

# Model-based Survey Weighting Using Logistic Regression

Seonghun Lee

Andrew Gelman

October 3, 2025

## Contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
<b>2 Methods</b>	<b>3</b>
2.1 Survey Design and Design-Based Weighting . . . . .	3
2.2 Model-Based Weighting . . . . .	5
2.3 Practical Issues . . . . .	9
<b>3 Empirical Results</b>	<b>10</b>
3.1 Simulation Study . . . . .	10
3.2 Applications to FFCWS Study . . . . .	14
<b>4 Asymptotic Properties</b>	<b>15</b>
<b>5 Discussion</b>	<b>18</b>
5.1 Key Findings and Limitations . . . . .	18
5.2 Future Research . . . . .	19

*Abstract.* Equivalent weights from regression models provide a bridge between design-based and model-based survey inference. We introduce a closed-form construction of logistic regression equivalent weights, defined by matching the weighted sample mean of the outcome to the population mean predicted under the logistic model. Although logistic weighting is nonlinear, a first-order expansion shows that outcome dependence is negligible, allowing the weights to be treated as functions of covariates alone. We establish asymptotic properties of these weights, including consistency, asymptotic linearity, and a plug-in variance estimator. The resulting estimators preserve efficiency relative to design-based alternatives, with population estimation properties compatible with those of linear weighting. Importantly, logistic weighting demonstrates superior preservation of effective sample size after raking and more accurate alignment with population benchmarks. Simulation evidence confirms that the asymptotic results hold well in finite samples of the scale typical in survey applications. An application to the Fragile Families and Child Wellbeing Study illustrates how logistic weighting modifies prevalence estimates while remaining consistent with simulation results. Together, these results extend regression-based weighting theory to the logistic case and provide new asymptotic justification for its use in survey practice.

**KEYWORDS:** Logistic equivalent weights; Effective sample size; Variance estimation; Asymptotic properties; Survey methodology

## 1 Introduction

Survey weights are central to making sample data representative of target populations. Constructing base weights or adjusting them involves multiple steps and subjective choices, often relying on auxiliary data to align estimates with census counts or other benchmarks. Yet, survey data frequently face challenges such as missingness and nonresponse, which complicate these procedures and introduce instability.

A substantial body of literature addresses nonresponse adjustment, design- and model-based inference, and small area estimation (Chapman et al., 1986; Little, 1986; Bethlehem et al., 1996; Chu and Goldman, 1997; Lu and Gelman, 2003; Little, 2015; Rao and Molina, 2015; Haziza and Beaumont, 2017; Skinner and Wakefield, 2017; Chen et al., 2017; Liu et al., 2023). At the same time, concerns remain regarding the construction of base weights themselves, where methods are often ad hoc, hindered by the limited availability of detailed and confidential census data (Carlson, 2008), and lacking systematic comparison. Moreover, the role of weights in regression remains controversial: they do not always reduce bias, and adjustments such as raking can yield extreme weights even after trimming (Little and Rubin, 1987; Deville and Särndal, 1992; Miller, 2011).

Traditional approaches to constructing base weights follow design-based guidelines (Chu and Goldman, 1997; Valliant et al., 2013; Valliant and Dever, 2018). Model-based alternatives have been

proposed, including regression-based constructions (Gelman, 2007), but their implementation and advantages over design-based methods remain underexplored. In particular, while linear regression has been considered for equivalent weights, logistic regression has not been fully developed as a basis for constructing weights, despite its natural appeal for binary or categorical sampling indicators.

This paper develops a model-based framework that constructs base weights via logistic regression, building on and extending Gelman (2007). We compare design-based, linear, and logistic weighting through both simulation and empirical analysis, providing practical guidance for researchers confronting challenges in weight construction. Beyond empirical evidence, we establish asymptotic properties: Theorem 1 and Corollary 1 show that the effective sample size ratio converges to a population limit, and Theorem 2 provides an asymptotic linear expansion with a consistent plug-in variance formula. These results demonstrate that logistic weights yield stable large-sample properties and offer a theoretical benchmark for interpreting finite-sample performance.

## 2 Methods

### 2.1 Survey Design and Design-Based Weighting

Our motivating application is the Future of Families and Child Wellbeing Study (FFCWS), a birth cohort survey of 4,898 children born in large U.S. cities (population over 200,000) between 1998 and 2000. The design followed a stratified multistage scheme, oversampling births to unmarried mothers at a 3-to-1 ratio and ensuring adequate representation of Black, Hispanic, and low-income families. Children and their families were followed across seven waves when the child was approximately ages 1, 3, 5, 9, 15, and 22. Poststratification in the FFCWS relied on four demographic variables—mother’s marital status, race/ethnicity, age, and education—along with city of birth. Population birth counts for these strata were derived from Centers for Disease Control and Prevention (CDC) natality data (Carlson, 2008). To mirror the FFCWS, our simulation study generates finite populations and base weights that replicate the demographic and geographic composition of the original survey.

In large-scale surveys, stratified multistage cluster sampling (SMCS) is commonly used to balance fieldwork feasibility with subgroup precision. In the FFCWS, strata corresponded to cities and

hospitals served as primary sampling units. For methodological clarity, however, our simulations employ stratified simple random sampling (SRS), where units are grouped by strata and selected independently within each group. Although SRS omits within-cluster correlation, two features motivate its use here: (i) SMCS typically reduces efficiency due to intra-cluster correlation, yielding larger standard errors; and (ii) once raking adjustments are applied, differences between SMCS- and SRS-based weights are expected to be minimal. Unless otherwise stated, we draw samples of size  $n = 3000$  from a finite population of  $N = 1,000,000$ , representing a setting where a moderate sample approximates a much larger population.

For a finite population and a probability sample, each unit  $i$  has a known selection probability  $\pi_i > 0$ , with the base weight defined as

$$w_i = \frac{1}{\pi_i}.$$

Selection probabilities are determined by stratum and demographic characteristics, ensuring that units within the same cell share identical  $\pi_i$ . These base weights are then adjusted in two standard ways. First, nonresponse adjustment multiplies weights by the inverse of the weighted response rate within adjustment cells (e.g., within each city). Second, raking aligns weighted sample margins with known population totals without requiring full cross-classification, thereby mitigating empty-cell problems. Nonresponse adjustments renormalize weights to the total population, while raking ensures consistency with auxiliary information (Deville and Särndal, 1992; Valliant and Dever, 2018). For recent applications of such adjustments to the FFCWS, see Vegetabile et al. (2020) and Lee and Gelman (2024). Table 1 reports base weights from a representative simulation.

Table 1: Summary of base weights (simulation example)

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Value	2.86	121.40	260.80	333.33	492.49	1396.96

When estimating a population mean, let  $X$  denote poststratification variables forming  $S$  cells with sizes  $N_s$  in the population and  $n_s$  in the sample. For an outcome  $y$ , the population mean is

$$\theta = \frac{\sum_{s=1}^S N_s \theta_s}{N},$$

with  $\theta_s$  the mean in stratum  $s$ . The sample analogue is the poststratified estimator

$$\hat{\theta}^{PS} = \frac{\sum_{s=1}^S N_s \hat{\theta}_s}{N}.$$

Equivalently, one may use the Horvitz–Thompson (HT) or Hájek estimator,

$$\bar{y}_{HT} = \frac{\sum_{i=1}^n w_i y_i}{N}, \quad \bar{y}_H = \frac{\sum_{i=1}^n w_i y_i}{\hat{N}}, \quad \hat{N} = \sum_{i=1}^n w_i,$$

which coincide in our setting since  $N$  is known. Importantly, the poststratified estimator  $\hat{\theta}^{PS}$  is algebraically equivalent to the Hájek estimator, a relationship that underpins the development of model-based equivalent weights in the next section.

