AI and Unstructured Data for Measurement and Estimation

Lecture 3: Downstream Regression

Stephen Hansen
University College London

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Outline

Motivation

The first two lectures discussed how to generate variables from AI algorithms.

But this is typically not the final goal for empirical economics.

Instead, we use the generated variables in downstream regressions.

We have seen large measurement error in well-known applications (e.g. EPU index).

Presents challenges for valid statistical inference.

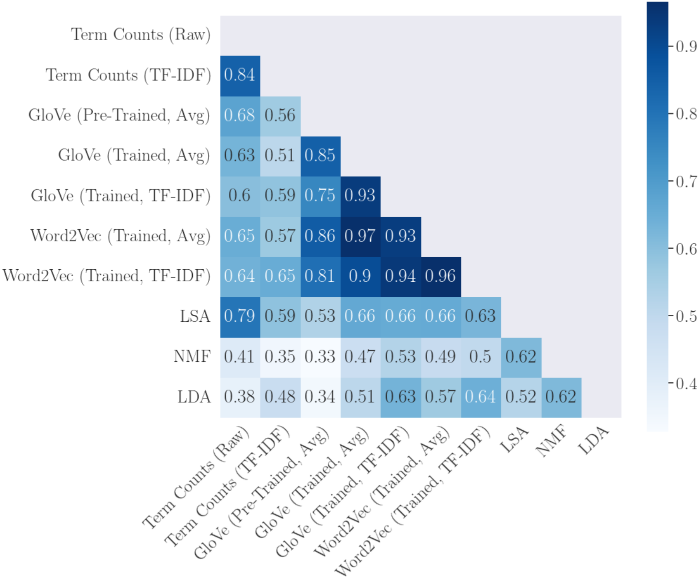
What is the Effect of Measurement Error?

Multiple algorithms for document similarity: bag-of-words, word2vec, BERT, etc.

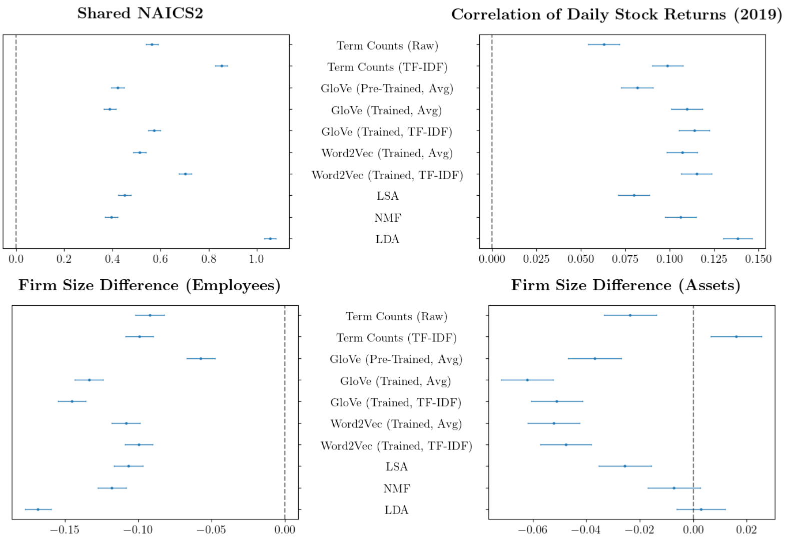
We compute document similarity in the context of 10-K risk factors using randomly sampled pairs from the universe of 2019 filing firms.

Keep data constant, and vary the algorithm used for similarity comparison.

