Remote Control: Debiasing Remote Sensing Predictions for Causal Inference

By Matthew Gordon, Megan Ayers, Eliana Stone, and Luke Sanford*

Draft: June 29, 2024

ABSTRACT

Advances in machine learning and the increasing availability of satellite imagery have led to the proliferation of social science research that uses remotely sensed measures of human activity or environmental outcomes to infer the impact of policy. However, prediction errors from machine learning models can lead to bias in the estimates of regression coefficients. In this paper, we show how this bias can arise, propose a test for detecting bias, and demonstrate the use of an adversarial machine learning algorithm in order to debias predictions. These methods are applicable to any setting where machine learned predictions are the dependent variable in a regression. We conduct simulations and empirical exercises using ground-truth data on forest cover in Africa. Using the predictions from a naive machine learning model leads to biased parameter estimates, while the predictions from the adversarial model recover the true coefficients. We then use the method to study gold mining-induced deforestation in Africa and find that using machine-learning predictions of deforestation cause us to underestimate the direct impacts of mining on deforestation, but overestimate the indirect impacts.

* Thanks to Eli Fenichel, Philipp Ketz, the Berkeley Political Methodology Seminar, the LSE/Imperial College Workshop, and participants at the AGU, ICLR, TWEEDS, and AERE conferences for helpful feedback and support. Gordon: Paris School of Economics; matthew.gordon@psemail.eu. Ayers: Yale Department of Statistics and Data Science; m.ayers@yale.edu. Stone: Yale School of the Environment; eliana.stone@yale.edu. Sanford: Yale School of the Environment; luke.sanford@yale.edu.

I. Introduction

Advances in machine learning and the increasing availability of satellite imagery have led to the proliferation of social science research that uses remotely sensed measures of human activity or environmental outcomes to infer facts about the world. However, when the machine learning models used to measure these outcomes minimize a standard loss function, the resulting predictions will often produce biased estimates when used to estimate regression parameters. If the measurement error in the outcome variable is correlated with policy variables or important confounders, as is the case for many widely used remote sensing data sets, estimates of the causal impacts of interventions will be biased. This bias can occur even in cases when researchers have a good instrument or an experimental research design.

In this paper, we show how this bias can arise, we describe a test for bias, and we propose methods to correct the bias, including an adversarial debiasing algorithm (Zhang, Lemoine and Mitchell, 2018) that can generate machine learning predictions that will result in less bias when used in regressions. Our algorithm has broad applications beyond satellite data to any setting in which researchers are using machine-learned outcome variables as the dependant variable in regressions.

Previous research has derived measures of economic output, air pollution, land-use change, and other variables at high resolution by using algorithms trained on satellite imagery and some ground-truth data (Henderson, Storeygard and Weil, 2011; Hansen et al., 2013; Meng et al., 2019). A typical approach is to train a machine learning model on satellite data using a limited number of ground-truth observations and then using the model predictions to impute outcomes for a larger population of interest. These measurements have then been widely used as a dependent variable in regressions to estimate the effects of various policies

on deforestation, GDP, pollution, and other variables (see e.g. Burgess et al. 2012; Alix-Garcia et al. 2013; Meng et al. 2019; Asher, Garg and Novosad 2020; Wren-Lewis, Becerra-Valbuena and Houngbedji 2020; Slough 2021; Sanford 2021; Jack et al. 2022). For example, the Hansen et al. (2013) estimates of deforestation have been cited more than 10,000 times.

While machine learning models can obtain high accuracy, there are widely documented biases in the predictions that can pose problems when they are used in regressions. Fowlie, Rubin and Walker (2019) showed that satellite estimates of pollutant concentrations show systematic attenuation bias toward the high end of the distribution. Bluhm and McCord (2022) demonstrated a similar bias in economic activity measured by night lights. Tropek et al. (2014) showed that the Hansen et al. (2013) predictions sometimes confuse tree plantations for forests. Although these biases were often ignored in early studies, this non-classical measurement error can violate the assumptions that are required for consistent estimation.

We show analytically and descriptively why this can be a problem, starting with the well-known result that the bias in a regression coefficient depends on the covariance between the covariate and the machine learning model's prediction error. Crucially, this bias can occur even in a randomized control trial or quasi-experimental setting where standard assumptions for causal identification typically hold, since treatment can induce differential measurement error. We give a number of examples of how this can occur in practice.

Given the formula for the bias of a regression coefficient, we observe that a simple test for bias is to regress prediction errors on the covariates of interest using a subsample of ground-truth data. This test can also be used to 'correct' regression coefficients estimated using the full sample. Furthermore, a power analysis of the regression can be used to determine how many observations researchers would

need to label to detect bias of a given size. In many cases, researchers may be able to collect this ground-truth data using high-resolution imagery.

Our second contribution is to demonstrate how the use of machine learning models with modified loss functions can eliminate biases in prediction errors and improve efficiency in some cases. An interesting version of this technique is adversarial debiasing — an algorithm that was originally developed to ensure that machine learning algorithms do not encode racial or other undesirable biases for decisions like hiring, admissions, or bail. Zhang, Lemoine and Mitchell (2018) propose a machine learning algorithm that uses a modified loss function to ensure that predictions are unbiased with respect to 'protected characteristics'. We directly borrow their approach and use the treatment variable as the protected characteristic to make sure that predictions are unbiased for use in causal estimation.

Intuitively this method can be understood as follows — a primary model attempts to minimize prediction error for the outcome of interest. The measurement errors are then passed to a secondary model (the adversary), that tries to predict the treatment status of an observation. When tuning the first model, a penalty term is added to the loss function that increases if the adversary's predictions improve. Thus the primary model attempts to minimize prediction error while also making errors uninformative about treatment status. In the special case where the adversary is a linear regression of prediction errors on regression covariates, we show that this loss function penalizes the covariance between prediction errors and the treatment variable. This can reduce or eliminate biases in regression coefficients estimated using these predictions.

We then demonstrate the effectiveness of these approaches by applying them to measurements of forest loss in Africa – a setting where we have access to ground-truth data on forest cover (Bastin, 2017; Guo, Zhu and Gong, 2022). As

a proof of concept, we simulate a scenario where prediction errors are correlated with the independent variable of interest. This leads to biased estimates of the target coefficient. The adversarial debiasing approach recovers the true parameter without requiring any knowledge of the sources of prediction error. We also demonstrate the utility of the bias test and bias correction approach.

Next, we conduct a descriptive exercise to measure forest cover as a function of distance to roads. Using standard machine learning model predictions, estimates of the relationship of interest are biased, but our proposed methods again recover the true relationship, without requiring knowledge of the sources of prediction error.

Finally, we use these techniques to study the effect of a gold mining boom in Africa on deforestataion. Our empirical strategy takes advantage of a rapid and dramatic expansion of the West-African gold mining industry. Between 2009 and 2012, gold exports in Burkina Faso, Ghana, and Mali increased by 1,250%. Benshaul-Tolonen (2019) finds that the wealth increases from this mining boom led to lower infant mortality rates in nearby communities. Some cross-sectional estimates have found large effects of the mines on deforestation (Giljum et al., 2022), but given the apparent tradeoffs between human health and ecological outcomes, careful measurement is critical in this context.

Using the Hansen et al. (2013) predictions of forest cover, we find significant negative impacts of the 2011 gold rush on forest cover in areas close to mines. When we test for bias in these estimates, however, a more nuanced story emerges. The Hansen et al. (2013) actually underestimates deforestation in areas very close to mines — our corrected estimates are six times larger than the estimates obtained from the predictions. Further from mines, we find that the Hansen et al. (2013) overestimates deforestation. Our corrected estimates find null impacts of mines on forest cover at ranges greater than 100 km, and a positive and significant

estimate betweeen 50 and 100 km. This could support a story similar to Foster and Rosenzweig (2003), who found that increases in economic growth lead to more forests in India. Our results indicate the Hansen et al. (2013) predictions from satellite data seem to 'smooth' forest loss over time and space, causing us to miss important heterogeneity on the ground.

This paper contributes to a growing literature that has begun to document the problem of non-random measurement error in machine-learning models (see Jain 2020 for a review[1]). A number of new papers propose econometric estimators that can correct for the non-classical measurement error in some cases. Alix-Garcia and Millimet (2022) proposes a misclassification model that requires users to specify the variables that may induce measurement error (e.g. cloud cover, satellite angle). Proctor, Carleton and Sum (2023) suggest a multiple imputation approach that may be sensitive to functional form specifications. Zhang (2021) and Fong and Tyler (2021) propose techniques that rely on strong assumptions, including that the labelled data be a random sample of the population, and that measurement error is orthogonal to the true outcome conditional on the machine learned measurements, as well as any treatment variables or covariates. Torchiana et al. (2023) is similar, although it trades off a set of stronger assumptions about the data-generating process for the advantage of not requiring any labelled data. In contrast, our bias estimation and correction method relies on a weak consistency assumption, and no knowledge of functional forms or sources of measurement error. We show that these methods can succeed in cases where multiple imputation fails.

Furthermore, the adversarial debiasing technique is fundamentally different from the above approaches. While the previous literature seeks to address measurement error in the estimation step, adversarial debiasing addresses it while

[1]For topic specific reviews, see Balboni et al. 2022 on deforestation, Gibson et al. 2021 and Bluhm and McCord 2022 on night lights, and Fowlie, Rubin and Walker 2019 on air pollution.

making the original machine learning predictions. While more demanding, in that it requires researchers build a custom machine-learning model for any given analysis, it is widely applicable, and it also improves efficiency in some of our experiments. The ability to build customized machine learning models for outcomes of interest may become increasingly important as researchers learn more about the shortcomings of off-the-shelf remotely sensed data products in certain contexts. Our results show that very simple machine learning models can be sufficient to obtain consistent parameter estimates, as long as the prediction errors are balanced with respect to the policy variable.

Our work builds on at least one previous attempt to customize machine learning predictions on satellite data for use in social science research. Ratledge et al. (2021) use a modified loss function when generating machine learning predictions in order to reduce attenuation biases in their estimates of household wealth. Our adversarial approach is more general and can be applied to attenuation bias as well as many other types of non-classical measurement errors. It also has broad applicability beyond satellite data. The same basic approach can be applied to machine learning predictions on text data, such patents or tweets for example.

