Online Appendix for Machine Learning Predictions as Regression Covariates

Christian Fong* Matthew Tyler†

July 14, 2020

*Ph.D. Candidate, Stanford Graduate School of Business. christianfong@stanford.edu
†Ph.D. Candidate, Department of Political Science, Stanford University. mdtyler@stanford.edu

Contents

A Technical Appendix

The method described in the main text is implemented as the function `predicted_covariates` in the companion `R` package. The following section explains the technical details of this function.

A.1 The GMM Estimator in a Closed-Form Expression

Let

$$\boldsymbol{B} = \begin{pmatrix} \frac{1}{n_v+n_t}\sum_{i=1}^{n}(v_i + t_i)x_i y_i \\ \frac{1}{n_v+n_p}\sum_{i=1}^{n}(p_i + v_i)z_i y_i \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} \frac{1}{n_v+n_t}\sum_{i=1}^{n}(v_i + t_i)x_i x_i^\mathsf{T} \\ \frac{1}{n_v+n_p}\sum_{i=1}^{n}(p_i + v_i)z_i z_i^\mathsf{T}\hat{\boldsymbol{\Gamma}}^\mathsf{T} \end{pmatrix}. \tag{1}$$

With this notation, we see that $g(b) = \boldsymbol{B} - \boldsymbol{A}b$ and $\boldsymbol{G} = \frac{\partial}{\partial b}g(b) = -\boldsymbol{A}$. This notation is useful since the first-order condition for the GMM optimization problem implies

$$2\boldsymbol{G}^\mathsf{T}\boldsymbol{W}g(\hat{\beta}) = 0 \tag{2}$$

$$\implies \boldsymbol{A}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} = \boldsymbol{A}^\mathsf{T}\boldsymbol{W}\boldsymbol{A}\hat{\beta} \tag{3}$$

$$\implies \hat{\beta} = (\boldsymbol{A}^\mathsf{T}\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{W}\boldsymbol{B}. \tag{4}$$

Computationally, this means that given $\hat{\boldsymbol{\Gamma}}$ and $\boldsymbol{W}$, $\hat{\beta}$ is just a few matrix operations away.

A.2 The Optimal GMM Weighting Matrix and Estimator Variance

This section assumes familiarity with a variety of asymptotic results from GMM and two-step estimation theory. Any implicitly used theorems can be found in Newey and McFadden (1994).

For $\hat{\beta}$ to achieve its optimal asymptotic efficiency, we need to find a sequence of positive definite matrices $\boldsymbol{W}$ that converge to Ξ^{-1}, where Ξ is the first-order asymptotic variance of $\sqrt{n_v}g(\beta)$. We scale by $\sqrt{n_v}$ because we base our asymptotics on $n_v \to \infty$. Let $\lambda_p = \lim_{n_v\to\infty}\frac{n_v}{n_v+n_p}$ and $\lambda_t = \lim_{n_v\to\infty}\frac{n_v}{n_v+n_t}$. There are no restrictions on λ_p and λ_t other than that they must lie in the interval $[0, 1]$.

To compute Ξ properly, we need to account for how much $g(\beta)$ deviates based on the error $\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}$. Using standard asymptotic theory, it can be shown that

$$\sqrt{n_v}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}) = -\frac{1}{\sqrt{n_v}} \sum_{i=1}^{n_v} v_i (x_i - \mathbf{\Gamma} z_i) z_i^\mathsf{T} \mathbb{E}(zz^\mathsf{T})^{-1} + o_p(1). \tag{5}$$

Fortunately, we only need to account for this error in the g_2 moment condition, since that is the only place where $\hat{\mathbf{\Gamma}}$ appears. Observe that

$$\begin{aligned}
\sqrt{n_v} g_2(\beta) &= \frac{\sqrt{n_v}}{n_p + n_v} \sum_{i=1}^{n} (v_i + p_i) z_i (y_i - (\hat{\mathbf{\Gamma}} z_i)^\mathsf{T} \beta) && (6)\\
&= \frac{\sqrt{n_v}}{n_p + n_v} \sum_{i=1}^{n} (v_i + p_i) \left[z_i (y_i - (\mathbf{\Gamma} z_i)^\mathsf{T} \beta) - z_i z_i^\mathsf{T} (\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})^\mathsf{T} \beta \right] && (7)\\
&= \frac{\sqrt{n_v}}{n_p + n_v} \sum_{i=1}^{n} (v_i + p_i) z_i (y_i - (\mathbf{\Gamma} z_i)^\mathsf{T} \beta) - \mathbb{E}(zz^\mathsf{T}) \sqrt{n_v} (\hat{\mathbf{\Gamma}} - \mathbf{\Gamma})^\mathsf{T} \beta + o_p(1) && (8)\\
&= \frac{\sqrt{n_v}}{n_p + n_v} \sum_{i=1}^{n} (v_i + p_i) z_i (y_i - (\mathbf{\Gamma} z_i)^\mathsf{T} \beta) + \frac{1}{\sqrt{n_v}} \sum_{i=1}^{n} v_i z_i (x_i - \mathbf{\Gamma} z_i)^\mathsf{T} \beta + o_p(1) && (9)\\
&= \frac{1}{\sqrt{n_v}} \sum_{i=1}^{n} \left[(v_i + p_i) \frac{n_v}{n_v + n_p} z_i (y_i - (\mathbf{\Gamma} z_i)^\mathsf{T} \beta) + v_i z_i (x_i - \mathbf{\Gamma} z_i)^\mathsf{T} \beta \right] + o_p(1) && (10)
\end{aligned}$$

Note that going forward we can use Slutsky's lemma to replace the sample size ratios with their corresponding λ values (we estimate the λ values with with their observed sample size ratios anyway, so in practice there is no error). This suggests a sequence of weighting matrices $\boldsymbol{W}$ given by

$$\begin{aligned}
\boldsymbol{W}^{-1} &= \frac{1}{n_v} \sum_{i=1}^{n} h_i h_i^\mathsf{T} && (11)\\
h_i &= \begin{pmatrix} (v_i + t_i) \lambda_t x_i (y_i - x_i^\mathsf{T} \tilde{\beta}) \\ (p_i + v_i) \lambda_p z_i (y_i - (\hat{\mathbf{\Gamma}} z_i)^\mathsf{T} \tilde{\beta}) + v_i z_i (x_i - \hat{\mathbf{\Gamma}} z_i)^\mathsf{T} \tilde{\beta} \end{pmatrix}, && (12)
\end{aligned}$$

which consistently estimates Ξ, so long as $\tilde{\beta}$ is some $\sqrt{n_v}$ consistent estimator of β (such as the labeled-only estimate) and the observations are independent (Newey and McFadden, 1994). Using

this $\boldsymbol{W}$ to compute $\hat{\beta}$, and denoting $\Delta = \text{plim}_{n_v \to \infty} G$, which is easily shown to be

$$\Delta = \begin{pmatrix} -\mathbb{E}(xx^\mathsf{T}) \\ -\mathbb{E}(zx^\mathsf{T}) \end{pmatrix}, \tag{13}$$

standard GMM results tell us that $\sqrt{n_v}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, (\Delta^\mathsf{T}\Xi^{-1}\Delta)^{-1})$.

Conveniently, Ξ can be written somewhat succinctly in terms of population moments. In particular,

$$\Xi = \begin{pmatrix} \Xi_{11} & \Xi_{21}^\mathsf{T} \\ \Xi_{21} & \Xi_{22} \end{pmatrix} \tag{14}$$

$$\Xi_{11} = \lambda_t \mathbb{E}(\epsilon^2 xx^\mathsf{T}), \quad \Xi_{21} = \lambda_p \lambda_t \mathbb{E}(\epsilon(\epsilon + \eta^\mathsf{T}\beta)zx^\mathsf{T}) + \lambda_t \mathbb{E}(\epsilon(\eta^\mathsf{T})\beta zx^\mathsf{T}) \tag{15}$$

$$\Xi_{22} = \lambda_p \mathbb{E}((\epsilon + \eta^\mathsf{T}\beta)^2 zz^\mathsf{T}) + 2\lambda_p \mathbb{E}((\epsilon + \eta^\mathsf{T}\beta)(\eta^\mathsf{T}\beta)zz^\mathsf{T}) + \mathbb{E}((\eta^\mathsf{T}\beta)^2 zz^\mathsf{T}). \tag{16}$$

We see that Ξ_{22} decreases (with respect to the convex cone of positive semidefinite matrices), and thus the GMM estimator achieves a lower variance, as $\lambda_p \to 0$. This tells us that for a fixed value of n_v, increasing the relative size of the primary sample decreases the GMM estimator variance. This is analogous to the discussion of Chen, Hong and Tamer (2005, §3.4), although our λ_p is defined somewhat differently than theirs. Note that it if one's goal is to determine the optimal split of labeled data into training and validation samples, it would not be reasonable (without additional arguments) to optimize Ξ as a function of λ_t and λ_p, since the number of training examples n_t also determines the distribution of z.

In practice, we find that it is often necessary to iterate between computing $\boldsymbol{W}$ and $\hat{\beta}$ several times. Let $\hat{\beta}^{(0)}$ be the labeled only estimator (i.e., OLS of y on x in the training and validation samples). This is a $\sqrt{n_v}$-consistent estimator of β, and thus allows us to consistently estimate Ξ. Beginning with this starting value, our algorithm proceeds as follows:

1. Compute $\boldsymbol{W}^{(t+1)}$ based on eq. (11) using $\hat{\beta}^{(t)}$.

2. Compute $\hat{\beta}^{(t+1)}$ based on eq. (4) using $\boldsymbol{W}^{(t+1)}$.

3. Repeat until convergence. In our algorithm, we declare convergence if, for each component $k = 1, \ldots, d_x$,

$$\frac{|\hat{\beta}_k^{(t+1)} - \hat{\beta}_k^{(t)}|}{|\hat{\beta}_k^{(t)}|} < 0.01. \tag{17}$$

That is, if each component changes less than 1%.

