Appendix A: Proofs of Theoretical Results

THEOREM 1. *Under Assumptions 1-2, for any two trees in the random forest, i and j ($i \neq j$),*

$$\lim_{n\to\infty} Cov(\widehat{X}^{(j)}, e^{(i)}) = 0.$$

Before proving the theorem, we first establish the following lemma that will be used later.

LEMMA 1. *Suppose that $\{X_n\}$ and $\{Y_n\}$ are two sequences of random variables and $X_n \xrightarrow{L^2} X$ and $Y_n \xrightarrow{L^2} Y$, where $\xrightarrow{L^2}$ means L^2 convergence (i.e., converge in quadratic mean). Furthermore, assume that $\{X_n\}$, $\{Y_n\}$, X, and Y all have finite expectations and variances. Then $\lim_{n\to\infty} Cov(X_n, Y_n) = Cov(X, Y)$.*

This lemma can be proved as follows. First, because L^2 convergence implies L^1 convergence, we know that $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$ and $\lim_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(Y)$. Next,

$$\begin{aligned}
\lim_{n\to\infty} Cov(X_n, Y_n) &= \lim_{n\to\infty} [\mathbb{E}(X_n Y_n) - \mathbb{E}(X_n)\mathbb{E}(Y_n)] \\
&= \lim_{n\to\infty} [\mathbb{E}(X_n Y_n)] - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \lim_{n\to\infty} [\mathbb{E}(X_n Y_n) - \mathbb{E}(XY)] + \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \\
&= \lim_{n\to\infty} [\mathbb{E}(X_n Y_n) - \mathbb{E}(XY)] + Cov(XY) \\
&= \lim_{n\to\infty} [\mathbb{E}(X_n Y_n) - \mathbb{E}(X_n Y) + \mathbb{E}(X_n Y) - \mathbb{E}(XY)] + Cov(XY) \\
&= \lim_{n\to\infty} [\mathbb{E}X_n(Y_n - Y)] + \lim_{n\to\infty} [\mathbb{E}Y(X_n - X)] + Cov(XY)
\end{aligned}$$

By Cauchy-Schwarz Inequality, we know that $|\mathbb{E}X_n(Y_n - Y)| \leq \sqrt{\mathbb{E}(X_n^2)\mathbb{E}(Y_n - Y)^2}$, and because $Y_n \xrightarrow{L^2} Y$ we have $\lim_{n\to\infty} \mathbb{E}(Y_n - Y)^2 = 0$. Combined with the fact that X_n have finite expectations and variances, we know $\lim_{n\to\infty} |\mathbb{E}X_n(Y_n - Y)| \leq 0 \Rightarrow \lim_{n\to\infty} \mathbb{E}X_n(Y_n - Y) = 0$. The same derivation (using L^2 convergence of X_n) would show that $\lim_{n\to\infty} \mathbb{E}Y(X_n - X) = 0$. Taken together, it follows that $\lim_{n\to\infty} Cov(X_n, Y_n) = Cov(X, Y)$.

Next, we return to the proof of our main theorem.

Under Assumptions 1-2, Theorem 1 of Scornet et al. (2015) shows that a random forest is L^2 consistent, and also note that a single tree in the random forest (i.e., a tree grown based on the CART algorithm, considering a random subset of all features at each split and using a random subsample of training data) is also L^2 consistent.[9] Therefore, for an arbitrary tree i in the forest, we have $\widehat{X}^{(i)} \xrightarrow{L^2} \mathbb{E}(X|\boldsymbol{f}) = \sum_{k=1}^{p} m_k(f_k)$. It follows that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}(\widehat{X}^{(i)} - \sum_{k=1}^{p} m_k(f_k))^2 = 0 \\
\Rightarrow &\lim_{n\to\infty} \mathbb{E}(X - e^{(i)} - \sum_{k=1}^{p} m_k(f_k))^2 = 0 \\
\Rightarrow &\lim_{n\to\infty} \mathbb{E}(\zeta - e^{(i)})^2 = 0
\end{aligned}$$

[9] In (Scornet et al. 2015, p. 1722), the authors state that "an easy adaptation of Theorem 1 shows that the CART algorithm is consistent under the same assumptions." This is because their proof of their Theorem 1 is independent of, and therefore holds for any values of, forest-specific parameters (including number of trees, number of features attempted at each split, and the size of the subsample employed by each tree). Consequently, their result is also applicable to any given tree in the forest. Consistency of trees has also been shown in other work (e.g., Györfi et al. (2006), Biau et al. (2008)) and discussed in Biau and Scornet (2016). We further note that tree consistency requires the number of training data points in a leaf node to go to infinity as $n \to \infty$, which is accommodated by Assumption 2. Instead, if the trees are fully grown (i.e., each leaf node contains only one training data point), then individual trees are inconsistent while the forest can remain consistent (Scornet et al. 2015, Theorem 2).

or equivalently, $e^{(i)} \xrightarrow{L^2} \zeta$. Finally, for two trees (i,j) in the random forest, using the lemma above, we have $\lim_{n\to\infty} Cov(\widehat{X}^{(j)}, e^{(i)}) = Cov(\sum_{k=1}^{p} m_k(f_k), \zeta) = 0$, because ζ is independent of $\sum_{k=1}^{p} m_k(f_k)$ by Assumption 1.

THEOREM 2. *The error rate of a random forest binary classifier decreases with* $\mathbb{E}_j\mathbb{E}_i Corr(|e^{(i)}|, |e^{(j)}|)$, *where* $e^{(i)}$ *and* $e^{(j)}$ *are prediction errors of tree* i *and tree* j $(i \neq j)$.

Breiman (2001) proves that the error rate of a random forest decreases with

$$\mathbb{E}_j\mathbb{E}_i \left[Corr(rmg(\widehat{X}^{(i)}), rmg(\widehat{X}^{(j)})) \right]$$

where $rmg(\widehat{X}^{(i)})$ represents the *raw marginal function* of tree i's predictions. Under binary classification, the raw marginal function is defined as $rmg(\widehat{X}^{(i)}) = \mathbb{I}(\widehat{X}^{(i)} = X) - \mathbb{I}(\widehat{X}^{(i)} \neq X)$, where $\mathbb{I}$ is an indicator function that checks vectors $\widehat{X}^{(i)}$ and X element-wise, and takes value 1 if the enclosed relationship is true or 0 otherwise. In other words, $\mathbb{I}(\widehat{X}^{(i)} = X)$ is a vector where correct predictions are marked with 1, and $\mathbb{I}(\widehat{X}^{(i)} \neq X)$ is a vector where incorrect predictions are marked with 1. Denote $\mathbf{1} = (1, \ldots, 1)$ as a vector of 1 with the same length as the vector of predictions. Clearly, we have $\mathbb{I}(\widehat{X}^{(i)} = X) = \mathbf{1} - \mathbb{I}(\widehat{X}^{(i)} \neq X)$, and $\mathbb{I}(\widehat{X}^{(i)} \neq X) = |e^{(i)}|$. Therefore, we know $Corr((rmg(\widehat{X}^{(i)}), rmg(\widehat{X}^{(j)})) = Corr(\mathbb{I}(\widehat{X}^{(i)} = X) - \mathbb{I}(\widehat{X}^{(i)} \neq X), \mathbb{I}(\widehat{X}^{(j)} = X) - \mathbb{I}(\widehat{X}^{(j)} \neq X)) = Corr(\mathbf{1} - 2\mathbb{I}(\widehat{X}^{(i)} \neq X), \mathbf{1} - 2\mathbb{I}(\widehat{X}^{(j)} \neq X)) = Corr(\mathbb{I}(\widehat{X}^{(i)} \neq X), \mathbb{I}(\widehat{X}^{(j)} \neq X)) = Corr(|e^{(i)}|, |e^{(j)}|)$.

