informs
http://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Vol. 32, No. 2, June 2021, pp. 462–480
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

Correcting Misclassification Bias in Regression Models with Variables Generated via Data Mining

Mengke Qiao,[a] Ke-Wei Huang[b]

[a] International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China;
[b] Department of Information Systems and Analytics, National University of Singapore, Singapore 117417

Contact: mengkeqiao@gmail.com, https://orcid.org/0000-0002-0554-7916 (MQ); huangkw@comp.nus.edu.sg, https://orcid.org/0000-0002-9932-6195 (K-WH)

Received: December 27, 2018
Revised: October 10, 2019; May 30, 2020
Accepted: August 10, 2020
Published Online in Articles in Advance: March 26, 2021

https://doi.org/10.1287/isre.2020.0977

Abstract. As a result of advances in data mining, more and more empirical studies in the social sciences apply classification algorithms to construct independent or dependent variables for further analysis via standard regression methods. In the classification phase of these studies, researchers need to subjectively choose a classification performance metric for optimization in the standard procedure. No matter which performance metric is chosen, the constructed variable still includes classification error because those variables cannot be classified perfectly. The misclassification of constructed variables will lead to inconsistent regression coefficient estimates in the following phase, which has been documented as a problem of measurement error in the econometrics literature. The pioneering discussions on the issue of estimation inconsistency because of misclassification in these studies have been provided. Our study attempts to investigate systematically the theoretical foundation of this problem when a newly constructed variable is used as the independent or dependent variable in linear and nonlinear regressions. Our theoretical analysis shows that consistent regression estimators can be recovered in all models studied in this paper. The main implication of our theoretical result is that researchers do not need to tune the classification algorithm to minimize the inconsistency of estimated regression coefficients because the inconsistency can be corrected by theoretical formulas, even when the classification accuracy is poor. Instead, we propose that a classification algorithm should be tuned to minimize the standard error of the focal regression coefficient derived based on the corrected formula. As a result, researchers can derive a consistent and most precise estimator in all models studied in this paper.

History: Ram Gopal, Senior Editor; Gautam Pant, Associate Editor.
Supplemental Material: The online appendices are available at https://doi.org/10.1287/isre.2020.0977.

Keywords: data mining • econometrics • measurement error • misclassification • statistical inference • performance metric

1. Introduction

There is a surge of interest in social science studies in applying data mining methods to construct variables for traditional regression analysis in various disciplines, including information systems, finance, and accounting. The increasing popularity of data mining methods results from the novel use of extracting information from unstructured data sources, such as text and images. For example, text classification has been applied to classify whether the review is subjective or objective (Ghose and Ipeirotis 2011, Ghose et al. 2012). The derived review subjectivity was used as an independent variable in the regression to examine its impact on review helpfulness or product sales. As another example, Chan and Wang (2014) applied computer vision technology to classify the worker's gender from a photo posted on a profile page. Given the classification output, the authors examined the impact of gender on hiring outcomes.

There are typically two stages in these *hybrid studies*, which are defined as using a first-stage classification algorithm to construct an independent or dependent variable for second-stage regression analysis. More specifically, in the first stage, researchers apply classification algorithms to a small set of observations with labels to build a classifier. This classifier is applied to the unlabeled set to construct a new categorical variable. In the second stage, this new categorical variable is used as an independent or dependent variable in a regression. Throughout this paper, *hybrid study* is used to refer to this kind of data analysis that has two stages.

In the classification stage, researchers can choose not only algorithms but also different hyper-parameter values to optimize a performance metric. Common performance metrics used in the classification literature include accuracy, F-measure, or area under the receiving operating characteristic curve (AUC) (Provost et al. 1998). No matter which performance metric is chosen, the constructed variable still includes classification error because the output variable cannot be classified perfectly. In the second regression stage, it has been well documented in the econometrics literature that the misclassification of the variable constructed in the first stage may affect the estimation results of any kind of regression analysis. Particularly, if the constructed variable is used as the independent variable in a linear regression or a generalized linear model, the misclassification of this independent variable may lead to inconsistency of all coefficients. Similarly, the classification error of a newly constructed dependent variable may lead to inconsistency of coefficients in a binary choice regression model (Hausman 2001, Greene 2012).

Yang et al. (2018) is the first paper to study the properties and solutions of hybrid studies. The authors adopted a simulation-based method named MC-SIMEX that was first published in Küchenhoff et al. (2006) to correct the regression inconsistency when the new variable is used as an independent variable in the regression. Our study complements Yang et al. (2018) in the following ways. First, our paper derives *theoretical* solutions to correct the second-stage regression inconsistency by modifying theoretical results in the econometrics literature. Specifically, when the newly constructed variable is used as the independent variable in generalized linear models, our proposed theoretical method is the theoretical benchmarking case for MC-SIMEX. Second, when the constructed variable is used as the dependent variable in a binary choice model, the misclassification will also lead to inconsistency of coefficients, which was not discussed in Yang et al. (2018). Last, but not least, we show that the consistency of the estimator can be recovered; therefore, the first stage classifier should be tuned to minimize the variance of the corrected coefficient.

This study investigates in detail three types of hybrid studies. For ease of exposition, this paper uses the *proxy variable* to refer to the predicted label of the classification algorithm. The proxy variable is also the newly constructed variable that may be used as an independent or dependent variable in a second-stage regression. We use the *true variable* to refer to the true label predicted by the proxy variable. In the first case analyzed in this study, the proxy variable is used as the independent variable in a multiple linear regression. For example, Balakrishnan et al. (2010) classified annual reports into outperforming or underperforming. The classified output was used to predict the future earnings of the same firm. For this case, the present study adopted the statistical estimator developed in Bound et al. (1994) to correct the inconsistency of coefficient estimates. One distinct feature of hybrid studies is that researchers can estimate the covariance between the independent variable and classification error on the labeled set by cross-validation, which is unobservable in many situations. With this additional information, this study shows that in the second stage, a consistent estimator[1] can be derived even when the classification accuracy is low.

In the second case analyzed in this paper, the proxy variable is used as the dependent variable in a binary regression model. For instance, Wang et al. (2013) classified the firm disclosures into action-oriented or not. The classified output was used as the dependent variable in a logistic regression to examine the factors that may impact firm disclosure decisions. For this case, the present study follows the proofs in Hausman et al. (1998) to derive the consistent estimator.

In the third case, the proxy variable is utilized as the independent variable in a generalized linear model (GLM). For example, Mousavi et al. (2015) classified the content of the answers provided in Yahoo! Answers as professional or not. Then they employed negative binomial regression to examine the effect of whether the first answer is professional or not on the number of subsequent answers. For this case, the present study adopts the estimator in Buonaccorsi (2010) to solve the problem. Similar to the second case, because researchers can estimate probability models to capture the relationship between true variable and classification error on the labeled set in hybrid studies, the likelihood function can be modified so that the new estimator is consistent.

Because the coefficient estimators can be corrected to be consistent by mathematical formulas in all three cases, researchers do not need to tune classification algorithms to minimize the inconsistency (the difference between the estimated coefficient and the true value) when the sample size is large enough. Instead, one of the most important findings of this study is that hyper-parameters should be tuned to minimize the standard error of the proposed estimator in the present study.

We also analyzed the estimator variance when researchers conduct regression directly on the smaller data set with true labels (labeled set). Obviously, there is no classification error by this approach, but the regression sample size may become too small. To address this tradeoff, we derive one threshold proportion of the size of the unlabeled data set to the size of the labeled data set. This value can help researchers decide whether they should conduct regression on the labeled data set directly or apply our approach to the unlabeled data set with the proxy variable.

The contribution of this study is threefold. First, we derive theoretical formulas of consistent estimators for hybrid studies with the second stage being generalized linear models. Second, we propose that the classification algorithm hyper-parameters should be chosen to minimize the standard error of the proposed estimator in this study. Third, we derive one threshold proportion of the unlabeled data to the labeled data, beyond which researchers should apply our approach to the unlabeled data with the proxy variable. These three findings can improve the consistency and precision of regression analysis in future hybrid studies.

The remainder of the paper is organized as follows. Section 2 provides the literature review. Section 3 reports the main theoretical findings for three cases. Sections 4 and 5 evaluate the proposed solutions by experimentation with two simulated data sets and a real-world application, respectively. Section 6 concludes this paper.

2. Literature Review

2.1. Hybrid Study Applications

Hybrid studies, especially those using text mining, have been gaining in popularity in the information systems (IS) field in recent years (Chen et al. 2012). Abundant hybrid studies have been published during recent years in the information systems discipline (Ghose and Ipeirotis 2011, Aggarwal et al. 2012, Ghose et al. 2012, Wang et al. 2013, Chan and Wang 2014, Goes et al. 2014, Gu et al. 2014, Moreno and Terwiesch 2014, Singh et al. 2014, Mousavi et al. 2015, Zhang et al. 2016, Kim and Park 2017). Text classification algorithms are the most commonly used data mining methods in hybrid studies, which are used to classify the textual contents in online platforms, such as consumer reviews (Ghose and Ipeirotis 2011, Ghose et al. 2012), answers in a Q&A community (Mousavi et al. 2015), and postings in virtual communities about investment (Gu et al. 2014). Image classification is also gaining popularity. It is used to classify the images to construct new variables in online platforms, such as a worker's gender in a profile page (Chan and Wang 2014), the facial expressions of emotion in crowdfunding platforms (Kim and Park 2017), and the room photo quality on the Airbnb platform (Zhang et al. 2016).

Hybrid studies have also been adopted in other disciplines outside the IS community. Plentiful studies have used text mining techniques in financial domains. Kumar and Ravi (2016) reviewed 89 research papers from 2000 to 2016 on the various applications of text mining in finance. For example, Huang et al. (2014) applied text classification to extract textual opinions from analyst reports. The computed opinions were used to predict abnormal returns and future earnings growth. There is also a surge of interest concerning text mining in accounting studies (Li 2010). Li (2010) surveyed around 70 research papers on the textual analysis of corporate disclosures.

2.2. Measurement Error of Variables in Traditional Studies

The measurement error of the continuous variable and the misclassification of the binary variable have been widely studied by econometricians (Greene 2012) and bio-statisticians (Buonaccorsi 2010). This section will review the case of measurement error in the independent variable, followed by the case of the dependent variable.

We first review the case of a linear regression with only one independent variable: $Y = \beta_0 + X\beta_1 + \epsilon$. Suppose the independent variable includes measurement error ($W = X + e$), the ordinary least squares (OLS) estimator would be biased and inconsistent. When the independent variable is continuous, two types of measurement errors are analyzed in the literature: classical error and nonclassical error. If the measurement error, e, is independent of both X and error term ϵ, such an error is defined as classical measurement error (Carroll et al. 2006). In the simple linear regression, the estimate is downward biased toward zero under the assumption of classical measurement error (see Greene 2012 for proof). Under the assumption of nonclassical error, the direction of the bias is difficult to characterize and could be upward, which is less favorable in empirical studies. When the independent variable is a binary variable, the classification error would also result in inconsistency of the estimator. The binary explanatory variable case was analyzed in Aigner (1973).