B Testing the Exclusion Restriction

Here we introduce a test of the exclusion restriction. It is important to note that this test treats the exclusion restriction as the null hypothesis:

$$H_0 : \mathbb{E}[z_u \epsilon] = 0 \tag{18}$$

and the alternative hypothesis is the violation of the exclusion restriction, $\mathbb{E}[\boldsymbol{z}_u \epsilon] \neq 0$.

To perform this test, we cast the exclusion restriction as another moment condition for $\boldsymbol{\beta}$ and use a standard over-identification test (Sargan, 1958; Hansen, 1982). In particular, given the exclusion restriction,

$$\mathbb{E}[k(y - x^\mathsf{T}\beta)] = 0, \tag{19}$$

where $k = (x, z_u)$. Using these $d_x + d_{z_u}$ moment conditions on the validation sample, we can create another GMM estimator of β, denoted $\bar{\beta}$, which can be used to test the null hypothesis.[1] The alternative hypothesis stipulates that at least some of the moment conditions are incorrect — but since we know that $\mathbb{E}(x(y - x^\mathsf{T}\beta)) = 0$ without loss of generality (by construction of β), the alternative hypothesis is equivalent to the exclusion restriction being violated by at least

[1] We could just estimate β on the validation sample without the z_u conditions, but that might decrease the power of the test since we are not using all the available moment conditions.

one of the elements of z_u. The companion `R` package automatically provides the p-value from the implied Sargan-Hansen over-identification test (a.k.a. Sargan's J-test) via the `gmm` package (Chaussé, 2010).[2]

C Including Auxiliary Data in the First Stage

In some cases, the researcher may wish to use an additional sample — which we will call the "auxiliary sample" — in the first-stage component of the 2SLS estimator but not include that sample in any other part of the GMM estimator. This is reasonable if the researcher suspects that the relationship between the covariates and the predictions is the same in the auxiliary sample as in the validation sample, but does not otherwise want to use the auxiliary sample to estimate the coefficients β. Appendix M provides an example application for the inclusion of auxiliary data: the researcher may wish to perform subgroup analyses because the relationship between the outcome and the covariate may vary across subgroups, even though the relationship between the covariate and the prediction is the same across subgroups. Including auxiliary data increases the power of the analysis.

In this case, we redefine the first-stage estimator $\hat{\Gamma}$ as

$$\hat{\Gamma} = \left(\sum_{i=1}^{n}(a_i + v_i)x_i z_i^{\mathsf{T}}\right)\left(\sum_{i=1}^{n}(a_i + v_i)z_i z_i^{\mathsf{T}}\right)^{-1}, \tag{20}$$

where a_i is the sample indicator for i belonging to the auxiliary sample. Accordingly, let $n_a = \sum_{i=1}^{n} a_i$. The new GMM estimator is the same as described in Section 3, except we use this new first-stage estimate of Γ.

Note, however, that some adjustments must be made to the formulas given in Online Ap-

[2]For those unfamiliar with this test, let $r_i = k_i(y - x_i^{\mathsf{T}}\bar{\beta})$, $\bar{r} = \frac{1}{n_v}\sum_{i=1}^{n} v_i r_i$, and $R = \frac{1}{n_v}\sum_{i=1}^{n} v_i r_i r_i^{\mathsf{T}}$. With the exclusion restriction as our null hypothesis, the statistic $n_v \bar{r}^{\mathsf{T}} R^{-1} \bar{r} \geq 0$ follows a χ^2 distribution with degrees of freedom equal to the length of z_u.

pendix A, which we now describe. It is useful to introduce some new notation:

$$o_i = v_i + t_i \tag{21}$$

$$f_i = v_i + a_i \tag{22}$$

$$s_i = v_i + p_i \tag{23}$$

with corresponding sample sizes n_o, n_f, and n_s defined appropriately. The new first stage estimator has the asymptotic expansion

$$\sqrt{n_f}(\hat{\Gamma} - \Gamma) = -\frac{1}{\sqrt{n_f}} \sum_{i=1}^{n} f_i (x_i - \Gamma z_i) z_i^\mathsf{T} \mathbb{E}[zz^\mathsf{T}]^{-1} + o_p(1). \tag{24}$$

This implies a new value for h_i from Equation 11, now updated to

$$h_i = \begin{pmatrix} o_i \frac{n_v}{n_o} x_i (y_i - x_i^\mathsf{T} \tilde{\beta}) \\ s_i \frac{n_v}{n_s} z_i (y_i - (\hat{\Gamma} z_i)^\mathsf{T} \tilde{\beta}) + f_i \frac{n_v}{n_f} z_i (x_i - \hat{\Gamma} z_i)^\mathsf{T} \tilde{\beta} \end{pmatrix}. \tag{25}$$

These adjustments are sufficient to fit the new GMM estimator and obtain its asymptotic variance for constructing standard errors and confidence intervals. This variant of our method is implemented as the function `predicted_covariates_aux_first` in the companion `R` package.

D Labeled Data with Sample Selection

The estimator introduced in the main text relies on the assumption that the sample indicators (p, v, t) are drawn independently of the other variables in the analysis. This is comparable to x_u being missing completely at random (MCAR). We will assume the linear projection of y on x in the primary population (hereafter just β) is the quantity of interest, since the linear projection in the labeled population can be estimated consistently using OLS in the labeled sample.

In general, without MCAR, the GMM estimator offered in the main text is not consistent for β. This is for two reasons: first, the OLS moment condition in the GMM is no longer relevant for

β, since the linear projection in the primary and labeled populations can be different. Second, the 2SLS moment condition requires the validation sample to estimate Γ, the linear projection of x on z in the primary population, in the first stage. However, the linear projection of x on z could be different in the primary and labeled populations, so the 2SLS estimator first stage is inconsistent (rendering the second stage inconsistent).

However, if the researcher is willing to assume that the linear projection of x on z is identical in the primary and validation populations, then some progress can usually be made. Suppose that

$$x = \Gamma z + \eta, \quad \mathbb{E}[z\eta^{\mathsf{T}} \mid p = 1] = \mathbb{E}[z\eta^{\mathsf{T}} \mid v = 1] = 0. \tag{26}$$

If this is the case, then we can use the 2SLS estimator to estimate β as normal. We must still disregard the OLS estimator, however, because we have not assumed the linear projection of y on x is the same in the primary population as in the labeled population.

What is the substantive meaning of assuming the linear projection of x on z is the same across validation and primary populations? It is most believable when we know that the conditional mean of x given z is linear - that is, when $\mathbb{E}[x \mid z, v = 1] = \mathbb{E}[x \mid z, p = 1] = \Gamma z$. Helpfully, the conditional mean is guaranteed to be linear if z is a binary variable. When the conditional mean of x given z is linear, then a sufficient condition for the two populations having the same conditional mean is that the sample indicators do not depend on x after z is conditioned on; that is, $p(v = 1 \mid x, z) = p(v = 1 \mid z)$. That is, the predictions z capture enough information about the sampling process vis-a-vis x that knowledge of x would not help us any further. This is an application of the idea of exogenous sampling versus endogenous sampling (Cameron and Trivedi, 2005, Ch. 24).

Estimating β in this scenario is simpler than the GMM estimator in the main text. There is now only one moment condition: $\mathbb{E}[z(y - (\Gamma z)^{\mathsf{T}}\beta)] = 0$. If we assume that the predictions z and the original covariates x are of the same dimension, then the number of moments is equal to the

dimension of β and there is a method-of-moments estimator is available in closed form:

$$\hat{\beta}_{\text{MM}} = \left(\sum_{i=1}^{n} p_i z_i z_i^{\mathsf{T}} \hat{\Gamma}^{\mathsf{T}} \right)^{-1} \left(\sum_{i=1}^{n} p_i z_i y_i \right) \tag{27}$$

where $\hat{\Gamma}$ is the linear projection of x on z estimated on the validation sample. Note that since $d_x = d_z$ and there is no other moment condition, the GMM estimator with any weighting matrix would always simplify to this method-of-moments estimator.

The convergence of this estimator depends on both n_v and n_p. The asymptotic theory for this estimator requires slight modifications to the formulas given in Online Appendix A. In particular, there is no longer an OLS component for the GMM estimator, nor is the validation sample included in the estimation of the second stage.

The asymptotic expansion of $\hat{\beta}_{\text{MM}}$ is found to be

$$\sqrt{n_v}(\hat{\beta}_{\text{MM}} - \beta) \tag{28}$$

$$= -\frac{1}{\sqrt{n_v}} \mathbb{E}[zz^{\mathsf{T}}\Gamma^{\mathsf{T}} \mid p = 1]^{-1} \sum_{i=1}^{n} \left\{ p_i \frac{n_v}{n_p} z_i (y_i - (\Gamma z_i)^{\mathsf{T}} \beta) + v_i z_i (x_i - \Gamma z_i)^{\mathsf{T}} \beta \right\} + o_p(1) \tag{29}$$

From this, it is easy to see that if $n_v / n_p \to 0$ then the error in the validation sample still dominates the error in $\hat{\beta}_{\text{MM}}$ even though the second stage only uses the primary sample. This expansion implies that the asymptotic distribution of $\hat{\beta}_{\text{MM}}$ is $\sqrt{n_v}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, (\Delta^{\mathsf{T}} \Xi^{-1} \Delta)^{-1})$ where

$$\Delta = \mathbb{E}[zz^{\mathsf{T}}\Gamma^{\mathsf{T}} \mid p = 1] \tag{30}$$

$$\Xi = \mathbb{E}[hh^{\mathsf{T}}] \tag{31}$$

$$h_i = p_i \frac{n_v}{n_p} z_i (y_i - (\Gamma z_i)^{\mathsf{T}} \beta) + v_i z_i (x_i - \Gamma z_i)^{\mathsf{T}} \beta \tag{32}$$

These quantities can be estimated by plugging in the already-computed estimators for Γ and β.

