Inference for Regression with Variables Generated by AI or Machine Learning

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

May 29, 2025

Outline

Motivation

Economists now routinely generate variables by AI/ML methods

- quantify unstructured data (text, images, ...)
- measure subtle concepts (uncertainty, sentiment, ...)
- generate variables previously too costly, labor-intensive, or infeasible to collect

The generated variables are inputs to downstream econometric models

Motivation

Economists now routinely generate variables by AI/ML methods

- quantify unstructured data (text, images, ...)
- measure subtle concepts (uncertainty, sentiment, ...)
- generate variables previously too costly, labor-intensive, or infeasible to collect

The generated variables are inputs to downstream econometric models

Almost all papers use a two-step strategy:

1. Generate estimate $\hat{\theta}_i$ of true variable θ_i using AI/ML algorithm
2. Plug estimates $(\hat{\theta}_i)$ into an econometric model, treating $\hat{\theta}_i$ as regular numeric data

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 $\hat{\theta}_i$

   (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 is AI/ML-generated variables in non-classical.

3. Shows empirical relevance in several empirical applications

Related Work

Recent work (mainly stats/poli sci) has pointed out potential for generated variables to cause problems

- 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)
- Variables generated by LLMs: Egami et al. (2023, 2024), Ludwig et al. (2025), Carlson and Dell (2025).

Related Work

Recent work (mainly stats/poli sci) has pointed out potential for generated variables to cause problems

- 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)
- Variables generated by LLMs: Egami et al. (2023, 2024), Ludwig et al. (2025), Carlson and Dell (2025).

These works propose bias corrections assuming a validation subsample in which $(Y_i, \boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i)$ are observed

- But one can estimate the model using $(Y_i, \boldsymbol{\theta}_i)$ in validation sample!

  AI/ML gen. vars only helpful for efficiency
- Not feasible in most economic use cases where $\boldsymbol{\theta}_i$ is truly latent (uncertainty, sentiment, ...)

Outline

1. Introduction
2. Setup and Use Cases
3. Two-Step Inference is Biased
4. How to Do Valid Inference
5. Application: Remote Work and Wage Inequality
6. Application: CEO Time Use and Firm Performance
7. Application: Central Bank Communication
8. Conclusion

Setup

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

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

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

Setup

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

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

- $\boldsymbol{\theta}_i$ 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 $\mathbf{x}_i$
2. Regress Y_i on $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{q}_i$. Perform inference treating $\hat{\boldsymbol{\theta}}_i$ as regular numeric data.

Example 1: AI/ML-Generated Labels

- ML algorithms often deployed to impute missing observations from unstructured data.

  Goldsmith-Pinkham and Shue (2023), Adams-Prassl et. al. (2023), Argyle et al. (2025), and Wu and Yang (2024)

- Leading use case: missing θ_i is binary (e.g., race indicator)

- 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

- ML algorithms often deployed to impute missing observations from unstructured data.

  Goldsmith-Pinkham and Shue (2023), Adams-Prassl et. al. (2023), Argyle et al. (2025), and Wu and Yang (2024)

- Leading use case: missing θ_i is binary (e.g., race indicator)

- 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: Dimensionality Reduction

- Obtain low-dimensional representation of unstructured data which is plugged into regression:
  - Text data: Hansen McMahon Prat (2018); Mueller and Rauh (2018); Larsen and Thorsrud (2019); Thorsrud (2020); Bybee Kelly Manela Xiu (2024); Ash Morelli Vannoni (2025)
  - Survey data: Bandiera Prat Hansen Sadun (2020); Draca and Schwarz (2020)
  - Network data: Nimczik (2017)

Example 2: Dimensionality Reduction

- Obtain low-dimensional representation of unstructured data which is plugged into regression:
  - Text data: Hansen McMahon Prat (2018); Mueller and Rauh (2018); Larsen and Thorsrud (2019); Thorsrud (2020); Bybee Kelly Manela Xiu (2024); Ash Morelli Vannoni (2025)
  - Survey data: Bandiera Prat Hansen Sadun (2020); Draca and Schwarz (2020)
  - Network data: Nimczik (2017)
- 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):

