Empirical Bayes shrinkage (mostly) does not correct the measurement error in regression

Jiafeng Chen, Stanford University and SIEPR
Jiaying Gu, University of Toronto
Soonwoo Kwon, Brown University

Abstract. In the value-added literature, it is often claimed that regressing on empirical Bayes shrinkage estimates corrects for the measurement error problem in linear regression. We clarify the conditions needed; we argue that these conditions are stronger than the those needed for classical measurement error correction, which we advocate for instead. Moreover, we show that the classical estimator cannot be improved without stronger assumptions. We extend these results to regressions on nonlinear transformations of the latent attribute and find generically slow minimax estimation rates.

Version: March 24, 2025. We thank Isaiah Andrews, Xiaohong Chen, Patrick Kline, Chris Walters, and Merrill Warnick for helpful discussions. Juan Yamin Silva provided excellent research assistance.

1. **Introduction**

Heterogeneity of individuals is a vital element in many important areas of inquiry within economics. Empirical Bayes methods (Robbins, 1956) are applicable in many such settings to denoise individual fixed effects from noisy data. These methods are increasingly widely applied: Researchers use them to estimate individual effects of teachers (Kane, Rockoff and Staiger, 2008; Chetty, Friedman and Rockoff, 2014a; Gilraine, Gu and McMillan, 2020), mobility of geographies (Chetty and Hendren, 2018), value-added of hospitals (Chandra, Finkelstein, Sacarny and Syverson, 2016), skill of patent examiners (Feng and Jaravel, 2020), quality of managers (Fenizia, 2022), and individual income dynamics (Gu and Koenker, 2017), among others.

Researchers also often hope to quantify how unobserved individual attributes, like a teacher value-added, predicts downstream economic outcomes, like long-term student outcomes. In such setting, there seems to be a common understanding in the empirical literature that empirical Bayes shrinkage provides a correction for attenuation caused by statistical noise in a linear regression estimator (Jacob and Lefgren, 2005a; Kane and Staiger, 2008; Chetty *et al.*, 2014a; Angrist, Hull and Walters, 2023). For instance, Angrist *et al.* (2023) write "shrinkage corrects measurement error in models that treat school value-added as a regressor," and this intuition appears widespread. On the other hand, the classical literature in errors-in-variable regression offers simple corrections for measurement error that apply in these settings (Fuller, 1987), and it is unclear how including shrinkage estimates on the right-hand side of a regression compares to the classical approach.

This paper clarifies the conditions under which the regress-on-shrinkage estimator corrects for measurement error. Suppose the researcher would like to compute an *infeasible regression* of Y_i on μ_i, but only has access to noisy measurements (X_i, σ_i), where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Commonly, researchers fit the regression of Y_i on $\hat{\mu}_i(X_i, \sigma_i)$, where $\hat{\mu}_i(X_i, \sigma_i)$ is a linear shrinkage estimate. Importantly, an often-ignored condition for the consistency of this estimator is that the degree of shrinkage—implicitly a function of the noise level σ_i—does not correlate with the outcome Y_i once μ_i is controlled. Under such a condition, adding flexible functions in σ_i to the infeasible regression would not change the coefficient. This is a condition much like *precision independence* (Walters, 2024; Chen, 2025) and can be economically unrealistic. On the other hand, in this setting, because the σ_i^2 are observed, the attentuation bias in the regression of Y_i on X_i is estimable—and thus can be directly corrected. Such classical measurement error corrections does not impose any precision independence assumption. This makes the classical corrections more robust and—in our view—preferable.

Second, we show that without stronger assumptions, this classical estimator is effectively the *only* reasonable estimator for the regression coefficient in the infeasible regression, up to asymptotic equivalence; it is thus (vacuously) semiparametrically efficient. Thus, the

deficiencies of the shrinkage estimator are not due to insufficiently flexible empirical Bayes modeling. No amount of flexible modeling—despite tools proposed by, e.g., Chen (2025); Gilraine *et al.* (2020); Kwon (2023)—can improve on the classical measurement error correction.

Third, we extend our results to more complex settings, where the infeasible regression involves known nonlinear transforms $f(\mu_i)$ (e.g., indicators like $\mathbb{1}(\mu_i > \mu_0)$ may represent "high value-added" teachers). Under strong assumptions—that both the infeasible regression and parametric empirical Bayes models are correctly specified—a regression of Y_i on $\mathbb{E}[f(\mu_i) \mid X_i, \sigma_i]$ recovers the infeasible regression coefficients. However, we caution that these strong assumptions are difficult to relax without relinquishing appealing features of the resulting estimator, due to the fundamental statistical difficulty of deconvolution (Cai and Low, 2011). To that end, we show that if the infeasible regression coefficient is estimable at polynomial ($n^{-\alpha}$) rates of convergence uniformly across data-generating processes that impose few assumptions, then $f(\cdot)$ must necessarily be an analytic function—an extreme smoothness requirement. Put differently, if $f(\cdot)$ is not smooth, then there is some data-generating process under which one would need exponentially-in-k larger sample sizes to reduce uncertainty by a factor of k.

We compare these methods in simulation and an empirical application. The simulation results confirm that the classical estimator performs well, and substantially better than the regress-on-shrinkage estimator, across a wide range of data generating processes. Regress-on-shrinkage estimates can be severely biased even under settings that are calibrated to real data. Moreover, we confirm that regressions involving nonlinear transformations $f(\mu_i)$ are indeed difficult, with widely dispersed estimates even with a reasonably large sample size. In an empirical application that revisits Feng and Jaravel (2020), we find potentially substantive economic differences when using the classical estimator rather than the regress-on-shrinkage estimator. We also find supportive evidence that the conditions needed for the regress-on-shrinkage estimator to provide reliable estimates are violated.

This paper is related to the classical errors-in-variable regression literature (Fuller, 1987; Bickel and Ritov, 1987) as well as a recent literature on generated regressors. We highlight and compare a few. Rose, Schellenberg and Shem-Tov (2022) study multidimensional teacher value-added (e.g., value-added on criminal justice event, on math performance, etc.); they advocate for estimating the variance-covariance matrix of teacher value-added by correcting for measurement error, rather than by taking the variance-covariance matrix of empirical Bayes posterior means. We show that the same extends to regressions of downstream variables on value-added. Deeb (2021) studies the regress-on-shrinkage estimator and proposes corrections to its standard error that accounts for the uncertainty in estimating empirical Bayes hyperparameters. In a similar setting, Xie (2025) establishes conditions under

which regression-on-shrinkage estimators automatically yield valid inference without requiring additional adjustments. Battaglia, Christensen, Hansen and Sacher (2024) study the generated regressor problem with machine learning predictions for μ_i. The setting is related but distinct—they do not impose the Gaussian structure commonly imposed in the empirical Bayes literature. To our knowledge, our efficiency and minimax rate results have not appeared in the literature.

This paper proceeds as follows. Section 2 studies regression on latent μ_i. Section 3 presents our results for regression on $f(\mu_i)$ with nonlinear $f(\cdot)$. Sections 4 and 5 illustrate our results using a simulation and an empirical application.

2. Linear regression coefficients

Suppose we observe $(Y_i, X_i, \sigma_i)_{i=1}^n$ for a sample of individuals. Throughout, we will refer to these individuals as teachers, but the applications extend beyond teacher value-added. Here, Y_i is some outcome variable, X_i is a noisy measure of an unobserved teacher attribute μ_i, and σ_i is the observed standard error for X_i. For instance, Y_i would be some attribute of a teacher i (e.g. teacher-level mean of student outcomes), X_i would be the estimated teacher value-added of i, μ_i is the true teacher value-added for teacher i, and σ_i is the estimated standard error for X_i. For expositional simplicity, our main results shall restrict to setups of the data that aggregate to the teacher level. Section 2.3 discusses implementations of analogous approaches with disaggregated data.

Following the empirical Bayes literature, we assume $X_i \mid Y_i, \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ is Normally distributed and unrelated to Y_i, collected in the following assumption:

Assumption 2.1. *We assume* $(Y_i, X_i, \mu_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0$. *We impose the following assumptions on* P_0*:*

(1) $X_i \mid Y_i, \mu_i, \sigma_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
(2) $\sigma_\mu^2 \triangleq \text{Var}(\mu_i) > 0$ *and* $\mathbb{E}[\sigma_i^2] > 0$.
(3) $P_0(\sigma_i > 0) = 1$.

Assumption 2.1(2) and (3) are standard regularity assumptions. Assumption 2.1(1) is common in the empirical Bayes literature. The Normality of X_i is motivated by the fact that X_i is typically an estimate of μ_i with micro-data within a teacher: For instance, X_i may be a teacher-level mean of student scores, and the central limit theorem provides a Gaussian approximation, where σ_i^2 is the estimated standard error (see Walters, 2024).

Assumption 2.1(1) also implies that the outcome Y_i does not predict the noise component of X_i, i.e., $X_i \perp\!\!\!\perp Y_i \mid \mu_i, \sigma_i^2$. This assumption may be violated if, for instance, X_i is the sample mean of student test scores at the teacher level, but Y_i is the average downstream outcome of that same set of students. Meanwhile, if X_i and Y_i come from different sets of students (and we assume student outcomes are independent) or if Y_i is an underlying characteristic

of teacher i,[1] then this assumption is reasonable. Section 2.3 considers general cases where Y_i needs not be independent of the noise in X_i.

Empirical researchers are often interested in the population *infeasible* regression of Y_i on μ_i:

$$Y_i = \alpha_0 + \beta_0 \mu_i + \eta_i, \tag{1}$$

where the population regression coefficient $\beta_0 = \frac{\text{Cov}(Y_i, \mu_i)}{\text{Var}(\mu_i)}$ by definition. Here, we do not treat Equation (1) as a linear restriction on $\mathbb{E}[Y_i \mid \mu_i]$, but simply as a definition for β_0. That is, β_0 is the OLS regression coefficient one would have obtained had one access to the true unobserved attribute μ_i and infinitely many observations. Regressions of this form appear in, among others, Chandra *et al.* (2016); Chetty *et al.* (2014a); Jacob and Lefgren (2005b); Jackson (2018); Warnick, Light and Yim (2025); Mulhern (2023).[2]

For instance, when Y_i is the mean student long-term outcome for those taught by teacher i and μ_i is a teacher value-added for a teacher experienced by student i, then β_0 is the coefficient of the best linear prediction the long-term outcome from true teacher value-added. Under appropriate identifying assumptions such that Y_i is unbiased for the mean potential outcome of students assigned to teacher i, β_0 admits a causal interpretation as the best linear approximation to the conditional mean of teacher causal effects on true teacher value added.

Immediately, (1) implies that β_0 is identified by the following formula.

Proposition 2.1. *Under Assumption 2.1, β_0 is equal to the following function of the joint distribution of the observed data (Y_i, X_i, σ_i):*

$$\beta_0 = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i) - \mathbb{E}[\sigma_i^2]} = \underbrace{\frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)}}_{\text{Regression coefficient of } Y_i \text{ on } X_i} \underbrace{\frac{\text{Var}(X_i)}{\text{Var}(X_i) - \mathbb{E}[\sigma_i^2]}}_{\text{Inflation factor}}. \tag{2}$$

Proof. The proposition follows immediately from the observations that (a) $\text{Cov}(Y_i, X_i) = \text{Cov}(Y_i, \mu_i)$ and (b) $\text{Var}(\mu_i) = \text{Var}(X_i) - \mathbb{E}[\sigma_i^2]$. □

[1] For instance, Chandra *et al.* (2016) consider a regression of hospital i market size (Y_i) on hospital quality (μ_i), which are then estimated from clinical outcome of patients (X_i). Here, it is reasonable to assume that hospital size is independent from the noise component of patient outcomes $X_i - \mu_i$.

