Measurement with Validation Samples: A Missing-Data Perspective

Melissa Dell Ashesh Rambachan

NBER Methods Lecture 2026

Slides posted on the NBER Methods Lecture website.

Observation

Definition

Observation entails **implementing a specified construct definition at scale**.

- **Deep neural networks** are state-of-the-art tools for large-scale feature extraction from unstructured, high-dimensional data (Goodfellow, Bengio and Courville, 2016).
- They are increasingly used by individual researchers to construct data for empirical research.

Observation with Deep Neural Networks

Neural networks will **not** generically produce unbiased predictions in finite samples.

Systematic biases can arise from:

- network architecture, the distribution of training data, and other implementation details;
- nonlinear transformations at each layer, and the frequent application to binary or multiclass classification.

These features **violate classical measurement error assumptions**.

Structured Data as Proxies

- Structured variables extracted from unstructured data are often treated as **proxies**, implicitly accepting the possibility of arbitrary measurement error.
- But in a world of inexpensive AI (*e.g.*, commercial LLMs), treating observation as the construction of proxies poses **serious challenges to the integrity of empirical work**.

Pitfalls of Treating Observed Measures as Proxies

1. The choice of different models, training data distributions, and other implementation details (unrelated to the construct definition itself) can produce different predictions, leading to **bias**, **interpretibility**, and **post-selection inference** concerns.

Pitfalls of Treating Observed Measures as Proxies

1. The choice of different models, training data distributions, and other implementation details (unrelated to the construct definition itself) can produce different predictions, leading to **bias**, **interpretibility**, and **post-selection inference** concerns.
2. **Sensitivity checks** cannot fully address these concerns, because biases are often systematically correlated across models.

Pitfalls of Treating Observed Measures as Proxies

1. The choice of different models, training data distributions, and other implementation details (unrelated to the construct definition itself) can produce different predictions, leading to **bias**, **interpretibility**, and **post-selection inference** concerns.
2. **Sensitivity checks** cannot fully address these concerns, because biases are often systematically correlated across models.
3. Economists frequently use proprietary neural networks, raising **reproducibility** concerns.

Pitfalls of Treating Observed Measures as Proxies

1. The choice of different models, training data distributions, and other implementation details (unrelated to the construct definition itself) can produce different predictions, leading to **bias**, **interpretibility**, and **post-selection inference** concerns.
2. **Sensitivity checks** cannot fully address these concerns, because biases are often systematically correlated across models.
3. Economists frequently use proprietary neural networks, raising **reproducibility** concerns.
4. It is often possible to improve neural network predictions through **costly investments**, but it is unclear how to assess which investments to make without a principled way to account for how prediction errors affect the estimation of target parameters.

Addressing These Challenges

Shift in perspective: Move from informally choosing plausible proxies to:

- **specifying the measurement aim**: a well-defined construct;
- **validation**: establishing how well that aim is defined and implemented at scale.

Statistical goals:

- A broadly applicable framework that **corrects biases** in observation, yielding target parameters robust to the choice of measurement instrument.
- Better measurements should also **improve precision**, clarifying the cost–benefit tradeoff of measurement investments.

Framing Observation as a Missing Data Problem

Framework. Foundational work on semiparametric inference with missing data, in particular Rubin's (1976) **missing-at-random (MAR)** mechanism, provides just such a framework.

Why missing data? Observation is framed this way because researchers often lack the low-dimensional summaries needed for statistical analysis.

Making this applicable to applied economics with AI predictions. Carlson and Dell (2025) extend this framework as MAR-S: **M**issing **A**t **R**andom **S**tructured Data.

Observation as a Missing Data Problem

Core idea

Use a **validation sample** to estimate bias in imputed structured data and adjust estimates accordingly.

Validation data. Come from an implementable construct definition, specified by the researcher and treated as given; they are obtained through non-scalable processes such as expert annotation or ground-station measurement.

Key assumption: missing at random. After adjusting for observables, annotated and unannotated data are comparable in their ground truth values. There are no unaccounted confounders determining whether an instance of unstructured data is annotated.

Why MAR Is Complementary to Neural Networks

Assumption-light. Semiparametric inference lets the data speak as much as possible; MAR makes **minimal assumptions about the neural network** itself.

Why not model the network instead? Its biases shift in complex ways with the unstructured inputs, and people have difficulty predicting how domain shift affects performance (Vafa, Rambachan and Mullainathan, 2024).

If MAR is not feasible: Researchers must impose **stronger restrictions** on the neural network or the data-generating process.

