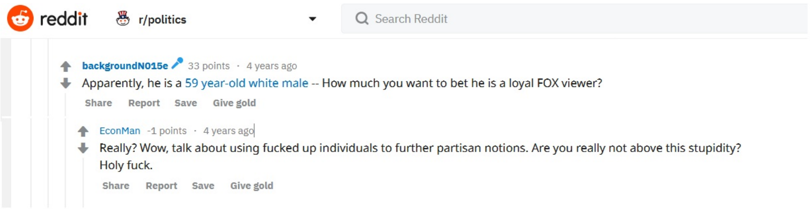
Machine Learning Predictions as Regression Covariates

Matt Tyler Christian Fong

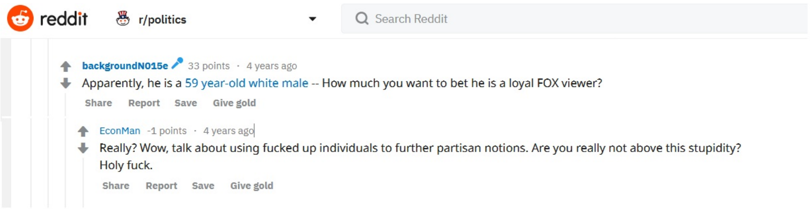
March 1, 2019

Studying Uncivil Political Dialogue



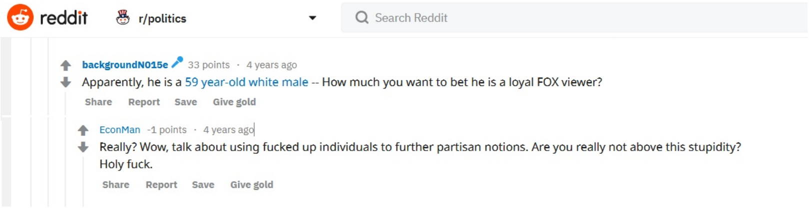
- **Task:** estimate β_0 from $y = x'\beta_0 + \epsilon$
  - y = score $\approx$ upvotes $-$ downvotes
  - x = (incivility, 1)

Studying Uncivil Political Dialogue



- **Task:** estimate β_0 from $y = x'\beta_0 + \epsilon$
  - y = score $\approx$ upvotes $-$ downvotes
  - x = (incivility, 1)
- **Challenge:** x_1 must be hand-coded, but there are millions of comments

Studying Uncivil Political Dialogue



- **Task:** estimate β_0 from $y = x'\beta_0 + \epsilon$
  - y = score $\approx$ upvotes $-$ downvotes
  - x = (incivility, 1)
- **Challenge:** x_1 must be hand-coded, but there are millions of comments
- **Solution?** hand-code a subset of comments,then predict x_1 with z_1 (using algorithm) for the remaining millions

This is a Common Challenge

This is a Common Challenge

- Statements by Russian politicians:

  Stewart & Zhukov (2009)

- Chinese social media:

  King, Pan, & Roberts (2013)

- Congressional press releases:

  Grimmer, Westwood, & Messing (2014)

- Arabic tweets:

  Jamal et al. (2015)

This is a Common Challenge

- Statements by Russian politicians:

  Stewart & Zhukov (2009)

- Chinese social media:

  King, Pan, & Roberts (2013)

- Congressional press releases:

  Grimmer, Westwood, & Messing (2014)

- Arabic tweets:

  Jamal et al. (2015)

- European tweets:

  Theocharis et al. (2016)

- Satellite imagery:

  Jean et al. (2016)

- Congressional pictures with constituents:

  Anastosopoulos et al. (2017)

- Supreme Court audio recordings:

  Knox and Lucas (2018)

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

- **Need $\mathrm{E}(z\tilde{\epsilon}) = 0$ for consistency!**

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

- **Need $\mathrm{E}(z\tilde{\epsilon}) = 0$ for consistency!**
  - Exclusion Restriction: $\mathrm{E}(z_1\epsilon) = 0$

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

- **Need $\mathrm{E}(z\tilde{\epsilon}) = 0$ for consistency!**
  - Exclusion Restriction: $\mathrm{E}(z_1\epsilon) = 0$
  - Classical Error: $\mathrm{E}(z(x_1 - z_1)') = 0$

