Model-based Survey Weighting Using Logistic Regression

Seonghun Lee

Abstract. Survey weighting is a crucial step in obtaining representative estimates, yet base weight construction often remains ad hoc. Traditional approaches rely on design-based weights derived from known selection probabilities, but these can be inefficient when auxiliary information is available. Gelman (2007) proposed a model-based method using linear regression to construct weights, conceptually aligning with the Hájek estimator when renormalizing observations for improved population inference. Building on this framework, we develop a model-based approach that employs logistic regression for weight construction. However, logistic regression lacks linear translation invariance. To address this, we approximate the weights using a Taylor series expansion. While model-based methods offer a structured framework, their practical advantages over traditional design-based approaches remain underexplored. We fill this gap by constructing base weights from scratch and comparing logistic method to design-based and linear regression methods. Our findings provide insights into the advantages and limitations of model-based weighting, offering guidance for researchers navigating survey weight construction in the presence of missingness and nonresponse.

1. Introduction

Constructing base weights or adjusting survey weights involves a series of intricate steps requiring user-defined choices to make samples representative of the target population. These steps often depend on leveraging auxiliary data to create weights that reproduce census counts or population values. However, survey data frequently encounter challenges such as missingness and non-response, which further complicate the process of weight adjustment.

The existing body of literature extensively covers topics such as survey weighting adjustments for non-response, design-based or model-based inference, and methods for 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; Hazia and Beaumont, 2017; Skinner and Wakefield, 2017, Chen et al; 2017; Liu et al, 2023). Despite this, there is limited research on the methods for constructing base weights from scratch or comparing these methods through rigorous simulation studies. This gap arises because the construction of base weights is often a messy process, filled with ad hoc approaches and lacking consensus among practitioners. Furthermore, accessing complete census data to learn population values across necessary demographic variables is challenging due to confidentiality concerns, adding to the complexity of creating reliable base weights (Carlson, 2008).

Moreover, many researchers across disciplines grapple with the question of whether using survey weights is beneficial in regression analysis (Miller, 2011). The use of weights does not always reduce bias, as adjustments for non-response and calibration through raking can lead to extreme weights even after trimming. In practice, forcing total weights to match population values at the end of the adjustment process can nullify earlier efforts and cause significant deviations from the original base weights (Deville and Särndal, 1992). This often leaves researchers questioning the efficacy of using weights in achieving more accurate or unbiased estimates (Little and Rubin, 1987).

Survey weights are traditionally constructed using design-based methods, and there is a body of literature providing guidelines for these approaches (Chu and Goldman, 1997; Valliant et al, 2013; Valliant and Dever, 2018). An alternative to these design-based methods involves model-based procedures for weight construct. For instance, Gelman (2007) proposed a model-based approach that utilizes linear regression to construct weights. While these methods provide a more structured framework for weight creation, their practical implementation and advantages over traditional design-based methods remain relatively underexplored.

In this paper, we aim to address this gap by constructing base weights from scratch and proposing a novel model-based approach that employs logistic regression for weight construction. Our work builds on Gelman (2007) to explicitly connect these procedures with practical applications in survey weighting. We then compare weights constructed by three different methods and present our findings through simulation studies. By doing so, we aim to provide deeper insights into the process of weight construction and offer practical guidance for researchers facing challenges in survey weighting.

1.1 Target Study Population and Design

The target survey population consists of live births occurring in in large U.S. cities with population over 200,000 between 1998 and 2000. This focus is motivated by the process of constructing base weights of the Future of Families and Child Wellbeing Study (FFCWS). The FFCWS sample is a stratified and multistage design with 4,898 children, oversampling births to unmarried mothers at a ratio of 3 to 1 with the inclusion of a large number of Black, Hispanic, and low-income families. Follow-up interviews were conducted across seven waves, when children were approximately ages 1, 3, 5, 9, 15, and 22. In constructing the FFCWS weights, four demographic variables were used for poststratification, and geographical information was used to estimate population birth counts using the Centers for Disease Control and Prevention (CDC) annual natality data. These variables are mother's marital status, race/ethnicity, age, education, and city of birth (Carlson, 2008). To mirror the FFCWS study, we generate data and base weights in our simulation analysis that replicate the characteristics of the FFCWS data.

