Inference for Regression with Variables Generated from Unstructured Data

Laura Battaglia (Oxford) Tim Christensen (Yale) Stephen Hansen (UCL) Szymon Sacher (Meta)

October 23, 2024

Outline

Motivation

Use of unstructured data (text, images, audio files, etc.) in applied work is growing rapidly

Almost all papers use a two-step strategy:

1. Estimate latent observations (θ_i) from unstructured data using an information retrieval model
2. Plug estimates ($\hat{\theta}_i$) into an econometric model, treating $\hat{\theta}_i$ as regular numeric data

Motivation

Use of unstructured data (text, images, audio files, etc.) in applied work is growing rapidly

Almost all papers use a two-step strategy:

1. Estimate latent observations (θ_i) from unstructured data using an information retrieval model
2. Plug estimates ($\hat{\theta}_i$) into an econometric model, treating $\hat{\theta}_i$ as regular numeric data

Pragmatic approach. But little is known about its statistical properties

- measurement error?
- generated regressors?
- analogy with FAR/FAVARs?

Examples

Supervised Learning (Impute a Missing Label)

- Baker Bloom Davis (2016): economic policy uncertainty measured from newspaper text
- Gorodnichenko Pham Talavera (2023): tone-of-voice measured from FOMC press conferences
- Adukia et. al. (2023): race and gender of children book characters

Unsupervised Learning (Learn Latent Representation)

- Hoberg Phillips (2016): latent industry type measured from corporate filings
- Hansen McMahon Prat (2018): policy deliberation measured from FOMC transcripts
- Magnolfi McClure Sorensen (2022): product differentiation measured from survey data
- Compiani Morozov Seiler (2023): substitutability measured from Amazon text + image data
- Gabaix Koijen Yogo (2023): firm characteristics measured from investor holdings

This Paper

1. Two-step strategy leads to invalid inference: CIs have right width but wrong centering (bias)

   Bias depends on relative importance of

   (a) Measurement error in upstream model

   (b) Sampling error in downstream model

   Valid inference requires (a) to vanish much faster than (b)

   Empirical interpretation: amount of unstructured data per observation swamps the sample size

   $\longrightarrow$ not the typical case in leading empirical applications

This Paper

1. Two-step strategy leads to invalid inference: CIs have right width but wrong centering (bias)

   Bias depends on relative importance of

   (a) Measurement error in upstream model

   (b) Sampling error in downstream model

   Valid inference requires (a) to vanish much faster than (b)

   Empirical interpretation: amount of unstructured data per observation swamps the sample size

   $\longrightarrow$ not the typical case in leading empirical applications

2. Solutions:

   (a) Bias correction

   (b) One-step estimation using likelihood of upstream and downstream components.

   (c) IV estimation?

This Paper

1. Two-step strategy leads to invalid inference: CIs have right width but wrong centering (bias)

   Bias depends on relative importance of

   (a) Measurement error in upstream model

   (b) Sampling error in downstream model

   Valid inference requires (a) to vanish much faster than (b)

   Empirical interpretation: amount of unstructured data per observation swamps the sample size

   $\longrightarrow$ not the typical case in leading empirical applications

2. Solutions:

   (a) Bias correction

   (b) One-step estimation using likelihood of upstream and downstream components.

   (c) IV estimation?

3. Shows empirical relevance in several applications (today: CEO time use).

Outline

Sentiment Regression

Suppose we wish to perform inference on γ_1 in the regression model

$$Y_i = \gamma_0 + \gamma_1 \theta_i + \varepsilon_i\,, \qquad \mathbb{E}[\varepsilon_i | \theta_i] = 0\,,$$

θ_i: latent 'sentiment' in month i.

We observe (X_i, C_i) where

$$X_i \sim \text{Binomial}(C_i, \theta_i)$$

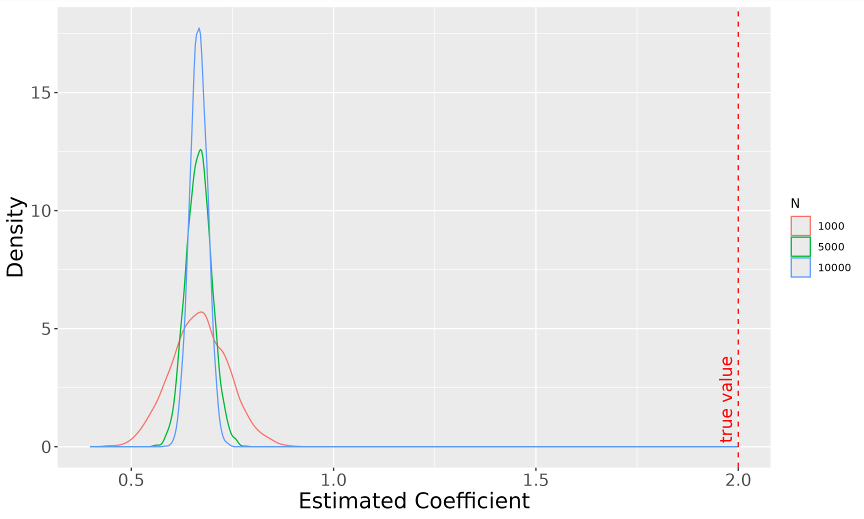
Two-step strategy:

