

Regression Calibration as a Baseline for Machine-Learned Covariates

October 2026

Abstract

Researchers increasingly use machine-learning predictions as covariates, including controls for confounding. Prediction error can then change the estimated effect of interest. Regression calibration offers an established baseline: use a labeled sample to predict the true covariate from its machine-generated surrogate and the other regression variables. We explain when this procedure identifies a linear model and what can go wrong when calibration must itself be learned. A population projection can suffice without a correct conditional-mean model, but estimated calibration must preserve the residual relationships that identify the downstream coefficients. We connect departures from those relationships to coefficient error, distinguish many calibration predictors from many mismeasured covariates, and separate estimation error from failures of support and transport. Exact examples show how omitting treatment or imposing shrinkage changes confounding adjustment; simulations illustrate these mechanisms. The resulting framework makes regression calibration a transparent benchmark for evaluating methods that use machine-learned covariates.

1 Introduction

A researcher uses a model to infer a confounder from text, images, or administrative records, then controls for that prediction in a regression. A good prediction need not be a good control. Its errors can leave part of the confounder associated with treatment, changing the estimated treatment effect. The practical question is which correction is justified by the labeled data and the downstream model.

Regression calibration deserves a place among the baselines. It replaces an error-prone covariate with a prediction of its true value conditional on available information. This is an established measurement-error method, not a new solution to machine-learning error ([Carroll and Stefanski, 1990](#); [Boe et al., 2023](#)). For a linear outcome model, a suitable calibration residual is orthogonal to the regressors used downstream. That fact explains both the attraction of calibration and its requirements. A machine-learning prediction may already satisfy the requirement; an additional calibration step is useful when it does not.

Work on machine-generated regressors has developed corrections tailored to the information available in validation samples ([Fong and Tyler, 2021](#); [Battaglia et al., 2025](#)). Prediction-powered inference and design-based approaches use labeled observations to correct downstream estimation ([Angelopoulos et al., 2023](#); [Egami et al., 2023](#)); [Kluger et al. \(2025\)](#) address imputed covariates

and nonuniform sampling explicitly. These methods need not share calibration’s outcome exclusion restriction or its target. A procedure for a population regression projection can remain meaningful when a structural linear model fails. Cross-prediction-powered inference concerns reuse of prediction-training observations and should not be conflated with a result specifically about covariate calibration (Zrnic and Candès, 2024).

The useful comparison begins with a distinction between population calibration and estimated calibration. At the population level, a linear projection onto a feature span containing the observed downstream regressors can be enough. It need not recover the full conditional mean. In practice, however, a labeled sample must estimate that projection or a more flexible function. Adding predictors can increase estimation variance; shrinkage can alter the required orthogonality; and observations outside the labeled sample’s support require extrapolation. If several covariates are mismeasured, their calibrated values may also lack the independent variation needed to identify their separate coefficients.

This note organizes these issues around the residual moments that determine downstream coefficient error. Its contribution is a formulation of the baseline and its failure modes for machine-learned covariates, with explicit implications for confounding adjustment. The calibration identity, two-stage uncertainty correction, and general concerns about predicted regressors are established (Boe et al., 2023; Haber et al., 2021). The examples clarify their consequences; the simulations illustrate them. They do not establish that calibration dominates competing correction methods.

2 What the baseline requires

Let $X \in \mathbb{R}^k$ contain the true, incompletely observed covariates. Let V contain an intercept and all fully observed downstream regressors, including treatment D when its coefficient is of interest. A predictor is trained independently and frozen. Write $H = (V, \hat{X}, A)$, where $\hat{X}$ denotes its output and A any additional calibration features. A labeled calibration sample $\mathcal{C}$ observes (X, H, Y) ; an independent target sample $\mathcal{U}$ observes (H, Y) . Calibration itself does not require Y in $\mathcal{C}$, but validation-only outcome regression does.

Assume the target population satisfies $Y = V^\top \beta + X^\top \gamma + \varepsilon$ and $\mathbb{E}_U[\varepsilon \mid H] = 0$. The latter restricts information carried by the prediction and calibration features, not just the true covariates. Standard outcome exogeneity $\mathbb{E}[\varepsilon \mid X, V] = 0$ alone does not imply it. A sufficient stronger condition is $\mathbb{E}[\varepsilon \mid X, H] = 0$. Text or image features that predict outcomes through omitted pathways can violate this restriction. Calibration does not resolve unmeasured confounding. A causal interpretation additionally requires a justified outcome model and treatment assignment assumptions; otherwise the coefficients require a noncausal interpretation.

