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

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–14
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

A Robust Optimization Approach to Reliable Statistical Inference with Variables Generated by Machine Learning

Aaron Schecter,[a,*] **Weifeng Li**[a]

[a] Department of Management Information Systems, University of Georgia, Athens, Georgia 30602
*Corresponding author
Contact: aschecter@uga.edu, https://orcid.org/0000-0002-3186-7788 (AS); weifeng.li@uga.edu, https://orcid.org/0000-0002-2105-3596 (WL)

Received: July 7, 2023
Revised: June 29, 2024; March 17, 2025; September 22, 2025; November 7, 2025
Accepted: November 15, 2025
Published Online in Articles in Advance: December 24, 2025

https://doi.org/10.1287/isre.2023.0340

Abstract. Leveraging supervised machine learning (SML) algorithms to operationalize constructs from unstructured data such as text or images is becoming increasingly common in practice and research. As a result, variables generated through SML are now used in traditional regression models to test hypotheses. However, algorithms are imperfect, and thus, the variables produced by SML have measurement errors relative to the underlying construct, potentially leading to biased coefficients and faulty inference. In this paper, we propose using robust optimization to reduce the negative impact of these errors and enable more accurate hypothesis testing. We leverage robust optimization techniques to fit a linear regression model in the presence of measurement errors of different magnitudes. We theoretically demonstrate the bias, variance, and hypothesis testing performance of the robust approach and propose a correction term to effectively reduce bias. Through experiments on simulated data sets and a case study of Amazon reviews, we demonstrate the effectiveness of our approach and identify conditions in which robust optimization likely outperforms other methods. We make recommendations for researchers leveraging machine learning–generated variables in causal inference.

History: Olivia Liu Sheng, Senior Editor; Huimin Zhao, Associate Editor.
Funding: The authors acknowledge support from the Terry Sanford Award from the University of Georgia.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2023.0340.

Keywords: robust optimization • machine learning • regression • measurement error correction • bias

1. Introduction

The information systems (IS) research community was among the first to enrich causal inference by operationalizing constructs of theoretical or practical importance from unstructured data (e.g., text and images) to test hypotheses. Increasingly, however, this practice is becoming ubiquitous—in fields such as biostatistics (risk scores) to environmental science (exposure indices) to economics/marketing (credit/propensity scores)—and researchers increasingly regress on proxy covariates produced by machine learning models. To incorporate such constructs into causal inference, past hybrid studies (Gu et al. 2007, Tirunillai and Tellis 2012, Zhang et al. 2022) establish a two-step estimation framework: the first step uses supervised machine learning (SML) methods to operationalize measures of constructs from unstructured data, such as text and images, and the second step includes these SML-based measures in a regression model (Qiao and Huang 2021). This two-step estimation framework broadens the scope of empirical research by allowing researchers to study phenomena and test theories in new and understudied research contexts. Two-step estimation with SML-generated variables is now routine in empirical research across text and image applications (e.g., Chan and Wang 2018, Lee et al. 2018).

Nonetheless, the variables generated by SML have measurement errors originating from SML methods' imperfect estimates of the target constructs (Yang et al. 2018, Qiao and Huang 2021). For instance, when using text classifiers 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 because of sarcasm. Measurement errors in the first step can attenuate or amplify the coefficient estimates of SML-based measures and further distort the estimation of the dependent variable in the second step, leading to biased regression coefficients (Carroll et al. 2006). For dependent variables measured by SML, the additional measurement error could affect the standard errors of the model coefficients. In both cases, measurement errors induced by SML methods compromise the statistical power of the resulting estimator by reducing the magnitude of test statistics, leading to higher false negative rates in hypothesis testing (Meijer et al. 2021). Unfortunately, the effects of measurement errors on test statistics are mostly studied in the simple linear regression context; in a multiple regression model with several (possibly

correlated) variables subject to SML errors, typical test statistics may yield inaccurate hypothesis testing results (Carroll et al. 2006). This problem is akin to the challenge of maximizing the signal-to-noise ratio, in which the true signal (unobserved construct) is partially obscured by the noise (SML errors).

Overall, using variables generated by SML in regression models can lead to both biased coefficients and inaccurate hypothesis testing. In response to the problem of bias, several correction techniques have been proposed, including simulation extrapolation (SIMEX) (Yang et al. 2018) and instrumental variables (IV) (Yang et al. 2022). Despite their performance in the unbiased estimation of coefficients affected by measurement errors, their implications for test statistics and power are less explicit. An alternative correction technique is the method of moments, which achieves consistent coefficient estimates through a function of the original coefficient, the covariance between the SML-based variable and other predictors, and the covariance between other predictors and the measurement error (Qiao and Huang 2021). Whereas this method provides an unbiased solution, it does not account for the influence of measurement errors on the test statistics and the subsequent impact on statistical power and hypothesis testing. In sum, existing techniques focus on the problem of biased coefficients, but the problem of inaccurate hypothesis testing is understudied.

Following the computational design science research paradigm (Abbasi et al. 2024), this paper introduces robust optimization as a general purpose error-correcting estimation technique that improves the accuracy of hypothesis testing on variables—independent, dependent, or both—generated by SML methods. This study provides salient design insights that complement prior research in two ways. First, we show that robust optimization can be used to conduct more accurate hypothesis testing using Wald inference when SML-induced errors are present. In particular, robust optimization amplifies the signal of the effect relative to the noise of measurement errors, resulting in enhanced test statistics. Second, we derive analytical solutions for the coefficients and their standard errors, which enables a thorough analysis of the bias–variance trade-off. In addition, we show that a small amount of labeled data can be used to create a consistent estimator without sacrificing significant variance. These contributions not only inform empirical IS research but also open pathways for more effective inference across disciplines.

2. Background

2.1. Measurement Errors in Regression

Measurement error arises when one or more independent or dependent variables are not measured accurately (Carroll et al. 2006). Errors in variables generated by an SML method are because of inaccuracies in a prediction made by the algorithm (Yang et al. 2018). For example, an SML method analyzing social media posts to determine customers' opinions about a company would potentially make erroneous predictions about posts that were sarcastic or used slang. Here, the method of translating the unstructured data is fundamentally imperfect; thus, the operationalization of the variables may not reflect the underlying construct.

To understand how these errors impact subsequent analyses, there are two types of measurement error models: the classical error model and the nonclassical error model. The classical error model assumes the measurement error to be additive and independent of the true measure and the residuals in the second step estimation, whereas the nonclassical model considers the measurement error as nonadditive or correlated with the true measure or residuals (Carroll et al. 2006). SML-induced measurement errors could be either classical or nonclassical (Yang et al. 2018), thus making it difficult to estimate—and, therefore, correct a priori—the direction and magnitude of bias in the resulting parameter estimates.

Measurement errors in an independent variable compromise the two-step estimation framework in two important ways. First, these errors can attenuate or amplify the resulting coefficient estimates, depending on the second step estimation model (Yang et al. 2018). In the case of simple linear regression with classical error, it is known (see Wooldridge 2010) that the effect of measurement error in simple linear regression is to attenuate the estimate of β in the direction of zero. In multiple regression and nonlinear settings, the sign and magnitude of bias can vary (Carroll et al. 2006, Yang et al. 2018).

Second, measurement errors can reduce the magnitude of the corresponding test statistics, which subsequently affects the testing of the hypotheses about the coefficients of erroneous variables as well as the other variables (Meijer et al. 2021). As the coefficient estimates are attenuated or amplified, the test statistics involving these biased coefficient estimates become inherently less reliable. Except for simple linear regression with classical error, most statistical tests could produce misleading results (Carroll et al. 2006). Consequently, the standard hypothesis testing procedure could suffer from a heightened level of false positives (type I errors) and false negatives (type II errors), leading to unreliable statistical inference.

2.2. Correcting Errors in Hybrid Studies

This section reviews the existing error correction methods in hybrid studies. To demonstrate the key ideas of these methods, we first introduce the formulation of a hybrid linear regression as the context: we assume that the dependent variable $\boldsymbol{Y} \in \mathbb{R}^n$ is an n-by-one vector of

continuous real numbers from n observations and that the independent variable $\tilde{\boldsymbol{X}} \in \mathbb{R}^n$ is measured by an SML algorithm. In particular, the SML-generated variable could be represented as the true value plus the SML-induced error $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$, where $\boldsymbol{X}$ is the error-free but unobservable true value of the estimated measure and $\Delta\boldsymbol{X}$ is the SML-induced measurement error. Hybrid studies aiming to estimate the regression model $\boldsymbol{Y} = \tilde{\boldsymbol{X}}\beta + \boldsymbol{\varepsilon}$ typically achieve a biased ordinary least squares (OLS) estimator $\hat{\beta} = \lambda\beta$, where λ is the reliability ratio $\lambda = \sigma_X^2/[\sigma_X^2 + \sigma_{\Delta X}^2]$, 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 proposes several techniques[1] as summarized in Table 1.