THEOREM 3. *For all* $i \in \{1, \ldots, M\}$, $Cov(e^{(i)}, X) < 0$.

We prove this theorem for a given sample with size N. For notational simplicity, we write the ground truth as $X = \{a_k\}_{k=1}^{N}$, and similarly write the prediction vector and error vector of tree i as $\widehat{X}^{(i)} = \{p_{ik}\}_{k=1}^{N}$, $e^{(i)} = \{e_{ik}\}_{k=1}^{N}$. Suppose the number of data points where $a_k = \alpha$ and $p_{ik} = \beta$ is $n_{\alpha\beta}$ $(\alpha, \beta \in \{0,1\})$. Clearly, $n_{00} + n_{01} + n_{10} + n_{11} = N$, and the relationship between X and $e^{(i)}$ is fully described as follows:

- There are n_{00} data points where $a_k = 0$ and $e_{ik} = 0$;
- There are n_{01} data points where $a_k = 0$ and $e_{ik} = 1$;
- There are n_{10} data points where $a_k = 1$ and $e_{ik} = -1$;
- There are n_{11} data points where $a_k = 1$ and $e_{ik} = 0$.

Next, we write $Cov(e^{(i)}, X) = \frac{1}{N^2}(N\sum e_{ik}a_k - \sum e_{ik}\sum a_k)$. Note that $\sum e_{ik}a_k = -n_{10}$, $\sum e_{ik} = n_{01} - n_{10}$, and $\sum a_k = n_{10} + n_{11}$. Therefore, we have $N\sum e_{ik}a_k - \sum e_{ik}\sum a_k = -(n_{00} + n_{01} + n_{10} + n_{11})n_{10} - (n_{01} - n_{10})(n_{10} + n_{11}) = -n_{00}n_{10} - 2n_{01}n_{10} - n_{01}n_{11} < 0$, and accordingly, $Cov(e^{(i)}, X) < 0$.

THEOREM 4. *For all* $i, j \in \{1, \ldots, M\}$ *and* $i \neq j$, $Cov(e^{(i)}, e^{(j)}) > 0$ *if and only if* $(p_{000}+p_{111})(p_{011}+p_{100}) + 2(p_{0\bullet\bullet} - p_{000})p_{100} + 2(p_{1\bullet\bullet} - p_{111})p_{011} + (p_{010} - p_{101})(p_{110} - p_{001}) > 0$.

Similarly, we prove this theorem for a given sample with size N. Again, we write the ground truth as $X = \{a_k\}_{k=1}^{N}$, and write the prediction vector and error vector of tree i as $\widehat{X}^{(i)} = \{p_{ik}\}_{k=1}^{N}$, $e^{(i)} = \{e_{ik}\}_{k=1}^{N}$. First, we lay out all possible value combinations of $a_k, p_{ik}, p_{jk}, e_{ik}, e_{jk}$ in the following table:

Next, write $Cov(e^{(i)}, e^{(j)}) = \frac{1}{N^2}(N\sum e_{ik}e_{jk} - \sum e_{ik}\sum e_{jk})$. Note that $\sum e_{ik}e_{jk} = n_4 + n_5$, $\sum e_{ik} = (n_2 + n_4) - (n_5 + n_7)$, and $\sum e_{jk} = (n_3 + n_4) - (n_5 + n_6)$. Then, $N\sum e_{ik}e_{jk} - \sum e_{ik}\sum e_{jk} = (n_1 + \cdots + n_8)(n_4 +$

a_k	p_{ik}	p_{jk}	Count	e_{ik}	e_{jk}	Abbr. Count Notation
0	0	0	$n_{000} = N \times p_{000}$	0	0	n_1
0	1	0	$n_{010} = N \times p_{010}$	1	0	n_2
0	0	1	$n_{001} = N \times p_{001}$	0	1	n_3
0	1	1	$n_{011} = N \times p_{011}$	1	1	n_4
1	0	0	$n_{100} = N \times p_{100}$	-1	-1	n_5
1	1	0	$n_{110} = N \times p_{110}$	0	-1	n_6
1	0	1	$n_{101} = N \times p_{101}$	-1	0	n_7
1	1	1	$n_{111} = N \times p_{111}$	0	0	n_8

$n_5) - [(n_2+n_4) - (n_5+n_7)][(n_3+n_4) - (n_5+n_6)]$. Denote $A = (n_1 + \cdots + n_8)(n_4+n_5)$ and $B = [(n_2+n_4) - (n_5+n_7)][(n_3+n_4) - (n_5+n_6)]$, we calculate the two quantities separately as follows.

First, we rewrite

$$\begin{aligned} A &= (n_1 + \cdots + n_8)n_4 + (n_1 + \cdots + n_8)n_5 \\ &= (n_1+n_8)(n_4+n_5) + (n_2+n_3+n_4)n_4 + (n_5+n_6+n_7)n_4 + (n_2+n_3+n_4)n_5 + (n_5+n_6+n_7)n_5 \end{aligned}$$

Second, we rewrite

$$\begin{aligned} B &= (n_2n_3 + n_2n_4 + n_3n_4 + n_4^2) + (n_6n_7 + n_5n_6 + n_5n_7 + n_5^2) \\ &\quad - (n_2n_5 + n_4n_5 + n_2n_6 + n_4n_6) - (n_3n_5 + n_3n_7 + n_4n_5 + n_4n_7) \\ &= (n_2n_3 + n_6n_7 - n_2n_6 - n_3n_7) + (n_2+n_3+n_4)n_4 + (n_5+n_6+n_7)n_5 \\ &\quad - (n_5+n_6+n_7)n_4 - (n_2+n_3+n_4)n_5 \end{aligned}$$

Together, we have $N\sum e_{ik}e_{jk} - \sum e_{ik}\sum e_{jk} = A - B = (n_1+n_8)(n_4+n_5) + 2(n_5+n_6+n_7)n_4 + 2(n_2+n_3+n_4)n_5 + (n_2n_6 + n_3n_7 - n_2n_3 - n_6n_7) = (n_1+n_8)(n_4+n_5) + 2(n_5+n_6+n_7)n_4 + 2(n_2+n_3+n_4)n_5 + (n_2 - n_7)(n_6 - n_3)$. Using the original count notations, the right-hand-side is equivalent to $(n_{000}+n_{111})(n_{011}+n_{100}) + 2(n_{0\bullet\bullet} - n_{000})n_{100} + 2(n_{1\bullet\bullet} - n_{111})n_{011} + (n_{010} - n_{101})(n_{110} - n_{001})$. Therefore, we have $Cov(e^{(i)}, e^{(j)}) > 0 \Leftrightarrow \frac{1}{N^2}[(n_{000}+n_{111})(n_{011}+n_{100}) + 2(n_{0\bullet\bullet} - n_{000})n_{100} + 2(n_{1\bullet\bullet} - n_{111})n_{011} + (n_{010} - n_{101})(n_{110} - n_{001})] > 0 \Leftrightarrow (p_{000}+p_{111})(p_{011}+p_{100}) + 2(p_{0\bullet\bullet} - p_{000})p_{100} + 2(p_{1\bullet\bullet} - p_{111})p_{011} + (p_{010} - p_{101})(p_{110} - p_{001}) > 0$.