1. estimate θ_i with $\hat{\theta}_i = X_i / C_i \implies$ measurement error depends on C_i.
2. regress Y_i on $\hat{\theta}_i$. Perform standard OLS inference (treating $\hat{\theta}_i$ as data)

Case I: Large Sample Size

- Suppose we observe IID sample $(X_i, Y_i, C_i)_{i=1}^n$
- Take $n \to \infty$ so that sampling error in downstream model vanishes.
- Non-vanishing measurement error in $\hat{\theta}_i$ leads to inconsistent OLS estimates.

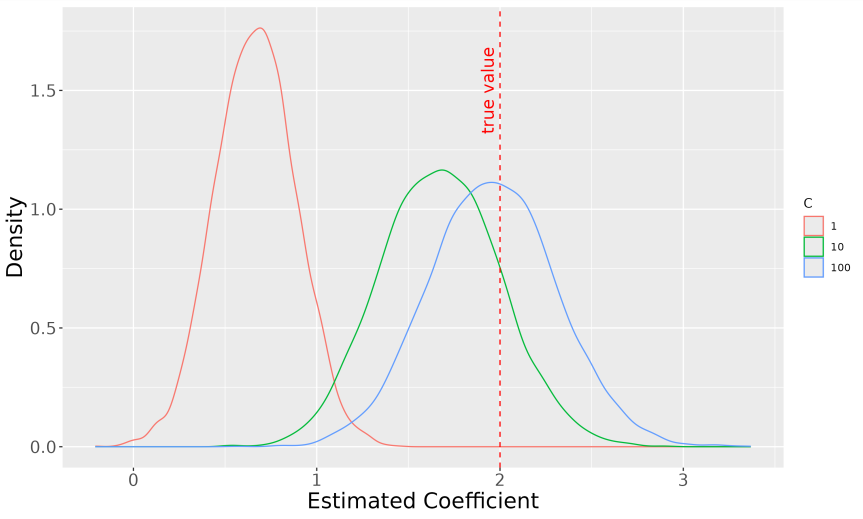
Effect of Increasing n with Fixed $C_i = C = 1$



Case II: Large Amount of Unstructured Data

- Suppose we again observe IID sample $(X_i, Y_i, C_i)_{i=1}^n$
- Now take $C_i \to \infty$ for each observation i.
- Implies that $\hat{\theta}_i \to \theta_i$ so that measurement error vanishes.
- OLS is unbiased in finite samples.

Effect of Increasing $C_i = C$ with Fixed $n = 100$



Case III: Both Forces Present

- In modern datasets, we have both large n and low measurement error.
- Challenge: how to develop asymptotic framework that reflects this?
- We consider sequential DGP where:
  1. Distribution of (Y_i, θ_i) is fixed with n.
  2. Conditional distribution of $\hat{\theta}_i$ given (Y_i, θ_i) varies with n.
  3. Along this sequence,
  $$\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa \in [0, \infty)$$
- κ measures importance of measurement error relative to sampling error.
- In spirit of small noise asymptotics of Chesher (1991), Evdokimov and Zeleneev (2024).

Implications for Inference

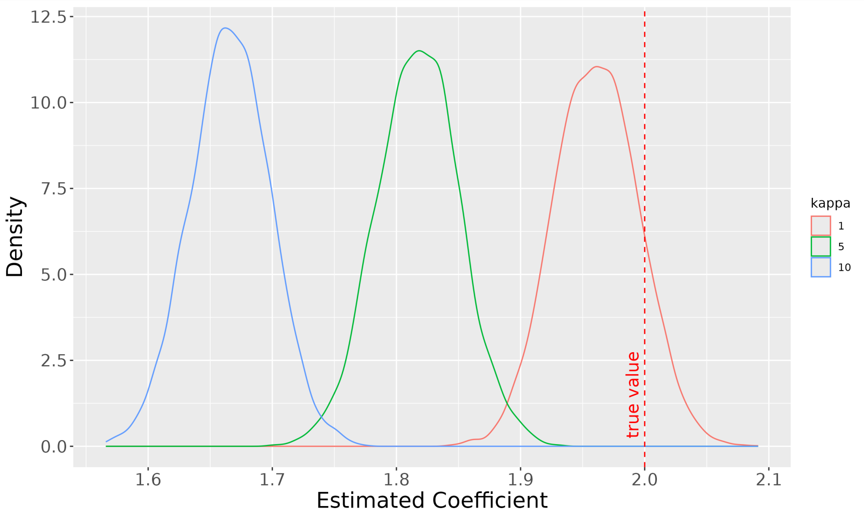
Proposition

Along this sequence of DGPs, we have

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) \to_d N\left(-\kappa\,\gamma_1 \frac{\mathbb{E}[\theta_i(1-\theta_i)]}{\mathrm{Var}(\theta_i)}, \frac{\mathbb{E}[\varepsilon_i^2(\theta_i - \mathbb{E}[\theta_i])^2]}{\mathrm{Var}(\theta_i)^2}\right)$$

