Optimal probability weights for inference with constrained precision

Michele Santacatterina*
Unit of Biostatistics, Institute of Environmental Medicine
Karolinska Institutet, Stockholm, Sweden
and
Matteo Bottai
Unit of Biostatistics, Institute of Environmental Medicine
Karolinska Institutet, Stockholm, Sweden

April 8, 2018

Abstract

Probability weights are used in many areas of research including complex survey designs, missing data analysis, and adjustment for confounding factors. They are useful analytic tools but can lead to statistical inefficiencies when they contain outlying values. This issue is frequently tackled by replacing large weights with smaller ones or by normalizing them through smoothing functions. While these approaches are practical, they are also prone to yield biased inferences. This paper introduces a method for obtaining optimal weights, defined as those with smallest Euclidean distance from target weights among all sets of weights that satisfy a constraint on the variance of the resulting weighted estimator. The optimal weights yield minimum-bias estimators among all estimators with specified precision. The method is based on solving

*The authors gratefully acknowledge the support provided by the KID grant from Karolinska Institutet, Sweden. The authors also acknowledge very helpful comments from the anonymous Reviewer and the Associate Editor.

a constrained nonlinear optimization problem whose Lagrange multipliers and objective function can help assess the trade-off between bias and precision of the resulting weighted estimator. The finite-sample performance of the optimally weighted estimator is assessed in a simulation study, and its applicability is illustrated through an analysis of heterogeneity over age of the effect of the timing of treatment-initiation on long-term treatment efficacy in patient infected by human immunodeficiency virus in Sweden.

Keywords: probability weights, weighted estimators, sampling weights, nonlinear constrained optimization, mathematical programming

1 Introduction

Probability weights have long been used in a variety of applications in many areas of research, from medical and social sciences to physics and engineering. For example, they are used when dealing with missing data, balancing distribution of confounders between populations being compared, or correcting for selection probability in complex survey designs. The increasing popularity of probability weights over several decades originates from their conceptual simplicity, modeling flexibility, and sound theoretical basis.

Our work was motivated by the abundant use of probability weights in studies on the effect of timing of initiation and switching of treatment for the human immunodeficiency virus (HIV-CAUSAL Collaboration et al., 2011; When To Start Consortium et al., 2009; Kitahata et al., 2009; Petersen et al., 2008). For example, the Writing Committee for the CASCADE Collaboration (2011) evaluated the relative benefits of early treatment initiation over deferral in patients with CD4-cell count less than 800 cells/μL. They analyzed time to death or first acquired-immunodeficiency-syndrome diagnosis with weighted survival curves and Cox regressions. The weights were obtained as the inverse of the probability of treatment initiation given a set of baseline covariates.

Probability weights are non-negative scalar values associated with each experimental unit that can be used by appropriate statistical methods. They may be known or estimated from observed data. When their distribution presents with long tails, the resulting inference may be highly imprecise (Rao, 1966; Kang and Schafer, 2007; Basu, 2011).

Methods have been proposed to alleviate the sometimes excessive imprecision of weighted inference, and the body of work on this topic is vast. In medical sciences the most frequent approach is weight trimming, or truncation, which consists of replacing outlying weights with less extreme ones. For example, values in the the top and bottom deciles may be replaced with the 90th and 10th centiles, respectively. Trimming reduces the variability of the weights and the standard error of the corresponding weighted estimator. Potter (1990) discussed techniques to choose the best trimming cutoff value, by assuming that the weights follow an inverted and scaled beta distribution. Cox and McGrath (1981) suggested finding the cutoff value that minimizes the mean squared error of the trimmed estimator evaluated empirically at different trimming level. Kokic and Bell (1994) provided an optimal cutoff for stratified finite-population estimator that minimized the total mean squared error of the trimmed estimator. Others suggested similar methods to obtain optimal cutoff points (among others Rivest et al. (1995); Hulliger (1995)).

Approaches other than trimming have also been considered. Pfeffermann and Sverchkov (1999) suggested modifying the weights by using a function of the covariates that minimized a prediction criteria. They later extended this approach to generalized linear models (Pfeffermann and Sverchkov, 2003). In the context of design-based inference, Beaumont (2008) proposed modeling the weights to obtain a set of smoothed weights that can lead to an improvement in statistical efficiency of weighted estimators. Fuller (2009) (Section 6.3.2) discussed a class of modified weights in which efficiency can be maximized. Kim and Skinner (2013) merged the ideas earlier proposed by Beaumont (2008) and Fuller (2009) and considered modified weights that were a function of both covariates and outcome variable of interest. Elliot and Little (2000) and Elliott (2008) provided a weight-pooling model averaging the estimates obtained from all different trimming points within the Bayesian framework. Elliott (2009) extended these results to generalized linear regression models. Beaumont et al. (2013) used the conditional biased to down-weigh the most influential units to obtain robust estimators. Other approaches, primarily based on likelihood, have been proposed. These provide efficient inference under informative sampling (Chambers, 2003; Pfeffermann, 2009, 2011; Scott and Wild, 2011, among others).

While weight trimming and smoothing reduce the variability in the weights and inferential imprecision, they can also introduce substantial bias. This paper describes a new method to obtain optimal weights, while controlling the precision of weighted estimators. The method is based on solving a constrained nonlinear optimization problem to find an optimal set of weights that minimizes a distance from target weights, while satisfying a constraint on the precision of the resulting weighted estimator. A similar approach was recently proposed by Zubizarreta (2015), who suggested obtaining stable weights by solving a constrained optimization problem to minimize the variance of the weights under the constraint that the mean value of the covariates remains within a given tolerance.

The following Section describes the constrained optimization problem that defines optimal weights and presents some of their properties. Section 3 discusses the choice of the variance constraint. Section 4 shows the results of a simulation study that contrasts optimal weights and trimmed weights with respect to mean squared error of the weighted marginal mean of a continuous variable and the parameters of a weighted regression. Section 5 illustrates the use of optimal weights to evaluate heterogeneity in the effect of timing of treatment initiation on long-term CD4-cell count. The data were extracted from a comprehensive register of patients infected by the human immunodeficiency virus in Sweden (Sönnerborg, 2016). Section 6 contains conclusions and some suggestions for the use of optimal weights in applied research.