Debiasing for Applied Economics in the AI Era

Debiasing is not new. Ground-truth annotation samples are central to large literatures on measurement error, semiparametric missing data, causal inference, and valid inference with black-box AI predictions.

MAR-S makes these ideas accessible to applied economists working wtih AI predictions

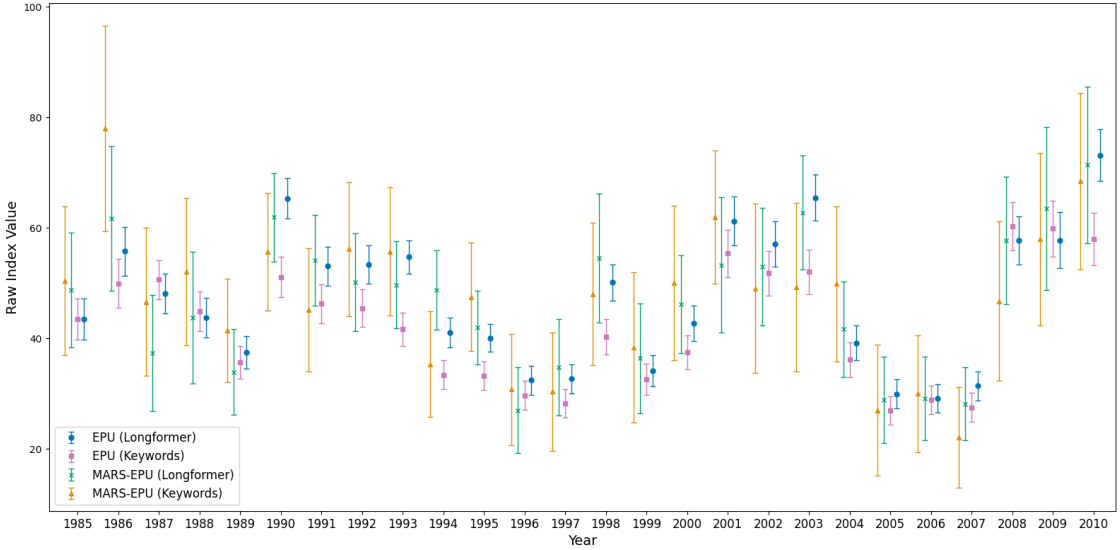
1. **Unifying** insights from these literatures in one applied framework geared towards economists;
2. **Identifying** estimators that are both unbiased and efficient;
3. **Extending** debiasing to common settings with limited guidance, including aggregated missing data.

Outline

Motivating Example: Baker, Bloom, and Davis (QJE 2016)

- Baker, Bloom and Davis (2016) construct an "economic policy uncertainty" (EPU) index by computing the share of articles in a set of newspapers that also satisfy a keyword query, and perform an audit study that collects ground truth on mentions of economic policy uncertainty.
- In the following figure, we present the original and MAR-S debiased EPU indices, that adjust for systematic differences between the predictions and the authors' validation sample. We also plot debiased and naive versions of the EPU index, imputed with a neural network fine-tuned on a split of their validation sample.
- We utilize 25% of their validation sample labels for debiasing, and use the rest for imputation.

Debiased EPU



The Debiased Estimation Literature

	PPI(++)	XPPI	ASI	RePPI	DSL	CDI	PTD	CEP-GMM	MAR-S
Estimands	M	M	M	M	M	M	Z	Z	Z
Assumptions	MAR+ PPI	MAR+ PPI	MAR+ PPI	MAR+ PPI	MAR+ PPI	MAR+ PPI	MAR	MAR	MAR
Allows unknown π?	No	No	No	No	No	No	No	Yes	Yes
Allows nonuniform π?	No	No	Yes	No	Yes	Yes	Yes	Yes	Yes
Semip. Efficiency?	No	No	No	Yes	No	No	No	Yes	Yes
Finetuning?	No	Yes	No	Yes	Yes	No	No	Yes	Yes
Sample Splitting?	No	Yes	No	Yes	Yes	No	No	No	Yes
High-Dim. Features?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Aggregation?	No	No	No	No	No	No	No	No	Yes

PPI(++): Angelopoulos et al. (2023); Angelopoulos et al. (2024); XPPI: Zrnic and Candès (2024b); ASI: Zrnic and Candès (2024a); RePPI: Ji, Lei and Zrnic (2025); DSL: Egami(2024); CDI: Gligoric et al., 2025); PTD: Kluger et al. (2025); CEP-GMM: Chen et al. (2008)

Missing At Random Structured data

- To perform robust and efficient inference with *high-dimensional, unstructured data*, we recast the problem as inference on *missing low-dimensional structured data*.
- Structured data, $M \in \mathcal{M}$, are low-dimensional data that can be used directly in estimating equations.
- Unstructured data, $U \in \mathcal{U}$, are high-dimensional and unsuitable for direct use in estimation

Missing At Random Structured data

- Low-dimensional, structured data are observed through "**annotation**".
- Because **criteria-based measurement is too expensive to scale**, we estimate a function to impute missing structured data $\hat{\mu}$.
- This allows the researcher to leverage the **full unstructured dataset**, often orders of magnitude larger than the annotated data.
- **Deep neural networks** increasingly serve as this imputation function.