[2] We take as given that the infeasible regression coefficient (1) is the target parameter. It is possible, however, that empirical researchers may prefer other estimands. For instance, since decisions are only functions of (X_i, σ_i), we might be interested instead in the statistical relationship between Y_i and functions of (X_i, σ_i). For instance, we might form a prediction $\hat{\mu}_i(X_i, \sigma_i)$ and we might want to assess how predictive *the predicted teacher value-added* is for Y_i, instead of how predictive *the true* μ_i is for Y_i. The former is more relevant, say, if we are more interested in assessing the quality of feasible predictions for teacher effects, rather than the inherent relationship between true teacher effects and outcomes.

Equation (2) suggests a simple analogue estimator for β_0 that replaces population covariances, variances, and expectations with their sample counterparts:[3]

$$\hat{\beta} = \frac{\mathrm{Cov}_n(Y_i, X_i)}{\mathrm{Var}_n(X_i) - \mathbb{E}_n[\sigma_i^2]}. \tag{3}$$

Under standard conditions, $\hat{\beta}$ is consistent and asymptotically Normal, whose asymptotic distribution can be consistently estimated by a nonparametric bootstrap given (Y_i, X_i, σ_i). This estimator is classical from the errors-in-variable regression literature (e.g., Section 3.1 of Fuller, 1987, (23) in Deaton, 1985) and more recently advocated by de Chaisemartin and Deeb (2024). Moreover, the consistency and Normality of $\hat{\beta}$ does not require X_i to be Normally distributed.

Our first main point is to advocate for this classical estimator $\hat{\beta}$ as opposed to common estimation approaches for (1). A popular—and standard—alternative estimator of β_0 is the regression coefficient of Y_i on estimated empirical Bayes posterior means $\hat{\mu}_i(X_i, \sigma_i)$, following a parametric empirical Bayes procedure:

$$\tilde{\beta} \triangleq \frac{\mathrm{Cov}_n(Y_i, \hat{\mu}_i)}{\mathrm{Var}_n(\hat{\mu}_i)} \tag{4}$$

$$\hat{\mu}_i(X_i, \sigma_i) \triangleq \frac{\sigma_i^2}{\sigma_i^2 + \hat{\sigma}_\mu^2}\hat{\mu} + \frac{\hat{\sigma}_\mu^2}{\sigma_i^2 + \hat{\sigma}_\mu^2}X_i \tag{5}$$

$$\text{where } \hat{\mu} \triangleq \mathbb{E}_n[Y_i] \text{ and } \hat{\sigma}_\mu^2 \triangleq \mathrm{Var}_n(Y_i) - \mathbb{E}_n[\sigma_i^2].$$

This approach is followed in, e.g., Jacob and Lefgren (2005b); Kane and Staiger (2008); Warnick *et al.* (2025); Jackson (2018); Bau and Das (2020); Angelova, Dobbie and Yang (2023). The estimator $\tilde{\beta}$ is widely thought to be consistent for β_0, which we argue rests strong assumptions, often implicitly imposed.[4] We also note that when $\sigma_i^2 = \sigma^2$ are constant for all i, then $\tilde{\beta} = \hat{\beta}$, but they are no longer equal in heteroskedastic settings.

To this end, we show that, first, the estimator $\tilde{\beta}$ is consistent for β_0 only under much stronger conditions than Assumption 2.1, rendering it less robust and less preferable to $\hat{\beta}$. Second, we show that it is impossible to improve upon $\hat{\beta}$ with any other procedure, including more flexible empirical Bayes procedures (Kwon, 2023; Gilraine *et al.*, 2020; Chen, 2025), at least without imposing stronger assumptions. Under Assumption 2.1, any consistent and

[3]We shorthand $\mathbb{E}_n[W_i] = \frac{1}{n}\sum_{i=1}^n W_i$, $\mathrm{Var}_n(W_i) = \mathbb{E}_n[W_i^2] - (\mathbb{E}_n[W_i])^2$, and $\mathrm{Cov}_n(W_i, Z_i) = \mathbb{E}_n[W_i Z_i] - \mathbb{E}_n[W_i]\mathbb{E}_n[Z_i]$.

[4]Jacob and Lefgren (2005b) (Appendix C) states that "one can easily show that using the EB estimates as an explanatory variable in a regression context will yield point estimates that are unaffected by the attenuation bias that would exist if one used simple OLS estimates."
Angrist *et al.* (2023) write that "shrinkage corrects measurement error in models that treat school value-added as a regressor. Putting the unbiased but noisy estimate [in our notation, X_i] on the right-hand side of a regression results in attenuation bias toward zero due to classical measurement error; the posterior mean introduces non-classical measurement error that corrects this so that a regression with $[\mu_i(X_i, \sigma_i)]$ on the right yields the same coefficient as using the true $[\mu_i]$."

asymptotically Normal estimator of β_0 is in fact asymptotically equivalent to $\hat{\beta}$. Thus, any empirical Bayes procedure can at most match the performance of $\hat{\beta}$.

2.1. **When is $\tilde{\beta}$ consistent?** Since $\mu \triangleq \mathbb{E}[\mu], \sigma_\mu^2 \triangleq \mathrm{Var}(\mu)$ are consistently estimable, let us restrict attention to an oracle counterpart of $\tilde{\beta}$ where μ, σ_μ^2 are known, namely

$$\tilde{\beta}^* \triangleq \frac{\mathrm{Cov}_n(Y_i, \mu_i^*)}{\mathrm{Var}_n(\mu_i^*)}, \text{ with population limit } \tilde{\beta}_0 \triangleq \frac{\mathrm{Cov}_{P_0}(Y_i, \mu_i^*)}{\mathrm{Var}_{P_0}(\mu_i^*)}.$$

We first formalize the folklore in the literature on the consistency of $\tilde{\beta}$.

Assumption 2.2. *For $\mu_i^* = \mu_i^*(X_i, \sigma_i)$ the oracle counterpart to (5), the distribution P_0 satisfies*

(1) (Forecast unbiasedness) $\mathrm{Cov}(\mu_i^*, \mu_i) = \mathrm{Var}(\mu_i)$

(2) (Exogeneity) $\mathbb{E}[\eta_i \mu_i^*] = 0$.

Proposition 2.2. *Under Assumptions 2.1 and 2.2, $\tilde{\beta}_0 = \beta_0$.*

Assumption 2.2(1) is often referred to as forecast unbiasedness (Chetty *et al.*, 2014a; Chetty, Friedman and Rockoff, 2014b; Stigler, 1990). It assumes that the empirical Bayes posterior means are reasonable predictors of the true unobserved attribute, in the sense that a hypothetical regression of the unobserved attribute μ_i on its predicted value μ_i^* returns a regression coefficient of 1.[5]

Assumption 2.2(2) is a key assumption that seems to be implicitly taken for granted in the literature. Importantly, the condition $\mathbb{E}[\eta_i X_i] = 0$—which we impose by definition of β_0 as the population projection coefficient—does not on its own justify Assumption 2.2(2).[6] The reason is that μ_i^* is a function of both X_i and σ_i, and we have made no assumptions on how σ_i interacts with η_i.

Under Assumptions 2.1 and 2.2, we review and formalize two arguments in the literature for Proposition 2.2, and point out where each of the assumptions in Assumption 2.2 is used. In particular, we have not found any work in the literature that explicitly mentions Assumption 2.2(2) or sufficient conditions for it, even though—as we shall see—Assumption 2.2(2) is crucial for the consistency of $\tilde{\beta}$.

First, many papers justify $\tilde{\beta}_0 = \beta_0$ using an argument from Jacob and Lefgren (2005b).

[5]When the empirical Bayes prior is well-specified, i.e. $\mu_i^* = \mathbb{E}[\mu_i \mid X_i, \sigma_i]$, we have that $\mathbb{E}[\mu_i \mid \mu_i^*] = \mu_i^*$ by the law of iterated expectations, and Assumption 2.2(1) is satisfied. For Gaussian-prior empirical Bayes, even if the empirical Bayes model is misspecified, Assumption 2.2(1) would be true if we only assume (6), meaning that σ_i does not predict the first two moments of μ_i.

[6]Under homoskedasticity, i.e. $\sigma_i = \sigma$ for all i, μ_i^* is simply a linear function of X_i, and $\mathbb{E}[\mu_i^* \eta_i] = 0$ holds by construction; however, the same is not true when σ_i are heterogeneous and can be correlated with Y_i.

Proof. (Jacob and Lefgren (2005b)'s argument for Proposition 2.2) Observe that for $v_i = \mu_i - \mu_i^*$, under Assumption 2.2(1),

$$\mu_i = \mu_i^* + v_i \quad \mathbb{E}_{P_0}[v_i \mu_i^*] = 0.$$

As a result, we can rewrite (1) as $Y_i = \alpha_0 + \beta_0 \mu_i^* + \beta_0 v_i + \eta_i$, $\mathbb{E}_{P_0}[\eta_i X_i] = 0$. By Assumption 2.2(2),

$$\mathbb{E}[(\beta_0 v_i + \eta_i)\mu_i^*] = \beta_0 \mathbb{E}[v_i \mu_i^*] + \mathbb{E}[\eta_i \mu_i^*] = \mathbb{E}[\eta_i \mu_i^*] = 0\,.$$

As a result, regressing Y_i on μ_i^* recovers the coefficient β_0. □

A second popular intuition justifies $\tilde{\beta}_0 = \beta_0$ by appealing to instrumental variables (Chetty *et al.*, 2014b, pp. 2639).

Proof. (IV argument for Proposition 2.2) We can write (Y_i, μ_i, μ_i^*) as a two-stage least-squares specification, where μ_i is an "endogenous treatment" and μ_i^* is an "exogenous instrument":

$$Y_i = \alpha_0 + \beta_0 \mu_i + \eta_i$$
$$\mu_i = \gamma_0 + \pi_0 \mu_i^* + v_i.$$

Assumption 2.2(1) implies that the population first-stage coefficient is one: $\pi_0 = 1$. Crucially, Assumption 2.2(2) is exactly the exogeneity and exclusion assumption, implying that μ_i^* is a valid instrument. Therefore, β_0 is equal to the population two-stage least-squares coefficient, which is further equal to the reduced-form coefficient $\tilde{\beta}_0$ since the first-stage coefficient $\pi_0 = 1$:

$$\beta_0 = \frac{\operatorname{Cov}(Y_i, \mu_i^*)}{\operatorname{Cov}(\mu_i^*, \mu_i)} = \frac{\overbrace{\operatorname{Cov}(Y_i, \mu_i^*)/\operatorname{Var}(\mu_i^*)}^{\text{Reduced-form coef. of } Y_i \text{ on } \mu_i^*}}{\underbrace{\operatorname{Cov}(\mu_i^*, \mu_i)/\operatorname{Var}(\mu_i^*)}_{\text{First-stage coef. of } \mu_i \text{ on } \mu_i^*,\ \pi_0 = 1}} = \frac{\operatorname{Cov}(\mu_i^*, \mu_i)}{\operatorname{Var}(\mu_i^*)} = \tilde{\beta}_0.$$

□

Assumption 2.2 is a set of high-level assumptions imposed to justify $\tilde{\beta}_0 = \beta_0$. The following assumptions are natural sufficient conditions for Assumption 2.2. Though they are stronger than Assumption 2.2, scenarios that satisfy Assumption 2.2 but violate Assumption 2.3 are economically knife-edge. Moreover, all assumptions in Assumption 2.3 are testable.

Assumption 2.3 (Sufficient conditions for Assumption 2.2)**.** *(1) (Precision independence)* $\sigma_i \perp\!\!\!\perp (\mu_i, Y_i)$ *under* P_0

(2) (Precision independence in first two moments) The distribution P_0 *satisfies*

$$\mathbb{E}_{P_0}[\mu_i \mid \sigma_i] = \mu \text{ and } \operatorname{Var}_{P_0}[\mu_i \mid \sigma_i] = \sigma_\mu^2, \tag{6}$$

(3) (Linearity of the conditional expectation function and exogeneity of σ*)* $\mathbb{E}_{P_0}[Y_i \mid \mu_i, \sigma_i] = \alpha_0 + \beta_0 \mu_i$.