2 Optimal probability weights

Let $\hat{\theta}_{w^*}$ be an unbiased estimator for a population parameter θ^* that uses weights $w^* = (w_1^*, \ldots, w_n^*)^T$, with $\mathbf{1}^T w^* = 1$ and $w^* \geq 0$. Throughout, the symbol $\mathbf{1}$ indicates an n-dimensional vector of ones. For example, $\hat{\theta}_{w^*} = y^T w^*$ is the weighted mean of a sample of

n observations $y = (y_1, \ldots, y_n)^T$. Let σ_{w^*} indicate the standard error of $\hat{\theta}_{w^*}$ and $\hat{\sigma}_{w^*}$ an estimator for it.

When w^* contains outliers, the standard error σ_{w^*} may be large and inference on θ^* inefficient. Instead of trimming the weights, we suggest deriving the weights $\hat{w}$ that are closest to w^* with respect to the Euclidean norm $\|w - w^*\|$, under the constraint that the estimated standard error $\hat{\sigma}_{\hat{w}}$ be less than or equal to a specified constant $\xi > 0$. The corresponding nonlinear constrained optimization problem can be written as follows,

$$\begin{aligned} \underset{w \in \mathbb{R}^n}{\text{minimize}} \quad & \|w - w^*\| && (1) \\ \text{subject to} \quad & \hat{\sigma}_w \leq \xi && (2) \\ & w \leq \epsilon && (3) \\ & w \geq 0 && (4) \end{aligned}$$

When a solution $\hat{w}$ to problem (1)-(4) exists, constraint (2) guarantees that the estimated standard error of the estimator with weights $\hat{w}$ is less than or equal to ξ. Constraints (3) and (4) guarantee that the optimal weights $\hat{w}$ are bounded and non-negative, respectively. The constant ϵ, with $0 < \epsilon \leq 1$, can be set close to 1 to improve the goodness of the asymptotic approximation of the the variance estimator $\hat{\sigma}_w$, as further discussed in the following Sections.

Throughout this paper we refer to $\hat{w}$ as the set of *optimal* weights and to w^* as the set of *target* weights. The following are some notable features of the optimal weights.

(i) *Consistency.* If the estimator of the standard error of the weighted estimator converges in probability to zero as the sample size tends to infinity for any set of weight w, then the probability that $\hat{\theta}_{\hat{w}} = \hat{\theta}_{w^*}$ converges to one,

$$\lim_{n \to \infty} P(\hat{\sigma}_w \leq \xi) = 1 \;\Rightarrow \lim_{n \to \infty} P(\hat{\theta}_{\hat{w}} = \hat{\theta}_{w^*}) = 1 \tag{5}$$

for any constant value $\xi > 0$. Property (5) holds because the target weights w^* are assumed to satisfy constraints (3) and (4). If they also satisfy constraint (2), then w^* is the optimum and $\hat{\theta}_{\hat{w}} = \hat{\theta}_{w^*}$.

(ii) *Minimum-bias estimator.* If the optimal weights are equal to the target weights, then the corresponding weighted estimators are equal to each other and unbiased for the target parameter θ^*,

$$\hat{w} = w^* \;\Rightarrow\; \hat{\theta}_{\hat{w}} = \hat{\theta}_{w^*} \;\Rightarrow\; E(\hat{\theta}_{\hat{w}}) = E(\hat{\theta}_{w^*}) = \theta^*$$

When the optimal weights are different from the target weights, then the optimally weighted estimator may or may not be biased. For example, suppose $\hat{\theta}_{w^*} = y^T w^*$ is an unbiased estimator for θ^*. The bias of the optimally weighted estimator $\hat{\theta}_{\hat{w}} = y^T \hat{w}$ with respect to the target parameter θ^* is

$$E\left[y^T\hat{w} - \theta^*\right] = E\left[y^T\hat{w} - y^T w^*\right] + \underbrace{E\left[y^T w^*\right] - \theta^*}_{=0} = E\left[y^T\left(\hat{w} - w^*\right)\right]$$

If the vectors y and $(\hat{w} - w^*)$ are orthogonal, then the optimally weighted estimator $\hat{\theta}_{\hat{w}}$ is unbiased for θ^*. Also, minimizing $\|\hat{w} - w^*\|$ is equivalent to minimizing the bias of the optimal estimator with respect to the target parameter.

More generally, suppose the target estimator $\hat{\theta}_{w^*}$ is the solution to a weighted equation for a given set of weights w^*

$$\sum_{i=1}^{n} w_i^* h_i(\hat{\theta}_{w^*}) = 0$$

where h_i is a known function of the parameter θ and sample data. A Taylor series expansion of $h_i(\hat{\theta}_{\hat{w}})$ around the parameter value $\hat{\theta}_{w^*}$ shows that the optimally weighted estimator is the solution to

$$\sum_{i=1}^{n} \hat{w}_i \left[h_i(\hat{\theta}_{w^*}) + h_i'(\hat{\theta}_{w^*})(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*}) + O((\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*})^2)\right] = 0$$

The remainder, O, converges to zero at a quadratic rate as $(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*})$ tends to zero. From the above equation, given that $E(\hat{\theta}_{w^*}) = \theta^*$ and ignoring the Taylor series remainder, the bias of the optimally weighted estimator with respect to the target parameter is approximately equal to

$$E(\hat{\theta}_{\hat{w}} - \theta^*) = E(\hat{\theta}_{\hat{w}} - \hat{\theta}_{w^*}) + E(\hat{\theta}_{w^*}) - \theta^* \approx -E\left[\frac{(\hat{w} - w^*)^T h(\hat{\theta}_{w^*})}{\hat{w}^T \nabla_w h(\hat{\theta}_{w^*})}\right]$$

where $h(\hat{\theta}_{w^*})$ denotes the stacked vector $(h_1(\hat{\theta}_{w^*}), \ldots, h_n(\hat{\theta}_{w^*}))^T$ and $\nabla_w h(\hat{\theta}_{w^*})$ its gradient. Similarly to the weighted mean estimator, if the vectors $h(\hat{\theta}_{w^*})$ and $(\hat{w} - w^*)$ are orthogonal, then the optimally weighted estimator $\hat{\theta}_{\hat{w}}$ is approximately unbiased for θ^*, bar the Taylor series remainder. Also, given property (5), minimizing $\|\hat{w} - w^*\|$ is equivalent to minimizing the bias of the optimal estimator with respect to the target parameter. The optimal weights $\hat{w}$ therefore yield the minimum-bias estimator among all weighted estimators with standard error less or equal to ξ.