    $\iff$ regression of error $x_1 - z_1$ on z_1, x_2 is zero

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

- **Need $\mathrm{E}(z\tilde{\epsilon}) = 0$ for consistency!**
  - Exclusion Restriction: $\mathrm{E}(z_1\epsilon) = 0$
  - Classical Error: $\mathrm{E}(z(x_1 - z_1)') = 0$

    $\iff$ regression of error $x_1 - z_1$ on z_1, x_2 is zero
  - If both somehow hold, Naive Estimator is consistent

Problems with Algorithmic Error

- Model: $y = x'\beta_0 + \epsilon$ where $\mathrm{E}(x\epsilon) = 0$
- True covariates: $x = (x_1, x_2)$, with x_2 always observed
- Predicted covariates $z = (z_1, z_2)$, with $z_2 = x_2$
- **Naive Estimator:** plug z in for x:

$$y = z'\beta_0 + \tilde{\epsilon}, \quad \tilde{\epsilon} = (x - z)'\beta_0 + \epsilon$$

- **Need $\mathrm{E}(z\tilde{\epsilon}) = 0$ for consistency!**
  - Exclusion Restriction: $\mathrm{E}(z_1\epsilon) = 0$
  - Classical Error: $\mathrm{E}(z(x_1 - z_1)') = 0$
    $\iff$ regression of error $x_1 - z_1$ on z_1, x_2 is zero
  - If both somehow hold, Naive Estimator is consistent
  - If just ER and $\mathrm{E}(x_2(x_1 - z_1)') = 0$, then Naive Estimator is substantially attenuated

Example

- score $= a + b \times \text{Uncivil} + \epsilon$

Example

- score $= a + b \times \text{Uncivil} + \epsilon$
- Suppose we classify a post as civil/uncivil with 80% accuracy, $P(x = z) = 0.8$

Example

- score $= a + b \times \text{Uncivil} + \epsilon$
- Suppose we classify a post as civil/uncivil with 80% accuracy, $P(x = z) = 0.8$
- With ER, the expected value of the Naive Estimator of b is $0.6b$

Example

- score = $a + b \times \text{Uncivil} + \epsilon$
- Suppose we classify a post as civil/uncivil with 80% accuracy, $P(x = z) = 0.8$
- With ER, the expected value of the Naive Estimator of b is $0.6b$
- If our accuracy is only 70%, then the expected value of the Naive Estiamtor of b is $0.4b$

Example

- score $= a + b \times \text{Uncivil} + \epsilon$
- Suppose we classify a post as civil/uncivil with 80% accuracy, $P(x = z) = 0.8$
- With ER, the expected value of the Naive Estimator of b is $0.6b$
- If our accuracy is only 70%, then the expected value of the Naive Estiamtor of b is $0.4b$
- These are substantial biases that limit hypothesis tests of $b = 0$

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
    - Classical Error is unlikely to hold in practice

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
    - Classical Error is unlikely to hold in practice
    - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**
- **With our method**: only need Exclusion Restriction

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**
- **With our method**: only need Exclusion Restriction
  - **No functional form assumptions**

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**
- **With our method**: only need Exclusion Restriction
  - **No functional form assumptions**
  - **Consistent with correct standard errors**

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**
- **With our method**: only need Exclusion Restriction
  - **No functional form assumptions**
  - **Consistent with correct standard errors**
  - Accounts for overfitting z to x in training data

Naive Estimator vs. Our Approach

- **Naive Estimator requires both Classical Error and Exclusion Restriction assumptions**
  - Classical Error is unlikely to hold in practice
  - Classical Error is impossible if covariates and predictions are categorical (Aigner 1973)
  - **Attenuation bias makes it hard to find effects**
  - **Inconsistency $\implies$ meaningless confidence intervals**
- **With our method**: only need Exclusion Restriction
  - **No functional form assumptions**
  - **Consistent with correct standard errors**
  - Accounts for overfitting z to x in training data
  - (Asymptotically) more efficient than just using data where x_1 is coded

