Unbiased Regression with Costly Item Labels

Gaurav Sood

Abstract

We study regression on per–row trait shares when item labels are costly. Rows are units (people, devices, firms), columns are items (domains, products, apps), and each item carries a latent binary trait. The statistical estimands are: (i) the vector of row shares $y = (y_1, \ldots, y_n)$ and its functionals (e.g., the mean), and (ii) the population OLS coefficient vector $\beta^* = (X^\top X)^{-1} X^\top y$. We use *item–sampled* Horvitz–Thompson (HT) estimators for row shares. HT delivers *row–wise* unbiasedness and therefore design–unbiased OLS under any sampling design independent of the unknown labels. We then make explicit what HT does *not* guarantee: because the same sampled items affect all rows, errors are shared across rows; the empirical distribution of estimated shares is a noisy convolution of the truth; and realized OLS variance is governed by the X–aligned component of the error. We propose two complementary, convex design objectives: *regression–SE control*, which targets the X–aligned error that moves OLS, and *row–SE control*, which guarantees per–row precision. Both admit prevalence–aware tightenings when at most an α fraction of items can be positive. We define concrete designs (HT–Uniform, HT–$\|g\|$, HT–A–Opt, and min–labels variants) and give a procedure to turn inclusion probabilities into an explicit list of items to label via balanced fixed–size sampling. Simulations show that A–optimal regression designs substantially reduce coefficient RMSE at a given budget; matching the same regression variance while enforcing per–row guarantees typically requires more labels; and balancing further lowers realized variance without sacrificing unbiasedness.

1 Setup, HT guarantees, and limits

Let $C = (c_{ij}) \in \mathbb{R}_+^{n\times m}$ be counts, $T_i = \sum_{j=1}^m c_{ij}$, and the row share

$$y_i \;=\; \sum_{j=1}^m \frac{c_{ij}}{T_i}\, a_j, \qquad a_j \in \{0,1\}.$$

Let $X \in \mathbb{R}^{n\times p}$ and define the OLS estimand

$$\beta^* \;=\; (X^\top X)^{-1} X^\top y.$$

We sample *items* with inclusion probabilities $\pi_j \in (0,1]$ and indicators $I_j \sim \text{Bernoulli}(\pi_j)$ that are independent of the unknown labels $a = (a_1, \ldots, a_m)$. The item–wise Horvitz–Thompson estimator of the share is (Horvitz and Thompson, 1952)

$$\widehat{y}_i \;=\; \sum_{j=1}^m \frac{I_j}{\pi_j} \frac{c_{ij}}{T_i}\, a_j, \qquad u := \widehat{y} - y.$$

Because $\mathbb{E}[I_j/\pi_j] = 1$ and T_i is fixed, $\mathbb{E}[\widehat{y}_i] = y_i$ for every i, hence

$$\mathbb{E}[\widehat{\beta}] = (X^\top X)^{-1} X^\top \mathbb{E}[\widehat{y}] = (X^\top X)^{-1} X^\top y = \beta^*.$$

HT is applied over items (the finite population), not rows; each y_i is a linear functional of a. Rows with $T_i = 0$ are excluded (or y_i defined and excluded from regression).

What HT does not guarantee. The same item draws enter every row, so errors are *shared across rows*:

$$u_i = \sum_{j=1}^{m} \left(\tfrac{I_j}{\pi_j} - 1\right) \frac{c_{ij}}{T_i}\, a_j, \qquad \mathrm{Cov}(u_i, u_k) = \sum_{j=1}^{m} \frac{1-\pi_j}{\pi_j} \frac{c_{ij}}{T_i} \frac{c_{kj}}{T_k}\, a_j^2.$$

Consequently, the empirical distribution of $\widehat{y}$ is the true distribution *convolved* with design noise; HT guarantees unbiased *means* and unbiased *OLS*, not unbiased *quantiles*. For $\widehat{\beta}$, only the X–aligned error matters:

$$\widehat{\beta} - \beta^* = (X^\top X)^{-1} X^\top u.$$

For inference, design–based SEs (or, conservatively, item–cluster robust SEs) should be used; see §6. If labels suffer misclassification (sensitivity/specificity $\neq 1$), HT is unbiased for the *noisy* trait; label–error corrections are then required for consistency in β^*.