E What if the Covariate is a Function of Several Predictions?

The baseline method supposes a one-to-one relationship between classifier outputs and predicted covariates. In some applications, the regression covariate is a function of several missing variables. For example, in **?**, the goal is to regress the proportion of an ethnicity in a politician's constituency on the proportion of that ethnicity in the politician's photos. But the ethnicity of a constituent in any given photo is unknown. **?** use a convolutional neural network to predict the ethnicity of the constituents in each photo and then average over these predictions to predict the proportion.

Our proposed GMM can be used for this task, but it requires careful planning on the part of the analyst. The missing covariate (in this case, the proportion of a given legislator's photos featuring a given ethnicity) must be observed for at least some observations. However, there are many photos for each legislator, so hand-labeling all of a legislator's photos is a time-consuming task which should be performed for as few legislators as possible. Fortunately, it is possible to train the classifier at the photo-level, preserving all legislators whose true averages are known for the validation sample. This can be achieved by the following procedure:

1. Hand-label a simple random sample of all photos. This simple random sample is the training sample.

2. Train the classifier using only photos from the training sample.

3. For a simple random sample of all legislators, hand-label all photos associated with that legislator. For each of these legislators, the proportion of photos featuring the desired ethnicity is the true label.

4. For legislators outside of the validation sample, estimate the proportion of photos featuring the desired ethnicity by taking the average prediction of all of the legislator's photos *excluding those in the training sample.*

5. Run the GMM fitting the OLS moment conditions on just the validation sample. The TSLS moment conditions should be fit on both the validation and primary samples, as usual.

Note that, in this procedure, the photos used to train the classifier are effectively excluded from the analysis after training. This does not affect bias or consistency, because the training photos are a simple random sample of all photos, and the proportion of a legislator's photos featuring a given ethnicity does not change once the training photos are removed from the sample.

F A Procedure for Non-Representative Training Samples

In some applications, one label is much rarer than the other. For example, in our application to Reddit from Section 5, only 21.0% of the training observations were uncivil posts, and the remaining 79.0% were civil posts. In other applications, the ratio could be even more skewed. Training an accurate classifier requires a sufficient number of examples of both labels, but if one of the labels is very rare, it would be wasteful to hand-label 10,000 observations just to get 500 examples of the rarer label.

For this reason, researchers sometimes desire a statistically unrepresentative training sample so that a higher proportion of the observations have the rare label. In the online incivility application, a researcher might want to oversample on brief comments if he belives that uncivil remarks are disproportionarely likely to be terse. This allows the researcher to gather more examples of incivility for a lower coding cost.

Similarly, researchers might want to use active learning to decide which observations to hand-label. Rather than hand-labeling many observations at once and then training the classifier, researchers might hand label some observations, train a classifier, hand-label the observations for which the classifier expresses the greatest uncertainty, and repeat.

Both of these approaches would of course preclude the use of the labeled-only estimator, because the hand-labeled observations would no longer be a simple random sample of the full sample. Moreover, these approaches could affect the plausibility of the exclusion restriction.

For oversampling, if the characteristics on which the researcher oversamples are correlated with the outcome, then the prediction error could be correlated with the outcome in violation of the exclusion restriction. Such a correlation could be induced automatically and unwittingly by an active learning procedure.

However, if the researcher can show that the procedure by which they construct their non-representative training sample does not induce a violation of the exclusion restriction, two challenges still remain. First, the linear projection within the training sample is no longer the same as in the full sample, because the training sample is a non-representative sample of the full sample. Second, drawing a non-representative training sample necessarily makes the either the validation sample or primary sample or both non-representative. Returning to the Reddit example, if the training sample has shorter posts than the full sample average, then either the validation sample, the primary sample, or both must have longer posts than the full sample average.

Fortunately, these issues can be avoided through appropriate planning and a slight adjustment to the GMM. A researcher who intends to oversample or use active learning should start by drawing a training sample completely at random from the full sample. This ensures that the validation sample and primary samples are still drawn from the same distribution as the full sample, and hence the linear projection of y on $\boldsymbol{x}$ is the same in them as it is in the full sample. Then, within the representative training sample, the researcher can hand-label whichever observations they like. They can determine which observations to hand-label by oversampling, active learning, or any other procedure. Then, after labeling all of the validation sample and generating predictions for the validation and primary samples, the researcher must exclude the training sample from the GMM. Formally, the first set of moment conditions, g_1, should be changed to

$$g_1(\boldsymbol{b}) = \frac{1}{n_v} \sum_{i=1}^{n} v_i \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^\top \boldsymbol{b}),$$

This is accomplished in our `R` package by setting `t = 0` for all observations (telling the package

there is no training sample for the purpose of GMM calculations).

G Related Methods

Several authors have suggested related ideas to ours for different settings. The most similar is Lee and Sepanski (1995), who show that, given $\mathbb{E}[y \mid \boldsymbol{x}] = m(\boldsymbol{x}, \boldsymbol{\beta})$ for some general (nonlinear) function m, a valid proxy for $m(\boldsymbol{x}, \boldsymbol{b})$ is the linear projection of $m(\boldsymbol{x}, \boldsymbol{b})$ on $\boldsymbol{z}$, to be estimated on validation data. When m is linear, their approach reduces to the 2SLS component of our GMM estimator. Our method could be modified to account for nonlinear $m(\boldsymbol{x}, \boldsymbol{\beta})$ as well; however, this would require estimating a separate first-stage parameter $\boldsymbol{\Gamma}(\boldsymbol{b})$ for each $\boldsymbol{b}$ value the GMM considers. Since most applications in political science use linear regression, we relegate the nonlinear version of our GMM estimator to future work. However, Lee and Sepanski (1995) are motivated by a (more general) scenario in which the primary sample data is drawn from a different population than the validation sample, and therefore do not consider the optimal combination of the primary sample with the validation sample as in our GMM estimator. Additionally, Lee and Sepanski (1995) do not consider the overfitting problem inherent to machine learning applications. They make no distinction between training and validation samples — since in their context measurement error comes from having a poor, exogenously-given measure of $\boldsymbol{x}$ rather than an endogenously-determined algorithmic prediction $\boldsymbol{z}_u$.

Our approach is similar to Lee and Sepanski (1995) because both approaches require making an exclusion restriction, but there are also errors-in-variables corrections that relax the exclusion restriction. Chen, Hong and Tamer (2005) propose a nonparametric series GMM estimator (CEP-GMM) that only requires the conditional distribution of $\boldsymbol{x}$ given $(y, \boldsymbol{z})$ remain constant across the primary and validation populations. Instead of predicting $\boldsymbol{x}$ with $\boldsymbol{z}$, the CEP-GMM estimator nonparametrically estimates $\mathbb{E}[\boldsymbol{x}(y - \boldsymbol{x}^\mathsf{T}\boldsymbol{b}) \mid y, \boldsymbol{z}]$ and replaces the $g_2(\boldsymbol{b})$ component of our GMM with the primary sample average of the estimates $\widehat{\mathbb{E}}[\boldsymbol{x}(y - \boldsymbol{x}^\mathsf{T}\boldsymbol{b}) \mid y = y_i, \boldsymbol{z} = \boldsymbol{z}_i]$. Like us, Chen, Hong and Tamer (2005) provide a GMM estimator that combines the primary and validation

sample data when they are drawn from the same population, but like Lee and Sepanski (1995) they do not consider overfitting and thus do not have a training sample. Chen, Hong and Tarozzi (2008) shows the CEP-GMM estimator achieves the semiparametric efficiency bound, given the aforementioned assumption of equal conditional distributions. Unfortunately, in our simulations in Online Appendix H, we found the CEP-GMM did not offer meaningful efficiency improvements over the labeled-only estimator.

A popular literature in statistics would cast this not as a measurement error problem, but as a more general problem of missing data. In our case, the values of $\boldsymbol{x}$ in the primary sample and $\boldsymbol{z}$ in the training sample are treated as missing. There are numerous model-based corrections for missing data that would try to improve upon the efficiency of the labeled-only estimator (which corresponds to complete-cases analysis or listwise deletion for regressing y on $\boldsymbol{x}$). Modern approaches include multiple imputation, full information maximum likelihood, and fully Bayesian approaches that treat missing data as if they are unknown parameters; see Ibrahim et al. (2005) for a review of model-based missing data approaches in the context of generalized linear models (GLMs). The parametric versions of these methods, which are by far the most used in practice, often require correctly specifying a parametric model for the unknown data generating process (DGP). While our GMM approach also uses and estimates Euclidean parameters, we are careful to never make assumptions about the DGP. The linear models of y on $\boldsymbol{x}$ and x on z are all defined without loss of generality (linear projections always exist), and the proof of our GMM estimator's consistency and the derivation of its variance make no assumptions about the distribution of errors ϵ or $\boldsymbol{\eta}$ other than cross-sectional independence. Our simulation experiments in Online Appendix H suggest that methods like parametric multiple imputation are highly sensitive to assumptions made about the DGP.

A developing literature on nonparametric multiple imputation that makes fewer assumptions about the DGP may eventually sidestep these concerns, but for now these methods are not in wide use among the applied community because of the required expertise and computational resources (e.g., Murray and Reiter, 2016; Yoon, Jordon and van der Schaar, 2018). Additionally,

nonparametric methods require more validation data than parametric methods to learn the relationship between the observed and unobserved variables, but due to budget constraints the labeled sample (and thus the validation sample) is often small relative to the primary sample. Our proposed GMM estimator strikes a balance by not making strong assumptions about the DGP while only requiring the small validation sample to estimate a Euclidean parameter $\mathbf{\Gamma}$ with least squares.