Lemma 2.3. *Under Assumption 2.1, Assumption 2.3(1) implies Assumption 2.2; Assumption 2.3(2) implies Assumption 2.2(1), and Assumption 2.3(3) implies Assumption 2.3(2).*

A strong sufficient condition for Assumption 2.2 is directly that σ_i is completely independent of the joint distribution of (Y_i, μ_i).

Assumption 2.3(2)–(3) are in some respects weaker. Assumption 2.3(2) imposes that σ_i does not predict μ_i at least in its first two conditional moments. This is a moment version of the precision independence condition discussed in Chen (2025), who show examples in which Assumption 2.3(2) fails to hold. This precision independence assumption alone does not involve Y_i, and is thus insufficient to justify $\tilde{\beta}_0 = \beta_0$.

Assumption 2.3(3) imposes two strong assumptions on P_0. The first is that the conditional expectation function of Y_i given μ_i is in fact linear in μ_i, viewing β_0 instead as the slope of that linear relationship rather than the population best linear approximation of $\mathbb{E}[Y_i \mid \mu_i]$. Second, Assumption 2.3 imposes that σ_i does not have additional predictive power over Y_i given μ_i. This is the analogue of a precision independence assumption for the outcome variable Y_i, and can fail due to similar reasons outlined in Chen (2025). This assumption is rejected in our empirical application.

Since $\hat{\beta}$ is consistent for β_0 without imposing Assumption 2.2 or Assumption 2.3 at all, it should be preferred over the estimator $\tilde{\beta}$ on robustness grounds. Nevertheless, motivated by Assumption 2.3(2), we might wonder whether the defects of $\tilde{\beta}$ is due to insufficiently flexible empirical Bayes modeling. Perhaps the Normality assumption in Assumption 2.1 can be exploited to yield efficiency benefits. The next subsection answers this question in the negative. Specifically, we show that under Assumption 2.1, not only is $\hat{\beta}$ a semiparametrically efficient estimator for β_0, but all well-behaved consistent estimators for β_0 must be asymptotically equivalent to $\hat{\beta}$. This provides strong justification for $\hat{\beta}$ as the *only* estimator for β_0 asymptotically.

2.2. **Efficiency and uniqueness of $\hat{\beta}$.** Let $\mathcal{P}_0$ be the set of distributions P_0 that satisfy Assumption 2.1 as well as some technical conditions, stated in Appendix B. Let $\mathcal{P}$ be the set of distributions on the observed data (Y_i, X_i, σ_i) that is induced by members of $\mathcal{P}_0$.

Following standard semiparametric theory (van der Vaart, 2000; Tsiatis, 2006), we restrict our attention to *regular and asymptotically linear* (RAL) estimators, defined formally in Appendix B. RAL estimators are asymptotically Normal along all local perturbations of P_0 within $\mathcal{P}_0$. Indeed, most $\sqrt{n}$-consistent and asymptotically Normal estimators are RAL, and *semiparametrically efficient* estimators are exactly the optimal estimators *among the class* of RAL estimators.

We also recall Definition 2.2 in Chen and Santos (2018) on local just identification, again formally stated in Appendix B. Local just identification is the semiparametric analogue to just identification in parametric GMM models. In a just identified GMM model, there are

no additional moment conditions to exploit, and all GMM weightings yield the same estimator. Exactly analogously, in just identified semiparametric models, there are no (local) testable restrictions to exploit. Just like all weighted GMM estimators are exactly the same under just identification, under local just identification, all RAL estimators are asymptotically equivalent. They have the same asymptotic distribution, and they are all (vacuously) semiparametrically efficient.

The next theorem shows that many members in $\mathcal{P}$ are locally just identified by $\mathcal{P}$. The results of Chen and Santos (2018) then imply that all RAL estimators for the regression coefficient β_0 are asymptotically equivalent to each other and to $\hat{\beta}$, which we also verify in the following theorem. This also implies that $\hat{\beta}$ is semiparametrically efficient, though in a vacuous sense.

Theorem 2.4. *Let $P \in \mathcal{P}$ and let $P_0 \in \mathcal{P}_0$ be the corresponding distribution over the complete data, defined formally in Appendix B. Then P is locally just identified by $\mathcal{P}$. As a result,*

(1) Any RAL estimator $\check{\beta}$ for

$$\beta_0 \triangleq \beta_0(P) \triangleq \frac{\mathrm{Cov}_P(X, Y)}{\mathrm{Var}_P(X) - \mathbb{E}_P[\sigma^2]}$$

is asymptotically equivalent to the analogue estimator $\hat{\beta}$: $\sqrt{n}(\check{\beta} - \hat{\beta}) = o_P(1)$.

(2) The semiparametric efficiency bound for $\beta_0(P)$ is equal to the asymptotic variance of $\hat{\beta}$ at P.

Theorem 2.4, an application of Example 25.35 in van der Vaart (2000), rules out efficiency gains from alternative estimators under the minimal assumptions Assumption 2.1. As a result, $\hat{\beta}$ is strongly justified for estimating β, since it is in fact the unique estimator for β in an asymptotic sense. The implications of Theorem 2.4 are not limited to the regression coefficient β_0; indeed, any regular parameter of P admits unique RAL estimators in the sense of Theorem 2.4(1).

Imposing stronger assumptions than Assumption 2.1 amounts to shrinking the model $\mathcal{P}$. Doing so would in general result in testable overidentification restrictions and weakly decrease the efficiency bound. Under Assumption 2.3(1) or similar assumptions, for instance, the estimator $\tilde{\beta}$ is indeed more efficient than $\hat{\beta}$ (Sullivan, 2001), though yet more efficient estimators exist (Bickel and Ritov, 1987). As discussed, Assumption 2.3 is much stronger than Assumption 2.1.

One assumption that is arguably reasonable is the conditional Normality structure of Y_i, given that i is a teacher and sometimes Y_i is a sample mean of student outcomes at the teacher level. In particular, we might consider imposing $Y_i \mid \theta_i, X_i, \mu_i, \sigma_i, \nu_i \sim \mathcal{N}(\theta_i, \nu_i^2)$ for some unknown θ_i and known ν_i^2, and assume that $(Y_i, \theta_i, \nu_i, X_i, \mu_i, \sigma_i) \overset{\text{i.i.d.}}{\sim} P_0^*$ under the additional

conditional Normality assumption. Simple modification to the proof of Theorem 2.4 shows that such a restriction does not alter the conclusion of Theorem 2.4.

2.3. **Implementation.** We close this section with a discussion of implementing $\hat{\beta}$ in richer environments. First, given teacher-level data $(Y_i, X_i, Z_i, \sigma_i)$, we may want to include covariates Z_i in the infeasible regression:

$$Y_i = \alpha_0 + \beta_0 \mu_i + \gamma_0' Z_i + \epsilon_i.$$

Suppose Z_i does not predict the noise component of X_i, then by Frisch–Waugh–Lovell,

$$Y_i = \beta_0(\mu_i - \text{proj}(\mu_i \mid Z_i)) + \epsilon_i$$

and $X_i - \text{proj}(X_i \mid Z_i) \sim \mathcal{N}(\mu_i - \text{proj}(\mu_i \mid Z_i), \sigma_i^2)$, where $\text{proj}(\cdot \mid Z_i)$ is the population linear projection of a random variable onto Z_i. Thus, our analysis above applies to the partialled out teacher effects $\mu_i - \text{proj}(\mu_i \mid Z_i)$ and its measurement $X_i - \text{proj}(X_i \mid Z_i)$.

More generally, it is common practice to consider a student-level regression. For j a student associated with teacher i, we consider the infeasible regression

$$Y_{ij} = \alpha_0 + \beta_0 \mu_j + \gamma_0' Z_{ij} + \epsilon_{ij} \tag{7}$$

where X_{ij} is unbiased for μ_j.[7] Here, we do not assume X_{ij}, Y_{ij}, Z_{ij} are uncorrelated within a teacher i, and so Y, Z may predict the noise component of X. This infeasible regression coefficient $\theta_0 = (\alpha_0, \beta_0, \gamma_0')'$ can be written as

$$\theta_0 = \mathbb{E}_{P_0}\left[\sum_{j=1}^{N_i} W_{ij} W_{ij}'\right]^{-1} \mathbb{E}_{P_0}\left[\sum_{j=1}^{N_i} W_{ij} Y_{ij}\right].$$

The quantities $\mathbb{E}_{P_0}\left[\sum_{j=1}^{N_i} W_{ij} W_{ij}'\right], \mathbb{E}_{P_0}\left[\sum_{j=1}^{N_i} W_{ij} Y_{ij}\right]$ involves infeasible terms such as $\mathbb{E}\sum_j \mu_i^2$, $\mathbb{E}\sum_j \mu_i Z_{ij}$, $\mathbb{E}\sum_j \mu_i Y_{ij}$. Fortunately, they can be similarly estimated from X_{ij}: For instance,

[7] One plausible and precise sampling process is as follows. Suppose

$$(N_i, \mu_i, (Y_{ij}, Z_{ij}, X_{ij})_{j=1}^{N_i}) \overset{\text{i.i.d.}}{\sim} P_0$$

and P_0 is such that $\mathbb{E}_{P_0}[X_{ij} \mid \mu_i, N_i] = \mu_i$. We also assume that conditional on μ_i, N_i, the student-level variables $(Y_{ij}, Z_{ij}, X_{ij})_{j=1}^{N_i}$ are uncorrelated across j.
To connect the aggregate and disaggregated setups, suppose $Z_{ij} = Z_i$ is not a function of the student. Then the population OLS regression (7) is equivalent to the teacher-level weighted least squares regression

$$Y_i = \alpha_0 + \beta_0 \mu_j + \gamma_0' Z_i + \epsilon_{ij}$$

where $Y_i = \frac{1}{N_i}\sum_j Y_{ij}$ and observations are weighted by N_i.

for $X_i = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$,

$$\mathbb{E}\left[\sum_{j=1}^{N_i} \mu_i Z_{ij}\right] = \mathbb{E}\left[\sum_{j=1}^{N_i} X_i Z_{ij}\right] - \mathbb{E}\left[\underbrace{\frac{1}{N_i - 1}\sum_{j=1}^{N_i} Z_{ij}(X_{ij} - X_i)}_{\text{Empirical covariance between } X_{ij} \text{ and } Z_{ij}}\right].$$

Thus, we may use $\sum_{j=1}^{N_i} X_i Z_{ij} - \frac{1}{N_i - 1}\sum_{j=1}^{N_i} Z_{ij}(X_{ij} - X_i)$ to substitute for $\sum_j \mu_i Z_{ij}$, and similarly for other terms in θ_0. This construction is very similar to $\hat{\beta}$, since it essentially debiases an empirical moment with X_i by adjusting for the impact of second moments. In a slightly different context, recent work by de Chaisemartin and Deeb (2024) proposes this estimator as well.

Even more conveniently, it turns out that such an analogue can be simply implemented by instrumenting for X_{ij} in following regression with a *leave-one-out instrument* $X_{i,-j} = \frac{1}{N_i - 1}\sum_{k \neq j} X_{ik}$

$$Y_{ij} = \alpha_0 + \delta_0 X_{ij} + \gamma_0' Z_{ij} + \epsilon_{ij}.$$

Devereux (2007) shows[8] that these two estimates are in fact numerically equivalent. This equivalence means that researchers can conveniently implement the measurement error correction by constructing the leave-one-out instrument and use off-the-shelf routines, without other multistep procedures.

3. Regression coefficients involving nonlinear functions of μ_i

Sometimes empirical researchers are interested in the projection coefficient on some known nonlinear function $f(\cdot)$ of the latent quantities μ_i:

$$Y_i = \rho_0 + \tau_0 f(\mu_i) + \eta_i.$$

For instance, we might be interested in how being a "good teacher" predicts outcomes, and being a good teacher is defined as $\mathbb{1}(\mu_i > \mu_0)$ for some threshold μ_0. For an example of setting where β_0 is of interest under such specification, see Section 3.2.2 of Bruhn, Imberman and Winters (2022).