(iii) *Convex optimization problem.* The objective function in (1) is convex, and constraints (3) and (4) are linear. In general, if the constraint in (2) is also convex, then the optimization problem (1)-(4) admits one unique solution. Computational algorithms to solve nonlinear constrained optimization problems exist. In our simulation and data analysis we used the primal-dual interior point algorithm implemented in the R package “Ipoptr” (Wächter and Biegler, 2005), which can solve general large-scale nonlinear constrained optimization problems. The “MA57” sparse symmetric system (HSL, 2016) was used as a line-search method within “Ipoptr”.

(iv) *Multiple constraints.* The optimization problem (1)-(4) can be extended to include multiple equality and inequality constraints. For example, when the weighted estimator of interest is a vector, each element of the estimated variance matrix can be constrained separately. This is further discussed in the simulation study in and the real-data application presented in Sections 4 and 5, respectively.

3 The precision constraint

The optimal probability weights $\hat{w}$, solution to the optimization problem (1)-(4), depend on the value ξ specified by constraint (2). The value ξ directly sets the standard error, $\hat{\sigma}_{\hat{w}}$, and the precision, $1/\hat{\sigma}_{\hat{w}}$, of the estimate $\hat{\theta}_{\hat{w}}$. Smaller values of ξ induce greater precision and larger values of the objective function (1). The latter generally imply larger bias of the estimator $\hat{\theta}_{\hat{w}}$ with respect to the target parameter θ_{w^*}. As shown in Section 4, substantial gains in precision can often be traded at slight bias.

An example may help to interpret the trade-off between precision and bias. Imagine that the target weights aim to balance the distributions of covariates between two treatment groups in an observational study, thus mimicking the conditions of a randomized experiment. The target weights contain outliers, and the resulting estimate of the treatment effect is excessively imprecise. Suppose a specified level ξ for the standard error of the treatment effect is considered acceptable. The optimally-weighted estimate $\hat{\theta}_{\hat{w}}$ would have the specified precision, but it would be biased for θ_{w^*}.

In practice, what precision level may be considered acceptable is for the analyst to determine. Sometimes, the desired precision of the estimate of interest is known or can be bounded within a reasonably small range. When it cannot be determined to any degree of accuracy, it is recommendable that different values be explored within reason.

Evaluating the magnitude of the Lagrange multipliers may be useful when contrasting precision and bias. Suppose that the vector of target weights w^* satisfies constraints (3) and (4). If w^* also satisfies constraint (2), then w^* is the unique optimum and the objective function (1) at the optimum is zero. If w^* does not satisfy constraint (2), then $\hat{w} \neq w^*$, $\hat{\sigma}_{\hat{w}} = \xi$, and $\nabla_w f(\hat{w}) = -\lambda \nabla_w g(\hat{w})$, where $\nabla_w f$ and $\nabla_w g$ are the gradients of the objective function (1) and constraint (2), respectively, and the scalar constant λ is the Lagrange multiplier. A small multiplier at the optimal solution $\hat{w}$ indicates that a decrease in ξ would cause a small increase in objective function (1) and in the bias with respect to the target parameter θ_{w^*}. Conversely, a large multiplier indicates that a decrease in ξ would cause a large increase in the objective function and bias. This point is further discussed in the real-data application in Section 5.

Determining an acceptable precision is similar to determining the detectable difference in sample size and power calculations. As in sample size and power calculations, the desired precision level may be defined before a study is initiated, and relative precision levels, such as effect sizes, may be more easily determined than absolute levels.

The trade-off between bias and precision is not peculiar of the optimal-weight method proposed in Section 2 above. For example, selecting the cutoff value in the traditional trimming approach also amounts to deciding precision and bias of the resulting weighted estimator, albeit less efficiently then the optimal-weights estimator.

4 Simulations

We examined the performance of the proposed optimal weights in different simulated scenarios. A first set of scenarios considered weighted means (Section 4.1), and a second set weighted least-squares estimators (Section 4.2). In all simulated samples, we compared the proposed optimal weights with trimmed weights with respect to bias, variance, and mean squared error of the weighted estimator (2).

4.1 Weighted mean

This Section describes setup and results of the simulation for the weighted mean estimator.

4.1.1 Simulation's setup

In each scenario we pseudo-randomly generated 1,000 samples each of which comprised 500 observations from a normally-distributed variable under the following model

$$y_i \sim N(20 + 8x_i, 4) \tag{6}$$

where $i = 1, \ldots, 500$, and $x_i \sim beta(x_i \mid \alpha_0, \beta_0)$, a beta distribution with parameters α_0 and β_0. The target weights were defined as

$$w_i^* = \frac{beta(x_i \mid \alpha_1, \beta_1)}{beta(x_i \mid \alpha_0, \beta_0)} \tag{7}$$

The weights defined in equation (7) can be interpreted as sampling weights (Quatember, 2015). The density $beta(x_i \mid \alpha_0, \beta_0)$ can be seen as the distribution of the variable x in the sampled population, while $beta(x_i \mid \alpha_1, \beta_1)$ the distribution in the population that represent the inferential target.

To evaluate the performance of the proposed optimal weights, we considered fifty different scenarios, constructed by combining the following parameter values: $\alpha_1 = \{1, 2, 3, 4, 5\}$, $\beta_1 = \{1, 2, 3, 4, 5\}$, and $(\alpha_0, \beta_0) = \{(2, 5), (5, 5)\}$. When $(\alpha_0, \beta_0) = (\alpha_1, \beta_1)$, the weights w_i^* were all equal to one. The farther away the two sets of parameters were from each other, the more extreme the weights became and the less efficient the resulting weighted inference was.

In each simulated sample we considered two estimators for the weighted mean: the optimal estimator $\hat{\theta}_{\hat{w}} = y^T \hat{w}$ and the trimmed estimator $\hat{\theta}_{\overline{w}} = y^T \overline{w}$. The cutoff value for calculating the trimmed weights, $\overline{w}$, was computed by using the method proposed by Cox and McGrath (1981) and Potter (1990). We considered a set of truncation thresholds on a grid of values. For each truncation threshold, we estimated the mean squared error as if the true value of the target parameter was known, not estimated from the data. This cannot be done in real-data settings, but it allowed us to compare our proposed approach with truncation at its best-possible performance. The estimation was performed by generating 100 Monte Carlo samples. In practical applications the true target parameter is unknown, and the optimal truncation threshold needs to be estimated. Recently, Borgoni et al. (2012) suggested estimating it efficiently through the bootstrap.