For a candidate calibration $m(H)$, define $R_m = (V^\top, m(H)^\top)^\top$, $\theta = (\beta^\top, \gamma^\top)^\top$, and $Q_m = \mathbb{E}_U[R_m R_m^\top]$. Assume finite second moments and nonsingular Q_m . Then

$$\mathbb{E}_U\{R_m(Y - R_m^\top \theta)\} = \mathbb{E}_U\{R_m(X - m)^\top\} \gamma. \quad (1)$$

Thus the sufficient calibration condition is $\mathbb{E}_U[R_m(X - m)^\top] = 0$. It is a condition on residual relationships with the entire downstream design. The choice $m_0(H) = \mathbb{E}_U[X \mid H]$ satisfies it by

iterated expectations. The residual is conditionally mean zero, not necessarily independent of H .

A second sufficient choice is the population linear projection of each component of X on a finite feature vector $B(H)$ whose span contains every component of V . Because m also lies in this span, projection residuals are orthogonal to both V and m . The conditional mean can be nonlinear and the result still holds. This observation makes low-dimensional linear calibration a particularly useful baseline. It also connects this construction to the familiar projection logic of linear instrumental-variable estimation; it does not introduce a new identification argument. As always, the calibrated downstream design must have full rank.

For the baseline, calibration and target observations are independent draws from the same population, conditional on the frozen predictor. With external calibration data, applying $\mathbb{E}_C[X | H]$ requires its equality to $\mathbb{E}_U[X | H]$ on target support and enough labeled support to learn it. A restricted linear projection is more sensitive: its coefficients depend on the distribution of H , so conditional-mean stability alone does not transport a misspecified projection. Appendix A gives a numerical example.

3 Learning calibration changes the problem

Conditional on the calibration sample, let $\hat{m}$ be fixed and let $\theta_U(\hat{m})$ be the coefficient from an infinite target population. Equation (1) gives the exact discrepancy

$$\theta_U(\hat{m}) - \theta = Q_m^{-1} \mathbb{E}_U \{ R_{\hat{m}}(X - \hat{m})^\top \} \gamma. \quad (2)$$

This separates three quantities: calibration residual moments, the outcome relevance of the mis-measured covariates, and the inverse downstream Gram matrix. Prediction MSE alone does not determine coefficient error. Residual error aligned with treatment can matter more than larger error in an irrelevant direction. A nearly singular Gram matrix can amplify small residual moments. The target-sample estimator adds sampling error around this population coefficient.

If $\hat{m}$ converges in target L_2 to a valid calibration m , with fixed dimension and a nonsingular limiting Gram matrix, the plug-in is consistent under the usual moment and sampling conditions. That is not a guarantee that an arbitrary flexible learner has sufficiently small error for ordinary confidence intervals. Calibration error generally enters at first order. Sample splitting prevents data reuse but does not make this plug-in score Neyman-orthogonal; orthogonal scores and cross-fitting address distinct issues (Chernozhukov et al., 2018).

Which variables must be included? Consider independent Gaussian innovations with $X \sim N(0, 2)$, $D = X/2 + \eta$, $\eta \sim N(0, 1/2)$, a raw surrogate $W = X + u$, $u \sim N(0, 1)$, and $Y = D/2 + X + \varepsilon$, $\varepsilon \sim N(0, 1)$. Here X is a pretreatment confounder and the treatment effect is $1/2$. Calibration using only W gives $\mathbb{E}[X | W] = 2W/3$. Regressing Y on D and this calibrated value yields treatment coefficient 1, not $1/2$. Including D gives $\mathbb{E}[X | W, D] = (W + D)/2$ and recovers the effect. The issue is compatibility with the downstream regression, even when the smaller calibration is an exact conditional expectation.

Many predictors. Adding calibration features increases the dimension of the function to be learned. Unpenalized calibration can become unstable or nonunique. Regularization addresses that fitting problem by changing the projection. Even leaving D unpenalized does not generally preserve the desired downstream coefficient. In the Gaussian example, residualize W and X on $(1, D)$, standardize the residual of W , and minimize mean squared prediction error plus λ times the squared coefficient. The resulting population calibration is $m_\lambda = (1 - b)D + bW$, where $b = 1/[2(1 + \lambda)]$. The downstream treatment coefficient is $1/2 - \lambda$. This is regularization bias, not a defect that vanishes with a larger target sample. Appendix A derives the result. Tuning or a shrinking penalty can change this tradeoff; no fixed penalty is uniformly appropriate.