- $\kappa = 0$: two-step inference is valid because sampling error dominates measurement error
- $\kappa \in (0, \infty)$: two-step inference is biased (CIs under-cover), bias proportional to κ

$C_i = C \in \{10, 20, 100\}$ and $n = 10,000$



Standard Deviations	0.032	0.033	0.034

Outline

General Model

We consider the linear regression model

$$Y_i = \boldsymbol{\gamma}^T \boldsymbol{\theta}_i + \boldsymbol{\alpha}^T \mathbf{q}_i + \varepsilon_i, \qquad \mathbb{E}\left[\varepsilon_i | \boldsymbol{\theta}_i, \mathbf{q}_i\right] = 0 \tag{1}$$

- $\boldsymbol{\theta}_i$ is a vector of latent variables of economic interest
- $\mathbf{q}_i$ are standard numeric covariates
- We focus on inference for $\boldsymbol{\gamma}$
- $\boldsymbol{\alpha}$ of interest in other applications (Avivi 2024)

Unstructured dataset available for estimating $\boldsymbol{\theta}_i$.

Two-Step Strategy

(i) Estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ obtained from unstructured data using an upstream information retrieval model.

(ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Inference is performed treating $\hat{\boldsymbol{\theta}}_i$ as regular numeric data.

Two-Step Strategy

(i) Estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ obtained from unstructured data using an upstream information retrieval model.

(ii) Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Inference is performed treating $\hat{\boldsymbol{\theta}}_i$ as regular numeric data.

$$\boldsymbol{\xi}_i = \begin{bmatrix} \boldsymbol{\theta}_i \\ \mathbf{q}_i \end{bmatrix}, \qquad \hat{\boldsymbol{\xi}}_i = \begin{bmatrix} \hat{\boldsymbol{\theta}}_i \\ \mathbf{q}_i \end{bmatrix}.$$

The OLS estimator of $\boldsymbol{\psi} = [\gamma, \boldsymbol{\alpha}]^T$ in the two-step strategy is given by

$$\hat{\boldsymbol{\psi}} = \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i Y_i \right). \tag{2}$$

Example 1: ML Generated Labels

- ML algorithms often deployed to impute missing observations from unstructured data, for example when labelling the full data set is prohibitively costly or otherwise infeasible
- Leading use case: missing θ_i is binary (e.g., race indicator)
- Generate estimate $\hat{\theta}_i$ using unstructured data $\mathbf{x}_i$ (e.g., photograph)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$

Example 1: ML Generated Labels

- ML algorithms often deployed to impute missing observations from unstructured data, for example when labelling the full data set is prohibitively costly or otherwise infeasible
- Leading use case: missing θ_i is binary (e.g., race indicator)
- Generate estimate $\hat{\theta}_i$ using unstructured data $\mathbf{x}_i$ (e.g., photograph)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$

Example 1: ML Generated Labels

- ML algorithms often deployed to impute missing observations from unstructured data, for example when labelling the full data set is prohibitively costly or otherwise infeasible
- Leading use case: missing θ_i is binary (e.g., race indicator)
- Generate estimate $\hat{\theta}_i$ using unstructured data $\mathbf{x}_i$ (e.g., photograph)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$
- Measurement error arises due to **classification error**.
- Let $p_i = \Pr\left[\theta_i = 1 \mid \mathbf{x}_i, \mathbf{q}_i\right]$ and $\pi_i = \Pr\left[\hat{\theta}_i = 1 \,\middle|\, \mathbf{x}_i, \mathbf{q}_i\right]$.
- False positive rate is $(1 - p_i)\pi_i$; false negative rate is $p_i(1 - \pi_i)$.

Example 1: ML Generated Labels

- ML algorithms often deployed to impute missing observations from unstructured data, for example when labelling the full data set is prohibitively costly or otherwise infeasible
- Leading use case: missing θ_i is binary (e.g., race indicator)
- Generate estimate $\hat{\theta}_i$ using unstructured data $\mathbf{x}_i$ (e.g., photograph)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$
- Measurement error arises due to **classification error**.
- Let $p_i = \Pr\left[\theta_i = 1 \mid \mathbf{x}_i, \mathbf{q}_i\right]$ and $\pi_i = \Pr\left[\hat{\theta}_i = 1 \,\middle|\, \mathbf{x}_i, \mathbf{q}_i\right]$.
- False positive rate is $(1 - p_i)\pi_i$; false negative rate is $p_i(1 - \pi_i)$.
- Used as part of the two-step strategy by: Baker Bloom Davis (2016); Imai Khanna (2016); ... ; Bybee (2024); Boxell Conway (2024).