Unfortunately, estimating τ_0 involves some unpleasanat tradeoffs for the analyst, as we no longer have access to a simple estimator like $\hat{\beta}$. On the one hand, imposing a strong assumption does allow us to recover τ_0 by regressing Y_i on the *correctly specified* parametric empirical Bayes posterior means for $f(\mu_i)$. The downside of this approach is that the assumptions involved can be quite strong, and are unlikely to be justified. On the other hand, without these assumptions, estimating τ_0 is fundamentally difficult. This difficulty is in the

[8]We thank Patrick Kline for this reference.

sense that the the error of the best possible estimator contracts at subpolynomial rates in the sample size, meaning that reducing uncertainty requires exponentially large sample sizes.

We begin with the first approach and document a simple estimator under strong assumptions on model specification.

Assumption 3.1 (Correct specification of outcome model)**.** $\mathbb{E}[Y_i \mid \mu_i, \sigma_i] = \rho_0 + \tau_0 f(\mu_i)$. *Equivalently,* $\mathbb{E}[\eta_i \mid \mu_i, \sigma_i] = 0$.

Proposition 3.1. *Under Assumptions 2.1 and 3.1, the population coefficient τ_0 is equal to the regression coefficient of Y_i on the* correctly specified empirical Bayes posterior mean *of $f(\mu_i)$ on X_i, σ_i*

$$\tau_0 = \frac{\mathrm{Cov}(Y_i, \mathbb{E}[f(\mu_i) \mid X_i, \sigma_i])}{\mathrm{Var}(\mathbb{E}[f(\mu_i) \mid X_i, \sigma_i])}. \tag{8}$$

Proof. We can write

$$Y_i = \rho_0 + \tau_0 \mathbb{E}[f(\mu_i) \mid X_i, \sigma_i] + \tau_0(f(\mu_i) - \mathbb{E}[f(\mu_i) \mid X_i, \sigma_i]) + \eta_i.$$

Under Assumption 3.1, $\mathbb{E}[\eta_i \mid X_i, \sigma_i] = 0$. By law of iterated expectations,

$$\mathbb{E}[\tau_0(f(\mu_i) - \mathbb{E}[f(\mu_i) \mid X_i, \sigma_i]) \mid X_i, \sigma_i] = 0.$$

Therefore, τ_0 is equal to the population regression coefficient of Y_i on $\mathbb{E}_{G_i}[f(\mu_i) \mid X_i, \sigma_i]$. □

Operationizing (8) requires us to estimate $\mathbb{E}[f(\mu_i) \mid X_i, \sigma_i]$ (importantly, distinct from $f(\mathbb{E}[\mu_i \mid X_i, \sigma_i])$). Traditional empirical Bayes methods often specify a parametric model, e.g., $\mu_i \mid \sigma_i \sim \mathcal{N}(\mu, \sigma_\mu^2)$. Regressing Y_i on the estimated empirical Bayes posterior means under such models can be interpreted as estimating (8). Nevertheless, the price of such an interpretation is the strong assumptions imposed: Assumption 3.1 and well-specified parametric prior.

Without these strong assumptions, τ_0 remains identified, as it is a function of the joint distribution of (Y_i, μ_i). This joint distribution is identified from the joint distribution of (Y_i, X_i, σ_i) via deconvolution. However, as is typical with deconvolution, the problem of estimating τ_0 becomes prohibitively difficult without making parametric restrictions (Fan and Truong, 1993). We illustrate this point with a minimax lower bound for τ_0.

Let us first define the class of distributions the lower bound is over. Fix some function $f : [-1, 1] \to \mathbb{R}$ such that it is bounded ($\|f\|_\infty < \infty$) and nonconstant in the sense that $V(f) \triangleq \mathrm{Var}_{U \sim \mathrm{Unif}[-1,1]}(f(U)) > 0$. Let $\mathcal{Q}_0$ collect all distributions Q_0 for $(Y_i, \mu_i, X_i, \sigma_i)$ satisfying Assumption 2.1 where (a) μ_i and Y_i are supported within the interval $[-1, 1]$, (b) $\mathrm{Var}_{Q_0}(f(\mu_i)) > \frac{1}{2}V(f)$ to avoid degeneracy, and (c) $\sigma_i \leq 1$.[9] Let $\mathcal{Q} = \left\{P^{\mathrm{obs}}(P_0) : P_0 \in \mathcal{Q}_0\right\}$. Relative to f, let $\tau_0(Q_0) = \frac{\mathrm{Cov}_{Q_0}(Y, f(\mu))}{\mathrm{Var}_{Q_0}(f(\mu))}$. The *minimax risk* of estimating $\tau_0(Q_0)$ is the

[9]The restrictions made on $\mathcal{Q}_0$ is for convenience. Note that the minimax rate over a larger set of distributions is necessarily bounded below by the minimax rate over $\mathcal{Q}_0$.

worst-case squared error of a given estimator T over $\mathcal{Q}_0$,

$$R_n(\mathcal{Q}_0, f) = \inf_T \sup_{Q_0 \in \mathcal{Q}_0} \mathbb{E}_{Q_0}\left[\{T(Y_{1:n}, X_{1:n}, \sigma_{1:n}) - \tau_0(Q_0)\}^2\right],$$

optimized over choices of all estimators.

The minimax rate measures the difficulty of estimating $\tau_0(Q_0)$ over Q_0. One way to interpret R_n is as the value of a zero-sum game where an analyst moves first and chooses and estimator T, and an adversary moves second and chooses a distribution $Q_0 \in \mathcal{Q}_0$. If different distributions $Q_0, Q_1 \in \mathcal{Q}_0$ produce very different τ_0 but very similar data, then the analyst would suffer large losses as they would be unable to distinguish these scenarios.

For well-behaved estimands (e.g. when $f(\mu) = \mu$), $R_n = O(1/\sqrt{n})$ contracts at the familiar parametric rate.[10] In sharp contrast, when $f(\cdot)$ is not an analytic function, the minimax rate vanishes slower than any polynomial in n. To reduce the uncertainty in our estimates proportionally by $t \in (0, 1)$, we therefore would require sample sizes exponential in $1/t$.

Theorem 3.2. *Under this setup, if f is not an analytic function,[11] then for any $\alpha > 0$, R_n contracts at a rate slower than $n^{-\alpha}$: $\limsup_{n\to\infty} n^\alpha R_n(\mathcal{Q}_0, f) = \infty$.*

Analytic functions are infinitely differentiable and admit Taylor expansions everywhere. However, many—if not most—transformations of interest are not analytic as they are typically non-smooth. For these functions, Theorem 3.2 is then a negative result, in the sense that the regression coefficient τ_0 associated with them is fundamentally difficult to estimate without assuming further structure.

The proof to Theorem 3.2 uses Le Cam's two-point method by constructing $Q_1, Q_0 \in \mathcal{Q}_0$ that generate similar data but very different $\tau_0(Q)$. A key step of the proof constructs Q_1, Q_0 so as to reduce the problem of estimating the regression coefficient τ_0 to the problem of estimating a mean $\mathbb{E}_Q[f(\mu)]$. Minimax rates for the latter problem then follow from the techniques developed by Cai and Low (2011) and presented in Wu and Yang (2020).[12] The proof to Theorem 3.2 similarly applies to regression problems where $f(\mu)$ appears on the left-hand side. That is, the minimax rate for the population regression coefficient τ_0 in the infeasible specification

$$f(\mu_i) = \rho_0 + \tau_0 W_i + \eta_i$$

suffers from similar subpolynomial rates if f is non-analytic.

[10] To wit, the restrictions on $\mathcal{Q}_0$ implies that $\tau_0(Q_0)$ is bounded by some $M > 0$. Thus, we can consider the estimator $T = \max(\min(\hat{\beta}, M), -M)$. The truncation at M is so that expectations always exist.

[11] A real-valued function $f : [-1, 1] \to \mathbb{R}$ is *analytic* if there exists an extension of f on an open subset of $\mathbb{C}$, $\tilde{f} : U \to \mathbb{C}$ where $U \subset \mathbb{C}$ is open, such that $\tilde{f} = f$ on $[-1, 1]$ and $\tilde{f}$ is complex analytic (i.e. complex differentiable).

[12] Cai and Low (2011) specifically study estimating $\mathbb{E}|\mu|$. Their proof technique extends to $\mathbb{E}[f(\mu)]$ by applying some results in approximation theory known as Bernstein's and Jackson's theorems. We have not seen these results stated in the statistical literature. Our proof of Theorem 3.2 makes it precise.

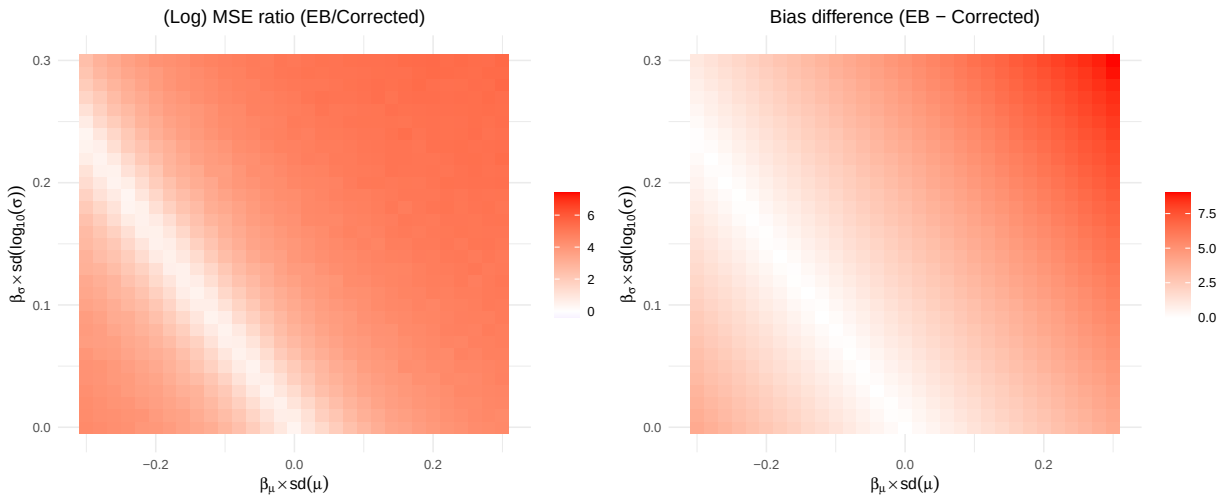


FIGURE 1. Simulation results for linear-in-μ regressions.

Notes: The left figure shows a heat map of the (natural) log of the ratio of the mean squared error (MSE) of $\hat{\beta}$ and $\tilde{\beta}$, over 1,000 simulation draws. The right figure shows a similar heat map, but for the difference in the bias of the two estimators.

4. **Simulation**

We follow Chen (2025) and calibrate our data generating process (DGP) to the Opportunity Atlas (Chetty, Friedman, Hendren, Jones and Porter, 2020), which provides (unshrunk) economic mobility estimates X_i and their standard errors σ_i. One measure of economic mobility μ_i of tract i they consider is the probability that a Black individual becomes relatively high-income (i.e., has family income in the top 20 percentiles nationally) after growing up relatively poor in tract i. (i.e., with parents at the 25th percentile nationally). We have 10,058 tracts in the data.