Finally, we make an interesting connection between the previous work on machine learning measurement error described above, and the literature on algorithmic bias and adversarial debiasing (Kleinberg, Mullainathan and Raghavan, 2016; Zhang, Lemoine and Mitchell, 2018; Kleinberg et al., 2018; Liang, Lu and Mu, 2023). We directly adapt some of the results regarding algorithmic bias in decision making to solve a common estimation problem. While adversarial models have been used in econometrics research previously, for example Chernozhukov et al. (2020) studied the use of an adversarial model to obtain debiased estimates of heterogeneous treatment effects, this paper describes a novel use for these algorithms in addressing non-classical measurement error.

In the following section, we lay out an analytical framework that shows why machine learned predictions can result in biased estimates of regression coefficients, and we describe our bias test. We then show how bias correction and adversarial debiasing can solve these problems. In section III we show that the methods work with simulated data and in a simple cross-sectional descriptive regression to recover the relationship between roads and forest cover. In IV we apply these methods to a time-series, causal application: the effect of a gold mining boom in Africa on deforestation.

II. Framework

For simplicity, consider a univariate regression. We seek to estimate the relationship between changes in some variable X on an outcome Y according to the model:

$$Y_i = \alpha + \beta X_i + e_i. \tag{1}$$

In this case, our parameter of interest is β - the marginal effect of X on Y. However we do not have access to the true Y_i. Instead we have predictions $\widehat{Y_i}$ based on a machine learning model. We can model $\widehat{Y}(k_i) = Y_i + \nu_i$, where k are the features we use to train our machine learning model, and ν_i is the measurement error for a given observation. When we estimate β from (1) using OLS with $\widehat{Y}$ as the dependent variable instead of Y_i, our estimate of β will equal in expectation:

$$\mathbb{E}[\widehat{\beta}] = \beta + \frac{Cov(e, X)}{Var(X)} + \frac{Cov(\nu, X)}{Var(X)}. \tag{2}$$

The middle term in the equation is standard endogeneity bias, and must equal

zero to recover the true parameter in any setting. For what follows, we ignore this source of potential bias as it is not the focus of our analysis. The final term is the result of using machine learned proxies for Y_i instead of the true values, and is the focus of this paper. In words, the measurement error from our machine learning model will bias our estimate if it is correlated with the treatment variable.

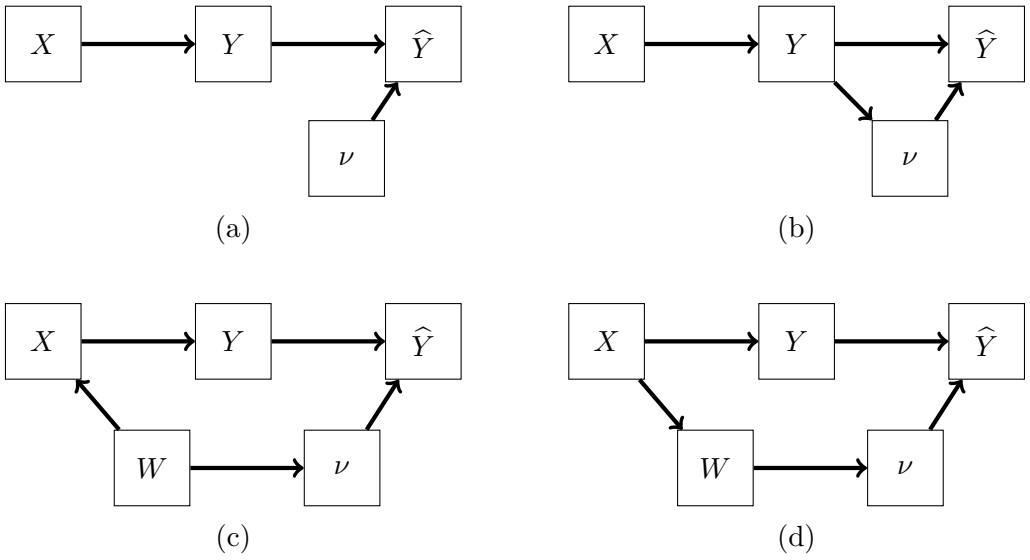


Figure 1. : Four Directed-Acrylic-Graphs illustrating potential relationships between treatment (X), outcomes (Y), measurement error (ν), machine learning model predictions ($\widehat{Y}$), and unobserved variables (W). (a) is classical measurement error, (b) shows outcome-induced bias, (c) shows confounder-induced bias, and (d) shows treatment-induced bias.

This situation can arise in several ways. First, measurement error can be correlated with the true values, Y_i, as Fowlie, Rubin and Walker (2019) demonstrated with air pollution predictions showing attenuation bias at higher concentrations. We call this outcome-induced bias. This is depicted in a Directed-Acrylic-Graph (DAG) in Figure 1.b. In this case, we can model the measurement error as a function of Y_i plus an idiosyncratic component ϵ_i:

$$\nu_i = f(Y_i) + \epsilon_i$$

$$\nu_i = f(\beta X_i + e_i) + \epsilon_i.$$

Thus, when $f(\beta X_i + e_i)$ is not constant with respect to X_i, $\widehat{\beta}$ may be biased, even in an RCT setting where $Cov(e, X) = 0$. Intuitively, if $\widehat{Y}$ doesn't change over some domain of the ground truth variable, estimates of the treatment effect cannot learn about effects in that part of the distribution of Y. Another example of this type of measurement error occurs when Y_i is binary, since in this case, errors are always negatively correlated with the true value of Y (Aigner, 1973).

Alternatively, assume measurement error is a function of some other variable W_i that is correlated with both treatment and outcomes, and some idiosyncratic component. For example, if we want to estimate the effect of a payment-for-ecosystem-services program on deforestation, some unobserved W such as soil moisture might both affect the probability that a parcel is enrolled in the program and cause forest cover to be over-estimated because of the increased ambient vegetation — in other words there is selection into treatment that is correlated with measurement error. See Figure 1.c for a graphical representation. In this case, measurement error can be modelled as a function of W:

$$\nu_i = g(W_i) + \epsilon_i.$$

If $Cov(X, g(W)) \neq 0$, again our estimates will be biased. Even controlling linearly for W_i in the regression is typically not sufficient to ensure $Cov(X, g(W)) = 0$.

Finally, in certain cases, treatment can induce measurement error. For example, take a researcher studying the effects of a cash transfer program on deforestation.

The treatment causes recipients to invest in irrigation and high-yielding crops, which may be more often confused for forests than the previous landcover. This would result in overestimation of the post-treatment forest cover for the treated group. We refer to this as treatment-induced bias. See Figure 1.d for a graphical representation.

These examples show that even if researchers have an experimental or quasi-experimental source of exogenous variation, estimates may still be biased when the dependent variable contains prediction error. Instrumental variables can be useful for measurement error in X. Even a valid and relevant instrument will not guarantee an unbiased estimate when Y contains measurement error, however.

A. Biased Predictions: Why do they occur?

If measurement error is systematic with respect to treatment, as in the cases described above, this raises the question of why the machine learning model didn't generate better measurements in the first place? Fong and Tyler (2021) claim that these types of errors are unlikely to occur with machine learning, since if the measurement error correlates with X, it does so at the expense of predicting Y. Despite this claim, biased measurements can arise for several reasons.

Machine learning models are typically trained by choosing a set of model parameters, ω, that minimize a loss function in a training data set — for example, the mean squared error between model predictions and true values. When choosing ω, a more complex model will better fit the training data, however, out of sample predictions will have greater variance. On the other hand, a simple model may be right on average, but biased in certain regions of the feature space. Navigating this bias-variance tradeoff is at the core of modern machine learning methods.

Still bias can arise for at least two reasons. First, limited or unrepresentative training data means that certain regions of the feature space can be given less

weight in the training process. Similarly, if data in certain parts of the feature space contains less information about the target variable of interest, then that will have less influence on the ω. This could occur if some of the data is of poorer resolution, for example. Liang, Lu and Mu (2023) formalize these ideas, showing that unless an algorithm's inputs satisfy a particular type of balance, the algorithm faces a tradeoff between accuracy and fairness (equivalent to unbiasedness in our context). In some cases they show that adding the group variable (X in our case) can even increase the bias of predictions.

B. Detecting and Correcting Bias

Unlike for omitted variable bias, an estimate of measurement error bias is directly obtainable if researchers have access to or can generate some ground-truth values of Y. Let $j \in J$ index observations in the labelled set. Consider the regression:

$$\nu_j = X_j\gamma + u_j, \tag{3}$$

with X as a vector of independent variables, and $\nu_j = \widehat{Y} - Y$ is the prediction error. Our estimate of γ will be $\widehat{\gamma} = (X'X)^{-1}X'\nu$. This is exactly the multivariate analog of the bias term in equation 2. This shows that a very simple regression coefficient can be used to estimate bias, under the consistency assumption that $\widehat{\gamma} \to_p \gamma$, where γ is the prediction error bias in the unlabelled population.

In practice, researchers may be able to obtain a number of such labels by visually interpreting high-resolution satellite imagery, for example. In this case, it should be easy to make sure that J is representative of the broader population, in which case consistency should hold. In cases where the labelled data suffers from selection bias or is non-representative in some other way, selection on observ-

ables techniques using the machine learning model inputs (k) may be a promising approach (Imbens, 2004).

Estimates of the standard errors of $\widehat{\gamma}$ can be used to test whether the bias is significantly different from zero, or to rule out biases greater than a certain size, though standard errors likely need to be adjusted for spatial or serial correlation, and/or bootstrapped in many cases.

It is also simple to adapt standard power calculations to estimate a 'minimum detectable bias' given a certain number of observations, and an estimate of the standard error of γ. Researchers can then estimate the number of labeled observations which will likely be necessary to rule out some amount of measurement error bias. We demonstrate this procedure in our first two applications.