Example 2: Topic Models

- Unstructured obs i is a V-dim vector of feature counts $\mathbf{x}_i$
- Factor structure on multinomial probabilities (as in probabilistic latent semantic analysis/LDA):

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  - $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$, each $\boldsymbol{\beta}_k \in \Delta^{V-1}$ is a topic
  - observation-specific topic weights $\boldsymbol{\theta}_i \in \Delta^{K-1}$

Example 2: Topic Models

- Unstructured obs i is a V-dim vector of feature counts $\mathbf{x}_i$
- Factor structure on multinomial probabilities (as in probabilistic latent semantic analysis/LDA):

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  - $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$, each $\boldsymbol{\beta}_k \in \Delta^{V-1}$ is a topic
  - observation-specific topic weights $\boldsymbol{\theta}_i \in \Delta^{K-1}$
- Estimate latent $\boldsymbol{\theta}_i$ from $\mathbf{x}_i$ e.g. via LDA $\rightarrow \hat{\boldsymbol{\theta}}_i$.
- Measurement error from **sampling error**.

Example 2: Topic Models

- Unstructured obs i is a V-dim vector of feature counts $\mathbf{x}_i$
- Factor structure on multinomial probabilities (as in probabilistic latent semantic analysis/LDA):

$$\mathbf{x}_i | (C_i, \boldsymbol{\theta}_i) \sim \text{Multinomial}(C_i, \mathbf{B}^T \boldsymbol{\theta}_i)$$

  - $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$, each $\boldsymbol{\beta}_k \in \Delta^{V-1}$ is a topic
  - observation-specific topic weights $\boldsymbol{\theta}_i \in \Delta^{K-1}$
- Estimate latent $\boldsymbol{\theta}_i$ from $\mathbf{x}_i$ e.g. via LDA $\rightarrow \hat{\boldsymbol{\theta}}_i$.
- Measurement error from **sampling error**.
- Used as part of two-step strategy by:
  - Text data: Hansen McMahon Prat (2018); Mueller Rauh (2018); Larsen and Thorsrud (2019); Thorsrud (2020); Bybee Kelly Manela Xiu (2020); Ash Morelli Vannoni (2022)
  - Survey data: Bandiera Prat Hansen Sadun (2020); Draca Schwarz (2020)
  - Network data: Nimczik (2017)

Example 3: Index Built from Classified Labels

- Suppose that each observation i has C_i observed labels, e.g. the number of classified newspaper articles observed in month i.

- Let $p_i = \Pr\left[\theta_{i,j} = 1 \mid \mathbf{q}_i\right]$, e.g. all articles have independent probability of discussing policy uncertainty given economic conditions.

- Suppose the realization of the observed label $\hat{\theta}_{i,j}$ depends only on true label $\theta_{i,j}$:
  $\longrightarrow \pi_1 = \Pr\left[\hat{\theta}_{i,j} = 1 \,\middle|\, \theta_{i,j} = 1\right]$ and $\pi_0 = \Pr\left[\hat{\theta}_{i,j} = 0 \,\middle|\, \theta_{i,j} = 0\right]$.

- Then the distribution of $x_{i,1} \equiv \sum_{j=1}^{C_i} 1\left(\hat{\theta}_{i,j} = 1\right)$ is

$$x_{i,1} \sim \text{Binomial}(C_i, p_i\pi_1 + (1 - p_i)\pi_0)$$

Topic model with $K = 2$ and topic-feature distributions $(\pi_1, 1 - \pi_1)$ and $(\pi_0, 1 - \pi_0)$.

Asymptotics: General Case

Consider a sequence of DGPs for $(Y_i, \boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i, \mathbf{q}_i, \mathbf{x}_i)_{i=1}^n$ indexed by sample size n, in which

$$\sqrt{n}\left[\frac{1}{N}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T\right] \to_p \kappa\,\boldsymbol{\Omega},$$

where $\kappa \geq 0$ measures the importance of measurement error relative to (downstream) sampling error

Asymptotics: General Case

Consider a sequence of DGPs for $(Y_i, \boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i, \mathbf{q}_i, \mathbf{x}_i)_{i=1}^n$ indexed by sample size n, in which

$$\sqrt{n}\left[\frac{1}{N}\sum_{i=1}^n \hat{\boldsymbol{\theta}}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)^T\right] \to_p \kappa\,\boldsymbol{\Omega},$$

where $\kappa \geq 0$ measures the importance of measurement error relative to (downstream) sampling error

Theorem: Two-Step Inference is Invalid Unless $\kappa = 0$

1. OLS estimator $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\alpha}})$ of $\boldsymbol{\psi} = (\boldsymbol{\gamma}, \boldsymbol{\alpha})$ from regressing Y_i on $\hat{\boldsymbol{\xi}}_i = (\hat{\boldsymbol{\theta}}_i, \mathbf{q}_i)$ has asy dist