Pearson Correlation Between Similarity Scores Across Pairs



Downstream Regression Results



Formal Setup

Want: perform inference on $\boldsymbol{\gamma}$ and/or $\boldsymbol{\alpha}$ in the model

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

- $\boldsymbol{\theta}_i$ is a latent variable of economic interest
- $\mathbf{q}_i$ are observed numeric covariates
- Unstructured/high-dim dataset $\mathbf{x}_i$ available for estimating $\boldsymbol{\theta}_i$

Formal Setup

Want: perform inference on γ and/or α in the model

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

- $\boldsymbol{\theta}_i$ is a latent variable of economic interest
- $\mathbf{q}_i$ are observed numeric covariates
- Unstructured/high-dim dataset $\mathbf{x}_i$ available for estimating $\boldsymbol{\theta}_i$

Two-Step Strategy:

1. Estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ obtained from unstructured data $\mathbf{x}_i$ and ML/AI model.
2. Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Perform inference treating $\hat{\boldsymbol{\theta}}_i$ as regular numeric data.

Existing Solutions

Suppose some observations have observed $\boldsymbol{\theta}_i$: $(Y_i, \hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i, \mathbf{q}_i)_{i=1}^{m}$ validation sample

Unlabeled data is $\boldsymbol{\theta}_i$: $(Y_i, \hat{\boldsymbol{\theta}}_i, \mathbf{q}_i)_{i=m+1}^{n+m}$ main sample

Growing literature proposes using the validation sample to bias-correct regression estimates from the main sample (e.g. 2SLS, GMM):

- General ML-generated variables: Fong and Tyler (2021), Allon et al. (2023), Angelopoulos et al. (2023a, 2023b), Zhang et al. (2023), Zrnic and Candès (2024), and Miao and Lu (2024)
- LLM-generated variables: Egami et al. (2023, 2024), Ludwig et al. (2025), Carlson and Dell (2025).

Valid inference requires $\frac{n}{m} \to c$.

Validation Data in Economics

Validation dataset typically requires human labels.

But use of ML/AI typically motivated:

1. Large n
2. Large cost of human labels

Unclear how a validation dataset of appropriate size can be constructed in this environment.

Deeper question: unclear that a "true" $\boldsymbol{\theta}_i$ can be observed.

Validation Data in Economics

Validation dataset typically requires human labels.

But use of ML/AI typically motivated:

1. Large n
2. Large cost of human labels

Unclear how a validation dataset of appropriate size can be constructed in this environment.

Deeper question: unclear that a "true" $\boldsymbol{\theta}_i$ can be observed.

Need for valid inference without (much) validation data!

Contributions of [Battaglia et al., 2025]

1. Asymptotic framework where measurement error $\downarrow$ as sample size $\uparrow$.

   Two-step CIs have right width but wrong centering (bias) which depends on relative importance of

   (a) measurement error in $\hat{\theta}_i$

   (b) sampling error in downstream model

   Valid two-step inference requires (a) $\ll$ (b)

   This is not the case in most leading applications

2. Two solutions: bias correction + one-step strategy
   NB: Measurement error in AI/ML-generated variables is non-classical.

3. Shows empirical relevance in several empirical applications

Outline

Example 1: AI/ML-Generated Labels

- Leading use case: missing θ_i is binary (e.g., race indicator): Goldsmith-Pinkham and Shue (2023), Adams-Prassl et. al. (2023), Argyle et al. (2025), and Wu and Yang (2024)
- Generate estimate $\hat{\theta}_i$ of θ_i using unstructured data $\mathbf{x}_i$ (e.g., voter registration data)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$

Example 1: AI/ML-Generated Labels

- Leading use case: missing θ_i is binary (e.g., race indicator): Goldsmith-Pinkham and Shue (2023), Adams-Prassl et. al. (2023), Argyle et al. (2025), and Wu and Yang (2024)
- Generate estimate $\hat{\theta}_i$ of θ_i using unstructured data $\mathbf{x}_i$ (e.g., voter registration data)
- Regress Y_i on $\hat{\theta}_i$ and controls $\mathbf{q}_i$
- Measurement error due to misclassification error:

$$\Pr(\theta_i = 1 | \mathbf{x}_i, \mathbf{q}_i) \neq \Pr(\hat{\theta}_i = 1 | \mathbf{x}_i, \mathbf{q}_i)$$