In large national surveys, stratified multistage cluster sampling (SMCS) is a common approach because it balances logistical challenges with the need to obtain sufficiently precise estimates for key subgroups. In this context, clusters, also referred to as primary sampling units (PSUs), are often used to simplify fieldwork. For example, the stratum is the city and the PSU is the hospital for cities selected with certainty in the FFCWS study. A simpler alternative is stratified simple random sampling (SRS), where each unit in the population is assigned a group label, and SRS is conducted within each group. These groups typically correspond to demographic strata or geographical regions. In our simulation study, we choose to construct design-based weights using stratified simple random sampling (SRS) rather than stratified multistage cluster sampling (SMCS), as intra-cluster correlation in SMCS reduces efficiency by limiting the independent information gained from additional sampled units within the same cluster. This leads to higher standard errors and lower overall precision. Moreover, we expect small differences in weight estimation between SMCS and SRS, particularly after raking.

1.2 Design-based Weighting

Base weights are computed when the sample is a probability sample drawn from a finite population. A probability sample is realized under the following conditions: the set of all possible samples that can be selected from a finite population is well-defined given a specified sampling procedure, and each possible sample is associated with a known probability of selection. Moreover, every unit in the target population has non-zero probability of being select with selection occurring via a random mechanism (Valliant and Dever, 2018). In our simulation study, the first step in constructing base weights is to assign the selection probability π_i to each unit in the population. The selection probability is determined based on factors such as the city of birth and demographic characteristics. These factors are used to form cells, where units within the same cell share an identical probability of selection. The base weight for unit i is then computed as the inverse of the unit's selection probability, $w_i = 1/\pi_i$.

The next step is to adjust the weight upward for survey non-respondents. The nonresponse adjustment is calculated as the inverse of the weighted response rate within each city, which is then is multiplied by the base weight of respondents to renormalize their weights, ensuring they sum to the total population weight. Given the extensive body of existing literature on nonresponse adjustment techniques, this study does not delve in to these methods in detail. For the latest methods on adjusting for non-response in a two-stage process using the FFCWS data, including a state-of-the-art approach utilizing an optimally balanced Gaussian process (Vegetabile et al, 2020), refer to Lee and Gelman (2024). Finally, we adjust the weights using raking to poststratify them, ensuring that the weighted sample counts align with the known population counts within each raking cell. Although the adjustment occurs within individual raking cells, the raking process only requires known marginal population totals for a single variable, rather than for each crossed combination of variables. Consequently, we do not need to be concerned with empty cells that may arise from crossing all the variables used in the poststratification process.

Table 1. Summary of base weights

Min	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Max**
12.59	82.12	172.42	333.33	386.64	6167.84

1.3 Estimating Population Mean from a Sample

We have the poststratification variables X whose joint distribution in the population is known along with the survey response of interest y that we aim to estimate for the population. The possible strata of X form the poststratification cells, denoted as s, with population sizes N_s and sample sizes n_s. The total population size is $N = \sum_{s=1}^{S} N_s$ and the sample size is $n = \sum_{s=1}^{S} n_s$. The population mean of the survey response y is then defined as the weighted average of the means within each stratum:

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

where θ_s is the mean of y for strata s. The corresponding estimate is

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

When estimating the population mean from sample survey data, the Horvitz-Thompson estimator or the Hájek estimator can be used.

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

$$\bar{y}_{HJ} = \frac{\sum_{i=1}^{n} w_i y_i}{\hat{N}}$$

where $\hat{N} = \sum_{i=1}^{n} w_i$. In our context, where N is known and $N = \hat{N}$, the Horvitz-Thompson estimator and the Hájek estimator are equivalent, and both serve as unbiased estimators for the population mean (Horvitz and Thompson, 1952; Hájek, 1971). Notice that the poststratified estimator $\hat{\theta}^{PS}$ is mathematically equivalent to the Hájek estimator when estimating the population mean. We will leverage this relationship to estimate equivalent unit weights using model-based approaches.