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \to_d N\Big(\kappa \times \textbf{bias}(\boldsymbol{\Omega}, \boldsymbol{\gamma}, \mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T])\,,\, \underbrace{\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T]^{-1}\mathbb{E}[\varepsilon_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T]\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T]^{-1}}_{=:V}\Big)$$

2. Eicker–Huber–White standard errors are consistent for all $\kappa \geq 0$:

$$\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)\left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1} \to_p V$$

Asymptotics: Examples

- ML-generated binary labels:

$$\sqrt{n} \times \mathbb{E}[FP_i] \to \kappa, \qquad \textbf{bias} = -\mathbb{E}\left[\xi_i \xi_i^T\right]^{-1} \begin{bmatrix} \gamma \\ \mathbf{0} \end{bmatrix}$$

- Topic models:

$$\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa, \qquad \textbf{bias} = \text{(complicated)}$$

- Further applications: ML-generated indices; similarity measures; VARs; ...

Implications

- $\kappa \in (0, \infty)$: two-step inference is **biased**
  - degree of bias is increasing in κ (relative importance of measurement vs sampling error)
  - no variance distortion, unlike generated regressors
- $\kappa = 0$: two-step inference is **valid**
  - can treat $\hat{\boldsymbol{\theta}}_i$ as if they are the true latent $\boldsymbol{\theta}_i$
  - analogy with Factor-Augmented Regressions (Bai Ng 2006):
    - impute latent factor $\mathbf{F}_t$ from N-dim cross-section of predictors $\mathbf{x}_t \to \hat{\mathbf{F}}_t$
    - Bai-Ng condition for valid two-step inference: $\sqrt{T}/N \to 0$
    - analogous to $\kappa = 0$: n analogous to T, $\mathbb{E}[C_i^{-1}]$ analogous $1/N$,

Implications

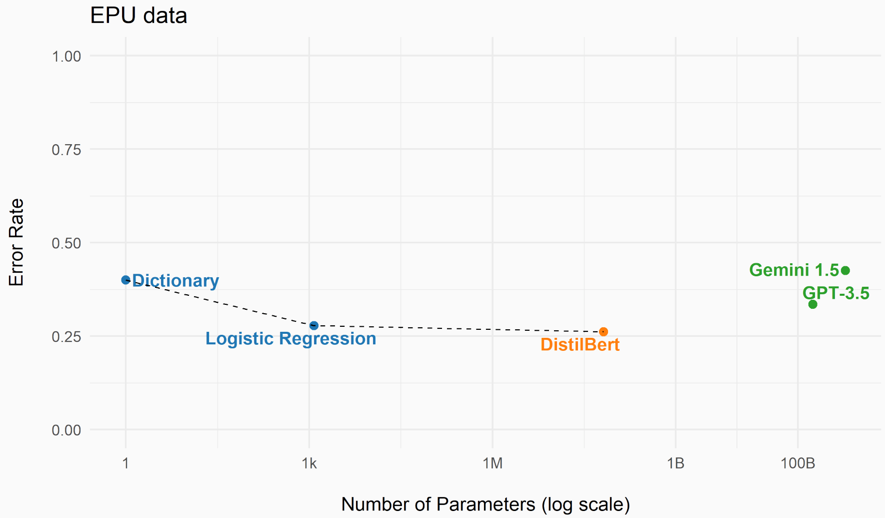
- $\kappa \in (0, \infty)$: two-step inference is **biased**
  - degree of bias is increasing in κ (relative importance of measurement vs sampling error)
  - no variance distortion, unlike generated regressors
- $\kappa = 0$: two-step inference is **valid**
  - can treat $\hat{\boldsymbol{\theta}}_i$ as if they are the true latent $\boldsymbol{\theta}_i$
  - analogy with Factor-Augmented Regressions (Bai Ng 2006):
    - impute latent factor $\mathbf{F}_t$ from N-dim cross-section of predictors $\mathbf{x}_t \to \hat{\mathbf{F}}_t$
    - Bai-Ng condition for valid two-step inference: $\sqrt{T}/N \to 0$
    - analogous to $\kappa = 0$: n analogous to T, $\mathbb{E}[C_i^{-1}]$ analogous $1/N$,
- Practical take-away: if κ is large, use resources for improving precision of $\hat{\boldsymbol{\theta}}_i$ (not increasing n)

Relevance of Measurement Error

Confusion Matrix from Baker, Bloom, and Davis (2016).

	Classification Labels	
Human Labels	0	1
0	1486	474
1	825	802

Errors Remain with Modern Algorithms



Relevance of Sampling Error

For several popular datasets, we can compute an empirical analogue of $\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right]$.