Appendix B: Formal Theoretical Setup

In this appendix, we provide a formal setup of the measurement error problem as well as the instrumental variable approach to address it. We begin by providing a standard setup of the econometric estimation problem. We assume that in a certain population, the relationship of interest is captured by

$$\boldsymbol{Y} = \boldsymbol{X}\beta_X + \boldsymbol{Z}\beta_Z + \varepsilon, \tag{3}$$

where ε is the random error term. We observe $\{(y_i, x_i, z_i)_{i=1,\ldots,n_1}\}$, a sample of n_1 independent and identically distributed units from the population of interest, such that

$$\begin{aligned} &y_i = x_i\beta_X + z_i\beta_Z + \varepsilon_i \quad \text{for } i = 1, \ldots, n_1 \\ &\boldsymbol{Y} = \boldsymbol{X}\beta_X + \boldsymbol{Z}\beta_Z + \boldsymbol{\varepsilon} \quad \boldsymbol{Z} \in \mathbb{R}^{n_1 \times k} \text{ and } \boldsymbol{Y}, \boldsymbol{\varepsilon}, \boldsymbol{X} \in \mathbb{R}^{n_1 \times 1} \end{aligned} \tag{4}$$

Moreover, we will let $\beta = [\beta_X, \beta_Z]$, $\boldsymbol{A} = [\boldsymbol{X}, \boldsymbol{Z}]$, and for simplicity assume that the $k+1$ explanatory variables represented by $\boldsymbol{A}$ have zero mean, in addition to the following standard linear regression assumptions:

(Condition A1) $\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{A}] = 0$,

(Condition A2) $rank(\mathbb{E}[\boldsymbol{A'A}]) = k+1$.

With (A1) and (A2), we have that the Ordinary Least Squares (OLS) estimator $\hat{\beta}_{OLS} = (\boldsymbol{A'A})^{-1}\boldsymbol{A'Y}$ is unbiased and consistent with respect to β.

In our context, $\hat{\beta}_{OLS}$ can be estimated using the labeled data (i.e., $n_1 \equiv |D_{label}|$). In addition, we are able to observe another sample $\{(y_i, \widehat{x_i}, z_i)_{i=n_1+1,\ldots,n_1+n_2}\}$ of n_2 independent and identically distributed units from the population of interest, where $\widehat{x_i}$ (i.e., vector $\widehat{\boldsymbol{X}}$) is an imperfect measurement of x_i (i.e., vector $\boldsymbol{X}$). In our context, $\boldsymbol{X}$ represents the true values of the data-mined variable, and $\widehat{\boldsymbol{X}}$ represents imperfect predictions generated by the machine learning model. For n_1 samples, we obtain the true value of $\boldsymbol{X}$ (i.e., the ground truth labels), but it is prohibitive (e.g., in cost or time) to obtain labels for the remaining n_2 samples (i.e., $n_2 \equiv |D_{unlabel}|$). Typically, $n_2 >> n_1$. Given its large size, there is a clear desire to utilize the information contained in the n_2 samples for inference. We make the following additional assumptions about the imperfect measurement, $\widehat{\boldsymbol{X}}$:

(Condition A3) $\widehat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{e}$,
(Condition A4) $\mathbb{E}[\boldsymbol{\varepsilon' e}] = \boldsymbol{0}$,
(Condition A5) $\mathbb{E}[\boldsymbol{X'e}] = \boldsymbol{0}$, and $\mathbb{E}[\boldsymbol{Z'e}] = \boldsymbol{0}$

Attempting to estimate (4) simply by replacing $\boldsymbol{X}$ with $\widehat{\boldsymbol{X}}$ is known to result in a biased and inconsistent estimate of $\hat{\beta}_{OLS}$ because $\widehat{\boldsymbol{X}}$ is endogenous:

$$\begin{aligned} &y_i = \widehat{x_i}\beta_X + z_i\beta_Z + [\varepsilon_i - e_i\beta_X] \quad \text{for } i = n_1+1,\ldots,n_1+n_2, \\ &\boldsymbol{Y} = \widehat{\boldsymbol{X}}\beta_X + \boldsymbol{Z}\beta_Z + [\boldsymbol{\varepsilon} - \boldsymbol{e}\beta_X] \quad \boldsymbol{Z} \in \mathbb{R}^{n_2 \times k} \text{ and } \boldsymbol{Y}, \boldsymbol{\varepsilon}, \boldsymbol{e}, \widehat{\boldsymbol{X}} \in \mathbb{R}^{n_2 \times 1}, \end{aligned} \tag{5}$$

where endogeneity derives from the fact that $\mathbb{E}\left(\widehat{\boldsymbol{X}}'\boldsymbol{e}\right) = Var(\boldsymbol{e})$. Note that $[\cdot]$ in the regression equations above and below is notation used (for emphasis) to represent the (unobserved) error term.

Instrumental variable regression is a common approach to resolve issues of endogeneity. It begins with the assumption that we actually observe $(y_i, \widehat{x_i}, z_i, w_i)_{i=n_1+1,\ldots,n_1+n_2}$, where w_i is a d-dimensional row-vector of (presumed) instrumental variables, with $d \geq 1$. Estimation on n_2 samples is carried out in a two-stage least squares (2SLS) regression, of the form:

$$\begin{aligned} &\widehat{\boldsymbol{X}} = \boldsymbol{W}\boldsymbol{\Lambda_W} + \boldsymbol{Z}\boldsymbol{\Lambda_Z} + \boldsymbol{u} \quad \boldsymbol{W} \in \mathbb{R}^{n_2 \times d}, \\ &\boldsymbol{Y} = \widetilde{\boldsymbol{X}}\beta_X + \boldsymbol{Z}\beta_Z + \overbrace{[\boldsymbol{\varepsilon} + \beta_X(\widetilde{\boldsymbol{u}} - \boldsymbol{e})]}^{\boldsymbol{r}}, \end{aligned} \tag{6}$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{H}_W\widehat{\boldsymbol{X}}$ represents the projection of $\widehat{\boldsymbol{X}}$ onto the column space of $\boldsymbol{B}$, where $\boldsymbol{B} = [\boldsymbol{W}, \boldsymbol{Z}]$ and $\Lambda_B = [\Lambda_W, \Lambda_Z]$, $\boldsymbol{H}_W = \boldsymbol{B(B'B)^{-1}B'}$; and where $\widetilde{\boldsymbol{u}} = \widehat{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}$. Denote $\boldsymbol{C} = \left[\widetilde{\boldsymbol{X}}, \boldsymbol{Z}\right]$, then the 2SLS estimator $\hat{\beta}_{2SLS} = (\boldsymbol{C'C})^{-1}\boldsymbol{C'Y}$, equates to

$$\begin{aligned} \hat{\beta}_{2SLS} =& \beta + \left(\frac{\boldsymbol{C'C}}{n_2}\right)^{-1}\left(\frac{\boldsymbol{C'}r}{n_2}\right) \\ \rightarrow^p \;& \beta + plim_{n_2\to\infty}\left(\frac{\boldsymbol{C'C}}{n_2}\right)^{-1} plim_{n_2\to\infty}\left(\frac{\boldsymbol{C'}r}{n_2}\right) = \beta \end{aligned} \tag{7}$$

Therefore, $\hat{\beta}_{2SLS}$ is a consistent estimator of β under additional standard instrumental variable assumptions:

(Condition A6) $\mathbb{E}[\boldsymbol{B'\varepsilon}] = \mathbf{0}$,

(Condition A7) $\mathbb{E}[\boldsymbol{B'e}] = \mathbf{0}$,

(Condition A8) $rank(\mathbb{E}[\boldsymbol{B'B}]) = d + k$,

(Condition A9) $rank(\mathbb{E}[\boldsymbol{B'A}]) = k + 1$.

