

Robust Optimization for Inference on Machine Learning Generated Variables

Aaron Schechter
University of Georgia
aschechter@uga.edu

Weifeng Li
University of Georgia
weifeng.li@uga.edu

Abstract

Leveraging supervised machine learning (SML) algorithms to operationalize constructs from unstructured data like text or images is becoming common in practice and research. As a result, variables generated through SML are used in regression models to make inferences and test theories. However, variables produced by SML will have measurement errors compared to the underlying construct. We propose using robust optimization to reduce the negative impact of these errors and produce less biased coefficient estimates while conducting more accurate hypothesis testing. To extend the burgeoning literature on this issue, our proposed method focuses on the generalized research setting where a flexible number of dependent and independent variables are measured by SML algorithms. We combine recent robust optimization techniques to fit a linear regression model in the presence of uncertain measurement error. We theoretically demonstrate the consistency and efficiency of the robust approach. Through simulations, we demonstrate the effectiveness of our approach.

Keywords: Robust optimization, machine learning, statistical inference, regression.

1. Introduction

With the increasing availability of unstructured data such as text and images, the information systems research community has grown a significant interest in operationalizing constructs from unstructured data using supervised machine learning (SML) methods. SML methods estimate measures through learning the

underlying relationships between constructs of interests (e.g., sentiments) and unstructured data (e.g., customer reviews). Applications of SML methods include determining the quality of images (Zhang, Lee, Singh, & Srinivasan, 2022), classifying the sentiment of product reviews (Tirunillai & Tellis, 2012), predicting post quality in stock discussion boards (Gu, Konana, Rajagopalan, & Chen, 2007), and more. SML methods have shown great potential for providing reliable measurements for constructs of theoretical or practical importance.

To incorporate SML-based variables into econometric analysis, past hybrid studies often use a two-step estimation framework (Qiao & Huang, 2021): in the first step, SML methods are used to develop measures of interest from unstructured data such as text and images; in the second step, these SML-based measures are included in an empirical regression model. Such a two-step estimation framework expands the scope of the information systems research community by allowing researchers to examine phenomena and test theories in previously unquantifiable contexts. In the two-step estimation framework, the variables generated by SML generally have measurement errors, originating from SML methods' imperfect estimates of the target constructs (Yang, Adomavicius, Burtch, & Ren, 2018; Qiao & Huang, 2021). For example, when using text mining SML to predict customer satisfaction from textual reviews, a researcher might face the risk of mistaking a satisfied customer for an unsatisfied one or vice versa. Such measurement errors can cause biases in the second step estimation. Specifically, measurement errors in the first step can attenuate or amplify the coefficient estimates of SML-based variables and further distort the estimation of the dependent variable in the second step (Carroll, Ruppert, Stefanski, &

Crainiceanu, 2006). In response to measurement errors, several error correction techniques have been proposed, including the method of moments (Qiao & Huang, 2021), simulation extrapolation (Yang et al., 2018), instrumental variables (Yang, McFowland III, Burtch, & Adomavicius, 2022), and likelihood analysis (Qiao & Huang, 2021). Findings from the existing error correction techniques underscore the need for dealing with measurement errors that are often unknown, heteroscedastic, or significantly deviant from the normal distribution, which is common for SML-based variables.

This study proposes a prescriptive method capable of enhancing the two-step estimation framework through mitigating measurement errors induced by SML prediction. To complement prior research (Yang et al., 2018; Qiao & Huang, 2021; Yang et al., 2022), our study is specialized in the generalized scenario where a flexible number of variables, including dependent and independent, are generated by SML methods with heteroscedastic measurement errors. Robust optimization seeks to provide robust solutions without requiring knowledge about the precise distribution of the errors and can therefore find the best solution in the worst scenario (e.g., the measurement errors are as egregious as possible); we refer to these types of errors as adversarial. We propose a robust optimization model that incorporates an uncertainty set learned from labeled data (Hong, Huang, & Lam, 2021) to account for the unknown SML-induced measurement errors, and incorporates this uncertainty directly into the model. This labeled data is a necessary component of the supervised machine learning process, and is thus available to researchers applying this method. We theoretically demonstrate the consistency and efficiency of the robust method, and propose a correction term to further reduce bias. To evaluate our model with respect to the bias and standard errors of the estimated coefficients we conducted a series of simulation experiments. We found that applying our robust method with the correction term produces less biased estimates than OLS with respect to the true coefficients. Moreover, the uncorrected robust optimization model yields smaller standard errors and larger test statistics for the model coefficients, suggesting it may be an effective tool for increasing statistical power in noisy data.

2. Background

2.1. Measurement Error Bias

Measurement error bias, broadly speaking, arises when one or more independent variables are not

measured accurately. For example, when researchers use surveys to measure some unobservable value (e.g., job satisfaction, comfort with technology, etc.), there will inevitably be some error, i.e., survey questions do not perfectly capture the underlying construct (Bound, Brown, Duncan, & Rodgers, 1994). By contrast, errors in variables generated by machine learning are due to inaccuracies in a prediction made by an algorithm. For example, consider an algorithm that analyzes Twitter data to determine customers' opinions about a company. This SML tool would potentially make erroneous predictions about posts that were sarcastic or used slang. While the resulting errors might be functionally similar they are caused by different sources. In one case, the data does not sufficiently describe the construct. In the other, the method of translating the data is fundamentally imperfect.

In the case of simple linear regression with classical error, it is known that the effect of measurement error in simple linear regression is to attenuate the estimate of the slope parameter β in the direction of zero. In multiple regression analysis, the classical measurement error can affect the coefficient estimation of both the erroneous variable and other variables (Greene, 2018). Moreover, nonlinear second-step regression models can face not only underestimated coefficient estimates but also overestimated coefficient estimates (Carroll et al., 2006; Yang et al., 2018). This issue may be exacerbated by heteroskedastic measurement errors, such as those generated by the application of some SML algorithms.