- Minimum Data Set (MDS) for Nursing Homes
  - 24,000,000 patients
  - $\hat{\kappa} \approx 46$
- Lightcast (formerly Burning Glass) job postings data
  - 45,000,000 observations
  - $\hat{\kappa} \approx 20$
- Nielsen Homescan
  - 40,000 households

    $\hat{\kappa} \approx 3.8$
- US Patents in 2023
  - 315,000 filings
  - $\hat{\kappa} \approx 1$

Outline

How to Correct Bias

1. **Explicit Bias Correction**: use analytical expressions in Theorem to adjust two-step estimates

   Advantage: Simple and scalable

   Disadvantage: Not feasible in complex models; poor approximation with large κ

2. **One-Step Strategy**: MLE using joint likelihood for upstream IR model + regression model

   Advantage: General purpose and flexible

   Disadvantage: More computationally demanding and parametric assumptions

Explicit Bias Correction: ML-Generated Labels

- Idea: estimate bias term in asy. dist.; adjust two-step CI accordingly

$$\widehat{\kappa} \times \widehat{\mathbf{bias}} = -\sqrt{n}\,\widehat{fpr} \times \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\xi}}_i \hat{\boldsymbol{\xi}}_i^T\right)^{-1} \begin{bmatrix} \hat{\gamma} \\ \mathbf{0} \end{bmatrix}}_{\substack{\text{consistent under} \\ \text{cond'ns of theorem}}}$$

- Estimate false-positive rate from subsample of correctly labeled data of size $m \ll n$:

$$\widehat{fpr} = \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}_j (1 - \theta_j),$$

- Valid coverage of bias-corrected CIs provided $n/m^2 \to 0$ and $\sqrt{n}\,\mathbb{E}[\pi_i(1 - p_i)] \to \kappa \geq 0$

Explicit Bias Correction: Topic Models

- Bias estimator:

$$\hat{\kappa}\,\widehat{\textbf{bias}} = \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{C_i}\right)}_{\hat{\kappa}}\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix} \boldsymbol{S}\left(\mathbf{Q}_{\hat{\mathbf{B}}}\,\mathrm{diag}(\hat{\mathbf{B}}^T\bar{\vartheta}_n)\mathbf{Q}_{\hat{\mathbf{B}}}^T - \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\vartheta}}_i\hat{\boldsymbol{\vartheta}}_i^T\right)\boldsymbol{S}^T\hat{\gamma} \\ \mathbf{0} \end{bmatrix}}_{\widehat{\textbf{bias}}}$$

where $\mathbf{Q}_{\widehat{\mathbf{B}}} = (\widehat{\mathbf{B}}\widehat{\mathbf{B}}^T)^{-1}\widehat{\mathbf{B}}$

- Valid provided $\sqrt{n} \times \mathbb{E}[C_i^{-1}] \to \kappa \geq 0$

One-Step Strategy: Computation

- Joint likelihood: $f(Y_i, \mathbf{x}_i, \boldsymbol{\theta}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, \ldots)$

- Integrated likelihood in terms of observables only:

$$f(Y_i, \mathbf{x}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, \ldots) = \underbrace{\int_{\Delta^{K-1}} f(Y_i, \mathbf{x}_i, \boldsymbol{\theta}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, \ldots)\, d\boldsymbol{\theta}_i}_{\text{intractable}}$$

- Use Bayesian computation:
  - Integrates out $\boldsymbol{\theta}_i$ as part of the sampling algorithm
  - Resulting credible sets are valid frequentist confidence intervals for large n by BvM theorem

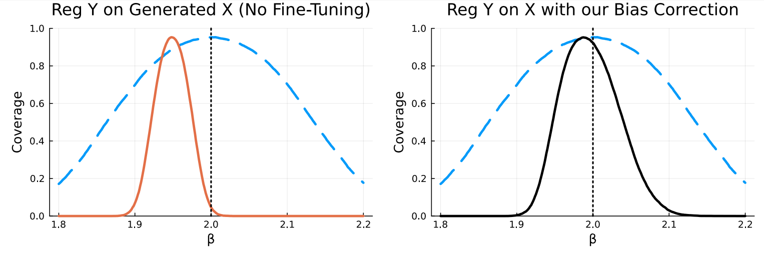
- Sampling: Hamiltonian MC implemented in probabilistic programming language NumPyro
  $\Rightarrow$ allows for estimation of models on large scale

Outline

1. Introduction
2. Warmup Example
3. Full Model
4. How to Correct Bias
5. Empirical Evidence: Simulations
6. Empirical Evidence: CEO Time Use
7. Conclusion

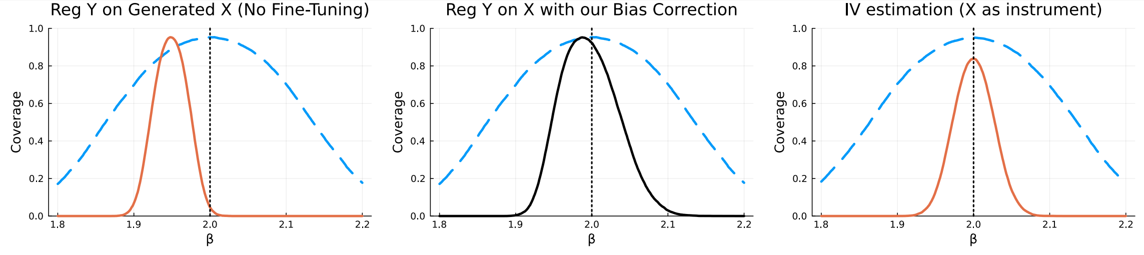
ML-Generated Labels: Coverage in a Small Simulation