Example 2: Indices

- Several influential works generate indices by classifying documents + aggregating: Baker Bloom Davis (2016), Caldara and Iacoviello (2022), Gorodnichenko Pham Talavera (2023).
- Each month observe C_i documents (e.g., set of newspapers)
- Of these, X_i are classified as pertaining to concept (e.g., policy uncertainty)
- Latent true uncertainty $\theta_i \in [0, 1]$
- Naive estimator: $\hat{\theta}_i = X_i / C_i$ (cf. BBD's EPU measure)

Example 2: Indices

- Several influential works generate indices by classifying documents + aggregating: Baker Bloom Davis (2016), Caldara and Iacoviello (2022), Gorodnichenko Pham Talavera (2023).
- Each month observe C_i documents (e.g., set of newspapers)
- Of these, X_i are classified as pertaining to concept (e.g., policy uncertainty)
- Latent true uncertainty $\theta_i \in [0, 1]$
- Naive estimator: $\hat{\theta}_i = X_i / C_i$ (cf. BBD's EPU measure)
- Natural model is $X_i | C_i, \theta_i \sim \text{Binomial}(C_i, \theta_i \beta_1 + (1 - \theta_i)\beta_0)$ where β_x is the probability that a document with true label x is classified a one.
- Measurement error in $\hat{\theta}_i$ arises from misclassification error ($\boldsymbol{\beta}$) and sampling error (C_i).

Example 2: Indices — Simulation Calibrated to Gorodnichenko et al. (2023)

Configuration	Bias 1	Bias 2	Bias 3	RMdSE 1	RMdSE 2	RMdSE 3	Coverage 1	Coverage 2	Coverage 3
				$n = 200$					
2-Step	-0.433	-0.218	-0.037	0.048	0.025	0.018	0.378	0.824	0.931
Joint	-0.003	0.007	0.004	0.024	0.020	0.018	0.945	0.948	0.938
				$n = 800$					
2-Step	-0.215	-0.041	0.084	0.024	0.010	0.012	0.507	0.942	0.894
Joint	0.004	-0.006	-0.006	0.011	0.010	0.010	0.956	0.950	0.950
				$n = 3200$					
2-Step	-0.042	0.085	0.158	0.006	0.009	0.017	0.887	0.739	0.353
Joint	-0.005	-0.002	-0.003	0.005	0.005	0.005	0.942	0.941	0.943

Outline

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

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

  (expressions are DGP-specific)

- Scalar $\kappa \geq 0$ measures the importance of measurement error relative to sampling error

- Positive κ allows both sampling error and measurement error to play a role

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

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

  (expressions are DGP-specific)

- Scalar $\kappa \geq 0$ measures the importance of measurement error relative to sampling error
- Positive κ allows both sampling error and measurement error to play a role
- Example 1 expression is

$$\sqrt{n} \times \underbrace{\mathbb{E}\left[\hat{\theta}_i (1 - \theta_i)\right]}_{\text{false-positive rate}} \to \kappa, \qquad \boldsymbol{\Omega} = 1$$

- Reflects prevailing trend: increasingly large data sets + increasingly accurate algorithms

Theorem on Two-Step Inference

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

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

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \to_d N\left(-\kappa \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1} \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{\psi}\,,\ \underbrace{\mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1} \mathbb{E}[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T] \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1}}_{=:\mathbf{V}}\right)$$

where $\boldsymbol{\xi}_i = (\boldsymbol{\theta}_i, \mathbf{q}_i)$ are the "true" covariates

Theorem on Two-Step Inference

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

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

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \rightarrow_d N\left(-\kappa \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1} \begin{pmatrix} \boldsymbol{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{\psi}\,,\ \underbrace{\mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1} \mathbb{E}[\varepsilon_i^2 \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T] \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T]^{-1}}_{=:\mathbf{V}}\right)$$

   where $\boldsymbol{\xi}_i = (\boldsymbol{\theta}_i, \mathbf{q}_i)$ are the "true" covariates

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

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

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**: treat $\hat{\boldsymbol{\theta}}_i$ as if they are the true latent $\boldsymbol{\theta}_i$
- Take-away: if κ is large, consider using resources to improve precision of $\hat{\boldsymbol{\theta}}_i$

