

Which Prediction Errors Matter?

Finite Covariate-Correction Paths for Downstream Regression

October 2026

Abstract

Correcting errors in machine-learned covariates can change a treatment-effect estimate in ways that separate corrections do not predict. Each correction changes the regression’s adjustment for other variables, so its effect can depend on which errors have already been corrected. Existing methods use verified measurements to correct inference and guide validation. We study the joint effects of finite corrections on a fitted coefficient. Our diagnostic replaces sets of proxy covariates with verified values and compares the resulting refits with the sum of isolated correction effects. Exact least-squares identities locate the source of nonadditivity in changes to the regression’s cross-product matrix. Correction paths describe how the coefficient moves as more measurements are corrected, and their dispersion describes how much the choice of correction set matters. Established Shapley methods attribute the endpoint change. Simulations illustrate these distinctions. Applications to housing and text-coded incivility compare refits with three additive approximations: isolated corrections track small random correction sets closely but misstate the full correction. The resulting diagnostic measures both the consequences of individual errors and the additional changes produced by correcting them together.

1 Introduction

Correcting a covariate changes the regression’s adjustment for confounding. Consequently, the effect of correcting one observation can depend on which other observations have already been corrected. This matters when researchers use machine-learning predictions as covariates: an assessment of errors one at a time can misrepresent the change in a treatment coefficient when those errors are corrected together. We study these joint correction effects and develop a diagnostic that measures them through regression refits.

Predicted covariates allow researchers to incorporate information that would otherwise be expensive to measure, including attributes coded from text and variables missing from linked datasets. Their errors can distort coefficient estimates and conventional standard errors (Fong and Tyler, 2021; Battaglia et al., 2024). Prediction accuracy does not summarize that distortion because it does not record how errors relate to treatment, outcomes, and the other regressors. When a predicted variable serves as a confounder, its measurement quality is also distinct from the assumptions that identify a causal effect (Egami et al., 2022). The practical question is how the observed errors change the coefficient a researcher wants to interpret.

Our diagnostic compares the effect of correcting a set of measurements with the sum of the effects of correcting each measurement separately. All observations remain in the regression, and the outcome and specification stay fixed. Only the selected proxy covariates are replaced with verified

values. The discrepancy between the joint and summed effects measures nonadditivity for that set. Repeating the comparison across sets reveals which corrections have effects that depend on the surrounding measurements. A correction path summarizes the coefficient as the number of corrected measurements increases; its dispersion records how different choices at the same correction budget produce different results.

For ordinary least squares, the source of this dependence is explicit. Disjoint corrections have additive effects on the estimating equation evaluated at the proxy fit, but they also change the cross-product matrix that maps those changes into coefficient movements. Its inverse depends on the whole correction set. Thus even exact individual refits need not add up to the joint refit. An algebraic example gives a jointly corrected coefficient of $1/2$ where summing the isolated effects predicts zero (Appendix B). This is a property of finite regression corrections, established without simulation. Numerical examples then illustrate how the discrepancy varies with error magnitude and its relation to treatment.

Verified measurements already support methods for correcting inference with predicted variables. These include methods for predicted outcomes, prediction-powered inference and its efficiency refinements, and design-based supervised learning (Wang et al., 2020; Angelopoulos et al., 2023a,b; Egami et al., 2023). Extensions explicitly accommodate imputed covariates and nonuniform verification samples (Kluger et al., 2025). These procedures address estimation of a population parameter. We examine how finite replacements change a specified fitted coefficient, conditional on the available verified measurements. A partially corrected fit is a diagnostic comparison; it does not inherit the inferential properties of a bias-corrected estimator.

Target-specific verification also has a substantial literature. Two-phase sampling allocates expensive measurements using information collected on a larger sample (Neyman, 1938; Amorim et al., 2021). For regression and causal parameters, allocation may depend on influence functions or on residual influence after using auxiliary information (Chen and Lumley, 2022; Wang et al., 2023). Active statistical inference projects gold-minus-proxy score discrepancies through an inverse Hessian, while adaptive multiwave M-estimation uses influence contrasts to guide subsequent verification (Zrnic and Candès, 2024; Kluger and Bates, 2026). Those methods already ask which errors matter for a target. Our comparison asks how well additive summaries describe the consequences of correcting specific sets. A discrepancy from a fixed influence approximation does not establish a failure of an adaptive procedure that updates its scores or uses different gold and proxy fits.

Local influence analysis studies how measurement error affects regression estimates (Lee and Zhao, 1996). Simulation-extrapolation adds noise and extrapolates toward the error-free case (Cook and Stefanski, 1994). Infinitesimal-jackknife methods provide refit approximations and error bounds, including higher-order expansions (Giordano et al., 2019a,b). Influence functions and Shapley values also explain predictions and value training observations (Koh and Liang, 2017; Ghorbani and Zou, 2019; Karlaš et al., 2024). We apply baseline-replacement Shapley values to a regression coefficient (Sundararajan and Najmi, 2020). The mean path under independent random corrections is Owen’s multilinear extension evaluated at a common inclusion probability (Owen, 1972). These tools summarize and attribute finite coefficient changes; the regression diagnostic supplies their interpretation here.