Estimating the bias in this way can be a useful diagnostic, but it can also be used to perform a 'bias correction' on estimates of $\widehat{\beta}$ from equation 2. We define a bias corrected estimator as:

$$\widehat{\beta}_c = \widehat{\beta} - \widehat{\gamma}. \tag{4}$$

In expectation, this will converge to the true value of β when there is no endogeneity bias and the consistency assumption holds.

An estimate of the standard error of $\widehat{\beta}_c$ can be produced with a bootstrap procedure. Note that $\widehat{\beta}$ and $\widehat{\gamma}$ are calculated in different datasets (the unlabelled and labelled data respectively). Given the simple structure of our experiments below, we assume independence in the draws from each dataset, but this may not always hold in practice.

There are two main benefits to this approach. First, $\widehat{\beta}$ and $\widehat{\gamma}$ can be generated with off-the-shelf predictions and some ground-truth data. This should make the approach much easier to implement for those who do not want to generate a

custom machine learning model for their research question as we describe in the following section. Second, we expect this approach to outperform the adversarial approach described below in settings with relatively few ground truth observations since typically less training data will be required to estimate γ than to train a machine learning model with many parameters.

There are some shortcomings of this estimator, however. The estimation uncertainty around $\widehat{\gamma}$ will generally inflate the standard errors around $\widehat{\beta}_c$. In the case of independence, the standard errors of $\widehat{\beta}$ and $\widehat{\gamma}$ will be additive.

Secondly, this approach takes a set of $\widehat{Y}$ predictions as given. The joint distributions of $\widehat{Y}, Y$, and X potentially limit the precision of $\widehat{\beta}_c$. In the next section, we describe how to obtain the 'best possible' predictions of $\widehat{Y}$ for a given estimation problem.

C. Adversarial Debiasing

A machine learning model's predictions are a function of k, input features, and ω, model weights:

$$\widehat{Y} = f(k, w). \tag{5}$$

Typically model weights are chosen to minimize some loss function — for example the sum of squared prediction error in the labelled data: $\sum_j (\widehat{Y}_j(k_j, w) - Y_j)^2$. If the model is a linear regression, for example, then the model weights are the regression coefficients. Given our analysis above, however, we can also add a penalty to the loss function if the prediction errors are correlated with X in the labelled data. We thus train the model such that its parameters, w^*, are chosen to satisfy:

$$\omega^* = \arg\min_{\omega} L_p(\widehat{Y}(\omega), Y, k) \quad \text{such that } Cov(X, Y - \widehat{Y}(\omega)) \approx 0. \tag{6}$$

where L_p is a standard loss function, such as mean squared error. One way of thinking about the constraint on the loss function, is as a requirement that the measurement errors should contain as little 'information' as possible about X. Thus a second machine learning model should not be able to predict X from the errors. Call such a model the adversary and define its loss function as $L_a(\hat{X}(\gamma), X, Y, \hat{Y}(\omega), k)$, with model weights γ.[2] This setup is called adversarial debiasing, and was first proposed by Zhang, Lemoine and Mitchell (2018) to debias machine learning model predictions with respect to race or gender.

Adversarial debiasing models are trained to minimize L_p, while maximizing L_a, subject to the adversary choosing γ in such a way as to minimize L_a. Formally, this can be written as the following objective function:

$$\min_{\omega} L_p(\hat{Y}(\omega), Y, k) - \alpha L_a(X, Y, \hat{Y}(\omega), \gamma, k) \quad \text{subject to: } \gamma \in \text{argmin } L_a(X, Y, \hat{Y}(\omega), \gamma, k), \tag{7}$$

where α controls the weight on the adversary's loss function, and must be chosen by the researcher (e.g. using cross fitting). Now consider if the adversary model is a linear regression of the exact form of the bias test above:

$$\nu_j = X_j \gamma + \epsilon_j \tag{8}$$

[2] γ is intentionally chosen to be the same variable as the coefficient in the bias test above for reasons that are about to become clear.

with loss function as the sum of squared errors. The loss function is of this adversary is minimized with respect to γ when $\gamma = (X'X)^{-1}X'\nu$. The primary model will try to choose ω such that the prediction errors maximize the adversary's loss function:

$$L_a = \sum_j (\nu_j - X(X'X)^{-1}X'\nu)'(\nu_j - X(X'X)^{-1}X'\nu).$$

Take two sets of prediction errors: ν and $\tilde{\nu}$. If overall accuracy is equal (e.g. $\nu'\nu = \tilde{\nu}'\tilde{\nu}$, then $L_a(\nu) > L_a(\tilde{\nu}) \iff |(X'X)^{-1}X'\nu| < |(X'X)^{-1}X'\tilde{\nu}|$ in the case where X is univariate (Proof in Appendix A).

Thus, if α is sufficiently high, choosing ω to optimize equation 7 implies setting $\widehat{\gamma} = (X'X)^{-1}X'\nu$ in the labelled data close to zero. When the goal of maximizing the adversary's loss is balanced with the goal of overall prediction accuracy, the model attempts to minimize $\nu'\nu$ while also minimizing bias.

If the relationship between X and ν is similar in the training data to the rest of the sample, this will minimize prediction error bias in our estimate of β. This method is also applicable to a multivariate X, or an instrumental variables application, with slight modifications (details in Appendix B).

Another simple approach is to penalize the covariance of X and ν directly. For example, the model's objective function could be:

$$\min_{\omega} L_p(\hat{Y}(\omega), Y, k) - \alpha \Big| \sum_j (x_j - \bar{x_j})\nu_j \Big|. \tag{9}$$

In our applications, we experiment with both approaches. For both methods, the choice of α is important. With too low of an α the model does not effectively debias the results, minimizing squared prediction error instead of maximizing the adversarial loss. However, when α is too high the model may produce random

measurements to inflate the adversary's loss. The typical approach is to use cross fitting within the labelled data, such that overall prediction error and bias can be examined for different choices of α.

One downside of the approach detailed here is that it requires a unique machine learning model for each downstream estimation task. This suggests that, for example, measuring tree cover to estimate the effect of property rights on deforestation is different from measuring tree cover to estimate the effect of wealth shocks on deforestation. While this may be taxing for researchers, it also suggests that failing to build a measurement strategy for any individual task risks biasing that task.

For the researchers that are willing to build these models, adversarial debiasing has a few advantages over existing approaches. It does not require a researcher to know the source of the measurement error – the debiasing procedure will eliminate differential measurement error without specifying the precise source. Secondly, researchers may be able to use more sophisticated adversaries than a linear regression, for example, modelling ν non-parametrically as a function of both X and k. In this case, the model will choose ω to minimize the amount of information contained in the measurement errors, which may make it less likely that systematic relationships between ν and X will emerge outside of the training data.

Finally, it may allow the researcher to achieve unbiased estimates of treatment effects using very simple machine learning models. State of the art models that aim to maximize accuracy can be computationally demanding especially when used over large areas. In our examples below, we are able to recover accurate treatment effects using a logistic regression as our primary machine learning model, when used in conjunction with adversarial debiasing.

A Note on Standard Errors

When using predicted outcomes as the dependent variable in a linear regression, the variance of the estimated coefficients can be expressed as:

$$
\begin{aligned}
Var(\widehat{\beta}) &= Var\left((X'X)^{-1}X'\widehat{Y}\right) = Var\left((X'X)^{-1}X'(X\beta + e + \nu)\right) \\
&= Var\left((X'X)^{-1}X'e\right) + Var\left((X'X)^{-1}X'\nu\right) + \\
&\qquad 2Cov\left((X'X)^{-1}X'e, (X'X)^{-1}X'\nu\right)
\end{aligned} \tag{10}
$$

Holding the covariance term constant, larger prediction errors will inflate the variance of estimates of β through the second term. This highlights the potential efficiency gains from adversarial debiasing. Since the variance $\widehat{\beta}$ increases in the variance of the prediction errors, minimizing prediction errors, subject to an unbiasedness constraint, will ensure that predictions are optimized for a particular regression. On the other hand, our bias correction method takes the set of prediction errors as given, from a possibly non-optimized model, and adds additional error (u from equation 3). Formally, if adversarial model prediction errors are ν_a, and standard model prediction errors are ν:

$$
\begin{aligned}
Var(\widehat{\beta}) \quad &< \quad \text{Var}(\widehat{\beta}_c) \iff \\
Var\left((X'X)^{-1}X'\nu_a\right) + \quad & \qquad Var\left((X'X)^{-1}X'\nu\right) + \\
2Cov\left((X'X)^{-1}X'e, (X'X)^{-1}X'\nu_a\right) \quad &< \quad 2Cov\left((X'X)^{-1}X'e, (X'X)^{-1}X'\nu\right) + \\
& \qquad Var\left((X'X)^{-1}X'u\right) + \\
& \qquad 2Cov\left((X'X)^{-1}X'e, (X'X)^{-1}X'u\right) + \\
& \qquad 2Cov\left((X'X)^{-1}X'\nu, (X'X)^{-1}X'u\right)
\end{aligned} \tag{11}
$$

While the first two lines on the left are likely to be somewhat larger than the first two lines on the right, due to the constraint (although not always, as we will show in the examples), as long as this difference is less than the additional variance from estimating the bias term, the adversarial predictions will result in smaller standard errors. Intuitively, as the adversarial model learns to make predictions that are precise and minimize bias, it may learn features that are predictive of measurement error and use those to produce high-quality measurements that are appropriate for the causal task.

We can use the typical OLS standard errors or heteroskedasticity-consistent standard errors to estimate the variance of $\widehat{\beta}$ under-sampling uncertainty after performing adversarial debiasing. However, this practice doesn't take into account uncertainty about the prediction model itself. If our training set was another sample from the same source, it might lead to a different set of measurements, and a different estimate of $\widehat{\beta}$.