Taking this mobility measure as our measure of interest, we estimate the conditional mean function $\mathbb{E}[\mu_i \mid \sigma_i] = \mathbb{E}[X_i \mid \sigma_i]$ and the conditional variance function $\mathrm{Var}(\mu_i \mid \sigma_i) = \mathrm{Var}(X_i \mid \sigma_i) - \sigma_i^2$ via local linear regression implemented by Calonico, Cattaneo and Farrell (2019); denote the estimates of the conditional mean and conditional variance functions as $\hat{m}(\cdot)$ and $\hat{s}^2(\cdot)$, respectively. We use these estimates to draw μ_i's from the distribution $\mu_i \mid \sigma_i \sim \mathcal{N}(\hat{m}(\sigma_i), \hat{s}^2(\sigma_i))$. The outcome variable Y_i for the regression of interest is generated by

$$Y_i = \beta_\mu \mu_i + \beta_\sigma \log_{10} \sigma_i + u_i,$$

where $u_i \sim \mathcal{N}(0, 1)$. Under this DGP, the linear projection coefficient—in a hypothetical regression of Y_i on μ_i—we wish to estimate is given by

$$\beta_0 = \beta_\mu + \frac{\mathrm{Cov}(\mu_i, \log_{10} \sigma_i)}{\mathrm{Var}(\mu_i)} \beta_\sigma.$$

We vary $(\beta_\mu, \beta_\sigma)$[13] and compare the performance of our proposed estimator $\hat{\beta}$, given in (3), with the regression-on-shrinkage estimator $\tilde{\beta}$, given in (4).

Figure 1 summarizes the results of our simulation exercise in this setting by comparing the MSE and bias of the two estimators across the different DGPs. As expected, the MSE of $\hat{\beta}$ is smaller than the $\tilde{\beta}$ in almost all specifications, and this improvement is substantial in a wide range of simulations. For example, when $\text{Var}(\mu)^{1/2}\beta_\mu = .2$ and $\text{Var}(\log_{10}(\sigma))^{1/2}\beta_\sigma = .05$, which is a setting where, roughly speaking, $\log_{10}(\sigma)$ has a low explanatory power compared to μ, the log MSE ratio is around 4.36, which indicates that the MSE of $\tilde{\beta}$ is about 78 times greater than that of $\hat{\beta}$. Similarly, the bias plot shows that $\hat{\beta}$ having a smaller bias than $\tilde{\beta}$ across all DGPs, which is expected given that $\hat{\beta}$ is unbiased across all DGPs we consider.

For the simulations regarding nonlinear transformations of μ_i, we use a simpler DGP for μ_i. Specifically, we take an estimate of the unconditional mean and variance implied by the previous DGP, $\hat{m} = n^{-1}\sum_{i=1}^n \hat{m}(\sigma_i)$ and $\hat{s}^2 = n^{-1}\sum_{i=1}^n \hat{s}^2(\sigma_i) + n^{-1}\sum_{i=1}^n (\hat{m}(\sigma_i) - \hat{m})^2$ and draw $\mu_i \sim G$, where G is the distribution function of $\mathcal{N}(\hat{m}, \hat{s}^2)$. Then, we generate the outcome variable as

$$Y_i = \tau_0 \mathbb{1}(\mu_i > \mu_0) + u_i,$$

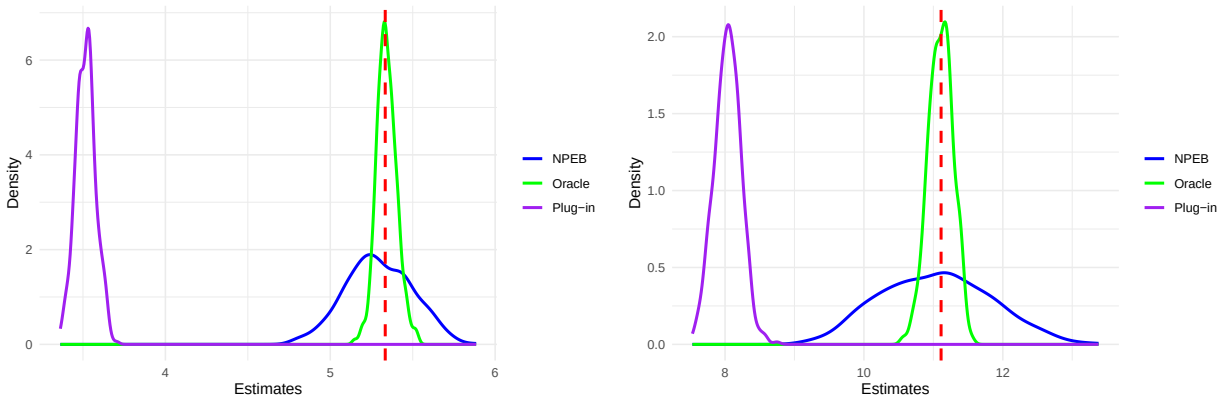
where $u_i \sim N(0,1)$ is independent of (μ_i, σ_i), and μ_0 is some fixed threshold assumed to be known. We vary τ_0 and μ_0 to learn about the performance of the different estimators we describe below.

We consider three distinct estimators. First is the oracle estimator $\hat{\tau}_0^{\text{oracle}}$ obtained by regressing Y_i on the true $\mathbb{E}_G[\mathbb{1}(\mu_i > \mu_0) \mid X_i, \sigma_i]$ using knowledge of the true G. By Proposition 3.1, $\hat{\tau}_0^{\text{oracle}}$ is consistent at the usual $n^{-1/2}$-rate. The second estimator we consider is the nonparametric empirical Bayes (NPEB) estimator $\hat{\tau}_0^{\text{NPEB}}$ obtained by regressing Y_i on $\mathbb{E}_{\hat{G}}[\mathbb{1}(\mu_i > \mu_0) \mid X_i, \sigma_i]$, where $\hat{G}$ is obtained by using nonparametric maximum likelihood as in Gilraine *et al.* (2020). This estimator is consistent, but potentially at a slower rate, as implied by Theorem 3.2. Finally, we also consider a plug-in estimator $\hat{\tau}_0^{\text{plug-in}}$ which is obtained by regressing Y_i on $\mathbb{1}(\mathbb{E}_G[\mu_i \mid X_i, \sigma_i] > \mu_0)$.[14] This estimator mimics the rather common empirical practice of plugging in the shrunk estimates $\hat{\mu}_i$ for μ_i in various downstream analyses.

Figure 2 shows the simulation results for two DGPs where we set τ_0 so that $\text{Var}(\mathbb{1}(\mu_i > \mu_0))^{1/2}\tau_0 = 1$. We consider two threshold levels for μ_0, the .75-quantile and .90-quantile of G. The plug-in estimator $\hat{\tau}_0^{\text{plug-in}}$ is substantially biased, as expected since $\hat{\tau}_0^{\text{plug-in}}$ plugging the posterior of μ_i into nonlinear transformation does not in general correct the measurement

[13] Specifically, we vary $(\beta_\mu, \beta_\sigma)$ so that $\text{Var}(\mu)^{1/2}\beta_\mu$ ranges from $-.3$ to .3 and $\text{Var}(\log_{10}(\sigma))^{1/2}\beta_\sigma$ from 0 to .3. Combined with the fact that we take $\text{Var}(u_i) = 1$, this results in a DGP where the regressors have realistic explanatory power.

[14] We use an "infeasible" plug-in estimator in the sense that we take the true G to calculate $\mathbb{E}_G[\mu_i \mid X_i, \sigma_i]$. Hence, the results should be considered an upper bound on how well a plug-in rule can do.



(A) μ_0 =.75-quantile of G (B) μ_0 =.9-quantile of G

FIGURE 2. Simulation results for regressions on nonlinear transformations of μ.

Notes: The figures show density plots of three different estimates across 500 simulation draws. The left figure shows results for a DGP where the threshold level μ_0 is set at the .75-quantile of G. The right figure shows the same for a DGP where the threshold level μ_0 is set at the .90-quantile of G. The red dashed line shows the true parameter value.

error. On the other hand, we see that $\hat{\tau}_0^{\text{oracle}}$ is unbiased, as expected by Proposition 3.1. Note that $\hat{\tau}_0^{\text{oracle}}$ shows higher accuracy when μ_0 is set at the .75-quantile of G, because this results in regressors with higher variance. Finally, while the estimates generated by $\hat{\tau}_0^{\text{NPEB}}$ are centered around the true parameter value, the distribution is quite dispersed and substantially noisier than $\hat{\tau}_0^{\text{oracle}}$. Given that the sample size is moderately large at 10,058, this demonstrates the slow convergence rate predicted by Theorem 3.2. The simulation results for other specifications of τ_0 and μ_0 are qualitatively similar.

5. **Empirical application: impact of examiner on patent outcome**

We revisit how patent examiners leniency (i.e., leniency in patent crafting) predicts the outcomes of the patents they decide to grant (Feng and Jaravel, 2020). As background, for each patent application, an examiner from a specific art unit is randomly assigned to assess the application and determine its merit for patent granting. There is a significant interaction between patent applicants and examiners during the revision of claims until the patent is either granted or the application is withdrawn. All other factors being equal, a stricter examiner may raise numerous questions regarding prior art, appropriate citations, and required clarifications before granting the patent, while a lenient examiner may provide minimal feedback for similar patents. This crafting process could influence the quality and clarity of the patent, should it be granted, which in turn could affect its market value, litigation propensity, and citations.

In particular, we consider the following infeasible regression:

$$Y_j = \beta_0 \mu_{i(j)} + a_{ut(j)} + \epsilon_j$$

where j indexes the patent, i the examiner, u the art unit, and t the filing year. The parameter β_0 represents the effect of examiner i's leniency during the crafting process, denoted as $\mu_{i(j)}$, on the patent outcome Y_j. Following Feng and Jaravel (2020), we use two measures of leniency. The first measure accounts for the average percentage change in the number of words per claim between the application and post-grant stages. The second measure records the percentage change in the total number of claims. These measures are constructed using the pre- and post-grant publication data[15].

An examiner who demands an increase in the length of the claims—typically to clarify distinctions from existing grants and enhance precision—is considered more careful and stringent. Similarly, an examiner who requests a reduction in the number of claims to limit the scope of the patent is also regarded as more stringent. As expected, these measures are, at best, noisy estimates of the true leniency μ_i.

We focus on the citation outcomes of patents within three years of grant. Additionally, we report results on the probability of purchase and litigation by a Patent Assertion Entity (PAE). In this case, the outcome variable takes a value of 1 if the granted patent is purchased by a PAE and is also litigated in court for infringement. We obtain the set of patents litigated by PAE firms from the Stanford NPE Litigation Database, which categorizes assertors into several categories. We classify acquired patents, failed startups, individual-inventor-stated companies, and individuals as Patent Assertion Entities (PAEs). As documented in Feng and Jaravel (2020), these entities acquire patents from third parties and generate revenue by asserting them against alleged infringers, commonly known as patent trolls. We also include additional cases from the Unified Patent and Lex Machina databases[16]. The citation data is publicly available from the USPTO and includes all citations received by a patent since its grant, updated annually.

The discussion below excludes the art unit and year fixed effects. In practice, we first purge these fixed effects from all other variables as discussed in Section 2.3.[17] Let the measure of leniency for each patent j reviewed by examiner i be denoted by E_{ij} and N_i represents the number of patents granted by examiner i. Then $Y_i = \frac{1}{N_i} \sum_{j:i(j)=i} Y_{ij}$ is the

[15]These data are publically available at https://patentsview.org/download/pg_claims and https://patentsview.org/download/claims.

[16]We thank Dr. Tommaso Alba for providing these data.

[17]We consider only those examiners who serve in a single art unit and have granted at least 10 patents, leaving us with 4,615 examiners.

examiner-level mean outcome[18], and $E_i = \frac{1}{N_i}\sum_{j:i(j)=i} E_{ij}$ is the examiner's leniency, for which $E_i \mid \mu_i \sim N(\mu_i, \sigma_i^2)$. where σ_i^2 is proportional to the inverse of N_i.