Many mismeasured covariates. With k components of X , the calibrated components must contain k independent directions after removing V . For linear calibration on a finite feature span, the residualized feature span must therefore have dimension at least k . Several accurate predictions can still identify only a common factor. This is different from fitting too many predictors: more labels cannot repair a population rank deficiency. Counting raw proxies is insufficient for nonlinear calibration, since distinct nonlinear functions of one proxy can span several directions.

Support and interpolation. Neither interpolation among observed labels nor a small within-sample error guarantees accuracy elsewhere. If a target region has no labeled support, different calibration functions can agree on every possible labeled observation yet differ in that region. Appendix A gives a two-world construction. It concerns learning calibration from labeled covariates and target predictors; it does not rule out identification from an additional outcome model using target Y . Checking marginal feature ranges can reveal obvious gaps, but does not establish joint overlap in many dimensions. More labels in the same restricted region do not solve this problem.

4 Illustration and inference

The simulation uses the Gaussian confounding model above. A ridge prediction of X from W is trained once on 1,000 independent observations and frozen. Its affine form deliberately isolates the calibration problem. We compare target oracle OLS, naive substitution, validation-only OLS, small OLS calibration using $(1, D, \hat{X})$, expanded OLS calibration, and expanded ridge calibration. Expanded calibrations add $q \in \{0, 20, 60, 100\}$ independent standard-normal noise predictors. Ridge leaves $(1, D)$ unpenalized and uses the standardized residual-feature penalty $\lambda = 0.1$. This is a specified shrinkage example, not an optimized ridge benchmark.

There are 1,000 target observations, either 80 or 300 calibration observations, and 400 replications per cell. Methods share data within each replication. Figure 1 displays median absolute treatment-coefficient error and its interquartile range across replications. Table A1 gives median signed error, descriptive run RMSE, calibration MSE, and failure counts for the endpoints. Adding noise features isolates estimation burden; all samples still have the same population support. At $n_c = 80, q = 100$, expanded OLS has no unique calibration coefficient and is reported as a failure. At $n_c = 300$, small calibration has median absolute error 0.065, compared with 0.057 for validation-only OLS;

expanded OLS reaches 0.341 with 100 noise features. Thus calibration corrects much of the naive error here, but using the larger target sample does not automatically improve on directly analyzing the labeled sample. The simulation does not establish which features an application should omit or how nonlinear text and image learners perform.

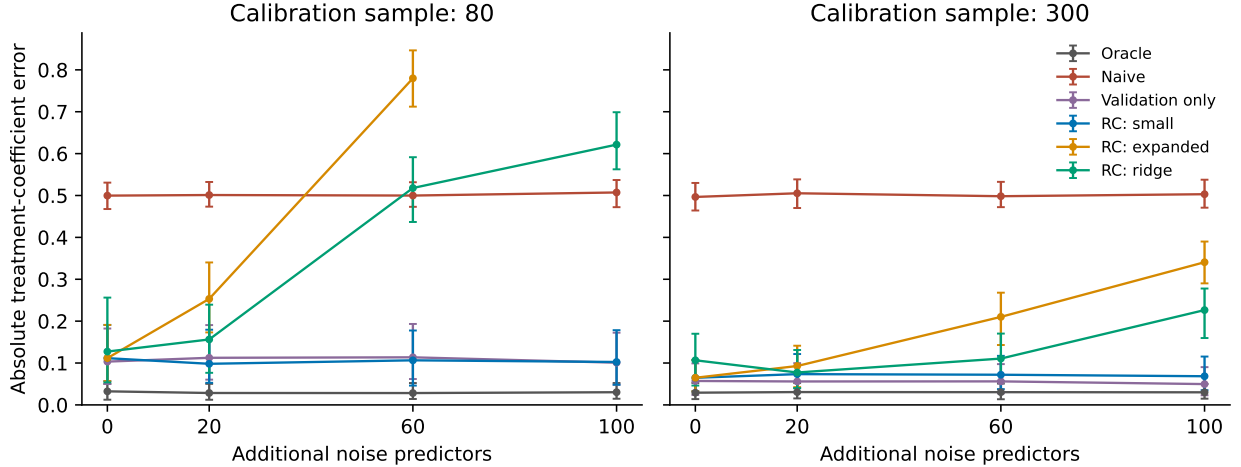


Figure 1: Treatment-coefficient error as calibration gains irrelevant predictors. Points show median absolute error; bars show the 25th–75th percentiles across replications, not confidence intervals or Monte Carlo standard errors. The expanded OLS curve stops where its calibration fit is nonunique. The oracle observes target labels and is an infeasible reference.