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**: treat $\hat{\boldsymbol{\theta}}_i$ as if they are the true latent $\boldsymbol{\theta}_i$
- Take-away: if κ is large, consider using resources to improve precision of $\hat{\boldsymbol{\theta}}_i$
- To the extent empirical papers flag concerns about 2-step inference, usually about std errors
- Common intuition is wrong: problem is measurement error not standard errors

Outline

How to Do Valid Inference

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

   Advantage: Simple and scalable

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

2. **Joint Estimation**: MLE using joint likelihood for upstream IR model + regression model

   Advantage: General purpose and flexible

   Disadvantage: More computationally demanding

Bias Correction

- First-order asymptotic bias of OLS estimator $\hat{\boldsymbol{\psi}}$ is

$$-\kappa\,\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\right]^{-1}\begin{pmatrix}\boldsymbol{\Omega} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\boldsymbol{\psi}$$

- Given estimators $\hat{\kappa}$ and $\hat{\boldsymbol{\Omega}}$ of κ and $\boldsymbol{\Omega}$, can construct bias-corrected estimators:

Additive $$\hat{\boldsymbol{\psi}}^{bca} = \left(\mathbf{I} + \frac{\hat{\kappa}}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\boldsymbol{\Omega}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)\hat{\boldsymbol{\psi}}$$

Multiplicative $$\hat{\boldsymbol{\psi}}^{bcm} = \left(\mathbf{I} - \frac{\hat{\kappa}}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^T\right)^{-1}\begin{bmatrix}\hat{\boldsymbol{\Omega}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)^{-1}\hat{\boldsymbol{\psi}}$$

- Bias-corrected CIs: center at $\hat{\boldsymbol{\psi}}^{bca}$ or $\hat{\boldsymbol{\psi}}^{bcm}$ and use 2-step std errors

Theory for Bias-Correction

Validity of Bias-Corrected Inference

If $\hat{\kappa} \to_p \kappa$ and $\hat{\mathbf{\Omega}} \to_p \mathbf{\Omega}$, the under conditions of previous theorem, have

1. Bias-corrected estimators are asymptotically equivalent and correctly centered

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bcm} - \boldsymbol{\psi}\right) = \sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bca} - \boldsymbol{\psi}\right) + o_p(1) \to_d N\left(\mathbf{0}\,, \mathbf{V}\right)$$

Theory for Bias-Correction

Validity of Bias-Corrected Inference

If $\hat{\kappa} \rightarrow_p \kappa$ and $\hat{\boldsymbol{\Omega}} \rightarrow_p \boldsymbol{\Omega}$, the under conditions of previous theorem, have

1. Bias-corrected estimators are asymptotically equivalent and correctly centered

$$\sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bcm} - \boldsymbol{\psi}\right) = \sqrt{n}\left(\hat{\boldsymbol{\psi}}^{bca} - \boldsymbol{\psi}\right) + o_p(1) \rightarrow_d N\left(\mathbf{0}\,,\mathbf{V}\right)$$

2. Bias-corrected CIs have correct coverage:

$$\lim_{n\to\infty} \Pr\left(\psi_i \in \hat{\psi}_i^{bc} \pm 1.96\sqrt{\frac{\hat{\mathbf{V}}_{ii}}{n}}\right) = 0.95.$$

Bias Correction: Labels Example

- Here need to estimate $\kappa = \sqrt{n} \lim_{n\to\infty} \mathbb{E}\left[\hat{\theta}_i(1-\theta_i)\right]$
- Just need an estimate of FPR from an external sample (as in Bursztyn Chaney Hassan Rao (2024))

$$\hat{\kappa} = \sqrt{n}\widehat{FPR}, \qquad \widehat{FPR} = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i(1-\theta_i)$$

- We show $\hat{\kappa} \to_p \kappa$ provided $n/m^2 \to 0$ (small validation data!)
- We also provide finite-sample correction to standard errors (complex expression).