When the measurement error occurs in a regression with more than one covariate, this issue becomes more serious. In multiple linear and nonlinear regressions, even when the other variables are measured precisely, the measurement error of one independent variable can cause the coefficients of all variables to be inconsistent in unknown directions (Buonaccorsi 2010; Carroll et al. 2006).

By contrast, a mis-measured continuous dependent variable under classical assumptions does not lead to inconsistent estimators in linear regression. The adverse effect is that estimated regression coefficients will have larger standard error (Greene 2012). However, the misclassification of the binary dependent variable in binary choice regression (i.e., logit or probit regression) can lead to inconsistent estimators when using standard specifications (Hausman et al. 1998).

Measurement error in hybrid studies has distinct features for the following reasons. First, in hybrid studies, measurement error is binary and can be estimated in the labeled set by cross-validation, which is unobservable in many empirical studies. Specifically, the variance of measurement error can be obtained by comparing model predictions from cross-validation with true values in the labeled data set. Second, researchers can choose and tune algorithms to affect the classification error, which cannot be altered in economic applications. Therefore, it is important to investigate how to tune the classification algorithm for a better estimation in the regression stage.

2.3. Measurement Error of Variables in Hybrid Studies

Yang et al. (2018) is the pioneering study that investigated measurement error in hybrid studies in information systems field. The authors analyzed the coefficient estimation inconsistency when the independent variable is either a continuous or discrete variable, both analytically and empirically. To correct the estimation inconsistency, they adopted two simulation-based methods, SIMEX and MC-SIMEX, which are applied to continuous variables with additive measurement error and discrete variables with misclassification (Cook and Stefanski 1994, Küchenhoff et al. 2006), respectively. The simulation-extrapolation (SIMEX) method was proposed by Cook and Stefanski (1994) to correct the estimation inconsistency when the continuous variable has additive measurement error (i.e., $W = X + e$) in regressions where the error variance σ_e^2 is known or estimated. The SIMEX method consists of two parts: simulation and extrapolation. In the simulation part, multiple versions of measurement error ($e(\lambda_k)$) are simulated with increasing error variance, $(1 + \lambda_k)\sigma_e^2$ (λ_k is a positive real number, e.g., 1, 2 ,. . ., and m). Then, multiple versions of W are simulated as $W(\lambda_1)$,$W(\lambda_2)$,... $W(\lambda_m)$, where $W(\lambda_k) = X + e(\lambda_k)$. In other words, the method generates variables with increasingly larger measurement error variance. Given $W(\lambda_k)$ with various degrees of measurement error, different versions of coefficient estimates $\theta(\lambda_k)$ can be obtained. In the extrapolation step, SIMEX first estimates one parametric model $\theta(\lambda)$, which fits the relationship between the magnitude of the measurement error (λ_k) and the coefficient $\theta(\lambda_k)$. Next, the method extrapolates $\theta(\lambda)$ to $\theta(-1)$, which can approximate the coefficient estimates under zero measurement error. MC-SIMEX was developed following a logic similar to SIMEX (Küchenhoff et al. 2006). Specifically, in the simulation step, $W(\lambda_k)$ is simulated by changing the values of W based on the λ_k^{th} power of the misclassification matrix. The extrapolation step is similar to that in SIMEX. One advantage of these two methods is their wide applicability to any second-stage regression model specification.

The effectiveness of SIMEX and MC-SIMEX has been demonstrated using both simulations and real-world applications in Yang et al. (2018). In most cases, coefficient inconsistency was reduced, and the corrected coefficients were close to the true values. However, Yang et al. (2018) also showed that the correction effectiveness of SIMEX and MC-SIMEX varied with (1) the amount of measurement error and (2) the model specification. Specifically, when the amount of measurement error increased, the correction effectiveness typically became worse. In addition, the correction effectiveness for linear and logit models were better than that for probit and Poisson models.

Our study complements Yang et al. (2018) in the following ways. First, our theoretical results provide the benchmark case for MC-SIMEX. Theoretically, the effectiveness of our approach does not vary with the amount of measurement error and model specification, which addresses the limitations of MC-SIMEX. The weakness of our approach is the generalizability to other types of regression models. If the regression model is not a generalized linear model, only MC-SIMEX is applicable. Second, we consider the case in which the proxy variable is used as the dependent variable in a regression model, which was not studied in Yang et al. (2018). Last, this study proposes tuning the classification algorithm to minimize the standard error of the corrected regression coefficient so that the adverse effects of the classification error are minimized. A summary of the differences between this paper and Yang et al. (2018) is provided in Table 1.

2.4. Classification Performance Metrics

Witten et al. (2016) is an excellent textbook that explains the basic concepts of classification and supervised learning in detail. Briefly put, in supervised learning algorithms, researchers need to manually label records in a small data set, which is randomly drawn from the full data set. In other words, researchers know the true labels in this subset of samples. This labeled subsample is used for training and evaluating the classifier. Specifically, the labeled subsample will be divided into two portions: one portion of the data (training set) is used to train the classifier and the remaining data (test set) is used to evaluate the classifier performance by comparing its predictions with the labels. Finally, this classifier is applied to the unlabeled set to construct a new categorical variable. Cross-validation is a more advanced evaluation method, wherein the labeled set is partitioned into K folds, and classifiers are iteratively trained on different sets of $K - 1$ folds and evaluated using the remaining fold. In this way, the labeled set is used for both training and testing.

For ease of exposition, let the label be binary: NO(0) or YES(1). Table 2 illustrates the general *confusion matrix* from the classification results for the test set. The total sample size is (TP + FP + FN + TN). The most intuitive performance metric is accuracy, which is defined as (TP + TN)/(TP + FP + FN + TN). In addition to accuracy, there exist many other performance metrics in the data mining literature, and researchers need to pick one metric subjectively. The *best performance metric* seems to be context dependent, and this remains an ongoing research topic. For example, Caruana and Niculescu-Mizil (2004) compared nine classification performance metrics: accuracy, lift, F-score, AUC, average precision, precision/recall, break-even point, squared error, cross entropy, and probability calibration.

Table 1. Comparison with Recent Literature

	Yang et al. (2018)	This paper
Solutions for continuous independent variable	Yes	No
Solutions for binary independent variable	Yes	Yes
Solutions for binary dependent variable	No	Yes
Closed-form solutions	No (simulation)	Yes
Can solutions correct the estimation inconsistency?	Partially	Completely
Can solutions be applied to all the regression models?	Yes	Only GLM
Regression precision as a new classification performance metric	No	Yes

Definitions of these nine metrics can be found in the appendix of Caruana and Niculescu-Mizil (2004).

In hybrid studies, most authors may follow the norms of choosing the widely used accuracy or AUC to select the best classifier. However, in hybrid studies, estimating regression coefficients in the second stage is the ultimate goal, which will be greatly influenced by the first-stage classification results. Therefore, the best classification performance metric should be a formula that depends on the information from both stages so that the regression estimation could be more accurate, which may not be achieved by using accuracy or AUC.

3. Theoretical Analysis

3.1. Case 1: Proxy Variable as the Independent Variable in Linear Regression

We first analyze the case in which the proxy variable in the first stage is used as the independent variable in linear regression in the second stage. Throughout this paper, we use the classification stage to refer to the first stage and regression stage for the second stage. The dependent variable in the regression stage is denoted by Y, which is an N-by-1 vector, and N is the sample size. The focal independent variable (true label) is denoted by X, which is an N-by-1 vector. X is observable only in the labeled set. In the classification stage, researchers build a classifier to predict X and the predicted variable is denoted by W, which is also an N-by-1 vector. As a result, W is observable in both labeled and unlabeled sets. The ideal regression is specified as

$$Y = X\beta + \boldsymbol{Z}\gamma + \varepsilon, \tag{1}$$

where β is the focal regression coefficient for estimation. We use the term *focal regression coefficient* because the primary finding of most hybrid studies is built on hypothesis testing regarding whether β is statistically different from 0. $\boldsymbol{Z}$ denotes the control variable, which is an N-by-K matrix and K is the number of control variables. γ is the regression coefficient vector of control variables, and ε is an N-by-1 vector of residuals with standard assumptions under which $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 \times I$. This paper inherits all OLS assumptions (e.g., A1 to A6 in section 2.3 of Greene 2012). It is common in the literature to model the relationship between X and W as

$$W = X + \mu, \tag{2}$$

where μ is the classification error, which is an N-by-1 vector with values being $-1, 0,$ or 1. Because only W is observable, most of the hybrid studies estimate the following regression model:

$$Y = W\hat{\beta} + \boldsymbol{Z}\hat{\gamma} + \varepsilon. \tag{3}$$

The term $\hat{\beta}$ is called the *naive estimator* throughout this paper. It has been well documented that the OLS estimator of Equation (3) is inconsistent. This matter gets worse in a multiple regression model because one misclassified independent variable causes all inconsistent coefficient estimates (Bound et al. 1994).

Table 2. General Confusion Matrix

	Actual (yes)	Actual (no)
Predicted (yes)	TP	FP
Predicted (no)	FN	TN

3.1.1. Consistency Analysis.

To solve this problem, we have reviewed the solutions in the literature. The pioneering solution is provided in Aigner (1973), which derives the closed-form solution of the inconsistent OLS estimator in Equation (3) under the assumption that the measurement error is uncorrelated with $\boldsymbol{Z}$. However, this assumption is violated in our problem because $\boldsymbol{Z}$ and μ are likely to be correlated with X, so that $\boldsymbol{Z}$ will be correlated with μ.[2] Relaxing this assumption, Bound et al. (1994) proposed the following consistent estimator for β and γ:

$$\begin{bmatrix} \gamma^* \\ \beta^* \end{bmatrix} = \left[I - \begin{bmatrix} a_{zz} & \boldsymbol{a}_{zw} \\ \boldsymbol{a}'_{zw} & a_{ww} \end{bmatrix}^{-1} \begin{bmatrix} 0 & \boldsymbol{a}_{z\mu} \\ 0 & a_{w\mu} \end{bmatrix} \right]^{-1} \begin{bmatrix} \hat{\gamma} \\ \hat{\beta} \end{bmatrix}. \tag{4}$$

Without loss of generality, all variables will be measured as deviations from their respective mean values. The terms $\hat{\beta}$ and $\hat{\gamma}$ are the naive OLS estimators of Equation (3). The term $a_{w\mu}$ is the covariance between

W and classification error μ. Similarly, $\boldsymbol{a}_{z\mu}$ and $\boldsymbol{a}_{zw}$ are the covariance vectors between $\boldsymbol{Z}$ and μ or W, respectively. The term a_{zz} is variance-covariance matrix of $\boldsymbol{Z}$. The term a_{ww} is the variance of W. In hybrid studies, although $a_{w\mu}$ and $\boldsymbol{a}_{z\mu}$ cannot be calculated on the unlabeled data set, they can be calculated on the labeled set (test set) by cross-validation, which will converge to the population covariance values when the sample size of the labeled set is large enough. When the labeled set is randomly divided into two portions for training and testing, $a_{w\mu}$ and $\boldsymbol{a}_{z\mu}$ can only be calculated on a portion of labeled set (test set). The theoretical results are the same for both evaluation methods.[3] To simplify the terminology, in Section 3, we conduct the theoretical analysis based on cross-validation as the evaluation method. Online Appendix A provides detailed proof of the solution.