The following theorem formally establishes that, given an endogenous, mis-measured covariate (in this case, a vector of predictions from a single tree within the random forest), as well as a set of *other* mis-measured covariates (in this case, vectors of predictions obtained from other trees comprising the same random forest), in the absence of correlation between the error vector (actual − prediction) associated with the endogenous covariate and the set of other mis-measured covariates, one can obtain consistent estimates of the coefficients of interest.

Theorem 5. *Let the matrix* $\boldsymbol{P} = [\boldsymbol{p_1}, \ldots, \boldsymbol{p_M}] = \boldsymbol{X} + [\boldsymbol{e_1}, \ldots, \boldsymbol{e_M}]$*, where* $\forall j \in \{1, \ldots, M\}$ $\boldsymbol{p_j} \in \mathbb{R}^{n_2 \times 1}$ *is a column vector measure (with error as defined in (A3) - (A5)) of variable* $\boldsymbol{X}$ *under the specification defined in* (5) *with corresponding assumptions (A1)-(A2). Additionally, let* $\boldsymbol{E} = \boldsymbol{P} - \boldsymbol{X} = [\boldsymbol{e_1}, \ldots, \boldsymbol{e_M}]$ *be the matrix of measurement errors,* $S_j \subseteq \{1, \ldots, M\} \setminus j$ *be a subset of cardinality* d *such that* $\boldsymbol{P_{S_j}}, \boldsymbol{E_{S_j}} \in \mathbb{R}^{n_2 \times d}$ *are subsets of the column vectors of* $\boldsymbol{P}$ *and* $\boldsymbol{E}$ *respectively defined by the column indices in* S_j*. If* $\mathbb{E}\left[\boldsymbol{E'_{S_j}} \boldsymbol{e_j}\right] = \mathbf{0}$ *then using* P_{S_j} *as instruments for* P_j *in 2SLS would provide consistent estimates of the population parameters* β*, as defined in* (3).

From (5) we therefore know that $\boldsymbol{p_j}$ is endogenous because its measurement error $\boldsymbol{e_j}$ is captured by the unobserved error term. While from (6) and (7) we know that given a matrix $\boldsymbol{W}$ of instrumental variables, $\hat{\beta}_{2SLS} \to^p \beta$ when *(A6)-(A9)* are satisfied. Therefore, it suffices to show that *(A6)-(A9)* are satisfied when we let $\boldsymbol{W} = \boldsymbol{P_{S_j}}$.

$$\begin{aligned} \textit{(A1), (A3)} - \textit{(A5)} &\implies \mathbb{E}[\boldsymbol{Z'\varepsilon}] = \mathbf{0},\ \mathbb{E}\left[\boldsymbol{P'_{S_j}\varepsilon}\right] = \mathbf{0} \\ &\implies \mathbb{E}[\boldsymbol{B'\varepsilon}] = \mathbf{0} \textit{(A6)}, \end{aligned}$$

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{E'_{S_j}} \boldsymbol{e_j}\right] = \mathbf{0} &\implies \mathbb{E}\left[\boldsymbol{P'_{S_j}} \boldsymbol{e_j}\right] = \mathbf{0} \\ &\implies \mathbb{E}\left[\left[\boldsymbol{X_{S_j}} + \boldsymbol{E_{S_j}}, \boldsymbol{Z}\right]' \boldsymbol{\varepsilon}\right] = \mathbf{0} \\ &\implies \mathbb{E}[\boldsymbol{B'\varepsilon}] = \mathbf{0} \textit{(A7)}, \end{aligned}$$

and finally *(A8)-(A9)* follow directly from *(A1)-(A5)*, recognizing that $\mathbb{E}\left[\frac{\boldsymbol{p'_j p_l}}{n_2}\right] = Var[X]$ $j, l \in \{1, \ldots, M\}$.

Appendix C: Simulation Results on Alternative Designs of ForestIV With Bike Sharing Data

C.1. Sample Splitting

An alternative way of constructing candidate instrumental variables is sample splitting, which is a general strategy in econometrics to leverage multiple independent samples for estimation and inference (e.g., Angrist and Krueger 1995, Chernozhukov et al. 2016, Athey and Imbens 2017). Consider splitting the training data into two independent subsets, and build one random forest model on each subset. Each forest is asymptotically consistent (as the size of training data goes to infinity), and its prediction error is characterized as a random

and independent noise term (Scornet et al. 2015). Therefore, predictions from one random forest could, in principle, serve as an instrument for predictions from the other. The potential appeal of this sample splitting approach is that it (to some degree) preserves the superior predictive performance of a random forest over an individual decision tree, and thus may reduce the extent of the measurement error problem up front.

We explore the correction performance of this alternative approach with a simulation experiment on the Bike Sharing data. The basic setup is the same as in Section 3.1 of the main manuscript, except that we randomly split the training data into two equally-sized samples (each with 500 instances), and train one random forest model with 100 trees on each sample. We refer to the two forests as Forest#1 and Forest#2. We then use Forest#1's predictions on the unlabeled data as the endogenous covariate, and use Forest#2's predictions on the unlabeled data as the instrumental variable. Because there is only one instrument in this case, no instrument selection procedure is needed. We report the average estimates of this sample splitting approach over 100 simulation runs, together with the biased, unbiased, and ForestIV average estimates, in Table 6.

Table 6 Sample Splitting: Results on Bike Sharing Data

	True	Biased	Unbiased	ForestIV	Sample Splitting
Intercept	1.0	0.708 (0.093)	0.999 (0.162)	0.958 (0.118)	0.553 (0.115)
$lnCnt$	0.5	0.565 (0.019)	0.500 (0.034)	0.511 (0.024)	0.599 (0.023)
Z_1	2.0	2.000 (0.003)	1.999 (0.010)	2.000 (0.003)	2.000 (0.003)
Z_2	1.0	1.000 (0.002)	1.001 (0.005)	1.000 (0.002)	1.000 (0.001)
RMSE		0.314	0.166	0.128	0.472

Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

We find that the estimates produced by the sample splitting approach are in fact even more biased than the biased estimates (taken directly from the machine learning model without any correction), which is a strong indication that the instrument is invalid (Murray 2006). Indeed, after examining the predictions from both random forests on the testing data, we find that, while Forest#1's predictions are strongly correlated with Forest#2's predictions (average correlation is 0.96, indicating strong instrument relevance), Forest#1's prediction *errors* are *not* weakly correlated with Forest#2's predictions (average correlation is 0.30, indicating systematic violations of the instrument exclusion requirement). Therefore, although sample splitting may be a theoretically feasible approach to construct instruments, it does not appear to be effective in this case. We further repeat the simulations for a few additional sizes of training data, and summarize the results in Table 7. We again observe that ForestIV consistently outperforms sample splitting.