## 2.2 Model-Based Weighting

Model-based approaches construct equivalent unit weights by projecting population quantities onto auxiliary variables through regression. Let  $X$  denote the  $n \times k$  design matrix of auxiliary variables for the sample, and let  $X^{pop}$  denote the  $S \times k$  design matrix for the population poststratification cells, with  $n$  the sample size,  $S$  the number of cells, and  $k$  the number of auxiliary variables (e.g., demographic or geographic characteristics).

**Linear regression weights.** Following Gelman (2007), if the survey outcome  $y$  has a linear relationship with the auxiliary variables, then for sample design matrix  $X$  and population cell matrix  $X^{pop}$ , the regression coefficients are

$$\hat{\beta} = (X^\top X)^{-1} X^\top y,$$

and the fitted population cell means are

$$\hat{\theta}_s = X_s^{pop} \hat{\beta}, \quad s = 1, \dots, S.$$

The poststratified estimate of the population mean is therefore

$$\hat{\theta}^{PS} = \frac{1}{N} \sum_{s=1}^S N_s \hat{\theta}_s = \frac{1}{N} (N^{pop})^\top X^{pop} (X^\top X)^{-1} X^\top y.$$

Define  $\hat{w}$  as the equivalent unit weights such that

$$\hat{\theta}^{PS} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i y_i, \quad \sum_{i=1}^n \hat{w}_i = n.$$

Solving for  $\hat{w}$  gives

$$\hat{w} = \frac{n}{N} (N^{pop})^\top X^{pop} (X^\top X)^{-1} X^\top.$$

These weights depend only on the auxiliary variables  $X$  and not on the outcome  $y$ , which is an advantage in practice since demographic and geographic covariates are typically observed with minimal missingness. The formula produces  $S$  unique weights, one for each cell. Renormalization to sum to the population size ensures equivalence to the Hájek estimator:

$$\hat{\theta}^{PS} = \frac{\sum_{i=1}^n \hat{w}_i y_i}{\hat{N}}, \quad \hat{N} = \sum_{i=1}^n \hat{w}_i.$$

Table 2 summarizes the distribution of linear regression weights in a representative simulation.

Table 2: Summary of linear equivalent weights (simulation example)

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Value	0.0022	0.15	0.82	1.00	1.52	7.00

**Logistic regression weights.** To extend beyond the linear setting, consider a binary outcome  $y = (y_1, \dots, y_n)$  modeled by logistic regression,

$$\Pr(y_i = 1 \mid X_i) = p_i, \quad \text{logit}(p_i) = X_i^\top \beta,$$

with fitted probabilities  $\hat{p} = \sigma(X\hat{\beta})$ . Equivalent weights  $\hat{w}$  are defined by requiring balance between the weighted sample mean and the poststratified population mean:

$$\frac{1}{n} \sum_{i=1}^n \hat{w}_i y_i = \frac{1}{N} \sum_{s=1}^S N_s \sigma(X_s^{pop} \hat{\beta}). \quad (1)$$

In survey practice it is standard to assume weights depend only on  $X$ , not on  $y$ . Because logistic regression is nonlinear, we assess the impact of this assumption by a first-order Taylor expansion. Differentiating (1) with respect to  $y$  yields

$$(\nabla \hat{w})y + \hat{w} = \left( \frac{n}{N} (N^{pop})^\top \text{diag} \left[ \sigma(X^{pop} \hat{\beta}) \odot (1 - \sigma(X^{pop} \hat{\beta})) \right] X^{pop} \left( \frac{d\hat{\beta}}{dy} \right)^\top \right)^\top, \quad (2)$$

where  $\odot$  denotes the Hadamard product and

$$\frac{d\hat{\beta}}{dy} = (X^\top W X)^{-1} X^\top, \quad W = \text{diag}(\hat{p} \odot (1 - \hat{p})).$$

Here  $X^\top W X$  is the Hessian of the log-likelihood in logistic regression, capturing the curvature of the likelihood function around the maximum likelihood estimate (MLE), and the diagonal entries of  $W$  represent the conditional variances of the predicted probabilities for the sample units. Substituting this expression back into the right-hand side shows how the first-order correction term  $(\nabla \hat{w})y$  depends on the curvature of the likelihood surface. In practice, however, our empirical results demonstrate that this derivative term is negligible, which justifies treating the equivalent weights as functions of the auxiliary variables alone.

For reference, the left-hand side of the estimating equation can be written as

$$(\nabla \hat{w})y + \hat{w} = J(y)y + \hat{w}(y),$$

where  $J(y)$  is the  $n \times n$  Jacobian matrix with entries  $[J(y)]_{ij} = \frac{\partial \hat{w}_i(y)}{\partial y_j}$ .

**Worked illustration of the Taylor approximation.** To make the derivative term transparent, we expand the estimating equation one unit at a time. Let  $\star$  denote the left-hand side of the equation

at observed  $y$ . For example,

$$\star = \hat{w}(1, 1, 0, \dots, 1) + \frac{dw}{dy_1}(1, 1, 0, \dots, 1) + \frac{dw}{dy_2}(1, 1, 0, \dots, 1) + \dots + \frac{dw}{dy_n}(1, 1, 0, \dots, 1)$$

Substituting  $y_i = 1$  by  $y_i = 0$  for each unit gives:

$$\hat{w}(0, 1, 0, \dots, 1) + \frac{dw}{dy_2}(0, 1, 0, \dots, 1) + \dots + \frac{dw}{dy_n}(0, 1, 0, \dots, 1) \quad (1)$$

$$\hat{w}(1, 0, 0, \dots, 1) + \frac{dw}{dy_1}(1, 0, 0, \dots, 1) + \dots + \frac{dw}{dy_n}(1, 0, 0, \dots, 1) \quad (2)$$

$\vdots$

$$\hat{w}(1, 1, 0, \dots, 0) + \frac{dw}{dy_1}(1, 1, 0, \dots, 0) + \frac{dw}{dy_2}(1, 1, 0, \dots, 0) + \dots + \frac{dw}{dy_n}(1, 1, 0, \dots, 0) \quad (n_1)$$

and the Taylor approximation gives

$$(\star) - (1) \approx \frac{dw}{dy_1}, \quad (\star) - (2) \approx \frac{dw}{dy_2}, \quad \dots, \quad (\star) - (n_1) \approx \frac{dw}{dy_{n_1}}$$

so that

$$(\nabla \hat{w})y \approx (\star) - (1) + (\star) - (2) + \dots + (\star) - (n_1).$$

**Empirical evidence on derivative term.** Across simulations and the FFCWS application, the derivative term  $\Delta(\hat{w})(y) = (\nabla \hat{w})y$  is extremely small: values range from about  $-1.4 \times 10^{-4}$  to  $4.2 \times 10^{-5}$ , with mean  $-4.0 \times 10^{-6}$  and median  $-2.4 \times 10^{-6}$ . In a separate check, perturbing a single outcome changed  $(\nabla \hat{w})y + \hat{w}$  by at most 0.02 (median 0.002). These findings confirm that outcome perturbations have negligible effect on the weights, justifying the assumption that weights depend only on  $X$ .

