# Machine Learning Predictions as Regression Covariates

**Christian Fong and Matthew Tyler**

Stanford Graduate School of Business and Stanford University

## 1. Current Practice for Regression with an Unobserved Covariate

### Setup

- Task: estimate $\theta_0 = (\beta_0, \alpha_0)$ for $y = x\beta_0 + v'\alpha_0 + \epsilon$
- Challenge: $x$ unobserved, but can be hand-coded
- Examples: text, images, videos, voter files, merging surveys

### Naive Estimator (Common in the Literature)

| Outcome | Covariate | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|
| $y_1$ | | $v_1$ | $q_1$ |
| $y_2$ | | $v_2$ | $q_2$ |
| $y_3$ | | $v_3$ | $q_3$ |
| $y_4$ | | $v_4$ | $q_4$ |

Data is missing a covariate

| Outcome | Covariate | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|
| $y_1$ | $x_1$ | $v_1$ | $q_1$ |
| $y_2$ | $x_2$ | $v_2$ | $q_2$ |
| $y_3$ | | $v_3$ | $q_3$ |
| $y_4$ | | $v_4$ | $q_4$ |

Hand-label covariate for some observations

| Outcome | Covariate | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|
| $y_1$ | $x_1$ | $v_1$ | $q_1$ |
| $y_2$ | $x_2$ | $v_2$ | $q_2$ |
| $y_3$ | $z_3$ | $v_3$ | $q_3$ |
| $y_4$ | $z_4$ | $v_4$ | $q_4$ |

Use machine learning to predict covariate for unlabeled observations from v and q

Effectively fits $y = z\beta_0 + v'\alpha_0 + (\epsilon + e\beta_0)$

### Assumptions Required for Consistency of the Naive Estimator

**Assumption 1** *The prediction $z$ is uncorrelated with the residual $\epsilon$.*

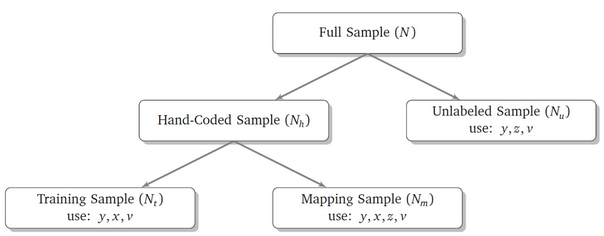
**Assumption 2** *The prediction error $e$ is uncorrelated with the other covariates $v$.*

**Assumption 3** *The prediction error $e$ is mean zero.*

**Assumption 4** *The prediction $z$ is uncorrelated with the prediction error $e$.*

## 2. A GMM-Based Correction That Only Needs (Testable) Assumption 1

### Proposed Sample Splitting




### Two-Stage Least Squares for Unlabeled Data (Use $z$ as an instrument)

| Outcome | Covariate | First-Stage Estimate | Prediction | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|---|---|
| $\mathbf{y_1}$ | $x_1$ | | | $\mathbf{v_1}$ | $\mathbf{q_1}$ |
| $y_2$ | $x_2$ | | | $v_2$ | $q_2$ |
| $y_3$ | | | | $v_3$ | $q_3$ |
| $y_4$ | | | | $v_4$ | $q_4$ |

Train machine learning algorithm in training sample to use $v$ and $q$ to predict $x$

| Outcome | Covariate | First-Stage Estimate | Prediction | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|---|---|
| $y_1$ | $x_1$ | | $z_1$ | $v_1$ | $q_1$ |
| $\mathbf{y_2}$ | $x_2$ | | $\mathbf{z_2}$ | $\mathbf{v_2}$ | $\mathbf{q_2}$ |
| $y_3$ | | | $z_3$ | $v_3$ | $q_3$ |
| $y_4$ | | | $z_4$ | $v_4$ | $q_4$ |