In addition, we extend the simulations to 13 different sizes of training data: 200, 250, 300, 350, 400, 450, 500, 750, 1000, 1250, 1500, 1750, and 2000. For each size of training data, we apply both ForestIV and sample splitting, then calculate the average squared bias, variance, and RMSE associated with each estimator across 100 simulation runs. In Figure 7, we plot the squared bias, variance, and RMSE of the two estimators across different sample sizes, both for the coefficient on $lnCnt$ (i.e., the variable generated by machine learning) and for all coefficients in the regression model. We also plot the ratios of squared bias, variance, and RMSE

Table 7 Sample Splitting with Different Sizes of Training Data

	True	$\lvert D_{train}\rvert$	250	500	1000	2000
		$\lvert D_{train}\rvert/2$	125	250	500	1000
Intercept	1.0	ForestIV	0.951 (0.377)	0.872 (0.242)	0.958 (0.118)	0.903 (0.096)
		Sample Splitting	0.259 (0.283)	0.421 (0.180)	0.553 (0.115)	0.631 (0.090)
$lnCnt$	0.5	ForestIV	0.514 (0.079)	0.528 (0.050)	0.511 (0.024)	0.522 (0.021)
		Sample Splitting	0.664 (0.056)	0.627 (0.035)	0.599 (0.023)	0.581 (0.018)
Z_1	2.0	ForestIV	1.999 (0.003)	2.000 (0.003)	2.000 (0.003)	2.000 (0.002)
		Sample Splitting	2.000 (0.003)	1.999 (0.002)	2.000 (0.003)	2.000 (0.003)
Z_2	1.0	ForestIV	1.000 (0.002)	1.000 (0.002)	1.000 (0.002)	1.000 (0.001)
		Sample Splitting	1.000 (0.002)	1.000 (0.002)	1.000 (0.002)	1.000 (0.002)
RMSE		ForestIV	0.392	0.272	0.128	0.127
		Sample Splitting	0.811	0.621	0.472	0.389

Standard errors in parentheses. $|D_{train}|$ represents the size of training data, and $|D_{train}|/2$ is the size of each split sample. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

of the two estimators. Empirically, we find that ForestIV has uniformly smaller bias, variance, and RMSE than sample splitting, and the ratios between the two are always smaller than 1. Moreover, we observe a downward trend in the bias ratio and RMSE ratio of ForestIV over sample splitting as the size of training data gets larger. This empirical evidence indicates that ForestIV may be a more efficient estimator than sample splitting. With a fixed size of training data, the former achieves smaller bias and smaller variance (i.e., better estimation) than the latter, and the advantage of ForestIV seems to become even larger as the size of training data increases.[10]

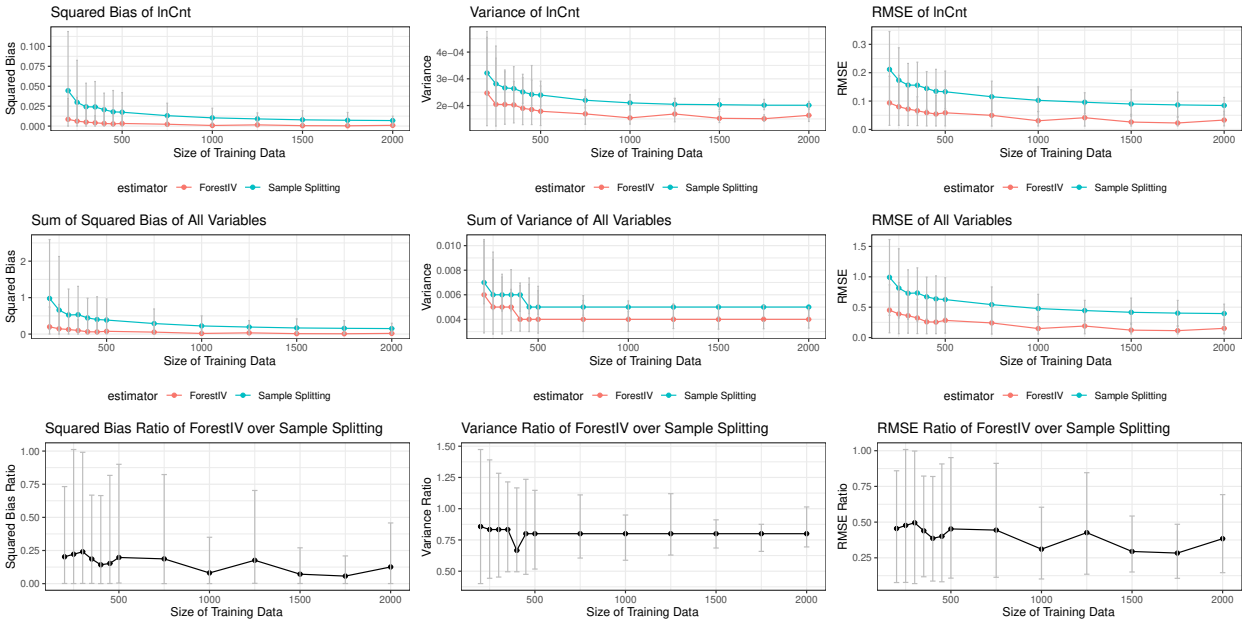
C.2. Using Subset of Trees as Endogenous Covariate and Instruments

While we propose to use a single tree in the random forest as the endogenous covariate and, under this restriction, select instruments for it, it is certainly possible to use a subset of trees in the forest to construct the endogenous covariate. The potential benefit of this approach is that aggregating over multiple trees can produce more accurate predictions, suggesting an up-front reduction of the measurement error problem. However, this approach poses a computational challenge, because there are $O(2^M)$ possible subsets of trees in a forest comprised of M trees, which are generally infeasible to enumerate and process exhaustively.

Nonetheless, we conduct a preliminary exploration of this strategy. Rather than enumerating all possible subsets, we randomly sample $q\%$ of all trees and use their averaged predictions as the endogenous covariate. Then, from the remaining $(1-q)\%$ of trees, we select the proper instruments (each instrument is still the predictions from a single tree) to estimate the regression. The tree sampling and instrument selection procedure is repeated 100 times, which produces 100 tuples of endogenous and instrumental variables. Note that we only generate 100 such tuples, again because it is computationally burdensome to examine all possible tree subsets even for a fixed $q\%$ (e.g., there are $\binom{100}{30} \approx 3 \times 10^{25}$ ways to sample 30 trees from a random forest of 100 trees). As with ForestIV, we conclude by selecting the tuple that minimizes the empirical MSE. We use the same basic simulation configurations under the Bike Sharing dataset, and vary $q\%$ over three levels,

[10] We acknowledge that, if the size of training data were significantly larger, performance of the sample splitting approach could potentially further improve. However, once again, we must reiterate that, in the presence of more training data, there is no need to "mine" covariates via machine-learning techniques in the first place.

Figure 7 Squared Bias, Variance, and RMSE Comparisons between ForestIV and Sample Splitting



Note. Vertical bars represent 95% confidence intervals. The three plots in the first row show the squared bias, variance, and RMSE of $lnCnt$, the variable generated by machine learning. The three plots in the second row show the squared bias, variance, and RMSE of all variables in the model. The three plots in the third row show the ratios of squared bias, variance, and RMSE between the two estimators, when all variables are considered.

30%, 50%, and 70%, respectively representing low, medium, and high levels of subset aggregation. For each choice of q%, we report the average estimates (along with the biased, unbiased, and ForestIV estimates) across 100 simulation runs, in Table 8.

Table 8 Using Subset of Trees: Results on Bike Sharing Data

	True	Biased	Unbiased	ForestIV	30% Sampling	50% Sampling	70% Sampling
Intercept	1.0	0.708 (0.093)	0.999 (0.162)	0.958 (0.118)	0.757 (0.083)	0.718 (0.080)	0.763 (0.064)
$lnCnt$	0.5	0.565 (0.019)	0.500 (0.034)	0.511 (0.024)	0.552 (0.018)	0.562 (0.016)	0.553 (0.014)
Z_1	2.0	2.000 (0.003)	1.999 (0.010)	2.000 (0.003)	2.000 (0.003)	1.999 (0.003)	1.999 (0.003)
Z_2	1.0	1.000 (0.002)	1.001 (0.005)	1.000 (0.002)	1.000 (0.002)	1.000 (0.002)	1.000 (0.001)
RMSE		0.314	0.166	0.128	0.309	0.407	0.363

Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

We find that randomly selecting a subset of trees to construct the endogenous covariate exhibits limited effectiveness in bias correction, as compared to our proposed ForestIV approach. Across three different sampling ratios, the estimates achieve only small improvements over the biased estimates. Note that we have explored only a highly limited number of subsets, compared to all $O(2^{100})$ possible subsets, and we have restricted each instrument to be the predictions from a single tree (rather than the aggregated predictions from a subset of trees). Fully exploring all possible pairs of endogenous and instrumental variables (each consisting of tree subsets) is clearly infeasible. Future research might further develop this avenue of inquiry,

studying potential heuristic or optimization methods to reduce the computational burden of extensive tree subset enumeration.