2. Model-based Weighting Methods

2.1 Linear Regression Weight

Let X denote $n \times k$ design matrix of raking variables for the samples, and let X^{pop} denote the $S \times k$ design matrix for the population poststratification cells. If the survey response of interest y has a linear relationship with the raking variables, then $\hat{\theta}_s = X_s^{\text{pop}} \hat{\beta}$, where the estimated vector of linear regression coefficients $\hat{\beta} = (X^t X)^{-1} X^t y$. Then, the poststratified estimate of the population mean

$$
\begin{aligned}
\hat{\theta}^{PS} &= \frac{1}{N}\sum_{s=1}^{S} N_s\left(X_s^{\text{pop}}\hat{\beta}\right) \\
&= \frac{1}{N}(N^{pop})^t X^{\text{pop}}(X^t X)^{-1}X^t y
\end{aligned}
$$

Now, define $\widehat{w}$ as the equivalent unit weights such that $\hat{\theta}^{PS} = \frac{\sum_{i=1}^{n}\widehat{w}_i y_i}{n}$ and $\sum_{i=1}^{n}\widehat{w}_i = n$. Solving for $\widehat{w}$ yields:

$$
\widehat{w} = (\frac{n}{N}(N^{pop})^t X^{\text{pop}}(X^t X)^{-1}X^t)^t.
$$

When solving for $\widehat{w}$, the survey response y cancels out and therefore does not contribute to the weight estimation. This is a significant advantage, as it eliminates the need to account for y, which often contains missing values due to non-participation or non-response. In contrast, raking variables, such as demographic and geographical information, are typically available with minimal missingness, making them reliable for constructing the weights. The formula produces S unique equivalent unit weights, which sum to n. Note that rescaling the estimated weights–by multiplying any constant–does not affect the estimation of the population mean. Thus, the weights can be renormalized to sum to the known population size, ensuring that $\sum_{i=1}^{n}\widehat{w}_i = \widehat{N}$. After this renormalization, the population estimate becomes $\hat{\theta}^{PS} = \frac{\sum_{i=1}^{n}\widehat{w}_i y_i}{\widehat{N}}$, which corresponds to a form of the Hájek estimator.

Table 2. Summary of linear equivalent weights

Min	**1st Quantile**	**Median**	**Mean**	**3rd Quantile**	**Max**
0.0044	0.12	0.79	1.00	1.50	7.35

2.2 Logistic Regression Weight

Unlike linear regression, weight estimation using logistic regression does not have a closed-form solution, as the linear assumption that enables such solutions, like a closed-form expression for $\hat{\beta}$, no longer apply. Consequently, it becomes necessary to approximate the weights. One approach we use is to compute the derivative using a Taylor series approximation to simplify the estimation process. Let $y = \{y_1, \ldots, y_n\}$ be a vector of any binary survey response, so the population parameter of interest, θ, represents the population prevalence, and the sample estimates $\hat{\theta}_s$ correspond to cell-level prevalences. The regression model is specified as $\text{logit}\left(\frac{p}{1-p}\right) = X\beta$ where $p = \text{P}(y = 1|X)$. The predicted probabilities are then $\hat{p} = \sigma\left(X\hat{\beta}\right)$ where $\sigma(x) = \frac{\exp(x)}{1+\exp(x)}$ is a sigmoid function. Using this model, the poststratified estimate of the population prevalence θ^{PS} is defined as:

$$\begin{aligned}\hat{\theta}^{PS} &= \frac{1}{N}\sum_{s=1}^{S} N_s\,\hat{p}_s \\ &= \frac{1}{N}\sum_{s=1}^{S} N_s\,\sigma\left(X_s^{pop}\hat{\beta}\right)\end{aligned}$$

where N_s represents the population counts for each poststratification cell s, and X_s^{pop} is the design matrix for the population. Now, define $\hat{w}$ as the equivalent unit weights such that $\hat{\theta}^{PS} = \frac{\sum_{i=1}^{n}\hat{w}_i y_i}{n}$. Equating the two forms of the estimator, we have:

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

where $\widehat{w}$ represents the equivalent unit weights for the sample units. This equation establishes the relationship between the logistic regression model for prevalence and the calculation of equivalent weights.