Our Method in Three Steps

1. Divide the Coded Data into Training and Validation Subsets

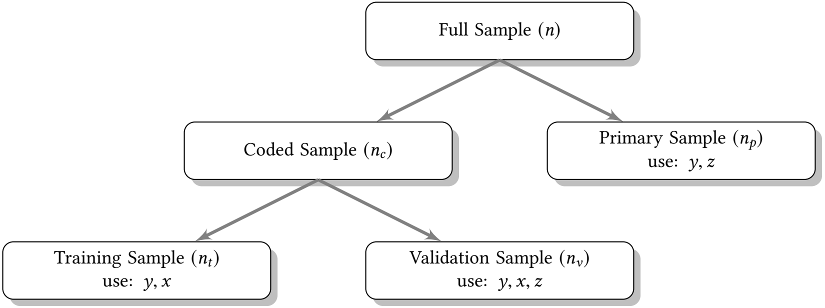
Our Method in Three Steps

1. Divide the Coded Data into Training and Validation Subsets
2. First-Stage Regression using Validation Subset

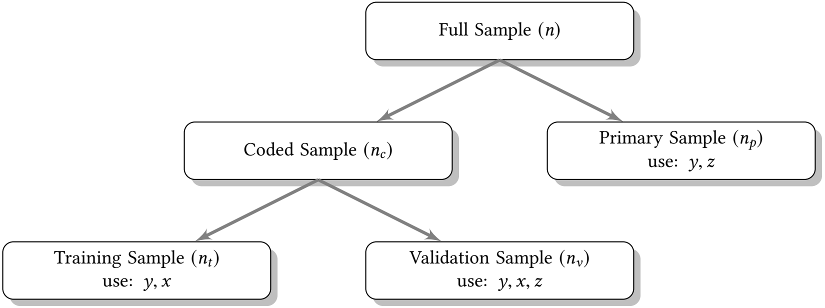
Our Method in Three Steps

1. Divide the Coded Data into Training and Validation Subsets
2. First-Stage Regression using Validation Subset
3. Combine Coded and Uncoded Data using Generalized Method of Moments (GMM)

Step 1: Divide the Coded Data into Training and Validation Subsets



Step 1: Divide the Coded Data into Training and Validation Subsets



$\rightsquigarrow$ Training sample to fit z-algorithm, but then hide its z values from Steps 2 & 3 to avoid overfitting bias

Step 2: First-Stage Regression

- We can always find $\Lambda_0 = \begin{bmatrix} \lambda_{10} & \lambda_{20} & \cdots \end{bmatrix}$ so that

$$\begin{aligned} x_1 &= z'\lambda_{10} + \eta_1 \\ x_2 &= z'\lambda_{20} + \eta_2 \\ \vdots &\quad \vdots \end{aligned}$$

Step 2: First-Stage Regression

- We can always find $\Lambda_0 = \begin{bmatrix} \lambda_{10} & \lambda_{20} & \cdots \end{bmatrix}$ so that

$$\begin{aligned} x_1 &= z'\lambda_{10} + \eta_1 \\ x_2 &= z'\lambda_{20} + \eta_2 \\ \vdots &\quad \vdots \end{aligned}$$

- e.g., if $x = (\text{Uncivil}, 1)$, and $z = (z_1, 1)$, then

$$\begin{aligned} \text{Uncivil} &= (z_1, 1)\lambda_{10} + \eta_1 \\ 1 &= (z_1, 1)\lambda_{20} + \eta_2 \end{aligned}$$

Step 2: First-Stage Regression

- We can always find $\Lambda_0 = \begin{bmatrix} \lambda_{10} & \lambda_{20} & \cdots \end{bmatrix}$ so that

$$\begin{aligned} x_1 &= z'\lambda_{10} + \eta_1 \\ x_2 &= z'\lambda_{20} + \eta_2 \\ \vdots & \quad \vdots \end{aligned}$$

- e.g., if $x = (\text{Uncivil}, 1)$, and $z = (z_1, 1)$, then

$$\begin{aligned} \text{Uncivil} &= (z_1, 1)\lambda_{10} + \eta_1 \\ 1 &= (z_1, 1)\lambda_{20} + \eta_2 \end{aligned}$$