The optimal weights $\hat{w}$ were obtained by solving the optimization problem (1)-(4). The constant ξ was set equal to the estimated standard error of the weighted sample mean using the trimmed weights. In each simulated sample we computed the estimated standard error of the weighted mean as described in Cochran (1977). The value of ϵ was set to be equal to the 0.999-quantile of the distribution of w^*.

In a secondary analysis, we evaluated the performance of the proposed weights at different values of ξ in (2). We considered the following two scenarios

$$w_i^* = \frac{beta(x_i \mid \alpha_1 = 3, \beta_1 = 3)}{beta(x_i \mid \alpha_0 = 2, \beta_0 = 5)}, \; w_i^* = \frac{beta(x_i \mid \alpha_1 = 4, \beta_1 = 2)}{beta(x_i \mid \alpha_0 = 5, \beta_0 = 5)}. \tag{8}$$

The target population was skewed in the first scenario and symmetric in the second. We set ξ equal to values on a grid from 50% to 100% of the variance observed when using the target weights w_i^* defined in (8). We set ϵ equal to values from the 0.99-quantile of the distribution of w^* when ξ was equal to 50% of the observed weighted variance, and to the maximum of w^* when ξ was equal to 100% of the observed weighted variance.

4.1.2 Simulation's results

The left-top panel in Figure 1 shows the scenarios where the target population is skewed with $\alpha_0 = 2$ and $\beta_0 = 5$. The curve indicates the ratio between the observed mean squared error of the weighted mean estimator with trimmed weights and that with optimal weights with respect to the target paramter θ^*,

$$\text{MSE ratio} = \frac{\sum_{i=1}^{1000} (\hat{\theta}_{\overline{w}}^{(i)} - \theta^*)^2}{\sum_{i=1}^{1000} (\hat{\theta}_{\hat{w}}^{(i)} - \theta^*)^2} \tag{9}$$

where $\hat{\theta}_{\overline{w}}^{(i)}$ and $\hat{\theta}_{\hat{w}}^{(i)}$ denote the two estimated parameters in the i-th simulated dataset. The curve was smoothed with a B-spline with 4 degrees of freedom. The bottom-left panel is identical to the top-left one, except the target population is symmetric with $\alpha_0 = 5$ and $\beta_0 = 5$.

The optimally weighted estimator $\hat{\theta}_{\hat{w}}$ performed better in all simulated scenarios. The larger gain in mean squared error was observed when the target weights w^* were larger. For example, with target weights $w_i^* = beta(x_i \mid \alpha_1 = 3, \beta_1 = 3)/beta(x_i \mid \alpha_0 = 2, \beta_0 = 5)$, indicated by the dot in top-left panel, the mean squared error of the trimmed estimator was about 1.4 times as large as that one of the optimal estimator. A ratio greater than 1.5 was observed when the target weights were $w_i^* = beta(x_i \mid \alpha_1 = 4, \beta_1 = 2)/beta(x_i \mid \alpha_0 = 5, \beta_0 = 5)$, as indicated by the dot in bottom-left panel. As expected, when the target weights were small, the difference between trimmed and optimal estimators was small, too.

The lines in the right-hand-side panels in Figure 1 depict mean squared error (solid line), variance (dotted), and bias (dashed) for different percentages of the variance of the weighted estimator observed when using w_i^* as defined in (8). The lines show that high precision could be obtained with relatively low bias.

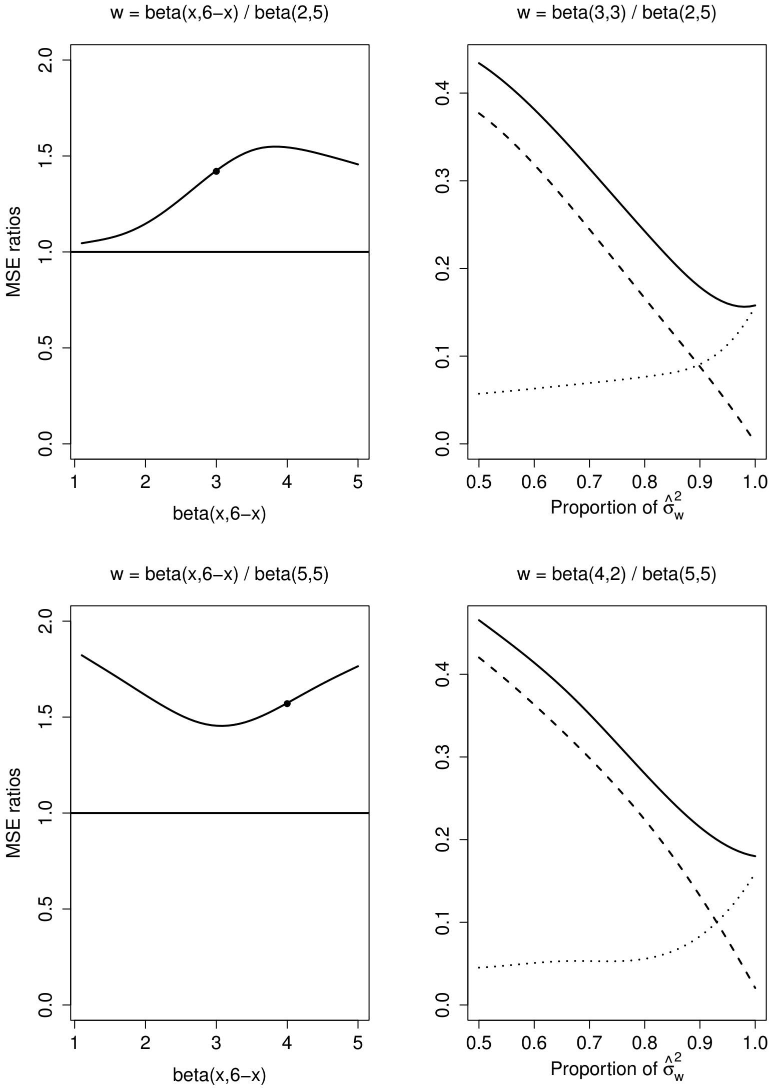

Figure 1: Left-hand-side panels: mean squared error ratio between trimmed and optimally weighted estimators. Right-hand-side panels: mean squared error (solid line), variance (dotted), and bias (dashed) of the optimally weighted estimator $\hat{\theta}_{\hat{w}}$, for different proportions of the variance of the weighted estimator observed when using w_i^* defined in (8).

4.2 Weighted least-square estimator

This Section describes setup and results of the simulation for the weighted least-squares estimator.