Theorem 1. *Let the labeled data set be independent and identically distributed (i.i.d.) randomly drawn from the population. Assume that* (1) *Equation* (1) *meets all standard OLS assumptions, and* (2) ε *is uncorrelated with measurement error; then,* $\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix}$ *is a consistent estimator of* $\begin{bmatrix}\gamma\\ \beta\end{bmatrix}$ *in Equation* (1).

3.1.2. Variance Analysis.

Similar to other econometric problems, the other important criterion for assessing the quality of the estimated coefficient is the variance of the coefficient. Particularly, in hybrid studies, researchers can estimate Equation (1) by using only the labeled set. In this case, the estimator is also consistent because there is no classification error, but the variance of the coefficient could become larger because of the small sample size of the labeled set. We need to show that the proposed estimator has smaller variance compared with conducting regression by using true labels from the labeled data set. Theoretically, the variance-covariance matrices estimated from labeled and unlabeled data are given by Equation (5) (from Greene 2012) and Equation (6) (see Online Appendix A for the proof):

$$\operatorname{Var}\begin{bmatrix}\gamma'_{labeled}\\ \beta'_{labeled}\end{bmatrix} = \frac{\sigma^2}{n}\times\underbrace{\begin{bmatrix}a_{zz} & \boldsymbol{a}_{zx}\\ \boldsymbol{a}'_{zx} & a_{xx}\end{bmatrix}^{-1}}_{A}. \tag{5}$$

$$\operatorname{Var}\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix} = \frac{\sigma^2}{N} \underbrace{\left[\begin{bmatrix}a_{zz} & \boldsymbol{a}_{zw}\\ \boldsymbol{a}'_{zw} & a_{ww}\end{bmatrix} - \begin{bmatrix}0 & \boldsymbol{a}_{z\mu}\\ \boldsymbol{a}'_{z\mu} & 2a_{w\mu}\end{bmatrix} + \begin{bmatrix}0 & 0\\ \boldsymbol{a}'_{z\mu} & a_{w\mu}\end{bmatrix}\begin{bmatrix}a_{zz} & \boldsymbol{a}_{zw}\\ \boldsymbol{a}'_{zw} & a_{ww}\end{bmatrix}^{-1}\begin{bmatrix}0 & \boldsymbol{a}_{z\mu}\\ 0 & a_{w\mu}\end{bmatrix}\right]^{-1}}_{B}. \tag{6}$$

From these two equations, we can find that the variance and covariance terms in matrix A in Equation (5) and matrix B in Equation (6) are sample estimates, which will converge to the population variance or covariance values when the sample size is large enough. Moreover, the estimated coefficient variances will decrease proportionally in sample size because of the first terms in Equations (5) and (6), whereas matrices A and B are approximately constant across different sample sizes. Theoretically, given any fixed n, it is trivial to show that there exists an N that is large enough so that the variance from Equation (6) can be smaller than the variance from Equation (5) because matrix B will converge to a constant matrix when N approaches infinity.

In practice, the sample size of the labeled data set (n) is typically small due to the high labeling costs, whereas the sample size of the unlabeled data set (N) is large because of the increasing availability of big data and unstructured data. However, the value of proportion N/n still matters. For example, if the proportion that makes our new estimator achieve smaller variance were 10,000, our new method would be practically useless because in most cases, the sample size of labeled data set is large enough for direct regression. A similar question is: how large is the sample size of the unlabeled set so that using the new estimator achieves smaller variance than conducting regression on the labeled set directly? The following theoretical analysis derives one equivalent proportion, where using either approach leads to approximately the same variance, that is, $\operatorname{Var}(\beta^*_{unlabeled}) \approx \operatorname{Var}(\beta'_{labeled})$. Our proposed new method produces smaller variance only when the sample size of the unlabeled set divided by the sample size of the labeled set is larger than that equivalent proportion.

Given only the labeled data set, we propose estimating the equivalent threshold proportion by

$$Proportion = \frac{\operatorname{Var}(\beta^*_{labeled})}{\operatorname{Var}(\beta'_{labeled})} = \frac{\operatorname{Std}(\beta^*_{labeled})^2}{\operatorname{Std}(\beta'_{labeled})^2}, \tag{7}$$

where $\operatorname{Var}(\beta^*_{labeled})$ and $\operatorname{Var}(\beta'_{labeled})$ are the variances estimated by applying Equations (6) and (5) to the labeled data set. Equation (7) is calculated based on β, not other coefficients γ, because β is the focal parameter of interest. The term γ will be used to calculate the proportion when the newly constructed variable is the control variable. This approach is new in that even though the proportion is derived by using only the labeled data set, it can be used to assess whether the sample size of unlabeled data set is large enough. Specifically, Equation (7) measures the proportion of element in matrix A to that in matrix B because of the same sample size for $\operatorname{Var}(\beta^*_{labeled})$ and $\operatorname{Var}(\beta'_{labeled})$. This proportion can approximate the counterpart of $\operatorname{Var}(\beta^*_{unlabeled})$ and $\operatorname{Var}(\beta'_{labeled})$ because the elements in matrices A and B are sample estimates of variance or covariance components. As a result, the threshold sample size of the unlabeled set is estimated as $n \times proportion$.

3.1.3. The Full Solution of Case 1. First, researchers use Equation (7) to estimate the threshold proportion of the unlabeled data to the labeled data. If the sample size of unlabeled data is smaller than the threshold sample size and it is not possible to collect more data, they should just conduct regression on the labeled data. Otherwise, they can proceed to use the proposed new method in hybrid studies. Second, given any classifier, researchers use Equation (4) to correct a naive OLS estimator on the right-hand side to a consistent estimator on the left-hand side of Equation (4). Third, they should use the standard error from Equation (6) as the performance metric to select the best classifier in the classification stage. The final estimator will be a consistent and most precise solution in Case 1.

3.2. Case 2: Proxy Variable as the Dependent Variable in Binary Choice Model

We start with the standard specification of the binary choice model given by

$$P(Y=1|X,\boldsymbol{Z}) = F(X\beta + \boldsymbol{Z}\gamma); \\ y = Y + \mu. \tag{8}$$

The definitions of Y, X, and $\boldsymbol{Z}$ are the same as those in Section 3.1. The term $F()$ denotes the link function such as logistic function. The main difference is that Y is predicted by the proxy variable y constructed in the classification stage, which is also an N-by-1 vector. In other words, similar to the first case, Y is unobservable except for records in the labeled set, whereas its proxy y is observable on all records. In hybrid studies, researchers typically estimate the following regression model:

$$P(y=1|X,\boldsymbol{Z}) = F(X\hat{\beta} + \boldsymbol{Z}\hat{\gamma}). \tag{9}$$

If y is used as the dependent variable, Hausman et al. (1998) proved that the estimated β and γ are inconsistent because the model does not account for the misclassification of Y.

3.2.1. Consistency Analysis. Hausman et al. (1998) provided a closed-form solution under the assumption that misclassification probabilities are independent with covariates X and $\boldsymbol{Z}$. Relaxing this assumption, we propose the following solution. Formally, the misclassification probabilities are denoted by

$$a_0 = P(y=1|Y=0,X,\boldsymbol{Z}), \\ a_1 = P(y=0|Y=1,X,\boldsymbol{Z}).$$

For $P(y|Y,X,\boldsymbol{Z})$, we propose that the logit model can be applied to the labeled data set to estimate the probability model; then, researchers apply the estimated model to the unlabeled data set to calculate the misclassification probability. Finally, we derive the conditional probability function of y and the associated log-likelihood function for maximum likelihood estimation (MLE):

$$P(y=1|X,\boldsymbol{Z}) = a_0[1 - F(X\beta + \boldsymbol{Z}\gamma)] \\ + (1-a_1)F(X\beta + \boldsymbol{Z}\gamma), \tag{10}$$

$$L(\beta|y) = \sum_{i=1}^{N}\{y \ln P(y=1|X,\boldsymbol{Z}) \\ + (1-y)\ln[1 - P(y=1|X,\boldsymbol{Z})]\}. \tag{11}$$

Hausman et al. (1998) showed that MLE with the misclassification correction in Equation (10) can be used to consistently estimate β and γ as long as a mild monotonicity condition holds: $a_0 + a_1 < 1$. This condition implies that the sum of misclassification probabilities cannot be larger than the counterpart of a random classifier, which is easy to be satisfied.

Theorem 2. *Let the labeled data set be i.i.d. randomly drawn from the population. Assume that $a_0 + a_1 < 1$. β and γ in Equation* (8) *can be consistently estimated by applying MLE to Equation* (10).

3.2.2. Variance Analysis. Following Hausman et al. (1998), the asymptotic variance-covariance matrices are given by

$$\mathrm{Var}\begin{bmatrix}\gamma'_{labeled}\\ \beta'_{labeled}\end{bmatrix} = \underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{f_i^2}{F_i(1-F_i)}\begin{bmatrix}\boldsymbol{z_i}\\ x_i\end{bmatrix}\begin{bmatrix}\boldsymbol{z_i} & x_i\end{bmatrix}\right]^{-1}}_{C} \\ \times\frac{1}{n}, \tag{12}$$

$$\mathrm{Var}\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix} = \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N}\frac{(1-a_{i0}-a_{i1})^2 f_i^2}{P_i(1-P_i)}\begin{bmatrix}\boldsymbol{z_i}\\ x_i\end{bmatrix}\begin{bmatrix}\boldsymbol{z_i} & x_i\end{bmatrix}\right]^{-1}}_{D} \\ \times\frac{1}{N}, \tag{13}$$

where F_i is defined by Equation (8), P_i is defined by Equation (10), $f_i = \frac{\partial F_i}{\partial(x_i\beta + z_i\gamma)}$ is the derivative of F_i, and n and N are the sample sizes of the labeled data set and unlabeled data set. Online Appendix A provides detailed proof. From Equations (12) and (13), we can find that the terms in matrices C and D are sample means, which will converge to the population means when the sample size is large enough. Therefore, given a fixed sample size of labeled data set n, there exists a sufficiently large sample size N for unlabeled data set so that the variance from Equation (13) is smaller than that from Equation (12). To estimate the threshold proportion of unlabeled set to labeled set, similar to Case 1, we can estimate $\mathrm{Var}(\beta^*_{labeled})$ and $\mathrm{Var}(\beta'_{labeled})$ by applying Equations (13) and (12) to the labeled set.

3.2.3. The Full Solution of Case 2. First, researchers use Equations (7), (12), and (13) to estimate the threshold proportion. If the sample size of unlabeled data is large enough, they proceed to use the new estimator in the regression stage. Otherwise, they conduct second-stage regression by the true labels from the labeled data set. Second, given any classifier, researchers can consistently estimate β and $\boldsymbol{\gamma}$ by Equation (10). Finally, they use Equation (13) to choose the classifier with the minimized standard error in the classification stage.