For fixed-dimensional, regular calibration fits, inference must account for both samples. Resampling calibration and target observations independently and refitting both stages is an established solution (Boe et al., 2023). The separate coverage experiment in Table A2 compares that procedure with second-stage HC1 intervals for the small OLS calibration. It uses 400 replications and 399 bootstrap resamples per interval. Its conclusions are conditional on the frozen prediction model. A random training stage would require a corresponding uncertainty analysis. The experiment does not validate bootstrap inference after arbitrary regularization, model selection, or weak identification.

5 Implications for practice

Regression calibration offers a transparent starting point for machine-learned covariates when the target is a justified linear model. Start with all observed downstream regressors in the calibration span, verify that the calibrated covariates retain independent variation, and account for the labeled sample’s estimation error. Assess target support separately from prediction fit. When the target is instead a population regression projection without the required exclusion restriction, or validation sampling supplies a different route to identification, moment-correction and design-based methods address a different problem.

For nonlinear outcome means, replacing a covariate by its conditional mean is generally an approximation; this limitation is also established (Carroll and Stefanski, 1990). It is not a universal impossibility result. Logistic curvature changes sign, so Jensen’s inequality does not give a global

direction of bias. With a log mean and a conditionally Gaussian calibration residual of constant variance, integration can change only the intercept while preserving slopes (Wang, 2012). The relevant comparison is therefore between assumptions, targets, and available information, rather than a blanket division between linear models that need calibration and nonlinear models that need something else.

The baseline’s value is that its requirements are visible. It forces a proposed improvement to explain whether it reduces calibration error, relaxes an outcome restriction, transports information more credibly, or improves efficiency for the same estimand.

Replication and assistance. Code, generated results, and the simulation specification are available at <https://github.com/finite-sample/calibration-baseline>. The specification was written during revision after inspecting the previous design; it is not a preregistration. AI tools assisted with the literature audit, manuscript revision, code, and checks.

References

- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671):669–674.
- Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025). Inference for regression with variables generated by AI or machine learning. arXiv:2402.15585v5. First circulated in 2024.
- Boe, L. A., Shaw, P. A., Midthune, D., Gustafson, P., Kipnis, V., Park, E., Sotres-Alvarez, D., and Freedman, L. (2023). Issues in implementing regression calibration analyses. *American Journal of Epidemiology*, 192(8):1406–1414. On behalf of the Measurement Error and Misclassification Topic Group (TG4) of the STRATOS Initiative.
- Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85(411):652–663.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.
- Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601.
- Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.
- Haber, G., Sampson, J., and Graubard, B. (2021). Bias due to Berkson error: issues when using predicted values in place of observed covariates. *Biostatistics*, 22(4):858–872.
- Kluger, D. M., Lu, K., Zrnic, T., Wang, S., and Bates, S. (2025). Prediction-powered inference with imputed covariates and nonuniform sampling. arXiv:2501.18577v3.
- Wang, C. Y. (2012). Robust best linear estimation for regression analysis using surrogate and instrumental variables. *Biostatistics*, 13(2):326–340.
- Zrnic, T. and Candès, E. J. (2024). Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121.

A Derivations and simulation details

A.1 Calibration and coefficient error

Substitute $Y = R_m^\top \theta + (X - m)^\top \gamma + \varepsilon$ into the population OLS normal equations. Since R_m is measurable with respect to H , exogeneity gives $\mathbb{E}_U[R_m \varepsilon] = 0$. Therefore

$$Q_m \{\theta_U(m) - \theta\} = \mathbb{E}_U[R_m(X - m)^\top] \gamma. \quad (\text{A1})$$

This proves both identities in the text. A conditional mean makes the matrix on the right zero. For a population projection onto B , each residual is orthogonal to every component of B . Since both V and m lie in that span, it is also zero. Neither argument guarantees nonsingularity of Q_m .