Replacing $\mu_{i(j)}$ with $E_{i(j)}$ introduces attenuation bias. Accounting for the fact that we aggregate the patent level regression to the examiner level, the population least-square objective is

$$\mathbb{E}\Big[\sum_{j:i(j)=i}(Y_{ij} - \delta_0 - \beta_0\mu_i)^2\Big] = \mathbb{E}[N_i(Y_i - \delta_0 - \beta_0\mu_i)^2] + \text{constant}$$

Therefore, the polulation coefficient β_0 is the weighted least square coefficient at the examiner level, with weights proportional to N_i. The natural sample analogue, correcting for measurement error, leads to:

$$\hat{\beta} = \frac{\sum_i W_i Y_i E_i - (\sum_i W_i E_i)(\sum_i W_i Y_i)}{\sum_i W_i E_i^2 - (\sum_i W_i E_i)^2 - \sum_i W_i \sigma_i^2} \quad W_i = N_i / \sum_i N_i.$$

An additional concern may be that the left-hand side variable is also a measurement of a latent effect and contains measurement error. In this case, it is possible that the measurement error in both Y_{ij} and E_{ij} are correlated. To model this, we consider the following model

$$\begin{pmatrix} Y_{ij} \\ E_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} \rho_i \\ \mu_i \end{pmatrix}, \Sigma\right) \quad \text{and} \quad \begin{pmatrix} Y_i \\ E_i \end{pmatrix} \sim N\left(\begin{pmatrix} \rho_i \\ \mu_i \end{pmatrix}, \Sigma_i\right)$$

where the second display is the model for the examiner-specific sample averages (Y_i, E_i) and $\Sigma_i = \Sigma/N_i$. The coefficient β_0 can be related to the population coefficient of a regression of Y_i on E_i by $\beta = \frac{\text{Cov}(Y_i,E_i)-\Sigma_{i,12}}{\text{Var}(E_i)-\Sigma_{i,22}}$. In the presence of heterogeneous N_i, the natural weighted least square estimator with the correction due to measurement error on both sides yields

$$\check{\beta} = \frac{\sum_i W_i Y_i E_i - (\sum_i W_i E_i)(\sum_i W_i Y_i) - \sum_i W_i \hat{\Sigma}_{i,12}}{\sum_i W_i E_i^2 - (\sum_i W_i E_i)^2 - \sum_i W_i \sigma_i^2}$$

where $\hat{\Sigma}_{i,12}$ is estimated based on patent level data (Y_{ij}, E_{ij}) and normalized by N_i.

In contrast to the two proposed estimators above, the common regress-on-shrinkage approach constructs the posterior mean as:

$$\overline{E}_i = \hat{\mu}_0 + \frac{\hat{\sigma}_e^2}{\sigma_i^2 + \hat{\sigma}_e^2}(E_i - \hat{\mu}_0)$$

where $\hat{\mu}_0$ and $\hat{\sigma}_e^2$ are often estimated via moments of E_i. The resulting weighted least square estimator is:

$$\tilde{\beta} = \frac{\sum_i W_i Y_i \overline{E}_i - (\sum_i W_i \overline{E}_i)(\sum_i W_i Y_i)}{\sum_i W_i \overline{E}_i^2 - (\sum_i W_i \overline{E}_i)^2}.$$

Table 1 reports the estimates for β_0 using different methods. The column with label FE represents the OLS estimation without correcting for measurement error in the leniency

[18]Here we index the patent level outcome by Y_{ij} rather than Y_j to clarify that the patent level variables are more granular than the examiner level.

measures. This estimator biases towards zero due to un-corrected measurement error. The next column reports the regress-on-shrinkage estimator. The last two columns (labeled as FE-Corrected and FE-Corrected-Twoside) report estimates based on the measurement error correction. FE-Corrected only considers measurement error in leniency measures while FE-Corrected-Twoside accounts for measurement error from both the left and right hand side variables. The standard errors in the bracket for our proposed method are obtained via bootstrap with 999 bootstrap repetitions.

TABLE 1. Patent Level Outcome Regression

	FE	Shrinkage	FE-Corrected	FE-Corrected-Twoside
		Citation within 3 year		
word change	-3.307^{***}	-3.950^{***}	-3.902^{***}	-3.786^{***}
	(0.536)	(0.614)	(0.805)	(0.841)
claims change	2.924^{***}	5.705^{***}	4.307^{*}	4.275^{*}
	(0.780)	(1.110)	(2.203)	(2.236)
		PAE Litigated		
word change	-0.0034^{***}	-0.0039^{***}	-0.0040^{***}	-0.0040^{***}
	(0.0007)	(0.0008)	(0.0009)	(0.0009)
claims change	0.0027^{**}	0.0043^{**}	0.0039^{**}	0.0037^{**}
	(0.0010)	(0.0014)	(0.0016)	(0.0017)
Note:				*p<0.1; **p<0.05; ***p<0.01.

For the PAE litigation outcome, there is little difference between the shrinkage method and the direct correction we proposed. However, for the citation outcome, the difference is noticeable. As reported in Table 2, individual variances σ_i appear to have a significant impact on the citation outcome, but not on the PAE litigation outcome. This suggests some evidence of a violation of the assumptions necessary to ensure the consistency of the regress-on-shrinkage estimator.

TABLE 2. Regression on logarithm of σ_i.

	Dependent variable:	
	Citation within 3 years	PAE Litigated
$\log_{10}(\sigma_i)$	-0.749^{***}	-0.0002
	(0.141)	(0.0002)
Constant	-1.346^{***}	-0.0004
	(0.256)	(0.0003)
Note:		*p<0.1; **p<0.05; ***p<0.01

6. **Conclusion**

This paper critically examines the widespread practice of using the empirical Bayes shrinkage estimator to correct for measurement errors in regression models incorporating individual latent effects. Our analysis reveals that this approach only provides reliable estimates for the regression coefficients under unnecessarily stringent conditions. We demonstrate that the classical correction, which we advocate, holds under weaker assumptions and cannot be asymptotically improved when the latent effect enters the regression model linearly. In cases where the latent effect enters non-linearly, empirical Bayes shrinkage leads to slower minimax estimation rates. These findings underscore the limitations of using the regress-on-shrinkage estimator as a method for correcting measurement error in regression models.

References

Angelova, V., Dobbie, W. S. and Yang, C. (2023). *Algorithmic recommendations and human discretion*. Tech. rep., National Bureau of Economic Research. 6

Angrist, J., Hull, P. and Walters, C. (2023). Methods for measuring school effectiveness. *Handbook of the Economics of Education*, **7**, 1–60. 2, 6

Battaglia, L., Christensen, T., Hansen, S. and Sacher, S. (2024). Inference for regression with variables generated by ai or machine learning. 4

Bau, N. and Das, J. (2020). Teacher value added in a low-income country. *American Economic Journal: Economic Policy*, **12** (1), 62–96. 6

Bickel, P. and Ritov, Y. (1987). Efficient estimation in the errors in variables model. *The Annals of Statistics*, **15** (2), 513–540. 3, 10

Bruhn, J., Imberman, S. and Winters, M. (2022). Regulatory arbitrage in teacher hiring and retention: Evidence from Massachusetts Charter Schools. *Journal of Public Economics*, **215**, 104750. 12

Cai, T. T. and Low, M. G. (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, pp. 1012–1041. 3, 14

Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019). nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference. *Journal of Statistical Software*, **91**, 1–33. 15

Chandra, A., Finkelstein, A., Sacarny, A. and Syverson, C. (2016). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review*, **106** (8), 2110–2144. 2, 5

Chen, J. (2025). Empirical bayes when estimation precision predicts parameters. *arXiv preprint arXiv:2212.14444*. 2, 3, 6, 9, 15

Chen, X. and Santos, A. (2018). Overidentification in regular models. *Econometrica*, **86** (5), 1771–1817. 9, 10, 27, 28

Chetty, R., Friedman, J. N., Hendren, N., Jones, M. R. and Porter, S. (2020). *The Opportunity Atlas Mapping the Childhood Roots of Social Mobility*. Tech. rep. 15

—, — and Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, **104** (9), 2593–2632. 2, 5, 7

—, — and — (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, **104** (9), 2633–79. 7, 8

— and Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility II: County-level estimates. *Quarterly Journal of Economics*, **133** (3), 1163–1228. 2

de Chaisemartin, C. and Deeb, A. (2024). Estimating treatment-effect heterogeneity across sites in multi-site randomized experiments with imperfect compliance. *arXiv e-prints*, pp. arXiv–2405. 6, 12

Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of econometrics*, **30** (1-2), 109–126. 6

Deeb, A. (2021). A framework for using value-added in regressions. *arXiv preprint arXiv:2109.01741*. 3

Devereux, P. J. (2007). Improved errors-in-variables estimators for grouped data. *Journal of Business & Economic Statistics*, **25** (3), 278–287. 12

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, vol. 303. Springer Science & Business Media. 31

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, pp. 1900–1925. 13

Feng, J. and Jaravel, X. (2020). Crafting intellectual property rights: Implications for patent assertion entities, litigation, and innovation. *American Economic Journal: Applied Economics*, **12** (1), 140–181. 2, 3, 17, 18

Fenizia, A. (2022). Managers and productivity in the public sector. *Econometrica*, **90** (3), 1063–1084. 2

Fuller, W. A. (1987). *Measurement error models*. John Wiley & Sons. 2, 3, 6

Gilraine, M., Gu, J. and McMillan, R. (2020). *A new method for estimating teacher value-added*. Tech. rep., National Bureau of Economic Research. 2, 3, 6, 16

Gu, J. and Koenker, R. (2017). Unobserved heterogeneity in income dynamics: An empirical bayes perspective. *Journal of Business & Economic Statistics*, **35** (1), 1–16. 2

Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, **126** (5), 2072–2107. 5, 6

Jacob, B. and Lefgren, L. (2005a). *Principals as Agents: Subjective Performance Measurement in Education*. Tech. Rep. w11463, National Bureau of Economic Research, Cambridge, MA. 2

— and — (2005b). What do parents value in education? an empirical investigation of parents' revealed preferences for teachers. 5, 6, 7, 8

Kane, T. J., Rockoff, J. E. and Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, **27** (6), 615–631. 2

— and Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. Tech. rep., National Bureau of Economic Research. 2, 6

Kwon, S. (2023). Optimal shrinkage estimation of fixed effects in linear panel data models. 3, 6

Lehmann, E. L. and Romano, J. P. (2008). *Testing statistical hypotheses*, vol. 3. Springer. 28

Mulhern, C. (2023). Beyond teachers: Estimating individual school counselors' effects on educational attainment. *American economic review*, **113** (11), 2846–2893. 5

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press: Berkeley, vol. I. 2

Rose, E. K., Schellenberg, J. T. and Shem-Tov, Y. (2022). *The effects of teacher quality on adult criminal justice contact*. Tech. rep., National Bureau of Economic Research. 3

Stigler, S. M. (1990). The 1988 neyman memorial lecture: a galtonian perspective on shrinkage estimators. *Statistical Science*, pp. 147–155. 7

Sullivan, D. G. (2001). A note on the estimation of linear regression models with heteroskedastic measurement errors. *Available at SSRN 295567*. 10

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*, vol. 4. Springer. 9

van der Vaart, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press. 9, 10, 26, 27, 28

Walters, C. (2024). Empirical bayes methods in labor economics. In *Handbook of Labor Economics*, vol. 5, Elsevier, pp. 183–260. 2, 4

Warnick, M., Light, J. and Yim, A. (2025). Instructor value-added in higher education. 5, 6

Wu, Y. and Yang, P. (2020). Polynomial methods in statistical inference: Theory and practice. *Foundations and Trends® in Communications and Information Theory*, **17** (4), 402–586. 14, 31

Xie, T. (2025). *Automatic Inference for Value-Added Regressions*. Tech. rep. 3

Appendix A. Proof of Lemma 2.3

We restate and prove the following result.