- $n = 25{,}000$, $m = 1{,}000$, t_{12} errors, $\gamma_1 = 2$, $\kappa = 2$ (fpr $\approx 1.2\%$)



ML-Generated Labels: Coverage in a Small Simulation

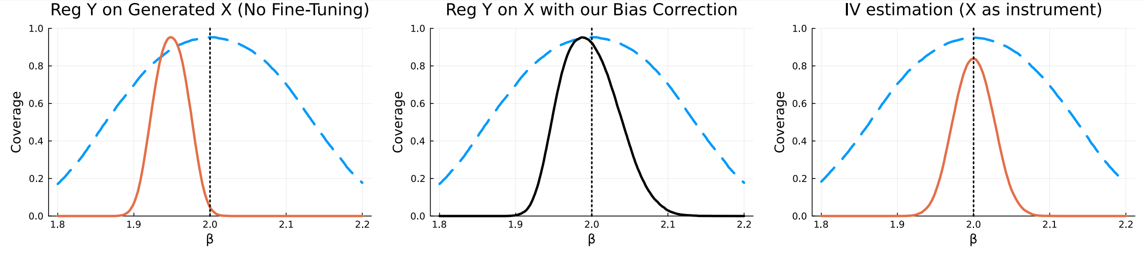
- $n = 25{,}000$, $m = 1{,}000$, t_{12} errors, $\gamma_1 = 2$, $\kappa = 2$ (fpr $\approx 1.2\%$)



- Several recent works (Fong Tyler 2021; Allon et al. 2023; Egami et al. 2023, Zhang et al. 2023) propose IV or GMM strategies based on using small subset w/ correct labels to estimate first stage
- Valid when $n/m \to c$, as in the literature on auxiliary data (Chen et al. 2005, 2008)

ML-Generated Labels: Coverage in a Small Simulation

- $n = 25{,}000$, $m = 1{,}000$, t_{12} errors, $\gamma_1 = 2$, $\kappa = 2$ (fpr $\approx 1.2\%$)



- Several recent works (Fong Tyler 2021; Allon et al. 2023; Egami et al. 2023, Zhang et al. 2023) propose IV or GMM strategies based on using small subset w/ correct labels to estimate first stage
- Valid when $n/m \to c$, as in the literature on auxiliary data (Chen et al. 2005, 2008)
- Not well suited to modern use cases where $n \gg m \Rightarrow$ coverage suffers

Supervised Topic Model with Covariates (STMC)

$$\left.\begin{aligned}\boldsymbol{\theta}_i &\sim \text{LogisticNormal}(\boldsymbol{\Phi}\mathbf{g}_i, \mathrm{I}_K\sigma_\theta^2)\\ \mathbf{x}_i &\sim \text{Multinomial}(C_i, \mathbf{B}^T\boldsymbol{\theta}_i)\\ Y_i &\sim \text{Normal}(\boldsymbol{\gamma}^T\boldsymbol{\theta}_i + \boldsymbol{\alpha}^T\mathbf{q}_i, \sigma_Y^2)\end{aligned}\right\} \quad \rightarrow \quad f(Y_i, \mathbf{x}_i, \boldsymbol{\theta}_i | C_i, \mathbf{q}_i, \mathbf{g}_i; \boldsymbol{\delta})$$

Parameters are $\boldsymbol{\delta} = (\mathbf{B}, \boldsymbol{\Phi}, \gamma, \boldsymbol{\alpha}, \sigma_Y, \sigma_\theta)$

Generalization of Structural Topic Model (Roberts et. al. 2014) and Bayesian Topic Regression for Causal Inference (Ahrens et. al. 2021).

Monte Carlo Design

- Simulate from STMC
- Configurations: $n = 10000$, $C_i = C \in \{10, 25, 200\} \rightarrow \kappa \in \{10, 4, 0.5\}$
- Compare: two-step, one-step, and two-step infeasible (regression on true latent $\boldsymbol{\theta}_i$)