Now, taking n partial derivatives with respect to y on both sides yields:

$$(\nabla\widehat{w})y + \widehat{w} = (\frac{n}{N}(N^{pop})^t\sigma\left(X^{pop}\hat{\beta}\right)\left(1-\sigma\left(X^{pop}\hat{\beta}\right)\right)^t X^{pop}\left(\frac{\mathrm{d}\hat{\beta}}{\mathrm{d}y}\right)^t)^t$$

The left-hand side equation is

$$(\nabla\widehat{w})y + \widehat{w} = \begin{pmatrix} \frac{dw_1(y)}{dy_1} & \cdots & \frac{dw_1(y)}{dy_n} \\ \vdots & \cdots & \vdots \\ \frac{dw_n(y)}{dy_1} & \cdots & \frac{dw_n(y)}{dy_n} \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} w_1(y) \\ w_2(y) \\ \vdots \\ \vdots \\ w_n(y) \end{pmatrix}$$

The term $\frac{\mathrm{d}\hat{\beta}}{\mathrm{d}y}$ on the right-hand side, derived using implicit differentiation, is expressed as:

$$\frac{\mathrm{d}\hat{\beta}}{\mathrm{d}y} = (X^tWX)^{-1}X^t$$

where X^tWX is the Hessian matrix of the log-likelihood function in logistic regression, capturing the curvature of the likelihood function around the maximum likelihood estimate (MLE). Here, W is a diagonal matrix with each element given by $\sigma\left(X\hat{\beta}\right)\left(1-\sigma\left(X\hat{\beta}\right)\right)$, representing the variance of the predicted probabilities for the sample units. The formula for $(\nabla\widehat{w})y + \widehat{w}$ produces S unique weights.

Next, we aim to estimate $(\nabla\widehat{w})y$ and subtract it to obtain the equivalent unit weights. Unfortunately, there is no closed-form solution for $(\nabla\widehat{w})y$, so we approximate it using a Taylor series expansion. It is important to note that $(\nabla\widehat{w})y = 0$ for $y_i = 0$, meaning we need to correct weights for units where $y_i = 1$, many of which have negative weights. After the correction, we observe that most of these negative weights turn positive, demonstrating that the Taylor series approximation provides a good estimate for $(\nabla\widehat{w})y$. However, additional adjustment is required to ensure that all negative weights are turned positive. For the Taylor series approximation, our strategy is to replace each $y_i = 1$ with $y_i = 0$ one by one and compute $\{(\nabla\widehat{w})y\}$ n_1 times, where n_1 represents the number of units for which $y_i = 1$. For example, let $y = \{1,1,0,\ldots,1\}$. Then, $\widehat{w} + (\nabla\widehat{w})y$ can be written as:

$$\widehat{w}(1,1,0,\ldots,1) + \frac{dw}{dy_1}(1,1,0,\ldots,1) + \frac{dw}{dy_2}(1,1,0,\ldots,1) + 0 + \ldots + \frac{dw}{dy_n}(1,1,0,\ldots,1) \qquad (\star)$$

Now, substituting $y_i = 1$ to $y_i = 0$ for each unit and compute $\widehat{w} + (\nabla\widehat{w})y$ one by one gives:

$$\begin{aligned}
&\widehat{w}(0,1,0,\ldots,1) + 0 + \frac{dw}{dy_2}(0,1,0,\ldots,1) + 0 + \ldots + \frac{dw}{dy_n}(0,1,0,\ldots,1) && (1)\\
&\widehat{w}(1,0,0,\ldots,1) + \frac{dw}{dy_1}(1,0,0,\ldots,1) + 0 + 0 + \ldots + \frac{dw}{dy_n}(1,0,0,\ldots,1) && (2)\\
&\qquad\qquad\vdots && \vdots\\
&\widehat{w}(1,1,0,\ldots,0) + \frac{dw}{dy_1}(1,1,0,\ldots,0) + \frac{dw}{dy_2}(1,1,0,\ldots,0) + 0 + \ldots + 0 && (n_1)
\end{aligned}$$

Now, the Taylor series approximation gives:

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

which means that:

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

Finally, we subtract $(\nabla\hat{w})y$ from $(\star)$ for only those units whose $y_i = 1$ gives $\hat{w}$. These weights are then renormalized to sum to n, ensuring that $\sum_{i=1}^{n} \hat{w}_i = n$.