2 Intuition: which items move OLS?

Define item j's row–normalized exposure, its projection on the regression space, and its OLS influence weight:

$$v_j = \frac{c_{\cdot j}}{T} \in \mathbb{R}^n, \qquad g_j = X^\top v_j \in \mathbb{R}^p, \qquad w_j = g_j^\top (X^\top X)^{-1} g_j = v_j^\top X (X^\top X)^{-1} X^\top v_j \geq 0.$$

Under independent item sampling,

$$\mathrm{Var}(u) = \sum_{j=1}^{m} \frac{1-\pi_j}{\pi_j} a_j^2\, v_j v_j^\top \;\preceq\; \sum_{j=1}^{m} \frac{1-\pi_j}{\pi_j}\, v_j v_j^\top.$$

We adopt the *whitened A–optimal* criterion (Kiefer, 1959)

$$\widetilde{\Delta}(\boldsymbol{\pi}) \;:=\; (X^\top X)^{-1/2} X^\top\, \mathrm{Var}(u)\, X\, (X^\top X)^{-1/2},$$

whose trace bounds as

$$\mathrm{tr}\; \widetilde{\Delta}(\boldsymbol{\pi}) \;\leq\; \sum_{j=1}^{m} \frac{1-\pi_j}{\pi_j}\, w_j = \sum_{j=1}^{m} \left(\frac{1}{\pi_j} - 1\right) w_j. \tag{1}$$

Interpretation. Items with large w_j are the ones whose noise projects strongly onto X; leaving them unlabeled inflates OLS variance. The fixed–budget optimum will therefore sample with $\pi_j \propto \sqrt{w_j}$.

Remark on metrics. One could minimize $\mathrm{tr}\,\mathrm{Var}(\widehat{\beta})$ directly, which weights items by $g_j^\top (X^\top X)^{-2} g_j$. We fix the *whitened* trace above for a consistent criterion across the paper; both choices yield square–root allocations and convex programs.

3 Two convex design objectives

Regression–SE control (target the X–aligned error). Two equivalent formulations:

$$\min_{\boldsymbol{\pi}} \sum_j \frac{w_j}{\pi_j} \quad \text{s.t.} \quad \sum_j \pi_j = K, \quad \pi_{\min} \le \pi_j \le 1, \qquad \Rightarrow \qquad \pi_j \;\propto\; \sqrt{w_j} \;\; (\text{clamp to } [\pi_{\min}, 1]),$$

or

$$\min_{\boldsymbol{\pi}} \sum_j \pi_j \quad \text{s.t.} \quad \sum_j \frac{w_j}{\pi_j} \;\le\; \rho^2 + \sum_j w_j, \quad \pi_{\min} \le \pi_j \le 1.$$

Both are convex; both admit heterogeneous label costs by minimizing $\sum_j c_j \pi_j$, which tilts the KKT solution to $\pi_j \propto \sqrt{w_j/c_j}$. The inclusion floor $\pi_{\min} > 0$ ensures HT is well–defined for any item that can affect the estimators (i.e., whenever some $q_{ij} > 0$ below).

Row–SE control (guarantee per–row precision). Let $q_{ij} = (c_{ij}/T_i)^2$. Under Poisson sampling,

$$\operatorname{Var}(u_i) \;\le\; \sum_{j=1}^{m} \frac{1-\pi_j}{\pi_j} q_{ij} = \sum_{j=1}^{m} \frac{q_{ij}}{\pi_j} - \sum_{j=1}^{m} q_{ij}.$$

Given tolerances $\varepsilon_i > 0$ and $\pi_{\min} > 0$,

$$\min_{\pi \in [\pi_{\min}, 1]^m} \sum_{j=1}^{m} \pi_j \quad \text{s.t.} \quad \sum_{j=1}^{m} \frac{q_{ij}}{\pi_j} \;\le\; \varepsilon_i^2 + \sum_{j=1}^{m} q_{ij} \quad \forall i, \tag{2}$$

which is convex because $1/\pi$ is convex and $q_{ij} \ge 0$. With costs $c_j > 0$, minimize $\sum_j c_j \pi_j$. The KKT shape (ignoring box constraints) is

$$\pi_j^\star \;\propto\; \sqrt{\frac{\sum_i \mu_i \, q_{ij}}{c_j}}, \qquad \mu_i \ge 0,$$

then clamp to $[\pi_{\min}, 1]$. Extremely small T_i can force large budgets for tight ε_i; choosing $\varepsilon_i \propto 1/\sqrt{T_i}$ equalizes effort per effective observation.