Finally, the form of our method is similar to that of Kane, Rouse and Staiger (1999), who use a GMM estimator to correct for non-classical measurement error. However, their use case requires *two* independent predictions of the true covariate, while ours requires only one prediction. Their estimator is likely much more efficient than ours (convergence in n_p rather than n_v), but as just stated their method requires more data collection than our method assumes. We do not think it plausible that most applied researchers can find multiple independent predictions satisfying the exclusion restriction.

H Simulations Against Competing Methods

Many seemingly promising approaches to addressing measurement error fail when the ratio of labeled to unlabeled data is large, as is typically the case in machine learning applications. Small violations of distributional assumptions propagate into massive errors in the resulting estimates. The table below presents simulations that underscore this point. These simulations are constructed according to those with a realistic (72% accurate) classifier and a low signal-to-noise ratio with non-Gaussian errors, as in the main simulations detailed in Section 4. $n_v = 1000$ and $n_t = 0$, to account for the fact that some of the competitors do not readily accommodate data where the relationship between the true label and the predicted label is different from the rest of the sample.

The three competitors presented are:

- Fully Bayesian: Specify a diffuse prior for all parameters and the conditional distribution

Estimator	Bias (n_p = 10K)	RMSE (n_p = 10K)	Bias (n_p = 1M)	RMSE (n_p = 1M)
Naive	-0.501	0.542	-0.559	0.560
Labeled-Only	0.005	0.686	0.005	0.686
GMM	0.022	0.411	0.005	0.084
Fully-Bayesian	0.011	0.393	0.458	3.449
Multiple Imputation	-0.731	0.743	-0.806	0.806
Mak-Li	-0.003	0.630	0.017	0.642
Oracle	0.008	0.203	-0.000	0.023

Table 1: A Comparison of Methods for Addressing Measurement Error: $n_v = 1K$ and $n_t = 0$.

of the variables, then sample from their conditional distribution (Ibrahim et al., 2005). Priors are given by $\beta \sim (N(0, 10), N(0, 10))$, $\sigma \sim$ Cauchy$(0, 2.5)$, $\zeta_{1,1} = Pr(x = 1|z = 1) \sim$ Uniform$(0, 1)$, $\zeta_{0,0} = Pr(x = 0|z = 0) \sim$ Uniform$(0, 1)$. The conditional distribution of y is modeled as a mixture over whether z is correct (with weight $\zeta_{1,1}$) or incorrect (with weight $\zeta_{0,0}$). β is estimated to be the mode of β's posterior distribution. This approach requires a distributional assumption on the error term; we use the Gaussian so that $y|x$ follows the normal distribution. Note that this choice means the model is misspecified, because the errors are not Gaussian. However, the errors distribution is close to normal (with a heavy right tail), so the misspecification is not too severe. The Fully Bayesian approach works extremely well when all of the distributional assumptions are correct, but the assumptions are not guaranteed to be correct in real data, which makes these results more informative.

- Multiple Imputation: Multiple imputation as implemented in the Amelia package (Honaker, King and Blackwell, 2011), using the default option of taking five imputed data sets and averaging over them.

- Mak-Li: A double-sampling estimator from statistics that is theoretically well-suited for the problem of misclassification error if the true label is known for some observations, (Mak and Li, 1988).

The results of 1000 simulations for each setup show that all three competitors perform poorly, not only when compared with the GMM but also when compared to the baseline of not using the unlabeled data at all. None of their performance improves in response to adding unlabeled data,

Estimator	Bias (n_p = 10K)	RMSE (n_p = 10K)	Bias (n_p = 1M)	RMSE (n_p = 1M)
Labeled-Only	-0.016	0.396	-0.016	0.396
GMM	-0.012	0.142	0.001	0.002
CHT	-0.019	0.384	-0.019	0.383

Table 2: Evaluation of the Parametric CHT Estimator

and the performance of the fully Bayesian estimator and multiple imputation degrades substantially.

The Bayesian estimator appears to offer better performance for small sample sizes, but this is because the errors are nearly Gaussian and hence the functional form assumption it requires is nearly correct. Using, for example, an error distribution of Gamma$(1, 1) - 1$, with $10K$ unlabeled observations, the Fully-Bayesian estimator has an RMSE of 0.127 to 0.041 for the GMM.

Table 2 compares the performance of the GMM to an estimator proposed by Chen, Hong and Tamer (2005), which theoretically should perform well at this task. This estimator is particularly attractive because it does not require an exclusion restriction. To evaluate the performance of this estimator, we implemented a parametric version that used logistic regression to estimate the conditional expected value of the label. Our simulation experiments give the CEP-GMM an additional advantage by giving it the correctly-specified conditional distribution of $\boldsymbol{x}$ given $(y, \boldsymbol{z})$ rather than having CEP-GMM estimate it nonparametrically. We designed simulation settings in which the parametric assumptions were correct, with $y \sim N(0, 10)$, $x_o = 1$, $z_u \sim$ Bernoulli(0.5) and $x_u \sim$ Bernoulli $\left(\frac{\exp(-1+2z)}{1+\exp(-1+2z)}\right)$. All simulations used 1,000 labeled observations, evenly divided between the training set and validation set. Unfortunately, Table 2 shows that it does not offer meaningful gains over simply using OLS in the hand-labeled data.

I Additional Simulations

Tables 3-11 present a more complete set of simulations. Unlike the simulations presented in Section 4 (which are a subset of the simulations in the tables), these simulations vary the size of the validation sample and the size of the training sample. They also include results for an

uninformative classifier with an accuracy of 0.52.

These simulations show that the performance of all estimators is improving as the size of the validation and training samples grows. The GMM performs about as well as the labeled-only estimator with the uninformative classifier, although estimation error leads to some bias when the primary sample is very large and results in undercoverage from the confidence intervals. It compensates for this bias with reduced variance. Tables 12 and 13 provide additional results to further illustrate the relationship between classifier accuracy and performance.

We also provide simulations for classifiers where the classes are imbalanced. For these simulations, the true label is equal to 1 20% of the time and equal to 0 80% of the time. The RMSE is generally higher for imbalanced classes. However, the relationships between parameters and performance are the same as in the simulations with balanced classes, as is the substantive relationship between different classifiers.

J Simulations with Exclusion Restriction Violated

The proposed GMM leverages the assumption that the predictions and the regression residual are uncorrelated. We have shown that a small violation of the exclusion restriction leads to a small bias, but this leaves open the practical question of what counts as small. To study this question empirically, we generate data according to the following procedure:

- $x = 0.5$ for exactly half of all observations and $z \sim \text{Bernoulli}(0.72x + 0.28(1 - x))$.
- Obtain ξ as the residual of the regression of z on x. This is the part of z that is uncorrelated with x.
- $\epsilon \sim \text{Normal}(0, 8) + \text{Bernoulli}(0.15) \times |\text{Normal}(0, 20)|$.
- $y = x + \gamma\xi + \epsilon$.

The data are divided into $1,000$ training observations, $1,000$ validation observations, and either $10,000$ or $1,000,000$ primary observations, depending on the setup. γ determines the severity of

Table 3: Bias: Uninformative Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.87	0.00	-0.00	0.00	-0.87	0.01	-0.01	0.01
		1000000	-0.96	0.00	-0.00	-0.00	-0.96	0.01	-0.18	-0.00
	1000	10000	-0.83	0.00	-0.00	0.00	-0.83	0.01	0.00	0.01
		1000000	-0.96	0.00	-0.00	-0.00	-0.96	0.01	-0.14	-0.00
1000	1000	10000	-0.80	-0.00	-0.00	0.00	-0.79	0.01	0.01	-0.00
		1000000	-0.96	-0.00	-0.00	-0.00	-0.96	0.01	-0.08	0.00
	2000	10000	-0.74	-0.00	-0.00	0.00	-0.75	-0.01	-0.01	-0.01
		1000000	-0.96	-0.00	-0.00	-0.00	-0.96	-0.01	-0.07	-0.00

Table 4: RMSE: Uninformative Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.87	0.06	0.06	0.02	0.89	0.73	0.73	0.20
		1000000	0.96	0.06	0.06	0.00	0.96	0.73	0.71	0.02
	1000	10000	0.83	0.05	0.05	0.02	0.85	0.58	0.57	0.20
		1000000	0.96	0.05	0.05	0.00	0.96	0.58	0.57	0.02
1000	1000	10000	0.80	0.04	0.04	0.02	0.82	0.46	0.46	0.19
		1000000	0.96	0.04	0.04	0.00	0.96	0.46	0.44	0.02
	2000	10000	0.74	0.04	0.04	0.02	0.77	0.40	0.39	0.19
		1000000	0.96	0.04	0.04	0.00	0.96	0.40	0.37	0.02

Table 5: Coverage of 95% Confidence Intervals: Uninformative Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.96	0.95	0.95	0.02	0.93	0.93	0.96
		1000000	0.00	0.96	0.95	0.95	0.00	0.93	0.77	0.94
	1000	10000	0.00	0.95	0.96	0.96	0.01	0.94	0.94	0.95
		1000000	0.00	0.95	0.95	0.95	0.00	0.94	0.79	0.95
1000	1000	10000	0.00	0.96	0.96	0.96	0.03	0.96	0.96	0.96
		1000000	0.00	0.96	0.96	0.95	0.00	0.96	0.88	0.96
	2000	10000	0.00	0.95	0.95	0.95	0.03	0.95	0.95	0.96
		1000000	0.00	0.95	0.95	0.95	0.00	0.95	0.91	0.95