Lemma A.1. *Under Assumption 2.1, Assumption 2.3(1) implies Assumption 2.2; Assumption 2.3(2) implies Assumption 2.2(1), and Assumption 2.3(3) implies Assumption 2.3(2).*

Proof. (1) Precision independence implies precision independence in the first two moments. Thus Assumption 2.3 implies Assumption 2.2(1) by part (2) of this lemma. For exogeneity, write

$$\mu^*(X_i, \sigma_i) = (1 - w(\sigma_i))\mu + w(\sigma_i)X_i$$

for $\mu = \mathbb{E}[\mu_i]$ By assumption, $\sigma_i \perp\!\!\!\perp (\mu_i, \eta_i)$. Then

$$\begin{aligned}\mathbb{E}[\eta_i \mu^*(X_i, \sigma_i)] &= \mathbb{E}[(1 - w(\sigma_i))]\mu \underbrace{\mathbb{E}[\eta_i]}_{0} + \mathbb{E}[w(\sigma_i)(\mu_i + \sigma_i\epsilon_i)\eta_i] \quad \epsilon_i \sim \mathcal{N}(0,1) \\ &= \mathbb{E}[w(\sigma_i)] \underbrace{\mathbb{E}[\mu_i\eta_i]}_{0} + \mathbb{E}[w(\sigma_i)\sigma_i] \underbrace{\mathbb{E}[\epsilon_i]}_{0} \mathbb{E}[\eta_i] = 0.\end{aligned}$$

(2) We compute

$$\begin{aligned}\operatorname{Cov}(\mu_i, \mu^*(X_i, \sigma_i)) &= \mathbb{E}\operatorname{Cov}(\mu_i, \mu^*(X_i, \sigma_i) \mid \sigma_i) + \operatorname{Cov}\left(\underbrace{\mathbb{E}[\mu_i \mid \sigma_i]}_{\text{constant } \mu}, \mathbb{E}[\mu^*(X_i, \sigma_i) \mid \sigma_i]\right) \\ &= \mathbb{E}\left[w(\sigma_i)\operatorname{Cov}(\mu_i, X_i \mid \sigma_i)\right] \\ &= \mathbb{E}\left[w(\sigma_i)\operatorname{Var}(\mu_i \mid \sigma_i)\right] \\ &= \mathbb{E}\left[\frac{1}{\sigma_\mu^2 + \sigma_i^2}\right]\sigma_\mu^4.\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}(\mu^*(X_i, \sigma_i)) &= \mathbb{E}\operatorname{Var}(\mu^*(X_i, \sigma_i) \mid \sigma_i) + \operatorname{Var}\left(\underbrace{\mathbb{E}\left[\mu^*(X_i, \sigma_i) \mid \sigma_i\right]}_{\text{constant } \mu}\right) \\ &= \mathbb{E}\left[w^2(\sigma_i)\operatorname{Var}(X_i \mid \sigma_i)\right] \\ &= \mathbb{E}\left[\left(\frac{\sigma_\mu^2}{\sigma_\mu^2 + \sigma_i^2}\right)^2 (\sigma_\mu^2 + \sigma_i^2)\right] \\ &= \mathbb{E}\left[\frac{1}{\sigma_\mu^2 + \sigma_i^2}\right]\sigma_\mu^4.\end{aligned}$$

Thus

$$\operatorname{Cov}(\mu_i, \mu^*(X_i, \sigma_i))/\operatorname{Var}(\mu^*(X_i, \sigma_i)) = 1.$$

(3) Finally, we have that $\mathbb{E}[\eta_i \mid \mu_i, \sigma_i] = 0$ by assumption. Thus,

$$\begin{aligned}
\mathbb{E}[\eta_i \mu^*(X_i, \sigma_i)] &= \mathbb{E}\left[\mathbb{E}[\eta_i \mu^*(X_i, \sigma_i) \mid \mu_i, \sigma_i]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\eta_i \left\{(1 - w(\sigma_i))\mu + w(\sigma_i)(\mu_i + \sigma_i \epsilon_i)\right\} \mid \mu_i, \sigma_i\right]\right] \\
&= \mathbb{E}\left[\sigma_i \mathbb{E}[\eta_i \epsilon_i \mid \mu_i, \sigma_i]\right] && (\mathbb{E}[\eta_i \mid \sigma_i, \mu_i] = 0) \\
&= 0.
\end{aligned}$$

where the last step follows since $\epsilon_i \mid (Y_i, \mu_i, \sigma_i)$ is mean zero.

□

Appendix B. Proof of Theorem 2.4

We first carefully define $P_0, \mathcal{P}_0, P, \mathcal{P}$. Given P_0 a joint distribution on the complete variables (Y_i, μ_i, σ_i), we define $P^{\text{obs}} = P^{\text{obs}}(P_0)$ as the induced distribution on the observed variables (Y_i, X_i, σ_i). Let $\mathcal{P}_0$ be the set of distributions P_0 that satisfy Assumption 2.1 and are dominated by some σ-finite product measure $\lambda_0 = \lambda_Y \otimes \lambda_{\mathbb{R}} \otimes \lambda_\mu \otimes \lambda_\sigma$, where $\lambda_{\mathbb{R}}$ is the Lebesgue measure and (iii) the support of $\mu \mid Y, \sigma$ under P_0 contains a nonempty interval almost surely. Similarly, let $\mathcal{P} = \left\{P^{\text{obs}}(P_0) : P_0 \in \mathcal{P}_0\right\}$, where $\mathcal{P}$ is dominated by $\lambda = \lambda_Y \otimes \lambda_{\mathbb{R}} \otimes \lambda_\sigma$. Let $\mathcal{Z} \subset \mathbb{R}^3$ denote the set of values that (Y_i, X_i, σ_i) takes. For a given $P \in \mathcal{P}$, define $L^2(P)$ as the Hilbert space of square-integrable functions $\mathcal{Z} \to \mathbb{R}$ under P and $L_0^2(P)$ as the Hilbert space of square-integrable and mean-zero functions $\mathcal{Z} \to \mathbb{R}$ under P.

We recall the following definitions from semiparametric theory (see, e.g., chapter 25 of van der Vaart, 2000).

Definition B.1 (Parametric submodel)**.** For a given $P \in \mathcal{P}$, a *smooth parametric submodel* is a set $\{P_t : t \in [0, \epsilon)\} \subset \mathcal{P}$ that is differentiable in quadratic mean at $t = 0$ and $P_{t=0} = P$: For some measurable function $g : \mathcal{Z} \to \mathbb{R}$,

$$\lim_{t \downarrow 0} \int \left\{ \frac{1}{t}(\sqrt{p_t} - \sqrt{p}) - \frac{1}{2} g \sqrt{p} \right\}^2 d\lambda = 0,$$

where $p_t = dP_t/d\lambda$ and $p = dP/d\lambda$ are densities with respect to the dominating measure. We refer to g as the *score* of the submodel.

Definition B.2 (Tangent space)**.** For a given $P \in \mathcal{P}$, the *tangent set* of $\mathcal{P}$ at P is defined as

$$\mathcal{T}(P) \triangleq \left\{g \in L_0^2(P) : \text{There exists a smooth parametric submodel with score } g\right\} \subset L_0^2(P).$$

The *tangent space* $\overline{\mathcal{T}}(P)$ at P is defined as the closure of the linear span of $\mathcal{T}(P)$ with respect to $L_0^2(P)$.

Definition B.3 (Regularity and asymptotic linearity)**.** For a given $P \in \mathcal{P}$ and a given parameter $\theta(P) \in \mathbb{R}$, an estimator $\hat{\theta}_n$ is *regular* if there exists a distribution L such that along all smooth parametric submodels P_t,

$$\sqrt{n}(\hat{\theta}_n - \theta(P_{1/\sqrt{n}})) \xrightarrow[P_{1/\sqrt{n}}]{d} L.$$

$\hat{\theta}_n$ is *asymptotically linear* at P if there exists some *influence function* $\psi \in L_0^2(P)$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i, X_i, \sigma_i; \theta(P)) + o_P(1).$$

Finally, we recall Definition 2.2 from Chen and Santos (2018).

Definition B.4 (Local just identification)**.** For a given $P \in \mathcal{P}$, if $\overline{\mathcal{T}}(P) = L_0^2(P)$, we say P is locally just identified by $\mathcal{P}$.

Loosely speaking, local just identification at P means that P can be perturbed in any direction within $\mathcal{P}$ and the model $\mathcal{P}$ is consistent with any (local) parametric model at P. Thus, there is no information that an analyst could exploit by imposing the model $\mathcal{P}$, since $\mathcal{P}$ makes no restrictions locally at P. This is the semiparametric analogue to just identification in parametric GMM models, where there are no additional moment conditions to exploit, and all GMM weightings yield the same estimator. Indeed, Chen and Santos (2018) (Theorem 3.1(i)) show that when P is locally just identified, all RAL estimators are asymptotically equivalent.

This section verifies the following theorem, under a technical assumption stated immediately after.

Theorem 2.4. *Let $P \in \mathcal{P}$ and let $P_0 \in \mathcal{P}_0$ be the corresponding distribution over the complete data, defined formally in Appendix B. Then P is locally just identified by $\mathcal{P}$. As a result,*

(1) Any RAL estimator $\check{\beta}$ for

$$\beta_0 \triangleq \beta_0(P) \triangleq \frac{\mathrm{Cov}_P(X, Y)}{\mathrm{Var}_P(X) - \mathbb{E}_P[\sigma^2]}$$

is asymptotically equivalent to the analogue estimator $\hat{\beta}$: $\sqrt{n}(\check{\beta} - \hat{\beta}) = o_P(1)$.

(2) The semiparametric efficiency bound for $\beta_0(P)$ is equal to the asymptotic variance of $\hat{\beta}$ at P.

Proof. The proof follows by applying Example 25.35 in van der Vaart (2000).

Fix $P_0 \in \mathcal{P}_0$, let Q_0 be the corresponding distribution of (Y_i, μ_i, σ_i) under P_0 and let $\mathcal{Q}_0$ collect all such distributions as P_0 ranges over $\mathcal{P}_0$. First observe that the tangent space at Q_0 relative to $\mathcal{Q}_0$

$$\overline{\mathcal{T}}(Q_0) = L_0^2(\mathcal{Q}_0).$$

Let $\boldsymbol{X} = (Y_i, X_i, \sigma_i) \sim P$ and $\boldsymbol{Z} = (Y_i, \mu_i, \sigma_i) \sim Q_0$. Note that the density of $\boldsymbol{X} \mid \boldsymbol{Z}$ (with respect to the dominating measure $\delta_y \otimes \lambda_{\mathbb{R}} \otimes \delta_\sigma$) is

$$p(\boldsymbol{X} \mid \boldsymbol{Z} = (y, \mu, \sigma)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

which is exponential family. The score space for P consists of the functions

$$(A_{Q_0} b)(\boldsymbol{x}) \triangleq \mathbb{E}_{Q_0}[b(\boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x} = (y, x, \sigma)] = \frac{\int_{-\infty}^{\infty} b(y, \mu, \sigma) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \, dQ_0(\mu \mid y, \sigma)}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \, dQ_0(\mu \mid y, \sigma)}$$

for b in the tangent space $\overline{\mathcal{T}}(Q_0)$. Following Example 25.35 in van der Vaart (2000), we show that such scores are dense in $L_0^2(P)$.

It suffices to show that the closure of the range of the *score operator* A_{Q_0} is $L_0^2(P)$, since $\overline{\mathcal{T}}(Q_0) = L_0^2(Q_0)$. This is further equivalent to showing the orthocomplement of the kernel $N(A_{Q_0}^*)$ is equal to $L_0^2(P)$. Thus it suffices to show that the kernel $N(A_{Q_0}^*)$ is trivial: That is,

$$0 = (A_\eta^* g)(\boldsymbol{z}) = \mathbb{E}[g(\boldsymbol{X}) \mid \boldsymbol{Z} = \boldsymbol{z} = (y, \mu, \sigma)] \; Q_0(\cdot \mid y, \sigma)\text{-a.s.} \implies g(y, \cdot, \sigma) = 0 \text{ a.e.}$$

By assumption, the support of $\mu \mid Y, \sigma$ under Q_0 contains an interval a.s., and thus the above display is true, for Q_0-almost all (Y, σ), by the completeness of Gaussian location models (Theorem 4.3.1 in Lehmann and Romano (2008)). This shows that members of $N(A_{Q_0}^*)$ are almost surely zero, and this completes the proof for the first statement.