Prevalence–aware tightening. If at most an α fraction of items are positive ($M = \lceil \alpha m \rceil$), replace sums by the *sum of the M largest* terms using the convex epigraph identity

$$\sum_{k=1}^{M} t_{(k)} = \min_{\tau \in \mathbb{R}} \left\{ M\tau + \sum_{j=1}^{m} (t_j - \tau)_+ \right\},$$

a standard trick in convex optimization (see Boyd and Vandenberghe, 2004). Use $t_j = (1/\pi_j - 1) w_j$ for regression–SE and $t_{ij} = (1/\pi_j - 1) q_{ij}$ for row–SE. This insures against the worst αm items while preserving convexity. If prior probabilities $\Pr(a_j = 1)$ are available, an *expected–risk* variant replaces a_j^2 by $\Pr(a_j = 1)$, yielding another convex program. For minimax/partial–ID intuition, see Manski (2003).

4 Designs used in experiments

All designs below use the *same estimator* (HT shares over items); they differ only in how π is chosen.

- HT–UNIFORM (fixed budget K): $\pi_j = K/m$ (clamped), then sample a fixed–size set of K items.
- HT–$\|g\|$ (fixed K): $\pi_j \propto \|g_j\|_2$; clamp and rescale so $\sum_j \pi_j = K$.
- HT–A–OPT (fixed K): $\pi_j \propto \sqrt{w_j}$; clamp and rescale so $\sum_j \pi_j = K$.
- MIN–LABELS (REG–SE CAP): solve $\min \sum_j \pi_j$ s.t. $\sum_j w_j/\pi_j \le \rho^2 + \sum_j w_j$ and $\pi_{\min} \le \pi_j \le 1$.
- MIN–LABELS (ROW–SE CAPS): solve (2) (optionally joint with the regression cap).
- PREVALENCE–AWARE variants: in either program, replace the relevant sum by the top–M aggregate via the epigraph.

5 From probabilities to a concrete list of items to label

Solving any program yields inclusion probabilities $\pi^\star = (\pi_1^\star, \ldots, \pi_m^\star)$. To produce an explicit labeling list:

1. **Deterministic picks.** Include all items with $\pi_j^\star \ge 0.99$. Let $K = \mathrm{round}(\sum_j \pi_j^\star)$ and $K_{\mathrm{rem}} = K - \#\{j : \pi_j^\star \ge 0.99\}$.
2. **Balanced fixed–size draw.** On the rest, draw exactly K_{rem} items with first–order inclusions $\pi^\star$ and auxiliaries g_j (or $[g_j; 1]$). A standard choice is the cube method (fixed–size phase) or any conditional–Poisson scheme with balancing (Deville and Tillé, 2004). This targets

$$\sum_j \left(\frac{I_j}{\pi_j^\star} - 1\right) g_j \approx 0,$$

   shrinking the realized $X^\top u$ for any a.
3. **Estimation.** Use HT weights $1/\pi_j^\star$ when computing shares; this preserves exact design–unbiasedness. (Optional: a post–sampling calibration step can further reduce variance at the cost of a negligible finite–sample bias (Deville and Särndal, 1992).)

Adaptive sampling caveat. If sampling proceeds in waves using already observed labels a_j, the final inclusion probabilities must reflect the adaptive design for HT to remain unbiased.

6 Inference and variance estimation

Under independent item sampling, $\widetilde{\Delta}(\boldsymbol{\pi})$ provides a conservative covariance bound for whitened coefficients; for without–replacement fixed–size designs, Sen–Yates–Grundy variance formulas with joint inclusions (π_{jk}) yield tighter design–based variance estimators for HT totals and, by the delta method, for $\widehat{\beta}$ (Sen, 1953; Yates and Grundy, 1953). In practice:

- Report design–based SEs using $\widetilde{\Delta}(\boldsymbol{\pi})$ or a SYG estimator adapted to the actual design.
- As a conservative check, use item–cluster robust SEs for OLS on $\widehat{y}$ (errors are shared across rows via items).

- Efficiency upgrade: generalized least squares (GLS) with an estimated $\Sigma \approx \mathrm{Var}(u)$, i.e.,

$$\widehat{\beta}_{\mathrm{GLS}} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \widehat{y},$$

  is unbiased (design–based) and can dominate OLS when shared–noise is strong; Σ can be approximated under the Poisson bound or via a fixed–size SYG approximation.