One common research goal in hybrid studies for causal inference is hypothesis testing, which leverages test statistics that are often functions of the mean and variance of the corrected estimator. Because bias corrections often increase variance, test statistics can be further weakened (Meijer et al. 2021), motivating methods that can manage the bias–variance trade-off (Carroll et al. 2006). In other words, as the noise of errors obscures the true signal of the construct's effect, attempts at reducing the noise by correcting the measurement error are also potentially associated with a diminished ability to detect the true signal.

However, maximizing the signal-to-noise ratio when correcting measurement errors is particularly important for achieving better hypothesis testing outcomes. Whereas the extant literature extensively focuses on correcting measurement bias, its implications for hypothesis testing are understudied. Moreover, few of the existing error correction methods are specifically designed to improve the signal-to-noise ratio in hypothesis testing. Specialized in maximizing the signal-to-noise ratio, robust optimization provides the methodological foundation that has the potential to address this important research gap.

2.3. Robust Optimization and Regression

Robust optimization is devised to find computationally tractable solutions to problems in which the data input is noisy, incomplete, or erroneous (Bertsimas et al. 2011). Recent advances have also coupled robust optimization with big data (Bertsimas et al. 2018, Hong et al. 2021) and machine learning (Bertsimas et al. 2019, Kuhn et al. 2019). In the case of two-step estimation with SML-generated variables, there is uncertainty because of measurement error in the generated variables. Accordingly, we propose that this branch of optimization, when applied to loss-function minimization (or maximum likelihood estimation) (Bertsimas and Nohadani 2019, Bertsimas et al. 2019, Schecter et al. 2022), lends itself well to the two-step estimation problem with SML variables. We focus on robust optimization for least squares with bounded uncertainty and then analyze the resulting variance, derive Wald tests, and provide a labeled-data correction.

3. Proposed Method

3.1. The Regression Problem with SML-Induced Measurement Error

We consider the most general case in which all variables are subject to some error because of the application of SML methods. The true regression model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a vector of random errors. We assume that the errors $\boldsymbol{\varepsilon}$ satisfy $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and $\mathbb{E}[\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}] = \sigma^2$ with each element of $\varepsilon \approx N(0, \sigma^2)$ as $n \to \infty$.[2] Each feature (i.e., independent variable) in the matrix X and the vector Y are mean-centered, and thus, the intercept can be ignored. The objective is to conduct hypothesis testing on whether each element of β is significantly different from zero. Specifically, the goal is to compare the null hypothesis $H_0 : \beta_j = 0$ and the alternative hypothesis $H_1 : \beta_j \neq 0$ to determine if H_0 should be rejected in favor of H_1 for a given variable j.

If the true values of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are not directly observable, one can use observable SML-generated proxy variables $\tilde{\boldsymbol{Y}}$ and $\tilde{\boldsymbol{X}}$ (Carroll et al. 2006, Fuller 2009, Buonaccorsi 2010) modeled as $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \Delta\boldsymbol{Y}$, where $\Delta\boldsymbol{Y}$ and $\Delta\boldsymbol{X}$ are the discrepancies between the true and observed values. The observed values $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ can be correlated

Table 1. Existing Approaches for Correcting SML-Based Measurement Errors

Technique	Corrected coefficient	Reference
Method of moments	$\boldsymbol{\beta}^{MM} = \frac{\hat{\boldsymbol{\beta}}}{\hat{\lambda}}$, where $\hat{\lambda} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_{\Delta X}^2}$.	Qiao and Huang (2021)
Instrumental variables	$\boldsymbol{\beta}^{IV} = \frac{cov(\boldsymbol{Y}, \boldsymbol{X}^{IV})}{cov(\boldsymbol{X}, \boldsymbol{X}^{IV})}$, where $\boldsymbol{X}^{IV}$ is the instrument correlated with X and uncorrelated with $\Delta\boldsymbol{X}$.	Fong and Tyler (2021), Yang et al. (2022)
Simulation extrapolation	$\boldsymbol{\beta}^{SIM} = \mathbb{E}(\hat{\boldsymbol{\beta}}\|\zeta = -1)$, where $\mathbb{E}(\hat{\boldsymbol{\beta}}\|\zeta)$ is a parametric model estimated on $\{\boldsymbol{\beta}^m, \zeta^m\}_{m=1}^M$, β^m is the OLS coefficient estimate of the regression model $\boldsymbol{Y} = \tilde{\boldsymbol{X}}(\zeta^m)\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and $\tilde{\boldsymbol{X}}(\zeta^m)$ is a simulated sample with measurement error variance $(1 + \zeta_m)\sigma_{\Delta X}^2$.	Yang et al. (2018)

with the measurement errors generated by the respective SML methods though this can be relaxed in practice (Bound et al. 1994).

We can solve the regression problem $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$ with $\hat{\boldsymbol{\beta}}$ estimated using OLS. The coefficient $\hat{\boldsymbol{\beta}}^{OLS}$ minimizes the norm of the residuals $\tilde{\boldsymbol{X}}b - \tilde{\boldsymbol{Y}}$ for the observable data and is equal to $\hat{\boldsymbol{\beta}}^{OLS} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$. Solving this problem results in bias because of the unobservable error (Bound et al. 1994, Wooldridge 2010). Further, the corresponding test statistic is smaller for a coefficient fit to the observed data $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ than for a coefficient fit to the true underlying data $\boldsymbol{X}$ and $\boldsymbol{Y}$. As a consequence, the statistical power of $\hat{\boldsymbol{\beta}}^{OLS}$—the probability that a nonzero effect is correctly identified—is reduced when only $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ are available. The additional noise associated with SML-induced errors makes identification of the true effect (the signal) more difficult, thereby hampering hypothesis testing.

3.2. The Robust Formulation

In response to this issue, we use robust optimization to estimate the regression coefficients. Robust optimization produces estimates that are less sensitive to perturbations in the data, making them more resilient to measurement errors. We borrow the notation from Section 3.1 and assume that our objective is to minimize the residuals with respect to the true data. Because we are only able to observe $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$, we rewrite the residuals as $\|\boldsymbol{X}\boldsymbol{b} - \boldsymbol{Y}\| = \|(\tilde{\boldsymbol{X}} - \Delta\boldsymbol{X})\boldsymbol{b} - (\tilde{\boldsymbol{Y}} - \Delta\boldsymbol{Y})\|$, where the entities $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Y}$ are the measurement errors between the observed and true values. We model the measurement errors as belonging to a finite and bounded uncertainty set of an arbitrary size. The uncertainty set is defined as $U = \{\Delta\boldsymbol{X}, \Delta\boldsymbol{Y} : \|\Delta\boldsymbol{X}, \Delta\boldsymbol{Y}\|_F \le \rho\}$ for the tolerance level $\rho \ge 0$, where $\|\cdot\|_F$ is the Frobenius norm. We assume that $\Delta\boldsymbol{X}, \Delta\boldsymbol{Y}$ are signed errors with finite second moments. Further, these errors can be correlated with the observed data, accounting for both classical and nonclassical measurement errors. The uncertainty set can be thought of as the level of error against which we want to hedge when fitting the regression model. In other words, it functions as a sort of budget for errors that we are allowing to enter into the estimation. When researchers pick a certain level of robustness, they are deciding how much they want to protect against data errors; larger values of ρ are a greater hedge against errors and, therefore, result in more conservative estimates. The resulting coefficient estimates then reflect the best balance between fitting the observed data and guarding against possible measurement errors within the specified uncertainty set.

Given the level of robustness, robust optimization commonly finds a solution that minimizes the objective function for any possible error within the uncertainty set. This solution is equivalent to finding the optimal solution for the greatest possible error and is typically referred to as an adversarial-robust solution. Alternatively, we can find a solution that minimizes the objective function with respect to the expected value of the errors, that is, the average error values $\mu_{\Delta X}, \mu_{\Delta Y}$ drawn from U.[3] This approach is typically referred to as distributionally robust optimization (Delage and Ye 2010, Chen and Paschalidis 2018). For simplicity we refer to the resulting solutions for these two cases as the worst case (WC) and average case (AC) estimators. It can be shown that, for either the WC or AC estimator, the closed-form solution to the least squares problem with error is

$$\hat{\boldsymbol{\beta}}^R = (\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}; \quad \gamma = \begin{cases} \alpha & \text{if WC} \\ \dfrac{n}{nk+2}\rho^2 & \text{if AC,} \end{cases} \tag{1}$$

where n is sample size and k is the number of variables. All subsequent theorems depend only on $\gamma > 0$ and, hence, apply to both WC and AC estimators; the choice between them changes the specific form of γ but not the form of the results. Details of the derivation are provided in the Online Appendix. The value ρ in both cases is the bound on the uncertainty set. For the worst case estimator, γ has the specific value of $\alpha = (\lambda - \tau)/\tau$; λ and τ are the unique solutions of the following second order conic program derived by El Ghaoui and Lebret (1997):

$$\min \lambda, \quad \text{subject to: } \|(\tilde{\boldsymbol{X}}\boldsymbol{b} - \tilde{\boldsymbol{Y}})\| \le \lambda - \tau,\ \rho\|[\boldsymbol{b}^T, 1]\| \le \tau. \tag{2}$$

The worst case and average case estimators can be interpreted as, respectively, the minimum guaranteed effect size assuming the data are subject to errors that are (i) as egregious as possible given the observed data or (ii) consistent in magnitude with the errors in the observed data.