Theorem 3.1(i) in Chen and Santos (2018) immediately implies (1), since (a) the tangent set $\mathcal{T}(P)$ is closed under linear combinations and (b) $\hat{\beta}$ is an RAL estimator. For (2), since $\hat{\beta}$ is an RAL estimator, the projection of $\hat{\beta}$'s influence function onto $\overline{\mathcal{T}}(P)$ is the efficient influence function. However, since $\overline{\mathcal{T}}(P) = L_0^2(P)$, the projection of $\hat{\beta}$'s influence function onto $\overline{\mathcal{T}}(P)$ is itself. This observation implies (2). □

Appendix C. Proof of Theorem 3.2

This section restates and proves the following theorem.

Theorem 3.2. *Under this setup, if f is not an analytic function,*[19] *then for any $\alpha > 0$, R_n contracts at a rate slower than $n^{-\alpha}$:* $\limsup_{n\to\infty} n^\alpha R_n(\mathcal{Q}_0, f) = \infty$.

Remark C.1. Lemma C.1 provides a construction linking functional estimation in Gaussian white noise models to estimation of $T(Q)$ and derives a lower bound using Le Cam's two-point method, given two priors G_{-1}, G_1 of μ. Theorem C.2 shows that certain worst-case choices of G_{-1}, G_1 connect to the problem of uniform approximation by polynomials. Theorem C.3

[19] A real-valued function $f : [-1, 1] \to \mathbb{R}$ is *analytic* if there exists an extension of f on an open subset of $\mathbb{C}$, $\tilde{f} : U \to \mathbb{C}$ where $U \subset \mathbb{C}$ is open, such that $\tilde{f} = f$ on $[-1, 1]$ and $\tilde{f}$ is complex analytic (i.e. complex differentiable).

is a result in approximation theory that shows that functions are well-approximated by polynomials if and only if they are analytic.

Theorem C.2 and Lemma C.1 links the minimax lower bound to polynomial approximation. Theorem C.3 shows that if f is not analytic, then the approximation rate cannot decay exponentially. The proof here puts things together and verifies that, *therefore*, the minimax rate cannot be polynomial. ■

Proof. By Lemma C.1 and Theorem C.2, we know that for any integer $k \geq 1$, we have that

$$R_n(\mathcal{Q}_0, f) \gtrsim_{\|f\|_\infty} E_k(f)^2 \left(1 - C \exp\left(-\frac{k+1}{2}(\log(k+1) - 1) + \frac{1}{2}\log n\right)\right)$$

where $E_k(f)$ is defined in Theorem C.2. If f is not analytic, by Theorem C.3, we have that[20]

$$\limsup_{k\to\infty} E_k(f)^{1/k} = 1. \tag{C.1}$$

Hence there exists a subsequence (k_ℓ) where $E_{k_\ell}(f)^{1/k_\ell} \to 1$ as $\ell \to \infty$. Note that we can take another subsequence (k_n) such that (i) $k_n \asymp \log n$ and (ii) infinitely many elements from (k_ℓ) features in (k_n).

For such a k_n,

$$R_n(\mathcal{Q}_0, f) \gtrsim_{\|f\|_\infty} E_{k_n}(f)^2.$$

Suppose, for contradiction, there is some α for which

$$\limsup_{n\to\infty} n^\alpha R_n(\mathcal{Q}_0, f) < \infty.$$

Then

$$\limsup_{n\to\infty} n^\alpha E_{k_n}(f)^2 \leq \infty.$$

This means that for some subsequence of (k_n) that contains infinitely many elements of k_ℓ, (k_{n_m}), we have that $\sup_m |n_m^\alpha E^2_{k_{n_m}(f)}| < M^2 < \infty$. However,

$$\begin{aligned} E_{k_{n_m}}(f)^{1/k_{n_m}} &= (n_m^{\alpha/2} E_{k_{n_m}}(f))^{1/k_{n_m}} n_m^{-\frac{\alpha}{2k_{n_m}}} \\ &\leq M^{1/k_{n_m}} \exp\left(-\frac{\alpha}{2k_{n_m}} \log n_m\right) \\ &\overset{m\to\infty}{\longrightarrow} \exp(-\alpha c) < 1 \end{aligned}$$

where $c \triangleq 2\lim_{m\to\infty} \frac{\log n_m}{k_{n_m}} > 0$ since $k_{n_m} \asymp \log n_m$. This contradicts (C.1). □

Lemma C.1. *In this problem, for any G_{-1}, G_1 supported on $[-1, 1]$,*

$$R_n(\mathcal{Q}_0, f) \gtrsim_{\|f\|_\infty} \left(\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]\right)^2 \left(1 - \frac{1}{2\sqrt{2}}\sqrt{n\chi^2(G_{-1} \star \mathcal{N}(0,1), G_1 \star \mathcal{N}(0,1))}\right)$$

[20] Note that $E_k(f)$ is bounded and decreasing, and so $E_k(f)^{1/k} \leq E_1(f)^{1/k} \to 1$.

where $\chi^2(Q_1, Q_2)$ is the χ^2-divergence between Q_1 and Q_2.

Proof. Fix some distributions G_{-1}, G_1 supported on $[-1, 1]$. Let G^*_{-1}, G^*_1 be the mixture

$$G^*_j = \frac{1}{2}G_j + \frac{1}{2}\,\mathrm{Unif}[-1, 1].$$

Note that $\mathrm{Var}_{G^*_j}[f(\mu)] \geq \frac{V(f)}{2}$. Choose distributions such that $\sigma = 1$ almost surely.

Consider induced Q_0, Q_1 where under Q_j,

$$\begin{aligned} Y &\sim_{Q_j} \text{Rademacher}(1/2) \text{ for } j = 0, 1 \\ \mu \mid Y &\sim_{Q_0} G^*_{-1} \\ \mu \mid Y = -1 &\sim_{Q_1} G^*_{-1} \\ \mu \mid Y = 1 &\sim_{Q_1} G^*_1. \end{aligned}$$

Thus, $Q_0, Q_1 \in \mathcal{Q}$. Note that

$$T(Q_0) = \frac{\frac{1}{2}\mathbb{E}_{G^*_{-1}}[f(\mu)] - \frac{1}{2}\mathbb{E}_{G^*_{-1}}[f(\mu)]}{\mathrm{Var}_{Q_0}(f(\mu))} = 0$$

and

$$\begin{aligned} T(Q_1) &= \frac{\frac{1}{2}\mathbb{E}_{G^*_1}[f(\mu)] - \frac{1}{2}\mathbb{E}_{G^*_{-1}}[f(\mu)]}{\mathrm{Var}_{G^*_1}(f(\mu))} \\ &= \frac{1}{2\,\mathrm{Var}_{G^*_1}(f(\mu))}\left(\mathbb{E}_{G^*_1}[f(\mu)] - \mathbb{E}_{G^*_{-1}}[f(\mu)]\right) \\ &= \frac{1}{4\,\mathrm{Var}_{G^*_1}(f(\mu))}\left(\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]\right) \end{aligned}$$

Let $P(Q) = P^{\mathrm{obs}}(Q)$ denote distribution of the observed data induced by $Q \in \mathcal{Q}_0$. Recall that Le Cam's two-point method states that for some absolute $c > 0$

$$R_n(\mathcal{Q}_0, f) \geq c(T(Q_1) - T(Q_0))^2 \left(1 - \mathrm{TV}\left(P(Q_1)^{\otimes n}, P(Q_0)^{\otimes n}\right)\right).$$

Note that

$$\begin{aligned} &1 - \mathrm{TV}\left(P(Q_1)^{\otimes n}, P(Q_0)^{\otimes n}\right) \\ &\geq 1 - \frac{1}{\sqrt{2}}\sqrt{\mathrm{KL}\left(P(Q_0)^{\otimes n}, P(Q_1)^{\otimes n}\right)} \\ &= 1 - \frac{1}{\sqrt{2}}\sqrt{n\mathrm{KL}\left(P(Q_0), P(Q_1)\right)} && \text{(Tensorization of KL)} \\ &= 1 - \frac{1}{\sqrt{2}}\sqrt{\frac{n}{2}\left(\mathrm{KL}\left(G^*_{-1} \star \mathcal{N}(0,1), G^*_{-1} \star \mathcal{N}(0,1)\right) + \mathrm{KL}\left(G^*_{-1} \star \mathcal{N}(0,1), G^*_1 \star \mathcal{N}(0,1)\right)\right)} \\ & && \text{(Conditional KL given } Y = 0, Y = 1) \\ &= 1 - \frac{1}{2}\sqrt{n\mathrm{KL}(G^*_{-1} \star \mathcal{N}(0,1), G^*_1 \star \mathcal{N}(0,1))} \end{aligned}$$

$$= 1 - \frac{1}{2}\sqrt{n\chi^2(G_{-1}^* \star \mathcal{N}(0,1), G_1^* \star \mathcal{N}(0,1))}$$
$$\geq 1 - \frac{1}{2\sqrt{2}}\sqrt{n\chi^2(G_{-1} \star \mathcal{N}(0,1), G_1 \star \mathcal{N}(0,1))}$$
(By convexity: $\chi^2(G_{-1}^* \star \mathcal{N}(0,1), G_1^* \star \mathcal{N}(0,1)) \leq \frac{1}{2}\chi^2\left(G_{-1} \star \mathcal{N}(0,1), G_1 \star \mathcal{N}(0,1)\right)$)

Putting things together, we have that

$$R_n(\mathcal{Q}_0, f)$$
$$\gtrsim \frac{1}{\mathrm{Var}_{G_1}(f(\mu))^2}\left(\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]\right)^2\left(1 - \frac{1}{2\sqrt{2}}\sqrt{n\chi^2(G_{-1} \star \mathcal{N}(0,1), G_1 \star \mathcal{N}(0,1))}\right)$$
$$\gtrsim_{\|f\|_\infty} \left(\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]\right)^2\left(1 - \frac{1}{2\sqrt{2}}\sqrt{n\chi^2(G_{-1} \star \mathcal{N}(0,1), G_1 \star \mathcal{N}(0,1))}\right).$$

□

Theorem C.2. *Fix bounded and measurable $f : [-1,1] \to \mathbb{R}$, let*

$$E_k(f) = \inf_{a_{0:k} \in \mathbb{R}^{k+1}} \|f(x) - a_0 - a_1 x^1 - \cdots - a_k x^k\|_\infty$$

be the best approximation error of f via polynomials. Then, for universal $c, C > 0$, for any $k \geq 1$, there exist G_{-1}, G_1 such that

$$|\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]| \geq cE_k(f)$$

and

$$\chi^2\left(G_1 \star \mathcal{N}(0,1), G_{-1} \star \mathcal{N}(0,1)\right) \leq C\exp\left(-(k+1)(\log(k+1) - 1)\right).$$

Proof. By Theorem 3.3.3 in Wu and Yang (2020), there exists G_1, G_{-1} that matches the first k moments and

$$\chi^2\left(G_1 \star \mathcal{N}(0,1), G_{-1} \star \mathcal{N}(0,1)\right) \leq C\exp\left(-(k+1)(\log(k+1) - 1)\right).$$

We can maximize $\mathbb{E}_{G_1}[f(\mu)] - \mathbb{E}_{G_{-1}}[f(\mu)]$ over moment-matching distributions G_1, G_{-1}. The dual of this linear program is the problem of uniform polynomial approximation. By (2.9) in Wu and Yang (2020), the optimal value of this problem is of the form $cE_k(f)$, where $E_k(f)$ is the approximation error. □

Theorem C.3. *A function $f : [-1,1] \to \mathbb{R}$ is analytic if and only if*

$$\limsup_{k\to\infty} E_k(f)^{1/k} < 1.$$

Proof. This is Theorem 8.1 in DeVore and Lorentz (1993). □