Balanced/fixed–size designs induce negative dependence among draws and reduce variance relative to Poisson; our Poisson–based caps are therefore conservative.

7 Simulation evidence (brief)

On synthetic data ($n = 400$, $m = 800$, $p = 6$), HT–A–Opt dominates HT–Uniform and slightly improves on HT–$\|g\|$ at the same budget K; e.g., at $K = 80$ it achieves lower coefficient RMSE and empirical tr $\widetilde{\Delta}$. In min–labels experiments, the regression–SE program yields a smooth budget–variance trade–off (e.g., $K \approx 79$ at a moderate cap, falling to ≈ 54 at a looser cap). In an iso–variance comparison (matching the empirical variance of a regression–SE design), the row–SE program required substantially more labels in our baseline instance, reflecting that protecting every row is stricter than protecting the X–aligned error alone. Balanced fixed–size selection further reduced realized variance while preserving unbiasedness.

8 Related work

Our setting transposes classical ideas from two adjacent literatures. In *two–phase (validation) designs for regression*, one optimizes a subsample of *units* to estimate regression parameters under cost constraints, often via influence–function–weighted allocations, GREG, or semiparametric efficient scores; see, e.g., Chen and Lumley (2022); McIsaac and Cook (2015). Here we optimize a subsample of *items* to construct a derived outcome and then regress, but the design logic (auxiliaries, calibration/balancing, convex programs) is analogous. In *balanced sampling and calibration*, the cube method and GREG aim to match auxiliary totals and reduce realized variance without changing first–order inclusions; that is exactly our goal when we drive $\sum_j (\frac{I_j}{\pi_j} - 1) g_j$ towards zero (Deville and Särndal, 1992; Deville and Tillé, 2004). Finally, *optimal experimental design* (A–optimality) motivates the square–root rule (Kiefer, 1959), and *randomized sketching* (leverage–score sampling) offers a useful contrast: those sample *rows of* X to approximate least–squares on full data, while we sample *columns (items)* to construct y itself (Drineas and Mahoney, 2016). For broader sampling foundations, see Neyman (1934).

9 Assumptions and caveats

Labels, once observed, are accurate; item selection is independent of unknown a_j but may depend on (C, X).[1] If $X^\top X$ is ill–conditioned, use $w_j = g_j^\top (X^\top X + \lambda I)^{-1} g_j$; convexity and numerics improve. Choose $\pi_{\min} > 0$ so that any item that can affect estimators has nonzero inclusion. Balanced/fixed–size designs reduce variance relative to Poisson; Poisson–based caps are conservative. If label misclassification is present, HT targets the noisy trait unless corrected.

[1] If adaptive designs use realized labels, HT remains unbiased only if final inclusion probabilities correctly reflect that adaptivity.

Software and reproducibility

A minimal, open–source implementation is available as the Python package `fewlab` (Sood, 2025). The repository includes a small API (`items_to_label`) and examples mirroring the design logic in this paper. See https://github.com/finite-sample/fewlab.

References

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. doi: 10.1017/CBO9780511804441.

T. Chen and T. Lumley. Optimal sampling for design-based estimators of regression models. *Statistics in Medicine*, 41(8):1482–1497, 2022. doi: 10.1002/sim.9300.

J.-C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992. doi: 10.1080/01621459.1992.10475217.

J.-C. Deville and Y. Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91(4): 893–912, 2004. doi: 10.1093/biomet/91.4.893.

P. Drineas and M. W. Mahoney. Randnla: Randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016. doi: 10.1145/2842602.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959. doi: 10.1111/j.2517-6161.1959.tb00340.x.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003. doi: 10.1007/b97478.

M. A. McIsaac and R. J. Cook. Adaptive sampling in two-phase designs: A biomarker study for progression in arthritis. *Statistics in Medicine*, 34(21):2899–2912, 2015. doi: 10.1002/sim.6523.

J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. doi: 10.2307/2342192.

A. R. Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127, 1953.

G. Sood. fewlab: Fewest items to label for unbiased ols on shares. https://github.com/finite-sample/fewlab, 2025. MIT License; Python package; accessed 2025-09-08.

F. Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society: Series B (Methodological)*, 15(2): 253–261, 1953.