3.3. Case 3: Proxy Variable as the Independent Variable in a Generalized Linear Model

Generalized linear models (GLMs) are formally defined by

$$P(Y|X,\mathbf{Z}) = G(X\beta + \mathbf{Z}\boldsymbol{\gamma}); \qquad (14)$$
$$W = X + \mu,$$

where $G()$ is the link function (Buonaccorsi 2010). The definitions of Y, X, and $\mathbf{Z}$ are the same as those in the previous sections. In this case, X is predicted by the proxy variable W constructed in the classification stage, which is also an N-by-1 vector. Therefore, similar to the first case, X is unobservable except for records in the labeled set, whereas its proxy W is observable on all records. When ignoring the classification error, researchers typically estimate the following regression model:

$$P(Y|W,Z) = G(W\hat{\beta} + \mathbf{Z}\hat{\boldsymbol{\gamma}}). \qquad (15)$$

However, the coefficients estimated from this model are not consistent because the probability function does not account for the relationship between X and W correctly (Buonaccorsi 2010).

3.3.1. Consistency Analysis. By probability theories, Buonaccorsi (2010) derived the following probability formula for $P(Y|W,\mathbf{Z})$ to account for the relationship between W and X:

$$P(Y|W,\mathbf{Z}) = \sum_X P(Y = y|X = x, W = w, \mathbf{Z} = z)$$
$$\times P(X = x|W = w, \mathbf{Z} = z).$$

Next, Buonaccorsi (2010) assumed that W provided no additional information about Y conditional on X. This assumption is imposed to greatly simplify the analysis and is a typical assumption imposed in statistics when solving similar problems. Given this assumption, it follows that

$$P(Y|W,\mathbf{Z}) = \sum_X P(Y = y|X = x, \mathbf{Z} = z)$$
$$\times P(X = x|W = w, \mathbf{Z} = z). \qquad (16)$$

The first term of the right-hand side in this equation is the same as that in Equation (14), which is the true model for estimation. For $P(X = x|W = w, \mathbf{Z} = z)$, the correction term of misclassification, we propose applying logit model to the labeled data set to estimate the probability model; then, researchers apply the model to an unlabeled data set to calculate the probability:

$$P(X = 1|W, \mathbf{Z}; \alpha, \boldsymbol{\theta}) = F(W\alpha + \mathbf{Z}\boldsymbol{\theta}). \qquad (17)$$

As a consequence, Equation (16) can be rewritten as

$$P(Y|W,\mathbf{Z}) = G(1 \times \beta + \mathbf{Z}\boldsymbol{\gamma})F(W\alpha + \mathbf{Z}\boldsymbol{\theta})$$
$$+ G(0 \times \beta + \mathbf{Z}\boldsymbol{\gamma})[1 - F(W\alpha + \mathbf{Z}\boldsymbol{\theta})]. \qquad (18)$$

This probability function can be used to derive the corrected likelihood function. Taking the binary model as an example, the log-likelihood function can be written as (Spiegelman et al. 2000)

$$\sum_{i=1}^{N}\left\{Y \ln\left[\sum_X G(X\beta + \mathbf{Z}\boldsymbol{\gamma})P(X|W,\mathbf{Z})\right] + (1 - Y)\right.$$
$$\left.\times \ln\left[1 - \sum_X G(X\beta + \mathbf{Z}\boldsymbol{\gamma})P(X|W,\mathbf{Z})\right]\right\}. \qquad (19)$$

Finally, researchers can consistently estimate β and $\boldsymbol{\gamma}$ by applying MLE to Equation (18).

Theorem 3. *Let labeled data set be i.i.d. randomly drawn from the population. Assume that W provides no additional information about Y conditional on X. Then β and $\boldsymbol{\gamma}$ in Equation* (14) *can be consistently estimated by applying MLE to Equation* (18).

3.3.2. Variance Analysis. We take logit and probit models as examples to illustrate the asymptotic variance-covariance matrices estimated from labeled and unlabeled data. Results are given by

$$\mathrm{Var}\begin{bmatrix}\gamma'_{labeled}\\ \beta'_{labeled}\end{bmatrix} = \left[\frac{1}{n}\sum_{i=1}^{n}\frac{g_i^2}{G_i(1-G_i)}\begin{bmatrix}\mathbf{z_i}\\ x_i\end{bmatrix}\begin{bmatrix}\mathbf{z_i} & x_i\end{bmatrix}\right]^{-1} \times \frac{1}{n}, \qquad (20)$$

$$\mathrm{Var}\begin{bmatrix}\gamma^*\\ \beta^*\end{bmatrix} = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_i(1-P_i)}\begin{bmatrix}(a_{i1}g_{i1} + a_{i0}g_{i0})\mathbf{z_i}\\ a_{i1}g_{i1}\end{bmatrix}\right.$$
$$\left.\begin{bmatrix}(a_{i1}g_{i1} + a_{i0}g_{i0})\mathbf{z_i} & a_{i1}g_{i1}\end{bmatrix}\right]^{-1} \times \frac{1}{N}, \qquad (21)$$

where G_i is defined by Equation (14), P_i is defined by Equation (18), $g_i = \frac{\partial G_i}{\partial(x_i\beta + z_i\gamma)}$ is the derivative of G_i, $g_{i1} = g(1 \times \beta + \boldsymbol{z_i\gamma})$, $g_{i0} = g(0 \times \beta + \boldsymbol{z_i\gamma})$, $a_{i1} = F(w_i\alpha + \boldsymbol{z_i\theta})$, and $a_{i0} = 1 - F(w_i\alpha + \boldsymbol{z_i\theta})$. Online Appendix A provides the proof. For other cases of GLM, the variance-covariance matrices have similar forms. Similar to the previous two cases, given a fixed n, there exists a large enough N such that using the new estimator obtains smaller variance than conducting regression by labeled set. The equivalent threshold proportion can be estimated by applying Equations (20) and (21) to the labeled data set.

3.3.3. The Full Solution of Case 3. First, researchers can use Equations (7), (20), and (21) to estimate the threshold proportion of the unlabeled set to the labeled set. If the unlabeled set is large enough, they proceed to use the new estimator. Second, researchers can consistently estimate β and $\boldsymbol{\gamma}$ by Equation (18). Finally, they choose the best classifier by Equation (21) to minimize the standard error of the focal regression coefficient.

4. Simulation Results

To assess the effectiveness of the new method discussed in Section 3, we conducted experimentation on two data sets. In the classification stage, we use two real-world data sets. Simulation is only used at the regression stage. Both data sets are downloaded from Kaggle.com, the largest data competition website. The first data set is about predicting review ratings by textual review content on Amazon (McAuley and Leskovec 2013). The original ratings have five categories, and we generate a new binary variable, *sentiment*, following Yang et al. (2018). Specifically, reviews with rating values four and five are labeled as positive, whereas reviews with ratings from one to three are labeled as negative. The proportion of positive reviews is around 50%, and therefore this classification problem has balanced label distribution. The second data set is about predicting comment toxicity by comment's content on Wikipedia (Wulczyn et al. 2016). The binary label is one if the comment is rude, disrespectful, or likely to make someone leave a discussion; otherwise, it is zero. The proportion of the toxic comments is around 33%, so the label distribution of this data set is slightly unbalanced.

For both data sets, the classification method is XGBoost (Chen and Guestrin 2016), which is the state-of-art classification algorithm that provides a parallel tree boosting and solves various data science problems in a fast and accurate way. The classification probability threshold is 0.5. As for the evaluation procedure, we randomly sample 2,000 rows as the labeled set and 30,000 rows as the unlabeled set while we know the actual labels of all rows. On the labeled data set, we use a fivefold cross-validation to build a classifier, which is used to predict the labels on the unlabeled set. The predicted labels in the 30,000 rows unlabeled set will be used in the regression stage to evaluate the performances of MC-SIMEX and our proposed method. The true labels can also be used in the regression stage as a first-best benchmarking case.

Moreover, we need a procedure to create several classifiers with different *accuracy* as the candidate classifiers. Our theories suggest that we can correct most of the inconsistency of the estimated coefficients given any classifier, including classifier with poor prediction accuracy. In this section, we will tune the nrounds hyper-parameter for XGBoost, which indicates the number of boosted trees to fit. We choose this parameter because it could be the most impactful parameter on classification accuracy, and it is a parameter that moderates underfitting and overfitting directly. We change nrounds from 3 to 300 with step value equaling to 3. Therefore, we will obtain 100 classifiers. The prediction performances of 100 classifiers for the review data and toxic comment data are shown in Figure 1, (a) and (b), respectively. In the figures, the x axis is the nrounds value divided by three, and the y axis is the performance metric value. We reported the values of accuracy, kappa, F1, and AUC in two figures. The results show that the prediction performance generally becomes better when the nrounds increases for two data sets.

In the second stage, we will use the predicted labels in the 30,000 rows unlabeled set as an independent or dependent variable, as described in Section 3. We will also simulate all other required variables in regression models. Last, we will estimate the regression in four ways: (1) without correction, (2) using MC-SIMEX, (3) using our method, and (4) running the regression on the labeled data set. The details of the simulation procedure and results will be discussed in the following three sections. A comparison of our metric versus commonly used traditional metrics, such as accuracy and AUC, is discussed in Section 4.4.

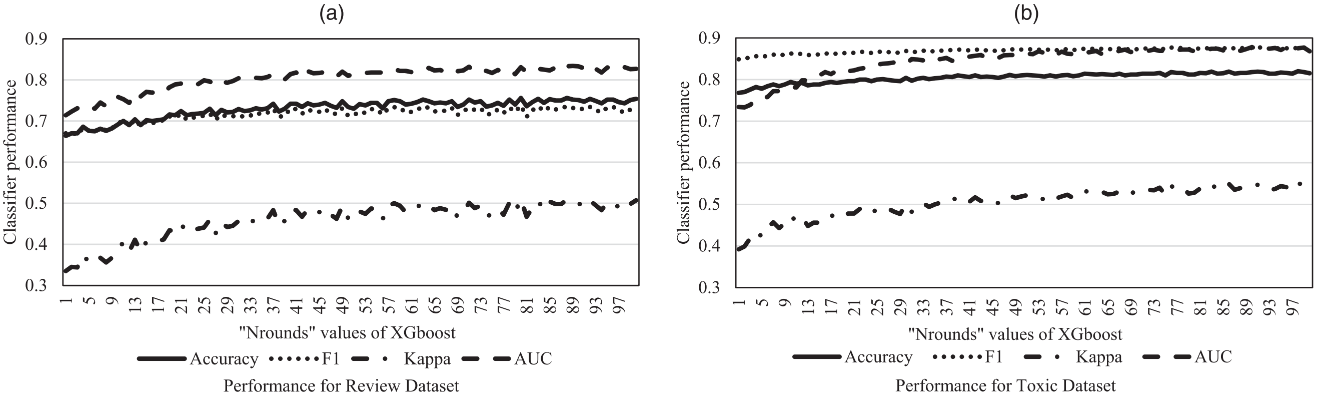
4.1. Case 1: Proxy Variable as the Independent Variable in Linear Regression

We simulate data by the following model with only two right-hand side variables:

$$Y = 0.2 + 2 \times X + Z + \varepsilon, \text{ where } Z = X + \epsilon.$$

First, recall that both X and its proxy variable, W, are observable from our data set; X is the review sentiment or comment toxicity; and Z is a control variable and is correlated with X, and this correlation is simulated by the second equation here. The two random noise terms are the only two variables that need to be simulated. ε and ϵ are drawn from a normal

Figure 1. Classification Performance for Two Data Sets



Performance for Review Dataset

Performance for Toxic Dataset

distribution with zero correlation. Then, the simulated values of Y and Z can be computed accordingly. These two equations also imply that the theoretical beta coefficients of X and Z are two and one, respectively.