Table 3. Difference between $(1,1,0,\ldots,1)$ and $(0,1,0,\ldots,1)$ in $(\nabla\hat{w})y + \hat{w}$

Min	1st Quantile	Median	Mean	3rd Quantile	Max
-0.05	-0.001	0.002	0.00	0.004	0.02

Table 4. Difference between $(1,1,0,\ldots,1)$ and $(0,1,0,\ldots,1)$ in $\hat{w}$

Min	1st Quantile	Median	Mean	3rd Quantile	Max
-1.02	-0.003	0.001	0.00	0.004	0.03

Tables 3 and 4 present the results of the simulation study under a specific case, demonstrating that the weights remain largely unchanged after perturbing a single element of y, indicating that the remainder term in the Taylor series is $o(1)$ when $y_i = 1 \to 0$.

Table 5. Summary of logistic equivalent weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
0.0005	0.0005	0.20	1.00	1.29	5.67

Given that the logistic equivalent weights contain some extreme values, we make a trivial adjustment by trimming the weights that exceed the 95th percentile and replace them with the 95th percentile value.

2.3 Negative Weights in Model-based Methods

Negative weights are not uncommon when constructing weights using model-based methods. This occurs because real-world surveys may fail to accurately represent the target population due to oversampling or undersampling of specific demographic groups, or due to limited sample sizes. This is expected, as one reason for constructing weights is to address the lack of representativeness in the sample. Mostly linear equivalent unit weights are positive when random samples are directly drawn from the target population, ensuring representativeness. However, since the samples in our simulation study were designed to mimic the FFCWS data, we needed to adjust 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, as this adjustment has minimal impact on the total weight count across demographic cells when renormalized. Ideally, negative weights would be adjusted within the same cells to preserve total counts across cells. This approach is preferable, as linear equivalent weights do not require further raking when all weights are positive. However, in practice, this is rarely feasible, as cells with negative weights often do not correspond to those with positive weights. Consequently, adjusting for negative weights results in small discrepancies in total counts relative to population benchmarks.

2.4 Missingness and Nonresponse

Notably, missing data is rarely an issue for demographic and geographical information in survey data. When missingness is minimal, imputation should be performed as a preliminary step before constructing base weights for both design-based and model-based methods. A key challenge arises when certain weighting variables are unavailable in auxiliary data. For example, city-level birth counts are not available in the CDC data for birth occurrences. However, population birth counts are available at the county level, and the CDC natality file includes variables indicating both the birth occurrence location and the mother's residency. This information can be used to link individual CDC birth records to specific geographical areas. To estimate total births at the city level, we leveraged available sample data alongside information from the CDC natality file. 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 total population counts by utilizing the population covariance matrix when estimating weights. For instance, the linear weighting method, when renormalized and free of negative weights, preserves the exact population count without requiring a separate adjustment, as the population covariate matrix ensures that the estimated weights sum to the correct total counts across cells. Though the logistic method also accounts for total population counts, it requires a separate adjustment to ensure they match the correct total.

3. Results

3.1 Simulation Study

We begin by comparing weight estimation methods based on their accuracy in estimating population counts. Model-based weights are renormalized to match the total population size. Table 6 and Table 7 present the population counts by ethnicity and education before raking. The results indicate that model-based linear weighting produces estimates closest to population counts, whereas both design-based and logistic weighting show notable deviations across certain ethnicity and education categories.

Table 6: Population counts by ethnicity before raking

Ethnicity	**White**	**Black**	**Hispanic**	**Other**	Total
Population	382,603	226,036	312,695	78,666	1,000,000
Design-based	366,720	333,986	285,190	14,104	1,000,000
Linear model	354,826	275,590	296,001	72,582	1,000,000
Logistic model	253,570	297,733	285,548	163,149	1,000,000

Table 7: 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	8,640	180,026	513,811	170,407	127,116	1,000,000
Linear model	89,851	191,678	322,830	188,271	207,370	1,000,000
Logistic model	83,455	195,490	312,833	202,808	205,414	1,000,000

To further assess accuracy, we compare the methods based on their ability to estimate population means. Table 8 and Table 9 display the population means of a binary outcome before and after raking. The linear model consistently yields estimates closer to the true population rates compared to the design-based and logistic methods. Notably, raking reduces discrepancies across all methods, but linear weighting preserves population structure more effectively.