C.3. Averaging across Multiple Estimates

We now examine whether averaging across multiple estimates in the ForestIV procedure may enhance bias correction. Specifically, in Step 3 of ForestIV, rather than selecting a single set of estimates that minimizes empirical MSE, we *average* across all sets of estimates that are not rejected by the Hotelling's T^2 test. The averaging estimates, along with the biased, unbiased, and ForestIV estimates, are reported in Tables 9.

Table 9 Averaging Estimates: Results on Bike Sharing Data

	True	Biased	Unbiased	ForestIV	Averaging ForestIV
Intercept	1.0	0.708 (0.093)	0.999 (0.162)	0.958 (0.118)	0.797 (0.086)
$lnCnt$	0.5	0.565 (0.019)	0.500 (0.034)	0.511 (0.024)	0.545 (0.018)
Z_1	2.0	2.000 (0.003)	1.999 (0.010)	2.000 (0.003)	2.000 (0.003)
Z_2	1.0	1.000 (0.002)	1.001 (0.005)	1.000 (0.002)	1.000 (0.002)
RMSE		0.314	0.166	0.128	0.318

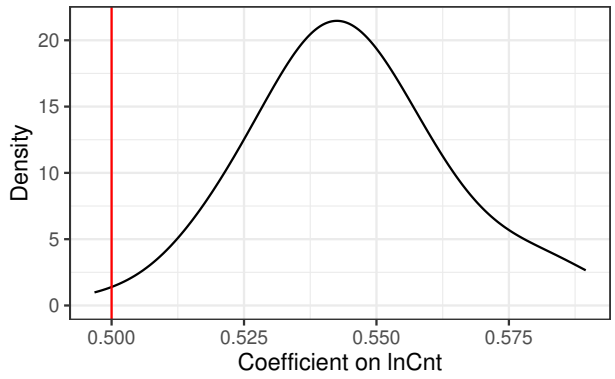
Standard errors in parentheses. RMSE contains the empirical RMSE associated with each set of estimates, averaged across 100 simulation runs.

We can see that, in this case, while averaging across multiple estimates does indeed continue to mitigate the bias to some extent, it is nonetheless not as effective as our baseline ForestIV approach, which employs the single tuple that minimizes the empirical MSE. The limited utility of averaging is because the estimates are not "symmetrically" distributed around the true values of coefficients (see Figure 8 for the distribution of the averaged coefficient estimate on $lnCnt$). Rather, the distribution of values systematically deviates from the true values in a single direction. This is unsurprising, to some extent, given that 2SLS estimation is known to be biased in finite samples (Nagar 1959, Buse 1992). Consequently, averaging produces worse correction results than picking the "best" tuple (the one that minimizes the empirical MSE). Note it is also possible that some tuples may ultimately fail to be rejected by the Hotelling's T^2 test, despite containing invalid or weak instruments (recall that a failure to reject the null hypothesis of equivalence does not imply that said null should be "accepted"). Intuitively, by choosing to focus on the "best" tuple, we are applying the most stringent p-value threshold possible, and thus mitigating the potential that we unintentionally retain poor instruments in our final estimation. To the extent that retained instruments are of low quality, they may bias the resulting estimations towards that of the biased OLS (Murray 2006, Wooldridge 2002), which is exactly what we observe here. Nonetheless, we believe future work can investigate whether averaging can be advantageous under certain conditions, or develop new methodologies to derive more robust estimations when some instruments may be invalid or weak (e.g., based on the work of Conley et al. 2012).

Appendix D: Sensitivity Analyses for Binary Endogenous Covariate with Bank Marketing Data

We repeat all the sensitivity analyses that have been carried out for the continuous case and report the results in this subsection. The sensitivity analyses again consist of three parts. We examine the performance of ForestIV estimations with respect to (1) the size of unlabeled data, (2) the size of labeled data, and (3)

Figure 8 Distribution of Averaged ForestIV Estimation on $lnCnt$ across 100 Simulation Runs

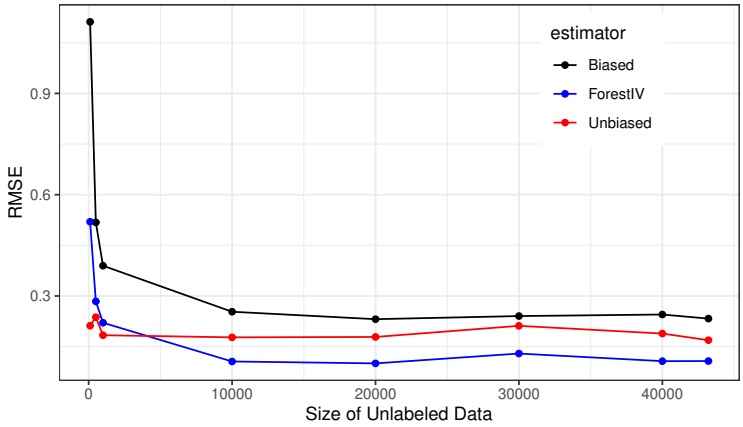


the predictive performance of the random forest (operationalized by changing the total number of trees). For all simulations in this section, we use the same "Bank Marketing" dataset (Moro et al. 2014) and, unless otherwise noted, the simulation setups are the same as in the main manuscript.

D.1. Size of Unlabeled Data

We repeat the simulation with 8 different sizes of unlabeled data, respectively 100, 500, 1,000, 10,000, 20,000, 30,000, 40,000, and 43,211 (i.e., all remaining instances). In Figure 9, we plot the RMSE of biased, unbiased, and ForestIV estimates under different sizes of unlabeled data.

Figure 9 RMSE of Biased, Unbiased, and ForestIV Estimates under Different Sizes of Unlabeled Data

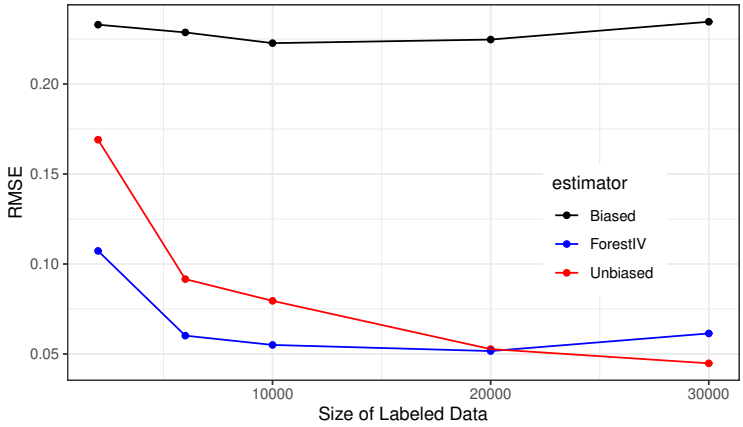


We observe the same patterns as in the case of a continuous endogenous covariate. First, the biased estimates have the highest RMSE regardless of the size of unlabeled data. On the other hand, ForestIV produces smaller RMSE than the unbiased estimates when sufficient unlabeled data is available.

D.2. Size of Labeled Data

We next repeat the main simulation with 5 different sizes of labeled data, respectively 2000, 6000, 10000, 20000, and 30000, while keeping the same $3:1$ training / testing split ratio as in the basic simulation setup. Other parameters are also kept the same. We plot the RMSE of biased, unbiased, and ForestIV estimates under different sizes of labeled data in Figure 10.