4.2.1 Simulation's setup

We pseudo-randomly generated 1,000 samples each of which comprised 500 observations on three variables (y_i, t_i, c_i) with the following model

$$y_i = -10 + \theta t_i + \gamma c_i + \varepsilon_i \tag{10}$$

with $\varepsilon_i \sim N(0,1)$, $c_i \sim N(10,1)$, and $t_i \sim Ber(\pi_i)$, with $\pi_i = \exp(c_i - 10)/(1 + \exp(c_i - 10))$. We considered 25 different values for the parameter γ from 0.1 to 5. The parameter θ was the inferential objective and was set at $\theta = 2$ in all scenarios.

We defined the target weights as $w^* = 1/\hat{\pi}_i$, where $\hat{\pi}_i$ was an estimator for π_i obtained from a logistic regression model with c_i as the only covariate with the "ipw" package in R (van der Wal and Geskus, 2011). We applied the target weights to the following weighted regression model

$$y_{i,w^*} = \beta_{1,w^*} + \beta_{2,w^*} t_i + \varepsilon_{i,w^*} \tag{11}$$

The setup described above reflects a common applied research settings where t_i represents a treatment, c_i a confounder, and y_i a response variable of interest. When estimating the treatment effect from observational data, inverse probability weights aim at balancing the distribution of covariates across treatment groups, thus mimicking a randomized experiment. The parameter γ in equation (10) determines the strength of the confounding effect of c_i.

In each scenario we estimated the optimally weighted estimator $\hat{\beta}_{2,\hat{w}}$ and the trimmed estimator $\hat{\beta}_{2,\overline{w}}$. The cutoff value for calculating the trimmed weights, $\overline{w}$, was computed as described in Section 4.1.1. The optimal weights were obtained by solving the following optimization problem

$$\begin{aligned} \underset{w \in \mathbb{R}^k}{\text{minimize}} \quad & \|w - w^*\| && (12) \\ \text{subject to} \quad & \hat{\sigma}_{1,w} \leq \xi_1 && (13) \\ & \hat{\sigma}_{2,w} \leq \xi_2 && (14) \\ & w \leq \epsilon && (15) \\ & w \geq 0 && (16) \end{aligned}$$

The above optimization problem has one constraint for each of the regression coefficients $\beta_{1,w}$ and $\beta_{2,w}$ in model (11). The level of precision ξ_2 for the coefficient $\beta_{2,w}$ was set equal to the estimated standard error of $\hat{\beta}_{2,\overline{w}}$, while the level of precision ξ_1 for the coefficient $\beta_{1,w}$ was set to a large number and the constraint (13) was inactive in all simulations. The sandwich estimator was used to compute $\hat{\sigma}_{1,w}$ and $\hat{\sigma}_{2,w}$ (Stefanski and Boos, 2002; Strutz, 2010, pag.109). The value of ϵ was choose to be equal to the 0.999-quantile of the distribution of w^*.

In a secondary analysis, we evaluated the performance of the proposed weights at different percentages of the variance of $\hat{\beta}_{2,w}$, when $\gamma = 4$. We set ϵ as described in Section 4.1.1.

4.2.2 Simulation's results

The left-hand-side panel in Figure 2 shows the ratio between the observed mean squared error of the trimmed weighted mean estimator $\hat{\beta}_{2,\overline{w}}$ and that of the optimally weighted estimator $\hat{\beta}_{2,\hat{w}}$ across values of γ, whose expression is analogous to (9). The lines in Figure 2 were smoothed using B-splines with 5 (left-hand-side panel) and 8 (right-hand-side panel) degrees of freedom, respectively. The optimally weighted estimator performed well in all scenarios. At high values of γ, the observed mean squared error of the trimmed estimator was more than 4 times as large as that of the optimally weighted estimator. The right-hand-side panel in Figure 2 shows mean squared error (solid line), variance (dotted), and bias (dashed) for different values of ξ_2 when $\gamma = 4$. With increasing values of ξ_2, the variance increases and the bias decreases.

5 Age at treatment initiation in HIV-infected patients

The human immunodeficiency virus (HIV) epidemic is a leading global burden with major economic and social consequences. Antiretroviral therapy is the current standard treatment for HIV-infected patients. Yet, several key questions still are unsolved, including when to initiate treatment. CD4-cell count is an indicator used to monitor the immune system, define the stage of the disease, and make clinical decisions. Once a patient is infected, the number of CD4 cells rapidly declines. Treatment is generally initiated when it falls below the threshold of 500 or sometimes 350 cells/μL. During treatment, the count rises again towards normal levels. Several observational studies have documented the prognosis for patients who started treatment at different CD4-cell count thresholds. Their findings are different and occasionally contrasting (HIV-CAUSAL Collaboration et al., 2011; When To Start Consortium et al., 2009; Kitahata et al., 2009).

5.1 Three target populations

After the introduction of antiretroviral therapy, mortality among treated patients has substantially declined, and medical and research interest has shifted from mortality to aging and long-term clinical outcomes (Wright et al., 2013). Recently, age at treatment initiation has received increasing attention as a potentially important modifier (Deeks and Phillips,