2.2. Correcting Errors in Hybrid Studies

To demonstrate the key ideas of the existing error correction methods in hybrid studies, we first introduce the formulation of a hybrid linear regression as the context: we assume that the dependent variable $Y \in R^N$ is an N -by-1 vector of continuous real numbers and that the independent variable $\tilde{X} \in R^N$ is measured by an SML algorithm. In particular, the SML-generated variable could be represented as the true value and the SML-induced error $\tilde{X} = X + \Delta X$ where X is the error-free but unobservable true value of the estimated measure and ΔX is the SML-induced measurement error. Hybrid studies aiming to estimate the regression model $Y = \tilde{X}\beta + \varepsilon$ typically achieve a biased OLS estimator $\hat{\beta} = \lambda\beta$, where λ is the reliability ratio $\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{\Delta X}^2}$, σ_X^2 is the variance of the true value, and $\sigma_{\Delta X}^2$ is the variance of the measurement error. To address this problem, existing research has proposed four approaches.

The first approach relies on the **method of moments**.

Here, the measurement error could be corrected by estimating the reliability ratio $\hat{\lambda}$ with moments such as the variance of SML measurement error $\sigma_{\Delta X}^2$ and the variance-covariance matrix of the independent variable σ_X^2 (Carroll et al., 2006). Specifically, the estimated reliability ratio can be calculated as $\hat{\lambda} = (\sigma_X^2 - \sigma_{\Delta X}^2)/\sigma_X^2$ and the corrected estimate becomes $\hat{\beta}/\hat{\lambda}$. The method of moments can be extended to find a consistent estimator for the regression models having one SML-generated independent variable and multiple error free control variables.

Second, researchers can use **simulation extrapolation**. The simulation extrapolation method aims to adjust the effect of measurement error through simulation experiments (Cook & Stefanski, 1994; Yang et al., 2018). The first step of the simulation extrapolation method is to simulate a large number of additional datasets, with M defined as the number of datasets. Each of these sets of data comes with a successively large measurement error variance $(1 + \zeta_m)\sigma_{\Delta X}^2$, where $0 = \zeta_1 < \zeta_2 < \dots < \zeta_M$. For each additional dataset, an OLS estimate $\hat{\beta}_m$ of the coefficient β is calculated. With dependent variable $\hat{\beta}_m$ and independent variable ζ_m gained from these additional simulated datasets, we can estimate a nonlinear regression model. The unbiased estimate of β can then be obtained by extrapolating to $\zeta = -1$. With accurate estimates of measurement error variance σ_u^2 , this method demonstrates competitive performance in correcting the OLS estimator β (Yang et al., 2018).

Third, The measurement error could be viewed as an endogeneity problem where the observed data \tilde{X} is correlated with the error term ε . Consequently, **instrumental variables** could be introduced to correct the estimation (Carroll & Stefanski, 1994). One candidate for the instrument X^{IV} could be the variable predicted by the SML model \tilde{X} (Fong & Tyler, 2021). Alternatively, (Yang et al., 2022) applies the random forest method to identify the instrument and shows that when the prediction from one individual tree serves as the SML-based variable in the linear regression, strong instruments could be selected from the predictions of other individual trees in the random forest.

Finally, the measurement errors can be directly modeled in the framework of **maximum likelihood estimation** (MLE). The likelihood analysis builds upon the specification of a parametric model for the data, which explicitly takes the measurement error into consideration. In particular, the errors in the parametric model are assumed to follow a specific probability distribution with a fixed mean and variance. For example, researchers could model the false positive

and false negative rates of SML misclassification as independent random variables following the Bernoulli distribution and subsequently incorporate these random variables into MLE (Qiao & Huang, 2021).

2.3. Robust Optimization and Regression

As an alternative to the described measurement error correction techniques, we introduce robust optimization. Robust optimization broadly refers to the collection of techniques devised for finding optimal solutions to problems in which the data input is noisy, incomplete, or erroneous (Ben-Tal, El Ghaoui, & Nemirovski, 2009; Bertsimas, Brown, & Caramanis, 2011). This branch of research provides computationally tractable methods for solving real world problems in which data or parameters are inherently uncertain. Recent advances have also coupled robust optimization with big data (Hong et al., 2021; Bertsimas, Gupta, & Kallus, 2018) and machine learning (Bertsimas, Dunn, Pawlowski, & Zhuo, 2019; Gabrel, Murat, & Thiele, 2014; Kuhn, Esfahani, Nguyen, & Shafieezadeh-Abadeh, 2019). In the case of two-step estimation with SML-based variables, there is uncertainty due to measurement error in the generated variables. Further, all or a subset of variables can be subject to error. The elements of a robust optimization problem include nominal or original data; an uncertainty set, which specifies the range over which the nominal data may vary; and an objective function, which is a deterministic combination of model data and parameters that the researcher would like to maximize or minimize.

3. The Proposed Method

3.1. The Regression Problem with Error

We consider the most general case, where all variables are subject to some error due to the application of SML techniques. We define the dependent variable $Y \in R$ as a vector of continuous real numbers; let $y \in R^N$ be an N -by-1 vector of observations. Further, let $X \in R^{N \times p}$ be an N -by- p matrix of independent variables, where p is the number of features. Assuming that both dependent and independent variables are observable with perfect accuracy, the true underlying regression model is $Y = X\beta + \varepsilon$ where $\beta \in R^p$ is the vector of regression coefficients and $\varepsilon \in R^N$ is a vector of random errors. We assume all of the standard OLS conditions (Greene, 2018; Wooldridge, 2010) and that the errors ε satisfy $E[\varepsilon] = 0$ and $E[\varepsilon^T \varepsilon] = \sigma^2$. In other words, the only potential bias in the model is assumed to be from the difference between the observed study variables and the true values, not some other form of error. For simplicity, we also assume that each feature

in the matrix X and the vector Y is mean centered and thus can ignore an intercept term. The objective of this regression problem is to conduct hypothesis testing on whether each entry of β is significantly different from 0.