- Compactly write $x' = z'\Lambda_0 + \eta'$ with $\text{E}(z\eta') = 0$

Step 2: First-Stage Regression on Validation Sample

- Given Exclusion Restriction, plug in $z'\Lambda_0$ for x':

$$y = z'\Lambda_0\beta_0 + \bar{\epsilon}, \quad \mathrm{E}(\Lambda_0' z\bar{\epsilon}) = 0$$

$\implies$ we can use 2SLS on the uncoded data to estimate β_0

Step 2: First-Stage Regression on Validation Sample

- Given Exclusion Restriction, plug in $z'\Lambda_0$ for x':

$$y = z'\Lambda_0\beta_0 + \bar{\epsilon}, \quad \mathrm{E}(\Lambda_0' z\bar{\epsilon}) = 0$$

$\implies$ we can use 2SLS on the uncoded data to estimate β_0

- Estimate first-stage Λ_0 on the validation sample

Step 2: First-Stage Regression on Validation Sample

- Given Exclusion Restriction, plug in $z'\Lambda_0$ for x':

$$y = z'\Lambda_0\beta_0 + \bar{\epsilon}, \quad \mathrm{E}(\Lambda_0' z\bar{\epsilon}) = 0$$

$\implies$ we can use 2SLS on the uncoded data to estimate β_0

- Estimate first-stage Λ_0 on the validation sample
- e.g.,

$$\hat{\lambda}_k = \left(\sum_{i=1}^{n} v_i z_i z_i'\right)^{-1} \left(\sum_{i=1}^{n} v_i z_i x_{ik}\right) \tag{1}$$

Step 3: Combining Coded and Uncoded Data

- Now we have two ways to estimate β_0

Step 3: Combining Coded and Uncoded Data

- Now we have two ways to estimate β_0
- OLS on the training & validation data ("labeled-only estimator"):

$$\hat{\beta}_1 \text{ such that } g_1(\hat{\beta}_1) = 0$$

$$\text{where } g_1(\beta) = \frac{1}{n_v + n_t} \sum_{i=1}^{n} (v_i + t_i) x_i (y_i - x_i' \beta)$$

Step 3: Combining Coded and Uncoded Data

- Now we have two ways to estimate β_0
- OLS on the training & validation data ("labeled-only estimator"):

$$\hat{\beta}_1 \text{ such that } g_1(\hat{\beta}_1) = 0$$

$$\text{where } g_1(\beta) = \frac{1}{n_v + n_t} \sum_{i=1}^{n} (v_i + t_i) x_i (y_i - x_i' \beta)$$

- 2SLS on the primary & validation samples:

$$\hat{\beta}_2 \text{ such that } g_2(\hat{\beta}_2) = 0$$

$$\text{where } g_2(\beta) = \frac{1}{n_v + n_p} \sum_{i=1}^{n} (v_i + p_i) \hat{\Lambda}' z_i (y_i - z_i' \hat{\Lambda} \beta)$$

Step 3: Combining Coded and Uncoded Data with GMM

- Efficiency gains by combining g_1 and g_2

Step 3: Combining Coded and Uncoded Data with GMM

- Efficiency gains by combining g_1 and g_2
- Generalized Method of Moments (GMM):

$$\hat{\beta}_{\mathrm{GMM}} = \underset{\beta}{\operatorname{argmin}}\ g(\beta)' W g(\beta), \quad g(\beta) = \begin{bmatrix} g_1(\beta) \\ g_2(\beta) \end{bmatrix}$$

Step 3: Combining Coded and Uncoded Data with GMM

- Efficiency gains by combining g_1 and g_2
- Generalized Method of Moments (GMM):

$$\hat{\beta}_{\text{GMM}} = \underset{\beta}{\operatorname{argmin}}\ g(\beta)' W g(\beta), \quad g(\beta) = \begin{bmatrix} g_1(\beta) \\ g_2(\beta) \end{bmatrix}$$