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

  - $\mathbf{B}^T = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K]$, each $\boldsymbol{\beta}_k \in \Delta^{V-1}$ is a topic
  - observation-specific topic weights $\boldsymbol{\vartheta}_i \in \Delta^{K-1}$
  - subset of interest: $\boldsymbol{\theta_i} = \boldsymbol{S}\boldsymbol{\vartheta}_i$

Example 2: Dimensionality Reduction

- Obtain low-dimensional representation of unstructured data which is plugged into regression:
  - Text data: Hansen McMahon Prat (2018); Mueller and Rauh (2018); Larsen and Thorsrud (2019); Thorsrud (2020); Bybee Kelly Manela Xiu (2024); Ash Morelli Vannoni (2025)
  - Survey data: Bandiera Prat Hansen Sadun (2020); Draca and Schwarz (2020)
  - Network data: Nimczik (2017)
- 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):

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

  - $\mathbf{B}^T = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]$, each $\boldsymbol{\beta}_k \in \Delta^{V-1}$ is a topic
  - observation-specific topic weights $\boldsymbol{\vartheta}_i \in \Delta^{K-1}$
  - subset of interest: $\boldsymbol{\theta_i} = \boldsymbol{S}\boldsymbol{\vartheta}_i$
- Measurement error due to upstream sampling error in $\hat{\boldsymbol{\theta}}_i$

Example 3: 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 3: 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)

- **Problem**: $\hat{\theta}_i$ is a signal of θ_i
- e.g., could change set of newspapers and get a different (but related) measure

Example 3: Indices

- Topic model representation:

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

  for $\mathbf{x}_i = (X_i, C_i - X_i)^T$,

$$\mathbf{B}^T = \underbrace{\begin{bmatrix} \beta_1 & \beta_0 \\ (1-\beta_1) & (1-\beta_0) \end{bmatrix}}_{\text{misclass. rates}}, \qquad \boldsymbol{\vartheta}_i = \begin{bmatrix} \theta_i \\ 1-\theta_i \end{bmatrix}$$

- Measurement error due to misclassification error and upstream sampling error

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

Asymptotics: κ and Ω

- ML-generated binary labels:

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

- Topic models:

$$\sqrt{n} \times \mathbb{E}\left[\frac{1}{C_i}\right] \to \kappa, \qquad \boldsymbol{\Omega} = \boldsymbol{S}(\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{B}\,\mathrm{diag}(\mathbf{B}^T\mathbb{E}[\boldsymbol{\vartheta}_i])\mathbf{B}^T(\mathbf{B}\mathbf{B}^T)^{-1}\boldsymbol{S}^T - \mathbb{E}\left[\boldsymbol{\theta}_i\boldsymbol{\theta}_i^T\right]$$

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

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

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

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

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

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

   Advantage: General purpose and flexible

   Disadvantage: More computationally demanding

NB: Measurement error is AI/ML-generated variables in non-classical.

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
- Estimate FPR from a subsample of size $m = 1,000$ postings, $\widehat{FPR} \approx 0.009$
- For 1-step: use 3 component Gaussian mixture for errors $\varepsilon_i|\theta_i$

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

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) $\longrightarrow \hat{\kappa} = 4.26$.

Similar Estimates in Full Sample

Table 1: Estimates of Impact of CEO Behavior on Firm Performance

Sample	Estimation Strategy		
	Two-Step	Bias Correction	Joint
Full	0.405	0.474	0.402
	[0.224, 0.585]	[0.294, 0.655]	[0.240, 0.603]
10% Subsample	0.227	1.054	0.439
	[-0.038, 0.492]	[0.789, 1.319]	[0.153, 0.711]

Difference in Subsample

Table 2: Estimates of Impact of CEO Behavior on Firm Performance

Sample	Estimation Strategy		
	Two-Step	Bias Correction	Joint
Full	0.405	0.474	0.402
	[0.224, 0.585]	[0.294, 0.655]	[0.240, 0.603]
10% Subsample	0.227	1.054	0.439
	[-0.038, 0.492]	[0.789, 1.319]	[0.153, 0.711]

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 1 and 2 step methods over 02/1995-06/2023

Central Bank Communication: One-Step Effect Size 3x Larger

	Estimation Strategy	
	Two-Step	One-Step
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 + one-step strategy
- Illustrate important differences in simulations + applications
- Works in progress: specific methods tailored to important use cases
    - VARs and impulse response analysis w/ Hansen and Shin