In summary, for a given tolerance level ρ, we can obtain a vector of robust regression coefficients for the adversarial error scenario efficiently by solving the optimization problem in (2) or, for the average error scenario by factoring in the number of observations and features, then using the formula in (1). This formula can also be modified to account for errors in only a subset of variables; we analyze this special case in the Online Appendix. In the following sections, we detail how to calculate standard errors and test statistics, how to adjust this estimator to reduce bias, and how to identify the appropriate tolerance level parameter ρ.

3.3. Theoretical Analysis of the Robust Estimator

Before conducting our theoretical analysis, we establish a set of assumptions that are necessary for the results to hold.

Assumption 1 (Sampling and Moments). *Observations* $(\tilde{x}_i, \tilde{y}_i)$ *are independent and identically distributed (i.i.d.) with* $\mathbb{E}\|\tilde{\boldsymbol{x}}\|^4 < \infty$ *and* $\mathbb{E}[\tilde{y}^2] < \infty$.

Assumption 2 (Exogeneity). The residuals satisfy $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \Delta\boldsymbol{X}, \Delta\boldsymbol{Y}] = 0$, *where* $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \Delta\boldsymbol{Y}$ *and* $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$.

Assumption 3 (Invertibility). *The matrix* $(\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})$ *is invertible.*

Unless stated otherwise, results apply to both worst case and average case estimators; the choice only affects the numerical value of γ. Assumption 2 implies that our method is not meant to address endogeneity beyond the measurement error induced by the SML-generated variables, which is a limitation of our study. Whereas homoscedasticity is not strictly required for our theoretical analysis, we assume that this condition holds though this assumption can be relaxed in cases of heteroscedasticity in which heteroscedastic-consistent (HC) standard errors can be used if additional conditions are met.

Given these assumptions, we can demonstrate the statistical performance of the robust optimization estimates (either WC or AC) by establishing a link between the robust solution $\hat{\boldsymbol{\beta}}^R$ and the true coefficients $\boldsymbol{\beta}$. Based on our definitions, we know that $\tilde{\boldsymbol{Y}} = (\tilde{\boldsymbol{X}} - \Delta\boldsymbol{X})\boldsymbol{\beta} + \Delta\boldsymbol{Y} + \boldsymbol{\varepsilon}$. Plugging in this expression, we can expand the formal definition of $\hat{\boldsymbol{\beta}}^R$ and then derive the theoretical mean and variance of the robust estimator with respect to the ground truth $\boldsymbol{\beta}$. Specifically, the expected value and variance are

$$\mathbb{E}[\hat{\boldsymbol{\beta}}^R] = \left(\frac{\gamma}{n}\boldsymbol{I} + \Sigma_{\tilde{X},\tilde{X}}\right)^{-1}(\Sigma_{\tilde{X},\tilde{X}} - \Sigma_{\tilde{X},\Delta X})\boldsymbol{\beta} + \left(\frac{\gamma}{n}\boldsymbol{I} + \Sigma_{\tilde{X},\tilde{X}}\right)^{-1}\Sigma_{\tilde{X},\Delta Y}; \tag{3}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^R) = \frac{\sigma^2}{n}\left(\frac{\gamma}{n}\boldsymbol{I} + \Sigma_{\tilde{X},\tilde{X}}\right)^{-1}\Sigma_{\tilde{X},\tilde{X}}\left(\frac{\gamma}{n}\boldsymbol{I} + \Sigma_{\tilde{X},\tilde{X}}\right)^{-1}. \tag{4}$$

Here, $\Sigma_{\tilde{X},\tilde{X}} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]$ is the variance–covariance matrix of the observed data $\tilde{\boldsymbol{X}}$ and $\Sigma_{\tilde{X},\Delta X} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\Delta\boldsymbol{X}]$, $\Sigma_{\tilde{X},\Delta Y} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\Delta\boldsymbol{Y}]$ are the covariance matrix of the observed data $\tilde{\boldsymbol{X}}$ and the SML-generated errors $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Y}$, respectively. We now proceed to analyze the theoretical attributes of the estimator.

3.3.1. Variance Analysis of the Robust Estimator.

We first consider the variance of the estimated coefficients. In particular, we are interested in the relative variance of the robust estimator compared with the naive OLS approach. We begin with the relation $\hat{\boldsymbol{\beta}}^R = (\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{OLS}$. Using the fact that $\mathrm{Var}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{x})\boldsymbol{A}^T$ holds for any matrix $\boldsymbol{A}$ and random variable $\boldsymbol{x}$ as well as assuming homoscedasticity, we can now reexpress the variance of $\hat{\boldsymbol{\beta}}^R$ as $\mathrm{Var}(\hat{\boldsymbol{\beta}}^R) = [(\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]\mathrm{Var}(\hat{\boldsymbol{\beta}}^{OLS})[(\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]^T$. The difference between the two estimators is equal to $\mathrm{Var}(\hat{\boldsymbol{\beta}}^{OLS}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}^R) = \sigma^2[\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]^{-1}[2\gamma\boldsymbol{I} + \gamma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}][\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]^{-1}$. Given this expression, we propose Theorem 1.

Theorem 1 (Relative Variance of Robust Estimator). *Under Assumptions 1–3 and homoscedasticity, that is,* $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega} = \sigma^2\boldsymbol{I}$, *then* $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})_j \geq \mathrm{Var}(\hat{\boldsymbol{\beta}}_R)_j$ *for all j, equivalently,* $\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_R) \geq 0$.

Remark 1. In the absence of homoscedasticity, we can replace variances by their HC analogs. Let $\boldsymbol{Z} = \tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}$ and $\boldsymbol{S} = \tilde{\boldsymbol{X}}^\top\boldsymbol{\Omega}\tilde{\boldsymbol{X}}$ with $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega}$; the HC variance can then be expressed as $\boldsymbol{Z}^{-1}\boldsymbol{S}\boldsymbol{Z}^{-1}$. If $\boldsymbol{Z}$ and $\boldsymbol{S}$ commute or are simultaneously diagonalizable (i.e., $\boldsymbol{Z}\boldsymbol{S} = \boldsymbol{S}\boldsymbol{Z}$), then $\mathrm{Var}_{HC}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}})_j \geq \mathrm{Var}_{HC}(\hat{\boldsymbol{\beta}}_R)_j$ for all j, equivalently, $\mathrm{Var}_{HC}(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}) - \mathrm{Var}_{HC}(\hat{\boldsymbol{\beta}}_R) \geq 0$. The commuting condition holds, for example, under homoscedasticity ($\boldsymbol{S} \propto \boldsymbol{Z}$) or when $\boldsymbol{\Omega}$ is a function of the projection $\boldsymbol{P} = \tilde{\boldsymbol{X}}(\boldsymbol{Z})^{-1}\tilde{\boldsymbol{X}}^\top$ (e.g., $\boldsymbol{\Omega} = \sigma^2(\boldsymbol{I} + \tau\boldsymbol{P})$). It typically fails under arbitrary observation-specific, clustered, serial, or spatial dependence misaligned with the design; in that case, bootstrapping can be used to verify the claim empirically.

3.3.2. Hypothesis Testing with the Robust Estimator.

Based on Theorem 1, we know that the robust optimization coefficients have standard errors that are bounded above by the standard errors of the OLS solution. However, our objective is not simply to produce smaller standard errors but to conduct more accurate hypothesis testing. Before comparing the robust estimator to OLS, we first need to establish that the robust solution is asymptotically normal and, therefore, usable for statistical inference. We, therefore, present the following theorem.

Theorem 2 (Asymptotic Normality and Wald Inference). *Under Assumptions 1–3,*

$$\sqrt{n}\,(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}_R^\star) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Psi}),$$

where $\boldsymbol{\beta}_R^\star = (\gamma\boldsymbol{I} + \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\mathbb{E}[\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{Y}}]$ *and* $\boldsymbol{\Psi}$ *is the covariance obtained from the joint central limit theorem of* $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}})$, *that is, the population limits of these covariances. A consistent estimator* $\hat{\boldsymbol{\Psi}}$ *is obtained by the previously defined formula (HC-robust if needed). Consequently, for any term j, the Wald statistic*

$$T_R(j) = \frac{(\hat{\beta}_R)_j - (\beta_0)_j}{\sqrt{\hat{\Psi}_{jj}}}$$

is asymptotically standard normal under $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

For proof of the theorem, please see the Online Appendix.