- Consistent and asymptotically normal for any (reasonable) weighting matrix W

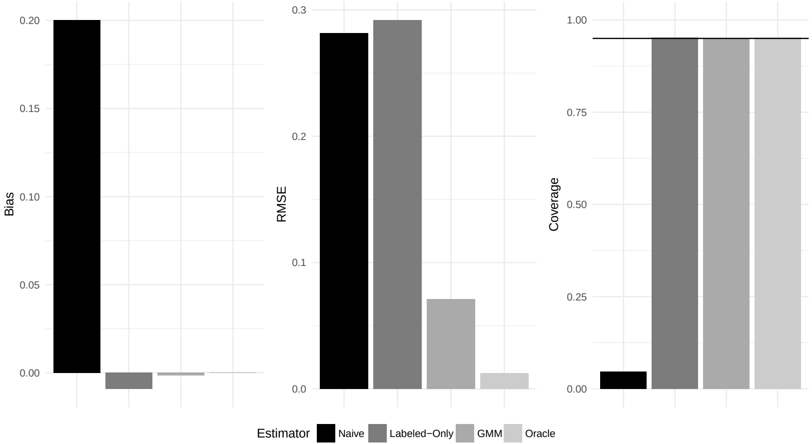
Step 3: Combining Coded and Uncoded Data with GMM

- Efficiency gains by combining g_1 and g_2
- Generalized Method of Moments (GMM):

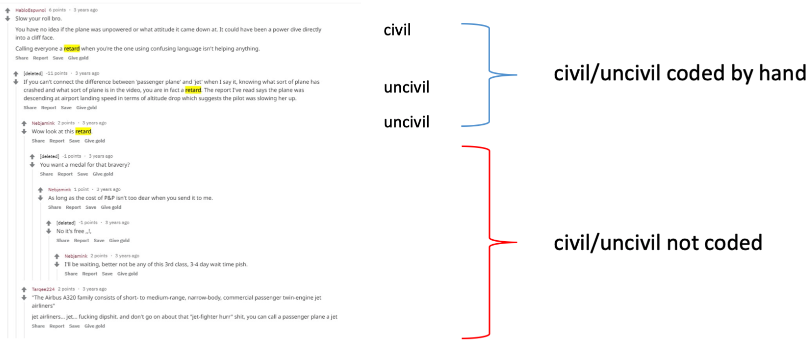
$$\hat{\beta}_{\text{GMM}} = \underset{\beta}{\operatorname{argmin}} \; g(\beta)' W g(\beta), \quad g(\beta) = \begin{bmatrix} g_1(\beta) \\ g_2(\beta) \end{bmatrix}$$

- Consistent and asymptotically normal for any (reasonable) weighting matrix W
- We derive W that makes $\hat{\beta}_{\text{GMM}}$ more efficient than $\hat{\beta}_1, \hat{\beta}_2$