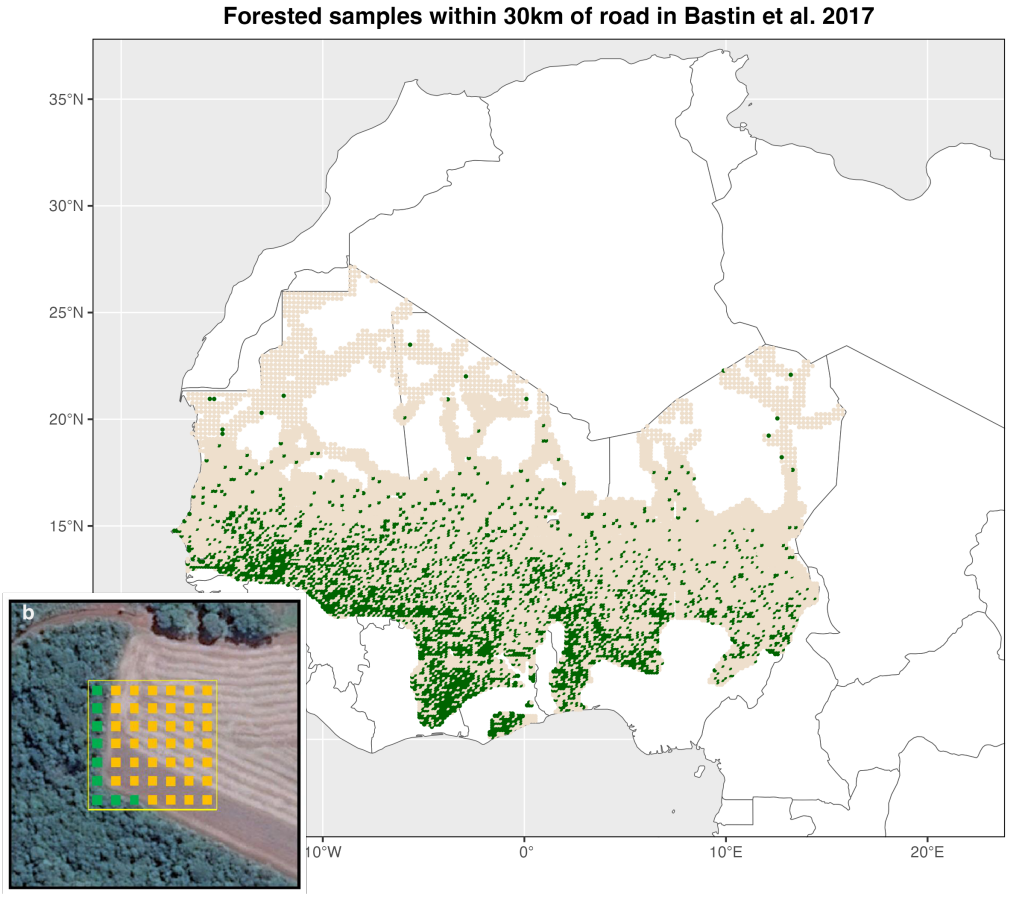
Most studies using machine-learning measurements as outcome variables do not account for prediction uncertainty in their reported standard errors. One approach to account for this source of uncertainty is to bootstrap the training of the machine learning model, drawing different samples from the training set to train the model. This approach is computationally intensive, since it requires bootstrapping the training of a potentially complex machine learning model, however, our simulations show that it can be important in practice.

In the sections that follow, we use hand-labeled datasets of forest cover in Africa to explore the biases generated by machine learning measurement error. We start with a simulation study of the cross-sectional relationship between roads and forest cover, and then estimate the relationship with real data. In section IV, we apply the above techniques to study the effects of a sudden wealth shock resulting from the West-African gold mining boom on local forest cover.

III. Roads and deforestation

To demonstrate the efficacy of our approach, we conduct simulations and empirical exercises on a common remotely-sensed outcome – forest cover – in settings where we also have access to ground truth data. For our first application, we use a hand-labeled dataset of 20,621 points in West Africa (Bastin et al. 2017) that are labelled with their percent treecover as our "true" measures of forest cover. This data is a cross-sectional sample of points from a grid covering an enormous region of dryland forest across much of West Africa. This data was collected in part to show the biases of the Hansen et al. (2013) data in dryland areas. Researchers used high-resolution imagery from between 2011 and 2015 to label the data — an example of which is shown in Figure 2. We use a sample of the data in West Africa within 30 km of a road and 100 km of a DHS cluster (see Figure 2).

As inputs to our machine learning models, we use data from the Landsat 7 ETM sensor. This sensor records the surface reflectance of light at several visible, near-infrared, and infrared wavelengths (called 'bands' in the remote sensing literature) at a 30-meter resolution. We generate three popular indices from these bands: Normalized Differenced Vegetation Index, Normalized Differenced Built Index, and Enhanced Vegetation Index. Over the course of a year, each location is observed up to 28 times (cloudiness obscures locations in some areas at some times). We take the 25th, 50th and 75th percentiles of each of the eight bands and three indices, and use those 24 variables as inputs. This feature-engineering strategy mirrors the approach in Hansen et al. (2013). With a simple 1-layer neural network (a logistic regression) we are able to predict forest cover using this data with 75% accuracy.



Bastin et al. (2017) Fig S16b.

Figure 2. : Colored areas show labelled pixels from Bastin (2017) — green for forested and beige for non-forested. Inset shows an example of how percent forested labels were generated from high resolution satellite data.

A. Simulation

All of the following empirical exercises take the following structure. First we divide the data into 3 folds. We use two of these folds to train a standard machine learning model, and predict on the last fold to generate $\widehat{Y}$. We repeat this across all three folds so we have a ground-truth value of Y and a prediction $\widehat{Y}$ for each point.

Then we add the adversarial constraint to a model with the same basic structure and get a new set of predictions using the exact same procedure. Finally, we estimate our regression of interest using the ground truth data, the baseline predictions, and adversarial model predictions. We test both the simple linear regression adversary described above (SLR adversary) as well as a model that simply penalizes the absolute value of $Cov(X, \nu)$ (the correlational adversary). We also test our bias correction method using the baseline model predictions, as well as the multiple imputation method recommended by Proctor, Carleton and Sum (2023). For both sets of predictions we bootstrap standard errors that include the uncertainty from training the model.

In this simulation and the next simple example, we focus on the cross-sectional relationship between roads and forest cover. Previous work has found that roads are an important driver of deforestation (Asher, Garg and Novosad, 2020). Roads and other infrastructure are non-randomly placed, so this cross sectional relationship is likely to generate confounder-induced bias (Figure 1.c). Consider X to be construction of a road, Y to be forest cover, and W to be some omitted geographic variable, like slope, that influences both measurement errors and the placement of roads.

Our first simulation follows this procedure:

1) Draw 20,000 observations of W from a Poisson distribution with shape parameter of 1.

2) Assign each observation $X \sim$ Bernoulli with $p = \max(1 - W/4, 0)$, so that treatment is more likely when W is lower.

3) Assign each observation a random forest cover Y and associated satellite data k from the Bastin points.

4) If $W > 0$, make the satellite data artifically 'greener' without changing

the label. In practice this is done by replacing the satellite data with the satellite data from a different point with a higher percent forest cover.

This gives a true treatment effect of zero, since the forest labels Y are assigned randomly. However the last step mimics a real source of bias — remotely sensed forest cover tends to be overestimated on steeper slopes than on flat land since images are taken from above and tend to capture more trees in a smaller spatial area when on a slope. Because of this bias, and selection into treatment, it will appear that roads are associated with lower forest cover. Note also that there are no "traditional" confounders here — nothing in the simulation is associated with both road proximity and true forest cover.

Figure 3 shows the distribution of coefficient estimates from 100 bootstrapped runs of each of the models and regressions using 10,000 training points. As predicted, using the baseline machine learning model results in a negative and significant estimate of the effect of X on Y, whereas using the ground truth estimates results in a fairly precise null. Furthermore, failing to bootstrap standard errors that include model uncertainty, as is done by every researcher that uses off the shelf satellite data, leads to dramatically overstated precision (the difference between the green and yellow distribtions in Figure 3. Both the bias correction method and the adversarial model result in coefficient distributions correctly centered around 0. Standard errors are fairly wide as they include the uncertainty from the machine learning step.

The performance of all of these methods could depend on the sample size of labelled points. Given this, we also run each of the models using a progressively increasing sample of labelled point. For each given sample size of labelled points J, we use the J labels to train the model, and then predict on the remaining points so that the sample size for the regression is always $N = 20,000$. For each J we bootstrap 100 different versions of the model to estimate standard errors as

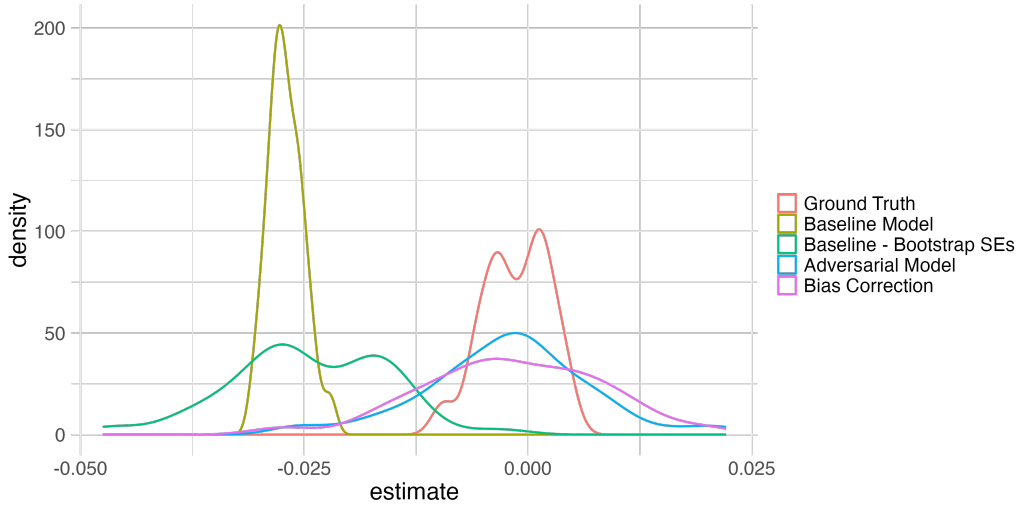


Figure 3. : Estimates from the baseline model vs the adversarial models with 10,000 labeled observations. Each distribution represents the distribution of the coefficients from each model after 100 runs on bootstrapped training data.

well.