First stage of TSLS: Regress $x$ on $z$ and $v$ in mapping sample to get $\hat{x}$

| Outcome | Covariate | First-Stage Estimate | Prediction | Other Covariate | Text, Pixels, Etc. |
|---|---|---|---|---|---|
| $y_1$ | $x_1$ | $\hat{x}_1$ | $z_1$ | $v_1$ | $q_1$ |
| $\mathbf{y_2}$ | $x_2$ | $\hat{\mathbf{x}}_2$ | $\mathbf{z_2}$ | $\mathbf{v_2}$ | $\mathbf{q_2}$ |
| $\mathbf{y_3}$ | | $\hat{\mathbf{x}}_3$ | $\mathbf{z_3}$ | $\mathbf{v_3}$ | $\mathbf{q_3}$ |
| $\mathbf{y_4}$ | | $\hat{\mathbf{x}}_4$ | $\mathbf{z_4}$ | $\mathbf{v_4}$ | $\mathbf{q_4}$ |

Second stage of TSLS: Regress $y$ on $\hat{x}$ and $v$ in mapping and unlabeled samples to get $\theta$.

### A GMM Estimator

OLS with training and mapping samples

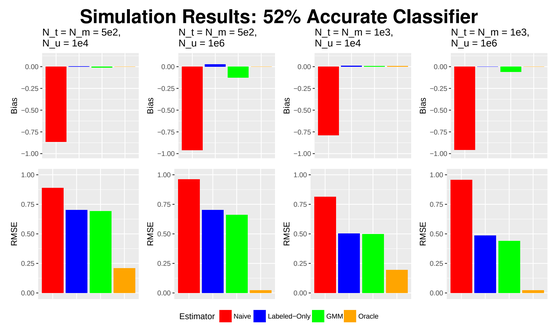
$$g_N(\theta, \hat{\psi}) = \begin{pmatrix} \sum_{i=1}^{N} (m_i + t_i)\begin{pmatrix} x_i \\ v_i \end{pmatrix}(y_i - \begin{pmatrix} x_i \\ v_i \end{pmatrix}'\theta) \\ \sum_{i=1}^{N} (m_i + u_i)\begin{pmatrix} \hat{x}_i \\ v_i \end{pmatrix}(y_i - \begin{pmatrix} \hat{x}_i \\ v_i \end{pmatrix}'\theta) \end{pmatrix}$$

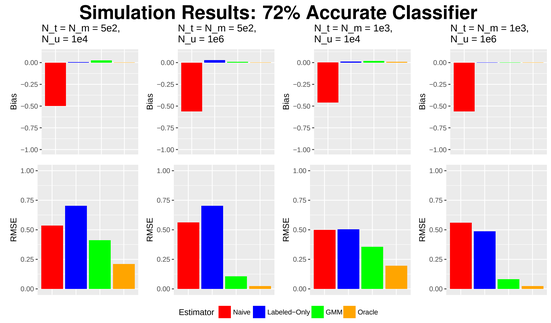
TSLS with mapping and unlabeled samples

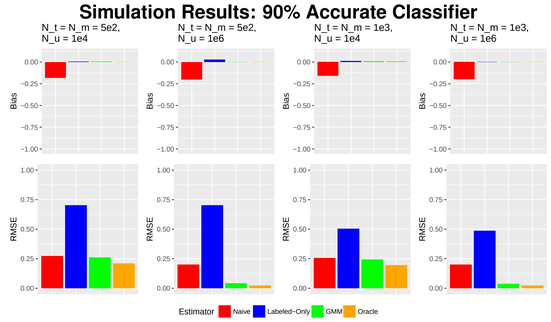
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, g_N(\theta, \hat{\psi})' W_N g_N(\theta, \hat{\psi})$$

## 3. Simulations

$$y_i = x_i + N(0, 0.8) + \text{Bernoulli}(0.15)|N(0, 20)|$$

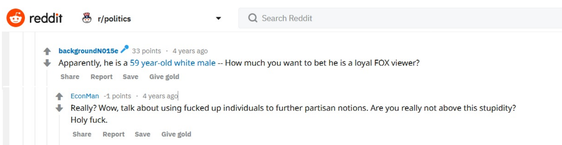



## 4. Application: Uncivil Political Dialog

- Munger (2018): Calling out uncivil tweeters decreases uncivil tweeting
- Do uncivil posts get called out in nature?
- 1,622,218 observations, 3,118 hand-labeled
- Split hand-labeled into 2,493 training and 625 mapping




### Effect of Incivility on Post Score