Potential Outcomes for Missing Data

MAR-S is closely linked to the Rubin Causal Model (Neyman, 1923; Rubin, 1974, 1978; Imbens and Rubin, 2015), and hence we incorporate the notation of potential outcomes into MAR-S

- We observe a random variable $M \in \mathcal{M} \subseteq \mathbb{R}$ that is subject to some data missingness, given by indicator $A \in \{0, 1\}$,

$$M = A\, M^{a=1} + (1 - A)\, M^{a=0}$$

- We set $M(a = 0) = 0$ w.p.1 WLOG, and, for notational convenience, define $M^* := M(a = 1)$, and so

$$M = AM^*.$$

- We call M^* the "ground truth" potential outcome.
- We call A the "annotation indicator."
- We also assume $P(A = 1) \gg 0$, which means that we have "annotated" some non-negligible count of our unstructured data.

Assumptions

Assumption (Consistency of potential outcomes)

For ground truth potential outcome of interest $M^* \in \mathcal{M}$*, observed outcome of interest* $M \in \mathcal{M} \times \{0\}$*, and annotation indicator* $A \in \{0, 1\}$*,*

$$M = AM^*.$$

- Annotation status needs to be well-defined, which will tend to hold trivially.
- The label for any given instance depends only on its own annotation status, not on the annotation status of other instances.

Assumptions

Assumption (Missing at random)

For ground truth potential outcome $M^* \in \mathcal{M}$, annotation indicator $A \in \{0, 1\}$, observed covariates $X \in \mathcal{X}$, and unstructured data $U \in \mathcal{U}$:

$$[(U, M^*) \perp\!\!\!\perp A] \mid X.$$

- After adjusting for observables X, annotated and unannotated data are comparable in their ground truth values.
- This is analogous to "selection on observables" in causal inference.
- If data are annotated at random, $(U, M^*) \perp\!\!\!\perp A$, the assumption is guaranteed to hold.

Assumptions

Assumption (Known, bounded annotation score function)

We define the annotation score function to be

$$\pi(x) := P(A = 1 \mid X = x) \tag{1}$$

We assume that $\pi(x)$ *is fixed, known, and bounded away from zero and one, i.e.,* $\pi(x) \in [\eta, 1 - \eta]$ *for* $0 < \eta \leq 1 - \eta < 1$ *and for all* $x \in \mathcal{X}$.

- This embeds the assumption of "strong overlap" often seen in observational causal inference settings.
- The annotation function will be known by the researcher who creates the annotated data, but this assumption can be relaxed if needed.

Decaying Overlap

- As Kallus and Mao (2024) make clear in their analysis of treatment effects, there is no need to change the MAR-S estimator itself under an asymptotic analysis where **the number of annotations is diverging but the ratio of labeled data to unlabeled data is converging to zero**.
- Appropriate asymptotic efficiency analysis shows the variance of the estimator is primarily driven by **the size of the labeled dataset**, as opposed to the full dataset size.
- A key difficulty explored in Kallus and Mao (2024) and Zhang et al. (2023) is estimating a propensity score in the decaying overlap setting. Doesn't apply if annotation score known.

Additional Assumption For Efficiency

Assumption (MSE consistency of the imputation function)

For function $\mu(\tilde{x})$, imputation function $\hat{\mu}(\tilde{x})$, and features $\tilde{X} \in \tilde{\mathcal{X}}$, we have that

$$E\left[\left(\hat{\mu}(\tilde{X}) - \mu(\tilde{X})\right)^2\right] = o(1).$$

- This is needed **only for efficiency**; it is **not necessary for unbiased inference.**
- Intuitively, we need the expected square error of our estimator to go to zero as the amount of data we train the estimator with goes to infinity.
- Mild in the context of deep neural networks. Recent theoretical work has shown that certain classes of deep neural networks learned with gradient descent are universally consistent (Drews and Kohler, 2024).