Table 6: Bias: Realistic Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.51	0.00	-0.00	0.00	-0.50	0.01	0.02	0.01
		1000000	-0.56	0.00	-0.00	-0.00	-0.56	0.01	0.01	-0.00
	1000	10000	-0.49	0.00	-0.00	0.00	-0.48	0.01	0.02	0.01
		1000000	-0.56	0.00	-0.00	-0.00	-0.56	0.01	0.01	-0.00
1000	1000	10000	-0.47	-0.00	-0.00	0.00	-0.46	0.01	0.01	-0.00
		1000000	-0.56	-0.00	-0.00	-0.00	-0.56	0.01	-0.00	0.00
	2000	10000	-0.43	-0.00	-0.00	0.00	-0.44	-0.01	-0.02	-0.01
		1000000	-0.56	-0.00	-0.00	-0.00	-0.56	-0.01	0.00	-0.00

Table 7: RMSE: Realistic Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.51	0.06	0.05	0.02	0.55	0.73	0.44	0.20
		1000000	0.56	0.06	0.05	0.00	0.56	0.73	0.10	0.02
	1000	10000	0.49	0.05	0.05	0.02	0.52	0.58	0.38	0.20
		1000000	0.56	0.05	0.04	0.00	0.56	0.58	0.10	0.02
1000	1000	10000	0.47	0.04	0.04	0.02	0.50	0.46	0.36	0.19
		1000000	0.56	0.04	0.04	0.00	0.56	0.46	0.08	0.02
	2000	10000	0.43	0.04	0.03	0.02	0.48	0.40	0.31	0.19
		1000000	0.56	0.04	0.03	0.00	0.56	0.40	0.08	0.02

Table 8: Coverage of 95% with the Confidence Intervals: Realistic Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.96	0.95	0.95	0.32	0.93	0.94	0.96
		1000000	0.00	0.96	0.95	0.95	0.00	0.93	0.96	0.94
	1000	10000	0.00	0.95	0.95	0.96	0.36	0.94	0.95	0.95
		1000000	0.00	0.95	0.94	0.95	0.00	0.94	0.96	0.95
1000	1000	10000	0.00	0.96	0.96	0.96	0.35	0.96	0.95	0.96
		1000000	0.00	0.96	0.95	0.95	0.00	0.96	0.94	0.96
	2000	10000	0.00	0.95	0.95	0.95	0.37	0.95	0.95	0.96
		1000000	0.00	0.95	0.95	0.95	0.00	0.95	0.94	0.95

Table 9: Bias: Best-Case Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.18	0.00	-0.00	0.00	-0.18	0.01	0.01	0.01
		1000000	-0.20	0.00	-0.00	-0.00	-0.20	0.01	0.00	-0.00
	1000	10000	-0.17	0.00	-0.00	0.00	-0.16	0.01	0.02	0.01
		1000000	-0.20	0.00	-0.00	-0.00	-0.20	0.01	0.00	-0.00
1000	1000	10000	-0.17	-0.00	-0.00	0.00	-0.17	0.01	-0.00	-0.00
		1000000	-0.20	-0.00	-0.00	-0.00	-0.20	0.01	0.00	0.00
	2000	10000	-0.15	-0.00	-0.00	0.00	-0.16	-0.01	-0.01	-0.01
		1000000	-0.20	-0.00	-0.00	-0.00	-0.20	-0.01	-0.00	-0.00

Table 10: RMSE: Best-Case Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.18	0.06	0.04	0.02	0.27	0.73	0.26	0.20
		1000000	0.20	0.06	0.03	0.00	0.20	0.73	0.04	0.02
	1000	10000	0.17	0.05	0.03	0.02	0.26	0.58	0.25	0.20
		1000000	0.20	0.05	0.03	0.00	0.20	0.58	0.04	0.02
1000	1000	10000	0.17	0.04	0.03	0.02	0.26	0.46	0.24	0.19
		1000000	0.20	0.04	0.02	0.00	0.20	0.46	0.04	0.02
	2000	10000	0.15	0.04	0.03	0.02	0.25	0.40	0.23	0.19
		1000000	0.20	0.04	0.02	0.00	0.20	0.40	0.04	0.02

Table 11: Coverage of 95% Confidence Intervals: Best-Case Classifier

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.96	0.94	0.95	0.87	0.93	0.95	0.96
		1000000	0.00	0.96	0.95	0.95	0.00	0.93	0.95	0.94
	1000	10000	0.00	0.95	0.97	0.96	0.89	0.94	0.95	0.95
		1000000	0.00	0.95	0.95	0.95	0.00	0.94	0.95	0.95
1000	1000	10000	0.00	0.96	0.97	0.96	0.88	0.96	0.96	0.96
		1000000	0.00	0.96	0.95	0.95	0.00	0.96	0.96	0.96
	2000	10000	0.00	0.95	0.96	0.95	0.86	0.95	0.95	0.96
		1000000	0.00	0.95	0.95	0.95	0.00	0.95	0.94	0.95

Table 12: Bias as a Function of Accuracy

		High Signal-to-Noise				Low Signal-to-Noise			
Accuracy	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
0.52	10000	-0.800	-0.000	-0.001	0.001	-0.790	0.007	0.006	-0.005
0.60	10000	-0.666	-0.000	-0.001	0.001	-0.659	0.007	0.014	-0.005
0.72	10000	-0.466	-0.000	-0.001	0.001	-0.462	0.007	0.014	-0.005
0.80	10000	-0.333	-0.000	-0.001	0.001	-0.334	0.007	0.004	-0.005
0.92	10000	-0.166	-0.000	-0.001	0.001	-0.170	0.007	-0.003	-0.005
0.52	1000000	-0.958	-0.000	-0.002	-0.000	-0.959	0.007	-0.080	0.000
0.60	1000000	-0.798	-0.000	-0.001	-0.000	-0.799	0.007	0.003	0.000
0.72	1000000	-0.559	-0.000	-0.000	-0.000	-0.559	0.007	-0.000	0.000
0.80	1000000	-0.399	-0.000	-0.001	-0.000	-0.399	0.007	-0.002	0.000
0.92	1000000	-0.200	-0.000	-0.001	-0.000	-0.199	0.007	0.001	0.000

Table 13: RMSE as a Function of Accuracy

		High Signal-to-Noise				Low Signal-to-Noise			
Accuracy	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
0.52	10000	0.800	0.043	0.043	0.017	0.816	0.465	0.463	0.193
0.60	10000	0.667	0.043	0.042	0.017	0.690	0.465	0.444	0.193
0.72	10000	0.466	0.043	0.038	0.017	0.503	0.465	0.355	0.193
0.80	10000	0.334	0.043	0.034	0.017	0.387	0.465	0.297	0.193
0.92	10000	0.167	0.043	0.027	0.017	0.258	0.465	0.240	0.193
0.52	1000000	0.958	0.043	0.043	0.002	0.960	0.465	0.442	0.022
0.60	1000000	0.798	0.043	0.042	0.002	0.800	0.465	0.178	0.022
0.72	1000000	0.559	0.043	0.036	0.002	0.560	0.465	0.083	0.022
0.80	1000000	0.399	0.043	0.030	0.002	0.400	0.465	0.055	0.022
0.92	1000000	0.200	0.043	0.021	0.002	0.200	0.465	0.037	0.022

the violation of the exclusion restriction; the larger γ, the larger the violation of the exclusion restriction. We provide simulations with $\gamma \in \{-0.05, -0.10, -0.20, -0.40\}$. γ is negative because the naive estimator is, in this case, biased downwards; a positive γ would partially offset that bias and make the naive estimator seem more attractive than it really is. These lead the correlation between z and ϵ to be about 0.04, 0.08, 0.16, and 0.32 times as strong as the correlation between x and y, respectively, and therefore range from minor violations of the exclusion restriction to severe violations.

Figures 1-3 present the results of these simulations. Small violations of the exclusion restriction lead to relatively small biases, and the improved variance of the estimator compensates for this bias. For these small violations, the coverage of the 95% confidence intervals for the GMM

Table 14: Bias: Uninformative Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.92	0.00	-0.00	0.00	-0.91	0.03	0.01	0.00
		1000000	-0.97	0.00	-0.01	0.00	-0.97	0.03	-0.22	0.00
	1000	10000	-0.89	0.00	-0.00	0.00	-0.88	0.01	-0.00	0.01
		1000000	-0.97	0.00	-0.00	0.00	-0.97	0.01	-0.18	0.00
1000	1000	10000	-0.87	0.00	-0.00	0.00	-0.86	0.02	0.02	0.00
		1000000	-0.97	0.00	-0.00	0.00	-0.97	0.02	-0.10	0.00
	2000	10000	-0.83	0.00	-0.00	0.00	-0.82	0.00	-0.00	-0.02
		1000000	-0.97	0.00	-0.00	0.00	-0.97	0.00	-0.08	0.00

Table 15: RMSE: Uninformative Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.92	0.08	0.08	0.02	0.94	0.84	0.83	0.26
		1000000	0.97	0.08	0.08	0.00	0.97	0.84	0.81	0.03
	1000	10000	0.89	0.06	0.06	0.02	0.91	0.72	0.71	0.26
		1000000	0.97	0.06	0.06	0.00	0.97	0.72	0.70	0.03
1000	1000	10000	0.87	0.06	0.06	0.02	0.89	0.61	0.60	0.25
		1000000	0.97	0.06	0.06	0.00	0.97	0.61	0.58	0.03
	2000	10000	0.83	0.04	0.05	0.02	0.85	0.50	0.50	0.24
		1000000	0.97	0.04	0.05	0.00	0.97	0.50	0.49	0.03

Table 16: Coverage of 95% Confidence Intervals: Uninformative Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.97	0.96	0.95	0.01	0.95	0.95	0.96
		1000000	0.00	0.97	0.95	0.95	0.00	0.95	0.80	0.95
	1000	10000	0.00	0.96	0.96	0.95	0.01	0.96	0.96	0.94
		1000000	0.00	0.96	0.95	0.95	0.00	0.96	0.81	0.94
1000	1000	10000	0.00	0.95	0.95	0.95	0.01	0.95	0.95	0.95
		1000000	0.00	0.95	0.94	0.95	0.00	0.95	0.89	0.95
	2000	10000	0.00	0.95	0.95	0.95	0.02	0.94	0.94	0.94
		1000000	0.00	0.95	0.95	0.96	0.00	0.94	0.90	0.95