Figure 4 shows the results of this exercise. The baseline model learns to measure forest cover with relatively few training observations but generates biased estimates of the relationship between roads and forest cover across all sample sizes. The adversarial models have much higher variances at small sample sizes but are consistently centered on $\beta = 0$, and precision increases as the sample size of labelled points increases. This bias correction method works well at smaller sample sizes, and consistently recovers the true null effect. In this case, using just the ground truth points works best at all sample sizes, since the machine learning labels simply add noise.

This setting should be a difficult case for the adversarial debiaser because the satellite data contains no information about the source of the bias, since it has been replaced by imagery from randomly drawn pixels with higher forest cover. This means that the adversarial model has to sacrifice predictive accuracy in order

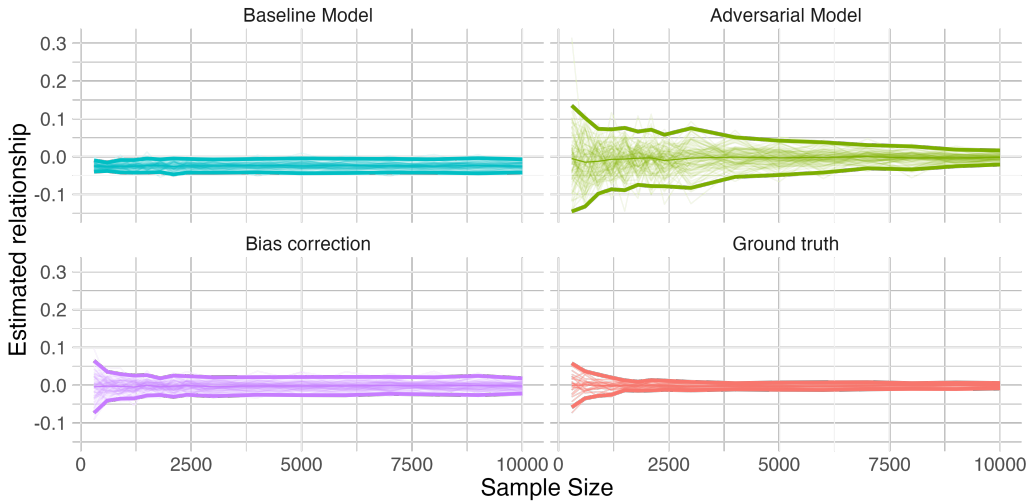


Figure 4. : Estimates from the baseline model vs the adversarial models across sample sizes. Each light colored line is an individual training run where researchers label progressively more observations. The thick lines represent the mean and two standard deviations from the mean of the runs.

to debias the predictions. More specifically, it has to do a worse job predicting the low W, high Y observations so that the measurement error is balanced across X. Note that the adversary never has access to W, yet is still able to adjust for W-induced measurement error.

Finally, we conduct power analyses of our bias test to estimate the minimum detectable bias (MDB) at different sample sizes. This could be crucial for a researcher using an off-the-shelf satellite data source deciding how many points to label in order to rule out large biases in their estimates. The results are presented in Figure 5. Each line in represents a different random draw of points to label. At each sample size, for each set of points, we estimate the standard error of γ and use that to perform a standard power calculation using 0.8 power and 95% statistical significance. Given that the true magnitude of the bias is 0.025 in this simulation, researchers would need to label more than 2,500 points to detect this bias as statistically different from zero 80% of the time. This may

be conservative, however, since in practice, researchers should care more about the confidence intervals around their estimates of bias, which will be large at low sample sizes, rather than whether it is statistically different from zero at the 95% level of confidence.

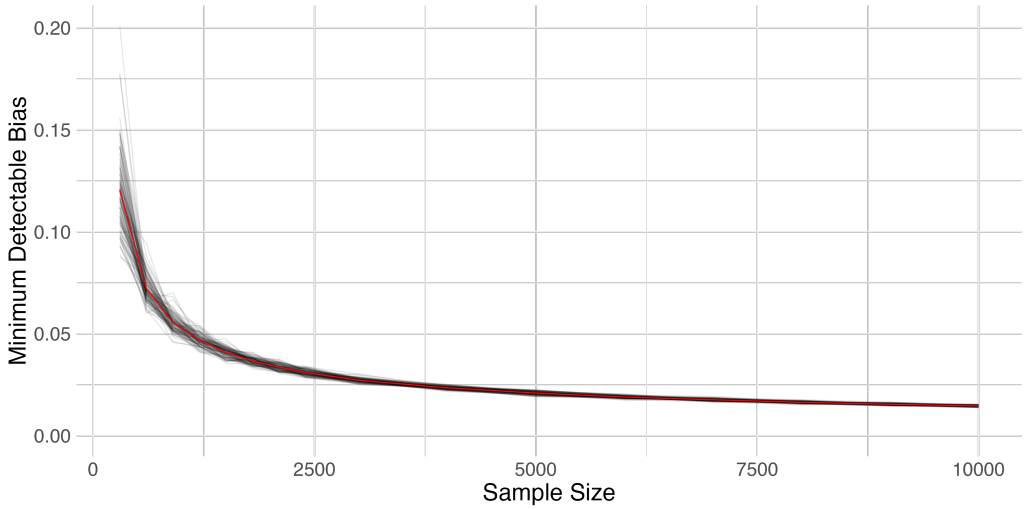


Figure 5. : Minimum detectable bias (MDB) across sample sizes at power of 0.8 and $\alpha = 0.05$. Each black line represents an estimate of MDB using a different random samples of labeled data, red line is the true MDB using standard errors estimated with the whole dataset.

B. Descriptive Exercise: Roads and Forest Cover

Next we use the true Bastin (2017) data, and data on the African road network from Meijer et al. (2018), to estimate the gradient of forest cover with respect to distance to the nearest road. Whereas before the independent variable was binary and the outcome was continuous, now our independent variable is continuous, log distance to the nearest road, and our outcome is binary (forested or not). We apply the same cross-fitting procedure as above for both a standard model and an adversarial model using a 3-layer neural network that gives then estimated

probability of forest cover as output. We then run the same regressions as in Section III.A. The results are shown in Figure 6.

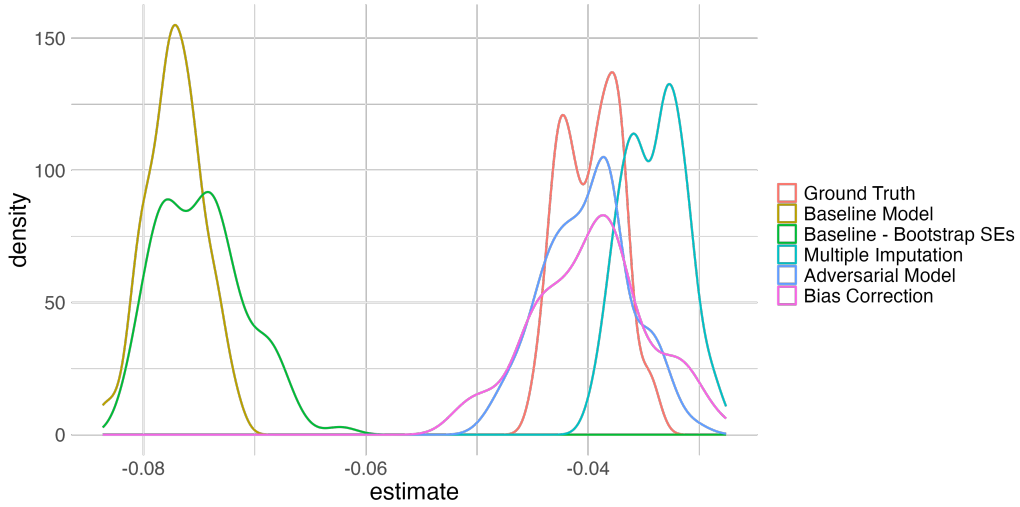


Figure 6. : Estimates from the baseline model vs the adversarial models with 10,000 labeled observations. Each distribution represents the distribution of the coefficients from each model after 100 runs on bootstrapped training data.

Now in a real-world setting, we see that the standard machine-learning model over-estimates the negative relationship between proximity to roads and forest cover. Clearly there are omitted variables in this context – topography, aridity, and others – that influence both the prediction errors, and the location of roads. Once again, however, the adversarial model and the bias correction method are both able to generate measurements that recover the true estimate, without any knowledge of these omitted variables. Furthermore, the standard error of the adversarial model is a bit smaller than the bias correction method, indicating that the model has been able to reduce bias without greatly increasing prediction error.

In Figure 7 we plot the mean measurement error at each decile of distance from a road for both the adversarial model and the standard model. While average error

for both models is close to zero, the standard model exhibits a strong measurement error - distance gradient that results in the biases we see in the regressions. The positive prediction error at close distances indicates the model is more likely to generate false positives (i.e. predict forest where there is no forest) close to roads, and the negative mean prediction error at far distances indicates the model tends to generate more false negatives at that distance. In contrast, the mean error is close to zero at every decile of distance for the adversarial model, indicating a balance of false positives and false negatives at all deciles of distance.

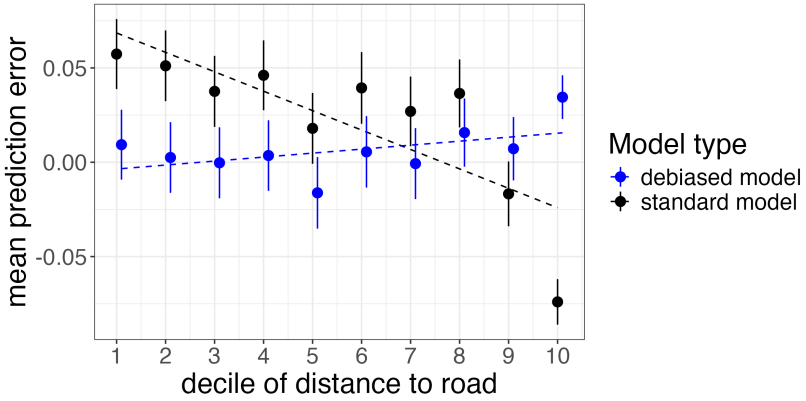


Figure 7. : Measurement error across deciles of distance to road

Figure 6 also shows estimates of the coefficients derived from a multiple imputation approach as recommended by Proctor, Carleton and Sum (2023) in Figure 6. While this approach improves upon the naive estimates, the resulting estimates are statistically different from the ground truth estimates.