If the annotation score function is estimated

- We lose some robustness, and the researcher now needs to **assume that the convergence of the imputation and annotation score functions occurs at a sufficiently fast rate** (Chernozhukov et al., 2018; Kennedy, 2023; Wager, 2024), e.g., at order $n^{-1/4}$ rates.
- While classic semiparametric theory yields "curse of dimensionality" results for learning high-dimensional (nonparametric) conditional expectation functions, in practice, nonparametric regression using deep neural networks appears to exhibit fast rates of convergence (Klaassen et al., 2024)

Outline

Identification of a Mean with Missing Data

We follow Chen et al. (2008), which provides general results on semiparametric efficient estimation for parameters identified by moment conditions with missing data

The moment function for means is simply given by

$$m(W;\theta) = M^* - \theta.$$

It is straightforward to derive that:

$$\hat{\theta} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \left(\frac{A_i}{\pi(X_i)} [M_i - \hat{\mu}(\tilde{X}_i)] + \hat{\mu}(\tilde{X}_i) \right).$$

where $\mathcal{I}$ is the set of indices of the data allocated to the "estimation" partition and $\hat{\mu}$ is an approximation of μ learned independently of the estimation sample (fixed, or fit in the tuning split)

This is the **AIPW estimator** (M^* is just a potential outcome)

Debiased mean estimation

Suppose that $\pi(X_i) = |\mathcal{I} \cap \mathcal{J}|/|\mathcal{I}|$, where $\mathcal{J}$ is the set of indices corresponding to annotated data points. Then we see that

$$\hat{\theta} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}(\tilde{X}_i)) \right] \tag{AIPW}$$

$$= \underbrace{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{\mu}(\tilde{X}_i)}_{A} + \underbrace{\frac{1}{|\mathcal{I} \cap \mathcal{J}|} \sum_{i \in \mathcal{I} \cap \mathcal{J}} (M_i^* - \hat{\mu}(\tilde{X}_i))}_{B} \tag{PPI}$$

Term A is our imputation of $E[M_i^*]$ in the estimation sample, and term B is a bias correction term–an estimate of the measurement error of our imputation function in the annotated sample. This expression is reproduced in recent work on "prediction-powered inference" (Angelopoulos et al., 2023)

Debiased mean estimation

Moreover:

$$\hat{\theta} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}(\tilde{X}_i)) \right] \tag{AIPW}$$

$$= \underbrace{\frac{1}{|\mathcal{I} \cap \mathcal{J}|} \sum_{i \in \mathcal{I} \cap \mathcal{J}} M_i^*}_{C} + \underbrace{\left(\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{\mu}(\tilde{X}_i) - \frac{1}{|\mathcal{I} \cap \mathcal{J}|} \sum_{i \in \mathcal{I} \cap \mathcal{J}} \hat{\mu}(\tilde{X}_i) \right)}_{D}. \tag{FRA}$$

Term C is our estimate using only ground truth data. We then leverage the imputation function $\hat{\mu}$ as a form of nonparametric regression adjustment in term D, to adjust for systematic differences between our large unlabeled sample and our small annotated ground truth sample as summarized by $\hat{\mu}$. The expression (FRA) is reported in recent work on "flexible regression adjustment" (List et al., 2024).

Linear Regression With a Missing Outcome

The relevant moment condition is:

$$m(W;\theta) := R(M^* - R^{\mathrm{T}}\theta)$$

for some vector of real-valued regressors R. $\tilde{X} = (X, U, R)$.

We can derive MAR-S estimator:

$$\hat{\theta} = \left(\sum_{i\in\mathcal{I}} R_i R_i^{\mathrm{T}}\right)^{-1} \sum_{i\in\mathcal{I}} R_i \left(\frac{A_i}{\pi(X_i)}[M_i - \hat{\mu}(\tilde{X}_i)] + \hat{\mu}(\tilde{X}_i)\right).$$

This estimator resembles the standard OLS estimator, where the outcome has been replaced with what is sometimes called a "pseudo-outcome"

The Efficient Imputation Function

- Importantly, the imputation function $\hat{\mu}$ is a function of context specific variables, i.e., $\tilde{X} := (X, U, R)$.
- This is known as wide and deep learning.
- This, to our knowledge, has not been emphasized in the existing literature, which treats the imputation function as an arbitrary black box.