Figure 10 RMSE of Biased, Unbiased, and ForestIV Estimates under Different Sizes of Labeled Data

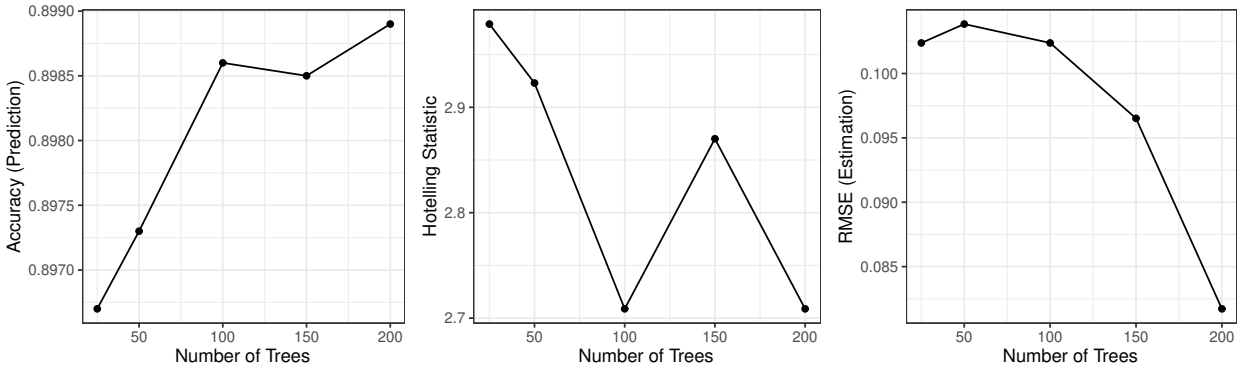


We find that ForestIV estimates achieve smaller RMSE than the unbiased estimates when the labeled dataset is relatively small, but the opposite is true when the size of labeled dataset grows larger than 20000 (or roughly 44% of all data). Again, these two sensitivity analyses indicate that ForestIV is most advantageous when the labeled dataset is much smaller than the unlabeled dataset.

D.3. Predictive Performance of Random Forest

We repeat the simulation with 5 different choices of the total number of trees in the random forest, M: 25, 50, 100, 150, and 200. In Figure 11, we plot (1) the prediction accuracy of the forest, (2) the Hotelling T^2 statistic of ForestIV estimates (averaged over 100 simulation runs), and (3) the RMSE of ForestIV estimates, across different number of trees.

Figure 11 Prediction Accuracy, Hotelling T^2 statistic, and Estimation RMSE under Different umber of Trees



In this particular example, different choices of M are associated with very similar accuracy on average. Nonetheless, we observe an alignment between estimation RMSE and the Hotelling T^2 statistic. In particular, $M = 200$ produces estimates with smallest RMSE as well as the lowest Hotelling T^2 statistic.

Appendix E: Limitation of SIMEX

While SIMEX is a generally reliable approach to dealing with measurement error and has been demonstrated to perform well in various types of econometric models (Yang et al. 2018), we have identified an important limitation that causes SIMEX to produce problematic correction results under certain conditions.

Again, consider the regression model $Y \sim \widehat{X}\beta_X + \boldsymbol{Z}\boldsymbol{\beta_Z}$, where $\widehat{X}$ is measured with additive error, i.e., $\widehat{X} = X + e$. Suppose that the measurement error component, e, is correlated with one of the precisely-measured control variables in the model, e.g., $\exists Z^* \in \boldsymbol{Z}, Cov(e, Z^*) \neq 0$. The SIMEX-corrected coefficient on Z^* under this setup can be even more biased than it is in the absence of correction.

This represents a realistic scenario that can arise when combining machine learning and econometric modeling. As an example, consider a credit scoring prediction model. An older person's credit score might be easier to predict (with less error) than a younger person's, as more data on historical consumption and repayment will be available in the former case. The error in credit score prediction may thus be correlated with age, which would plausibly appear as a control variable in many econometric models. Another example that has attracted heated discussion is Propublica's critique of the COMPAS risk tool (Angwin et al. 2016), a predictive model used in the U.S. criminal justice system to assess a defendant's risk of recidivism. The prediction errors that COMPAS produces are correlated with race; the algorithm has been shown to be more inaccurate for (i.e., biased against) African-American defendants than for white defendants. Similar racial disparity in predictive performance has been documented in gender classification (Buolamwini and Gebru 2018), where researchers found that image classifiers are less accurate for darker-skinned people.

In the following theorem, we prove this limitation of SIMEX in the simple case of a linear regression with additive independent (i.e., classical) measurement error. Formally, consider a population regression equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_1 is measured with additive independent error, i.e., $\widehat{X_1} = X_1 + e$, and $Cov(X_1, e) = 0$. Suppose that X_2 is correlated with the measurement error, i.e., $Cov(X_2, e) = \sigma_{2e} \neq 0$. We still assume the model error term, ε, is exogenous, i.e., $Cov(X_1, \varepsilon) = Cov(X_2, \varepsilon) = Cov(e, \varepsilon) = 0$. For simplicity, we also assume $Cov(X_1, X_2) = 0$.[11] Further denote $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$, and $Var(e) = \sigma_e^2$. Finally, denote the true coefficient, the biased (i.e., uncorrected) coefficient, and the SIMEX-corrected coefficient on X_2 as β_2, $\widehat{\beta_2}$, and $\widehat{\beta_2}^{SIMEX}$ respectively.

Theorem 6. *The SIMEX-corrected coefficient will become even more biased than in the absence of correction, that is,* $|\widehat{\beta_2}^{SIMEX} - \beta_2| > |\widehat{\beta_2} - \beta_2|$*, if and only if* $|\sigma_2^2\sigma_1^2 - \sigma_{2e}^2| < |\sigma_2^2(\sigma_1^2 + \sigma_e^2) - \sigma_{2e}^2|$.

Consider OLS estimation of the biased regression of Y on $\{\widehat{X_1}, X_2\}$. Using the regression anatomy method (Angrist and Pischke 2008), the estimated coefficient associated with X_2 is $\widehat{\beta_2} = \frac{Cov(Y, \widetilde{X_2})}{Var(\widetilde{X_2})}$, where $\widetilde{X_2}$ is the residual from regressing X_2 on $\widehat{X_1}$, i.e., $X_2 = r_0 + r_1\widehat{X_1} + \widetilde{X_2}$. We know that $r_1 = \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}$, and therefore $\widetilde{X_2} = X_2 - r_0 - \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}\widehat{X_1}$. It follows that $\widehat{\beta_2} = \frac{Cov(Y, \widetilde{X_2})}{Var(\widetilde{X_2})} = \frac{Cov\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad X_2 - r_0 - \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}\widehat{X_1}\right)}{Var\left(X_2 - r_0 - \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}\widehat{X_1}\right)}$. Dropping the constant terms β_0, r_0 and the exogenous ε, the above expression simplifies to $\widehat{\beta_2} = \frac{Cov\left(\beta_1 X_1 + \beta_2 X_2, \quad X_2 - \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}\widehat{X_1}\right)}{Var\left(X_2 - \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})}\widehat{X_1}\right)} = \frac{-\beta_1 \frac{Cov(X_2, \widehat{X_1})}{Var(\widehat{X_1})} Cov(X_1, \widehat{X_1}) + \beta_2\sigma_2^2 - \beta_2 \frac{Cov^2(X_2, \widehat{X_1})}{Var(\widehat{X_1})}}{\sigma_2^2 - \frac{Cov^2(X_2, \widehat{X_1})}{Var(\widehat{X_1})}} = \beta_2 - \beta_1 \frac{Cov(X_2, \widehat{X_1})Cov(X_1, \widehat{X_1})}{\sigma_2^2 Var(\widehat{X_1}) - Cov^2(X_2, \widehat{X_1})}$. Given that $Var(\widehat{X_1}) = \sigma_1^2 + \sigma_e^2$, $Cov(X_1, \widehat{X_1}) = \sigma_1^2$, $Cov(X_2, \widehat{X_1}) = \sigma_{2e}$, we have $\widehat{\beta_2} = \beta_2 - \beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2(\sigma_1^2 + \sigma_e^2) - \sigma_{2e}^2}$. This implies that the absolute bias in the coefficient associated with X_2, $|\widehat{\beta_2} - \beta_2| = \left|\beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2(\sigma_1^2 + \sigma_e^2) - \sigma_{2e}^2}\right|$.