Parameter	Value	Description
	(a) Data Simulation	
V	300	Number of distinct features
K	2	Number of latent types
True ϕ	**1**	Effect of a covariates on un-normalized type shares
True γ	**5**	Effect of topic shares on numerical outcomes
True α	$(0, 1, 1, 1)$	Effect of additional covariates on numerical outcomes
g_i	$\sim N(0, \frac{\log(3)}{1.96})$	Covariate affecting type shares
$q_{i,m}$ $\forall m \in (1, 2, 3)$	$\sim N(0, 3)$	Additional covariates affecting outcome
σ_Y^2	16	SD of the numeric outcome's residual
σ_θ^2	1	SD of residual of the un-normalized type shares
η	0.2	Dirichlet concentration parameter
	(b) Hyperparameters	
$p(\phi_1)$	$N(0, 4)$	Prior for ϕ_1, i.e. $\sigma_\phi^2 = 4$
$p(\gamma_1)$	$N(0, 100)$	Prior for γ_1, i.e. $\sigma_\gamma^2 = 100$
$p(\alpha)$ $\forall m \in (0, 1, 2, 3)$	$N(0, 100)$	Prior for α, i.e. $\sigma_\alpha^2 = 100$
$p(\sigma_Y)$	Gamma$(1, 10)$	Prior for σ_Y, i.e. $s_0 = 1$ and $s_1 = 10$

Performance of One-Step Strategy

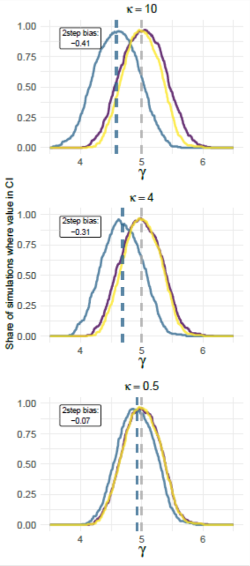


Table 1: Coverage Rates of 95% CIs

	2-Step	1-Step	Infeas
κ	Coverage for γ		
10	0.575	0.955	0.955
4	0.635	0.965	0.955
0.5	0.910	0.960	0.955

Performance of Bias Correction

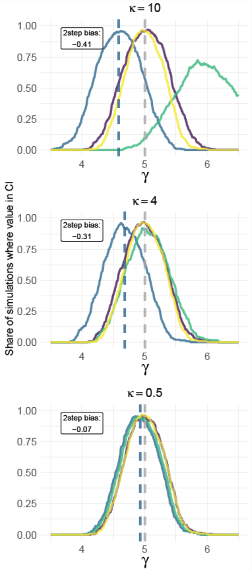


Table 2: Coverage Rates of 95% CIs

κ	2-Step	2-Step BC	1-Step	Infeas
	Coverage for γ			
10	0.575	0.095	0.955	0.955
4	0.635	0.915	0.965	0.955
0.5	0.910	0.935	0.960	0.955

Outline

Bandiera Hansen Prat Sadun (JPE, 2020)

- Time-use survey data for 916 CEOs
- 654 combinations of activities (e.g., meeting with suppliers) in 15min intervals
- LDA with $K = 2$: 2 types of CEO behaviors $\boldsymbol{\beta}_1$ (leaders) and $\boldsymbol{\beta}_2$ (managers).
- Two-step strategy: regress log sales Y_i on leader weight $\hat{\theta}_{i,1}$ and firm characteristics $\mathbf{q}_i$.

Bandiera Hansen Prat Sadun (JPE, 2020)

- Time-use survey data for 916 CEOs
- 654 combinations of activities (e.g., meeting with suppliers) in 15min intervals
- LDA with $K = 2$: 2 types of CEO behaviors $\boldsymbol{\beta}_1$ (leaders) and $\boldsymbol{\beta}_2$ (managers).
- Two-step strategy: regress log sales Y_i on leader weight $\hat{\theta}_{i,1}$ and firm characteristics $\mathbf{q}_i$.

Original Paper: $\hat{\kappa} = 0.44$ (average $C_i = 88.4$).

Modified Sample: draw 10% of activities for each CEO (without replacement) $\longrightarrow \hat{\kappa} = 4.26$.

Observed Activities High: Similar Coefficient Estimates

	Dependent variable: Log(sales)			
	Full Sample		10% Subsample	
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.400	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)

Observed Activities High: Similar Confidence Interval Widths

	Dependent variable: Log(sales)			
	Full Sample		10% Subsample	
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.400	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)

Observed Activities Low: Two-Step Coefficient Estimate Falls

	Dependent variable: Log(sales)			
	Full Sample		10% Subsample	
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.400	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)

Observed Activities Low: Similar Confidence Interval Widths

	Dependent variable: Log(sales)			
	Full Sample		10% Subsample	
	(1) 2-Step	(2) 1-Step	(3) 2-Step	(4) 1-Step
CEO Index	0.400	0.402	0.211	0.439
	(0.219, 0.572)	(0.240, 0.603)	(-0.028, 0.449)	(0.153, 0.711)
Log Employment	1.212	1.198	1.239	1.199
	(1.159, 1.268)	(1.154, 1.248)	(1.186, 1.29)	(1.148, 1.26)

Comparison of $\hat{\theta}_{i,1}$ from One-Step v Two-Step Strategies

(a) Full Sample

(b) 10% Sample

Outline

Conclusion

- Empirical work increasingly uses unstructured data to recover latent variables of economic interest
- We show: dominant two-step strategy leads to invalid inference in most empirical settings
- We propose two solutions: bias correction + one-step strategy
- Illustrate important differences in simulations + applications