If the true values of Y and X are not directly observable, it is common to use observable proxies for these variables \tilde{Y} and \tilde{X} (Carroll et al., 2006; Buonaccorsi, 2010) modeled as:

$$\tilde{X} = X + \Delta X; \quad \tilde{Y} = Y + \Delta Y \quad (1)$$

where $\Delta Y \in R^N$ is an N -by-1 vector of discrepancies between the true and observable values of Y , and $\Delta X \in R^{N \times p}$ is an N -by- p vector of discrepancies between the true and observable values of X . In the context of using machine learning to estimate variables, \tilde{X} could be the estimated sentiment values of online reviews, X the true sentiment, and ΔX the error from the machine learning algorithm. Note that we assume error can be introduced from both the left and right side of the regression equation (Bound et al., 1994; Stefanski, 2000). We assume that the observed values \tilde{X} and \tilde{Y} are correlated with the errors generated by the respective machine learning methods, ΔX and ΔY ; in other words, $E[\tilde{X}^T \Delta X] \neq 0$ and $E[\tilde{Y}^T \Delta Y] \neq 0$. We also assume that the errors associated with the independent variables are correlated with the errors associated with the dependent variable, i.e., $E[\tilde{X}^T \Delta Y] \neq 0$ and $E[\tilde{Y}^T \Delta X] \neq 0$. This assumption could be relaxed for simplicity; this correlation is likely close to zero with large sample sizes (Bound et al., 1994). However, we allow the correlation to be non-zero in our theoretical analyses.

Given these observable values, it is straightforward to solve the regression problem $\tilde{Y} = \tilde{X}\beta + \varepsilon$ with β estimated using ordinary least squares. The ordinary least squares approach would be to find the point estimate $\hat{\beta}^{OLS}$ that minimizes the Euclidean norm of the residuals $\tilde{X}b - \tilde{Y}$ for the observable data. This estimator has a closed form solution, given as $\hat{\beta}^{OLS} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$. As pointed out in the literature, solving this problem results in a biased estimator due to the unobservable error (Wooldridge, 2010; Bound et al., 1994). This is broadly referred to as measurement error bias.

3.2. The Robust Formulation

To formulate the robust optimization problem, we borrow the notation from Section 3.1 and assume that our objective is to minimize the residuals with respect to the true (nominal) data. In other words, we want to find a point estimate b that minimizes $\|Xb - Y\|$. Because we are only able to observe \tilde{X} and \tilde{Y} , we

can rewrite this expression as $\|Xb - Y\| = \|(\tilde{X} - \Delta X)b - (\tilde{Y} - \Delta Y)\|$ where the entities ΔX and ΔY are the differences between the observed and true data values. If the distribution of these errors were known with precision we could minimize this quantity directly. Instead, we model the measurement errors as belong to a finite and bounded set, $[\Delta X, \Delta Y] \in U$, where U is specified by the researcher. A common choice of uncertainty set is the ellipsoidal case, where the norm of the errors are finitely bounded by some constant. This choice has been used numerous times in the robust optimization literature, particularly in adversarial cases (i.e., worst errors, rather than distributions of errors) (Bertsimas & Nohadani, 2019; Xu, Caramanis, & Mannor, 2008; Bertsimas et al., 2018; Ben-Tal et al., 2009). The advantage of the ellipsoidal uncertainty set is that it is typically less conservative than methods such as linear intervals while also making the robust optimization problem more tractable. Other methods such as distributional robust optimization (Delage & Ye, 2010; Xu, Caramanis, & Mannor, 2012) use different forms of uncertainty sets, however those methods require distributional assumptions about the errors. In the present research, the bound on the uncertainty set is approximated using labeled data.

Given this uncertainty set, we can therefore rewrite the uncertainty set as $\|\Delta X, \Delta Y\|_F \leq \rho$ for $\rho \geq 0$, where $\|\cdot\|_F$ is the Frobenius norm. Our objective now is to find the parameters b which minimize the squared residuals while accounting for any perturbation within the uncertainty set U . Therefore, we want to optimize the function based on the *biggest errors* possible in the uncertainty set. The robust regression problem is then

$$\hat{\beta}^R = \min_b \max_{\|\Delta X, \Delta Y\|_F \leq \rho} \|(\tilde{X} - \Delta X)b - (\tilde{Y} - \Delta Y)\| \quad (2)$$

This type of formulation is known as an adversarial robust optimization problem, and has been studied specifically in the context of minimizing the squared loss function (El Ghaoui & Lebret, 1997; Eldar, Ben-Tal, & Nemirovski, 2004; Chandrasekaran, Golub, Gu, & Sayed, 1998). If there are no errors, then the uncertainty set is simply the empty set and the problem is equivalent to ordinary least squares. In the case that there are errors (i.e., the machine learning procedure is not perfect), we can obtain a closed-form solution for $\hat{\beta}^R$ following prior work.

