table 3 are very similar to those in Table 2. The SMI's superior performance is due in part to its biased imputed values, which results in more true values being disclosed. The SMI's advantage is also due to its ability to minimize imputation error on strategically missing data. We have explained this in detail when discussing the results of Table 2. The SMI method also outperforms the Mean, LS, ML, and WV methods in the *RMAD* measure. It appears that the SMI's better performance is also owing to the SMI's abilities to incent disclosure of true values and to minimize imputation errors, which subsequently result in better prediction results. With the other methods, the errors in the *RMAD* measure are larger because fewer true values are disclosed or imputation errors are larger.

The second set of our experiments compares the performances of the SMI and SMILE methods. As described earlier in the previous section, SMI preserves all reduced models that

have consistent coefficient signs with those of the full model, while SMILE keeps only m important reduced models. Therefore, we expect SMI to perform better than SMILE regarding imputation and prediction accuracies; at the same time, SMI will be computationally more expensive than SMILE. For a data set with m optional attributes, we selected $k = 1, 2, \dots, m$ attributes and then tested the performance when the values of selected attributes are missing. We ran SMI and SMILE for all possible $\binom{m}{k}$ combinations of missing attributes and reported the average results for k . The results of the first experiment with five ε values show that in most cases SMI performed best when $\varepsilon = 0.15$. So we have run the experiment with $\varepsilon = 0.15$ on all data.

The average *IMAD*, *RMAD*, and time results with SMI and SMILE are shown in Table 4. In general, SMI outperforms SMILE in both *IMAD* and *RMAD* measures. This is expected because SMI includes much more reduced models than SMILE. So, for certain combinations of optional attributes, it is very likely that SMI has the exact reduced models to estimate the missing values while SMILE does not and has to use reduced models with fewer input attributes for estimation. On the other hand, SMILE is faster than SMI, particularly when the number of missing attributes is large (e.g., the Earnings data). As the number of missing attributes increases, the *IMAD* and *RMAD* values with both SMI and SMILE increase, but SMILE has greater advantage over SMI in run time. The results of the Earnings data show that as the number of missing attributes becomes larger, SMILE's advantage over SMI in run time becomes more evident. Therefore, SMILE may be a good alternative for SMI when the number of attributes with missing values is large and the test set (number of new records to be evaluated) is large.

5. Discussion

Data providers often disclose their information selectively in hope of gaining favorable but perhaps undeserved decision outcomes. Decision-makers can thwart such strategic behaviors of data providers with the proposed approach. On the basis of the proposed SMI method, the decision maker uses imputation values that are likely to be unfavorable to data providers but still fair in that the values fit the decision models in use. Thus, most data providers are not likely to be rewarded by hiding information strategically. Even if hiding is beneficial for some data providers, their gains are minimized using the proposed SMI approach. We have demonstrated in experiments using very relevant real data that the proposed SMI approach can be used by managers and policy makers to effectively incent data providers to disclose information and handle information concealment problems.

Both decision-makers and honest data providers can benefit from the proposed approach. When information is selectively reported, it can be costly for managers or decision-makers to find the missing information. Sometimes it is impossible for decision-makers to get the right information. On the other hand, data providers can save effort, time, and cost by not providing certain information. Without an appropriate incentive-compatible mechanism, data providers who honestly disclose all information may actually be in a disadvantageous position. Our proposed approach helps to limit such an unfair situation in practice.

Our approach can improve the quality of the data collected by organizations, resulting in well-supported decision models. Decision-makers build a decision model based on the data provided by individuals or entities. The model can be biased if it is built on strategically manipulated data. With the proposed SMI method, the strategically missing values in the new records have no impact on the decision-maker's existing decision model. Managers are able to preserve the existing decision model from being manipulated by strategic data providers. Consequently, decision outcomes based on the SMI approach are also not affected by the strategic behaviors of the data providers.