We next propose the following two corollaries.

Corollary 1 (Asymptotic Power Comparison Under Variance Ordering). *Maintain Assumptions 1–3. If, in addition, the variance ordering of Theorem 1 holds (either homoscedastic or heteroscedastic with $\boldsymbol{ZS} = \boldsymbol{SZ}$), then for any coefficient j, the noncentrality parameter of $T_R(j)$—for the WC or AC estimator—is greater than or equal to that of the OLS-based Wald statistic. Hence, the robust test is asymptotically weakly more powerful for such alternatives.*

Corollary 2 (Asymptotic Limits to Test Statistics). *Maintain Assumptions 1–3. Fix a coefficient j. Let $T_R(j)$ be the Wald statistic based on the WC or AC robust estimator using $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ and let $T_{\text{oracle}}(j)$ be the Wald statistic from the linear model fit to error-free $(\boldsymbol{X}, \boldsymbol{Y})$. Then, the asymptotic noncentrality parameter $(T_R(j))$ is less than or equal to $(T_{\text{oracle}}(j))$. Hence, the robust test is asymptotically weakly less powerful than the oracle test that uses error-free covariates.*

The proofs of both corollaries are in the Online Appendix. Finally, we integrate these corollaries to propose Theorem 3.

Theorem 3 (Robust Hypothesis Testing). *Maintain Assumptions 1–3 and assume that Corollaries 1 and 2 hold. Suppose we are studying the theoretical model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ and want to test the hypothesis that $\beta_j \neq 0$ for some variable j. Now, suppose we have observable data $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \Delta\boldsymbol{Y}$ containing errors caused by machine learning algorithms. Let $\hat{\boldsymbol{\beta}}^{OLS}$ be the OLS estimator fit to $\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}$ and $\overline{\boldsymbol{\beta}}$ be the OLS estimator fit to $\boldsymbol{X}, \boldsymbol{Y}$, that is, the oracle estimator. Then, the robust estimator (WC or AC with $\gamma > 0$) $\hat{\boldsymbol{\beta}}^R$ fit to $\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}$ has the following asymptotic properties: when H_0 is $\beta_j = 0$, it correctly rejects H_0 in favor of $H_1 : \beta_j \neq 0$ as or more often than $\hat{\boldsymbol{\beta}}^{OLS}$. Further, it correctly fails to reject H_0 as or more often than $\overline{\boldsymbol{\beta}}$.*

Theorem 3 has important implications for researchers using SML-generated variables. The ability to conduct hypothesis testing with regression coefficients accurately, even in the presence of SML-induced errors, should lead to more reliable outcomes. Indeed, as Theorem 3 implies, researchers can be confident that they are committing fewer type II errors (i.e., have greater power; Corollary 1) and also not overcommitting type I errors (Corollary 2). This finding underscores the purpose of robust optimization, to amplify the signal amid the noise, but only if there is a signal at all.

3.3.3. Correcting the Bias of the Robust Estimator.

We use a similar approach to Qiao and Huang (2021) and estimate $\Sigma_{\tilde{X},\tilde{X}} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}]$, $\Sigma_{\tilde{X},\Delta X} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\Delta\boldsymbol{X}]$, and $\Sigma_{\tilde{X},\Delta Y} = \mathbb{E}[\tilde{\boldsymbol{X}}^T\Delta\boldsymbol{Y}]$ 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. Given these estimated covariance matrices and our knowledge of the tolerance level ρ, the robust estimator differs from β by a multiplicative factor that we can directly calculate. Thus, we can construct a consistent estimator of β by applying a correction term, similar to the method-of-moments approach. We summarize these observations in Theorem 4.

Theorem 4 (Consistency of Bias-Corrected Robust Estimator). *Let the unlabeled sample size be n_u and the labeled sample size be n_ℓ. Maintain Assumptions 1–3 and assume that the labeled-sample moment estimators $\hat{s}_{\tilde{X},\Delta X}$ and $\hat{s}_{\tilde{X},\Delta Y}$ are consistent for their population counterparts. Define*

$$\boldsymbol{A} \equiv \frac{\gamma}{n_u}\boldsymbol{I} + s_{\tilde{X},\tilde{X}}, \quad \hat{\boldsymbol{V}} \equiv \boldsymbol{I} - \boldsymbol{A}^{-1}\left(\frac{\gamma}{n_u}\boldsymbol{I} + \hat{s}_{\tilde{X},\Delta X}\right),$$
$$\hat{\boldsymbol{W}} \equiv \boldsymbol{A}^{-1}\hat{s}_{\tilde{X},\Delta Y},$$

and let $\hat{\boldsymbol{\beta}}^R = \boldsymbol{A}^{-1}s_{\tilde{X},\tilde{Y}}$ denote the robust estimator fit on the unlabeled data. If $\hat{\boldsymbol{V}}$ is positive definite (and, thus, invertible) with probability tending to one, then the bias-corrected estimator

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{V}}^{-1}(\hat{\boldsymbol{\beta}}^R - \hat{\boldsymbol{W}})$$

is consistent for $\boldsymbol{\beta}$ as $n_u, n_\ell \to \infty$. This holds for either the WC or AC estimator with $\gamma > 0$.

Remark 2 (Variance). If $\boldsymbol{V}$ and $\boldsymbol{W}$ are treated as population quantities (or if the labeled sample is large enough that $\sqrt{n_u}/\sqrt{n_\ell} \ll 1$ so first stage noise is negligible), then $\text{Var}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{V}^{-1}\,\text{Var}(\hat{\boldsymbol{\beta}}^R)\,\boldsymbol{V}^{-1}$.

The proof of the theorem follows from the limit definition of the robust estimator (3) and the assumed unbiasedness of the cross-validation estimates. Importantly, because the value of ρ is known exactly, the correction term should effectively counteract the bias regardless of the level of robustness.

To summarize, we find that robust optimization can identify a solution to the regression problem that, whereas more biased than OLS, has more statistical power, that is, is more likely to detect a statistically significant effect when variables are subject to error. This advantage is because of the relatively smaller standard errors associated with the robust coefficients. The robust estimator also does not overstate the presence of effects relative to the ground truth. Finally, if researchers want to obtain a consistent estimate of the underlying parameters $\boldsymbol{\beta}$, for example, if the magnitude of an effect is important, then an explicit correction can be applied to the robust estimator.

3.4. Selecting the Tolerance Level ρ

In practice, it is not always straightforward to determine the best value of ρ as the optimal value can be

dependent on the analytical goals, for example, accurate predictions or unbiased coefficients. One option is to find a ρ^* that maximizes the probability of covering out-of-sample errors, and this is referred to as chance-constrained optimization (Bertsimas et al. 2021, Hong et al. 2021). This choice of tolerance leads to the WC estimator being potentially very conservative but unlikely to overstate the hypothesized effects. The AC estimator could then also be calculated using the same tolerance to identify a somewhat less conservative estimate. The value of ρ^* that achieves the desired coverage can be approximated empirically using bootstrap resampling, which we discuss in the Online Appendix.

The choice of ρ^* has implications for the bias–variance trade-off; specifically, a small choice of ρ^* generally produces coefficients that are less biased but with larger standard errors. Researchers may want to explore the relationship between different values of ρ^*, the coefficient estimates, and the corresponding hypothesis testing results. We recommend fitting the robust model with incremental values of ρ^* and plotting the resulting estimates against the tolerance level. This procedure shows how the underlying effect size may differ as greater levels of uncertainty are introduced.

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 covariates $\boldsymbol{X} = (X_1, X_2, X_3)^\top \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.1$ for $i \neq j$, fix the true parameter $\boldsymbol{\beta} = (1,1,1)^\top$, and draw $\varepsilon \sim \mathcal{N}(0,1)$ independently of $(\boldsymbol{X}, \Delta\boldsymbol{X}, \Delta\boldsymbol{Y})$ so that $\boldsymbol{Y} = \boldsymbol{X}^\top\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}, \Delta\boldsymbol{X}, \Delta\boldsymbol{Y}] = 0$. Measurement errors are signed: each coordinate of $\Delta\boldsymbol{X}$ and $\Delta\boldsymbol{Y}$ is i.i.d. Unif$[-1,1]$ (independent across observations and coordinates unless stated otherwise). We form the observed (noisy) data $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \Delta\boldsymbol{Y}$. In each experiment, we draw an unlabeled sample of size 9,000 (only $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ observed) and a labeled sample of size 1,000 (both $(\boldsymbol{X}, \boldsymbol{Y})$ and $(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ observed). Cross-moments needed for the bias–variance correction are estimated from the labeled sample via K-fold cross-validation, yielding $\hat{s}_{\tilde{X},\Delta X}$ and $\hat{s}_{\tilde{X},\Delta Y}$; the robustness tolerance ρ is selected by a bootstrap over labeled folds (we use the 95th percentile by default). The robust estimator is then fit on the unlabeled sample. Further details of the general simulation procedure can be found in the Online Appendix. We report bias, standard errors, approximate t-statistics, and root mean squared error (RMSE); the full simulation protocol is in the appendix. We repeat each design 100 times and report means.