For a fixed estimator b , let $r(\tilde{X}, \tilde{Y}, b, \rho) = \max_{\|\Delta X, \Delta Y\|_F \leq \rho} \|(\tilde{X} - \Delta X)b - (\tilde{Y} - \Delta Y)\|$ be the value of the worst-case residual vector norm for the observed data. This value can also be thought of

as the largest possible value of the root-mean square error. The worst-case residual norm has a closed form expression $r(\tilde{X}, \tilde{Y}, b, \rho) = \|\tilde{X}b - \tilde{Y}\| + \rho\sqrt{\|b\|^2 + 1}$ for a fixed b , which can be demonstrated using the triangle inequality (El Ghaoui & Lebret, 1997). Thus, the robust optimization problem is equivalent to solving the nonlinear program $\min_b \|\tilde{X}b - \tilde{Y}\| + \rho\sqrt{\|b\|^2 + 1}$. As noted by (Xu et al., 2008), this form establishes a link between regularized regression and robust optimization.

Following (El Ghaoui & Lebret, 1997) and (Chandrasekaran et al., 1998), this problem has a closed form solution:

$$\begin{aligned}\hat{\beta}^R &= \min_b \|\tilde{X}b - \tilde{Y}\| + \rho\sqrt{\|b\|^2 + 1} \\ &= (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}\end{aligned}\quad (3)$$

Here α has the specific value of $\alpha = \rho(\lambda - \tau)/\tau$; λ and τ are the unique solutions of the following second-order conic program:

$$\begin{aligned}\min \lambda, \quad \text{subject to: } & \|\tilde{X}b - \tilde{Y}\| \leq \lambda - \tau \\ & \rho\|b\| \leq \tau\end{aligned}\quad (4)$$

This optimization problem is equivalent to solving the original robust problem in Equation (2), and it can be solved in polynomial time (El Ghaoui & Lebret, 1997). If $\alpha = 0$, $\hat{\beta}^R = \tilde{X}^+ \tilde{Y}$ where \tilde{X}^+ is the Moore-Penrose pseudo-inverse of \tilde{X} . The case where $\alpha = 0$ only occurs when $\tilde{X}b = \tilde{Y}$ exactly for some vector b . For the rest of this paper we consider the case that $\alpha \neq 0$.

In summary, for a given tolerance level ρ , we can obtain a vector of robust regression coefficients by solving the optimization problem in Equation (4) and using the formula in Equation (3). This problem can be solved in polynomial time, and in practice can be run efficiently even on very large datasets. In the following sections we detail how to correct this estimator to reduce bias, and how to calculate standard errors.

3.3. Consistency Analysis of the Robust Estimator

We first consider whether the robust solution in Equation (3) is biased with respect to the true value of the estimator β . We assume that there are N_1 observations where the true values are known exactly (Δ is known), and $N_2 = N - N_1$ observations where the errors are unknown. In other words, there is some subset of the data with exact labels. While this might not be reasonable for some scenarios (such as using surveys to measure latent constructs), in the case of training a supervised machine learning algorithm there will always

be some labeled data to train the model with. It is assumed that $N_2 \gg N_1$, though this is not necessary for the following theoretical results. Based on our definitions, we know that $\tilde{Y} = (\tilde{X} - \Delta X)\beta + \Delta Y + \varepsilon$. Plugging in this expression, we can expand the formal definition of $\hat{\beta}^R$ as follows:

$$\begin{aligned}\hat{\beta}^R &= (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T (\beta \tilde{X} - \beta \Delta X + \Delta Y + \varepsilon) \\ &= (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{X} \beta \\ &\quad - (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \Delta X \beta \\ &\quad + (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \Delta Y \\ &\quad + (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \varepsilon\end{aligned}\quad (5)$$

From Equation (5) we can make two observations. First, when $\alpha > 0$ and/or $[\Delta X, \Delta Y] \neq 0$ the robust solution $\hat{\beta}^R$ is a biased estimator of β . Of course, the OLS solution is also biased; as a result, we are interested in how biased the robust solution is relative to OLS. To make this comparison, we first establish that $\hat{\beta}^R = (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{X} \hat{\beta}^{OLS}$. From this relation, we conclude that for $\alpha > 0$ the robust estimator will be an attenuation of the OLS estimator. However, as the sample size increases, this multiplicative factor will converge to the identity matrix. In other words, for very large samples, the robust estimator will be about as accurate as the naive OLS approach.

Our second observation is that $E[\hat{\beta}^R]$ can be expressed as an exact function of β .

$$\begin{aligned}E[\hat{\beta}^R] &= \beta - E\left[(\alpha I + \tilde{X}^T \tilde{X})^{-1} (\alpha I + \tilde{X}^T \Delta X) \beta\right. \\ &\quad \left. + (\alpha I + \tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \Delta Y\right] \\ &= \beta \left[I - \left(\Sigma_{\tilde{X}, \tilde{X}} \right)^{-1} \left(\Sigma_{\tilde{X}, \Delta X} \right) \right] \\ &\quad + \left(\Sigma_{\tilde{X}, \tilde{X}} \right)^{-1} \Sigma_{\tilde{X}, \Delta Y}\end{aligned}\quad (6)$$

Here $\Sigma_{\tilde{X}, \tilde{X}} = E[\tilde{X}^T \tilde{X}]$ is the variance-covariance matrix of the observed data \tilde{X} , and $\Sigma_{\tilde{X}, \Delta X} = E[\tilde{X}^T \Delta X]$, $\Sigma_{\tilde{X}, \Delta Y} = E[\tilde{X}^T \Delta Y]$ are the covariance matrix of the observed data \tilde{X} and the machine learning generated errors ΔX and ΔY respectively. Finally, we note that the last term disappears because $E[\varepsilon] = 0$ by assumption.