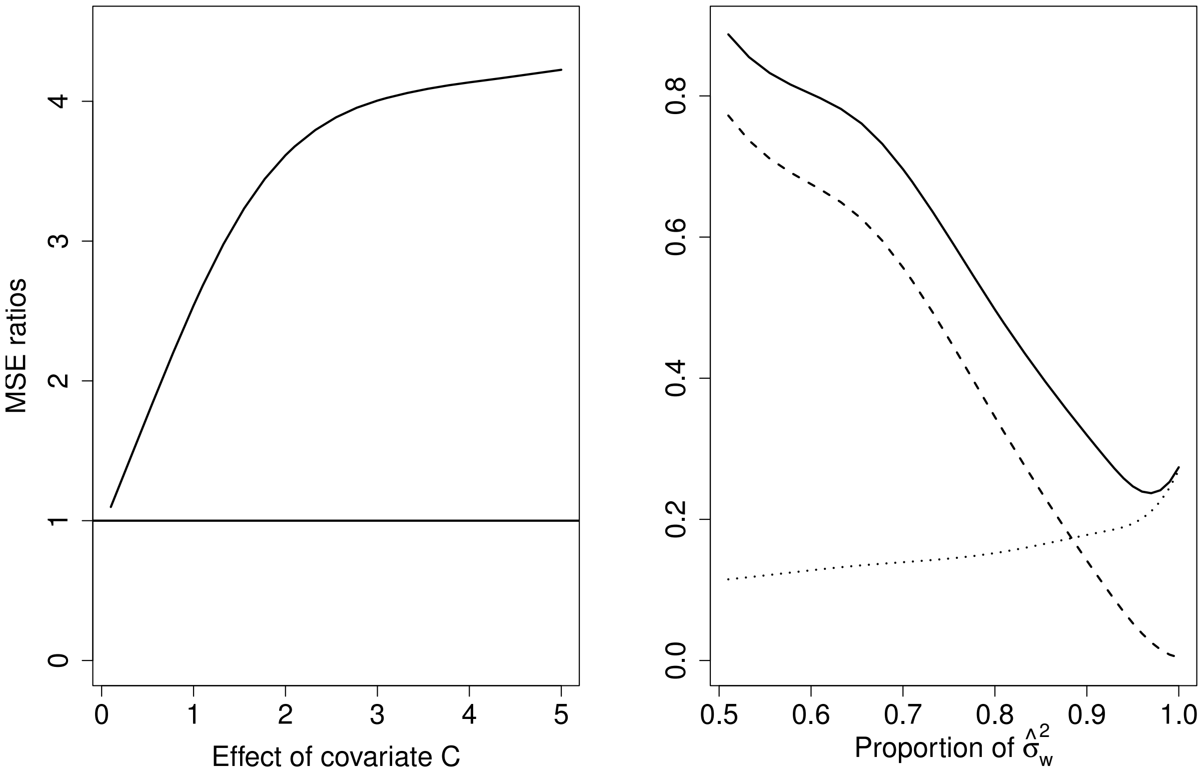

Figure 2: Left-hand-side panel: ratio between the observed mean squared error of the trimmed weighted mean estimator $\hat{\beta}_{2,\overline{w}}$ and that of the optimally weighted estimator $\hat{\beta}_{2,\hat{w}}$ across values of γ. Right-hand-side panel: mean squared error (solid line), variance (dotted), and bias (dashed) for different values of for different values of ξ_2 when $\gamma = 4$.

2009).

We investigated the association between CD4-cell count at treatment initiation and that at five years after initiation across groups of patients starting treatment at different ages. We used data on 500 subjects from the Swedish Infcare HIV database, which has collected data from all known HIV-infected patients in Sweden continuously for decades. CD4-cell count at treatment initiation was classified in two categories, 351−500 and 501+ cells/μL.

Instead of stratifying the analysis by possible age groups, we defined three target populations as normal densities centered at age 27, 36, and 44 years. These values correspond to the 25th, 50th, and 75th percentile, respectively, of the marginal distribution of age in our sample. Specifically, the target weights for the k-th target population, $k = 1, 2, 3$, were calculated as

$$w_i^* = \frac{\phi((\text{age}_i - \mu_k)/\sqrt{2})}{\hat{f}(\text{age}_i)} \tag{17}$$

where ϕ is the standard normal density function, $\mu_1 = 27$, $\mu_2 = 36$, $\mu_3 = 44$, and $\hat{f}$ is a non-parametric density estimator. For the latter we used the "density" function available in R.

5.2 Optimal weights

For each target population, we obtained optimal weights $\hat{w}$ by solving the following problem

$$
\begin{aligned}
\underset{w \in \mathbb{R}^k}{\text{minimize}} \quad & \|w - w^*\| & (18)\\
\text{subject to} \quad & \hat{\sigma}_{1,w} \leq \xi_1 & (19)\\
& \hat{\sigma}_{2,w} \leq \xi_2 & (20)\\
& w \leq \epsilon & (21)\\
& w \geq 0 & (22)
\end{aligned}
$$

The symbols $\hat{\sigma}_{i,w}, i = 1, 2$ denote the estimated standard errors of the coefficients of the following weighted linear regression model

$$E(\text{CD4}_5 \mid \text{CD4}_0) = \beta_{1,w} + \beta_{4,w} I_{\text{CD4}_0 \geq 501} \tag{23}$$

where CD4_5 is the count at five year after treatment initiation, CD4_0 is the count at treatment initiation, β's are the regression coefficients to be estimated, and I_A is the indicator function of the event A. When the target weights were applied, the standard errors of the regression coefficients ranged from 20 to 99, making inference very imprecise. We therefore constrained the standard errors at three different sets of values: (1) half the values observed for the weighted estimator with target weights, i.e., $\xi_2 = \hat{\sigma}_{2,w^*}/2$; (2) 30% standard error reduction and (3) 10% standard error reduction for the weighted estimator with target weights, i.e., $\xi_2 = \hat{\sigma}_{2,w^*}$. The standard error of the intercept $\hat{\sigma}_{1,w}$ was left unconstrained in all analyses. The constant ϵ was set as described in Section 4.1.1.

5.3 Results

Table 1 shows the Lagrange multipliers λ_1 and λ_2 associated with constraints (19), and (20), respectively, the square root of objective function $\sqrt{n\|\hat{w} - w^*\|}$, which can be interpreted as the quadratic mean difference between optimal and target weights per observation, and the estimated optimal weighted coefficients in model (23) at the optimal weights $\hat{w}$.

Age	**27**					
ξ_2	43		60		77	
λ_1	0.0		0.0		0.0	
λ_2	12.4		2.6		0.4	
$\sqrt{n}\lVert\hat{w}-w^*\rVert$	0.4		0.2		0.1	
$\hat{\beta}_{1,\hat{w}}, \quad \hat{\sigma}_{1,\hat{w}}$	547	(29)	543	(36)	540	(40)
$\hat{\beta}_{2,\hat{w}}, \quad \hat{\sigma}_{2,\hat{w}}$	101	(43)	131	(60)	151	(77)

Table 1: Lagrange multipliers λ_1 and λ_2 (multiplied by 100) associated with constraints (19), and (20), respectively, the square root of objective function multiplied by the sample size $\sqrt{n}\lVert\hat{w}-w^*\rVert$, and the estimated optimal weighted coefficients with standard errors in brackets for model (23) at the optimal weights $\hat{w}$ in the population of patients starting treatment at about 27 years of age.

In patient starting treatment at about 27 years of age, constraining the standard error $\hat{\sigma}$ to be no greater than 43, i.e. $\xi_2 = 43$, half the values observed for the target estimator, resulted in large multiplier and average distance between optimal and target weights, $\sqrt{n}\lVert\hat{w}-w^*\rVert = 0.4$. When the standard errors were constrained at $\xi_2 = 60$, the multipliers and the average distance between optimal and target weights were all small. Further increasing the standard errors resulted in a small change in the objective function and negligible changes in the multipliers. In patient starting treatment at about 27 years of age, it appeared that the precision of the estimates for the regression coefficients of scientific interest could be reduced with no major expected loss in bias.