Data-Based Simulation: GMM is Equal to Coding 16 Times More Posts

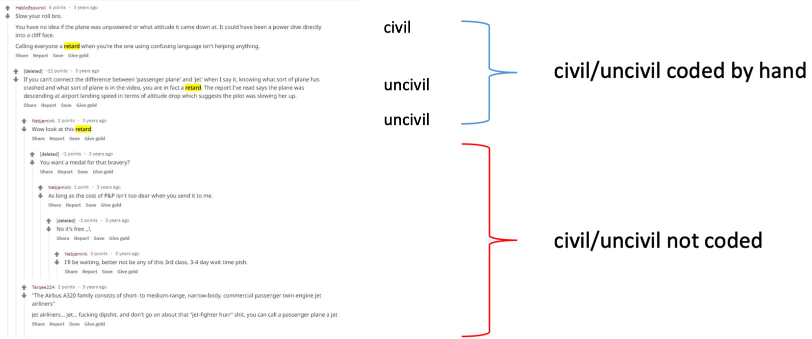


Applying the GMM Estimator



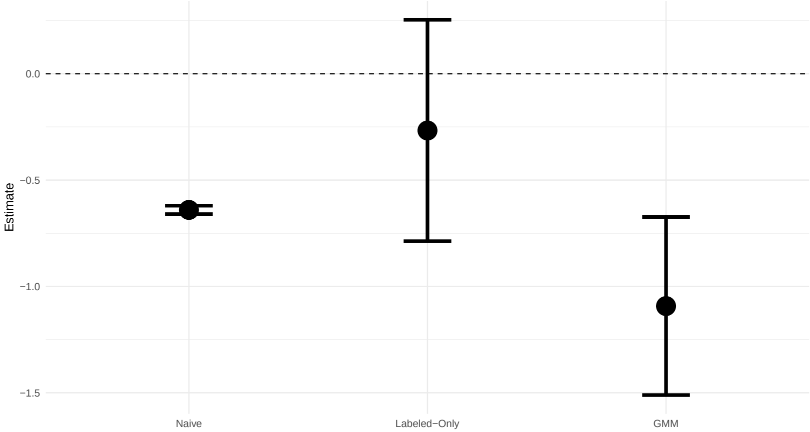
- n =1,622,118 split into $n_v = 625$; $n_t = 2{,}493$; & $n_p =$ 1,619,000

Applying the GMM Estimator



- n =1,622,118 split into $n_v = 625$; $n_t = 2{,}493$; & $n_p =$ 1,619,000
- 5,000 most common words + LASSO $\implies$ z

GMM Corrects for Naive Estimator's Attenuation Bias



GMM: uncivil posts on average have ≈ 1.1 (0.2) lower score

What About the Exclusion Restriction?

- Relative to the Naive Estimator, the GMM Estimator bypasses most measurement error assumptions

What About the Exclusion Restriction?

- Relative to the Naive Estimator, the GMM Estimator bypasses most measurement error assumptions
- Remaining assumption, the Exclusion Restriction (ER), can be tested in the validation sample (see paper) as $n_v \to \infty$

What About the Exclusion Restriction?

- Relative to the Naive Estimator, the GMM Estimator bypasses most measurement error assumptions
- Remaining assumption, the Exclusion Restriction (ER), can be tested in the validation sample (see paper) as $n_v \to \infty$
- Unlike in IV, z is fine-tuned to predict x: should expect only minor violations of ER

What About the Exclusion Restriction?

- Relative to the Naive Estimator, the GMM Estimator bypasses most measurement error assumptions
- Remaining assumption, the Exclusion Restriction (ER), can be tested in the validation sample (see paper) as $n_v \to \infty$
- Unlike in IV, z is fine-tuned to predict x: should expect only minor violations of ER
- Bias from violating ER decays as error $x - z$ gets smaller (appendix)

What About the Exclusion Restriction?

- Relative to the Naive Estimator, the GMM Estimator bypasses most measurement error assumptions
- Remaining assumption, the Exclusion Restriction (ER), can be tested in the validation sample (see paper) as $n_v \to \infty$
- Unlike in IV, z is fine-tuned to predict x: should expect only minor violations of ER
- Bias from violating ER decays as error $x - z$ gets smaller (appendix)
- Nonparametric methods that do not lean on ER (Chen, Hong, & Tamer 2005) are not RMSE-competitive in realistic scenarios (appendix)

Summary

- Highlighted an algorithmic theme in recent empirical work across subdisciplines of political science

Summary

- Highlighted an algorithmic theme in recent empirical work across subdisciplines of political science
- Many practitioners are plugging-in algorithmic predictions or using ad-hoc adjustments

Summary

- Highlighted an algorithmic theme in recent empirical work across subdisciplines of political science
- Many practitioners are plugging-in algorithmic predictions or using ad-hoc adjustments
- We provide a robust solution that does not depend on classical measurement error assumptions

Summary

- Highlighted an algorithmic theme in recent empirical work across subdisciplines of political science
- Many practitioners are plugging-in algorithmic predictions or using ad-hoc adjustments
- We provide a robust solution that does not depend on classical measurement error assumptions
- No functional form or homoskedasticity assumptions

Summary

- Highlighted an algorithmic theme in recent empirical work across subdisciplines of political science
- Many practitioners are plugging-in algorithmic predictions or using ad-hoc adjustments
- We provide a robust solution that does not depend on classical measurement error assumptions
- No functional form or homoskedasticity assumptions
- More efficient than just using hand-coded data!