If we can determine exact values for the expectation of $[\Delta X, \Delta Y]$, and subsequently the covariance matrices, we could then recover the true value β directly from the robust estimator. While this is generally not feasible, we can obtain good approximations for these expectations using the labeled data. We use a similar approach to (Qiao & Huang, 2021) and estimate $\Sigma_{\tilde{X}, \tilde{X}} = E[\tilde{X}^T \tilde{X}]$, $\Sigma_{\tilde{X}, \tilde{X}} = E[\tilde{X}^T \Delta X]$, and $\Sigma_{\tilde{X}, \tilde{X}} = E[\tilde{X}^T \Delta Y]$ using the labeled portion of the data. Specifically, let $s_{\tilde{X}, \tilde{X}}$ be the empirical variance-covariance matrix, $s_{\tilde{X}, \Delta X}$ be the empirical covariance matrix in the labeled independent variable data, and $s_{\tilde{X}, \Delta Y}$ be the empirical covariance matrix between the features and dependent variable errors. Using cross validation, these matrices are unbiased estimators of the true values. Then, we can construct a *consistent* estimator of β by applying an explicit correction term. We summarize these observations in the following theorem.

Theorem 1: Consistency of Robust Estimator.

Let the labeled data be i.i.d. samples drawn from the same population as the complete data set. Further, assume that (a) the underlying data follows the standard OLS assumptions, (b) the machine learning errors are uncorrelated with ε , and (c) there is a non-empty subset of data where the ground truth values for X and Y are known. Then, for $\hat{\beta}^R$ defined as Equation (3) with $\alpha \neq 0$, the limit $\lim_{n \rightarrow \infty} \Pr(|\hat{\beta}^R - \hat{\beta}^{OLS}| > \epsilon) = 0$ for any $\epsilon > 0$. Further, the estimator $\beta^ = \left[I - s_{\tilde{X}, \tilde{X}}^{-1} s_{\tilde{X}, \Delta X} \right]^{-1} \left(\hat{\beta}^R - s_{\tilde{X}, \tilde{X}}^{-1} \Sigma_{\tilde{X}, \Delta Y} \right)$ is a consistent estimator of β .*

The proof of the theorem follows from the limit definition of the robust estimator Equation (6) and the unbiasedness of the cross validation estimates. We make a few additional observations regarding the estimator in Theorem 1. First, when there is zero covariance between the observed data and the errors, the estimator is equivalent to the true coefficient vector β in the limit. In other words, if the machine learning errors are truly random noise, the robust solution will be the same as the underlying true solution. Second, if the dependent variable is assumed to be measured without error (i.e., Y is not estimated with machine learning) or if we assume $E[\tilde{X}^T \Delta Y] = 0$, we can still recover a consistent estimator.

3.4. Variance Analysis of the Robust Estimator

Aside from the consistency of the estimator $\hat{\beta}^R$, it is also important to consider the variance of the estimator, as the standard errors of coefficients are needed to conduct hypothesis testing. In particular, we are interested in the *relative efficiency* of the robust estimator, compared to the naive OLS approach. To measure the relative efficiency, we determine whether or not the variance of $\hat{\beta}^R$ is less than that of the naive OLS estimator $\hat{\beta}^{OLS}$ computed on the observed data as the sample size increases. We begin with the relation $\hat{\beta}^R = \left[\alpha I + \tilde{X}^T \tilde{X} \right]^{-1} \tilde{X}^T \tilde{X} \hat{\beta}^{OLS}$. Using the fact that $Var(Ax) = AVar(x)A^T$ for any matrix A and random variable x , we can now derive the variance of $\hat{\beta}^R$ and compare it to the variance of $\hat{\beta}^{OLS}$.

Because these are covariance matrices, we are specifically interested in whether $Var(\hat{\beta}^{OLS}) \succeq \hat{\beta}^R$; if so, then the variance of the OLS coefficients will be greater than those of the robust model. We now take the difference of the variances between the two estimators: $Var(\hat{\beta}^{OLS}) - Var(\hat{\beta}^R) = \sigma^2[\alpha I + \tilde{X}^T \tilde{X}]^{-1} [2\alpha I + \alpha^2(\tilde{X}^T \tilde{X})^{-1}][\alpha I + \tilde{X}^T \tilde{X}]^{-1}$. We know that $\sigma^2 > 0$ and $\alpha > 0$ by assumption. Further, the diagonal values of $\tilde{X}^T \tilde{X}$ are non-negative. Thus, the difference in variance matrices must have all non-negative diagonal values, implying it is a positive semi-definite matrix. More formally, we conclude that $Var(\hat{\beta}^{OLS}) - Var(\hat{\beta}^R) \succeq 0$. We summarize this finding in Theorem 2.

Theorem 2: Efficiency of Robust Estimator.

Let the labeled data be i.i.d. samples drawn from the same population as the complete data set. Further, let $\hat{\beta}^{OLS}$ be the standard OLS estimate fit to the observed data, and $\hat{\beta}^R$ be the robust estimator also fit to the observed data with $\alpha > 0$. Then, the inequality $Var(\hat{\beta}^{OLS}) \succeq Var(\hat{\beta}^R)$ holds.