Table 8: Population means of a binary outcome before raking

Education	**<8 grade**	**Some HS**	**HS**	**Some College**	**College+**
Population rate	0.29	0.28	0.32	0.25	0.30
Design-based	0.16	0.23	0.31	0.23	0.25
Linear model	0.26	0.28	0.34	0.21	0.26
Logistic model	0.38	0.28	0.31	0.20	0.27

Table 9: Population means of a binary outcome after raking

Education	**<8 grade**	**Some HS**	**HS**	**Some College**	**College+**
Population rate	0.29	0.28	0.32	0.25	0.30
Design-raked	0.30	0.28	0.33	0.19	0.26
Linear model	0.27	0.28	0.33	0.20	0.26
Logistic model	0.34	0.28	0.28	0.22	0.25

We also examine the pooled mean squared error (MSE) of regression coefficients across one hundred resampled datasets, as shown in Figure 1. The linear weighting method achieves the lowest MSE, supporting its stability in regression modeling. Additionally, Figure 2 presents the effective sample size (ESS) of the weights across one hundred resampled datasets, demonstrating that linear weighting retains a higher ESS post-raking compared to design-based and logistic weighting methods. ESS is calculated as $(\sum_{i=1}^{n} w_i)^2 / \sum_{i=1}^{n} w_i^2$.