Tables 2 and 3 report the results in patient starting treatment at about 36 and about 44 years of age. The multipliers and objective function showed similar patterns to the population of 27-year-olds.

The magnitude of the regression coefficients varied across the three target populations,

Age	**36**					
ξ_2	(20)		(28)		(37)	
λ_1	0.0		0.0		0.0	
λ_2	70		18.3		2.8	
$\sqrt{n\|\hat{w} - w^*\|}$	0.4		0.2		0.1	
$\hat{\beta}_{1,\hat{w}}, \quad \hat{\sigma}_{1,\hat{w}}$	620	(14)	636	(19)	650	(24)
$\hat{\beta}_{2,\hat{w}}, \quad \hat{\sigma}_{2,\hat{w}}$	33	(20)	26	(28)	22	(37)

Table 2: Lagrange multipliers λ_1 and λ_2 (multiplied by 100) associated with constraints (19), and (20), respectively, the square root of objective function multiplied by the sample size $\sqrt{n\|\hat{w} - w^*\|}$, and the estimated optimal weighted coefficients with standard errors in brackets for model (23) at the optimal weights $\hat{w}$ in the population of patients starting treatment at about 36 years of age.

indicating that age modified the effect of CD4-cell count at treatment initiation on that at five years after initiation. The point estimates of the regression coefficients at the smallest precision were different from those obtained with unconstrained precision. However, they were all within the confidence intervals of the unconstrained estimates. Corroborated by the results from the simulation study described in Section 4, this led us to believe that the inference from the models with standard errors constrained at values smaller than those observed for the target estimator had high precision and acceptable bias. In all three age populations mean CD4 count at 5 years was larger at increasing levels of baseline CD4 count.

6 Conclusions

Statistical methods that use probability weights are widely popular in many areas of statistics. Unbiased weighted estimators, however, often show excessively low precision. This paper presents optimal weights that are solution to an optimization problem and yield

Age	**44**					
ξ_2	(49)		(69)		(88)	
λ_1	0.0		0.0		0.0	
λ_2	12.6		2.3		0.3	
$\sqrt{n\|\hat{w} - w^*\|}$	0.4		0.2		0.1	
$\hat{\beta}_{1,\hat{w}}, \quad \hat{\sigma}_{1,\hat{w}}$	626	(25)	629	(29)	630	(31)
$\hat{\beta}_{2,\hat{w}}, \quad \hat{\sigma}_{2,\hat{w}}$	158	(49)	189	(69)	204	(88)

Table 3: Lagrange multipliers λ_1 and λ_2 (multiplied by 100) associated with constraints (19), and (20), respectively, the square root of objective function multiplied by the sample size $\sqrt{n\|\hat{w} - w^*\|}$, and the estimated optimal weighted coefficients with standard errors in brackets for model (23) at the optimal weights $\hat{w}$ in the population of patients starting treatment at about 44 years of age.

minimum-bias estimators among all estimators with specified precision.

Unlike the traditional trimmed weights, which differ from the target weights only at the tails of their distribution, the optimal weights are uniformly closest to the target weights. This feature explains the considerable advantage of optimal weights over trimmed weights observed across all the scenarios in our simulation study. The simulation study also showed that sizable precision could often be gained at the cost of negligible bias.

The Euclidean distance utilized in this paper has an intuitive interpretation, but other measures could be used instead, such as the Bregman divergence, which includes the popular Kullback-Leibler divergence and Mahalanobis distance (Bregman, 1967). With any given set of data and inferential objective, these alternative measures may be preferable to the Euclidean distance.

In applied settings, researchers may consider the following analytic steps: (1) estimate the parameter of interest with the target weights; (2) if the precision is acceptable no further steps are necessary; (3) otherwise, constrain the precision and obtain optimal weights as described in this paper; (4) investigate the choice of ξ as suggested in Section 3.

When weights are used to identify causal quantities, the presence of extreme probability weights is related to the violations of the positivity assumption. In this situation, instead of optimizing the weights, the first thing to do is verify that the causal quantity of interest is identifiable for any possible combination of the covariates. If not, one should think about redefining the quantity before moving to the estimation step. An approach for responding to violations in the positivity assumption is to identify the corresponding observations which cause extreme weights, exclude them from the analysis for positivity violation and acknowledge the estimation does not apply to those subjects. More on the diagnosis to violations in the positivity assumption can be found in Petersen et al. (2012).

The large-sample variance estimator we used in constraint (2) is very popular. In the presence of extreme outlying probability weights and comparatively small sample sizes, however, its large-sample approximation may prove inadequate. In our study we found that constraining all weights by an upper limit, defined in equation (3), satisfactorily improved this approximation. In practical settings we generally suggest to set ϵ equal to the maximum value of the target weights w^*. When high precision is desired, we recommend to remove the most extreme target weights by setting ϵ equal to the 0.99-quantile of the distribution of w^*.

In many real settings, the probability weights are not known and fixed, but rather they are estimated from the available data. The inherent sampling error of the estimated weights can be taken into account when estimating the standard error of the resulting weighted estimator, and the variance of the two-step estimator can be used in constraint (2) (Carroll et al., 1988; Murphy and Topel, 2002; Zou et al., 2016).

References

Basu, D. (2011). An Essay on the Logical Foundations of Survey Sampling, Part One. In A. DasGupta (Ed.), *Selected Works of Debabrata Basu*, Selected Works in Probability and Statistics, pp. 167–206. Springer New York.

Beaumont, J.-F. (2008, September). A New Approach to Weighting and Inference in Sample Surveys. *Biometrika 95*(3), 539–553.

Beaumont, J.-F., D. Haziza, and A. Ruiz-Gazen (2013, September). A unified approach to robust estimation in finite population sampling. *Biometrika 100*(3), 555–569.

Borgoni, R., D. Marasini, and P. Quatto (2012). Handling nonresponse in business surveys. *Survey Research Methods 6*(3), 145–154.

Bregman, L. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *{USSR} Computational Mathematics and Mathematical Physics 7*(3), 200 – 217.

Carroll, R. J., C. F. J. Wu, and D. Ruppert (1988). The effect of estimating weights in weighted least squares. *Journal of the American Statistical Association 83*(404), 1045–1054.