Table 17: Bias: Realistic Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.66	0.00	-0.00	0.00	-0.66	0.03	-0.00	0.00
		1000000	-0.70	0.00	-0.00	0.00	-0.70	0.03	0.01	0.00
	1000	10000	-0.64	0.00	0.00	0.00	-0.62	0.01	0.02	0.01
		1000000	-0.70	0.00	0.00	0.00	-0.70	0.01	0.01	0.00
1000	1000	10000	-0.62	0.00	-0.00	0.00	-0.61	0.02	0.02	0.00
		1000000	-0.70	0.00	0.00	0.00	-0.70	0.02	0.00	0.00
	2000	10000	-0.59	0.00	-0.00	0.00	-0.59	0.00	-0.01	-0.02
		1000000	-0.70	0.00	-0.00	0.00	-0.70	0.00	0.00	0.00

Table 18: RMSE: Realistic Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.66	0.08	0.07	0.02	0.69	0.84	0.58	0.26
		1000000	0.70	0.08	0.06	0.00	0.70	0.84	0.15	0.03
	1000	10000	0.64	0.06	0.06	0.02	0.66	0.72	0.51	0.26
		1000000	0.70	0.06	0.05	0.00	0.70	0.72	0.15	0.03
1000	1000	10000	0.62	0.06	0.05	0.02	0.64	0.61	0.48	0.25
		1000000	0.70	0.06	0.05	0.00	0.70	0.61	0.11	0.03
	2000	10000	0.59	0.04	0.04	0.02	0.63	0.50	0.42	0.24
		1000000	0.70	0.04	0.04	0.00	0.70	0.50	0.11	0.03

Table 19: Coverage of 95% with the Confidence Intervals: Realistic Classifier with Imbalanced Classes

			Easy				Hard			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.97	0.96	0.95	0.14	0.95	0.95	0.95
		1000000	0.00	0.97	0.95	0.95	0.00	0.95	0.96	0.95
	1000	10000	0.00	0.96	0.96	0.95	0.17	0.96	0.96	0.94
		1000000	0.00	0.96	0.95	0.95	0.00	0.96	0.96	0.94
1000	1000	10000	0.00	0.95	0.95	0.95	0.18	0.95	0.95	0.95
		1000000	0.00	0.95	0.95	0.95	0.00	0.95	0.96	0.95
	2000	10000	0.00	0.95	0.96	0.95	0.17	0.94	0.94	0.94
		1000000	0.00	0.95	0.95	0.96	0.00	0.94	0.96	0.95

Table 20: Bias: Best-Case Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	-0.31	0.00	-0.00	0.00	-0.31	0.03	-0.01	0.00
		1000000	-0.33	0.00	-0.00	0.00	-0.33	0.03	-0.00	0.00
	1000	10000	-0.30	0.00	0.00	0.00	-0.29	0.01	0.01	0.01
		1000000	-0.33	0.00	0.00	0.00	-0.33	0.01	0.00	0.00
1000	1000	10000	-0.29	0.00	-0.00	0.00	-0.28	0.02	0.00	0.00
		1000000	-0.33	0.00	-0.00	0.00	-0.33	0.02	-0.00	0.00
	2000	10000	-0.27	0.00	-0.00	0.00	-0.28	0.00	-0.02	-0.02
		1000000	-0.33	0.00	-0.00	0.00	-0.33	0.00	-0.00	0.00

Table 21: RMSE: Best-Case Classifier with Imbalanced Classes

			High Signal-to-Noise				Low Signal-to-Noise			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.31	0.08	0.05	0.02	0.39	0.84	0.36	0.26
		1000000	0.33	0.08	0.04	0.00	0.34	0.84	0.07	0.03
	1000	10000	0.30	0.06	0.05	0.02	0.37	0.72	0.35	0.26
		1000000	0.33	0.06	0.04	0.00	0.33	0.72	0.07	0.03
1000	1000	10000	0.29	0.06	0.04	0.02	0.36	0.61	0.32	0.25
		1000000	0.33	0.06	0.03	0.00	0.33	0.61	0.05	0.03
	2000	10000	0.27	0.04	0.03	0.02	0.36	0.50	0.31	0.24
		1000000	0.33	0.04	0.03	0.00	0.34	0.50	0.05	0.03

Table 22: Coverage of 95% Confidence Intervals: Best-Case Classifier with Imbalanced Classes

			Easy				Hard			
n_v	n_t	n_p	NV	LAB	GMM	ORCL	NV	LAB	GMM	ORCL
500	500	10000	0.00	0.97	0.96	0.95	0.73	0.95	0.95	0.95
		1000000	0.00	0.97	0.96	0.95	0.00	0.95	0.96	0.95
	1000	10000	0.00	0.96	0.96	0.95	0.76	0.96	0.95	0.94
		1000000	0.00	0.96	0.95	0.95	0.00	0.96	0.96	0.94
1000	1000	10000	0.00	0.95	0.96	0.95	0.78	0.95	0.95	0.95
		1000000	0.00	0.95	0.96	0.95	0.00	0.95	0.96	0.95
	2000	10000	0.00	0.95	0.96	0.95	0.74	0.94	0.96	0.94
		1000000	0.00	0.95	0.97	0.96	0.00	0.94	0.95	0.95

estimator is approximately correct if the amount of unlabeled data is small enough. However, as the violation grows more severe, the bias grows worse, eventually becoming so large that it overwhelms the improvement in variance.

Additionally, for any given violation of the exclusion restriction, as the size of the primary sample relative to the validation sample increases, the resulting bias grows more severe. In these simulations, increasing the size of the primary sample by a factor of 100 doubled the bias.

Unfortunately, the overidentification test of the exclusion restriction presented in Online Appendix B detected the violation of the bias in only a small number of cases. It rejected the null hypothesis that the exclusion restriction was not violated at the 90% level only about 10% of the time. Thus, researchers should not take failure to reject the null hypothesis as confirmation that the exclusion restriction is satisfied. Researchers worried about violations of the exclusion restriction can increase the power of the test by labeling more observations and adding them to the validation sample.

These simulations emphasize that the exclusion restriction is a critical assumption. However, the simulation results of competing methods that do not rely on the exclusion restriction in Online Appendix H shows that it is a necessary assumption to extract meaningful returns from unlabeled data.[3] Researchers who for theoretical or empirical reasons are unwilling to commit to this assumption should consider eschewing machine learning altogether and conducting their analyses using only a simple random sample of hand-labeled data.

K Semi-Synthetic Application: Vote Choice and Homeownership in the 2016 Election

The simulations and theoretical analysis show that the naive estimator is biased and inconsistent and that our proposed GMM estimator is consistent and more efficient than restricting analysis

[3]The Fully Bayesian estimator, like the GMM, is sensitive to violations of the exclusion restriction. In all of the simulation setups reported in this appendix, its bias and RMSE are between 50% and 1000% larger than the bias and RMSE for the GMM.

Figure 1: Bias with Exclusion Restriction Violated

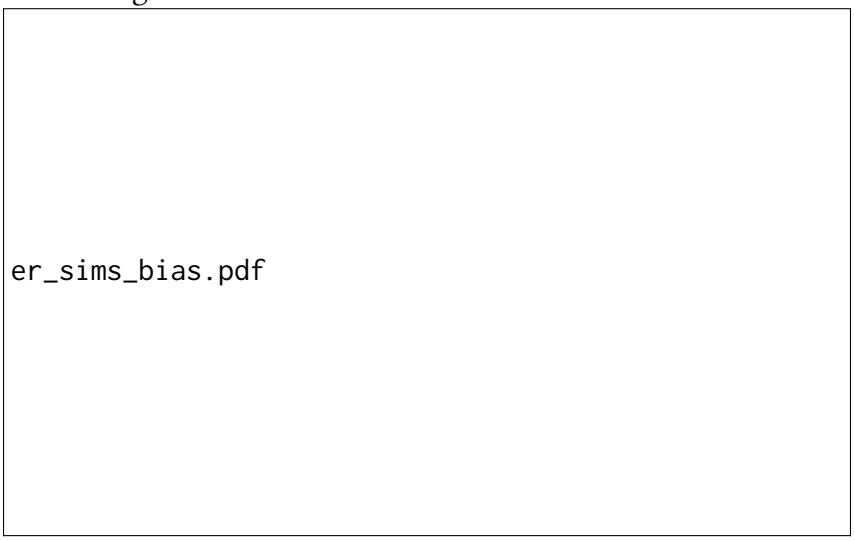


Figure 2: RMSE with Exclusion Restriction Violated

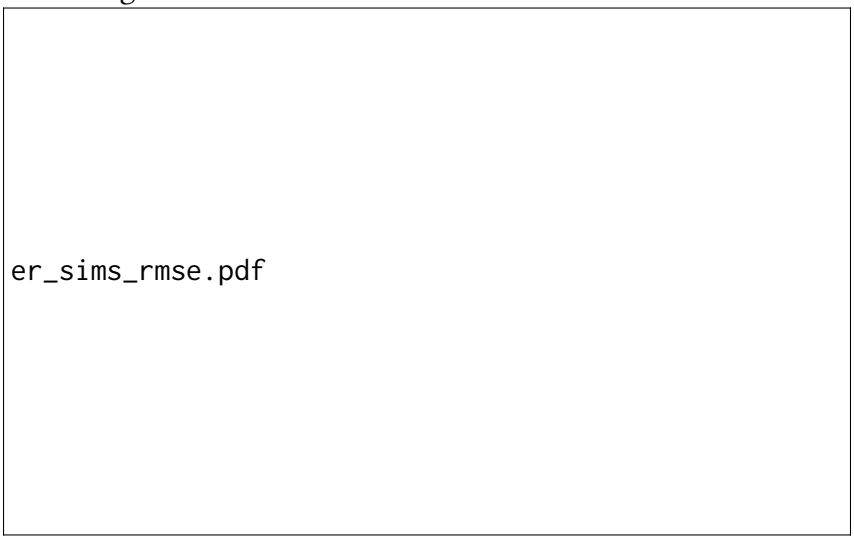


Figure 3: Coverage of 95% Confidence Intervals with Exclusion Restriction Violated

to the hand-labeled sample. However, these arguments do not address whether the purportedly strong assumptions required for the naive estimator are usually satisfied in practice, whether the resulting bias and inconsistency is large enough to affect substantive conclusions, and whether the efficiency gains of the GMM lead to different conclusions than the labeled-only estimator. Accordingly, we consider a validation of our proposed method based on its performance on a real data set answering a descriptive research question.