The analysis separates three quantities: the endpoint coefficient change, nonadditivity of corrections, and dispersion across correction sets. They answer different questions. In particular, a straight mean path need not imply that the choice of corrections is inconsequential. We report all three and compare full refits with isolated corrections and two influence approximations. The numerical examples hold prediction RMSE fixed across error patterns to illustrate how error alignment and

concentration affect these quantities. Two replication archives then show how the comparisons behave with actual predictions and reference measurements. In housing, we correct adjustment covariates; in a text application, we correct the regressor of interest. Both applications concern descriptive associations, and neither evaluates a prospective sampling policy.

2 A game of verified replacements

Let X be the proxy design matrix, including an intercept, and y the outcome. Let $G_1, \dots, G_m$ be disjoint sets of rows with verified design values. A player is one such correction unit; it can be a single observation or a batch. For $S \subseteq \{1, \dots, m\}$, replace precisely the rows belonging to players in S , giving X_S . Define the value of that correction set as

$$v(S) = e_t^\top (X_S^\top X_S)^{-1} X_S^\top y, \quad (1)$$

where t indexes the coefficient of interest. Every design used in this game must have full column rank. We use least-squares refits and report a failure if a coalition is rank deficient; a pseudoinverse would silently redefine the target.

The baseline $v(\emptyset)$ is the proxy fit. The endpoint $v(M)$ corrects the declared pool M , not any unverified observations outside it. Both are finite-sample regression coefficients. Even if the gold regression identifies a causal parameter in the population, their difference is not itself an estimate of population measurement-error bias without additional assumptions.

For a given correction set, define its departure from isolated corrections as

$$R(S) = v(S) - v(\emptyset) - \sum_{i \in S} \{v(\{i\}) - v(\emptyset)\}. \quad (2)$$

This discrepancy is in coefficient units. The correction game is additive exactly when $R(S) = 0$ for every set S . A zero discrepancy for one set alone does not establish additivity, since interactions within it can cancel.

2.1 Mean paths and correction-set dispersion

Independently include player i with probability p_i . The expected coefficient is the multilinear extension of the finite game (Owen, 1972):

$$F(\mathbf{p}) = \sum_{S \subseteq M} v(S) \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i). \quad (3)$$

The common-probability path is $\phi(p) = F(p, \dots, p)$. Let S_k be uniform over all subsets of exactly k players and $\mu_k = \mathbb{E}[v(S_k)]$. Then

$$\phi(p) = \sum_{k=0}^m \binom{m}{k} p^k (1-p)^{m-k} \mu_k. \quad (4)$$

The Bernoulli design has a random correction count; the fixed-budget design does not. Their variances differ because the former also varies that count. We report both, and store the entire finite distribution at each fixed budget. These are distributions over correction sets conditional on the data, not sampling distributions of a population estimator.

If $v(S) = v(\emptyset) + \sum_{i \in S} a_i$, then both mean paths interpolate between the endpoints. Curvature therefore rules out additivity, but a straight path does not establish it: average interactions can cancel. A straight path also need not have small dispersion. Accordingly, we report errors of additive approximations at the coalition level as well as mean curvature and fixed-budget dispersion.

2.2 Attribution along the path

Let $S_{-i}(p)$ independently include every player other than i with probability p , and define

$$d_i(p) = \mathbb{E}[v(S_{-i}(p) \cup \{i\}) - v(S_{-i}(p))]. \quad (5)$$

The derivative satisfies $\phi'(p) = \sum_i d_i(p)$. The Shapley attribution is

$$s_i = \int_0^1 d_i(p) dp \quad (6)$$

$$= \frac{1}{m} \sum_{k=0}^{m-1} \mathbb{E}[v(S \cup \{i\}) - v(S) \mid |S| = k, i \notin S], \quad (7)$$

so $\sum_i s_i = v(M) - v(\emptyset)$. This integral representation is established (Owen, 1972; Okhrati and Lipani, 2021). Uniformly averaging over all subsets of other players instead gives $d_i(1/2)$, the unnormalized Banzhaf attribution, which generally does not sum to the endpoint change. The implementation distinguishes the two; its tests use a three-player unanimity game, for which the respective per-player values are $1/3$ and $1/4$.

The choice of game matters (Sundararajan and Najmi, 2020). These values attribute movement in a regression coefficient, not prediction accuracy or social welfare. Positive contributions are not inherently beneficial, since improvement depends on a separate target and loss. Batch-game values also depend on grouping and generally do not equal sums of individual-row Shapley values.

3 What an influence approximation omits

Write $\hat{\beta} = (X^\top X)^{-1} X^\top y$, $r = y - X\hat{\beta}$, and $A = X^\top X$. For a coalition's design change $\Delta = X_S - X$, define

$$B_S = X^\top \Delta + \Delta^\top X + \Delta^\top \Delta, \quad (8)$$

$$q_S = \Delta^\top r - X^\top \Delta \hat{\beta} - \Delta^\top \Delta \hat{\beta}. \quad (9)$$

Subtracting the old normal equations from the new ones gives the exact identity

$$\hat{\beta}_S - \hat{\beta} = (A + B_S)^{-1} q_S. \quad (10)$$

This is ordinary least-squares algebra, not a new influence-function result. For disjoint row corrections, q_S is the sum of the individual paired-score changes. Dependence of the inverse matrix on S can still make their coefficient effects nonadditive.