**Closed-form expression for logistic weights.** Since the derivative contribution  $(\nabla \hat{w})y$  is empirically negligible, we assume that the equivalent weights depend only on the auxiliary weighting variables. Accordingly, we set  $(\nabla \hat{w})y = 0$  in the estimating equation. This yields a closed-form expression for the logistic equivalent weights that depend only on the auxiliary variables  $X$ , consistent with survey practice. The resulting logistic weights are then renormalized to sum to  $n$ , regardless of whether negative weights are present. Importantly, the formula produces  $S$  unique equivalent

weights, one for each poststratification cell.

$$\hat{w}(\hat{\beta}) = \left( \frac{n}{N} (N^{\text{pop}})^{\top} W^{\text{pop}}(\hat{\beta}) X^{\text{pop}} ((X^{\top} W(\hat{\beta}) X)^{-1} X^{\top})^{\top} \right)^{\top},$$

Table 3 summarizes the distribution of logistic equivalent weights from our simulations. While the weights are generally well-behaved, a small fraction attain extreme values. To mitigate their influence, we trim weights above the 95th percentile by replacing them with the 95th percentile value, a standard adjustment in survey practice (DevilleSarnal1992, Valliant2013).

Table 3: Summary of logistic equivalent weights (simulation example)

	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Value	0.0024	0.12	0.83	1.00	1.53	7.01

## 2.3 Practical Issues

**Negative Weights in Model-based Methods.** Negative weights are not uncommon when constructing weights using model-based approaches. They arise when real-world surveys fail to accurately represent the target population due to oversampling or undersampling of particular demographic groups, or because of limited sample sizes. This is expected, as one of the motivations for constructing weights is precisely to correct for imbalances in representativeness.

In most cases, linear equivalent unit weights are positive when samples are randomly drawn from the population. However, in our simulation study designed to mimic the FFCWS data, we encountered negative weights for both linear and logistic equivalent unit weights. To address this, we replace negative weights with the minimum value of the positive weights, an adjustment that has negligible impact on the total weight count across demographic cells once weights are renormalized. Ideally, negative weights would be adjusted within the same cells to preserve cell-specific totals, but in practice this is rarely feasible, as cells with negative weights often do not overlap with cells containing positive weights. Consequently, adjusting for negative weights results in only minor discrepancies in total counts relative to population benchmarks.

**Missingness and Nonresponse.** In practice, missing data in demographic and geographic variables used for survey weighting is rare, but it should be assessed and addressed before constructing

weights. When auxiliary variables are missing, constructing base weights becomes more challenging. For example, city-level birth counts are not included in the CDC natality data, which only provides county-level totals. However, population birth counts can still be estimated at the city level by combining sample information with the CDC natality file. In our application, we linked birth occurrence location with maternal residence to construct city-level estimates.

For nonresponse adjustment, non-respondents are excluded from the weight construction process. Design-based methods require an explicit nonresponse adjustment step, whereas model-based methods inherently account for population totals by incorporating auxiliary information directly into the estimation. For example, the linear weighting method, when free of negative weights, preserves both the sample count and the total population count, since rescaling does not affect the estimate. By contrast, the logistic weighting method also aligns the estimated weights with known population totals, but in practice additional trimming and renormalization are often required to ensure exact agreement because of the nonlinearity of the logistic model.

## 3 Empirical Results

### 3.1 Simulation Study

To assess the finite-sample performance of our methods, we conduct a simulation study designed to mirror realistic survey conditions. We set the sample size at  $n = 3000$ , drawn from a finite population of size  $N = 1,000,000$ , reflecting settings where the sample is small relative to the population but large enough for asymptotic theory to be informative. This design allows us to evaluate how well the asymptotic properties established in Section 4 approximate finite-sample behavior.

We begin by comparing weighting methods in terms of their ability to reproduce known population counts. Table 4 reports estimated population totals by education before raking. The design-based weights show substantial distortions across most categories. In contrast, both model-based approaches yield estimates that closely track the true population distribution. The logistic model performs slightly better than the linear model, with smaller deviations across most categories. Linear regression weighting reduces to an algebraic adjustment based only on  $X$ , whereas logistic regression weighting requires model fitting with fitted probabilities entering the weight formula,

yielding more balanced weights across categories.

Table 4: Population counts by education before raking

Education	< 8 grade	Some HS	HS	Some College	College+	Total
Population	97,409	188,417	300,059	189,532	224,583	1,000,000
Design-based	8,733	179,550	534,609	163,598	113,510	1,000,000
Linear model	90,234	192,799	320,665	188,261	208,041	1,000,000
Logistic model	90,290	190,412	315,998	191,782	211,517	1,000,000

Next, we assess weighting methods by their ability to recover population means of a binary outcome. Table 5 presents estimates by education before raking. Both model-based approaches perform far better than the design-based method, with estimates much closer to the true population rates. Between the two model-based approaches, the logistic model yields the most accurate results, producing slightly smaller deviations across categories than the linear model.

Table 5: Population means of a binary outcome by education before raking

Education	< 8 grade	Some HS	HS	Some College	College+
Population rate	0.47	0.43	0.40	0.23	0.35
Design-raked	0.33	0.40	0.34	0.18	0.28
Linear model	0.34	0.45	0.41	0.22	0.34
Logistic model	0.34	0.45	0.40	0.22	0.34

After raking, all three methods produce estimates that are much closer to the true population rates and broadly similar to one another in Table 6. Some discrepancies persist in certain education categories, however, indicating that residual bias is not fully removed even with raking.

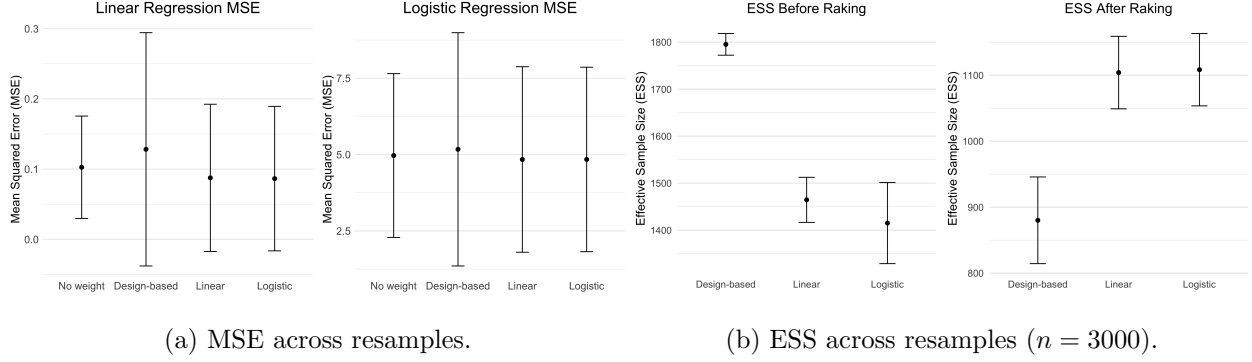
Table 6: Population means of a binary outcome by education after raking

Education	< 8 grade	Some HS	HS	Some College	College+
Population rate	0.47	0.43	0.40	0.24	0.35
Design-raked	0.33	0.45	0.41	0.23	0.38
Linear model	0.35	0.46	0.42	0.23	0.36
Logistic model	0.36	0.46	0.42	0.23	0.36

We also examine the pooled mean squared error (MSE) of regression coefficients across 300 independent resamples, shown in Figure 1a. On average, both model-based approaches achieve lower MSE than the design-based and unweighted regressions, indicating improved accuracy and more stable performance. Figure 1b reports the effective sample size (ESS) across the same replications. Before raking, the design-based method yields the highest ESS, followed by linear and

logistic weighting. After raking, however, the model-based approaches, particularly logistic weighting, retain ESS more effectively, while the design-based method experiences a sharper reduction. Notably, logistic weighting achieves slightly higher ESS than linear weighting after raking.