Linear regression with a missing regressor

Define $R^* := (M^*, R^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d_r+1}$, and $\theta = (\beta, \gamma^{\mathrm{T}})^{\mathrm{T}} \in \Theta \subseteq \mathbb{R}^{d_r+1}$. Let $\mu_2(\tilde{X}) := E[M^{*2} \mid \tilde{X}] = E[M^2 \mid A = 1, \tilde{X}] \in \mathbb{R}$. We can derive the debiased estimator:

$$\hat{\theta} = \left[\sum_{i \in \mathcal{I}} \left\{ \frac{A_i}{\pi(X_i)} \begin{pmatrix} M_i^{*2} - \hat{\mu}_2(\tilde{X}_i) & \left(M_i - \hat{\mu}(\tilde{X}_i)\right) R_i^{\mathrm{T}} \\ R_i \left(M_i - \hat{\mu}(\tilde{X}_i)\right) & 0 \end{pmatrix} + \begin{pmatrix} \hat{\mu}_2(\tilde{X}_i) & \hat{\mu}(\tilde{X}_i) R_i^{\mathrm{T}} \\ R_i \hat{\mu}(\tilde{X}_i) & R_i R_i^{\mathrm{T}} \end{pmatrix} \right\} \right]^{-1} \cdot \left[\sum_{i \in \mathcal{I}} \left\{ \frac{A_i}{\pi(X_i)} \begin{pmatrix} M_i - \hat{\mu}(\tilde{X}_i) \\ 0 \end{pmatrix} + \begin{pmatrix} \hat{\mu}(\tilde{X}_i) \\ R_i \end{pmatrix} \right\} Y_i \right].$$

Linear IV with Missing Outcome

Under a linear IV model, the moment function for θ is given by

$$m(W;\theta) = Z(M^* - R^{\mathrm{T}}\theta)$$

for vector of instruments Z and vector or regressors R. $\tilde{X} = (X, U, R, Z)$.

We can derive MAR-S estimator:

$$\hat{\theta} = \left(\sum_{i \in \mathcal{I}} Z_i R_i^{\mathrm{T}} \right)^{-1} \sum_{i \in \mathcal{I}} Z_i \left(\frac{A_i}{\pi(X_i)} \left[M_i - \hat{\mu}(\tilde{X}_i) \right] + \hat{\mu}(\tilde{X}_i) \right).$$

Linear IV with a Missing Instrument

Revisiting TSLS, we can assume the structural equation

$$Y = D\beta + C^{\mathrm{T}}\gamma + \varepsilon, \quad R = \begin{pmatrix} D \\ C \end{pmatrix}, \quad \theta = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

and instrument $Z^* = \binom{M^*}{C}$, where D is the treatment and β is the parameter of ultimate interest. Then the moment condition for θ is

$$m(W;\theta) = Z^*(Y - R^{\mathrm{T}}\theta).$$

We can derive the MAR-S estimator:

$$\hat{\theta} = \left[\sum_{i\in\mathcal{I}} \begin{pmatrix} \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i)) + \hat{\mu}(\tilde{X}_i) \\ C_i \end{pmatrix} R_i^{\mathrm{T}}\right]^{-1} \cdot \left[\sum_{i\in\mathcal{I}} \begin{pmatrix} \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i)) + \hat{\mu}(\tilde{X}_i) \\ C_i \end{pmatrix} Y_i\right].$$

Differences in Conditional Means

Differences in conditional means arise in nonparametric difference-in-differences (Callaway and Sant'Anna, 2021), local sharp regression discontinuity designs (Cattaneo and Titiunik, 2022; Cattaneo, Idrobo and Titiunik, 2024), and average treatment effects in RCTs.

Generically, we may represent the moment condition for a conditional mean via the function

$$m_{K=k}(W;\theta_{K=k}) = p_{K=k}^{-1}\mathbf{1}_{K=k}M^* - \theta_{K=k}$$

assuming $p_{K=k} := P(K=k)$ can be treated as known and is bounded away from zero and $\mathbf{1}_{K=k}$ is the binary indicator of $\{K=k\}$, with K having finite support.

Differences in Conditional Means

The MAR-S estimator is:

$$\hat{\theta}_{K=k} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \frac{\mathbf{1}_{K=k,i}}{p_{K=k}} \left(\frac{A_i}{\pi(X_i)} [M_i - \hat{\mu}_{K=k}(\tilde{X}_i)] + \hat{\mu}_{K=k}(\tilde{X}_i) \right).$$

If the researcher is interested in the parameter

$$\theta := \theta_{K=k} - \theta_{\tilde{K}=\tilde{k}}$$

We may use the fact that the EIF of the difference between two parameters is the difference of their EIFs and conclude that

$$\hat{\theta} = \hat{\theta}_{K=k} - \hat{\theta}_{\tilde{K}=\tilde{k}}.$$

Conditional Means (Missing Conditioning Variable)