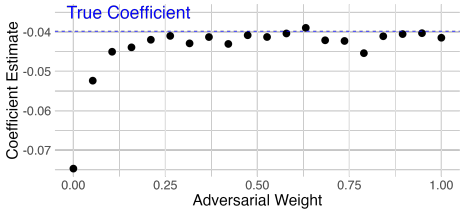
We also use this setting to run experiments on the α parameter: the researcher chosen weight on the adversary in the loss function. The results are summarized in Figure 8 for two different primary prediction models, a logistic regression, and a deep neural net (DNN). The left column shows that increasing the weight on the adversary from zero (standard model) to 1 quickly eliminates bias in the estimated

coefficients. Naively, we might think that this increasing weight would come at the cost of predictive accuracy, as described in the simulation, since an unconstrained model should be able to minimize MSE at least as well as a constrained model. This seems to be the case for the logistic regression, which shows a trend towards higher mean-squared error on predictions as the adversary's weight increases. For the DNN, however, increasing the adversary's weight actually improves prediction accuracy. This result can be understood as a kind of regularization effect. In some settings with many model parameters, for example when using LASSO with many features, it is well known that a regularization penalty can reduce overfitting and improve out-of-sample prediction performance (Tibshirani, 1996). While it is difficult to say precisely when adversarial models will improve prediction accuracy, the adversary seems to be serving a regularization function in this example.

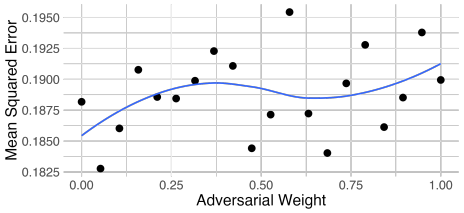
This application shows that estimates of a simple descriptive relationship can be biased by measurement error, but that the measurement error can be eliminated by including an adversarial debiaser as part of the model or by performing a bias correction. Furthermore, when using adversarial debiasing, researchers need not specify, or even be aware of, all of the possible sources of measurement error. In the following section, we study whether these methods remain important to study causal relationships using a research design that can rule out many traditional confounders.

IV. Mining and Deforestation

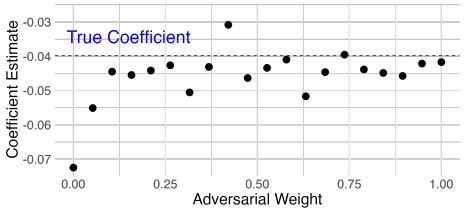
We now turn to a causal relationship of interest - the effect of mining activity on deforestation. Figure 9 shows the sudden increase in gold exports in Burkina Faso, Ghana, and Mali around 2011. In 2010, exports were negligible in all three countries, by 2012, all countries exported more than $1 billion a year. This sudden increase was a result of increased exploration in new gold fields due to rising global



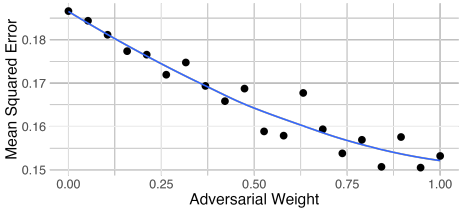
(a) Bias: Logistic Regression Primary Model



(b) Accuracy: Logistic Regression Primary Model



(c) Bias: DNN Primary Model



(d) Accuracy: DNN Primary Model

Figure 8. : Tuning alpha: Tradeoffs between bias and accuracy. Graphs show how coefficient bias (left column) and overall predictive accuracy (MSE - right column) change as the weight on the adversary increases. Top row shows results for a primary model that is a logistic regression. Bottom row shows results for a Deep Neural Net (DNN - 3 layers). Blue lines show Loess smoothed best fit curves.

prices. This sudden and unanticipated increase in gold production makes for a natural experiment that allows us to assess the impact of an extractive industry on deforestation.

While presumably the direct effect of the mines is to displace some forest, in a context where households rely on forest products, it is possible that the indirect effects of increased income from the mines could reduce deforestation through other channels (Foster and Rosenzweig, 2003). Furthermore, Benshaul-Tolonen (2019) shows that the increase in mining activity reduced infant mortality. This sharp tradeoff between environmental degradation and social benefits makes careful measurement important in this context. Giljum et al. (2022) finds a strong relationship between distance to mines and forest cover, but the results are cross-

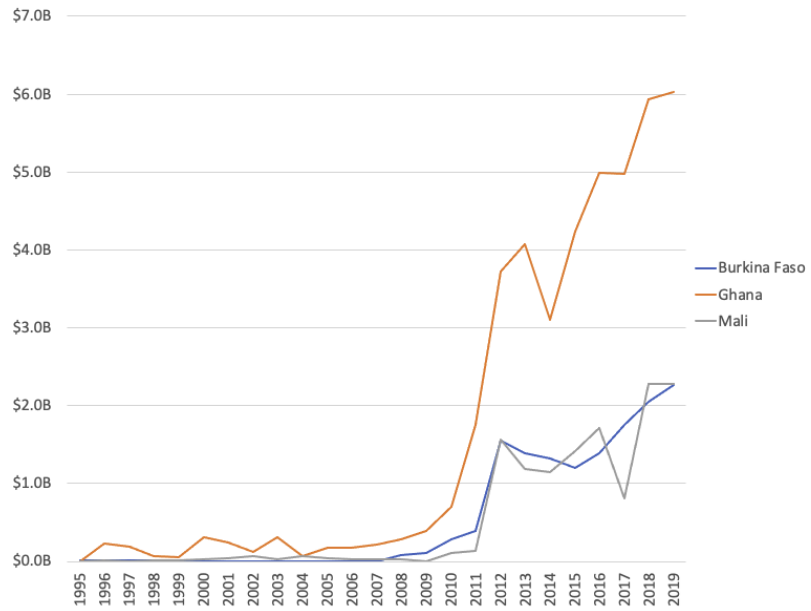


Figure 9. : Gold Exports in Three West-African Countries. Data from UN Comtrade.

sectional and difficult to interpret causally.

A. Data and Empirical Strategy

Since our application is time-series, we rely on the Guo, Zhu and Gong (2022) hand labelled dataset on forest cover change from 2000-2020 instead of the Bastin (2017) data. The Guo, Zhu and Gong (2022) data has fewer points, but it records the status of these points over time, and crucially, whether the forest cover changes during the period of observation. We drop points that change multiple times or have ambiguous forest status, so our results are best interpreted as the effect of mining activity on persistent forest-cover losses. The resulting panel contains 24,003 observations in Africa of 1,143 unique points, 14% of which change from forest to non-forest during the period.

We combine these labelled points with the Hansen et al. (2013) predictions of

forest cover for the same points. Using the Hansen et al. (2013) data instead of generating our own machine learning predictions allows us to evaluate the utility of our methods with satellite data predictions that are commonly used in research practice rather than our overly simple machine learning models. We also pull the Hansen et al. (2013) predictions associated with 250,000 random points within 250 km of a gold mine for every year from 2000 to 2020. This is our unlabelled dataset.

Our data on mines comes from Padilla (2021), and reflects active gold mines in the year 2018, of which there are 190 in Africa. We draw buffers around the mines at various distances, and assign each forest cover observation to the nearest mine and corresponding buffer. We are not aware of datasets that contain the opening dates of mines over time, however given the change in gold exports shown in Figure 9, we expect that most of these mines became active in 2011 or afterwards. Therefore our empirical strategy is a simple difference-in-differences estimator. We estimate the regression:

$$\widehat{Y}_{imbt} = \sum_{b \in [25,50,100]} \beta^b D_i^b \times Post2011_t + \alpha_m + \mu_b + \lambda_t + e_{imbt} \tag{12}$$

where $\widehat{Y}_i mbt$ is the Hansen et al. (2013) prediction of percent forest cover in pixel i, near mine m, buffer b, in year t. We interact dummies D_i^b that denote whether a pixel is within 25, 50, or 100 km from a mine with a $Post2011$ dummy for time periods after 2011. The control group is thus pixels between 100 and 250 km of a mine.

We control for mine, buffer, and year fixed effects. Treatment timing is the same for all units. Therefore the identification assumption is that there are no time-varying confounders associated with being close to a mine and changing

forest cover.

Finally, we then estimate another version of equation 12 in the labelled data set, but we replace the independent variable with $\nu_{imbt} = \widehat{Y}_{imbt} - Y_{imbt}$, the Hansen et al. (2013) prediction errors. The coefficients of this regression are our estimates of bias in equation 12, and can be used to correct those coefficients.

B. Results

Figure 10 plots the map of our study area with the Hansen et al. (2013) prediction errors — there are some clear geographic trends. Purple points show false positives, points predicted to have a high forest cover that actually have no forest cover. These are concentrated in the most densely forested regions of Africa. The surrounding greenery makes it more difficult to identify isolated non-forested pixels in the midst of a forest. Green points show false negatives — points that are predicted not to be forested but are. These are more common in dryland, less forested areas, it seems the Hansen et al. (2013) algorithim has difficulty identifying fragmented patches of forest in otherwise non-forested areas. Note that the areas with many false negatives contain many gold mines (orange circles represent 250 km buffers around the mines on the map).

Level effects in prediction error would not be a concern in our time-series application. If the amount of prediction error was constant over time, this could be captured by fixed effects. What would be problematic is if the presence of mining activity changed the amount of prediction error. This could occur, for example, if mines led to forest fragmentation, which made the remaining forested pixels more difficult to detect. This what we test for in our bias correction regression.