Based on Theorem 2, we know that the robust solution is at least as efficient as the standard OLS solution. However, when conducting statistical inference it is also relevant to know which estimator has greater *statistical power*. In other words, it is often desirable to have an estimator that correctly rejects the null hypothesis of no effect, even if it is slightly more biased than an alternative estimator. To determine which estimator has greater power, we formulate that test statistic for each. Define the test statistic for coefficient j as $t_j^* = \hat{\beta}_j / \sqrt{Var(\hat{\beta})_{jj}}$. Given this definition, we

advance the following corollary to Theorem 2:

Corollary: *For any given variable j , the robust solution has greater statistical power than the OLS solution, i.e., $t_j^*(\hat{\beta}^R) \geq t_j^*(\hat{\beta}^{OLS})$. Equality holds for simple linear regression, or if all non-diagonal entries of $\text{Var}(\hat{\beta}^{OLS})$ are 0. Otherwise, the inequality is strict.*

The proof for the corollary follows from the definitions of the OLS and robust estimators, as well as their covariance matrices. To summarize Sections 3.3 and 3.4, we find that robust optimization can find a solution to the regression problem that, while more biased than OLS, has more statistical power, i.e., is more likely to detect a statistically significant effect when variables are subject to error. This advantage is due to the relatively smaller standard errors associated with the robust coefficients. Further, the OLS advantage in terms of bias will tend to disappear as the sample size becomes larger. Finally, if researchers want to obtain a consistent estimate of the underlying parameters β , e.g., if the magnitude of an effect is more important than significance, then an explicit correction can be applied to the robust estimator. As the sample size increases, the corrected robust solution will converge to an unbiased estimate.

4. Simulation Experiments

We conduct a series of simulation experiments to validate our theoretical results with respect to the bias and standard errors of the estimated coefficients. We generate synthetic data in order to control the ground truth values for the data used in inference. Because our method is built to incorporate errors, we will also randomly generate perturbations in the values of the statistics. For each test, we will generate $N = 10,000$ observations and designate $n = 1,000$ as labeled, with the remaining observations considered unlabeled. The labeled data will be used to tune the robust tolerance parameter and obtain estimates of the covariance matrices. Then, we will fit the robust model to the unlabeled data and compare the resulting coefficient estimates to the ground truth. For sake of comparison, we will also fit OLS regression directly to the unlabeled data.

To test the general performance of our proposed estimator, we first generated three features X_1 , X_2 , and X_3 by drawing 10,000 observations from the multivariate normal distribution. The mean of each variable was 0, and the covariance matrix was such that $\sigma_{ii} = 1$ for $i = 1, 2, 3$ and $\sigma_{ij} = 0.1$ for all $i \neq j$; this allows for some minor collinearity

among the features. The ground truth parameters were $\beta_1^{TRUE} = \beta_2^{TRUE} = \beta_3^{TRUE} = 1$. The dependent variable was then computed as $Y = \beta_1^{TRUE} X_1 + \beta_2^{TRUE} X_2 + \beta_3^{TRUE} X_3 + \varepsilon$, where ε was also drawn from the standard normal distribution. All variables are mean centered to ensure consistency. These values of Y and X are considered the “true” or nominal values of the data, assuming no measurement error. To create the measurement error, we generate vectors ΔX_1 , ΔX_2 , ΔX_3 , and ΔY by drawing 10,000 observations from the uniform distribution with a range of -1 to 1 . Using these errors, we then created $\tilde{Y} = Y + \Delta Y$ and $\tilde{X} = X + \Delta X$.

Once the data were generated, we designated 1,000 observations as labeled data, i.e., we assumed that we had access to both the nominal values X and Y and the errors ΔX and ΔY only for those 1,000 observations. To obtain estimates for the covariance matrices $s_{\tilde{X}, \tilde{X}}$, $s_{\tilde{X}, \Delta X}$, and $s_{\tilde{X}, \Delta Y}$ we used 10-fold cross validation on the labeled data.

We next fit the robust model as defined in Equation (3) using the 9,000 “unlabeled” observations \tilde{X} and \tilde{Y} and the tolerance parameter $\hat{\rho}$. This estimate is designated $\hat{\beta}^R$ and has standard errors $SE(\hat{\beta}^R)$. We also fit the typical regression estimator to this unlabeled data, and designate the coefficients as $\hat{\beta}^{OLS}$ with standard errors $SE(\hat{\beta}^{OLS})$. Finally, we also applied the correction as detailed in Theorem 1; this corrected term is denoted β^* . We measured performance in three ways. First, we determined the overall bias as $bias.total(\hat{\beta}) = \|\hat{\beta} - \beta^{TRUE}\| / \|\beta^{TRUE}\|$ for all three estimates; we divide by the norm of the ground truth coefficient to ensure consistency across experiments. Second, we calculated the bias in each parameter β_k similarly: $bias.param(\hat{\beta}_k) = (\beta_k - \hat{\beta}_k^{TRUE}) / |\beta_k^{TRUE}|$. Finally, we compared the standard errors of the coefficient estimates for both the robust and OLS models, as well as the corresponding test statistics. This entire procedure was repeated 100 times and the results were averaged. Our findings are illustrated in Figure 1.

We find that on average, the robust solution yields a slightly more biased parameter vector (0.226) compared to OLS (0.218) fit to the testing data. Put another way, the overall magnitude of discrepancy between the robust solution and the true value was less than 22.6% of the true value, while the normal regression solution was off by 21.8%, a difference of less than 1% between models. However, the corrected model performed much better, with an average overall bias of 0.038, or only 3.8%. At the individual coefficient level, the robust model produced average standardized biases of -0.226 across for the three coefficients, while the OLS model produced average standardized biases of -0.218