For consistency, condition on the calibration sample and use independence of target observations. Target L_2 convergence of $\hat{m}$ to m implies convergence of $\mathbb{E}_U[R_{\hat{m}} R_{\hat{m}}^\top]$ and $\mathbb{E}_U[R_{\hat{m}} Y]$ by Cauchy–Schwarz, provided Y has a finite second moment. Uniformly controlled fourth moments of the fitted regressors and Y suffice for the corresponding triangular-array sample moments to converge as $n_u \rightarrow \infty$. Continuity of matrix inversion at nonsingular Q_m gives consistency. This fixed-dimensional argument is a sufficient condition, not a rate theorem for high-dimensional calibration.

For scalar X and a perturbation $m_t = m_0 + th$, where $m_0 = \mathbb{E}_U[X \mid H]$, the derivative of the residual moment at zero is

$$\left. \frac{d}{dt} \mathbb{E}_U[R_{m_t}(X - m_t)] \right|_{t=0} = -\mathbb{E}_U[R_{m_0} h]. \quad (\text{A2})$$

The derivative from changing R vanishes by conditional-mean orthogonality; the displayed term generally does not. This establishes the first-order sensitivity of the ordinary plug-in.

A.2 Omitted treatment and ridge shrinkage

In the example, $\text{Var}(D) = 1$, $\text{Var}(X) = 2$, $\text{Var}(W) = 3$, and every covariance involving D and either X or W equals 1; $\text{Cov}(X, W) = 2$. The population regression of Y on (D, W) has slopes $(1, 1/2)$. Rescaling W to $2W/3$ leaves the coefficient on D equal to 1. Replacing W by $(D + W)/2$ instead gives slopes $(1/2, 1)$.

Write $X = D + v$ using the Gaussian conditional representation, where v has variance 1 and is independent of D . Residualizing W on D leaves $v + u$, with variance 2. The standardized predictor is $(v + u)/\sqrt{2}$. Penalized population regression of v on this predictor has coefficient $1/[\sqrt{2}(1 + \lambda)]$. Hence $m_\lambda = (1 - b)D + bW$, $b = 1/[2(1 + \lambda)]$. Substituting $W = (m_\lambda - (1 - b)D)/b$ in the population projection $Y = D + W/2 + \text{residual}$ gives

$$\beta_D(\lambda) = 1 - \frac{1 - b}{2b} = \frac{1}{2} - \lambda, \quad \gamma(\lambda) = \frac{1}{2b} = 1 + \lambda. \quad (\text{A3})$$

An invertible affine prediction of W does not change these arguments when the intercept is included and residual predictors are standardized.

A.3 Transport of a restricted projection

Let $V = 1$ and $X = H^2$, with $H \sim \text{Uniform}[0, 1]$ in calibration and density $2h$ on the same interval in the target population. Both populations have exactly the same conditional mean and support. In calibration, the linear projection of X on $(1, H)$ is $m_C = H - 1/6$. In the target population, $\mathbb{E}_U[X] = 1/2$ and $\mathbb{E}_U[m_C] = 2/3 - 1/6 = 1/2$, but

$$\mathbb{E}_U[m_C(X - m_C)] = \frac{1}{90} \neq 0. \quad (\text{A4})$$

Indeed, $\text{Cov}_U(H, H^2) = 1/15$ and $\text{Var}_U(H) = 1/18$. If $Y = X$, target OLS of Y on $(1, m_C)$ has slope $6/5$, rather than 1. Estimating the target conditional mean H^2 or a target-weighted linear projection repairs this particular example. Stable conditional means and common support alone do not transport a restricted projection's orthogonality.

A.4 A support gap that more labels cannot close

Suppose labeled H is uniform on $[0, 1]$ and target H is uniform on $[0, 2]$. Consider $X = m_j(H)$, with $m_0(h) = h$ and $m_1(h) = h + (h - 1)_+$. The two worlds give identical distributions for calibration labels and target predictors. For any predictor $a(h)$ learned solely from these inputs,

$$\frac{(a - m_0)^2 + (a - m_1)^2}{2} \geq \frac{(m_1 - m_0)^2}{4}. \quad (\text{A5})$$

Integrating over target H gives average risk at least $1/24$, so at least one world has target MSE at least $1/24$, even with infinitely many calibration observations. The inequality also holds after averaging over any fitting randomness. Both candidate means are continuous. Additional structural restrictions, labels outside the original support, or an informative outcome model can change the conclusion. This example establishes a limit on learning calibration from the stated inputs, not a general impossibility of downstream inference.