We compare the exact target change with three additive baselines on the same correction sets. The principal influence baseline is $e_t^\top A^{-1} q_S$: it retains the finite quadratic score term while holding the inverse Gram matrix fixed. A feature-amplitude derivative additionally drops $-\Delta^\top \Delta \hat{\beta}$. Finally, the

single-repair baseline sums $v(\{i\}) - v(\emptyset)$ over the chosen players. It allows each isolated repair to change the inverse matrix, but ignores interactions among repairs.

Neither influence approximation uniformly dominates the other. Both generally have second-order error as feature-error amplitude goes to zero. Retaining one quadratic term does not provide a complete second-order expansion. A large failure of the paired-score baseline does not prove that established active inference procedures fail: they target sampling efficiency and valid inference, and may update their fitted scores or use different gold and proxy influence functions. Our comparison addresses a fixed-baseline approximation to the finite raw-replacement game.

4 Numerical illustrations

To illustrate the algebraic mechanism, we simulate independent standard Normal variables U, W, η, ϵ and set

$$D = \mathbf{1}\{0.8U + 0.3W + \eta > 0\}, \quad Y = 0.5D + U + 0.3W + 0.5\epsilon. \quad (11)$$

The verified design is $(1, D, U, W)$, and the proxy replaces U by $U + e$. Conditional exogeneity and the additive outcome model identify the population treatment coefficient 0.5 using the verified covariates. The realized gold coefficient need not equal 0.5; our comparisons always use the realized fit.

The main example uses seed 2024, 96 observations, eight randomly assigned equal batches, and proxy RMSE 0.75. Within the same dataset we compare independent diffuse errors, errors supported on the first batch only, and errors proportional to centered treatment. Each error vector is rescaled to have the same RMSE. The one-active-batch scenario is deliberately an additive control. The aligned scenario is a stress test of differential measurement error, not a claim about its prevalence in trained ML systems. The benchmark prescribes errors rather than training a predictor. In the aligned case the proxy confounder adds a multiple of treatment, so the proxy and gold designs span the same space while assigning different coefficients to treatment. Partial row replacement changes that relation, making this a transparent constructed stress test.

We enumerate all 2^8 correction sets. Coalition approximation RMSE weights each set equally, equivalently using a Bernoulli design with $p = 1/2$. Mean curvature is the maximum absolute departure of $\phi(p)$ from endpoint interpolation on a 101-point grid. We also report the SD at a fixed half-pool budget. The repeated benchmark uses seeds 0–19 and error RMSE values 0.05, 0.25, 0.75, 1.5, holding each seed’s dataset, batches and error direction fixed as amplitude changes. Its 240 designs were specified in the analysis plan before this benchmark was run; the overall project remains exploratory.

Error pattern	Proxy $\hat{\tau}$	Gold $\hat{\tau}$	Curvature	SD at $k = 4$	IF RMSE
Diffuse	0.781	0.344	0.057	0.071	0.064
One active batch	0.728	0.344	0.000	0.192	0.134
Treatment-aligned	-1.123	0.344	0.263	0.045	0.927

Table 1: Main benchmark, with identical proxy RMSE across rows. Curvature is the maximum departure from endpoint interpolation. SD is over the 70 sets of four batches. IF RMSE compares the paired-score approximation with full refits over all 256 subsets. All quantities except RMSE of the proxy itself are in treatment coefficient units.