Table 1 shows the results of our analysis. Column 1 shows that, using the Hansen et al. (2013) predictions we find significant changes in deforestation after 2011, especially in areas close to mines. Within the 25 km buffer, forest cover is

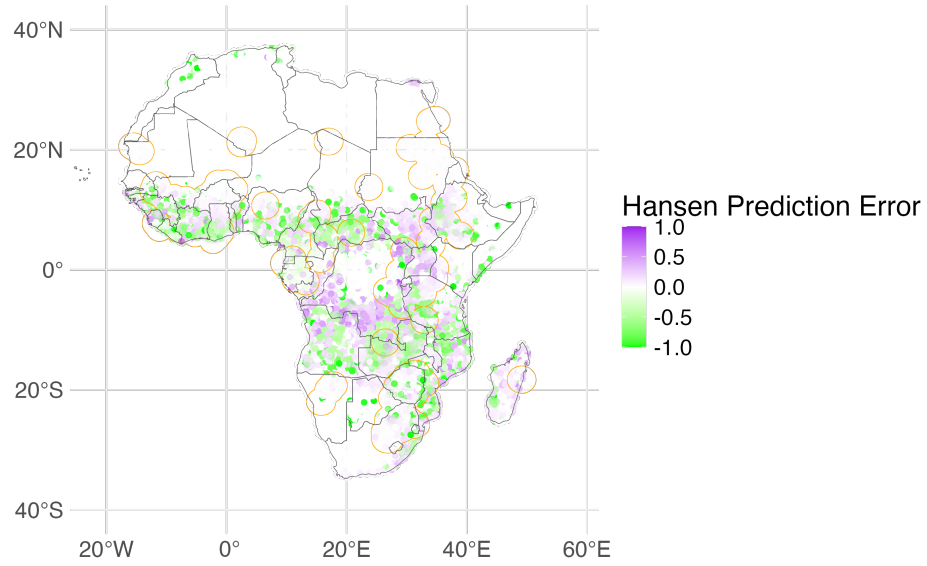


Figure 10. : Prediction Errors in the Hansen Data. Green points are False Negatives. Purple points are false positives. Orange circles show the locations of the 250 km buffers around gold mines.

0.47% lower relative to the 250 km buffer after 2011. Effects are felt as far away as the 150 km buffer, which sees 0.25% less forest cover post 2011 relative to the 250 km buffer.

In column 2, we test for bias in these coefficients by regressing the same set of independent variables and fixed effects on prediction error. Though noisily estimated, due to fewer observations close to mines, we find significant bias in several of our coefficients of interest that changes the story substantially. At the larger buffers (200, 150, and 100 km), we see an increase in false negatives after 2011, though we only find the bias to be significant at the 95% threshold for the 100 km buffer. On the other hand at the 50 km buffer, we find a significant increase in false positives, as if deforestation was underestimated in these areas.

Correcting for this bias in column 3, we see a very different story than in column 1. At larger distances, we see a possible *increase* in forest cover after the gold

	Hansen et al. (2013) Forest Cover (1)	ν (2)	Forest Cover Adjusted (3)
200 km buffer × Post2011	-0.057	-1.731	1.674
	(0.028)	(1.032)	(1.061)
150 km buffer × Post2011	-0.249	-1.832	1.583
	(0.033)	(1.067)	(1.100)
100 km buffer × Post2011	-0.410	-1.497	1.086
	(0.039)	(0.322)	(0.361)
50 km buffer × Post2011	-0.378	2.281	-2.659
	(0.036)	(1.031)	(1.067)
25 km buffer × Post2011	-0.472	-0.697	0.225
	(0.034)	(0.745)	(0.779)
Num. Obs.	4,121,442	23,835	
R2	0.77	0.37	
FEs	Year + Mine + Buffer	Year + Mine + Buffer	
Cluster SEs	Mine + Buffer	Mine + Buffer	

Table 1—: Regression results of equation 12. First two columns show results of regressing buffer x Post2011 dummies on Hansen et al. (2013) predictions and prediction errors respectively. Third column shows the results of correcting the Column 1 estimates for bias, assuming independent standard errors. All regressions contain year, mine and buffer fixed effects and clustered standard errors by mine and buffer.

rush, more consistent with a Foster and Rosenzweig (2003) type story of wealth increases causing an increase in forested areas, though this is only significantly different from zero at the 100 km buffer. Closer to the mines, we find an even larger effect on deforestation, possibly as large as a 2.7% decrease in forested areas. This probably captures the direct effects of deforestation resulting from mine infrastructure.

This heterogeneity is smoothed over by the Hansen et al. (2013) data. Our results seem to show that the Hansen et al. (2013) predictions show important time-varying biases. The underestimates of deforestation at the closer buffers could be a result of predictions that missed forests in the dryland regions of Africa even before the mining booms. The overestimates of deforestation at larger buffers could be due to forest fragmentation making it difficult for the Hansen et al. (2013) model to find the remaining patches of forest cover. Correcting for these biases shows a much more nuanced picture of mining-induced deforestation.

V. Conclusion

Advances in machine learning represent a tremendous opportunity for social science research. Satellite data now makes it possible to measure land use changes at unprecedented scale and resolution. Beyond satellite data, machine learning techniques can be used to measure difficult to quantify concepts from text and other unstructured data. As this field advances, however, researchers need to be wary of the non-classical measurement error generated by these techniques.

In this paper, we demonstrate how non-classical measurement error from machine learning algorithms can bias coefficient estimates. We also demonstrate several general and widely-applicable techniques to test for biases and correct these issues.

We demonstrate the usefulness of these techniques in several simulations and empirical exercises studying forest cover in Africa. We find that across applications, standard machine learning models produce measurements which bias the downstream estimation tasks, and that both the bias correction and adversarial debiasing methods are able to recover the true parameters estimated with ground-truth data.

In addition to the many practical applications of this technique, several theoretical questions present themselves for future work. In particular, the question of how to correct standard errors in regressions using machine learned proxies to account for prediction uncertainty is tremendously important. Furthermore, when choosing which points to label, researcher may be able to choose labels more efficiently than using random sampling. Progress on these questions will allow researchers to be better prepared to exploit improvements in the availability of data and machine learning algorithms for causal research.

REFERENCES

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

Alix-Garcia, Jennifer, and Daniel Millimet. 2022. "Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation." *Journal of the Association of Environmental and Resource Economists*. Publisher: The University of Chicago Press.

Alix-Garcia, Jennifer, Craig McIntosh, Katharine R. E. Sims, and Jarrod R. Welch. 2013. "The Ecological Footprint of Poverty Alleviation: Evidence from Mexico's Oportunidades Program." *The Review of Economics and Statistics*, 95(2): 417–435.

Asher, Sam, Teevrat Garg, and Paul Novosad. 2020. "The Ecological Impact of Transportation Infrastructure." *The Economic Journal*, 130(629): 1173–1199.

Balboni, Claire, Aaron Berman, Robin Burgess, and Benjamin Olken. 2022. "The Economics of Tropical Deforestation." *Working Paper*.

Bastin, Jean-François et al. 2017. "The extent of forest in dryland biomes." *Science*, 356(6338): 635–638. Publisher: American Association for the Advancement of Science.

Benshaul-Tolonen, Anja. 2019. "Local Industrial Shocks and Infant Mortality." *The Economic Journal*, 129(620): 1561–1592.

Bluhm, Richard, and Gordon C. McCord. 2022. "What Can We Learn from Nighttime Lights for Small Geographies? Measurement Errors and Heterogeneous Elasticities." *Remote Sensing*, 14(5): 1190. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Burgess, Robin, Matthew Hansen, Benjamin A. Olken, Peter Potapov, and Stefanie Sieber. 2012. "The Political Economy of Deforestation in the Tropics*." *The Quarterly Journal of Economics*, 127(4): 1707–1754.

Chernozhukov, Victor, Whitney Newey, Rahul Singh, and Vasilis Syrgkanis. 2020. "Adversarial Estimation of Riesz Representers." arXiv:2101.00009 [cs, econ, stat].

Fong, Christian, and Matthew Tyler. 2021. "Machine Learning Predictions as Regression Covariates." *Political Analysis*, 29(4): 467–484. Publisher: Cambridge University Press.

Foster, Andrew D., and Mark R. Rosenzweig. 2003. "Economic Growth and the Rise of Forests." *The Quarterly Journal of Economics*, 118(2): 601–637. Publisher: Oxford University Press.

Fowlie, Meredith, Edward Rubin, and Reed Walker. 2019. "Bringing Satellite-Based Air Quality Estimates Down to Earth." *AEA Papers and Proceedings*, 109: 283–288.

Gibson, John, Susan Olivia, Geua Boe-Gibson, and Chao Li. 2021. "Which night lights data should we use in economics, and where?" *Journal of Development Economics*, 149: 102602.

Giljum, Stefan, Victor Maus, Nikolas Kuschnig, Sebastian Luckeneder, Michael Tost, Laura J. Sonter, and Anthony J. Bebbington. 2022. "A pantropical assessment of deforestation caused by industrial mining." *Proceedings of the National Academy of Sciences*, 119(38): e2118273119. Publisher: Proceedings of the National Academy of Sciences.

Guo, Jing, Zhiliang Zhu, and Peng Gong. 2022. "A global forest reference set with time series annual change information from 2000 to 2020." *International Journal of Remote Sensing*, 43(9): 3152–3162.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science*, 342(6160): 850–853. Publisher: American Association for the Advancement of Science.

Henderson, Vernon, Adam Storeygard, and David N. Weil. 2011. "A Bright Idea for Measuring Economic Growth." *American Economic Review*, 101(3): 194–199.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *The Review of Economics and Statistics*, 86(1): 4–29.

Jack, B. Kelsey, Seema Jayachandran, Namrata Kala, and Rohini Pande. 2022. "Money (Not) to Burn: Payments for Ecosystem Services to Reduce Crop Residue Burning."

Jain, Meha. 2020. "The Benefits and Pitfalls of Using Satellite Data for Causal Inference." *Review of Environmental Economics and Policy*, 14(1): 157–169. Publisher: Oxford Academic.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." *AEA Papers and Proceedings*, 108: 22–27.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." arXiv:1609.05807 [cs, stat].