In general, we may represent the moment condition for binary conditioning indicator $M^* \in \{0, 1\}$ and outcome $Y \in \mathbb{R}$, and $\theta = (\beta, \gamma)$ as

$$m(W; \theta) = \begin{pmatrix} \gamma^{-1} M^* Y - \beta \\ M^* - \gamma \end{pmatrix}.$$

Where

$$\gamma = E[M^*] = P(M^* = 1), \quad \beta = E[\gamma^{-1} M^* Y] = E[Y \mid M^* = 1],$$

We can derive:

$$\hat{\gamma} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \left[\frac{A_i}{\pi(X_i)} \left(M_i - \hat{\mu}(\tilde{X}_i) \right) + \hat{\mu}(\tilde{X}_i) \right],$$

$$\hat{\beta} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} \frac{\left[\frac{A_i}{\pi(X_i)} \left(M_i - \hat{\mu}(\tilde{X}_i) \right) + \hat{\mu}(\tilde{X}_i) \right]}{\hat{\gamma}} Y_i.$$

The General Case of Semiparametric Estimation (Chen et al., 2008)

- Suppose the parameter of interest $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ is identified by $E_P[m(W;\vartheta)] = 0 \quad \text{iff} \quad \vartheta = \theta$, where $W = (M^*, V)$ and $m(\cdot;\vartheta)$ is a vector of moments with dimension $d_m \geq d_\theta$. V is a set of context specific random variables that influence the estimation of our quantity of interest, such as controls in an ordinary least squares regression.

- Assume $[V \perp\!\!\!\perp A] \mid (X, U, M^*)$ so that, by Assumption 2, $[(U, M^*, V) \perp\!\!\!\perp A] \mid X$.
- Define $\tilde{X} := (X, U, V)$, where X are covariates observed by the researcher.

- Then under MAR, $[M^* \perp\!\!\!\perp A] \mid \tilde{X}$. which implies identification through $E_P\left[E_P[m((M, V);\vartheta) \mid A = 1, \tilde{X}]\right] = 0 \quad \text{iff} \quad \vartheta = \theta.$

- Define the conditional moment and Jacobian: $\mathcal{E}(\tilde{X};\theta) := E_P[m(W;\theta) \mid \tilde{X}], \quad \mathcal{J}_\theta := \frac{\partial}{\partial\theta} E_P[m(W;\theta)].$

Efficient Influence Function: Just-Identified

Suppose assumptions 1-3 hold, $d_m = d_\theta$, and $\mathcal{J}_\theta$ has full column rank
Then the efficient influence function is

$$\varphi(M, A, \tilde{X}; \theta) = -\mathcal{J}_\theta^{-1} \left(\frac{A}{\pi(X)} [m(W; \theta) - \mathcal{E}(\tilde{X}; \theta)] + \mathcal{E}(\tilde{X}; \theta) \right).$$

(Chen et al., 2008, Thm. 2).

The EIF itself defines a valid moment condition: $E_P[\varphi(M, A, \tilde{X}; \theta)] = 0$.

Under appropriate local product structure, this moment condition is Neyman orthogonal, yielding **double robustness.** We can replace $\mathcal{E}(\tilde{X}; \theta)$ with any parameter-dependent function $\eta(\tilde{X}; \theta)$ and still retain identification, although the efficient choice is $\eta(\tilde{X}; \theta) = \mathcal{E}(\tilde{X}; \theta)$.

Sample splitting allows η to be treated as fixed. Standard GMM asymptotics apply. Practical implication: **Learn $\mathcal{E}(\tilde{X}; \theta) = E_P[m(W; \theta) \mid \tilde{X}]$ as accurately as possible in the held-out sample.**

Outline

Issues that Arise in Large Datasets

The bedrock assumption. Debiasing frameworks assume ground-truth data are available for the **imputed variables** used in the estimating equation.

Why this often fails. Ground truth is available at the **granular** level of individual texts or images, whereas the variable of interest is a (potentially non-linear) function or functional of the granular missing data.

Inference with Aggregated and Transformed Missing Data

For the framework to be practically useful, it must **address this mismatch**.

Linear Models with MAR-S First-Step Measurement Error

Consider the linear model

$$Y = X^*\beta + \varepsilon,$$

In particular, we consider the case that X_i^* is not actually observed, but has been estimated with a MAR-S first-step, X_i.

$$X_{k,i} = X_{k,i}^* + \eta_{k,i}, \quad (\varepsilon_i, X_{k,i}^*) \perp\!\!\!\perp \eta_{k,i}, \quad \eta_{k,i} \overset{d}{\approx} N(0, \sigma_{\eta,k}^2),$$

$$\sigma_{\eta,k}^2 := |\mathcal{I}_k|^{-1}\mathrm{Var}(\varphi_k)$$

where φ_k is the efficient influence function associated with X_k.