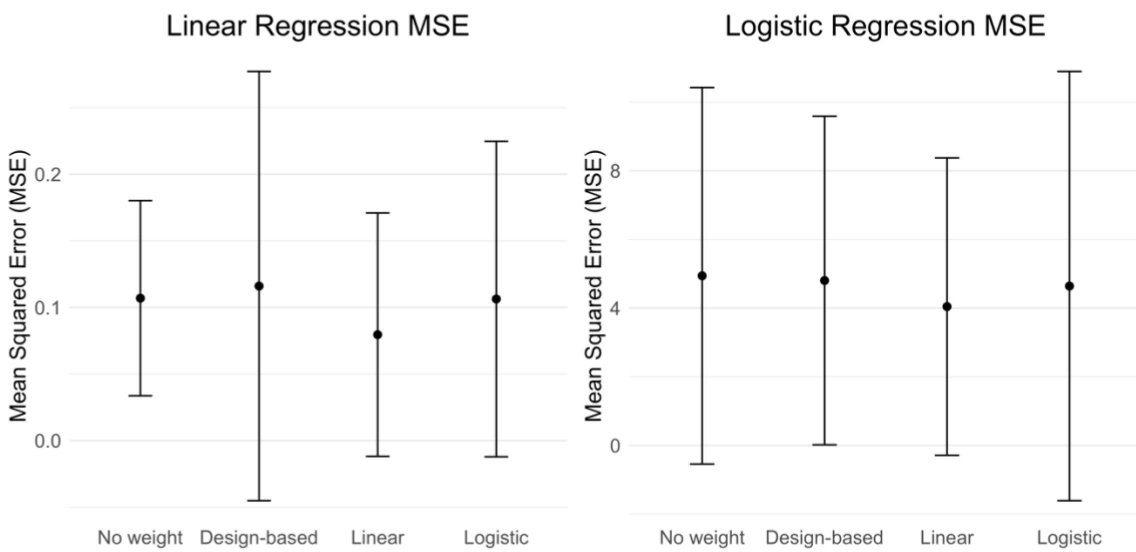

Figure 1. Mean and one standard error of mean squared error (MSE) after raking

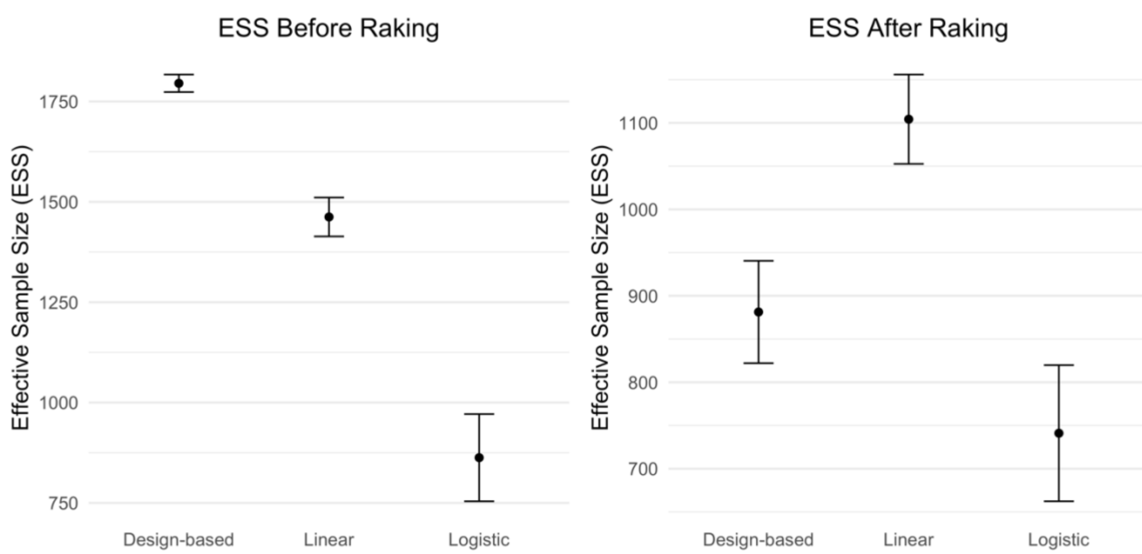

Figure 2. Mean and one standard error of effective sample size (ESS) for survey weights (n=3000)

3.2 Applications to FFCWS Study

We now apply our methods to the FFCWS study and present the results. 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 10), we find that the linear model closely approximates population counts, whereas the logistic model exhibits notable deviations. Effective sample size (ESS) comparisons in Table 11 further demonstrate that linear weighting preserves a higher ESS, while both design-based and logistic weighting results in greater reductions post-raking.

Table 10: 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	253,466	1,131,308
Linear model	73,232	267,082	367,476	242,049	181,469	1,131,308
Logistic model	171,522	382,311	303,227	202,286	71,962	1,131,308

Table 11: Effective Sample Size (ESS) for FFCWS Weights (n=3442)

Design-based	**Linear**	**Logit**	**Linear-raked**	**Logit-raked**
527	1273	1092	1082	392

To assess the accuracy of estimated population means, we evaluate Child Protective Services (CPS) contact prevalence across cities (Figure 3). Considering 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 (recent paper). The design-based and linear weighting methods produce similar adjustments but with noticeable differences in cities like 9, 14, and 15. The logistic weighting method also shows some variability compared to the linear approach—while it lowers estimates in some cities, it pulls them higher in others, particularly in cities like 10, 15, and 16. These results suggest that the choice of weighting method can influence the estimated population means.