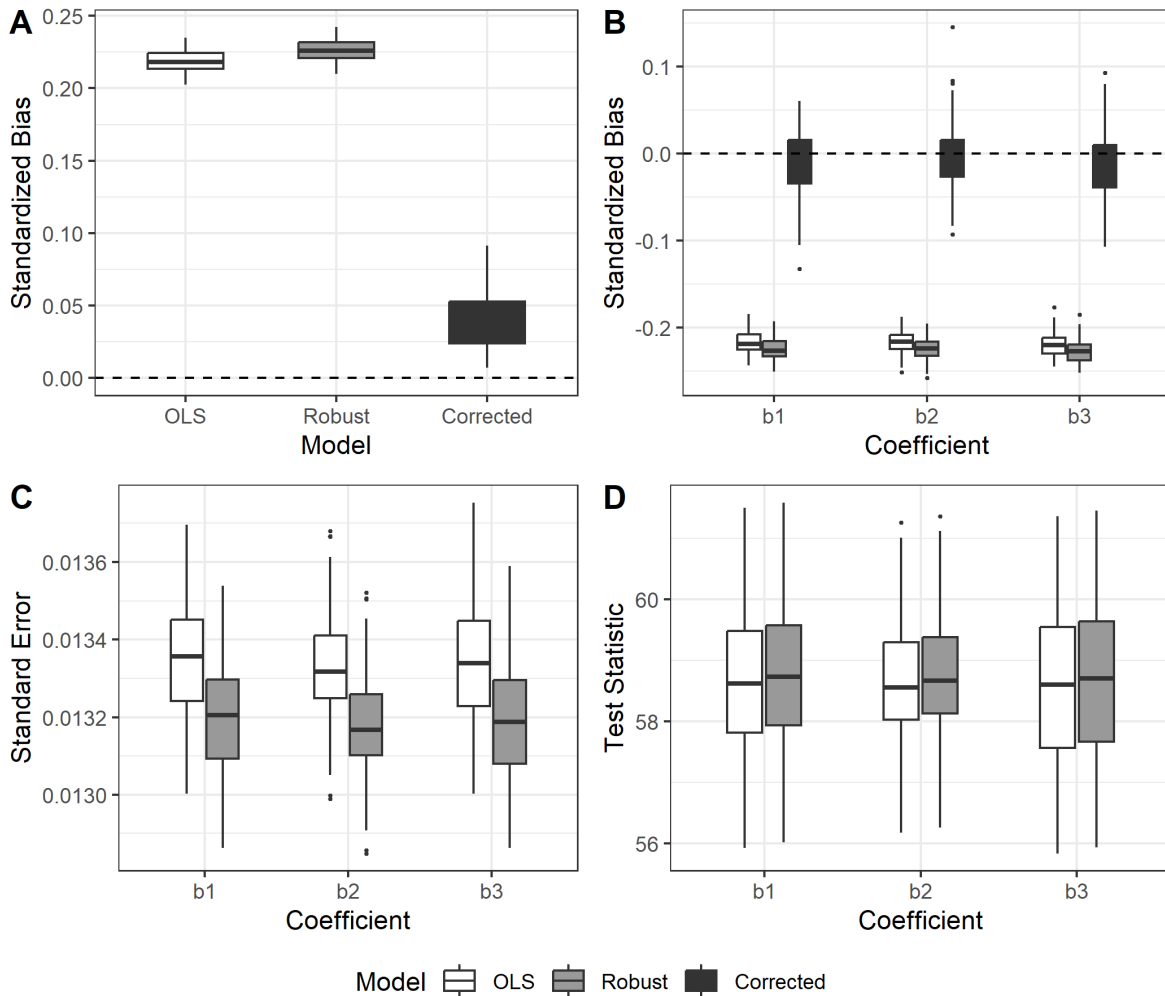


Figure 1. Simulation Results for Experiment

Notes. Results of 100 simulations. (A) shows the average standardized bias for the entire coefficient vector, normalized by the ground truth vector. (B) shows the distribution of bias values by each variable. Values are standardized by dividing by the absolute value of the ground truth. (C) shows the distribution of standard error values for each variable. (D) shows the test statistic (coefficient / std. error).

for the same coefficients. The negative values indicate underestimation of the true values, or an attenuation bias as expected. The corrected model produced an average bias of only -0.012 across all three coefficients. Turning to Figure 1C, we find that the standard errors for the robust coefficients are consistently smaller than those of the OLS coefficients. Further, in Figure 1D we see that the test statistics (e.g., t-value) for the coefficients are higher for the robust model than the OLS model. This result is consistent with our theoretical analysis.

Overall, Figure 1A-B demonstrate that 1) the robust model is nearly as effective as OLS at producing unbiased estimates, and 2) the correction term eliminates nearly all of the bias in the robust estimator. Additionally, Figure 1C-D indicate that the uncorrected

robust model will produce larger test statistics - and subsequently higher statistical power - compared to OLS. Thus, we conclude that using the robust model can improve hypothesis testing, while the corrected solution is highly accurate in terms of the ground truth magnitudes.

5. Discussion

In this study, we address the issue of bias introduced into regression models when one or more independent variables are generated by supervised machine learning. This two-step estimation framework – one step of SML predictions followed by econometric analysis – is an increasingly viable means of testing hypotheses and

making decisions in a variety of business settings. However, a drawback of this analytical approach is the potential for errors to be introduced by the SML methods. Indeed, no algorithm is perfect, and some degree of error is unavoidable. Our paper contributes to the burgeoning literature on correcting such errors (Qiao & Huang, 2021; Yang et al., 2018, 2022) by introducing an alternative method, robust optimization. Robust optimization is an established technique in the field of operations research (Bertsimas et al., 2011), but recently it is gaining some traction in inferential statistics (Bertsimas et al., 2019; Bertsimas & Nohadani, 2019; Chen & Paschalidis, 2018; Schechter, Nohadani, & Contractor, 2022). We extend this work by deriving a robust optimization solution to the least squares regression problem and applying it to the two-step econometric setting.