Next, this study conducts regression analysis on the 30,000 rows unlabeled review data set. This study first conducts linear regression using true review sentiment (X) to derive the empirical coefficients. The estimated coefficients for X and Z are 2.049 and 0.992, respectively. For each nrounds value, we can derive the proxy variable W for X. We first run the naive regression by using the proxy variable (W) as the independent variable (without correction), following Equation (3) in Section 3.1. After we repeat the regression by using 100 different proxy variables based on different nrounds values, we can draw a scatter plot of nrounds values versus regression coefficients. The coefficient results of X and Z are shown in Figure 2, (a) and (b), by the solid lines, respectively. In the figures, the x axis is the nrounds value divided by three, and the y axis is the coefficient value. Next, we use Equation (4) in Section 3.1 to derive the corrected coefficients of X and Z, which are shown in Figure 2, (a) and (b), as dotted lines. In other words, the dotted lines depict the results of our proposed solutions. Similarly, coefficient results by MC-SIMEX are reported in Figure 2, (a) and (b), as dash-dotted lines. Last, we run the regression by using the true sentiment (Equation (1)) from the labeled data set and results of X and Z are reported as dashed lines.

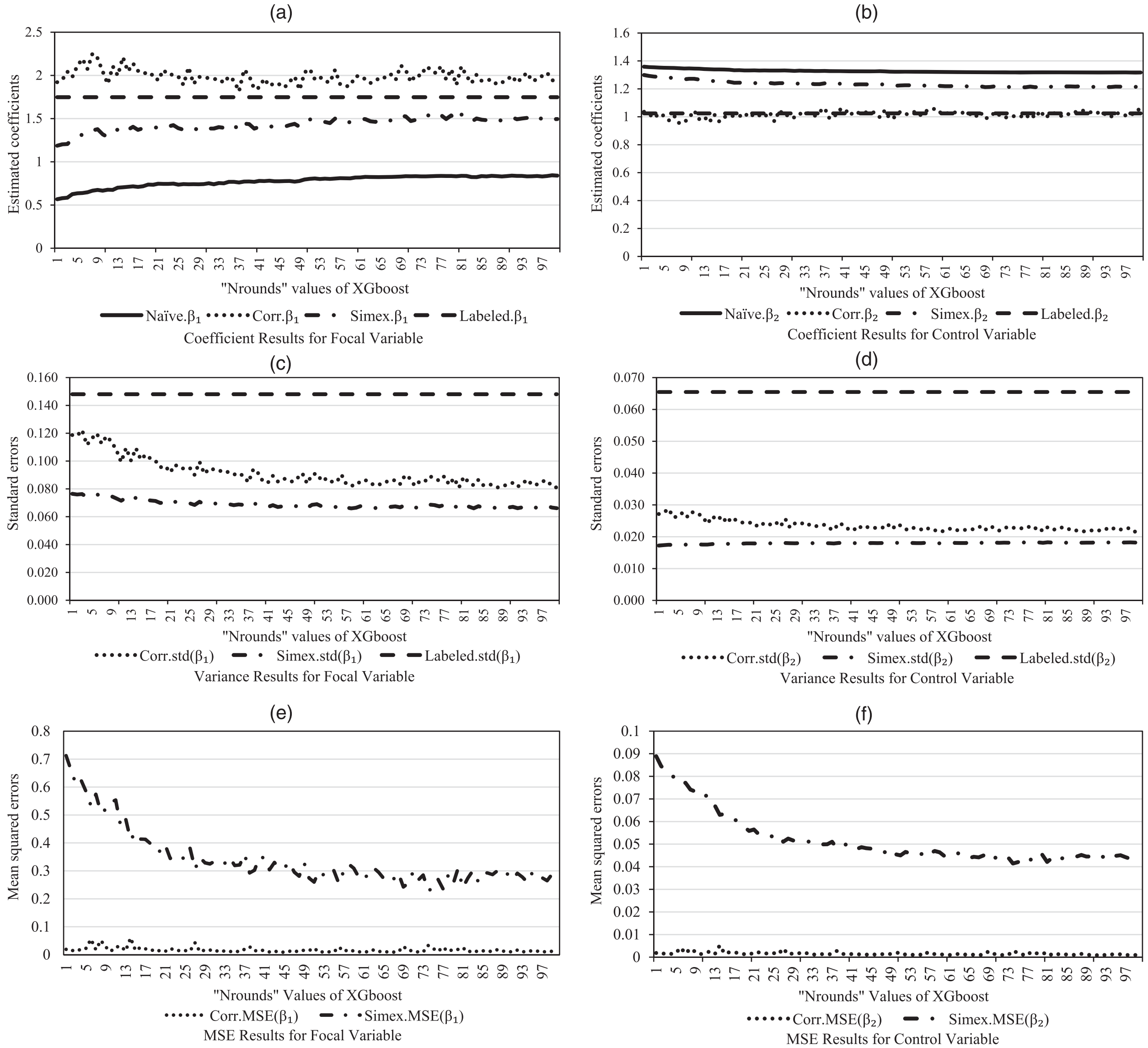
From the figures, we can find that the naive method used in the existing literature performs poorly and produces a much underestimated coefficient for X and an overestimated coefficient for Z. Second, the coefficients obtained by the regression using the labeled data are close to the theoretical values. Third, MC-SIMEX can only partially correct the estimation inconsistency. Finally, our method can correct the estimation of the regression coefficients and is not very sensitive to classifier performance, whereas the correction effectiveness of MC-SIMEX varies with classifier performance.

Before we compare the standard errors of coefficients, we check whether the sample size of unlabeled data is larger or smaller than the threshold sample size. Specifically, this study first conducts the linear regression using true review sentiment in the labeled set and the standard error of the sentiment is shown in Figure 3 as a solid line. Then, we derive the standard error of the sentiment by applying Equation (6) (our method) to the labeled set, which is shown in Figure 3 as a dotted line. The square ratio between these two lines is the threshold proportion, which is around 5 to 10 and smaller than the real value, 15 (30,000/2,000). This result indicates that applying our method to the unlabeled set can achieve smaller variance than running the regression on the labeled set. We also verify that this proportion is a good approximation. We randomly sample *threshold proportion* × 2,000 rows as the unlabeled set and derive the standard error by using our method, which is shown in Figure 3 as a dashed line. The results show that our estimated threshold proportion indeed achieves nearly the same variance as running the regression on the labeled set.

Last, we report the standard error results of X and Z in Figure 2, (c) and (d). The results show that, generally, our method and MC-SIMEX applied to the unlabeled set achieve smaller variance than running regression on the labeled set. Besides, given the same sample size, MC-SIMEX can achieve smaller variance than our method. The standard error obtained by our method will change following classification performance while MC-SIMEX has relatively stable standard error.

Even though MC-SIMEX has smaller variance, the estimated coefficient from MC-SIMEX has a larger bias from the true coefficient compared with our method. Estimator bias can be used to measure the

Figure 2. Simulation Results for Case 1



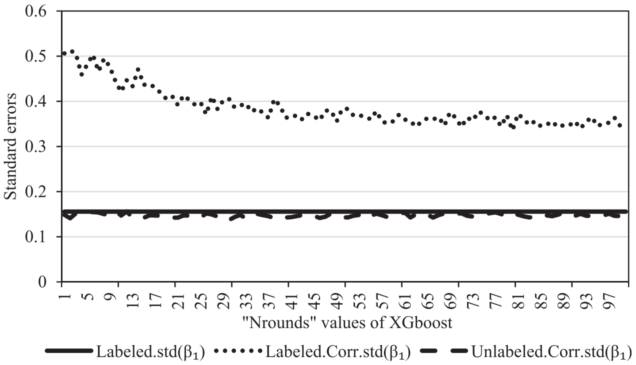
Notes. The term *Naive.*β_1 refers to the coefficient of X without correction; *Corr.*β_1, *Corr.std*(β_1), and *Corr.MSE*(β_1) refer to the coefficient, standard error, and MSE of X by our method, respectively. The terms *Simex.*β_1, *Simex.std*(β_1), and *Simex.MSE*(β_1) refer to the coefficient, standard error, and MSE of X by MC-SIMEX, respectively. The terms *Labeled.*β_1 and *Labeled.std*(β_1) refer to the coefficient and standard error of X after running the regression on the labeled set, respectively. The term β_2 refers to the coefficient of Z.

difference between the estimated coefficient and the true coefficient when the sample size is finite.[4] Because of the bias-variance tradeoff of two methods, the mean squared error (MSE) from the classic bias-variance decomposition theory is used to compare our method and MC-SIMEX comprehensively in our experiment.[5] In theory, MSE of an estimator can be decomposed into squared bias and variance (Geurts 2009),

$$\mathrm{MSE}(\hat{\beta}) = \mathrm{E}(\hat{\beta} - \beta)^2 = \left(\mathrm{E}(\hat{\beta}) - \beta\right)^2 + \mathrm{E}\left(\hat{\beta} - \mathrm{E}(\hat{\beta})\right)^2 = \mathrm{Bias}^2(\hat{\beta}) + \mathrm{Var}(\hat{\beta}). \qquad (22)$$

To calculate the sample MSE of the estimator, we replicate the second stage simulation 50 times. Then we apply our method and MC-SIMEX to derive the coefficients for each simulated data set. As a result, we can obtain 50 coefficients corrected by our method and those corrected by MC-SIMEX given each nrounds value. Finally, we can obtain MSE results of two methods. We report the MSE results of X and Z in Figure 2, (e) and (f). The results show that our method indeed achieves lower MSE than MC-SEIMX. Moreover, our method has relatively stable MSE while MSE of MC-SIMEX generally decreases when the classifier has better performance. The results from the

Figure 3. Proportion Results for Case 1



Note. The term *Labeled.std*(β_1) refers to the standard error of X by running regression on the labeled set; *Labeled.Corr.std*(β_1) refers to the standard error of X by applying our method to labeled set; and *Unlabeled.Corr.std*(β_1) refers to the standard error of X by applying our method to the unlabeled set with threshold sample size.

toxic comment data are qualitatively the same and, for brevity, are provided in Online Appendix C.

4.2. Case 2: Proxy Variable as the Dependent Variable in Binary Choice Model

We simulate data by the following logit model:

$$P(Y=1|X,Z)=\frac{exp(0.2+2\times X+Z)}{1+exp(0.2+2\times X+Z)}\text{, where}$$
$$X=0.5\varepsilon \text{ and } Z=0.5\epsilon+X.$$

The term Y is the review sentiment or comment toxicity, which is predicted by y. The terms ε and ϵ are drawn from a normal distribution with zero correlation. Given the simulated values of ε and ϵ, the values of X, Z, and $P(Y=1|X,Z)$ can be computed. Then, the binomial function is used to generate a Y value with probability $P(Y=1|X,Z)$. Next, we randomly match a record in review data or toxic comment data with the simulated Y, X, and Z so that the true label of review sentiment or comment toxicity equals simulated Y. The equations also imply that the theoretical beta coefficients of X and Z are two and one, respectively.