$$\text{ESS} = \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}.$$



To further connect the finite-sample evidence with the asymptotic theory, we provide two additional checks. To evaluate Theorem 1 and Corollary 1, We compare the empirical ESS ratio  $\bar{w}_n(\hat{\beta})^2/\overline{w_n^2}(\hat{\beta})$  with the population target  $\mathbb{E}[w]^2/\mathbb{E}[w^2]$  using 300 replications. In practice we compute the effective sample size ratio using sample proportions,

$$\frac{\left(\sum_{s=1}^S \hat{\pi}_s w(x_s; \beta_0)\right)^2}{\sum_{s=1}^S \hat{\pi}_s w(x_s; \beta_0)^2}, \quad \hat{\pi}_s = \frac{n_s}{n},$$

as a finite-sample analogue of the population quantity

$$\frac{\mathbb{E}[w(X; \beta_0)]^2}{\mathbb{E}[w(X; \beta_0)^2]} = \frac{\left(\sum_{s=1}^S \pi_s w(x_s; \beta_0)\right)^2}{\sum_{s=1}^S \pi_s w(x_s; \beta_0)^2}, \quad \pi_s = \frac{N_s}{N}.$$

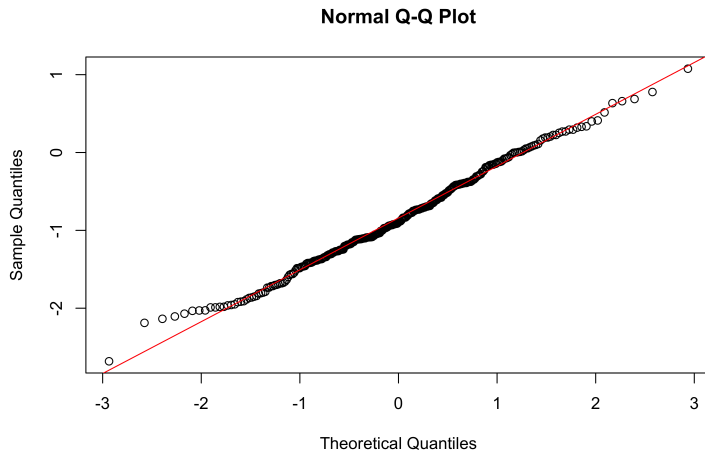
Since  $\hat{\pi}_s \rightarrow \pi_s$  as  $n \rightarrow \infty$ , the sample-based ratio converges to the population limit in Corollary 1. Table 7 reports the mean and interquartile range of two finite-sample analogues of the ESS ratio across 300 replications. The first row (“empirical ESS ratio”) is computed directly from the individual weights as  $\bar{w}_n(\hat{\beta})^2/\overline{w_n^2}(\hat{\beta})$ . The second row (“cell-based ESS ratio”) evaluates the population-style formula using sample proportions  $\hat{\pi}$  in place of  $\pi$ . Both estimators yield nearly

identical ESS ratios, with the cell-based version slightly lower on average. This close agreement demonstrates that the asymptotic limit in Corollary 1 provides an accurate approximation even in finite samples.

Table 7: Finite-sample ESS/ $n$  ratios: mean and IQR

	Mean	IQR [Q1, Q3]
Empirical ESS ratio $(\bar{w}_n(\hat{\beta})^2/\overline{w_n^2}(\hat{\beta}))$	0.38	[0.37, 0.39]
Cell-based ESS ratio (with $\hat{\pi}_s$ )	0.37	[0.36, 0.38]

To evaluate the finite-sample performance of Theorem 2, we computed the centered and scaled estimator  $\sqrt{n}(\hat{\theta} - \theta_0)$  across 300 replications. For asymptotic variance estimation, we use the unnormalized raw equivalent weights when computing the gradient term  $\hat{A}$ , while the normalized weights are used in the influence function  $\hat{\psi}_i$ . This normalization ensures that  $\hat{\theta}$  is expressed on the  $1/n$  scale required in Theorem 2. The distribution of the centered and scaled estimator closely follows the Gaussian limit, as shown in the Q–Q plot in Figure 2, with only mild deviations in the tails. The plug-in variance estimator  $\hat{V}$  from Theorem 2 was stable across replications, with mean 0.55 and interquartile range 0.53 to 0.57, and its values aligned closely with the empirical variance of 0.54. Taken together, the variance comparison and the Q–Q plot provide strong evidence that the asymptotic linear representation in Theorem 2 is accurate and that the plug-in variance estimator provides accurate finite-sample approximations to the asymptotic variance.



	Value
Empirical variance	0.54
Mean $\hat{V}$	0.55
IQR of $\hat{V}$	(0.53, 0.57)

Table 8: Comparison of empirical and plug-in variance estimates.

Figure 2: Normal Q–Q plot of the centered and scaled estimator  $\sqrt{n}(\hat{\theta} - \theta_0)$  across 300 replications.

### 3.2 Applications to FFCWS Study

We now apply our methods to the FFCWS study, using  $n = 3,442$  sampled units representing a population of 1,131,308 individuals. Both model-based methods produce negative weights, which we address by replacing them with the smallest positive weight. For the logistic model, the father’s interview status is used as the outcome variable since it has no missing values. When raking the model-based weights, we use the same four demographic variables as in the national FFCWS weights, excluding the city variable. We refer to the national FFCWS weights as the design-based weights, which are already raked in the FFCWS data.

Analyzing population counts of live births by education level (Table 9), we find that both the linear and logistic models closely approximate the population totals, with the logistic model showing slightly better alignment across categories. Effective sample size (ESS) comparisons in Table 10 show that the logistic model yields the highest ESS, both before and after raking. After raking, both model-based approaches preserve ESS more effectively than the design-based method, with the logistic model showing the minimal reduction and thus the strongest preservation of ESS. Overall, these empirical results are consistent with our simulation study, where model-based approaches, and logistic weighting in particular, demonstrated improved accuracy and greater stability relative to design-based weighting.

Table 9: Population counts of live births by education before raking

Education	<8 grade	Some HS	HS	Some College	College+	Total
Design-based	111,324	211,988	340,211	214,319	253,467	1,131,309
Linear model	73,232	267,083	367,476	242,049	181,469	1,131,309
Logistic model	60,214	207,301	316,517	258,784	288,491	1,131,309

Table 10: Effective Sample Size (ESS) for FFCWS Weights ( $n = 3442$ )

	Design-based	Linear	Logit	Linear-raked	Logit-raked
ESS	527	1273	1435	1082	1301

To assess the estimated population means, we evaluate Child Protective Services (CPS) contact prevalence across cities (Figure 3). Considering that the expected national prevalence of CPS contact at age 18 ranges from 0.3 to 0.6 across cities, the weighting methods generally lower the unweighted estimates, pulling them closer to the national range (Jung et al., 2025). Before raking,

the linear and logistic approaches differ across a few cities but overall display similar city-level patterns, while after raking their results become much closer. The design-based weights, once raked, also yield broadly comparable estimates, though with modest differences relative to the model-based approaches. Across both stages, logistic weighting generally produced the most accurate alignment with population benchmarks, reinforcing its advantage observed in the simulation study.