With the proposed approach, managers and decision-makers can get useful message from missing data. If a data provider does not disclose his information, then either the imputed values are beneficial to the provider or the data are indeed missing. If this occurs to a significant number of new records, then it is likely that the decision model does not fit new data well or there is not much data available for the attribute. In this case, the decision-maker should consider modifying the decision model.

In addition to the main SMI method, we have also proposed a simplified version called SMILE. Both SMI and SMILE provide imputation values that motivate data providers to truthfully report their information while keeping the decision models unaffected by the reported or imputed values. The main difference between the two methods is that SMI attempts to maintain a large number of reduced models unaffected whereas SMILE considers only a few critically reduced models. Consequently, SMI will generally perform better than SMILE in terms of imputation and prediction accuracies, but SMILE will run considerably faster than SMI. When choosing between the two methods, managers and decision-makers should consider the trade-off between accuracy and runtime and select an appropriate method based on their priority.

6. Conclusions and Extensions

We have proposed a novel SMI approach to impute strategically missing data for regression prediction. Our idea of using a “biased” value for hidden data imputation is new and innovative. Our method offers incentives for data providers to disclose true information and subsequently results in outcomes that are less biased than those with the traditional “unbiased” imputation methods. With our approach, the decision models are not affected by the strategically provided information. We have shown that the SMI approach minimizes the imputation errors of strategically missing data under some reasonable conditions. The experimental study has demonstrated that the proposed SMI methods significantly outperform traditional imputation methods such as mean substitution, least squares, and maximum likelihood imputation.

Although it is desirable to keep the decision model relatively stable, the model should also be able to digest new information and to adapt to changes. The proposed approach can actually capture new patterns. Suppose with the current model, there is a certain proportion (p) of the training records falling outside the margins. After applying the model to a series of new records, if the proportion of the new records with missing data are significantly larger or smaller than p , then this signals that the new data may follow different distributions from the

old data, and the decision-maker may need to adjust the model accordingly by including new data. It is interesting to study how to adapt the model to such changes as an extension to this work. In this study, we have assumed linear kernels for the SVR model. In many real-world problems, nonlinear kernels are needed to make more accurate predictions. Our method can be extended to solve nonlinear regression problems by using nonlinear kernels in the SVR model.

This study focuses on decision-making problems with strategically missing data, but not intentionally distorted data. In many real-world decision problems, there is a clear distinction between not reporting information and providing incorrect information. The right to remain silent is well-known in legal systems. “No comment” is one of the most popular responses that public officials give to journalistic inquiries. We have also provided many applications where hiding is acceptable but lying is not, in terms of ethical and/or legal implications. Therefore, it is clear that data concealment and data distortion are different in nature and should be handled differently. That being said, we believe some ideas, such as providing incentive for data providers to truthfully disclose their information, can be extended to deal with data distortion problems. It is interesting and challenging to develop an imputation approach for the strategic data distortion problem in a regression context for future research.

In some circumstances, information is missing not because it is concealed but because the data provider indeed does not have the information to report. For example, a college applicant will not be able to report an SAT Subject test score if the applicant has not taken this test. The imputed test score may be unfavorable to the applicant, but this will motivate the applicant to take the SAT Subject test and report the score, even though it is optional. As a result, there will be fewer missing values in the data collected. In practice, if a large number of values for an attribute may be truly missing, the decision-maker may also consider removing the attribute from the decision model. The problem where intentionally and unintentionally missing data are both present deserves more extensive research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to the area editor, associate editor, and the three anonymous reviewers for their insightful comments and suggestions that have improved the paper considerably. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

Funding: The research by X.-B. Li was supported in part by the National Library of Medicine of the National Institutes of Health [Grant R01LM010942].

References

- Aydilek IB, Arslan A (2013) A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Inform. Sci* 233: 25–35.
- Ba S (2001) Establishing online trust through a community responsibility system. *Decision Support Systems* 31:323–336.