The question we seek to answer is whether homeownership helps predict presidential vote choice in the 2016 election even after we control for party and race. Scholars have become increasingly interested in the relationship between homeownership and political behavior (e.g., Hall and Yoder, 2019; Marble and Nall, 2020). In this example, we will examine the correlation between homeownership and vote choice conditional on party and white identity.

We obtain the data to answer our question from the 2016 Cooperative Congressional Election Study (Ansolabehere and Schaffner, 2017). We use the CCES measures of respondents' two-party vote choice in the 2016 election (dropping non-two-party voters), their party, self-reported race and ethnic identity (quantified dichotomously as white or non-white), family income (quantified in 2016 US Census quintiles), age, and whether the respondent owns their own home. Statistically, we seek to answer our question using the following regression:

$$\text{Voted for Trump} = \beta_1 \text{Homeowner} + \beta_2 \text{Republican} + \beta_3 \text{White} + \beta_4 + \text{Residual} \tag{33}$$

With β_1 the coefficient of primary interest. A positive value of β_1 implies homeowners were more likely to vote for Trump than Clinton than non-homeowners — even after accounting for party and racial/ethnic identity. A negative value of β_1 would imply the opposite.

Before examining the data, it is not obvious what we should expect the true value or even the sign of β_1 to be. On one hand, we might expect $\beta_1 = 0$ since party identification and racial identity might already capture the average policy preferences of homeowners and non-homeowners. Alternatively, we might think $\beta_1 \neq 0$ due one of the candidates having made a better appeal to

homeowners or non-homeowners to deviate from their usual party preferences. And, of course, we should not forget that homeownership might just be a confounder for another variable (e.g., region) that caused voters to stray from their usual party preferences.

The CCES collects the homeowner variable for every respondent, but many data sets are not so generous. In many administrative data sets, such as voter files, important covariates like homeownership are missing altogether. Accordingly, we conduct an exercise to see how our GMM could help us in these more difficult data sets.

Our goal is to mimic a scenario in which the researchers sought to add homeownership to a political survey that did not ask about homeownership. They have the ability to reinterview respondents from the political survey, but doing so is costly. The researchers also have access to exogenously-generated predictions, perhaps from a commercial vendor, or perhaps from a larger exercise dedicated specifically to predicting homeownership from supplemental data like zip code and income.

To mimic this scenario, we randomly choose 90% of the CCES to be the primary sample ($n_p = 32,634$) and force the homeowner variable to be missing for these respondents. Of the remaining 10%, we randomly assign nine-tenths to the training sample ($n_t = 3,315$) and one-tenth to the validation sample ($n_v = 380$). Next, we fit a random forest classifier on the training set to predict homeownership and apply the algorithm to predict homeownership on the validation and primary samples. We chose a random forest for this task because it is known to perform well in political science applications with a relatively small number of input variables (for an introduction to random forest methods, see Montgomery and Olivella, 2018). We used the randomForest function with the default classification settings from the randomForest `R` package (Liaw and Wiener, 2002). We used income, age, and race as inputs into the classifier to generate our predictions for homeownership status; we did not use party as a classifier input to avoid violating the exclusion restriction (party and vote choice are highly correlated). With a prevalence of homeownership in our CCES subset of about 70%, the random forest classifier we fit has an accuracy of 0.76, a precision of 0.77, a recall/sensitivity of 0.93, and a specificity of 0.36. These

Table 23: Comparing Different Methods

	Voted for Trump [1] instead of Clinton [0]			
	Naive	**Labeled-Only**	**GMM**	**Oracle**
	Model 1	**Model 2**	**Model 3**	**Model 4**
Homeowner	0.023*	0.0001	0.045*	0.051*
	(0.005)	(0.040)	(0.009)	(0.004)
Republican	0.745*	0.695*	0.744*	0.741*
	(0.004)	(0.037)	(0.004)	(0.004)
White	0.095*	0.126*	0.094*	0.094*
	(0.005)	(0.045)	(0.005)	(0.005)
Constant	0.087*	0.130*	0.076*	0.073*
	(0.005)	(0.042)	(0.006)	(0.004)
N	33014	380	33014	36329

statistics suggest the classifier is informative about homeownership, but also that there is still significant prediction error.

After generating the classifier's predictions on the validation and primary samples, we drop the training sample from subsequent analyses. This is to mimic a scenario in which the predictions are derived exogenously from another source, such as in surname-based prediction of race and ethnicity on a voter file (Imai and Khanna, 2016).[4] Thus, our training data is used simply to generate realistic but imperfect predictions, and we suppose that the analyst wishing to run the regression can access only the validation and primary samples. The validation sample consists of a set of homeowners that the researchers reinterviewed to ask about homeownership; the primary sample consists of the rest of the observations.

Table 23 reports the coefficient estimates from four different estimators of eq. (33). Column (1) reports the naive estimates, found by plugging in predicted homeownership for actual homeownership in both the primary and validation samples. Column (2) reports the labeled-only estimates, found by using actual homeownership but only on the validation sample (which is the entire la-

[4]We do not replicate the Imai and Khanna (2016) example as-is because we find that the exclusion restriction is violated in that case, and thus our GMM estimator would not help. The exclusion restriction is violated in that example because non-whites with predicted white surnames in Florida were apparently more likely to have voted in 2008 than other non-whites in Florida.

beled sample, since the training set is dropped). Column (3) reports the GMM estimates based on a combination of OLS in the validation sample and 2SLS in the validation and primary samples.[5] We feel comfortable providing the GMM estimates since the exclusion restriction test p-value is only 0.48, which means that the exclusion restriction cannot be rejected. If the p-value had been much lower, say less than the conventional level of 0.05, we would need to reconsider using the GMM estimator, since the exclusion restriction would would have been rejected. Finally, Column (4) gives the oracle estimates which use actual homeownership for every respondent in the validation and primary samples, which we use as a benchmark for validating each estimate (even though it would be inaccessible in a real application).

Although it is merely one instance of the situations we have been discussing in this paper, we see that in this case the GMM estimates are the most useful. Taking the oracle estimates as a baseline, we see that the naive estimate of the partial relationship between vote choice and homeownership has been severely attenuated and only reaches about half the magnitude of either the GMM or oracle estimates.[6] In other words, the naive estimate understates the rate at which homeowners voted for Trump and overstates the rate at which non-homeowners voted for Trump. This difference in magnitude is substantively meaningful, given that an extra few percentage points in two-party vote share might have changed the outcome of the close 2016 election in pivotal swing states. Next, the labeled-only estimate is very imprecise, and cannot detect whether β_1 is positive, negative, or zero. Finally, the GMM estimate is the closest to the oracle estimate, it is statistically quite distinct from zero, and it achieves a much higher precision than the labeled-only estimate. In fact, the improvement in GMM estimator variance is (in expectation) equivalent to having collected approximately 20 times more validation data ($n_{\text{gmm}} = 7600$ vs. $n_{\text{labeled-only}} = 380$).

[5] Predicted homeownership is significant in the first stage regression with a coefficient of 0.44 (std. error of 0.07).

[6] Based on the sensitivity of 0.93 and the specificity of 0.36, the expected value of the naive estimator of β_1 in the absence of the other two covariates would be $0.28\beta_1$ (Aigner, 1973, eq. 11). The actual estimates are not far off from their hypothetical expected values: the naive estimate is 0.45 times the oracle estimate.

L Reddit-Like Simulations

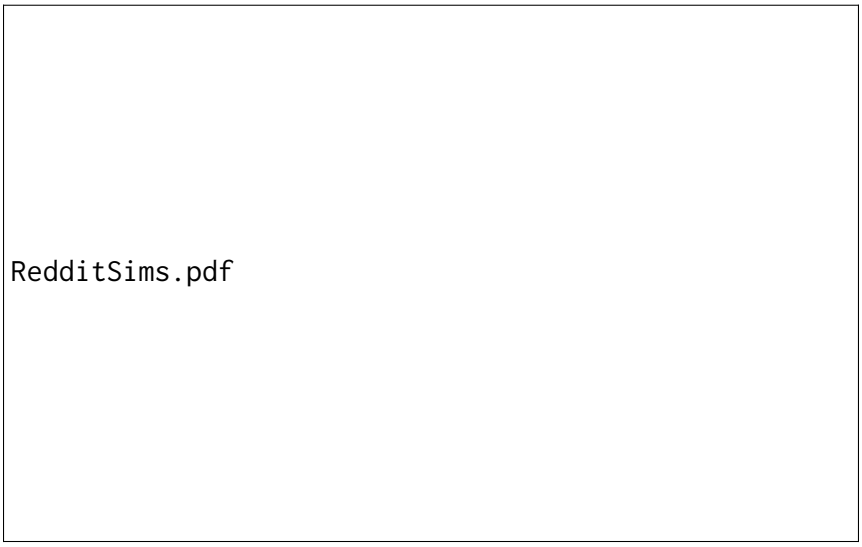
The simulations in Appendix I shows that the GMM estimator generally performs quite well, but it is still worthwhile to test whether it performs well in settings like the Reddit application from Section 5. To address this concern, we test the performance of the GMM estimator in data designed to mimic the data in the Reddit application. In particular, we simulate data where $z_u \sim$ Bernoulli(0.173) and generate x with $\pi_{11} = 0.352$ and $\pi_{00} = 0.863$. $y \sim \mathcal{N}(\beta x, 4.02)$, where $\beta_1 \sim \mathcal{N}(2.82, 0.08)$, $\beta_2 \sim \mathcal{N}(-0.24, 0.18)$, $x = (x_o, x_u)$, $x_o = 1$, and $x_u \sim$ Bernoulli$(\pi_{11}z + (1-\pi_{0,0})(1-z))$. All of these parameter values are drawn from the real Reddit data; the distribution of β comes from the asymptotic distribution implied by the labeled estimate and the error variance for y comes from the mean of the squared residuals in the labeled-only regression.