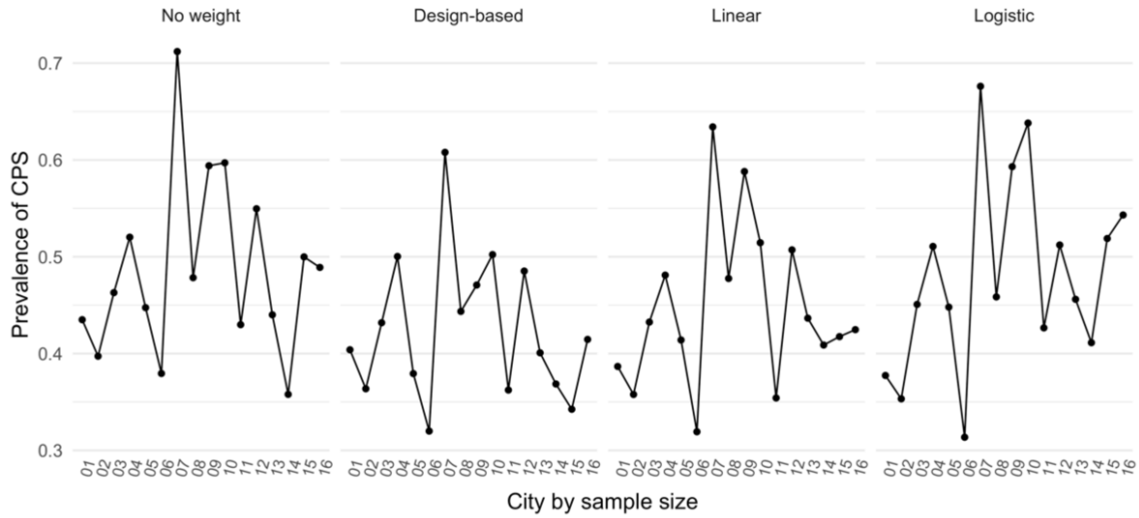

Figure 3. Prevalence of Child Protective Services (CPS) contact by city before raking

4. Discussion

4.1 Key Findings and Limitations

Our study provides a detailed comparison of model-based and design-based weighting methods, highlighting their respective advantages and limitations in terms of accuracy, efficiency, and effective sample size (ESS) retention. Design-based weighting, while widely used, often involves complex multistage steps that rely on auxiliary data to estimate selection probabilities. Model-based methods, on the other hand, do not require auxiliary data for probability estimation but use them as prior information to guide the modeling strategy. This makes model-based weighting more flexible and computationally simpler, especially when auxiliary information is incomplete or inconsistent.

Linear weighting, among the model-based approaches, consistently demonstrates superior performance, achieving the lowest mean squared error (MSE) in both regression analyses. Linear equivalent unit weights are derived from a closed-form solution, ensuring reliable estimates and efficiency. In contrast, logistic weighting, while capturing nonlinear relationships, lacks translation invariance and introduces higher variability. Notably, both model-based weightings outperform design-based weighting in terms of MSE.

However, logistic weighting exhibits challenges, especially in binary outcome contexts. These weights, derived from nonlinear regression, often require additional adjustments, such as Taylor series approximations and weight raking, which can be computationally intensive and may require parallel computing for resampling purposes. Our simulation study demonstrated that the prevalence of the binary outcome affects weight estimation. A higher prevalence resulted in less extreme weights and a larger effective sample size (ESS). For instance, when the binary outcome prevalence was around 0.3 across the resampled datasets, we flipped the 1 and 0 values to increase the prevalence to 0.7. This adjustment improved weight estimation, aligning with our goal of optimizing the weight calculation.

Raking often resulted in substantial reductions in effective sample size (ESS), particularly with design-based and logistic weighting, whereas linear weighting retained more of its efficiency post-raking. Additionally, both design-based and logistic methods can distort the final weighted sample even after raking, while linear weighting preserves population counts, leading to more stable post-raking distributions. Despite the advantages of model-based linear weighting, logistic weighting can still be useful in settings requiring explicit probability modeling. Yet, its higher variability and need for adjustments highlight potential limitations in its practical use.

4.2 Future Research

Future research could explore ways to improve the integration of population priors in both Bayesian logistic regression and non-parametric models like Gaussian Processes (GPs). In

Bayesian logistic regression, incorporating population margins as priors while maintaining computational efficiency presents a challenge. One promising direction is the development of hierarchical Bayesian models that integrate these population priors directly, enhancing both the robustness of weight estimates and the representativeness of the weighted sample. However, the complexity of hierarchical structures with numerous demographic and geographical categories may introduce computational and convergence issues.

On the other hand, non-parametric models such as GPs offer flexibility in capturing complex relationships between covariates and weights. However, their ability to incorporate population priors without losing predictive accuracy or becoming overly parameterized remains an open question. Researchers could explore methods for adapting GPs by introducing priors reflecting population margins, potentially leveraging kernel methods or structured priors that connect observed data to population totals.

Furthermore, future studies could examine the integration of resampling techniques like bootstrapping to assess uncertainty in weight estimates while aligning them with population characteristics. This approach could refine the models and improve weight estimation in practice, providing more robust methods for addressing variability and bias in survey weighting.