- Barreno M, Nelson B, Sears R, Joseph AD, Tygar JD (2006) Can Machine Learning Be Secure? (ACM, New York).
- Bartov E, Mohanram P, Seethamraju C (2002) Valuation of Internet stocks-An IPO perspective. *J. Accounting Res* 40(2):321–346.
- Bernat A, Grimson E (2011) Doctoral program rankings for U.S. computing programs: The National Research Council strikes out. *Comm. ACM* 54(12):41–43.
- Boylu F, Aytug H, Koehler GJ (2010) Induction over strategic agents. *Inform. Systems Res* 21(1):170–189.
- Brand R, Domingo-Ferrer J, Mateo-Sanz J (2002) Reference data sets to test and compare SDC methods for protection of numerical microdata. Accessed January 15, 2015, <http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf>.
- Braun S, Dwenger N, Kubler D (2010) Telling the truth may not pay off: An empirical study of centralized university admissions in Germany. *B. E. J. Econom. Anal. Policy* 10(1):1–35.
- Cecchini M, Aytug H, Koehler GJ, Pathak P (2010) Making words work: Using financial text as a predictor of financial events. *Decision Support Systems* 50(1):164–175.
- Clarke M (2002) Quantifying quality: What can the U.S. news and world report rankings tell us about the quality of higher education? Ed. *Policy Anal. Arch* 10(16):1–20.
- Conlin M, Dickert-Conlin S, Chapman G (2013) Voluntary disclosure and the strategic behavior of colleges. *J. Econom. Behav. Organ* 96:48–64.
- Dalvi N, Domingos P, Mausam, Sanghai S, Verma D (2004) Adversarial classification. *Proc. 10th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining (ACM, Seattle)*, 99–108.
- Degeratu AM, Rangaswamy A, Wu J (2000) Consumer choice behavior in online and traditional supermarkets: The effects of brand name, price, and other search attributes. *Internat. J. Res. Marketing* 17(1):55–78.
- Fabozz FJ, ed. (2008) *Handbook of Finance: Valuation, Financial Modeling, and Quantitative Tools*, vol. 3 (John Wiley & Sons, New York).
- Goldberg S (2010) Young job-seekers hiding their Facebook pages. Accessed January 15, 2015, <http://www.cnn.com/2010/TECH/03/29/facebook.job-seekers/index.html>.
- Greene WH (2011) *Econometric Analysis* (Granite Hill Publishers, Upper Saddle River, NJ).
- Harrison D, Rubinfeld DL (1978) Hedonic housing prices and the demand for clean air. *J. Environ. Econom. Management* 5(1):81–102.
- Higgins K (2012) Initial public offerings, forecasts, and selective disclosure. *Insight J.* 26(9):11–17.
- Ho C-H, Lin C-J (2012) Large-scale linear support vector regression. *J. Machine Learn. Res* 13(1):3323–3348.
- Honghai F, Guoshun C, Cheng Y, Bingru Y, Yumei C (2005) A SVM regression based approach to filling in missing values. Khosla R, Howlett RJ, Jain LC, eds. *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, vol. 3683 (Springer, Berlin), 581–587.
- Insure.com (2010) Can you hide smoking from life insurance companies? Accessed January 15, 2015, <http://www.insure.com/articles/lifeinsurance/smoking.html>.
- Kim M, Ritter JR (1999) Valuing IPOs. *J. Finance Econom* 53(3): 409–437.
- Lipsett A (2008) Blunder lets universities ‘hide’ poor researchers in rankings. Accessed January 15, 2015, <http://www.theguardian.com/education/2008/oct/29/researchassessmentexercise-hefcea-hesa-abandonedleaguetables>.
- Little RJA (1992) Regression with missing X’s: A review. *J. Amer. Statist. Assoc* 87(420):1227–1237.
- Lowd D, Meek C (2005) *Adversarial Learning* (ACM Press, New York).
- Morse R, Flanigan S (2013) How U.S. News calculated the 2013 best colleges rankings. Accessed January 15, 2015, <http://www.usnews.com/education/best-colleges/articles/2013/09/09/how-us-news-calculated-the-2014-best-colleges-rankings?page=3>.
- Platt J (1998) *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* (MIT Press, Cambridge, MA).
- Pukthuanthong-Le K, Varaiya N (2007) IPO pricing, block sales, and long-term performance. *Financial Rev.* 42(3):319–348.