[11] Relaxing this assumption makes the derivation more elaborate without changing the underlying mechanism. When $Cov(X_1, X_2) \neq 0$, our statement is still true under slightly more strict conditions

Now consider the SIMEX correction procedure (Cook and Stefanski 1994). In the simulation step, SIMEX creates $\widehat{X_1}^{(\lambda)} = \widehat{X_1} + \sqrt{\lambda}z = X_1 + e + \sqrt{\lambda}z$, where $z \sim N(0, \sigma_e^2)$, and thereby introduces more measurement error. Note that $Var(\widehat{X_1}^{(\lambda)}) = \sigma_1^2 + (1+\lambda)\sigma_e^2$, and $Cov(X_2, \widehat{X_1}^{(\lambda)}) = \sigma_{2e}$ because z is independently generated. Following the same derivation above, we know that regressing Y on $\{\widehat{X_1}^{(\lambda)}, X_2\}$, we would have $\widehat{\beta_2}^{(\lambda)} = \beta_2 - \beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2\left(\sigma_1^2+(1+\lambda)\sigma_e^2\right)-\sigma_{2e}^2}$, or equivalently, $|\widehat{\beta_2}^{(\lambda)} - \beta_2| = \left|\beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2\left(\sigma_1^2+(1+\lambda)\sigma_e^2\right)-\sigma_{2e}^2}\right|$. In the extrapolation step, SIMEX estimates $\widehat{\beta_2}^{(-1)}$, i.e., the coefficient that would be obtained had there been no measurement error (note that $\widehat{\beta_2}^{(-1)} \equiv \widehat{\beta_2}^{SIMEX}$). Accordingly, $|\widehat{\beta_2}^{(-1)} - \beta_2| = \left|\beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2\sigma_1^2-\sigma_{2e}^2}\right|$.

Finally, we compare $|\widehat{\beta_2}^{(-1)} - \beta_2| = \left|\beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2\sigma_1^2-\sigma_{2e}^2}\right|$ and $|\widehat{\beta_2} - \beta_2| = \left|\beta_1 \frac{\sigma_{2e}\sigma_1^2}{\sigma_2^2(\sigma_1^2+\sigma_e^2)-\sigma_{2e}^2}\right|$, or equivalently, compare $\frac{1}{\left|\sigma_2^2\sigma_1^2-\sigma_{2e}^2\right|}$ and $\frac{1}{\left|\sigma_2^2(\sigma_1^2+\sigma_e^2)-\sigma_{2e}^2\right|}$. Therefore, the condition $|\sigma_2^2\sigma_1^2 - \sigma_{2e}^2| < |\sigma_2^2(\sigma_1^2+\sigma_e^2) - \sigma_{2e}^2|$ implies that $\frac{1}{\left|\sigma_2^2\sigma_1^2-\sigma_{2e}^2\right|} > \frac{1}{\left|\sigma_2^2(\sigma_1^2+\sigma_e^2)-\sigma_{2e}^2\right|} \Rightarrow |\widehat{\beta_2}^{(-1)} - \beta_2| > |\widehat{\beta_2} - \beta_2|$, i.e., the SIMEX corrected coefficient on X_2 becomes even more biased than it would be in the absence of correction.

Remark. We note that the condition $|\sigma_2^2\sigma_1^2 - \sigma_{2e}^2| < |\sigma_2^2(\sigma_1^2+\sigma_e^2) - \sigma_{2e}^2|$ is quite easily satisfied. The right-hand-side of the inequality is equivalent to $|\sigma_2^2\sigma_1^2 - \sigma_{2e}^2 + \sigma_2^2\sigma_e^2|$. Therefore, a sufficient (but not necessary) condition for the inequality to hold is $\sigma_2^2\sigma_1^2 - \sigma_{2e}^2 \geq 0$. Because $\sigma_2^2\sigma_1^2 - \sigma_{2e}^2 \geq 0 \Leftrightarrow \frac{\sigma_1^2}{\sigma_e^2} \geq \frac{\sigma_{2e}^2}{\sigma_2^2\sigma_e^2} = \rho_{2e}^2$, we can see that the inequality always holds if $\sigma_1^2 \geq \sigma_e^2$, which means that the variance of measurement error is no larger than the variance of true covariate. This is typically true unless the measurement error is exceedingly large.

Essentially, the SIMEX approach relies on the implicit assumption that degree of measurement error is *positively* related to degree of bias. This assumption is violated when a precisely-measured covariate is correlated with the measurement error, as we have shown in the above theorem. As a result, unless special modifications are made to the SIMEX procedure, it produces incorrect results on the precisely-measured covariate. However, our ForestIV approach does not suffer from this issue, because it does not rely on the same implicit assumption. Instead, the identified instruments should mitigate estimation bias on the error-prone covariate, without introducing additional bias to the estimates of precisely-measured covariates.

Appendix F: Example Research that Combines Machine Learning and Econometric Modeling for Causal Inference

In the following Table 10, we list a number of papers that first build certain predictive machine learning model to derive a variable of interest, and then incorporate that variable into an econometric model for estimation and inference.

Table 10 Example Research that Combines Machine Learning and Econometric Modeling for Causal Inference

Citation	Research Question	Variables Generated by Predictive Machine Learning Models
Goh et al. (2013)	The impact of user- and marketer-generated content on consumer spending.	Text sentiment.
Gu et al. (2007)	The trade-offs between information quantity and quality in virtual investment communities.	The "quality" of online postings (noise, neutral, or signal).
Ghose and Ipeirotis (2010)	The economic impact of product reviews on Amazon.com.	Objectivity of product reviews (objective vs. subjective).
Ghose et al. (2012)	Using user-generated content to design better ranking systems for search engines.	Objectivity of customer reviews (objective vs. subjective).
Tirunillai and Tellis (2012)	Impact of user-generated content on firms' stock market performance.	Text sentiment.
Liu et al. (2012)	The relationship between lending activity and lender motivation in online microfinance sites.	Lender motivation (based on 10 predefined classes).
Zhu et al. (2011, 2012)	Identifying and investigating the impact of shared leadership on Wikipedia contribution behaviors.	Shared leadership types (based on 4 pre-defined classes).
Aggarwal et al. (2012)	Influence of electronic word-of-mouth on venture capital financing.	Text sentiment.
Lu et al. (2013)	Opinion leadership in online community.	Objectivity of product reviews (objective vs. subjective).
Moreno and Terwiesch (2014)	The effect of reputation system on market outcomes.	Text sentiment.
Singh et al. (2014)	Dynamics of blog reading behavior.	Text sentiment.
Wang et al. (2013)	Relationship between risk factor disclosure and future breach announcement.	Breach announcement (yes vs. no, predicted based on textual disclosure).
Gu et al. (2014)	Whether social media participation exhibits homophily.	Text sentiment.