4.1. Baseline Results

To test the general performance of our proposed estimator, we first report the average results for the robust model, OLS, corrected robust model, and the unbiased (true) model (OLS fit to data with no error). Our findings across the trials are summarized in Figure 1.

We find that, on average, the robust estimator (worst or average case) yielded a more biased estimate of the coefficients compared with OLS fit to the testing data although the AC solution was significantly less biased than the WC solution. The estimates were all under the ground truth value of one, indicating an attenuation effect as expected. The corrected models performed much better with an average overall estimate of 1.003 or only 0.3% bias. We find that the standard errors for the robust coefficients are consistently smaller than those of the OLS coefficients. Further, we observed that the Wald test statistics for the coefficients were higher for the robust model than the OLS model but less than the Wald statistics for the coefficients fit to the data without error. This result was consistent with our theoretical analysis. Finally, we observed that the corrected robust model has a lower RMSE than the regular robust or the regular OLS model, indicating it provided the best balance between bias and variance.

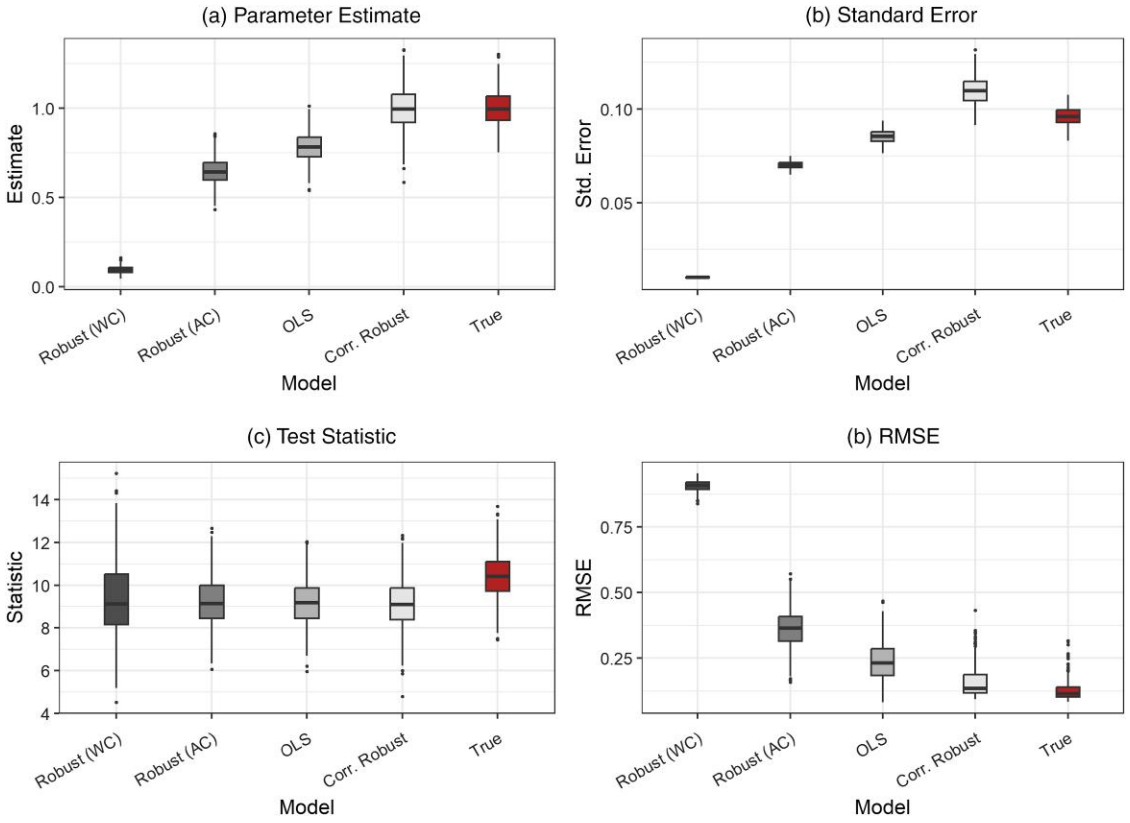
To further assess the performance of the models with respect to hypothesis testing, we estimate the probability that the models reject the null hypothesis of $\beta = 0$ with significance level $\alpha = 0.05$. We simulate data with effect sizes ranging from zero to one and determine the proportion of results with a test statistic over the critical threshold. The results are presented in Table 2.

Note that we do not include effect sizes greater than 0.6 as all models correctly rejected the null 100% of the time. When the true effect was zero, the ground truth model had the highest false positive rate (7.3%), the WC robust estimator had a 5.3% false positive rate, and the AC robust estimator and OLS model had a 5% false positive rate. This indicated that a robust model correctly fails to reject the null hypothesis more often than a model trained to the true data as posited in Theorem 3. For all other nonzero effect sizes, we observed that the true model rejected the null most frequently (as expected). Further, the WC robust estimator correctly rejected the null more frequently than the AC robust estimator for effect sizes greater than 0.1 and rejected the null more frequently than OLS for all effect sizes. Again, this is consistent with Theorem 3. We, therefore, conclude that, when no effect is present, the robust model does not produce excess false positives (type I errors) and, for nonzero effects, commits fewer false negatives (type II errors) than OLS.

4.2. Sensitivity Analysis

We further explore performance of the robust estimator (worst case, average case, and corrected) relative

Figure 1. (Color online) Simulation Results for Baseline Experiment



to OLS in a variety of circumstances. In Table 3, we list each of the tests and the corresponding figures. For brevity, we detail the simulation procedures in the appendix and additional results in the Online Appendix.

To summarize, we find that the baseline results are generally consistent when errors are larger in the unlabeled data set than the labeled data set when the effect sizes are small, when SML errors are small, for different choices of the robust tolerance, and when relaxing the typical OLS assumptions. 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 data set, the robust model tended to overcorrect, thus leading to greater bias. Similarly, for very small errors or very large sample sizes, the relative advantage of the robust model was diminished. We also find that when a control variable is subject to SML error, rather than the main independent variable, hypothesis testing for the focal effect was accurate. The robust estimator was also effective for testing the difference between two coefficients but did not correctly determine if a coefficient was different than a fixed scalar value. When we relaxed some of the assumptions about properties of our data—that the residuals are normally distributed and independent of the machine learning errors—we further found no significant changes to the results. Finally, we compared the robust model to the alternatives of SIMEX (Yang et al. 2018) and random forest (RF)–IV (Yang et al. 2022). We found that the robust model has comparable or superior performance to these methods in terms of bias and hypothesis testing.

Table 2. Probability of Rejecting Null by Model and Effect Size

	Effect size						
Model	0	0.1	0.2	0.3	0.4	0.5	0.6
WC robust	0.0533	0.1445	0.4524	0.8040	0.9628	0.9968	0.9999
AC robust	0.0500	0.1401	0.4377	0.7891	0.9567	0.9958	0.9998
Naive OLS	0.0500	0.1389	0.4334	0.7846	0.9548	0.9954	0.9998
Corrected	0.0433	0.1392	0.4377	0.7828	0.9539	0.9955	0.9998
Ground truth	0.0733	0.1636	0.5258	0.8663	0.9844	0.9993	1.0000

Table 3. Summary of Additional Simulation Experiments

Test	Results	Support
Different unlabeled error magnitudes	Online Appendix H.2	Partial
Different tolerance levels	Online Appendix H.3	✓
Different effect sizes	Online Appendix H.4	✓
Different SML error magnitudes	Online Appendix H.5	✓
Error in control variable	Online Appendix H.6	✓
Different comparisons	Online Appendix H.7	Partial
Relaxing assumptions	Online Appendix H.8	✓
Benchmark comparison	Online Appendix H.9	✓

5. Applying the Robust Regression Method

Now that we have validated the robust regression model, we outline a step-by-step process for applying the robust optimization solution to hypothesis testing on SML-generated variables. This process is detailed in an accompanying GitHub repository with all accompanying code (https://github.com/aschecter/RobustReg) and visualized in Appendix B. First, researchers should separate their complete set of data into two components: the fully labeled component and the partially labeled component (i.e., data without ground truth for the SML variables). Using the fully labeled data, identify the tolerance parameter ρ and obtain cross-validation estimators of the error variances. Subsequently, train the robust regression model using the formula in (1); the WC or AC estimator can be used depending on the desired level of conservatism. The standard errors for the coefficients can be obtained following the definitions in Section 3.3.1. These parameter estimates and standard errors should be used for hypothesis testing by calculating the Wald test statistic $T_R = \hat{\beta}^R / SE(\hat{\beta}^R)$ and p-value p^R. To obtain an unbiased estimator, apply the solution in Theorem 4 by using the estimated covariance matrices from the labeled data; this provides the consistent estimator $\hat{\boldsymbol{\beta}}^*$ and standard error $SE(\hat{\boldsymbol{\beta}}^R)$. Researchers can then interpret the coefficients for the SML variable(s) according to the template laid out in Table 4.