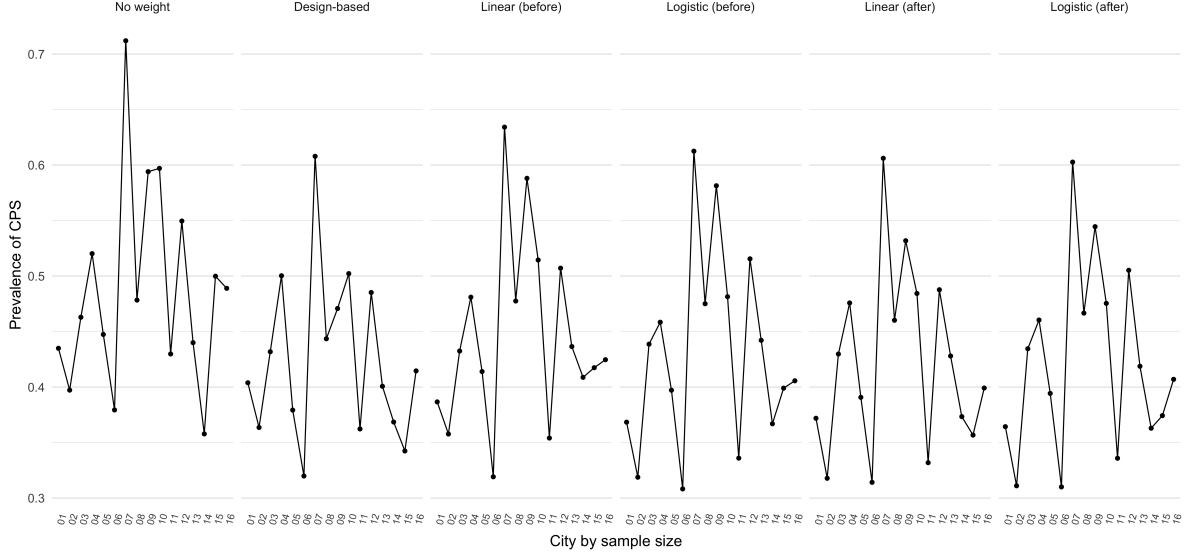


Figure 3: Prevalence of Child Protective Services (CPS) contact by city before and after raking

Beyond comparing weighting methods in terms of population alignment and effective sample size, we also illustrate how the logistic equivalent weights can be used for substantive inference in the FFCWS. Using these weights, the estimated prevalence of CPS contact is  $\hat{\theta} = 0.80$  with plug-in variance  $\hat{V} = 0.42$ , yielding a 95% confidence interval  $[0.78, 0.82]$  based on Theorem 2. This provides a direct empirical demonstration of our theoretical results: the plug-in variance formula delivers valid standard errors and confidence intervals from a single dataset, without requiring resampling or replication.

## 4 Asymptotic Properties

**Lemma 1** (Uniform constant envelope for logistic equivalent weights with categorical  $X$ ). *Let  $X$  consist entirely of categorical variables encoded as dummies, so that each  $X_i$  takes values in a finite*

set  $\mathcal{X}$  and  $\sup_{x \in \mathcal{X}} \|x\| < \infty$ . Define the closed-form logistic equivalent weights (as a function of  $\beta$ )

$$\hat{w}(\beta) = \left( \frac{n}{N} (N^{\text{pop}})^\top W^{\text{pop}}(\beta) X^{\text{pop}} ((X^\top W(\beta) X)^{-1} X^\top)^\top \right)^\top,$$

where  $W(\beta) = \text{diag}(\sigma(X\beta) \odot (1 - \sigma(X\beta)))$  and  $X^{\text{pop}}, N^{\text{pop}}$  are fixed finite matrices/vectors. Assume there exists a neighborhood  $\mathcal{N}$  of  $\beta_0$  and a constant  $c > 0$  such that the weighted information matrix  $H(\beta) := X^\top W(\beta) X$  satisfies  $\lambda_{\min}(H(\beta)) \geq c$  for all  $\beta \in \mathcal{N}$ . Then there exists  $C < \infty$  such that, for all  $\beta \in \mathcal{N}$  and all  $i$ ,

$$|\hat{w}_i(\beta)| \leq C.$$

**Lemma 2** (Continuity of logistic equivalent weights in  $\beta$ ). *Under the same assumptions as Lemma 1, the map  $\beta \mapsto \hat{w}(\beta)$  is continuous on  $\mathcal{N}$  (coordinatewise and in any matrix norm).*

**Theorem 1** (Plug-in LLN for logistic equivalent weights with categorical  $X$ ). *Let  $\{X_i\}_{i=1}^n$  be i.i.d. with law  $F$ . Let  $\hat{w}_i(\beta)$  denote the closed-form logistic equivalent weights defined in Lemma 1. Assume:*

1. (Envelope) *The assumptions of Lemma 1 hold, so there exists a constant  $C < \infty$  such that  $|\hat{w}_i(\beta)| \leq C$  for all  $\beta \in \mathcal{N}$  and all  $i$  (hence  $g(x) \equiv C$  is an integrable envelope with  $\mathbb{E}g < \infty$  and  $\mathbb{E}g^2 < \infty$ );*
2. (Continuity) *The assumptions of Lemma 2 hold, so  $\beta \mapsto \hat{w}(\beta)$  is continuous on the neighborhood  $\mathcal{N}$  of  $\beta_0$ ;*
3. (Compactness)  *$\mathcal{N}$  is compact and  $\beta_0 \in \text{int}(\mathcal{N})$ .*
4.  *$\hat{\beta} \xrightarrow{P} \beta_0$ ;*

Define

$$\bar{w}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \hat{w}_i(\beta), \quad \overline{w^2}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \hat{w}_i(\beta)^2.$$

Then

$$\bar{w}_n(\hat{\beta}) \xrightarrow{P} \mathbb{E}[\hat{w}_1(\beta_0)], \quad \overline{w^2}_n(\hat{\beta}) \xrightarrow{P} \mathbb{E}[\hat{w}_1(\beta_0)^2].$$

**Corollary 1** (ESS Limit). *Let  $\text{ESS}_n = (\sum_{i=1}^n w(X_i; \hat{\beta}))^2 / \sum_{i=1}^n w(X_i; \hat{\beta})^2$ . Under the assumptions of Theorem 1,*

$$\frac{\text{ESS}_n}{n} = \frac{\bar{w}_n(\hat{\beta})^2}{\bar{w}_n^2(\hat{\beta})} \xrightarrow{p} \frac{(\mathbb{E}[w(X; \beta_0)])^2}{\mathbb{E}[w(X; \beta_0)^2]}.$$

*Proof.* Apply Theorem 1 to  $\bar{w}_n(\hat{\beta})$  and  $\bar{w}_n^2(\hat{\beta})$ , then invoke the continuous mapping theorem.  $\square$

*Remark 1* (Scale invariance). If  $\tilde{w} = cw$  for any constant  $c > 0$  (e.g., normalization), then  $\text{ESS}_n/n$  is unchanged since both numerator and denominator are multiplied by  $c^2$ .