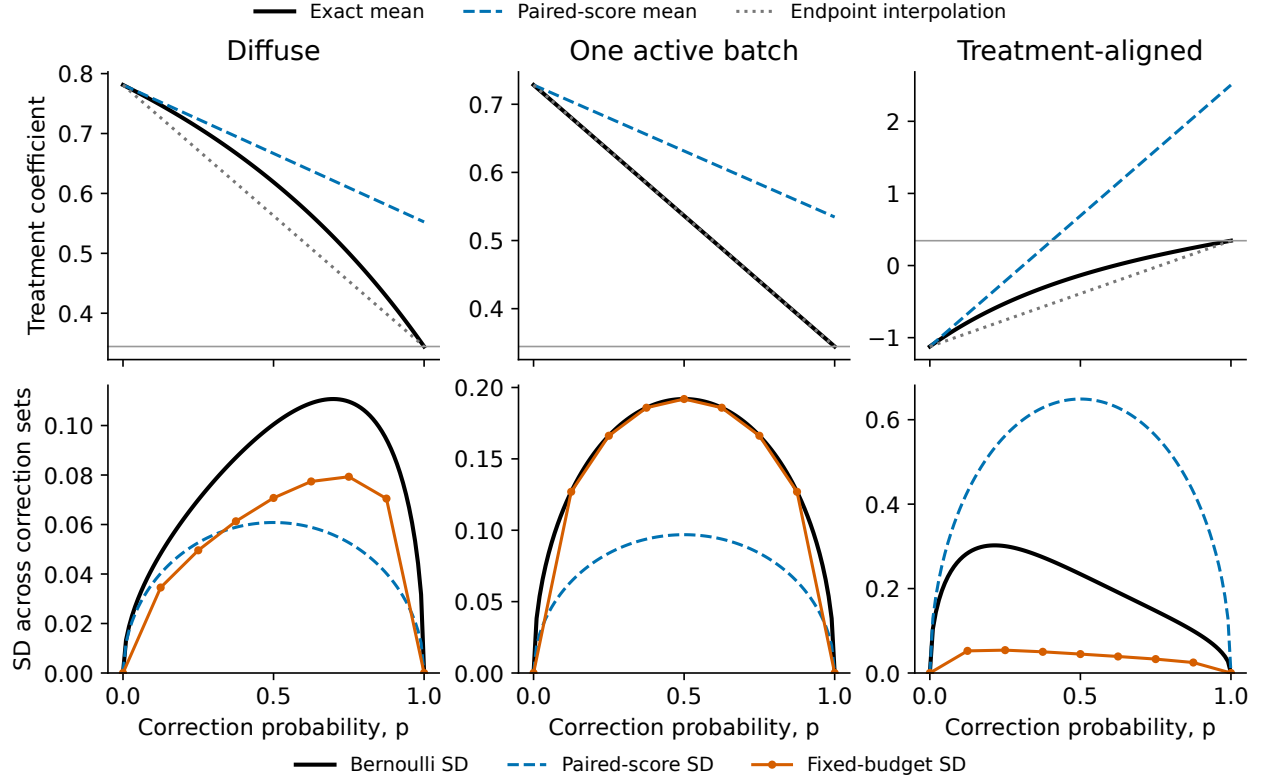


Figure 1: Mean correction paths (top) and dispersion (bottom). The exact Bernoulli path includes each batch independently with probability p . Fixed-budget SD uses exactly k batches, plotted at $k/8$. The horizontal gray line is the verified endpoint. Vertical scales vary by panel to show each case’s structure; numerical magnitudes appear in Table 1. Dispersion describes correction sets, not uncertainty about the causal parameter.

The aligned case has curvature 0.263 and paired-score approximation RMSE 0.927. The diffuse case also departs from additivity, with curvature 0.057. The one-active-batch control has zero curvature and its single-repair approximation is exact, yet its half-budget SD is 0.192. This control demonstrates why a straight path is insufficient evidence of stability. Identical predictive RMSE does not fix either the endpoint gap or the distribution of intermediate corrections.

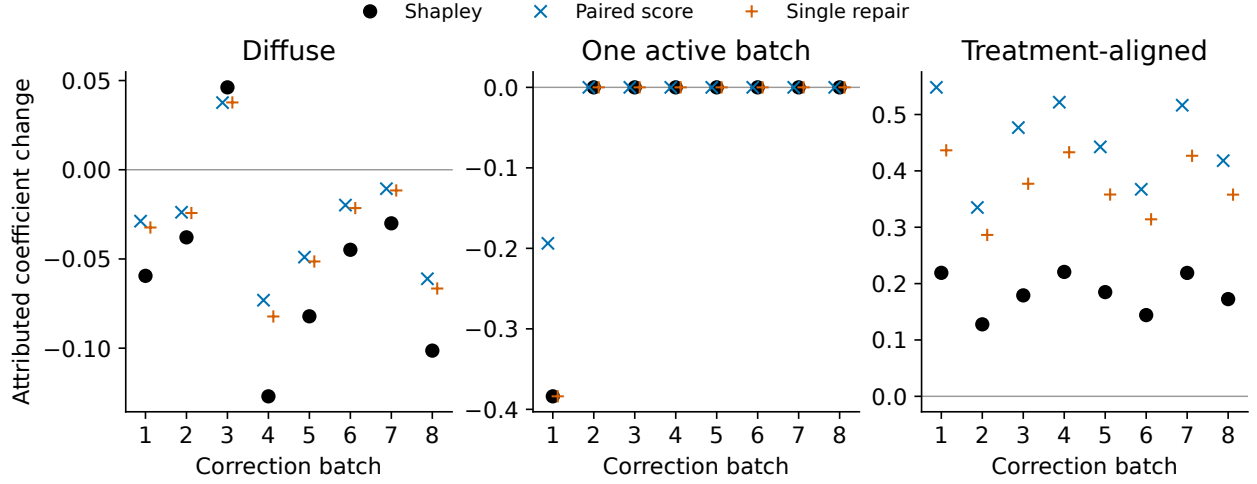


Figure 2: Attribution of the endpoint change across the eight declared correction batches. Shapley values average marginal changes over all correction orders. Paired scores and isolated repair changes are alternative additive summaries; neither is required to sum to the finite endpoint gap. Vertical scales vary.