Next, this study conducts the traditional logit regression on the 30,000 rows unlabeled data set. We first run the logit model using true review sentiment Y. The coefficients of X and Z are 2.045 and 0.972, respectively. Similar to Case 1, we construct 100 classifiers by tuning nrounds parameter. For each classifier, we run the naive regression by using the proxy variable (y) as the dependent variable, following Equation (9) in Section 3.2. Next, we use Equation (10) in Section 3.2 to derive the corrected coefficients of X and Z. Then, we use MC-SIMEX to derive the corrected coefficients. Last, we run the regression by using the labeled data set. Figure 4, (a) and (b), displays the results of the coefficients. Our proposed solution is also applicable to the probit model, and the results are reported in Online Appendix B. Results from both the logit and probit models show that the proposed method indeed has corrected the coefficients around the theoretical values and is not very sensitive to classifier performance, whereas the naive method used in the existing literature again performed poorly across all cases. MC-SIMEX still partially corrects the inconsistency and generally performs better when the classifier is more accurate. Another important observation in this case is that our method can fully correct the coefficients for both probit and logit models, whereas MC-SIMEX's performance is sensitive to the model specification. Moreover, standard error results for X and Z are shown in Figure 4, (c) and (d). The results show that, generally, our method and MC-SIMEX applied to the unlabeled set achieve smaller variance than running the regression on the labeled set. This finding is also consistent with the fact that the threshold proportion range is around 5 to 15, which is smaller than the real value 15. The estimated threshold proportion is effective as illustrated in Figure 5. Besides, the results also show that MC-SIMEX achieves the smaller variance than our method. To compare our method and MC-SIMEX comprehensively, we derive the MSE results for the two methods. The results in Figure 4, (e) and (f), show that our method achieves better performance than MC-SIMEX in terms of MSE.

4.3. Case 3: Proxy Variable as the Independent Variable in GLM

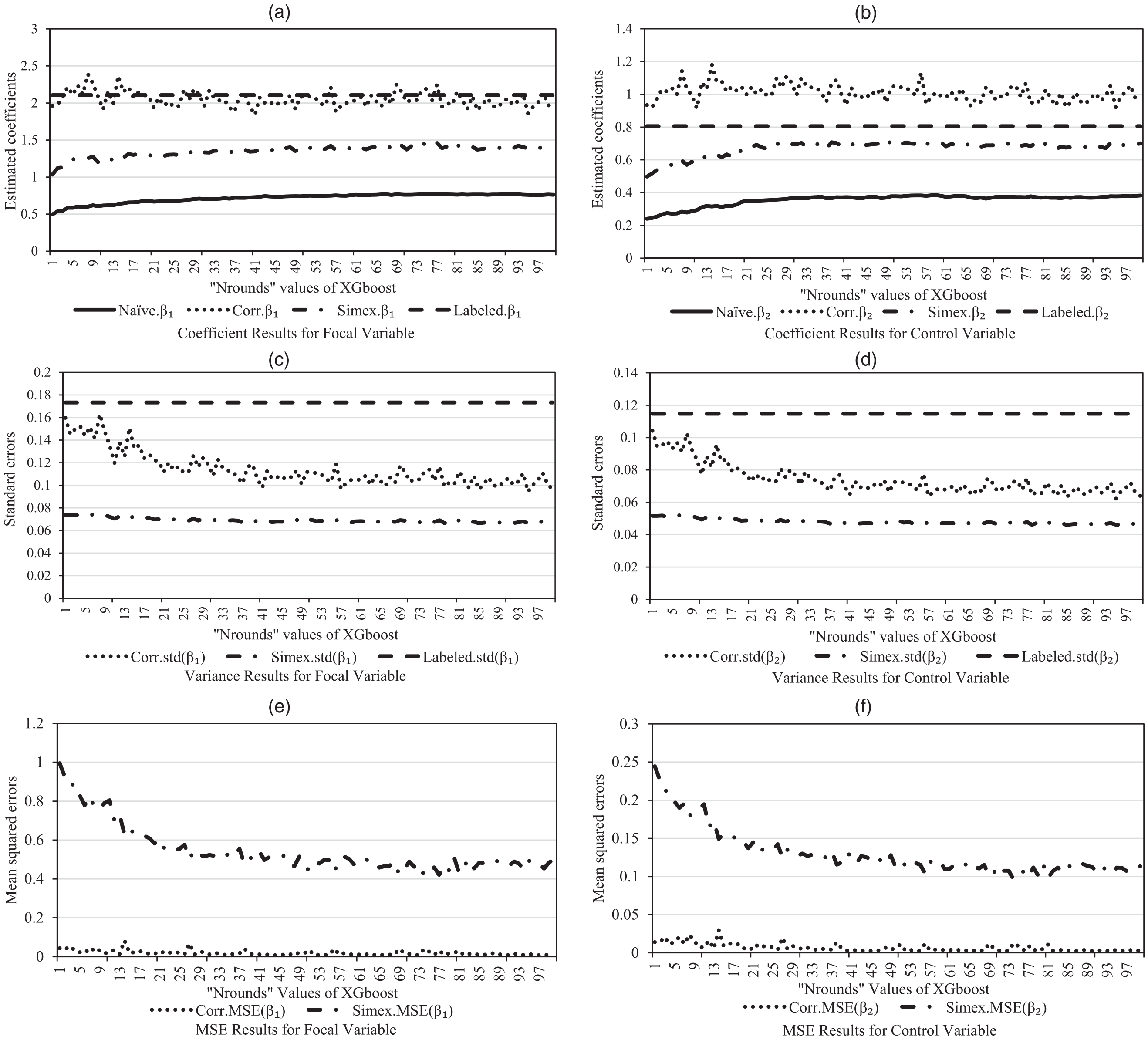
We simulate data by the following logit model:

$$P(Y=1|X,Z)$$
$$=\frac{exp(0.2+2\times X+Z)}{1+exp(0.2+2\times X+Z)},\text{ where } Z=X+\varepsilon.$$

The term X is review sentiment or comment toxicity, predicted by W; Z is a control variable and is correlated with X as specified in the second equation. ε is drawn from a normal distribution. Given the simulated values of ε and X, Z can be computed accordingly. Last, given the conditional probability function, the binomial function will be utilized to simulate Y. These two equations also imply that the theoretical beta coefficients of X and Z are two and one, respectively.

Next, we run logit regression using true review data to get the benchmark results. The coefficients for X and Z are 2.011 and 0.999, respectively. Similar to Case 1, we also construct 100 classifiers. For each classifier, we run the naive regression by using the proxy variable (W) as the independent variable, following Equation (15) derived in Section 3.3. Next, we use Equation (18) derived in Section 3.3 to obtain the

Figure 4. Simulation Results for Case 2 (Logit Model)



Coefficient Results for Focal Variable

Coefficient Results for Control Variable

Variance Results for Focal Variable

Variance Results for Control Variable

MSE Results for Focal Variable

MSE Results for Control Variable

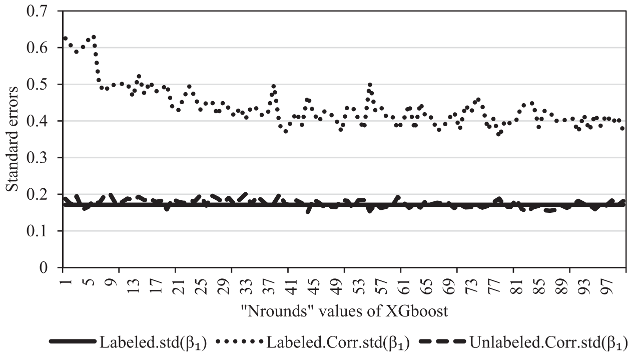
Note. The legend entries are the same as Figure 2.

corrected coefficients for X and Z. Then, we use MC-SIMEX to derive the corrected coefficients for X and Z. Last, we run the regression by using the true review sentiment in the labeled data set as the independent variable. Figure 6, (a) and (b), shows the coefficient results. In this case, we also evaluate our solution for the probit model. The results are included in Online Appendix B. All results are qualitatively the same as those in Cases 1 and 2. Moreover, standard error results shown in Figure 6, (c) and (d), also confirm that our method and MC-SIMEX achieve smaller variance than directly running the regression on the labeled data set. Figure 6, (e) and (f), shows the MSE results for our method and MC-SIMEX. We also evaluate the effectiveness of the threshold proportion. The threshold proportion range is around 4 to 9. Figure 7 shows that the threshold proportion can effectively estimate the threshold sample size for our method.

4.4. Classification Performance Metrics

This section compares the biases and variances of our corrected estimator given different classifiers selected by our metric and traditional metrics. We examine a comprehensive list of 13 metrics described in Table 3. Theoretically, our proposed correction method produces consistent estimates, which implies that the

Figure 5. Proportion Results for Case 2.



Note. The legend entries are the same as Figure 3.

estimator bias goes to zero asymptotically. Under this condition, researchers can focus on minimizing estimator variance. However, when the sample size of unlabeled data is finite, the estimator may be biased. Even though our metric minimizes the variance, it may inadvertently inflate the bias. Therefore, we will evaluate both biases and variances of our estimator for all the metrics.

The data simulation process is as follows. First, we repeat the simulation procedure of each case in the previous three sections 50 times. Then, for traditional metrics, we can directly select the best classifier from 100 classifiers because they only depend on classification outputs in the first stage. Given the classification output selected by each traditional metric, we can use our method to derive 50 corrected coefficients and standard errors from 50 simulated data sets. Next, we apply our method to 100 classification outputs for each simulated data set since our metric (min.std.error) depends on the estimation results in the second stage. Finally, we use our metric to select the best classifier from 100 classifiers for each data set. As a result, we can obtain 50 corrected coefficients and standard errors given 50 best classifiers selected by our metric. Given one simulated data set, the running time of traditional metrics in three cases is around 0.752, 1.424, and 1.816 seconds while the running time of our metric is around 8.866 seconds, 1.917 minutes, and 2.980 minutes. All experiments were conducted on an ordinary PC with Intel i7-6700 3.4-GHz CPU and 8 GB RAM. The running time of our metric is longer than that of traditional metrics because researchers only need to estimate the regression coefficient once for traditional metrics, whereas the regression needs to be estimated 100 times for our metric, as described in the simulation process. Even though the running time of our metric is longer, it is still acceptable in practice.

