AI-Generated Measurements for Identification and Inference with Missing Data: A Weak Shadow Variable Approach

1. Research Summary

Across business and social science applications, outcomes are often missing in ways that depend on the unobserved outcomes themselves. Consider a service platform that evaluates customer satisfaction through post-interaction surveys in which customers may rate their experience on a 1–5 scale. Such surveys usually have high missingness rates, and this nonresponse is typically *missing not at random* (MNAR), meaning that customers with more extreme experiences are more likely to leave ratings (Hu et al., 2017). As a result, the observed ratings would not be representative of the full customer population, and standard statistical estimators based only on the observed data are typically biased.

At the same time, rich unstructured data, such as users' interaction histories, are available for all observations and can be transformed by large language models (LLMs) into low-dimensional evaluations of service quality. These measurements offer a new source of information for addressing missing-data problems. However, LLM evaluations may not faithfully reflect human perceptions of service, while the MNAR structure makes it impossible to estimate this bias.

In this paper, we propose a new perspective to use such low-dimensional measurements. In particular, we will view them as *weak shadow variables*, which are defined to be outcome-informative proxies are conditionally independent of whether the outcome is observed, given the true outcome and observed covariates. Intuitively, it is like a "shadow" of the true outcome, which does not require the measurement to be accurate prediction. In fact, the measurements can be arbitrarily biased or only weakly informative. Under this new structure, we characterize sharp bounds on population quantities such as the average outcome through a pair of linear programs (LPs), whose width reflects the informativeness of the weak shadow variable. For estimation and inference, we propose a localized penalized LP that remains feasible under sampling error and a subsampling procedure that constructs asymptotically valid confidence regions.

Our contributions therefore are threefold: **(1)** We propose a new assumption-light perspective for using AI-generated measurements in MNAR settings. **(2)** We formulate the corresponding identification intervals as LPs and devise a localized penalized estimator for finite-sample estimation and inference. **(3)** We validate the framework using real customer-service dialogue data, demonstrating an 89% reduction in interval width compared with the case without weak shadow variables.

More broadly, our framework propose a new role for AI-generated measurements that can play in empirical research. Rather than treating an LLM evaluation as a substitute for the missing outcome—an approach whose validity depends on predictive accuracy—we use it only to restrict how the missing outcomes can be distributed. This distinction allows researchers to extract useful information from AI outputs even when they are noisy, systematically biased, or too weak to support direct imputation. The same idea can apply whenever unstructured data are available for both respondents and nonrespondents, including employee surveys, patient-reported outcomes, and online reviews.

2. Key Methodology

We use customer-service ratings as our running example. Let $Y \in \{1, \dots, M\}$ denote the true satisfaction rating, $R \in \{0, 1\}$ the response indicator (R=1 if the rating is observed), and $F \in \{1, 2, \dots, B\}$ an AI-generated evaluation of service quality. We also allow for observable covariates $X \in \mathbb{R}^m$. For n individuals, the observed i.i.d. data are $\{(R_i, F_i, X_i, R_i Y_i)\}_{i=1}^n$. Our goal is to estimate the population mean $\theta = \mathbb{E}[Y]$, the average true rating.

Instead of imposing strong assumptions on either the missingness mechanism or the predictive accuracy of the LLM evaluation F, we impose only a single assumption on the relationship between F and R.

Assumption 1 (Weak Shadow Variable)**.** *Conditional on the true outcome and observed covariates, missingness is independent of the LLM prediction:* $F \perp\!\!\!\perp R \mid Y, X$.

The classical *shadow variable* framework additionally requires completeness—equivalently here, A below has full column rank (Miao et al., 2016). We drop this condition and refer to F as a *weak shadow variable*. For example, $F = f(Y, X) + \epsilon$ obeys Assumption 1 whenever $\epsilon \perp\!\!\!\perp R \mid Y, X$.