Hence, we are in a **classical measurement error setting** and can proceed with the standard method-of-moments correction for classical errors-in-variables (Deaton, 1985; Fuller, 1987).

This setup nests the case where only some regressors are generated by a MAR-S first-step.

Measurement Error Structure

We therefore assume:

$$\eta_i \perp\!\!\!\perp Y_i, \quad \eta_i \overset{\text{iid}}{\sim} N(0, \Sigma)$$

where:

$$\Sigma := \text{diag}(\sigma^2_{\eta,1}, \ldots, \sigma^2_{\eta,K})$$

Reasonable given:

- Each MAR-S first-step regressor X_k is typically estimated on a separate sample.
- For large $\mathcal{I}_k$, the normal approximation for $\eta_{k,i}$ is accurate.

We consider Σ to be known, which is reasonable when $\mathcal{I}_k$ is sufficiently large.

Linear Models with MAR-S First-Step Measurement Error

The setup extends readily to:

- clustering;
- the panel data case (Deaton, 1985);
- various arm's-reach extensions: *e.g.*, heterogeneity in the variance of η_i across i, the outcome as a MAR-S first step, or non-normal measurement error distributions.

The MAR-S first step is **crucial** for making this measurement error model plausible.

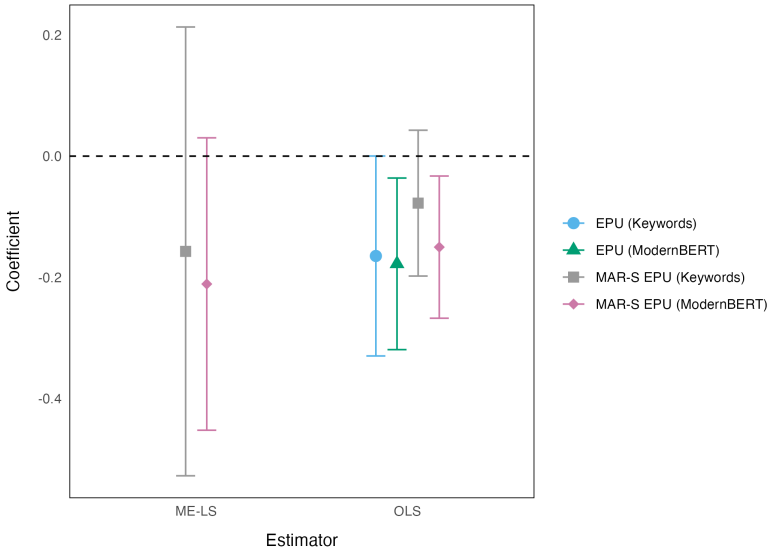
Motivating Example: Baker, Bloom, and Davis Firm-Level Regression

We reanalyze the following baseline regression from Table IV, column (5):

$$\Delta \text{Emp}_{it} = \beta \, \Delta \log \text{EPU}_t \times \text{intensity}_{it} + \gamma \, \Delta \frac{\text{Federal purchases}_t}{\text{GDP}_t} \times \text{intensity}_{it} + \alpha_i + \psi_t + u_{it}$$

- i: Firm index, t: Year index
- α_i, ψ_t: Firm and year fixed effects
- intensity_{it}: Firm-year policy exposure intensity
- ΔEmp_{it}: Employment growth
- β: Coefficient of interest

Results


0.2
0.0
-0.2
-0.4
Coefficient
ME-LS
OLS
Estimator
EPU (Keywords)
EPU (ModernBERT)
MAR-S EPU (Keywords)
MAR-S EPU (ModernBERT)

Outline

Keywords and the Annotation Function

- In large datasets, the structured data of interest may be a **rare event**.
- Social scientists often use **keyword-based filtering** to annotate text: texts with specific keywords get some probability of annotation, while all others are excluded.
- This **violates strong overlap** by assigning zero probability to some instances.

Intuitively, a language model's measurement error is plausibly **systematically correlated** with the terms that appear in the text.

Annotating Class-Imbalanced Data

The prescription. Dropping existing ML/stats work into the MAR-S framework shows we should **oversample instances that are harder to impute**.

The catch. This is infeasible directly, as it depends on knowing the ground truth, but the ML literature suggests proxies:

- model-based variance estimates;
- cross-validation residuals;
- ensemble disagreement.

How Many Labels Should Be Collected?