Liang, Annie, Jay Lu, and Xiaosheng Mu. 2023. "Algorithm Design: A Fairness-Accuracy Frontier." arXiv:2112.09975 [econ].

Meijer, Johan R., Mark A. J. Huijbregts, Kees C. G. J. Schotten, and Aafke M. Schipper. 2018. "Global Patterns of Current and Future Road Infrastructure." *Environmental Research Letters*, 13(6): 064006.

Meng, Jun, Chi Li, Randall V. Martin, Aaron van Donkelaar, Perry Hystad, and Michael Brauer. 2019. "Estimated Long-Term (1981–2016) Concentrations of Ambient Fine Particulate Matter across North America from Chemical Transport Modeling, Satellite Remote Sensing, and Ground-Based Measurements." *Environmental Science & Technology*, 53(9): 5071–5079. Publisher: American Chemical Society.

Padilla, Abraham D et al. 2021. "Compilation of Geospatial Data (GIS) for the Mineral Industries and Related Infrastructure of Africa." Type: dataset.

Proctor, Jonathan, Tamma Carleton, and Sandy Sum. 2023. "Parameter Recovery Using Remotely Sensed Variables." *NBER Working Paper*.

Ratledge, Nathan, Gabriel Cadamuro, Brandon De la Cuesta, Matthieu Stigler, and Marshall Burke. 2021. "Using Satellite Imagery and Machine Learning to Estimate the Livelihood Impact of Electricity Access."

Sanford, Luke. 2021. "Democratization, Elections, and Public Goods: The Evidence from Deforestation." *American Journal of Political Science*, n/a(n/a).

Slough, Tara et al. 2021. "Adoption of community monitoring improves common pool resource management across contexts." *Proceedings of the National Academy of Sciences*, 118(29): e2015367118. Publisher: Proceedings of the National Academy of Sciences.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1): 267–288. Publisher: [Royal Statistical Society, Wiley].

Torchiana, Adrian L., Ted Rosenbaum, Paul T. Scott, and Eduardo Souza-Rodrigues. 2023. "Improving Estimates of Transitions from Satellite Data: A Hidden Markov Model Approach." *The Review of Economics and Statistics*, 1–45.

Tropek, Robert, Ondřej Sedláček, Jan Beck, Petr Keil, Zuzana Musilová, Irena Šímová, and David Storch. 2014. "Comment on "High-resolution global maps of 21st-century forest cover change"." *Science*, 344(6187): 981–981. Publisher: American Association for the Advancement of Science.

Wren-Lewis, Liam, Luis Becerra-Valbuena, and Kenneth Houngbedji. 2020. "Formalizing land rights can reduce forest loss: Experimental evidence from Benin." *Science Advances*, 6(26): eabb6914.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. "Mitigating Unwanted Biases with Adversarial Learning." arXiv:1801.07593 [cs].

Zhang, Han. 2021. "How Using Machine Learning Classification as a Variable in Regression Leads to Attenuation Bias and What to Do About It." SocArXiv preprint.

Appendix A. Proof that adversarial debiasing penalizes the absolute value of γ

Define $P = X(X'X)^{-1}X'$ as the n x n symmetric and idempotent projection matrix, and $\mathbb{I}$ as the n x n identity matrix. The Adversary's loss function is:

$$\begin{aligned} L_a &= (\nu - P\nu)'(\nu - P\nu) \\ &= v'(\mathbb{I} - P)'(\mathbb{I} - P)\nu = v'(\mathbb{I} - P)\nu \end{aligned} \tag{A1}$$

by the properties of the projection matrix. Take two different vectors of prediction errors, ν and $\tilde{\nu}$, such that $|\gamma| = |(X'X)^{-1}X'\nu| > |\tilde{\gamma}| = |(X'X)^{-1}X'\tilde{\nu}|$. The difference in L_a for these two vectors is:

$$\nu'\nu - \nu'P\nu - [\tilde{\nu}'\tilde{\nu} - \tilde{\nu}'P\tilde{\nu}]. \tag{A2}$$

Assume $\nu'\nu = \tilde{\nu}'\tilde{\nu}$, i.e. the overall prediction error is the same. Then we want that the primary model will choose $\tilde{\nu}$, since $|\tilde{\gamma}|$ is smaller. Since we are minimizing $L_p(\cdot) - \alpha L_a(\cdot)$, we want that $L_a(\nu) < L_a(\tilde{\nu}) \iff \tilde{\nu}'P\tilde{\nu} < \nu'P\nu$. We know:

$$\begin{aligned} |(X'X)^{-1}X'\nu| > |(X'X)^{-1}X'\tilde{\nu}| \iff \\ \nu'X(X'X)^{-1}(X'X)^{-1}X'\nu > \tilde{\nu}'X(X'X)^{-1}(X'X)^{-1}X'\tilde{\nu} \end{aligned} \tag{A3}$$

Since X is univariate, both sides are scalars. Multiply both sides by $X'X$ which is a positive scalar, maintaining the inequality:

$$\begin{aligned} \nu'X(X'X)^{-1}(X'X)(X'X)^{-1}X'\nu > \tilde{\nu}'X(X'X)^{-1}(X'X)(X'X)^{-1}X'\tilde{\nu} \\ \nu'P'P\nu > \tilde{\nu}'P'P\tilde{\nu} \\ \nu'P\nu > \tilde{\nu}'P\tilde{\nu} \end{aligned}$$

Concluding the proof.

Appendix B. Debiasing with Control Variables and Instruments

Adding Control Variables

Assume we want to estimate β_1 in the regression

$$\widehat{Y}_i = \beta_1 x_1 + x_2\beta_2 + e_i \tag{B1}$$

where x_1 is an n x 1 vector of the treatment variable, and x_2 is an n x k matrix of control variables. By the Frisch-Waugh-Lovell theorem, we can write $\widehat{\beta}_1$ as:

$$\widehat{\beta}_1 = (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{\widehat{Y}} \tag{B2}$$

where $\tilde{X}_1$ are the residuals of the regression of X_1 on X_2, and $\tilde{\widehat{Y}}$ are the residuals of the regression of $\widehat{Y}$ on X_2. $\widehat{\beta}_1$ can thus be rewritten as follows:

$$\begin{aligned}\widehat{\beta}_1 &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')(Y+\nu) \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')Y + \\ &\quad (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'(\mathbb{I} - X_2(X_2'X_2)^{-1}X_2')\nu \\ &= (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{Y} + (\tilde{X}_1'\tilde{X}_1)^{-1}\tilde{X}_1'\tilde{\nu}\end{aligned} \tag{B3}$$

where $\tilde{\nu}$ are the residuals of the regression of ν on X_2. The expectation of this estimate is

$$\mathbb{E}[\beta_1] = \beta_1 + \frac{cov(\tilde{X}_1, \tilde{\nu})}{var(\tilde{X}_1)}. \tag{B4}$$

Intuitively this makes sense – if the residual variation in X_1 is correlated with the residual prediction error, after controlling for X_2 in both cases, our estimate will be biased. Thus following the same logic as above, we can make the adversary a linear regression of $\tilde{\nu}$ on $\tilde{X}_1$.

A similar argument can be extended to the instrumental variables case with controls. Following the two-stage least squares estimation procedure, we first use covariates X_2 and instruments Z to predict X_1:

$$\widehat{X}_1^{IV} = C(C'C)^{-1}C'X_1 \tag{B5}$$

where $C = [1, X_2, Z]$. Then, we regress the outcome against the predicted values of X_1 and the covariates X_2 and take the estimated coefficient for $\widehat{X}_1^{IV}$ in this second stage regression as our estimate of the true β_1. By the Frisch-Waugh-Lovell theorem, we have

$$\begin{aligned} \widehat{\beta}_1^{2SLS} &= (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{\widehat{Y}} \\ &= (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{Y} + (\tilde{\widehat{X}}_1{}'\tilde{\widehat{X}}_1)^{-1}\tilde{\widehat{X}}_1{}'\tilde{\nu} \end{aligned} \tag{B6}$$

where $\widetilde{\widehat{X}_1}$ are the residuals from regressing $\widehat{X}_1^{IV}$ on X_2, $\tilde{Y}$ are the residuals from regressing Y on X_2, and $\tilde{\nu}$ are the residuals from regressing ν on X_2.

In the case of a single instrument Z, this estimator of β_1 can be rewritten more simply following the indirect least-squares procedure. For this approach, we perform linear regressions for the models

$$\widehat{Y}_i = \gamma_0 + \gamma_1 w_i + \gamma_2' X_{2i} + w_i; \tag{B7}$$

$$\widehat{X}_{1i} = \alpha_0 + \alpha_1 Z_i + \alpha_2' X_{2i} + u_i \tag{B8}$$

This produces the following estimate of β_1, which coincides with $\widehat{\beta}_1^{2SLS}$ for this special case:

(B9)
$$\widehat{\beta}_1^{ILS} = \frac{\widehat{\gamma_1}}{\widehat{\alpha_1}} = \frac{(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\hat{\tilde{Y}}}{(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'\tilde{X_1}} = \frac{cov(\tilde{Z}, \hat{\tilde{Y}})}{cov(\tilde{Z}, \tilde{X_1})}$$
$$= \frac{cov(\tilde{Z}, \tilde{Y})}{cov(\tilde{Z}, \tilde{X_1})} + \frac{cov(\tilde{Z}, \tilde{\nu})}{cov(\tilde{Z}, \tilde{X_1})}$$

In the above expressions for $\widehat{\beta}_1^{2SLS}$ and $\widehat{\beta}_1^{ILS}$, we see that the coefficient estimates are the sum of the coefficient estimate that we would obtain from performing these procedures given Y without measurement error - which is consistent for β_1 given IV assumptions - and an additional bias term involving the measurement error ν. To minimize this bias, we propose an adversary in the form of a regression of $\tilde{\nu}$ on $\tilde{Z}$ for the single instrument case, or $\tilde{\nu}$ on $\hat{\tilde{X}}_1$ more generally.