Joint Estimation: Computation

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

- Integrated likelihood in terms of observables only:

$$f(Y_i, \mathbf{x}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, ...) = \underbrace{\int f(Y_i, \mathbf{x}_i, \boldsymbol{\theta}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, ...) \, d\boldsymbol{\theta}_i}_{\text{intractable}}$$

- 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

Joint Estimation: 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:

$$f(Y_i, \mathbf{x}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, \ldots) = \underbrace{\int f(Y_i, \mathbf{x}_i, \boldsymbol{\theta}_i | \mathbf{q}_i; \gamma, \boldsymbol{\alpha}, \ldots) \, d\boldsymbol{\theta}_i}_{\text{intractable}}$$

- 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

- Note: in examples, we are **not** attempting to specify a likelihood for the AI/ML algorithm

Outline

Hansen Lambert Bloom Davis Sadun Taska (WP, 2023)

- Consider $n = 16,315$ SD food+accom sector (NAICS code 72) job postings from January 2022
- Regress log wages Y_i on ML-generated remote work indicator $\hat{\theta}_i$
- Fixed effects for SOC code (job type) and full/part time

Hansen Lambert Bloom Davis Sadun Taska (WP, 2023)

- Consider $n = 16,315$ SD food+accom sector (NAICS code 72) job postings from January 2022
- Regress log wages Y_i on ML-generated remote work indicator $\hat{\theta}_i$
- Fixed effects for SOC code (job type) and full/part time
- For bias correction, use estimate $\widehat{FPR} \approx 0.009$.
- For joint estimation, use three-component Gaussian mixture for errors $\varepsilon_i | \theta_i$

Bias Correction with Minimal Human Effort

Advantage 1: Smaller Auxiliary Dataset

Existing papers: bias correction when m and n are comparable.

We estimate FPR with $m = 1000$. $n/m = 16$, $n/m^2 = 0.016$.

Bias Correction with Minimal Human Effort

Advantage 1: Smaller Auxiliary Dataset

Existing papers: bias correction when m and n are comparable.

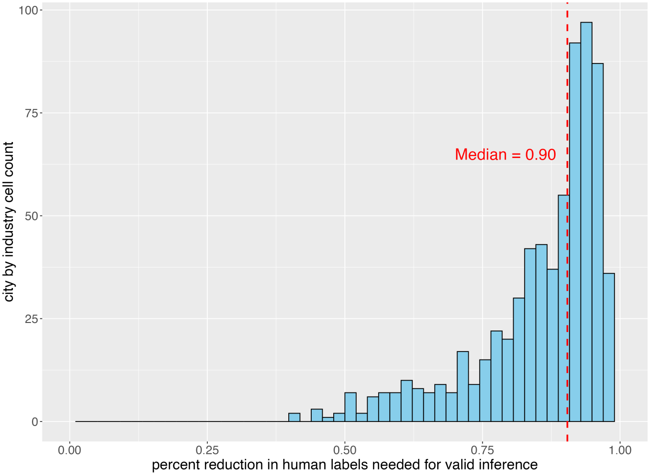
We estimate FPR with $m = 1000$. $n/m = 16$, $n/m^2 = 0.016$.

Advantage 2: Only Need Partial Labeling

Existing papers: build full validation dataset by inspecting each posting.

This paper: only examine labeled "ones", 26 in this dataset.

Distribution of Reduction in Human Labels



Two-Step Estimates Smaller

	No Fixed Effects			With Fixed Effects		
	Est.	Std Err	95% CI	Est.	Std Err	95% CI
OLS	0.648	0.024	[0.599, 0.697]	0.363	0.021	[0.321, 0.406]
BC	1.052	0.140	[0.777, 1.326]	0.641	0.099	[0.446, 0.836]
1-Step	0.563	0.016	[0.532, 0.595]	0.448	0.017	[0.415, 0.480]