Researchers can use the robust coefficient and p-value to identify the minimum effect size for errors of a given magnitude and test for whether the effect is nonzero. This step is essentially a binary test for whether a signal is present (but not what the signal is) given the presence of error. Because each entry $\hat{\beta}^R$ is a conservative estimate of β, the consistent estimator $\hat{\beta}^*$ should be used to interpret the economic effect (signal without noise).

We identify three outcomes of interest with respect to hypothesis testing. First, if an element of $\hat{\beta}^R$ is not significant, one concludes that, after accounting for the possible measurement errors, there is not enough evidence to conclude an effect is present. Second, if an element of $\hat{\beta}^R$ is significant but the corrected estimator $\hat{\beta}^*$ is not, then one concludes that (i) a nonzero effect is likely present even with measurement errors and (ii) we do not have enough evidence to identify an unbiased effect. This result likely emerges in cases of small sample sizes or with very small error magnitudes (in which case $\hat{\beta}^R$ may be close to unbiased on its own). Third, if both $\hat{\beta}^R$ and $\hat{\beta}^*$ are statistically significant, then we conclude that (i) an effect is present and (ii) an unbiased estimate is $\hat{\beta}^*$. Note that it is never the case that $\hat{\beta}^R$ is not significant but the corrected term is.

Our final consideration is when to use the robust approach discussed in this paper rather than other methods. Through the empirical examples, we find that the theoretical properties of the robust estimator were unaffected by effect size, error magnitude, or error distribution. However, the advantages of the robust method are less pronounced for very small effects or very small errors. Thus, if the observed errors in the labeled data are small (e.g., 5% in magnitude or less), it may not be necessary to use the robust methodology; of course, other existing error correction methods are also less useful. Finally, if the measurement errors are approximately normally distributed, we anticipate that OLS and SIMEX perform relatively well, albeit with larger standard errors. Yet, in all of our empirical examples, the corrected robust estimator had a lower RMSE than either

Table 4. Template for Interpreting Robust Optimization Outputs

Output	Interpretation
$\hat{\beta}^R$	The solution to the robust optimization problem with a given tolerance ρ. It is the minimum effect size for β that is identifiable if errors are as egregious as possible (WC) or if they are consistent with the observed errors (AC).
p^R	The probability that $\hat{\beta}^R \neq 0$ under the null hypothesis of no effect. It should be used to indicate if a nonzero effect is present.
$\hat{\beta}^*$	The unbiased estimate of β based on the empirical data.
p^*	The probability that $\hat{\beta}^* \neq 0$ under the null hypothesis of no effect. It should be used to indicate whether an unbiased estimate can be obtained with precision.

method, suggesting it performs well across a variety of scenarios.

6. Empirical Example: Amazon Reviews

Armed with our procedure for robust hypothesis testing, we next provide a concrete example of how to apply robust optimization to a real data set. Following recent work (Yang et al. 2018, Qiao and Huang 2021), we analyze Amazon Fine Food Reviews (McAuley and Leskovec 2013). The dependent variable of interest is helpfulness. The control variables are votes, which is the number of votes on a review (helpful or not helpful), and words, which is the length of the review text. We take the natural logarithm of these variables as well as the dependent variable. The main independent variable, sentiment, is the number of stars that the reviewer gave the product. This variable is our target to predict with SML, leveraging the text of the review itself. The entire data set contains 568,453 observations; however, we exclude reviews with fewer than five total votes. The final data set used in our analysis has 67,476 observations.

We constructed the data set using two techniques with varying levels of accuracy. First, we extracted sentiment scores from the review data using eight emotional dimensions (Mohammad and Turney 2010). The resulting values represent the relative strength of emotions such as fear, anger, and joy.[4] Then, we used a random forest model to predict the review rating using these eight dimensions. Using a random forest model enables us to apply the instrumental variable method introduced by Yang et al. (2022). Second, we used a state-of-the-art text-mining method—bidirectional encoder representations from transformers (BERT)—to predict sentiment from the text data. We trained both models on 25% of the data and then applied the tuned model to the remaining observations. These model predictions constitute our SML-generated variables with two degrees of accuracy. For additional details on the models and how they were tuned, refer to the Online Appendix. The random forest model had an RMSE of 1.047, whereas the BERT model had an RMSE of 0.823, which is significantly more accurate.

After generating the data, we then estimated the coefficients of the following regression model:

$$\log(\textit{helpfulness}) = \beta_0 + \beta_1 \times \log(\textit{votes}) + \beta_2 \times \log(\textit{words}) + \beta_3 \times (\textit{sentiment}).$$

To be consistent with our prior analyses, we separated 10% of the data as the labeled data set for which we know both the true and predicted values of sentiment. We tuned the robust model on this 10% of the data, applying 10-fold cross-validation. The remaining 90% of data were treated as unlabeled with only the predicted sentiment score being used. The WC and AC robust estimators were then fit to this data; we labeled these coefficients $\boldsymbol{\beta}^{WCR}$ and $\boldsymbol{\beta}^{ACR}$. Additionally, we computed a corrected robust estimator $\boldsymbol{\beta}^{*}$ following Section 3.3.3. To compare our approach with other state-of-the-art procedures, we applied the SIMEX method (Carroll and Stefanski 1994, Yang et al. 2018) and the random forest IV approach (Yang et al. 2022) (only for the random forest approach to generating data). We estimated the measurement error variance using the 10% labeled data. The fitted parameters were labeled $\boldsymbol{\beta}^{SIM}$ and $\boldsymbol{\beta}^{IV}$. As a baseline, we fit OLS directly to the unlabeled data, and designated these coefficients as $\boldsymbol{\beta}^{OLS}$. Finally, we fit a regression model to the same observations as the unlabeled data but using the true sentiment scores. We consider this parameter vector $\boldsymbol{\beta}^{TRUE}$ to be the ground truth. The results of our analysis using the random forest are presented in Table 5, whereas the results using BERT are presented in Table 6.

As can be observed from Table 5, both SIMEX and the random forest IV method improved the estimate of β_3 from 0.1580 to 0.2770 and 0.2766, respectively (ground truth estimate is 0.3226). The worst case robust estimator was much more biased (0.0175) though the standard error was significantly smaller. The average case robust estimator was somewhat less biased (0.0765) but still

Table 5. Regression Coefficients and Standard Errors for Alternative Models with Sentiment Estimated via Random Forest

	Model						
	Ground truth $\hat{\boldsymbol{\beta}}^{TRUE}$	Naive OLS $\hat{\boldsymbol{\beta}}^{OLS}$	SIMEX $\hat{\boldsymbol{\beta}}^{SIM}$	RF-IV $\hat{\boldsymbol{\beta}}^{IV}$	WC robust $\hat{\boldsymbol{\beta}}^{WCR}$	AC robust $\hat{\boldsymbol{\beta}}^{ACR}$	Corrected $\hat{\boldsymbol{\beta}}^{*}$
log(*votes*)	0.7136	0.6990	0.7012	0.7011	0.6966	0.6976	0.7171
	(0.0025)	(0.0027)	(0.0027)	(0.0029)	(0.0027)	(0.0027)	(0.0027)
log(*words*)	0.1212	0.1295	0.1171	0.1147	0.1444	0.1381	0.1266
	(0.0025)	(0.0027)	(0.0027)	(0.0029)	(0.0027)	(0.0027)	(0.0027)
sentiment	0.3226	0.1580	0.2770	0.2766	0.0175	0.0765	0.3306
	(0.0024)	(0.0027)	(0.0036)	(0.0053)	(0.0003)	(0.0013)	(0.0056)
RMSE	0.0042	0.1655	0.0477	0.0488	0.3065	0.2472	0.0123

Table 6. Regression Coefficients and Standard Errors for Alternative Models with Sentiment Estimated via BERT

	Model					
	Ground truth $\hat{\beta}^{TRUE}$	Naive OLS $\hat{\beta}^{OLS}$	SIMEX $\hat{\beta}^{SIM}$	WC robust $\hat{\beta}^{WCR}$	AC robust $\hat{\beta}^{ACR}$	Corrected $\hat{\beta}^{*}$
log(*votes*)	0.7136	0.7121	0.7119	0.6972	0.7092	0.7178
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
log(*words*)	0.1212	0.1105	0.1113	0.1442	0.1172	0.1264
	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)	(0.0025)
sentiment	0.3226	0.2993	0.2927	0.0171	0.2429	0.3331
	(0.0024)	(0.0025)	(0.0024)	(0.0001)	(0.0020)	(0.0028)
RMSE	0.0042	0.0261	0.0318	0.3068	0.0800	0.0133

more so than the alternative methods. The corrected robust coefficient estimate was 0.3306, which is significantly closer to the ground truth than all other methods. Overall, the corrected robust estimator demonstrated superior performance with an RMSE of 0.0123 (calculated with respect to the ground truth), which is significantly smaller than the random forest IV method (0.0488) and SIMEX (0.0477). In sum, the robust coefficient, although biased, allowed us to validate the hypothesis test for sentiment even after accounting for measurement error.[5] Further, the corrected robust estimator was less biased than all other methods and shows overall superior performance when accounting for bias and precision together.