**Lemma 3** (Differentiability and uniform derivative envelope). *Under the assumptions of Lemmas 1–2, the map  $\beta \mapsto \hat{w}(\beta)$  is continuously differentiable ( $C^1$ ) on  $\mathcal{N}$ . Moreover, there exists a finite constant  $C'$  (depending only on  $X, X^{\text{pop}}, N^{\text{pop}}, c$ ) such that*

$$\sup_{\beta \in \mathcal{N}} \|\nabla_{\beta} \hat{w}_i(\beta)\| \leq C' \quad \text{for all } i.$$

**Theorem 2** (Asymptotic linearity of  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{w}_i(\hat{\beta}) Y_i$ ). *Let  $\hat{w}_i(\beta)$  be the logistic equivalent weights. Assume:*

- (i) (Constant envelope for weights) Lemma 1 holds:  $|\hat{w}_i(\beta)| \leq C$  for all  $\beta \in \mathcal{N}$  and all  $i$ .
- (ii) ( $C^1$  and uniform derivative envelope) Lemma 3 holds:  $\beta \mapsto \hat{w}(\beta)$  is  $C^1$  on  $\mathcal{N}$  and there exists  $C' < \infty$  such that  $\sup_{\beta \in \mathcal{N}} \|\nabla_{\beta} \hat{w}_i(\beta)\| \leq C'$  for all  $i$ .
- (iii) (Logistic MLE regularity) The logistic model is correctly specified at  $\beta_0$  with  $p_0(x) = \sigma(x^\top \beta_0)$ ;  $\beta_0$  is an interior point of the parameter space;  $X$  has  $\mathbb{E}\|X\|^2 < \infty$ ; and the Fisher information  $\mathcal{I}(\beta_0) = \mathbb{E}[XX^\top p_0(X)\{1 - p_0(X)\}]$  is positive definite.

Define  $g(\beta) = \mathbb{E}[\hat{w}_1(\beta) Y_1]$  and

$$A := \nabla_{\beta} g(\beta) \big|_{\beta=\beta_0} = \mathbb{E}[\nabla_{\beta} \hat{w}_1(\beta_0) Y_1] \quad (\text{under correct logit spec., } = \mathbb{E}[\nabla_{\beta} \hat{w}_1(\beta_0) p_0(X_1)]).$$

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \underbrace{\left\{ \hat{w}_i(\beta_0) Y_i - \mathbb{E}[\hat{w}_1(\beta_0) Y_1] \right\}}_{\psi_1(Z_i)} + A^\top \sqrt{n}(\hat{\beta} - \beta_0) + o_p(1),$$

so that

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V), \quad V = \mathbb{E}[\psi(Z)^2],$$

with influence function

$$\psi(Z) = \{\hat{w}(\beta_0)Y - \mathbb{E}[\hat{w}(\beta_0)Y]\} + A^\top \mathcal{I}^{-1}X \{Y - p_0(X)\}.$$

A consistent plug-in variance is

$$\hat{V} = \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i^2, \quad \hat{\psi}_i = \hat{w}_i(\hat{\beta})Y_i - \bar{m}_n + \hat{A}^\top \hat{\mathcal{I}}^{-1}X_i \{Y_i - \hat{p}_i\},$$

where

$$\bar{m}_n = \frac{1}{n} \sum_{j=1}^n \hat{w}_j(\hat{\beta})Y_j, \quad \hat{p}_i = \sigma(X_i^\top \hat{\beta}), \quad \hat{\mathcal{I}} = \frac{1}{n} \sum_{j=1}^n X_j X_j^\top \hat{p}_j(1 - \hat{p}_j), \quad \hat{A} = \frac{1}{n} \sum_{j=1}^n \nabla_\beta \hat{w}_j(\hat{\beta}) \hat{p}_j.$$

## 5 Discussion

### 5.1 Key Findings and Limitations

Our study provides a detailed comparison of model-based and design-based weighting methods, evaluating their performance in terms of accuracy, efficiency, and effective sample size (ESS) retention. Across simulations, both model-based approaches—linear and logistic weighting—consistently outperformed design-based weighting in regression analyses, achieving lower mean squared error (MSE) of regression coefficients and more stable results across replications. A key distinction emerged in ESS preservation: before raking, the design-based method had the largest ESS, but after raking, both model-based approaches retained substantially more of their efficiency, with logistic weighting showing the strongest preservation.

In the application to the FFCWS, these patterns carried over. Logistic weighting produced the highest ESS both before and after raking, with linear weighting close behind, while the design-based weights showed a sharper decline. For substantive outcome, all weighting methods shifted unweighted prevalence estimates of Child Protective Services (CPS) contact toward the expected national range, but logistic weighting generally yielded the most accurate alignment with population

benchmarks. Before raking, linear and logistic approaches displayed some differences across cities, whereas after raking their results were nearly identical, with design-based weights producing broadly comparable but modestly different patterns. Taken together, the empirical results strongly reinforce the simulation findings: logistic weighting provides the greatest overall stability and accuracy among the methods considered.

These findings underscore the superiority of model-based weighting, particularly logistic regression weighting, relative to traditional design-based methods. Beyond empirical performance, our theoretical results show that the effective sample size (ESS) converges to a well-defined population limit and that the weighted estimator admits an asymptotic linear representation, yielding asymptotic normality with a consistent plug-in variance formula. The alignment of the simulation and application findings with the asymptotic theory highlights its usefulness in practice, demonstrating that the large-sample results provide meaningful guidance even in realistic finite-sample settings. To our knowledge, this is the first work to derive survey weights directly from logistic regression score equations, providing both methodological innovation and theoretical justification.

Several limitations warrant discussion. First, regression-based weights rely on models that may be strained when the number of poststratification cells approaches or exceeds the sample size, a situation common in large surveys. Second, our model-based approach does not incorporate richer calibration targets, such as auxiliary population totals or outcome benchmarks, which could further improve efficiency. This may also help explain the modest biases we observed in certain population statistics under logistic weights, as our simulations generated binary outcomes that depend on both weighting and non-weighting variables. Finally, although our asymptotic results provide theoretical justification, finite-sample performance remains sensitive to design choices, such as the definition of poststratification cells and the prevalence of binary outcomes.

## 5.2 Future Research

Our proposed approach does not rely on prior information about the population; weights are constructed solely from observed weighting variables without incorporating population margins as priors. Future research could therefore explore methods that extend beyond this framework by integrating population information more explicitly. One promising direction is Bayesian hierarchical models, where population margins could be incorporated as priors while maintaining computational

efficiency. Hierarchical Bayesian models that embed population priors directly may enhance the robustness of weight estimates and improve representativeness of the weighted sample. At the same time, the complexity of such models, particularly with numerous demographic and geographic categories, raises challenges in terms of computation and convergence.

Regression-based weights rely on fitted models that may be strained when the number of post-stratification cells approaches or exceeds the sample size, a situation common in large surveys. Future work could address this challenge through regularization or hierarchical pooling strategies that stabilize estimation in high-dimensional settings. A complementary direction involves nonparametric methods such as Gaussian Process (GP) regression, which provide natural regularization through their kernel structure while offering flexibility to capture nonlinear relationships between covariates and outcomes in the estimation of poststratification cells. An open question is how to incorporate population information in these models without losing predictive accuracy or risking overparameterization; one possible strategy is hierarchical modeling with structured priors that connect observed data to known population totals.

Finally, resampling techniques such as stratified bootstrapping for survey data could be further developed to quantify uncertainty in weight estimates while maintaining alignment with population characteristics. Such approaches may provide more robust tools for addressing variability and bias in survey weighting.