- Quacquarelli Symonds (2012) QS world university rankings 2012 report. Accessed January 15, 2015, <http://www.iu.qs.com/product/2012-qs-world-university-rankings-result-tables-excel-format/>.
- Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592.
- Ryu YU, Chandrasekaran R, Jacob V (2004) Prognosis using an isotonic prediction technique. *Management Sci.* 50(6):777–785.
- Schafer JL, Graham JW (2002) Missing data: Our view of the state of the art. *Psych. Methods* 7(2): 147–177.
- Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. *Statist. Comput* 14(3):199–222.
- Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 338:b2393. [PubMed: 19564179]
- Vapnik VN (1998) *Statistical Learning Theory* (John Wiley & Sons, Hoboken, NJ).
- Witten IH, Frank E, Hall MA (2011) *Data Mining: Practical Machine Learning Tools and Techniques* (Elsevier Science, Amsterdam).
- Yang Y, Liu H, Cai Y (2013) Discovery of online shopping patterns across websites. *INFORMS J. Comput* 25(1):161–176.
- Zhang J, Aytug H, Koehler GJ (2014) Discriminant analysis with strategically manipulated data. *Inform. Systems Res* 25(3):654–662.

Figure 1.
Support Vector Regression for Salary Offering Example

Table 1.

A Salary Offering Example

No.	Education (X_1), years	Experience (X_2), years	Salary (Y), \$
1	10	5	33,000
2	12	7	36,000
3	12	13	50,000
4	13	14	51,000
5	14	10	42,000
6	14	12	45,000
7	15	16	59,000
8	16	11	49,000
9	16	18	60,000
10	17	20	72,000
11	18	15	69,000
12	20	17	77,000

Table 2.

Error Results Based on Imputed Values

RMAD																			
IMAD																			
Percent of values actually missing ^a																			
	Data	Mean	LS	ML	WV	0.05	0.10	0.15	0.20	0.25	Mean	LS	ML	WV	0.05	0.10	0.15	0.20	0.25
0%	Housing	0.62	0.17	0.29	0	0.04	0.04	0.03	0.007	0.006	0.23	0.17	0.20	0.005	0.01	0.01	0.009	0.006	0.006
	Rankings	0.79	0.46	0.31	0	0.28	0.27	0.07	0.006	0.005	0.43	0.14	0.11	0.002	0.03	0.03	0.009	0.006	0.006
	Earnings	0.61	0.10	0.09	0	0.04	0.04	0.01	0.013	0.009	0.42	0.11	0.09	0.010	0.09	0.08	0.050	0.070	0.030
25%	Housing	0.62	0.17	0.29	0.21	0.11	0.11	0.03	0.10	0.10	0.23	0.17	0.20	0.20	0.12	0.12	0.02	0.11	0.11
	Rankings	0.79	0.46	0.31	0.47	0.29	0.27	0.07	0.14	0.15	0.43	0.14	0.11	0.18	0.10	0.10	0.03	0.10	0.10
	Earnings	0.61	0.10	0.09	0.08	0.06	0.06	0.03	0.05	0.05	0.42	0.11	0.10	0.11	0.10	0.09	0.05	0.06	0.07
50%	Housing	0.62	0.17	0.29	0.61	0.14	0.13	0.03	0.14	0.14	0.23	0.17	0.20	0.21	0.13	0.13	0.02	0.15	0.15
	Rankings	0.79	0.46	0.31	0.79	0.29	0.29	0.07	0.15	0.16	0.43	0.14	0.11	0.42	0.10	0.10	0.03	0.10	0.10
	Earnings	0.61	0.10	0.09	0.59	0.07	0.06	0.04	0.07	0.07	0.42	0.11	0.10	0.40	0.08	0.08	0.06	0.09	0.09
75%	Housing	0.62	0.17	0.29	0.66	0.14	0.14	0.03	0.15	0.15	0.23	0.17	0.20	0.23	0.13	0.13	0.02	0.15	0.15
	Rankings	0.79	0.46	0.31	0.80	0.30	0.29	0.08	0.16	0.17	0.43	0.14	0.11	0.46	0.10	0.10	0.03	0.10	0.10
	Earnings	0.61	0.10	0.09	0.60	0.07	0.07	0.03	0.07	0.07	0.42	0.11	0.10	0.41	0.10	0.09	0.04	0.09	0.09