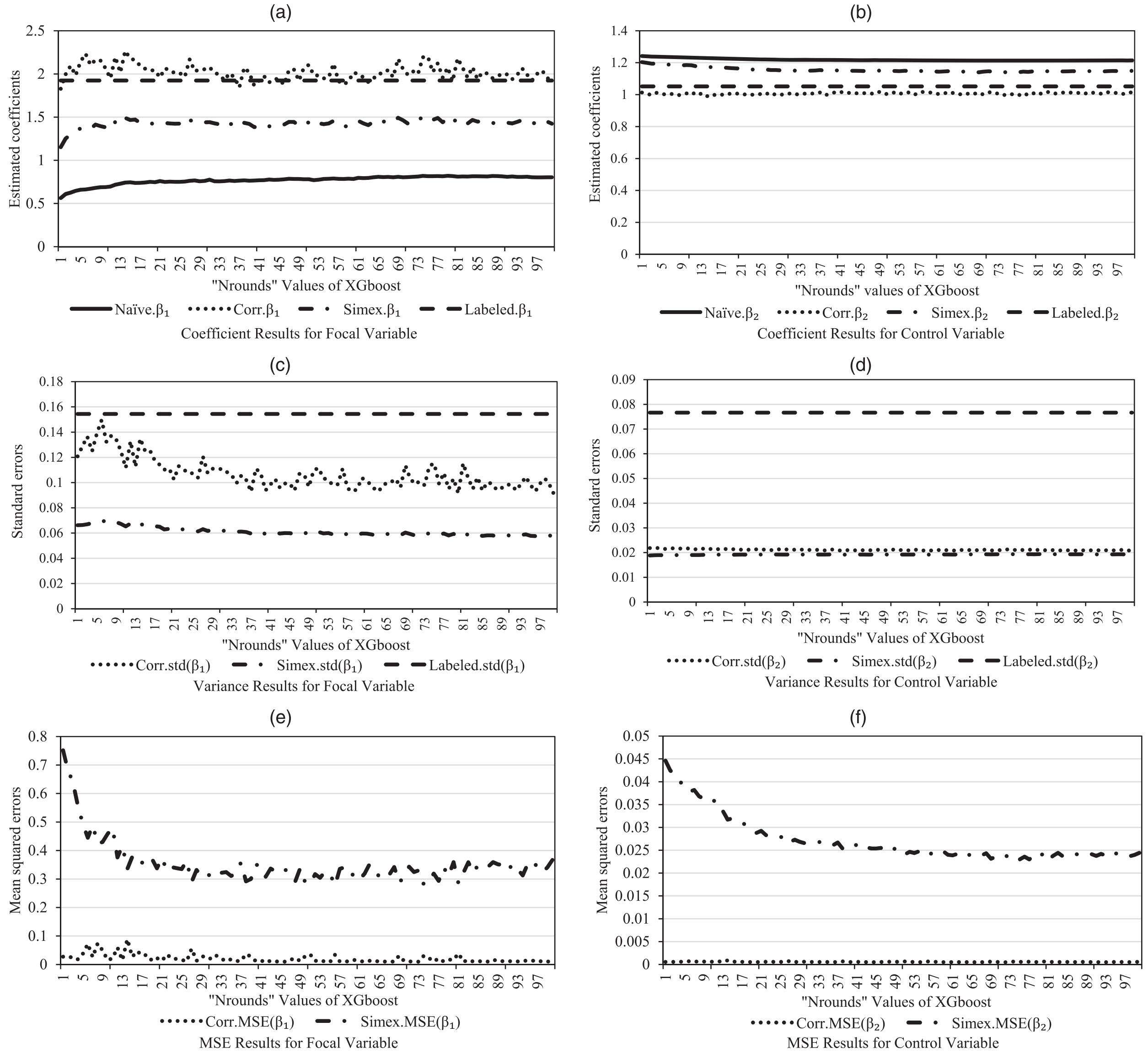
We report the squared bias and mean of the variance of the focal coefficient for each performance metric in Table 3. We also report the nrounds (Nrounds) value of XGBoost selected by each metric so we know whether different metrics select very different classifiers. MSE is the sum of the squared bias and variance columns. From the variance columns in Table 3, the following conclusions can be reached. First, our proposed performance metric (min.std.error) indeed produces the minimized variance. Second, among traditional metrics, except four metrics (sensitivity, specificity, AUC, and Brier score), all other metrics achieve nearly the same variance as our performance metric in all three cases. Third, sensitivity, AUC, and Brier score perform worse compared with the best traditional metrics. Last, specificity performs the worst among all the metrics. By our experiment, researchers should be more cautious in using AUC in hybrid studies because AUC is one of the most widely used metrics in classification problems. At least in our experiment, specificity should be avoided in hybrid studies.

From the bias results in Table 3, the following conclusions can be reached. First, the bias of our metric is smallest in Case 1 and Case 3. In Case 2, the bias is very close to the lowest case. Our metric does not inflate bias while it minimizes the variance of the estimated coefficient. Second, among traditional metrics, most metrics except three metrics (sensitivity, AUC, and Brier score) achieve great performance. Last, sensitivity, AUC, and Brier score produce higher bias compared with other metrics. In summary, from the bias-variance decomposition perspective, our experiment results suggest that our new metric can perform quite well. The most widely used AUC should be used with caution because its performance is surprisingly worse than the simplest metric such as accuracy in our experiment. Sensitivity and specificity should be avoided because they tend to choose very different classifiers judging by Nrounds. Those two metrics should be used only in special cases, such as cost sensitive classification.

5. Applications to Real-World Data Set

The Amazon review data described in Section 4 are used to conduct real-world applications for evaluating the solutions for linear and logit models. Because of the limited number of variables in the data set, we cannot evaluate the second case. We will examine the impact of review sentiment (positive or negative) on the review helpfulness with two control variables, review word count, and the total number of votes for review helpfulness. In linear regression (Case 1), review helpfulness is operationalized as the ratio of helpful votes to total votes received by a review, following Ghose and Ipeirotis (2011). In logit regression (Case 3), the dependent variable is one when

Figure 6. Simulation Results for Case 3 (Logit Model)



Note. The legend entries are the same as Figure 2.

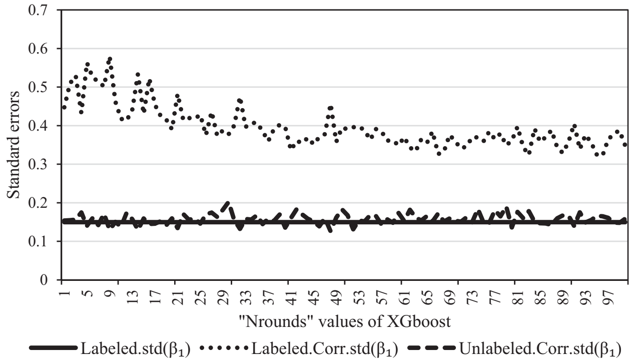
the review received at least one helpful vote and is zero otherwise. The model specifications are given as

$$\begin{aligned} &Log/Logit(Helpfulness)\\ &\quad = \beta_0 + \beta_1 \times Sentiment + \beta_2 \times \log(Votes)\\ &\qquad + \beta_3 \times \log(Words) + \varepsilon. \end{aligned}$$

This study first conducts two regression models using true review sentiment to derive the true coefficients, which are reported in Table 4. As shown in Table 4, the sentiment was positively associated with review helpfulness. Compared with positive reviews, negative reviews were less likely to receive helpful votes. In addition, longer reviews and reviews with more votes were more likely to be perceived as helpful. Next, we evaluate different methods. Similar to the cases in Section 4, for each of 100 classifiers, we run the naive regression by using the predicted sentiment as the independent variable (the typical worst case without any correction). Next, we use our solutions in Sections 3.1 and 3.3 to derive the corrected coefficients. We also use MC-SIMEX to derive the coefficients. Last, we run the regression on the labeled data set.

Figures 8 and 9 report the coefficient results for the linear and logit models, respectively. In these figures, the y axis shows the differences between the estimated coefficients by four different methods and the true coefficient. The terms β_1, β_2, and β_3 refer to the coefficient differences of the review sentiment, the number of votes, and word count. We also evaluate

Figure 7. Proportion Results for Case 3



Note. The legend entries are the same as Figure 3.

our solution in Case 3 for the probit model. Due to the page limit, the results are reported in Online Appendix D.

Comparing *Naive*.β_1 with 0 shows that, without any correction, the estimator is considerably biased downward in all the cases. Comparing *Naive*.β_2 and *Naive*.β_3 with 0 shows that the misclassification in the sentiment variable biased the coefficients of precisely measured variables to different degrees. Comparing *Labeled*.β_1, *Labeled*.β_2, and *Labeled*.β_3 with 0 shows that, the estimates obtained from the labeled data set have larger differences than those obtained from our method, possibly resulting from the large standard error of estimated coefficients. Comparing *Simex*.β_1, *Simex*.β_2, and *Simex*.β_3 with 0 shows MC-SIMEX can correct the inconsistent coefficients partially. Moreover, MC-SIMEX generally performs better when the classifier has better performance. MC-SIMEX tends to approach the correction performance of our method when the number of rounds of XGBoost is large enough.

Comparing *Corr*.β_1, *Corr*.β_2, and *Corr*.β_3 with 0 shows the effectiveness of our method in correcting estimation inconsistency, which is not very sensitive to classifier performance. Moreover, our method's performance does not vary with the model specifications. The slight difference between the corrected coefficient and true value may result from the omitted variable bias. For example, whether the product description is informative or not may be correlated with both review sentiment and review helpfulness. Moreover, the omitted variable bias from the true sentiment is different from that from the predicted sentiment because the covariance between true sentiment and the error term is different from the covariance between the predicted sentiment and the error term. The variance results are qualitatively the same as those in simulation cases and are included in Online Appendix D because of page limit.

6. Conclusions

Inspired by the pioneering work of Yang et al. (2018), our paper investigated the theoretical foundations for addressing the classification error in hybrid studies. This paper contributes to the literature in the following ways. First, we summarize and modify existing proofs in econometrics to derive theoretical formulas of consistent estimators for hybrid studies that are widely used in modern research in business disciplines. One surprising finding is that we can correct the estimation inconsistency even for classifiers with low accuracy. Second, we also examine the case of misclassified dependent variable, which was not studied in the literature. Third, because the consistency of estimated coefficients can be recovered by our formulas, we propose using *minimizing the standard error of the corrected beta coefficient* as the performance metric for hyperparameter tuning in the classification stage. Last,

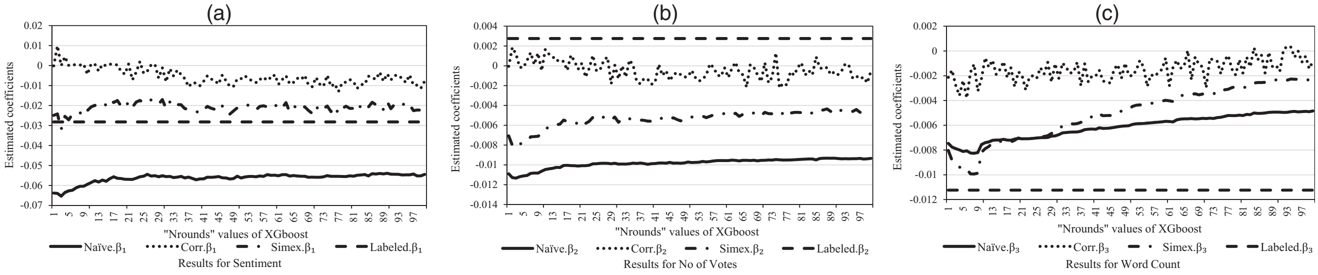
Table 3. Performance Metric Comparison Results for Review Data Set