## Appendix: Proofs

### Proof of Lemma 1

*Proof.* The logistic variance satisfies  $\sigma(t)\{1 - \sigma(t)\} \leq 1/4$  for all  $t$ , so  $\|W(\beta)\|_{\text{op}} \leq 1/4$ . The matrices  $X^{\text{pop}}$  and  $N^{\text{pop}}$  are fixed, hence have finite operator norms. Because  $H(\beta)$  is symmetric positive definite with  $\lambda_{\min}(H(\beta)) \geq c$ , its spectral norm satisfies

$$\|H(\beta)^{-1}\| = \frac{1}{\lambda_{\min}(H(\beta))} \leq \frac{1}{c}, \quad (3)$$

for all  $\beta \in \mathcal{N}$ . Since  $X_i \in \mathcal{X}$  takes values in a finite set of dummy vectors,  $\|X\|$  and  $\|X^\top\|$  are uniformly bounded. Thus in the general form one might write

$$|\hat{w}_i(\beta)| \leq C(1 + \|X_i\|^{m'}),$$

but because  $\|X_i\|$  is uniformly bounded this polynomial term can be absorbed into a constant, yielding the sharper bound

$$|\hat{w}_i(\beta)| \leq C',$$

for some finite  $C'$  depending only on  $(X, X^{\text{pop}}, N^{\text{pop}}, c)$ .  $\square$

## Proof of Lemma 2

*Proof.* The logistic map  $\sigma(t) = 1/(1 + e^{-t})$  is infinitely differentiable on  $\mathbb{R}$ , hence continuous. Therefore  $\beta \mapsto \sigma(X^{\text{pop}}\beta)$  and  $\beta \mapsto W(\beta)$  are continuous. Matrix multiplication and the Hadamard product are continuous operations, so  $\beta \mapsto X^\top W(\beta)X = H(\beta)$  is continuous. Now, because  $\lambda_{\min}(H(\beta)) \geq c > 0$  for all  $\beta \in \mathcal{N}$ , the set  $\{H(\beta) : \beta \in \mathcal{N}\}$  consists entirely of nonsingular matrices bounded away from singularity. On this set, the matrix inversion map  $A \mapsto A^{-1}$  is continuous. Therefore  $\beta \mapsto H(\beta)^{-1}$  is continuous on  $\mathcal{N}$ . Finally, composing continuous maps and multiplying on the left and right by the fixed matrices  $X^{\text{pop}}$  and  $(N^{\text{pop}})^\top$  preserves continuity. Hence  $\beta \mapsto \hat{w}(\beta)$  is continuous on  $\mathcal{N}$ .  $\square$

## Proof of Theorem 1

*Proof.* Write

$$\bar{w}_n(\hat{\beta}) - \mathbb{E}[\hat{w}_1(\beta_0)] = \underbrace{\left(\bar{w}_n(\hat{\beta}) - \mathbb{E}[\hat{w}_1(\hat{\beta})]\right)}_{(A)} + \underbrace{\left(\mathbb{E}[\hat{w}_1(\hat{\beta})] - \mathbb{E}[\hat{w}_1(\beta_0)]\right)}_{(B)}.$$

**(A) Uniform LLN.** Consider the class  $\mathcal{F} = \{\hat{w}(\cdot; \beta) : \beta \in \mathcal{N}\}$ . By Lemma 2,  $\beta \mapsto \hat{w}(\beta)$  is continuous; by Lemma 1,  $|\hat{w}_i(\beta)| \leq C$  for all  $\beta \in \mathcal{N}$ , so  $g \equiv C$  is an integrable envelope. With continuity and compact  $\mathcal{N}$ ,  $\mathcal{F}$  is totally bounded in  $L_1(F)$  and, with envelope  $g$ , is Glivenko–Cantelli.

By the uniform weak law of large numbers (Wald, 1949; van der Vaart, 1998),

$$\sup_{\beta \in \mathcal{N}} \left| \bar{w}_n(\beta) - \mathbb{E}[\hat{w}_1(\beta)] \right| \xrightarrow{P} 0.$$

Since  $\hat{\beta} \xrightarrow{P} \beta_0 \in \text{int}(\mathcal{N})$ , we have (A)  $\xrightarrow{P} 0$ .

**(B) Dominated convergence.** By Lemma 2,  $\hat{w}_1(\hat{\beta}) \rightarrow \hat{w}_1(\beta_0)$  almost surely as  $\hat{\beta} \rightarrow \beta_0$ , and by Lemma 1,  $|\hat{w}_1(\hat{\beta})| \leq g(X_1) \equiv C$  with  $\mathbb{E}[g] < \infty$ . Therefore, by dominated convergence,  $\mathbb{E}[\hat{w}_1(\hat{\beta})] \rightarrow \mathbb{E}[\hat{w}_1(\beta_0)]$ , so (B)  $\rightarrow 0$ .

Combining (A) and (B) gives  $\bar{w}_n(\hat{\beta}) \xrightarrow{P} \mathbb{E}[\hat{w}_1(\beta_0)]$ .

The argument for  $\overline{w^2}_n(\hat{\beta})$  is identical, applied to the class  $\{\hat{w}(\cdot; \beta)^2 : \beta \in \mathcal{N}\}$  with envelope  $g^2 \equiv C^2$  (and  $\mathbb{E}[g^2] < \infty$ ), yielding  $\overline{w^2}_n(\hat{\beta}) \xrightarrow{P} \mathbb{E}[\hat{w}_1(\beta_0)^2]$ .  $\square$

### Proof of Lemma 3

*Proof.* By Lemmas 1 and 2,  $\hat{w}(\beta)$  is bounded and continuous on  $\mathcal{N}$ . Since the logistic map  $\sigma$  is  $C^\infty$ , we have  $W(\beta) = \text{diag}(\sigma(X\beta) \odot (1 - \sigma(X\beta))) \in C^1$  and  $H(\beta) = X^\top W(\beta)X \in C^1$ . On  $\mathcal{N}$ ,  $\lambda_{\min}(H(\beta)) \geq c > 0$ , hence the inversion map is  $C^1$  with

$$D(A^{-1})[\Delta] = -A^{-1}\Delta A^{-1} \quad (\text{Fréchet derivative}). \quad (4)$$

**Step 1: Bound  $D(H^{-1})$ .** For a direction  $h \in \mathbb{R}^p$ ,

$$DH(\beta)[h] = X^\top DW(\beta)[h]X,$$

and, writing  $p_i = \sigma(x_i^\top \beta)$ ,

$$(DW(\beta)[h])_{ii} = p_i(1 - p_i)(1 - 2p_i)(x_i^\top h).$$

Hence  $\|DW(\beta)[h]\| \leq \frac{1}{4} \|X\| \|h\|$ , and by submultiplicativity

$$\|DH(\beta)[h]\| \leq C_X \|h\|, \quad C_X := \frac{1}{4} \|X^\top\| \|X\|^2. \quad (5)$$

Using (4),

$$D(H(\beta)^{-1})[h] = -H(\beta)^{-1} DH(\beta)[h] H(\beta)^{-1},$$

so

$$\|D(H(\beta)^{-1})[h]\| \leq \|H(\beta)^{-1}\|^2 \|DH(\beta)[h]\| \leq \|H(\beta)^{-1}\|^2 C_X \|h\|, \quad (6)$$

with

$$\|H(\beta)^{-1}\| = \frac{1}{\lambda_{\min}(H(\beta))} \leq \frac{1}{c}, \quad (7)$$

and combining (6)–(7) yields

$$\|D(H(\beta)^{-1})[h]\| \leq c^{-2} C_X \|h\|. \quad (8)$$

**Step 2: Bound  $D\hat{w}(\beta)[h]$ .** Write  $\hat{w}(\beta) = \frac{n}{N} T(\beta)^\top S(\beta)$  with

$$T(\beta) := H(\beta)^{-1} X^\top, \quad S(\beta) := X^{\text{pop}\top} W^{\text{pop}}(\beta) N^{\text{pop}}.$$