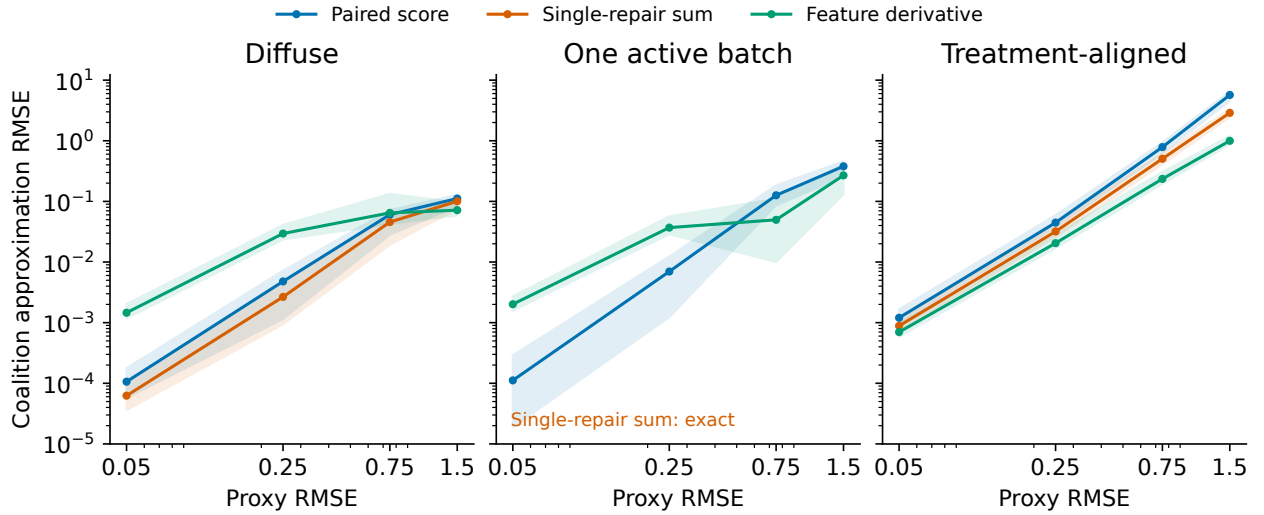


Figure 3: Repeated benchmark: median coalition approximation RMSE across 20 seeds, with 10th–90th percentile simulation ranges. These ranges are not confidence intervals. Both axes are logarithmic. The one-active-batch single-repair baseline is exact up to numerical error and is annotated instead of plotted at zero. All panels use the same vertical scale.

The repeated benchmark shows why a useful diagnostic must include the baselines rather than presume their failure. Approximation error generally shrinks with error amplitude, and the feature derivative can outperform the finite paired score. No single approximation is uniformly best across the prescribed designs. Every coalition in all 240 designs has full rank. The largest design-matrix condition number is 14.17; these findings therefore do not require near-singular matrices in this benchmark. This does not establish robustness to arbitrary conditioning or superiority in other data-generating processes.

Table 2: Full corrections in the paired archives.

	Housing	Reddit
Rows	46,418	613
Proxy coefficient	0.37184	-0.53414
Reference coefficient	0.37238	-0.76573
Actual correction	0.00054	-0.23159
Sum of isolated corrections	0.00249	-0.15550
Paired score	0.00249	-0.15222
Feature derivative	0.00831	-0.84858
$ R(N) /\text{reference HC0 scale}$	0.74	0.17

The three approximation rows predict the change from the proxy coefficient. Housing units are the archive’s housing-price outcome per standard deviation of income; Reddit units are comment-score points. $R(N)$ is the actual change minus the sum of isolated changes. The reference-fit HC0 scale supplies a descriptive normalization, not a population confidence interval; spatial or within-thread dependence is not accounted for.

5 Evidence from two replication archives

We examine two archives selected because they provide paired predictions and reference measurements. The first is the housing dataset distributed with [Kluger et al. \(2025\)](#): 46,418 observations with measured income and housing prices, and both predicted and reference values for nightlights and road length. We regress housing price on an intercept, income, nightlights, and road length, using the archive’s transformations. The target is the income coefficient per one standard deviation of income. A correction replaces both adjustment covariates for one row. Fixed centering and scaling improve numerical stability without changing the regression specification.

The second is the Reddit application of [Fong and Tyler \(2021\)](#). We reproduce the archive’s weighted support-vector classifier using its original 2,413-row training CSV and document-term matrix, then predict incivility for the separate 613-row validation sample. No validation outcomes or labels enter classifier training or tuning. In that validation sample, we regress comment score, clamped to $[-10, 10]$ as in the original application, on an intercept and incivility. Here a correction replaces the regressor of interest with its human label. This is a diagnostic of the validation sample, not a reproduction of the original paper’s primary-sample estimate. Both targets are descriptive associations; neither comparison identifies a causal effect.

Each row is a correction unit, including rows whose proxy already agrees with the reference. At seven budgets from 1% to 100% of rows, we compare direct least-squares refits with the same three additive approximations used above. For each interior budget, 200 random permutation prefixes give uniformly sampled correction sets; the full-correction endpoint is computed once. The comparisons therefore concern random corrections within these observed pools, not a policy for discovering previously unobserved errors.

Table 2 shows nonadditivity in both settings, with different magnitudes. In housing, summing isolated effects predicts a coefficient change 4.6 times the actual change. The ratio is large partly because the actual endpoint change is small; the table reports both in coefficient units. In Reddit, the isolated sum understates the magnitude of the endpoint change by 33%. The paired-score approximation is close to the isolated sum in both archives. Its discrepancy should not be confused with the much larger error of the feature derivative, which also omits finite-amplitude changes in