2.1. Partial Identification under Weak Shadow Variables

Assumption 1 does not guarantee point identification: even knowing the observed-data distribution exactly may not determine $\mathbb{E}[Y]$. We therefore consider every full-data distribution $\mathcal{P}$ of (R, F, X, Y) consistent with the observed distribution and collect its implied mean. The resulting sharp identification region is

$$I = \{\mathbb{E}_{\mathcal{P}}[Y] : \mathcal{P} \text{ is consistent with the observed data}\}.$$

For brevity, we omit X; with discrete X, we solve the LP by stratum and average its endpoints. Here, $y \in \{1, \ldots, M\}$ indexes outcomes, while $f \in \{1, \ldots, B\}$ indexes values of F. The scalar $w(y) = 1/\mathbb{P}(R = 1 \mid Y = y) - 1$ is the nonresponse-to-response odds for outcome y, and $\mathbf{w} = (w(1), \ldots, w(M))^\top$ collects these odds. The matrix $A = [\mathbb{P}(R{=}1, F{=}f, Y{=}y)]_{f,y} \in \mathbb{R}^{B \times M}$ records respondent probabilities by (f, y), and $\boldsymbol{\beta} = [\mathbb{P}(R{=}0, F{=}f)]_f$ records nonrespondent probabilities by f. Finally, $D = \mathrm{diag}(1, \ldots, M)$ is the diagonal matrix of outcome values, and $\mathbf{1}$ is the M-vector of ones.

Theorem 1 (Identification region)**.** *The sharp identification region I is an interval $[\theta_L, \theta_U]$ whose endpoints are given by the following linear program.*

$$\theta_L / \theta_U = \min_{\mathbf{w}} / \max_{\mathbf{w}} \ \mathbf{1}^\top A D(\mathbf{w} + \mathbf{1}) \quad \textit{s.t.} \quad A\mathbf{w} = \boldsymbol{\beta}, \ \ \mathbf{w} \geq \mathbf{0}, \tag{1}$$

The LP has a simple interpretation. A candidate $\mathbf{w}$ describes how the missing probability is distributed across outcomes, while $A\mathbf{w} = \boldsymbol{\beta}$ requires the implied probabilities to match the observed nonresponse separately for each value $F = f$. Thus, the role of the shadow variable F is to add B linear constraints, one for each possible value of F. By ruling out allocations that are inconsistent with the shadow-variable patterns, these constraints shrink the feasible region and tighten the identification interval. If A has full column rank, the allocation is unique and the interval collapses to a point, recovering classical point identification.

2.2. $\sqrt{n}$-Rate Estimation and Inference

In practice, A and $\boldsymbol{\beta}$ must be estimated from data. A natural plug-in approach replaces them with $\hat{A}_n$ and $\hat{\boldsymbol{\beta}}_n$ and solves the resulting sample LP. Although simple, this approach can create two problems. First, sampling error may make the estimated equality constraints mutually inconsistent, so the sample LP may have no feasible solution. Second, even when the sample LP is feasible, a small change in the estimated probabilities may change the binding constraints or the optimal solution, making the LP value nonsmooth and unstable. To address both problems, we use the following *localized penalized estimator* for a chosen penalty K:

$$B_K(\hat{A}_n, \hat{\boldsymbol{\beta}}_n) = \min_{0 \leq \mathbf{w} \leq K\mathbf{1}} \left\{ \mathbf{1}^\top \hat{A}_n D(\mathbf{w} + \mathbf{1}) + K \|\hat{A}_n \mathbf{w} - \hat{\boldsymbol{\beta}}_n\|_1 \right\}.$$

This estimator addresses the two problems directly. First, the penalty $K\|\hat{A}_n\mathbf{w} - \hat{\boldsymbol{\beta}}_n\|_1$ softens the estimated equality constraints, allowing small violations and keeping the problem feasible. Second, the box $[0, K]^M$ rules out solutions with very large weights and makes the objective value Lipschitz continuous. In the paper, we propose a data-driven procedure for choosing K, which result in a $\sqrt{n}$-consistent estimator.