With this procedure, we generate 2,413 training observations, 613 test observations, and 1,207,140 unlabeled observations for each of 1,000 simulated data sets. We compare the naive, labeled-only, GMM, and oracle estimators across the simulated data sets. The GMM estimator performs far better than the other two feasible estimators, and its analytic confidence interval has the correct coverage. This suggests that we are on firm ground drawing inferences from the two-stage estimator in the application.

M Reddit Subgroup Analysis

In Section 5's analysis of the Reddit data, the GMM proved useful because it was sufficiently powerful to detect a negative relationship between incivility and post score that was statistically different from -1 (which is the only way to rule out the hypothesis that only the target of incivility downvotes the post and other users are indifferent). One reasonable objection to this finding is that the scores may be biased towards 0, because some are in unpopular threads or buried deep in conversations and hence go unobserved by other users. By subsetting to the comments that were most likely to be seen by others, we might be able to observe an effect statistically smaller than -1 using the labeled-only estimator.

Figure 4: Performance of Estimators in Reddit-Like Simulations



Note: In data that closely mimics the Reddit application, the GMM estimator dramatically outperforms the other feasible estimators. Moreover, its confidence interval achieves the correct coverage.

To address this possibility, we present the results of subgroup analyses that divide the comments along three dimensions: (1) whether the comment was a reply to a top-level comment or a reply to a comment that was itself a reply, (2) whether the comment was a reply to a comment in a thread that was above or below the median thread score, and (3) whether the comment was a reply to a comment in a thread that was above or below the median number of comments in the thread. Note that features (2) and (3) are properties of the thread (which contains many comments), not the comment that the post in question is replying to.

However, dividing the sample in this way threatens to limit the statistical power of our tests. In particular, a naive subgroup analyses would divide the validation sample used to learn the first stage of the two-stage least squares in half, even though we expect the relationship between the classifier's prediction and the true incivility label to be the same in both subgroups. To avoid this loss in power, we use all observations to fit the first stage of the two-stage least squares (to learn the linear projection of x on z), but only the relevant subgroup to actually estimate the linear projection of y on x.

Figures 5-7 report the results.[7] If the difference between the labeled-only and GMM estimators were caused by attenuation due to low-visibility comments, we would expect that the labeled-only estimator would have a confidence interval entirely below -1 for replies to top-level comments, popular threads, or high-discussion threads. That is not what we find.

Rather, the labeled-only estimator's 95% confidence interval includes -1 for every subgroup. The GMM's 95% confidence interval is entirely below -1 for every subgroup except replies to top-level comments. But this is not consistent with the hypothesis that the pattern observed in the main results are attributable to attenuation from low-visibility, since the high-visibility posts have point estimates closer to 0 than the low-visibility posts.

[7] The results from the lower-level comments plot should be viewed with skepticism, as the test of the exclusion restriction yields a p-value of 0.107.

Figure 5: Subgroup Analysis Split by Depth of Comment

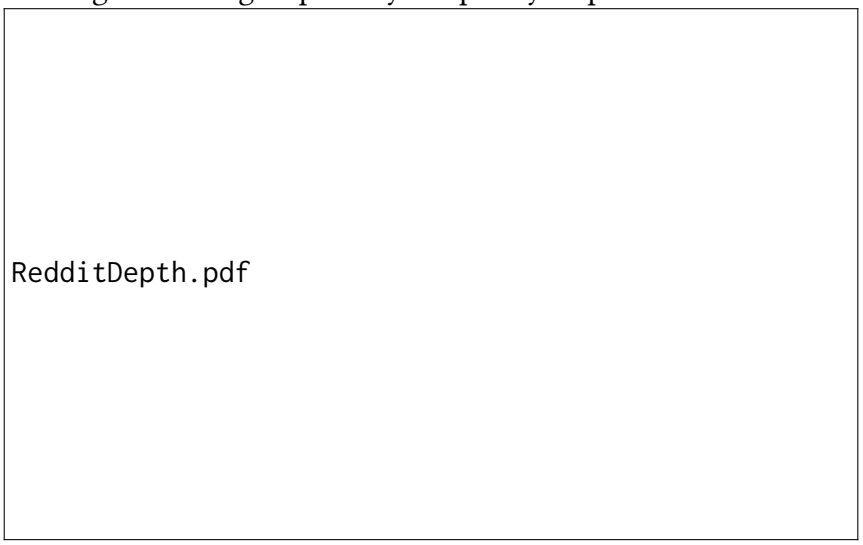


Figure 6: Subgroup Analysis Split by Score of Thread

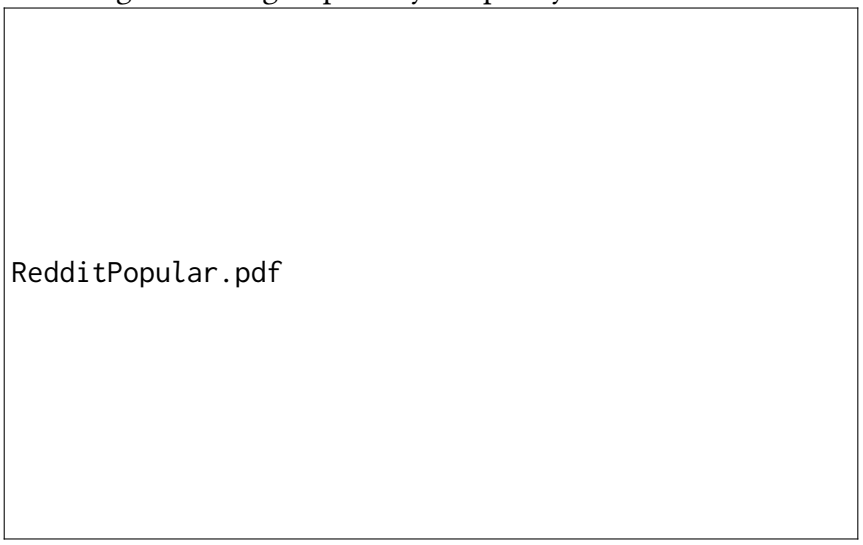


Figure 7: Subgroup Analysis Split by Number of Comments in Thread

References

Aigner, Dennis J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1(1):49–59.

Ansolabehere, Stephen and Brian F. Schaffner. 2017. "CCES Common Content, 2016.".
URL: *https://doi.org/10.7910/DVN/GDF6Z0*

Cameron, A Colin and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chaussé, Pierre. 2010. "Computing generalized method of moments and generalized empirical likelihood with R." *Journal of Statistical Software* 34(11):1–35.

Chen, Xiaohong, Han Hong and Alessandro Tarozzi. 2008. "Semiparametric efficiency in GMM models with auxiliary data." *Annals of Statistics* 36(2):808–843.

Chen, Xiaohong, Han Hong and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *The Review of Economic Studies* 72(2):343–366.

Hall, Andrew B and Jesse Yoder. 2019. "Does Homeownership Influence Political Behavior? Evidence from Administrative Data.".
URL: *https://www.andrewbenjaminhall.com/homeowner.pdf*

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50(4):1029–1054.

Honaker, James, Gary King and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.

Ibrahim, Joseph G, Ming-Hui Chen, Stuart R Lipsitz and Amy H Herring. 2005. "Missing-Data Methods for Generalized Linear Models: A Comparative Review." *Journal of the American Statistical Association* 100(469):332–346.

Imai, Kosuke and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24:263–272.

Kane, Thomas J, Cecilia Elena Rouse and Douglas Staiger. 1999. Estimating returns to schooling when schooling is misreported. Technical report National Bureau of Economic Research.

Lee, Lung-fei and Jungsywan H Sepanski. 1995. "Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data." *Journal of the American Statistical Association* 90(429):130–140.

Liaw, Andy and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.
URL: *https://CRAN.R-project.org/doc/Rnews/*

Mak, TK and WK Li. 1988. "A New Method for Estimating Subgroup Means under Misclassification." *Biometrika* 75(1):105–111.

Marble, William and Clayton Nall. 2020. "Where Self-Interest Trumps Ideology: Liberal Homeowners and Local Opposition to Housing Development.".
URL: *https://williammarble.co/docs/MarbleNallJOP.pdf*

Montgomery, Jacob M and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Murray, Jared S. and Jerome P. Reiter. 2016. "Multiple Imputation of Missing Categorical and Continuous Values via Bayesian Mixture Models With Local Dependence." *Journal of the American Statistical Association* 111(516):1466–1479.

Newey, Whitney K. and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, ed. R. F. Engle and D. L. McFadden. Vol. 4 pp. 2111–2245.

Sargan, John D. 1958. "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica: Journal of the Econometric Society* pp. 393–415.

Yoon, Jinsung, James Jordon and Mihaela van der Schaar. 2018. "GAIN: Missing Data Imputation using Generative Adversarial Nets." *Proceedings of the 35th International Conference on Machine Learning* .