How many **effective observations** the validation and imputed data correspond to will depend on the **asymptotic correlation between the ground truth only estimator and the imputation only estimator**, *e.g.*, in mean estimation (Broska et al., 2025):

$$n_0 := |\mathcal{I} \cap \mathcal{J}| \times \frac{|\mathcal{I}|}{|\mathcal{I} \cap \mathcal{J}| + |\mathcal{I} \cap \mathcal{J}^c| \left(1 - \max\{\tilde{\rho}^2, 0\}\right)},$$

where $\mathcal{J}$ is the set of annotated indices, $\mathcal{J}^c$ is the set of unannotated indices and $\tilde{\rho}$ is the asymptotic correlation between the ground truth only estimator and the imputation only estimator $\tilde{\rho} := \text{Cor}(M^*, \mu(\tilde{X}))$.

The more predictive the imputation function, the greater the effective sample size, which would equal the full size of the dataset if the imputation function were perfect.

Varying the Size of Annotated Data

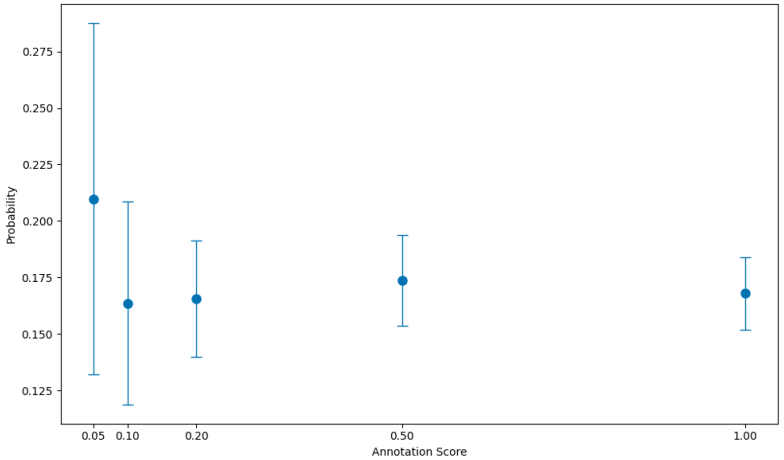


Figure: Estimating $P(\text{discussion of politics})$ with annotation score $\pi \in \left\{\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{2}, 1\right\}$.

Varying Imputation Accuracy

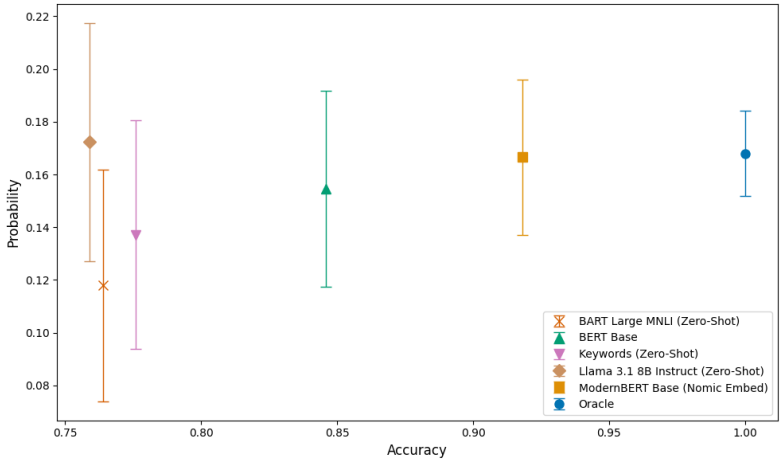


Figure: Estimating $P(\text{discussion of politics})$ across fine-tuned and zero-shot classifiers.

Outline

Conclusion

- Deep learning provides **powerful tools** for measurement.
- Accounting for imputation bias avoids many challenges that arise when imputed structured data are treated as **proxies**.
- MAR-S is accompanied by an **implementation package** at `https://github.com/jscarlson/mar-s`.
- MAR-S applies to many but by no means all common empirical scenarios. Notably, what if the **missing-at-random** assumption is impossible?

Bibliography

Baker, Scott R., Nicholas Bloom, and Steven J. Davis. 2016. "Measuring Economic Policy Uncertainty." *The Quarterly Journal of Economics*, 131(4): 1593–1636.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA:MIT Press.

Vafa, Keyon, Ashesh Rambachan, and Sendhil Mullainathan. 2024. "Do Large Language Models Perform the Way People Expect? Measuring the Human Generalization Function." Vol. 235 of *Proceedings of Machine Learning Research*, 48919–48937.