Corrected CIs to the right of Two-Step CIs

	No Fixed Effects			With Fixed Effects		
	Est.	Std Err	95% CI	Est.	Std Err	95% CI
OLS	0.648	0.024	[0.599, 0.697]	0.363	0.021	[0.321, 0.406]
BC	1.052	0.140	[0.777, 1.326]	0.641	0.099	[0.446, 0.836]
1-Step	0.563	0.016	[0.532, 0.595]	0.448	0.017	[0.415, 0.480]

Outline

Central Bank Communication

- Does written central bank communication drive long rates? Estimate

$$Y_i = \gamma\theta_i + \boldsymbol{\alpha}'\mathbf{q}_i + u_i$$

  - Y_i is the path factor from Gürkaynak, Sack, and Swanson (2005) (mkt perceptions of future rates)
  - θ_i is a hawkish/dovish index (cf. Gorodnichenko, Pham, Talavera (2023))
  - $\mathbf{q}_i$ are controls (including shadow short rate)

Central Bank Communication

- Does written central bank communication drive long rates? Estimate

$$Y_i = \gamma\theta_i + \boldsymbol{\alpha}'\mathbf{q}_i + u_i$$

  - Y_i is the path factor from Gürkaynak, Sack, and Swanson (2005) (mkt perceptions of future rates)
  - θ_i is a hawkish/dovish index (cf. Gorodnichenko, Pham, Talavera (2023))
  - $\mathbf{q}_i$ are controls (including shadow short rate)

- Hawkish/dovish index:
  - classify FOMC sentences as hawkish/dovish/neutral using fine-tuned BERT + aggregate
  - sentiment estimate

$$\hat{\theta}_i = \frac{N_i^H - N_i^D}{N_i^H + N_i^D}$$

Central Bank Communication

- Does written central bank communication drive long rates? Estimate

$$Y_i = \gamma\theta_i + \boldsymbol{\alpha}'\mathbf{q}_i + u_i$$

  - Y_i is the path factor from Gürkaynak, Sack, and Swanson (2005) (mkt perceptions of future rates)
  - θ_i is a hawkish/dovish index (cf. Gorodnichenko, Pham, Talavera (2023))
  - $\mathbf{q}_i$ are controls (including shadow short rate)

- Hawkish/dovish index:
  - classify FOMC sentences as hawkish/dovish/neutral using fine-tuned BERT + aggregate
  - sentiment estimate

$$\hat{\theta}_i = \frac{N_i^H - N_i^D}{N_i^H + N_i^D}$$

- Compare two-step and joint estimation over 02/1995-06/2023

Central Bank Communication: Joint Estimation Effect Size 3x Larger

	Estimation Strategy	
	Two-Step	Joint
Sentiment (θ_i)	0.039	0.114
	[0.012, 0.066]	[0.027, 0.198]
Policy Rate (q_i)	-0.004	-0.003
	[-0.011, 0.003]	[-0.011, 0.004]
β_0		0.009
		[0.001, 0.026]
β_1		0.676
		[0.585, 0.768]
Observations	200	200
R^2	0.0425	0.1429

Central Bank Communication: Material Misclassification Error

	Estimation Strategy	
	Two-Step	Joint
Sentiment (θ_i)	0.039	0.114
	[0.012, 0.066]	[0.027, 0.198]
Policy Rate (q_i)	-0.004	-0.003
	[-0.011, 0.003]	[-0.011, 0.004]
β_0		0.009
		[0.001, 0.026]
β_1		0.676
		[0.585, 0.768]
Observations	200	200
R^2	0.0425	0.1429

Outline

Conclusion

- Empirical work routinely uses AI/ML algorithms to generate new variables
- Common empirical practice leads to invalid inference
- We propose two solutions: bias correction + joint estimation. Neither requires validation data.
- Illustrate important differences in simulations + applications
- Packages: `ValidMLInference` (Python) and `MLBC` (R)
- Works in progress: specific methods tailored to important use cases, e.g. VARs and impulse response analysis.

References I

Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025).
Inference for Regression with Variables Generated by AI or Machine Learning.