^aThe percentage of the values of the optional attributes that are considered actually missing.

Table 3.

Error Results Based on All Data

		RMAD									
		IMAD					SMI $\epsilon =$				
		Mean	LS	ML	WV	0.05	0.10	0.15	0.20	0.25	0.25
Percent of values actually missing ^a	Data										
0%	Housing	0.31	0.09	0.15	0	0.01	0.01	0.01	0.001	0.001	0.001
	Rankings	0.40	0.22	0.16	0	0.09	0.08	0.02	0.001	0.001	0.001
	Earnings	0.33	0.08	0.08	0	0.04	0.03	0.02	0.033	0.032	0.017
25%	Housing	0.31	0.09	0.15	0.09	0.01	0.01	0.01	0.01	0.01	0.01
	Rankings	0.40	0.22	0.16	0.23	0.09	0.08	0.02	0.02	0.02	0.01
	Earnings	0.33	0.08	0.08	0.13	0.04	0.03	0.02	0.21	0.21	0.03
50%	Housing	0.31	0.09	0.15	0.32	0.01	0.01	0.01	0.01	0.01	0.02
	Rankings	0.40	0.22	0.16	0.41	0.08	0.08	0.02	0.03	0.03	0.02
	Earnings	0.33	0.08	0.08	0.33	0.04	0.03	0.01	0.03	0.04	0.05
75%	Housing	0.31	0.09	0.15	0.42	0.01	0.01	0.01	0.02	0.02	0.02
	Rankings	0.40	0.22	0.16	0.51	0.08	0.08	0.02	0.03	0.03	0.02
	Earnings	0.33	0.08	0.08	0.39	0.04	0.03	0.01	0.03	0.04	0.05

^aThe percentage of the values of the optional attributes that are considered actually missing.

Table 4.

Comparisons Between SMI and SMILE Methods ($\varepsilon = 0.15$)

Data	Number of missing attributes, k	IMAD			RMAD			Time (sec)		
		SMI	SMILE	SMI	SMI	SMILE	SMI	SMI	SMILE	SMILE
Housing	1	0.007	0.012	0.002	0.002	0.005	5	3		
	2	0.011	0.017	0.002	0.002	0.006	6	4		
Rankings	1	0.012	0.101	0.001	0.011	0.011	14	4		
	2	0.020	0.124	0.002	0.014	0.014	20	6		
Earnings	1	0.012	0.020	0.005	0.013	0.013	19	7		
	2	0.012	0.021	0.007	0.013	0.013	29	9		
	3	0.013	0.023	0.007	0.018	0.018	35	10		
	4	0.013	0.025	0.007	0.019	0.019	69	13		
	5	0.013	0.025	0.008	0.023	0.023	99	15		
	6	0.015	0.025	0.009	0.035	0.035	173	18		
	7	0.018	0.031	0.010	0.037	0.037	399	19		
	8	0.021	0.034	0.011	0.040	0.040	787	30		