By product rule,

$$D\hat{w}(\beta)[h] = \frac{n}{N} \left( DT(\beta)[h]^\top S(\beta) + T(\beta)^\top DS(\beta)[h] \right).$$

Using (8),

$$\|DT(\beta)[h]\| \leq c^{-2} C_X \|X^\top\| \|h\|, \quad \|T(\beta)\| \leq \|H(\beta)^{-1}\| \|X^\top\| \leq c^{-1} \|X^\top\|.$$

Since  $X^{\text{pop}}, N^{\text{pop}}$  are fixed and  $\sigma(u)(1 - \sigma(u)) \leq \frac{1}{4}$ , there exist finite constants  $C_1, C_2$  (depending only on  $X^{\text{pop}}, N^{\text{pop}}$  and logistic bounds) such that

$$\|S(\beta)\| \leq C_1, \quad \|DS(\beta)[h]\| \leq C_2 \|h\|.$$

Therefore, by submultiplicativity,

$$\|D\hat{w}(\beta)[h]\| \leq \frac{n}{N} \left( (c^{-2} C_X \|X^\top\|) C_1 + (c^{-1} \|X^\top\|) C_2 \right) \|h\| =: C' \|h\|. \quad (9)$$

**Step 3: From directional to gradient bound (each coordinate).** For the  $i$ -th coordinate  $f(\beta) = \hat{w}_i(\beta)$  (a scalar),

$$|D\hat{w}_i(\beta)[h]| \leq \|D\hat{w}(\beta)[h]\| \leq C' \|h\|.$$

Taking  $\sup_{\|h\|=1}$  gives

$$\|\nabla_{\beta}\hat{w}_i(\beta)\| = \sup_{\|h\|=1} |D\hat{w}_i(\beta)[h]| \leq C' \quad \text{for all } \beta \in \mathcal{N} \text{ and all } i,$$

which is the desired uniform derivative envelope. Finally, since all components used are  $C^1$  and  $H(\beta)^{-1}$  is  $C^1$  on  $\mathcal{N}$ , it follows that  $\beta \mapsto \hat{w}(\beta)$  is  $C^1$  as well (part (i)). This completes the proof of the lemma.  $\square$

## Proof of Theorem 2

*Proof.* Write

$$\hat{\theta} = P_n\{\hat{w}(\hat{\beta})Y\}, \quad \theta_0 = P\{\hat{w}(\beta_0)Y\},$$

hence

$$\hat{\theta} - \theta_0 = \underbrace{\{P_n - P\}[\hat{w}(\beta_0)Y]}_{(I)} + \underbrace{(P\hat{w}(\hat{\beta})Y - P\hat{w}(\beta_0)Y)}_{(II)} + \underbrace{\{(P_n\hat{w}(\hat{\beta})Y - P\hat{w}(\hat{\beta})Y) - (P_n\hat{w}(\beta_0)Y - P\hat{w}(\beta_0)Y)\}}_{(III)}.$$

(I) By Lemma 1,  $\hat{w}(\beta_0)Y$  has an integrable constant envelope; the i.i.d. CLT gives

$$\sqrt{n}(I) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\hat{w}_i(\beta_0)Y_i - P[\hat{w}(\beta_0)Y]\} \Rightarrow \mathcal{N}(0, \mathbb{V}[\hat{w}(\beta_0)Y]).$$

(II) By Lemma 3,  $\beta \mapsto \hat{w}(\beta)$  is  $C^1$  on  $\mathcal{N}$ ; therefore  $g(\beta) = P[\hat{w}(\beta)Y]$  is differentiable at  $\beta_0$  with gradient  $A = P[\nabla_{\beta}\hat{w}(\beta_0)Y]$ . A first-order expansion yields  $g(\hat{\beta}) - g(\beta_0) = A^{\top}(\hat{\beta} - \beta_0) + o_p(\|\hat{\beta} - \beta_0\|)$ .

By (iii) and the influence function representation of the MLE in logistic regression (Hampel et al.,

1986; van der Vaart, 1998, Thm. 5.23),

$$\begin{aligned}\sqrt{n}(II) &= A^\top \sqrt{n}(\hat{\beta} - \beta_0) + o_p(1) \\ &= A^\top \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathcal{I}^{-1} X_i \{Y_i - p_0(X_i)\} \right) + o_p(1).\end{aligned}$$

(III) By Lemma 1, the class  $\mathcal{F} := \{\hat{w}(\cdot; \beta)Y : \beta \in \mathcal{N}\}$  has a constant envelope. By Lemma 3,  $\beta \mapsto \hat{w}(\beta)$  is  $C^1$  on  $\mathcal{N}$  with  $\sup_{\beta \in \mathcal{N}} \|\nabla_\beta \hat{w}_i(\beta)\| \leq C'$ , so the parametrization  $\beta \mapsto \hat{w}(\cdot; \beta)Y$  is Lipschitz in  $\beta$ , uniformly on  $\mathcal{N}$ . Restricting to a compact neighborhood of  $\beta_0$ , the class is finite-dimensional, uniformly bounded, and  $L^2(P)$ -Lipschitz in  $\beta$ , and is therefore  $P$ -Donsker. Since we also have  $\|\hat{w}(\cdot; \hat{\beta}) - \hat{w}(\cdot; \beta_0)\|_{L^2(P)} \xrightarrow{P} 0$ , by (van der Vaart, 1998, Lemma 19.24),

$$\sqrt{n}(III) = \sqrt{n}(P_n - P)[\hat{w}(\hat{\beta})Y - \hat{w}(\beta_0)Y] = o_p(1).$$

Combining (I)–(III) gives the stated linear expansion and asymptotic variance. Consistency of  $\hat{V}$  follows by LLN and continuous mapping.  $\square$

## References

- Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Wiley, Hoboken.
- Carlson, B. L. (2008). *Fragile Families & Child Wellbeing Study: Methodology for Constructing Mother, Father, and Couple Weights for Core Telephone Public Survey Data Waves 1–4*. Princeton University, Technical Report.
- Chapman, D. W., Bailey, L. and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology* **12**, 161–180.
- Chu, A. and Goldman, J. (1997). Weighting procedures for USDA’s Continuing Survey of Food Intakes by Individuals (1994–96). In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 802–807.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**(418), 376–382.

- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science* **22**(2), 153–164.
- Hájek, J. (1971). Comment on “An essay on the logical foundations of survey sampling” by D. Basu. In V. P. Godambe and D. A. Sprott (eds.), *Foundations of Statistical Inference*, 236–242. Toronto: Holt, Rinehart and Winston.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
- 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.
- Haziza, D. and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science* **32**(2), 206–220.
- Jung, H.-M., Berger, L. M., Dickerson, T., Gelman, A., Lee, S., Thomas, M., and Waldfogel, J. (2025). Adjusting for underreporting of child protective services involvement in the Future of Families and Child Wellbeing Study and assessing its empirical implications through illustrative analyses of young adult disconnection. *Social Service Review*, forthcoming.
- Lee, S. and Gelman, A. (2024). Year 22 Future of Families Survey Weight Adjustment. Manuscript.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Springer, New York.
- Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20**(4), 595–601.
- Valliant, R., Dever, J. A. and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Chapman & Hall/CRC, Boca Raton.
- Valliant, R. and Dever, J. A. (2018). *Survey Weights: A Step-by-Step Guide to Calculation*. Stata Press, College Station.
- Skinner, C. J. and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science* **32**(2), 165–175.