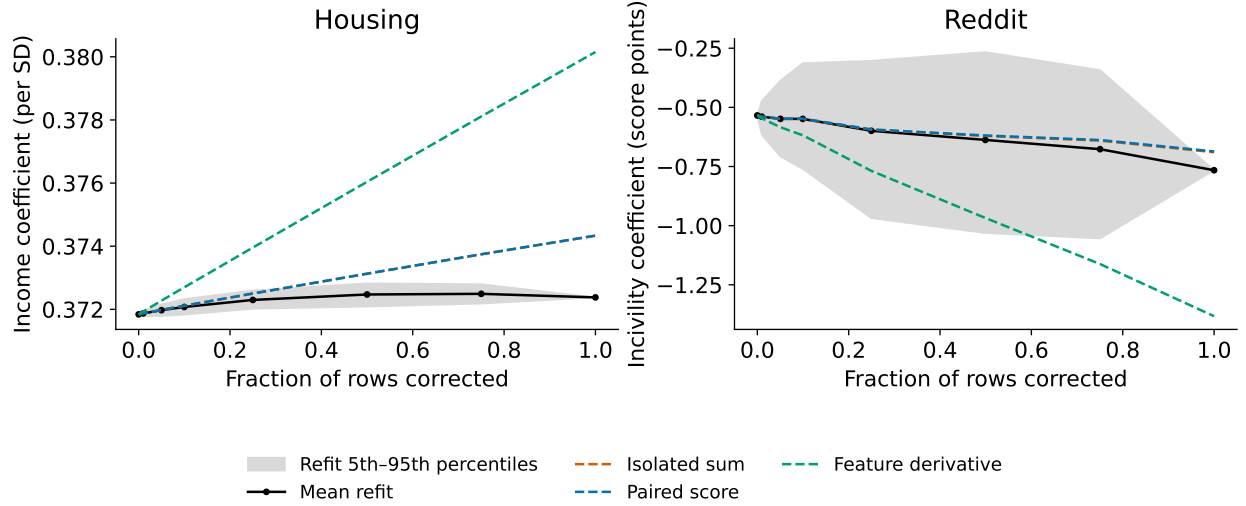


Figure 4: Correction paths in the paired archives. Points are mean refitted coefficients at the declared budgets. Shading shows the 5th–95th percentiles across sampled correction sets, not confidence intervals. Dashed lines show mean additive approximations on those same sets; isolated corrections and paired scores largely overlap. Lines connect the evaluated budgets.

the estimating equation.

Figure 4 places the endpoint comparisons in context. Isolated corrections and paired scores track refits closely for small random correction sets; their errors grow as corrections accumulate. Reddit also has substantial dispersion across correction sets despite relatively modest nonadditivity. The housing mean path flattens as the correction budget increases. Thus neither the endpoint gap nor the error of the feature derivative alone describes the diagnostic. These exploratory results show that the joint effects are measurable in actual prediction archives. They do not establish their prevalence across applications or that accounting for them changes a verification decision.

6 Discussion

With a fully verified benchmark or pilot pool, the method can reveal which correction budgets and sets leave a coefficient sensitive, whether influence approximations are adequate, and which declared units account for the endpoint change. This can inform hypotheses about a later validation design. It cannot by itself price the unseen errors of a new population. Learning an acquisition policy would require a separate pilot-to-target design, available sampling features, an objective such as estimation error or tail risk, and comparison with established influence-based allocation under a common label budget.

An exact counterexample in Appendix B makes the information limit concrete. Two possible completions of the unverified covariates give identical observed data whenever a random pilot misses two specific rows. One completion has no measurement error; the other has interacting corrections that an additive summary misrepresents. The probability of missing those rows is known from the sampling design. Thus an uneventful pilot alone cannot certify negligible interactions outside it. This is an indistinguishability argument, not a claim about how frequently such configurations occur in practice.

A prospective study must also distinguish choosing which measurements to verify from choosing which predictions to use. Filtered prediction-powered inference selects regions for prediction-based augmentation while retaining the full-population target (Xu and Sun, 2026). OPAL optimizes labeling policies for prediction-assisted inference using influence-function information (Ma and Candès, 2026). Power analysis for prediction-powered inference addresses labeled sample size through asymptotic variance and power calculations (Chen et al., 2026). These are relevant comparators for a future decision rule. Our conditional correction distributions do not establish better allocation, a smaller required sample, or valid selection of a low-distortion population.

Exact enumeration is exponential and is capped at 16 players in the implementation. For larger games, the code provides permutation Shapley estimates with numerical Monte Carlo standard errors. Each sampled order telescopes to the endpoint gap. That estimator does not recover the entire correction-set distribution. The archive applications instead sample fixed-size sets directly to estimate means and dispersion; rare-tail estimation would require further computation. Large batches change the question, so grouping is part of the diagnostic specification, not an innocuous speed optimization.

The present evidence supports a finite-sample diagnostic application. A stronger research contribution would require evidence that the path changes a consequential decision, or a prospective validation study showing better estimation under the same acquisition budget. The archive applications measure approximation discrepancies, but do not establish either decision benefit.

References

- Amorim, G., Tao, R., Lotspeich, S., Shaw, P. A., Lumley, T., and Shepherd, B. E. (2021). Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4):1368–1389.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023a). Prediction-powered inference. *Science*, 382(6671):669–674.
- Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2023b). PPI++: Efficient prediction-powered inference. arXiv:2311.01453.
- Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2024). Inference for regression with variables generated by AI or machine learning. arXiv:2402.15585.
- Chen, T. and Lumley, T. (2022). Optimal sampling for design-based estimators of regression models. *Statistics in Medicine*, 41(8):1482–1497.
- Chen, Y. T., Guo, M., and Li, S. (2026). Power analysis for prediction-powered inference. arXiv:2603.16041.
- Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.
- Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022). How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.
- 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.