Theorem 2 (Convergence rate, informal)**.** *Suppose $\hat{A}_n$ and $\hat{\boldsymbol{\beta}}_n$ are $\sqrt{n}$-consistent, the population LP has a finite optimum, and K is selected as in the paper. Then $B_{\widehat{K}_n}(\hat{A}_n, \hat{\boldsymbol{\beta}}_n) - \theta_L = O_p(n^{-1/2})$, i.e., the local penalized estimator preserves the usual $\sqrt{n}$ estimation rate.*

Because the estimator remains nonsmooth, the ordinary bootstrap and standard delta methods fail (Goff and Mbakop, 2024). We therefore use subsampling (Politis and Romano, 1994) that repeatedly recompute the estimator on $m = o(n)$ observations and calibrate it using centered subsample quantiles.

Theorem 3 (Confidence interval, informal)**.** *Under the stated regularity conditions and $m \to \infty$ with $m/n \to 0$, the subsampling confidence interval is asymptotically valid for the identification region. With $\sqrt{n}$-consistent $\hat{A}_n$ and $\hat{\boldsymbol{\beta}}_n$, its endpoints converge to the corresponding bounds at rate $O_p(n^{-1/2})$.*

3. Empirical Application: Customer Service Satisfaction

To evaluate the framework, we use the public USS dataset of $n = 3{,}300$ JD.com customer-service dialogues with manually annotated satisfaction ratings on a 1–5 scale (Sun et al., 2021). We generate semi-synthetic MNAR patterns using response probabilities predicted by four LLMs and construct weak shadow variables from 12 LLM evaluation methods covering the process, outcome, and overall service interaction.

We compare our method with four MNAR baselines: complete-case analysis (CCA, which uses only complete observations), naive imputation (NI, which uses the LLM evaluation as a replacement), the pattern-mixture model (Rubin, 1987), and the Heckman selection model (Heckman, 1979). We evaluate the MAE of the interval midpoint and the identification-interval width. Figure 1 shows that our estimator reduces MAE by approximately 40% and interval width by 90%. All four baselines also use the LLM evaluations as covariates; thus, the improvement comes from treating the evaluation as a weak shadow variable rather than from access to additional information. In this sense, AI does not replace human feedback; it makes the limited feedback that firms observe more informative.

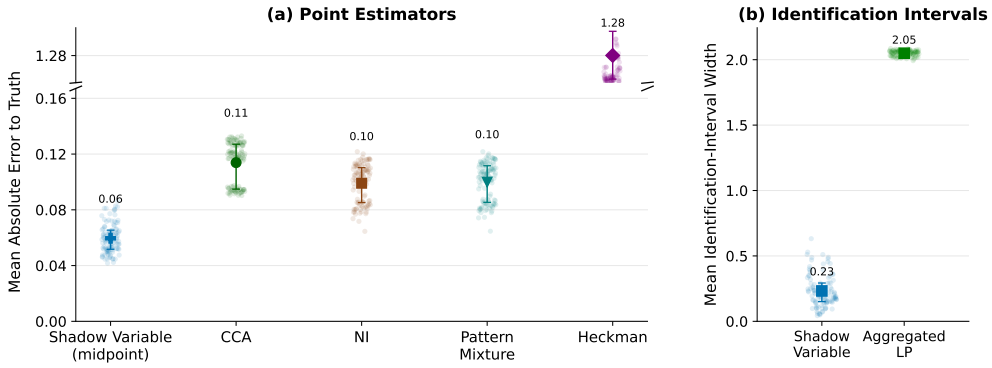

Figure 1: Estimator comparison in the semi-synthetic MNAR experiment.

4. References

L. Goff and E. Mbakop. Inference on the value of a linear program. Working paper, 2024.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

N. Hu, J. Zhang, and P. A. Pavlou. On self-selection biases in online product reviews. *MIS Quarterly*, 41(2):449–471, 2017.

W. Miao, P. Ding, and Z. Geng. Varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):453–468, 2016.

D. N. Politis and J. P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050, 1994.

D. B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987.

K. Sun et al. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proc. SIGDIAL*, 2021.