We found that the uncorrected robust optimization solution on average produced slightly more biased coefficients but smaller standard errors, leading to larger corresponding test statistics. We also found that the proposed corrected estimator β^* is much less biased than either the uncorrected robust or OLS solutions. These findings suggest that we can use a combination of β^R and β^* to both conduct more accurate hypothesis testing and identify point estimators with precision. The robust optimization approach however was not uniformly superior in our experiments. When the SML-induced errors in the unlabeled data were much smaller than in the labeled dataset, the robust model tended to overcorrect, thus leading to greater bias. Similarly, for very large errors, the relative advantage of the robust model is diminished. These results suggest that proper measurement of the SML error is key to maximizing the benefit of the robust model. This can be achieved if the labeled dataset is sufficiently large; otherwise, researchers could use techniques such as bootstrapping to achieve more reliable estimates.

The current research extends prior work in a number of ways. First, this study extends the method of moments approach (Qiao & Huang, 2021) by examining continuous variables, rather than binary outcomes. This change makes the proposed methodology more widely applicable. Further, we go beyond prior work by directly assessing the ability of our method to identify statistically significant effects, which is arguably more important than finding unbiased estimates when conducting inference. Second, we study a more generalized form of hybrid design where a flexible number of dependent and independent variables are measured by SML methods with heteroscedastic measurement errors, as compared to prior work examining a single corrupted variable (Yang et al.,

2018). Finally, our work differs from other studies on robust optimization and linear regression (Xu et al., 2008; Bertsimas & Copenhaver, 2018) in that we prove the consistency and efficiency of the estimator with respect to the true underlying parameters. By contrast, other studies have emphasized the equivalence (or lack thereof) between robust optimization and regularization, as well as the predictive accuracy of the model.

6. Conclusion

The practice of leveraging machine learning to operationalize constructs from large-scale unstructured data is becoming increasingly common in practice and in our research community. As a result, variables generated through machine learning are now leveraged in more traditional regression models to make inferences and test theories. However, algorithms are inevitably imperfect, and as a result measurement error is introduced into the regression models, leading to potentially biased estimates or incorrect hypothesis testing. In this paper we propose using robust optimization to reduce the negative impact of these errors and produce less biased coefficient estimates. We derive a novel robust optimization solution to the least squares regression problem with adversarial errors. Through experiments on simulated data we demonstrate the effectiveness of our approach and identify conditions where robust optimization will likely outperform other methods, such as ordinary least squares.

References

- Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization* (Vol. 28). Princeton university press.
- Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, 53(3), 464–501.
- Bertsimas, D., & Copenhaver, M. S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3), 931–942.
- Bertsimas, D., Dunn, J., Pawlowski, C., & Zhuo, Y. D. (2019). Robust classification. *INFORMS Journal on Optimization*, 1(1), 2–34.
- Bertsimas, D., Gupta, V., & Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167, 235–292.
- Bertsimas, D., & Nohadani, O. (2019). Robust maximum likelihood estimation. *INFORMS Journal on Computing*, 31(3), 445–458.

- Bound, J., Brown, C., Duncan, G. J., & Rodgers, W. L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3), 345–368.
- Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. CRC press.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.
- Carroll, R. J., & Stefanski, L. A. (1994). Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statistics in Medicine*, 13(12), 1265–1282.
- Chandrasekaran, S., Golub, G., Gu, M., & Sayed, A. H. (1998). Parameter estimation in the presence of bounded data uncertainties. *SIAM Journal on Matrix Analysis and Applications*, 19(1), 235–252.
- Chen, R., & Paschalidis, I. C. (2018). A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13).
- Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314–1328.
- Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3), 595–612.
- Eldar, Y. C., Ben-Tal, A., & Nemirovski, A. (2004). Robust mean-squared error estimation in the presence of model uncertainties. *IEEE Transactions on Signal Processing*, 53(1), 168–181.
- El Ghaoui, L., & Le Bret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4), 1035–1064.
- Fong, C., & Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4), 467–484. doi: 10.1017/pan.2020.38
- Gabrel, V., Murat, C., & Thiele, A. (2014). Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3), 471–483.
- Greene, W. H. (2018). *Econometric analysis*. Pearson Education.
- Gu, B., Konana, P., Rajagopalan, B., & Chen, H.-W. M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information systems research*, 18(1), 68–85.
- Hong, L. J., Huang, Z., & Lam, H. (2021). Learning-based robust optimization: Procedures and statistical guarantees. *Management Science*, 67(6), 3447–3467.
- Kuhn, D., Esfahani, P. M., Nguyen, V. A., & Shafieezadeh-Abadeh, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics* (pp. 130–166). INFORMS.
- Qiao, M., & Huang, K.-W. (2021). Correcting misclassification bias in regression models with variables generated via data mining. *Information Systems Research*, 32(2), 462–480.
- Schechter, A., Nohadani, O., & Contractor, N. (2022). A robust inference method for decision making in networks. *Management Information Systems Quarterly*, 46(2), 713–738.
- Stefanski, L. A. (2000). Measurement error models. *Journal of the American Statistical Association*, 95(452), 1353–1358.
- Tirunillai, S., & Tellis, G. J. (2012). Does chatter really matter? dynamics of user-generated content and stock performance. *Marketing Science*, 31(2), 198–215.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.
- Xu, H., Caramanis, C., & Mannor, S. (2008). Robust regression and lasso. *Advances in neural information processing systems*, 21.
- Xu, H., Caramanis, C., & Mannor, S. (2012). A distributional interpretation of robust optimization. *Mathematics of Operations Research*, 37(1), 95–110.
- Yang, M., Adomavicius, G., Burtch, G., & Ren, Y. (2018). Mind the gap: Accounting for measurement error and misclassification in variables generated via data mining. *Information Systems Research*, 29(1), 4–24.
- Yang, M., McFowland III, E., Burtch, G., & Adomavicius, G. (2022). Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*.
- Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2022). What makes a good image? airbnb demand analytics leveraging interpretable image features. *Management Science*, 68(8), 5644–5666.