We repeat this analysis using BERT to predict sentiment, and this produced much more accurate estimates of the target variable. The regression results are presented in Table 6; the RF-IV method is not included because the SML model was not tree-based. Model performance for OLS and SIMEX improved with the more accurate first stage model as anticipated. The worst case robust estimator was effectively unchanged though the standard error decreased further. Notably, the average case robust estimator was 0.2429, which is much closer to the SIMEX and OLS estimates. This result suggests that, for more accurate SML models, the AC robust estimator can be comparable in bias and RMSE to SIMEX even without adjustments. Finally, the corrected robust estimator still demonstrated the lowest bias in the key variable (0.3331 compared with 0.2993 for OLS) and the best overall performance with the lowest RMSE (0.0133 compared with 0.0261 for OLS and 0.0318 for SIMEX). The relatively small RMSE indicates that the corrected robust estimator's low bias compensates for the larger standard error and statistically dominates the other solutions.

7. Discussion

In this study, we address the issues of hypothesis testing and bias correction in linear regression models when one or more variables (independent or dependent) are generated by SML. This two-step estimation framework is fundamental to causal inference with machine learning in both the IS discipline and in empirical research more broadly. However, a drawback of this approach is the potential for errors to be introduced by the SML methods. No SML algorithm is perfect, and some degree of error is unavoidable. Drawing upon the computational design science paradigm (Abbasi et al. 2024), our paper contributes to the burgeoning literature on correcting such errors (Yang et al. 2018, 2022; Qiao and Huang 2021) by introducing an alternative method, robust optimization, into the two-step estimation framework to mitigate SML-induced measurement errors. We apply the robust least squares solution to the two-step causal machine learning setting and extend the model by considering two types of uncertainty—adversarial errors and expected errors—and deriving corresponding test statistics. We also leverage the empirical information available in the SML pipeline to compute a correction term that debiases the robust solution. Using multiple simulation experiments and a real data set, we demonstrate the advantages of this approach for inference.

A key contribution of this work is its design insights into how robust optimization can improve statistical power for hypothesis testing. In practice, noisy covariates tend to produce biased (often attenuated) coefficient estimates and inflated standard errors, leading to difficulties in detecting true effects. Robust optimization systematically amplifies the underlying signal relative to the noise, resulting in larger test statistics than naive OLS on noisy data. As such, the probability of rejecting the null hypothesis can be boosted when there is indeed a true effect without inflating false positives. This addresses a key gap in prior measurement error solutions, many of which do not explicitly aim to boost statistical power. Moreover, our introduction of the expected error, along with the adversarial error, provides researchers with two approaches to modeling SML-induced errors with varying levels of conservatism.

A second key theoretical contribution lies in deriving a closed-form solution for the robust regression coefficients, complete with exact formulas for their standard errors. Unlike approaches such as SIMEX, which correct bias but rely on numerical approximations of standard errors for inference, the analytic nature of this robust estimator allows for a more precise study of the bias–variance trade-off. Moreover, we provide an explicit bias-correction step that requires a relatively small labeled data set, enabling researchers to construct a consistent estimator even when only a fraction of data is labeled or ground truth is expensive. Importantly, the corrected robust model has the potential to statistically dominate other solutions given its combination of low bias and small standard errors.

In terms of practical impact, the robust optimization approach equips both researchers and practitioners with a step-by-step methodology (see Section 5) to mitigate measurement errors that inevitably arise when predicting constructs from text, images, or other unstructured data. The broad applicability of our approach is twofold. First, this approach avoids specific distributional assumptions about the errors. This generality can be a major advantage in practice given that ML errors may be non-Gaussian or exhibit complex patterns. Second, our proposed approach can handle a flexible number of SML-induced errors in independent variables, dependent variables, or both. Accordingly, our approach applies to hybrid studies in which multiple forms of unstructured data are turned into constructs. This approach is particularly relevant for strategic decision-making applications. Not only does the robust method help managers avoid overstating (or understating) the true impact of these predictive variables, but it also offers confidence in hypothesis testing results that inform critical organizational policies, marketing strategies, and investment choices.

7.1. Comparison of Robust Optimization to Other Methods

7.1.1. When Robust Optimization Helps—And When It Does Not. Our results should not be read as claiming universal superiority of the robust estimator. Rather, the estimator is one tool in a broader measurement error toolkit whose advantages depend on the error structure and available information. Robust optimization is especially well-matched to bounded or norm-constrained errors and to settings in which measurement error may be correlated with X and/or heteroscedastic (potentially depending on other covariates). In such cases, the uncertainty set formulation is a faithful abstraction and, under the variance-ordering conditions we establish, the associated Wald tests are (weakly) more powerful than OLS with noisy proxies. By contrast, when the assumptions behind alternatives hold—for example, SIMEX/MC-SIMEX with unbiased errors and reliable error–variance inputs or replicates (e.g., Yang et al. 2018) or forest IV/IV-style approaches with valid surrogates/instruments for the latent regressors (e.g., Yang et al. 2022)—those methods can be preferable. Our oracle bound formalizes this perspective: no method that observes only noisy proxies can asymptotically beat a test that observes the error-free covariates, whereas our variance-ordering result identifies when the robust estimator improves on OLS within the noisy-covariate regime.

7.1.2. Empirical Guidance and Comparisons with Existing Methods. To make these trade-offs concrete in our experiments, we (i) vary the magnitude and structure of measurement error (including bounded and correlated designs), (ii) report sensitivity to the robustness tolerance ρ (for WC/AC cases), and (iii) include head-to-head comparisons against SIMEX/MC-SIMEX and Forest IV alongside a labeled data bias correction that yields a consistent, low-RMSE variant of our estimator (similar to Qiao and Huang 2021). The emerging picture is practical: when domain knowledge implies bounded scores (e.g., indices in [0, 1]) or plausible correlation/heteroscedasticity in the errors, robust optimization is often the right default; use the corrected version when a labeled subset is available and report WC/AC sensitivity. When unbiased-error calibrations or valid instruments are credible, SIMEX/Forest IV may be more appropriate. Framed this way, our contribution delineates when robust optimization excels, provides theory that bounds its power relative to OLS and to an oracle, and supplies empirical evidence against leading alternatives so researchers can match methods to the data-generation process at hand.

7.2. Limitations and Future Work

Despite its advantages, the robust optimization approach faces several limitations that merit careful consideration. For instance, if the errors exhibit strong heteroscedasticity, autocorrelation, or other complexities, one must further refine the uncertainty set to match those patterns or introduce HC standard error estimates. More generally, the researcher chooses which variables are subject to contamination and to what extent they may vary, which could affect the outcome. To mitigate the bias that could be injected by this process, we recommend testing a variety of values for ρ as well as the average and worst case estimators.

Additionally, as with many measurement error corrections, the proposed solution depends on having a small but reliable labeled data set to tune the uncertainty bound and estimate correction terms. If labeled data are limited or unrepresentative, the bias-correction step may become unreliable. Further, if the costs of labeling sufficient data are too high, we recommend

exercising caution when interpreting the robust estimators and potentially using synthetic data to supplement the training information; future work could explore this approach. Further, whereas robust optimization can improve statistical power in detecting true effects, it does not, by itself, address other endogeneity concerns such as omitted variables or simultaneity, which may also bias inferences.

Acknowledgments

A. Schecter thanks the faculty at the University of Notre Dame Department of Information Technology, Analytics, and Operations for their helpful feedback. The authors thank the anonymous reviewers, associate editor, and senior editor for their constructive feedback.