- Erdős, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 4(1):49–61.
- Fong, C. and Tyler, M. (2021). Machine learning predictions as regression covariates. *Political Analysis*, 29(4):467–484.
- Ghorbani, A. and Zou, J. (2019). Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR.
- Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss Army infinitesimal jackknife. arXiv:1907.12116.
- Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019b). A Swiss Army infinitesimal jackknife. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1139–1147. PMLR.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(3):361–374.
- Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.
- Karlaš, B., Dao, D., Interlandi, M., Schelter, S., Wu, W., and Zhang, C. (2024). Data debugging with Shapley importance over machine learning pipelines. In *The Twelfth International Conference on Learning Representations*.
- Kluger, D. M. and Bates, S. (2026). M-estimation under two-phase multiwave sampling with applications to prediction-powered inference. arXiv:2602.16933.
- Kluger, D. M., Lu, K., Zrnic, T., Wang, S., and Bates, S. (2025). Prediction-powered inference with imputed covariates and nonuniform sampling. arXiv:2501.18577.
- Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.
- Lee, A. H. and Zhao, Y. (1996). Assessing local influence in measurement error models. *Biometrical Journal*, 38(7):829–841.
- Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.
- Ma, V. L. and Candès, E. J. (2026). Optimized labeling resource allocation for prediction-assisted inference via OPAL. arXiv:2606.03211.
- Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116.
- Okhrati, R. and Lipani, A. (2021). A multilinear sampling algorithm to estimate Shapley values. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7992–7999. IEEE.
- Owen, G. (1972). Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79.

- Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.
- Sundararajan, M. and Najmi, A. (2020). The many Shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278.
- Wang, R., Wang, Q., and Miao, W. (2023). A maximin optimal approach for sampling designs in two-phase studies. arXiv:2312.10596. Accepted for publication in *Statistica Sinica*.
- Wang, S., McCormick, T. H., and Leek, J. T. (2020). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275.
- Xu, S. and Sun, W. W. (2026). Do more predictions improve statistical inference? filtered prediction-powered inference. arXiv:2602.10464.
- Zrnic, T. and Candès, E. J. (2024). Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010. PMLR.

Appendix: Supplementary results and checks

A Linear outcome corrections as a boundary case

For a fixed linear treatment contrast, let ψ_g be the change from correcting cluster g , and $\hat{\tau}_P$ the proxy contrast. Raw replacement gives

$$\hat{\tau}_R(S_k) = \hat{\tau}_P + \sum_{g \in S_k} \psi_g, \quad \mathbb{E}\hat{\tau}_R(S_k) = \hat{\tau}_P + \frac{k}{G}(\hat{\tau}_G - \hat{\tau}_P). \quad (\text{A1})$$

The mean is mechanically linear, but the correction-set distribution need not be concentrated or approximately Normal. Finite-population limit theory provides conditions for Normal approximation (Erdős and Rényi, 1959; Hájek, 1960; Li and Ding, 2017); enumerating a finite distribution does not create a new sampling theory. The retained outcome benchmark holds RMSE and predictive R^2 fixed across three patterns. Its diffuse and concentrated patterns also share both endpoints, while their dispersions differ.

Error pattern	Gold ATE	Proxy ATE	Gap	RMSE	R^2	Top-two share
Diffuse	0.100	0.060	0.040	0.200	0.965	10%
Concentrated	0.100	0.060	0.040	0.200	0.965	100%
Offsetting	0.100	0.100	0.000	0.200	0.965	50%

Table A1: Retained outcome benchmark: 20 clusters of 50 observations. Top-two share is the share of absolute correction contributions in the two largest clusters.

A design-scaled correction, $\hat{\tau}_H(S_k) = \hat{\tau}_P + (G/k) \sum_{g \in S_k} \psi_g$, is the standard difference-estimation construction (Horvitz and Thompson, 1952; Särndal et al., 1992). Unlike the raw swap, its design expectation is the full-gold contrast at every positive budget. The following table describes this scaled estimator. Its logic should not be transferred to nonlinear covariate refits merely by multiplying a raw coefficient change by G/k .

Error pattern	Gold clusters	Mean	SD	5th	95th	Tail risk	Normal price
Diffuse	2	0.100	0.000	0.100	0.100	0%	0%
	5	0.100	0.000	0.100	0.100	0%	0%
	10	0.100	0.000	0.100	0.100	0%	0%
	20	0.100	0.000	0.100	0.100	0%	0%
Concentrated	2	0.100	0.083	0.060	0.260	19%	54%
	5	0.100	0.048	0.060	0.220	5%	29%
	10	0.100	0.028	0.060	0.140	0%	7%
	20	0.100	0.000	0.100	0.100	0%	0%
Offsetting	2	0.100	0.123	-0.100	0.300	35%	68%
	5	0.100	0.071	0.020	0.180	57%	48%
	10	0.100	0.041	0.020	0.180	14%	22%
	20	0.100	0.000	0.100	0.100	0%	0%

Table A2: Design-scaled outcome correction. Quantiles, SD and tail risk are over uniform fixed-size subsets. Tail risk is the probability of an absolute departure from the full-gold contrast greater than 0.050; “Normal price” is the corresponding Normal approximation using the exact variance. These are conditional benchmark probabilities, not confidence statements.