Chambers, R. L. (2003). Introduction to Part A. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, pp. 11–28. John Wiley & Sons, Ltd.

Cochran, W. (1977). *Sampling Techniques* (Third ed.). Wiley.

Cox, B. and D. McGrath (1981). An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of Survey Estimates. *Paper Presented at the 1981 Biometric Society ENAR Meeting*. Richmond, VA, U.S.A.

Deeks, S. G. and A. N. Phillips (2009). Clinical review: Hiv infection, antiretroviral treatment, ageing, and non-aids related morbidity. *Bmj 338*, 288–292.

Elliot, M. and R. Little (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics 16*(3), 191.

Elliott, M. R. (2008, December). Model Averaging Methods for Weight Trimming. *Journal of official statistics 24*(4), 517–540.

Elliott, M. R. (2009, March). Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models. *Journal of official statistics 25*(1), 1–20.

Fuller, W. A. (2009). Frontmatter. In *Sampling Statistics*, pp. i–xvi. John Wiley & Sons, Inc.

HIV-CAUSAL Collaboration, L. E. Cain, R. Logan, J. M. Robins, J. A. C. Sterne, C. Sabin, L. Bansi, A. Justice, J. Goulet, A. van Sighem, F. de Wolf, H. C. Bucher, V. von Wyl, A. Esteve, J. Casabona, J. del Amo, S. Moreno, R. Seng, L. Meyer, S. Perez-Hoyos, R. Muga, S. Lodi, E. Lanoy, D. Costagliola, and M. A. Hernan (2011, April). When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Annals of Internal Medicine 154*(8), 509–515.

HSL (2016). "HSL. A collection of Fortran codes for large scale scientific computation. ".

Hulliger, B. (1995). Outlier Robust Horvitz-Thompson Estimators. *21*(1), 79–87.

Kang, J. D. Y. and J. L. Schafer (2007, November). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science 22*(4), 523–539.

Kim, J. K. and C. J. Skinner (2013, February). Weighting in survey analysis under informative sampling. *Biometrika*, ass085.

Kitahata, M. M., S. J. Gange, A. G. Abraham, B. Merriman, M. S. Saag, A. C. Justice, R. S. Hogg, S. G. Deeks, J. J. Eron, J. T. Brooks, S. B. Rourke, M. J. Gill, R. J. Bosch, J. N. Martin, M. B. Klein, L. P. Jacobson, B. Rodriguez, T. R. Sterling, G. D. Kirk, S. Napravnik, A. R. Rachlis, L. M. Calzavara, M. A. Horberg, M. J. Silverberg, K. A. Gebo, J. J. Goedert, C. A. Benson, A. C. Collier, S. E. Van Rompaey, H. M. Crane, R. G. McKaig, B. Lau, A. M. Freeman, and R. D. Moore (2009, April). Effect of Early versus Deferred Antiretroviral Therapy for HIV on Survival. *New England Journal of Medicine 360*(18), 1815–1826.

Kokic, P. and P. Bell (1994). Optimal winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics 10*(4), 419.

Murphy, K. M. and R. H. Topel (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics 20*(1), 88–97.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research 21*(1), 31–54. PMID: 21030422.

Petersen, M. L., M. J. van der Laan, S. Napravnik, J. J. Eron, R. D. Moore, and S. G. Deeks (2008, October). Long-term consequences of the delay between virologic failure of highly active antiretroviral therapy and regimen modification. *AIDS (London, England) 22*(16), 2097–2106.

Pfeffermann, D. (2009, October). Inference under informative sampling. In D. Pfeffermann and C. R. Rao (Eds.), *Sample Surveys: Inference and Analysis*, pp. 455–487. Elsevier.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it. *Survey Methodology 37*(2), 115–136.

Pfeffermann, D. and M. Sverchkov (1999, April). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002) 61*(1), 166–186.

Pfeffermann, D. and M. Y. Sverchkov (2003). Fitting Generalized Linear Models under Informative Sampling. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, pp. 175–195. John Wiley & Sons, Ltd.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Volume 225230.

Quatember, A. (2015). The pseudo-population concept. In *Pseudo-Populations*, pp. 5–51. Springer International Publishing.

Rao, J. N. K. (1966, March). Alternative Estimators in PPS Sampling for Multiple Characteristics. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002) 28*(1), 47–60.

Rivest, L.-P., D. Hurtubise, and Statistics Canada (1995). On Searls' winsorized mean for skewed populations. In *Survey methodology*, pp. 107–116.

Scott, A. J. and C. J. Wild (2011, September). Fitting regression models with response-biased samples. *Canadian Journal of Statistics 39*(3), 519–536.

Sönnerborg, A. (2016). InfCare hiv dataset. `http://infcare.se/hiv/sv/`. Accessed: 2016-03-10.

Stefanski, L. A. and D. D. Boos (2002). The calculus of m-estimation. *The American Statistician 56*(1), 29–38.

Strutz, T. (2010). *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Germany: Vieweg and Teubner.

van der Wal, W. and R. Geskus (2011). ipw: An r package for inverse probability weighting. *Journal of Statistical Software 43*(1), 1–23.

Wächter, A. and L. T. Biegler (2005, April). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming 106*(1), 25–57.

When To Start Consortium, J. A. C. Sterne, M. May, D. Costagliola, F. de Wolf, A. N. Phillips, R. Harris, M. J. Funk, R. B. Geskus, J. Gill, F. Dabis, J. M. Mir, A. C. Justice, B. Ledergerber, G. Ftkenheuer, R. S. Hogg, A. D. Monforte, M. Saag, C. Smith, S. Staszewski, M. Egger, and S. R. Cole (2009, April). Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet (London, England) 373*(9672), 1352–1363.

Wright, S. T., K. Petoumenos, M. Boyd, A. Carr, S. Downing, C. C. O'Connor, M. Grotowski, M. G. Law, and Australian HIV Observational Database study group (2013, April). Ageing and long-term CD4 cell count trends in HIV-positive patients with 5 years or more combination antiretroviral therapy experience. *HIV medicine 14*(4), 208–216.

Writing Committee for the CASCADE Collaboration (2011, September). Timing of HAART initiation and clinical outcomes in human immunodeficiency virus type 1 seroconverters. *Archives of Internal Medicine 171*(17), 1560–1569.

Zou, B., F. Zou, J. J. Shuster, P. J. Tighe, G. G. Koch, and H. Zhou (2016). On variance estimate for covariate adjustment by propensity score analysis. *Statistics in Medicine 35*(20), 3537–3548.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.