	Case 1			Case 2			Case 3		
	Bias^2	Variance	Nrounds	*Bias*^{2}	Variance	Nrounds	*Bias*^{2}	Variance	Nrounds
Min.std.error	0.0000	0.0075	262	0.0003	0.0110	250	0.0001	0.0099	232
Accuracy	0.0001	0.0076	288	0.0000	0.0111	288	0.0005	0.0102	288
F1	0.0000	0.0076	222	0.0001	0.0111	222	0.0002	0.0101	222
Kappa	0.0001	0.0076	288	0.0000	0.0111	288	0.0005	0.0102	288
MCC	0.0001	0.0076	288	0.0000	0.0111	288	0.0005	0.0102	288
Sensitivity	0.0048	0.0081	228	0.0050	0.0122	228	0.0200	0.0127	228
Specificity	0.0003	0.0141	9	0.0000	0.0202	9	0.0022	0.0181	9
Bacc	0.0001	0.0076	288	0.0000	0.0111	288	0.0005	0.0102	288
Informedness	0.0000	0.0076	222	0.0001	0.0111	222	0.0002	0.0101	222
PPV	0.0000	0.0076	222	0.0001	0.0111	222	0.0002	0.0101	222
NPV	0.0007	0.0077	276	0.0005	0.0113	276	0.0030	0.0108	276
Markedness	0.0001	0.0076	288	0.0000	0.0111	288	0.0005	0.0102	288
AUC	0.0085	0.0081	285	0.0098	0.0124	285	0.0114	0.0117	285
Brier	0.0085	0.0081	285	0.0098	0.0124	285	0.0114	0.0117	285

Table 4. Regression Results Using True Review Sentiment

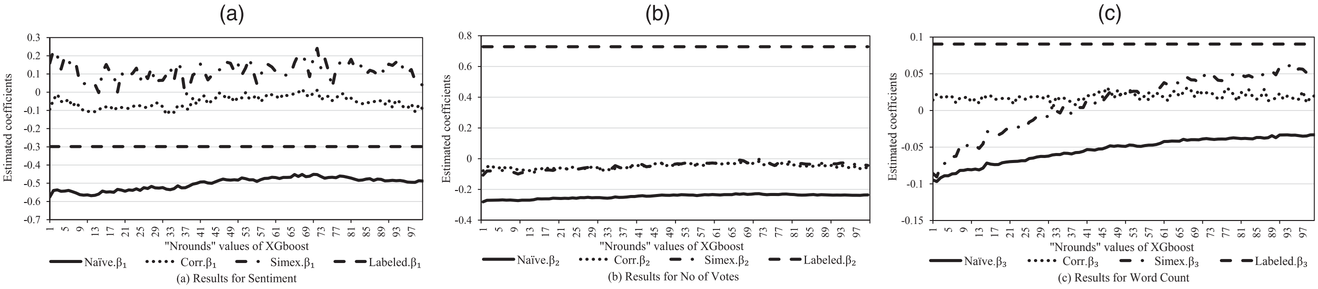
	Sentiment	Standard deviation	No. of votes	Standard deviation	Word count	Standard deviation
Linear model	0.105	(0.004)	0.234	(0.002)	0.027	(0.002)
Logit model	1.197	(0.054)	4.670	(0.064)	0.287	(0.035)

Figure 8. Results for Linear Regression



Note. The legend entries are the same as Figure 2.

Figure 9. Results for Logit Regression



Note. The legend entries are the same as Figure 2.

we derive one threshold proportion of the unlabeled data set to the labeled data set, beyond which our method is superior. In summary, our proposed new estimation procedure can produce a consistent and most precise estimator in hybrid studies.

We evaluate our solutions using both simulations and one real-world application. Our results confirm that the proposed method can adequately correct the inconsistency of estimated coefficients, whereas MC-SIMEX can partially correct the coefficients in all the cases. Theoretically, the performance of our method does not differ much with variations in the base rate of the data set, classifier type, and classifier's hyperparameter because our theoretical correction uses only the misclassification information from labeled data set to achieve consistency. The related additional simulation results are provided in Online Appendix B. In addition to consistency, our method indeed can minimize the variance of the corrected estimator. Most traditional metrics can produce similar estimation results in our experimentation. However, some metrics, including the widely used AUC, perform poorly in our experiment, and therefore researchers should be more cautious when using AUC to select classifiers in hybrid studies.

The first limitation of the present study is that our theoretical results are applicable only to linear regressions and generalized linear models but not to more complicated models, including panel models, time series models, and duration models. This also relates to the advantage of MC-SIMEX, which is one approach for any regression model. Theoretical results of other models could be derived, but it is not possible to cover all cases in one paper. Second, we propose tuning parameters to minimize the variance of one independent variable. If there are more than one independent variable constructed from different classifiers, then we may need a new criterion to optimize the regression estimation. Similarly, if the constructed variable is used in more than one regression equation, the current method is not applicable because one classifier may have different variances in two regressions. Third, even though our metric can minimize the variance, it may not achieve the least bias. The best solution may be using mean squared error of the coefficient as the performance metric. However, optimizing over MSE cannot be achieved practically because we cannot observe the true coefficient, which is the core element of MSE formula.[6] Last, there exist various methods of imputing missing variables to improve regression analysis and the most famous solution is the multiple imputation approach. It is possible that the regression estimation is better if we use predicted labels in the regression based on the values of dependent variables or other covariates in the second stage regression. More theoretical analysis is needed along this direction.

Acknowledgments

We thank the senior editor (Ram Gopal), the associate editor (Gautam Pant), the anonymous reviewers, and the participants at the 2018 Workshop on Information Technologies and Systems (WITS) for valuable comments and suggestions. The authors also thank Gedas Adomavicius, Mochen Yang, Yi Yang, Kunpeng Zhang, Sunil Mithas, Sean Xin Xu, D. J. Wu, Khim-Yong Goh, Jungpil Hahn, Nan Chen, and Tuan Quang Phan for valuable comments.

Endnotes

[1] The *consistent estimator* refers to the property that as the sample size of the data set increases indefinitely, the estimate of a parameter will converge in probability to the true value of the parameter.

[2] The term X correlates with μ because both are binary variables. X and Z are correlated because otherwise, the estimated regression coefficient of X will not be affected by the inclusion of Z.

[3] The terms $a_{w\mu}$ and $\boldsymbol{a}_{z\mu}$ estimated based on either evaluation method can be generalized to unlabeled data set.

[4] Estimator bias measures the difference between this estimator's expected value and the true value of the parameter.

[5] We thank the anonymous reviewer for providing this suggestion.

[6] One solution is that researchers calculate empirical MSE by replacing the unknown true coefficient values with the corresponding unbiased estimates from labeled set.

References

Aggarwal R, Gopal R, Gupta A, Singh H (2012) Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Inform. Systems Res.* 23(3-part-2):976–992.

Aigner DJ (1973) Regression with a binary independent variable subject to errors of observation. *J. Econometrics* 1(1):49–59.

Balakrishnan R, Qiu XY, Srinivasan P (2010) On the predictive ability of narrative disclosures in annual reports. *Eur. J. Oper. Res.* 202(3):789–801.

Bound J, Brown C, Duncan GJ, Rodgers WL (1994) Evidence on the validity of cross-sectional and longitudinal labor market data. *J. Labor Econom.* 12(3):345–368.

Buonaccorsi JP (2010) *Measurement Error: Models, Methods, and Applications* (CRC Press, Boca Raton, FL).

Carroll RJ, Ruppert D, Crainiceanu CM, Stefanski LA (2006) *Measurement Error in Nonlinear Models: A Modern Perspective* (Chapman and Hall/CRC, Boca Raton, FL).

Caruana R, Niculescu-Mizil A (2004) Data mining in metric space: An empirical analysis of supervised learning performance criteria. *Proc. 10th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 69–78.

Chan J, Wang J (2014) Hiring biases in online labor markets: The case of gender stereotyping. *Proc. 35th Internat. Conf. Inform. Systems (ICIS), Auckland, New Zealand*.

Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: from big data to big impact. *Management Inform. Systems Quart.* 36(4):1165.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proc. 22nd ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (Association for Computing Machinery, New York), 785–794.

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* 89(428):1314–1328.

Geurts P (2009) *Bias vs Variance Decomposition for Regression and Classification. Data Mining and Knowledge Discovery Handbook* (Springer, New York).

Ghose A, Ipeirotis PG (2011) Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10): 1498–1512.

Ghose A, Ipeirotis PG, Li B (2012) Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Marketing Sci.* 31(3):493–520.

Goes PB, Lin M, Yeung CMA (2014) "Popularity effect" in user-generated content: Evidence from online product reviews. *Inform. Systems Res.* 25(2):222–238.

Greene WH (2012) *Econometric Analysis* (Pearson, Boston).

Gu B, Konana P, Raghunathan R, Chen HWM (2014) Research note: The allure of homophily in social media: Evidence from investor responses on virtual communities. *Inform. Systems Res.* 25(3): 604–617.

Hausman JA (2001) Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *J. Econom. Perspective* 15(4):57–67.

Hausman JA, Abrevaya J, Scott-Morton FM (1998) Misclassification of the dependent variable in a discrete-response setting. *J. Econometrics* 87(2):239–269.

Huang AH, Zang AY, Zheng R (2014) Evidence on the information content of text in analyst reports. *Accounting Rev.* 89(6):2151–2180.

Kim J, Park J (2017) Does facial expression matter even online? An empirical analysis of facial expression of emotion and crowdfunding success. *Proc. 38th Internat. Conf. Inform. Systems (ICIS), Seoul, South Korea*.

Küchenhoff H, Mwalili SM, Lesaffre E (2006) A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics* 62(1):85–96.

Kumar BS, Ravi V (2016) A survey of the applications of text mining in financial domain. *Knowledge Base. Systems* 114:128–147.

Li F (2010) Textual analysis of corporate disclosures: A survey of the literature. *J. Accounting Literature* 29:143.

McAuley JJ, Leskovec J (2013) From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. *Proc. 22nd Internat. Conf. World Wide Web* (Association for Computing Machinery, New York), 897–908.

Moreno A, Terwiesch C (2014) Doing business with strangers: Reputation in online service marketplaces. *Inform. Systems Res.* 25(4):865–886.

Mousavi R, Raghu T, Frey K (2015) Assessing order effects in online community-based health forums. *Proc. 36th Internat. Conf. Inform. Systems (ICIS), Fort Worth, TX*.

Provost FJ, Fawcett T, Kohavi R (1998) The case against accuracy estimation for comparing induction algorithms. *Proc. 15th Internat. Conf. Machine Learn.* (Morgan Kaufmann, San Francisco), 445–453.

Singh PV, Sahoo N, Mukhopadhyay T (2014) How to attract and retain readers in enterprise blogging? *Inform. Systems Res.* 25(1):35–52.

Spiegelman D, Rosner B, Logan R (2000) Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J. Amer. Statist. Assoc.* 95(449):51–61.

Wang T, Kannan KN, Ulmer JR (2013) The association between the disclosure and the realization of information security risk factors. *Inform. Systems Res.* 24(2):201–218.

Witten IH, Frank E, Hall MA, Pal CJ (2016) *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, Cambridge, MA).

Wulczyn E, Thain N, Dixon L (2016) Wikipedia detox. figshare. Accessed February 23, 2017, http://doi.org/10.6084/m9.figshare.4054689.

Yang M, Adomavicius G, Burtch G, Ren Y (2018) Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Inform. Systems Res.* 29(1): 4–24.

Zhang S, Lee D, Singh PV, Srinivasan K (2016) How much is an image worth? An empirical analysis of property's image aesthetic quality on demand at AirBNB. *Proc. 37th Internat. Conf. on Inform. Systems* (ICIS, Dublin, Ireland).