B A validation sample can miss interacting errors

Consider a finite cohort of $N \geq 3$ observations. The outcome Y , treatment D , and proxy covariate $\tilde{U}$ are observed for everyone. Only a simple random sample of n distinct rows receives verification. The regression includes an intercept, D , and the covariate. Fix the observed data and two possible completions of its unverified values as follows:

Rows	D	Y	$\tilde{U}$	$U^{(0)}$	$U^{(1)}$
$N - 2$ ordinary rows	0	0	0	0	0
Row i	1	1	0	0	1
Row j	0	1	1	1	2

Table A3: Observed data and two possible covariate completions. Only the verified values of the pilot rows are observed. The last two columns agree on every ordinary row. Values are specified analytically.

These are stipulated values defining an algebraic example, not simulation output. Completion 0 has no measurement error. In completion 1 only rows i and j require correction. Write $v(S)$ for the coefficient on D after correcting rows in S . Every possible correction design has full column rank. Indeed, the ordinary rows imply a fitted intercept of zero, row j gives the covariate coefficient $1/u_j$, and row i gives $v(S) = 1 - u_i/u_j$, where $u_i \in \{0, 1\}$ and $u_j \in \{1, 2\}$ are the covariate values used in that fit. All outcomes are fitted exactly. Hence, in completion 1,

$$v(\emptyset) = 1, \quad v(\{i\}) = 0, \quad v(\{j\}) = 1, \quad v(\{i, j\}) = \frac{1}{2}. \quad (\text{A2})$$

The interaction is

$$v(\{i, j\}) - v(\{i\}) - v(\{j\}) + v(\emptyset) = \frac{1}{2}. \quad (\text{A3})$$

Adding the isolated correction effects predicts a jointly corrected coefficient of zero; the actual coefficient is $1/2$. The two calculations therefore fall on opposite sides of any substantive threshold strictly between zero and $1/2$. This concerns a regression coefficient; the example makes no claim of causal identification.

Let E be the event that the pilot misses both special rows. For $n \leq N - 2$,

$$q = \Pr(E) = \frac{\binom{N-2}{n}}{\binom{N}{n}} = \frac{(N-n)(N-n-1)}{N(N-1)}. \quad (\text{A4})$$

For $n > N - 2$, $q = 0$. On E , the entire observed record, including the full cohort's outcomes and proxies, the sampled identities, and their verified values, is identical in the two completions. Every correction available in the pilot is zero, although completion 1 has the interaction in (A2). For example, $N = 100$ and $n = 10$ give $q = 89/110$, about 81%.

This also bounds any procedure using these observations. Let $T \in \{0, 1\}$ flag consequential interactions, and suppose its false-positive probability in completion 0 is at most α . Allow the procedure its own randomization. The sampling law and observed data agree on E , so

$$\Pr_1(T = 1) = \Pr_1(T = 1, E) + \Pr_1(T = 1, E^c) \quad (\text{A5})$$

$$\leq \Pr_0(T = 1, E) + 1 - q \leq \alpha + 1 - q. \quad (\text{A6})$$

Thus its probability of failing to flag completion 1 is at least $q - \alpha$ when that quantity is positive. This elementary indistinguishability argument rules out a uniform guarantee over unrestricted covariate completions under the stated random-pilot design.

The example deliberately concentrates identifying variation in two observable rows. A design that targets those rows can avoid the problem; this is not a lower bound for every adaptive sampling policy. Nor does the calculation establish the prevalence of hidden interactions. It shows precisely why a random pilot’s absence of errors is insufficient on its own: extrapolation to unverified rows needs restrictions on the possible errors, a sampling design that covers consequential rows, or other information linking observed and unobserved corrections.

C Reproducibility and checks

The repository contains the analysis plan, complete coalition values for the main covariate examples, repeated-design summaries, generated tables and figures, and separate metadata for the benchmarks. Tests check the normal-equation identity against independent least-squares fits, exact Shapley against every permutation in a small game, its agreement with integrated marginal paths, partial-validation endpoints, the additive control, small-amplitude behavior, and explicit failures for rank-deficient coalitions. The historical implementation audit is retained in the research notes; its domain-selection and calibration claims are not used as evidence here.

The archive pilot fixes source-file checksums and uses the housing repository at commit `f4adf2d` and the Fong–Tyler Dataverse deposit, version 1.2.¹ The original Reddit training CSV has 2,413 rows, whereas Dataverse’s converted tabular file has 2,412. The document-term indices match the original CSV, which we use without reordering. Training and validation comment IDs do not overlap; the archive’s matrix-to-row ordering is retained. Human labels and other reference measurements are treated as fixed, not assumed to measure constructs without error.

The empirical specifications and the correction of the training-row count are recorded in the analysis plan. No downstream coefficients were inspected before specifying the two targets and budgets; this is a self-attested exploratory plan, not a preregistration. Saved outputs include every sampled-set coefficient, approximation errors, conditioning, numerical Monte Carlo standard errors for means and RMSE, and software versions. All sampled designs have full rank. Individual-row updates computed from sufficient statistics agree with independent least-squares refits on random, high-leverage, and large-update checks. Neither archive supplies the design needed here to interpret HC0 scales as valid population uncertainty; they are used only to put discrepancies on a common descriptive scale.

¹Housing: https://github.com/DanKluger/Predict-Then-Debias_Bootstrap. Reddit: <https://doi.org/10.7910/DVN/QQHBHY>. The machine-readable source manifest contains the full commit identifier and file-specific download URLs.