Appendix A. Simulation Setup

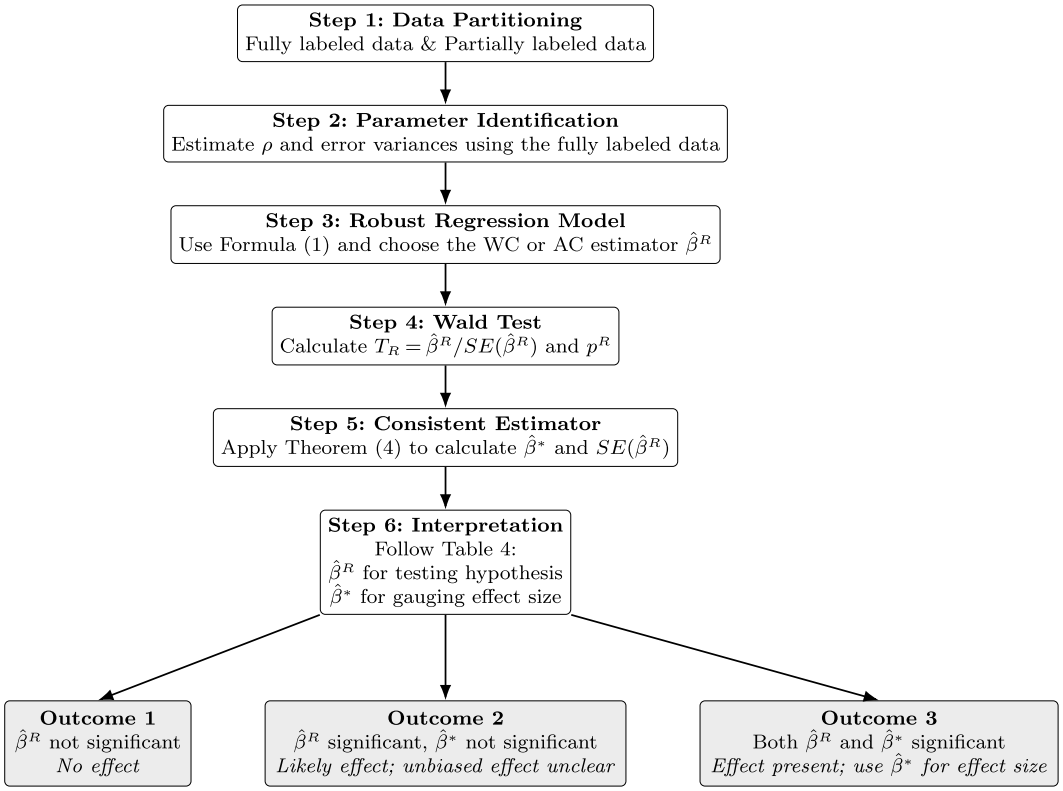
We first generated three features X_1, X_2, and X_3 by drawing from the multivariate normal distribution. The mean of each variable was zero, and the covariance matrix was such that $\sigma_{ii} = 1$ for $i = 1,2,3$ and $\sigma_{ij} = 0.1$ for all $i \neq j$; this allows for some minor collinearity among the features. The ground truth parameters were $\beta_1^{TRUE} = \beta_2^{TRUE} = \beta_3^{TRUE} = 1$. The dependent variable was then computed as $\boldsymbol{Y} = \beta_1^{TRUE}\boldsymbol{X}_1 + \beta_2^{TRUE}\boldsymbol{X}_2 + \beta_3^{TRUE}\boldsymbol{X}_3 + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ was also drawn from the standard normal distribution. All variables are mean centered to ensure consistency. These values of $\boldsymbol{Y}$ and $\boldsymbol{X}$ are considered the true or nominal values of the data, assuming no measurement error. To create the measurement error, we generate vectors $\Delta\boldsymbol{X}_1$, $\Delta\boldsymbol{X}_2$, $\Delta\boldsymbol{X}_3$, and $\Delta\boldsymbol{Y}$ by drawing 10,000 observations from the uniform distribution with a range of -1 to 1. Using these errors, we then created $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \Delta\boldsymbol{Y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \Delta\boldsymbol{X}$.

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 then also tuned the parameter ρ using the bootstrap procedure described in Section 3.4 with 1,000 bootstrap repetitions on the labeled data. The optimal parameter value $\hat{\rho}$ was determined to be the 95th percentile of the bootstrap values and was then used for testing.

We next fit the robust model as defined in (1) using the 9,000 unlabeled observations[6] $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$ and the tolerance parameter $\hat{\rho}$. We also fit the typical regression estimator $\hat{\boldsymbol{\beta}}^{OLS} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$ to this unlabeled data. The correction as detailed in Theorem 4 was applied to the robust estimator. Finally, we fit a linear regression model to the true values of the 9,000 observations used for testing; the resulting estimates can be considered true or unbiased coefficients.

Appendix B. Robust Regression Method Flowchart

Figure B.1. Flowchart of Robust Hypothesis Testing



Endnotes

[1] A detailed review of these techniques is provided in the Online Appendix.

[2] In practice, this assumption is not required for the results presented in this paper.

[3] Note that we do not need to know these values exactly; they can be approximated from the observed data for purposes of calibrating the tolerance level ρ.

[4] These emotions are captured using the package "syuzhet" in R (Jockers 2017).

[5] In both cases, the robust models had smaller p-values than OLS and larger p-values than the ground truth as hypothesized though the differences were negligible given the sample size.

[6] Note that, in practice, researchers may run the regression using all data, including the labeled observations. However, we exclude that here to ensure there is no influence of the labeled data on the final results.

References

Abbasi A, Parsons J, Pant G, Sheng ORL, Sarker S (2024) Pathways for design research on artificial intelligence. *Inform. Systems Res.* 35(2):441–459.

Bertsimas D, Nohadani O (2019) Robust maximum likelihood estimation. *INFORMS J. Comput.* 31(3):445–458.

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Rev.* 53(3):464–501.

Bertsimas D, Den Hertog D, Pauphilet J (2021) Probabilistic guarantees in robust optimization. *SIAM J. Optim.* 31(4):2893–2920.

Bertsimas D, Gupta V, Kallus N (2018) Data-driven robust optimization. *Math. Programming* 167:235–292.

Bertsimas D, Dunn J, Pawlowski C, Zhuo YD (2019) Robust classification. *INFORMS J. Optim.* 1(1):2–34.

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* (Chapman and Hall/CRC, New York).

Carroll RJ, Stefanski LA (1994) Measurement error, instrumental variables and corrections for attenuation with applications to meta-analyses. *Statist. Medicine* 13(12):1265–1282.

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

Chan J, Wang J (2018) Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Sci.* 64(7):2973–2994.

Chen R, Paschalidis IC (2018) A robust learning approach for regression models based on distributionally robust optimization. *J. Machine Learn. Res.* 19(13):1–48.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

El Ghaoui L, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.* 18(4):1035–1064.

Fong C, Tyler M (2021) Machine learning predictions as regression covariates. *Political Anal.* 29(4):467–484.

Fuller WA (2009) *Measurement Error Models* (John Wiley & Sons, New York).

Gu B, Konana P, Rajagopalan B, Chen HWM (2007) Competition among virtual communities and user valuation: The case of investing-related communities. *Inform. Systems Res.* 18(1):68–85.

Hong LJ, Huang Z, Lam H (2021) Learning-based robust optimization: Procedures and statistical guarantees. *Management Sci.* 67(6):3447–3467.

Jockers M (2017) Package "syuzhet." https://cran.r-project.org/web/packages/syuzhet.

Kuhn D, Esfahani PM, Nguyen VA, Shafieezadeh-Abadeh S (2019) Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research* (INFORMS, Catonsville, MD), 130–166.

Lee D, Hosanagar K, Nair HS (2018) Advertising content and consumer engagement on social media: Evidence from Facebook. *Management Sci.* 64(11):5105–5131.

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.

Meijer E, Oczkowski E, Wansbeek T (2021) How measurement error affects inference in linear regression. *Empirical Econom.* 60(1):131–155.

Mohammad S, Turney P (2010) Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Proc. NAACL HLT 2010 Workshop Comput. Approaches Anal. Generation Emotion Text* (Association for Computational Linguistics, Stroudsburg, PA), 26–34.

Qiao M, Huang KW (2021) Correcting misclassification bias in regression models with variables generated via data mining. *Inform. Systems Res.* 32(2):462–480.

Schecter A, Nohadani O, Contractor N (2022) A robust inference method for decision making in networks. *MIS Quart.* 46(2):713–738.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

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.

Yang M, McFowland E III, Burtch G, Adomavicius G (2022) Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS J. Data Sci.* 1(2):138–155.

Zhang S, Lee D, Singh PV, Srinivasan K (2022) What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Sci.* 68(8):5644–5666.