A.5 Simulation specification and reporting

The frozen learner is scikit-learn ridge with penalty 1 in its sum-of-squares convention and an intercept, trained on 1,000 observations. The calibration ridge penalty is instead 0.1 in a mean-squared-loss convention, with intercept and treatment unpenalized. Calibration-only residualization and scaling are applied to target features; no target labels enter a feasible estimator. Independent pseudo-random streams index each sample-size, dimension, and replication cell. The generator, seeds, versions, and frozen prediction coefficients are recorded in `results/metadata.json`; replicate-level results are in `results/replications.jsonl`.

The primary summaries are the median signed error, median absolute error, and absolute-error quartiles. Run RMSE is the square root of the average squared error among the realized replications; it is descriptive, not an estimate accompanied by a normal-theory Monte Carlo interval. All methods share samples within a replication. Coverage MC SE is $\sqrt{\hat{p}(1 - \hat{p})/R}$. Unexpected errors stop execution; rank failures are counted without substituting coefficients.

This reporting choice matters even in the simple Gaussian design. Write small fitted calibration as $\hat{m} = \hat{a}_0 + \hat{a}D + \hat{b}W$, and target OLS of Y on (D, W) as having slopes $(\hat{d}, \hat{c})$. The calibrated treatment estimate is $\hat{d} - \hat{c}\hat{a}/\hat{b}$. Conditional on a full-rank calibration design, $(\hat{a}, \hat{b})$ is nondegenerate normal. Its density at $\hat{b} = 0$ is positive, and the conditional second moment of $\hat{a}$ there is positive. Target data are independent, with $\hat{c} \neq 0$ almost surely. The resulting ratio has infinite second moment. Rare draws near zero can dominate squared error even though none appears in a modest simulation. Fixed-dimensional consistency and local asymptotic normality at a nonzero population slope do not imply finite moments at every sample size. The initial replacement design used RMSE error bars; mathematical review after its first run prompted the change to quantiles recorded in `analysis_plan.json`. No observations or replications were trimmed.

Table A1: Calibration complexity and treatment-coefficient error.

n_c	q	Method	Median error	Run RMSE	X MSE	Failed
80	0	Oracle	-0.001	0.044	0.000	0
80	0	Naive	0.500	0.500	0.668	0
80	0	Validation only	-0.003	0.160	—	0
80	0	RC: small	-0.001	0.173	0.523	0
80	0	RC: ridge	-0.104	0.229	0.526	0
80	100	RC: small	0.001	0.163	0.520	0
80	100	RC: expanded	—	—	—	400
80	100	RC: ridge	0.621	0.634	1.240	0
300	0	Oracle	-0.003	0.044	0.000	0
300	0	Naive	0.497	0.501	0.668	0
300	0	Validation only	-0.004	0.086	—	0
300	0	RC: small	0.001	0.092	0.505	0
300	0	RC: ridge	-0.101	0.142	0.508	0
300	100	RC: small	0.003	0.097	0.505	0
300	100	RC: expanded	0.341	0.349	0.761	0
300	100	RC: ridge	0.226	0.239	0.691	0

Each cell uses 400 independent replications. The full grid is in the replication files. At $q = 0$, expanded OLS equals small RC. q counts added calibration predictors; the downstream regression always contains an intercept, treatment, and one true, predicted, or calibrated covariate. X MSE is evaluated on target observations. Validation-only OLS makes no imputation. Failed denotes nonunique OLS fits; no pseudoinverse coefficient is substituted. Oracle OLS uses target labels unavailable to the feasible methods. Run RMSE is descriptive: small OLS calibration has infinite finite-sample population MSE in this design.

Table A2: Uncertainty from estimating the small calibration model.

n_c	Interval	Coverage	MC SE	Mean width
80	Second stage HC1	0.580	0.025	0.265
80	Two-sample bootstrap	0.932	0.013	0.635
300	Second stage HC1	0.825	0.019	0.263
300	Two-sample bootstrap	0.940	0.012	0.384

Nominal coverage is 0.95; 400 independent replications and 399 resamples per bootstrap interval. Both samples are independently resampled and both regressions refit. Second-stage HC1 treats estimated calibration as fixed. No fits failed. These results concern only the fixed, three-parameter OLS calibration.