View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Toulouse Capitole Publications

N° TSE –1000

March 2019

"Improving the Estimation of the Odds Ratio in Sampling Surveys using Auxiliary Information"

Camelia Goga and Anne Ruiz-Gazen

Improving the Estimation of the Odds Ratio in Sampling Surveys using Auxiliary Information

Camelia Goga
Institut de Mathématiques de Besançon
Université de Bourgogne Franche-Comté
Dijon, France
and Anne Ruiz-Gazen
Toulouse School of Economics
Université Toulouse 1 Capitole

March 19, 2019

Abstract

The odds-ratio measure is widely used in Health and Social surveys where the aim is to compare the odds of a certain event between a population at risk and a population not at risk. It can be defined using logistic regression through an estimating equation that allows a generalization to continuous risk variable. Data from surveys need to be analyzed in a proper way by taking into account the survey weights. Because the odds-ratio is a complex parameter, the analyst has to circumvent some difficulties when estimating confidence intervals. The present paper suggests a nonparametric approach that can take advantage of some auxiliary information in order to improve on the precision of the odds-ratio estimator. The approach consists in B-spline modelling which can handle the nonlinear structure of the parameter in a flexible way and is easy to implement. The variance estimation issue is solved through a linearization approach and confidence intervals are derived. Two small applications are discussed.

Keywords: B-spline functions, estimating equation, influence function, linearization, logistic regression, survey data.

1 Introduction

In health and social surveys, the Odds Ratio (OR) is used to quantify the association between the levels of a response variable Y and a risk variable X. The value taken by Y is y_i and the value taken by X is x_i for the i-th individual in a population $U = \{1, \ldots, N\}$. Let $p_i = P(Y = 1|X = x_i)$ and the logistic regression

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_i, \quad i \in U$$

implying that $p_i = \exp(\beta_0 + \beta_1 x_i)(1 + \exp(\beta_0 + \beta_1 x_i))^{-1}$. The odds ratio is defined as [Agresti, 2002]:

$$\text{OR} = \frac{\text{odds}(Y = 1|X = x_i + 1)}{\text{odds}(Y = 1|X = x_i)} = \exp \beta_1, \tag{1}$$

where $\text{odds}(Y = 1|X = x_i+1) = P(Y = 1|X = x_i+1)/P(Y = 0|X = x_i+1)$.

The estimator of the parameter β_1 is obtained as a solution of a population estimating equation. Then, the method suggested in Binder [1983] can be used to estimate β_1 with survey data. In the context of surveys, Korn and Graubard [1999] and Heeringa et al. [2017] give details and examples of estimating an odds ratio but without taking into account auxiliary information. Concerning auxiliary information, Korn and Graubard [1999], p. 169-170, advocate the use of weighted odds ratios and Rao et al. [2002] suggest using poststratification information to estimate parameters of interest obtained as solutions of estimating equations. In the present paper, we propose to study the estimation of the odds ratio parameter when auxiliary information is available. Results are derived from Goga and Ruiz-Gazen [2014] who use auxiliary information to estimate nonlinear parameters through nonparametric methods. The solutions of estimating equations are particular nonlinear parameters but Goga and Ruiz-Gazen [2014] give few details for such estimators.

In Section 2, we propose a B-spline nonparametric estimator for the odds-ratio. In Section 3, we use linearization to derive the asymptotic variance of the estimator under broad assumptions. We also suggest a variance estimator

and give asymptotic normal confidence intervals. In Section 4, we illustrate our approach on two real data sets and conclude in Section 5 with a short discussion.

2 Odds ratio estimation in surveys using B-spline regression

2.1 Maximum likelihood estimation at the population level

In order to estimate the parameter OR, we estimate first the regression coefficient $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ by $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$, where $'$ denotes the transpose, and obtain the estimator $\widehat{\mathrm{OR}} = \exp \hat{\beta}_1$. The estimators of the regression parameters β_0 and β_1 are obtained by maximization of the population likelihood:

$$L(y_1, \ldots, y_N; \boldsymbol{\beta}) = \prod_{i \in U} p_i^{y_i} \, (1 - p_i)^{1 - y_i}.$$

Let $\mathbf{x}_i = (1 \quad x_i)'$ and $\mu(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))^{-1}$. Under the logistic regression model, the maximum likelihood estimator of $\boldsymbol{\beta}$ satisfies:

$$\sum_{i \in U} \mathbf{x}_i (y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta})) = 0 \tag{2}$$

or $\sum_{i \in U} \mathbf{t}_i(\boldsymbol{\beta}) = 0$ with $\mathbf{t}_i(\boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))$.This equation is called also the score equation and $\mathbf{t}_i(\boldsymbol{\beta})$ the score function. The regression estimator of $\boldsymbol{\beta}$ is defined as an implicit solution of the estimating equation (2) and we use iterative methods such as the Newton-Raphson algorithm to compute it.

2.2 Estimation at the sample level using B-spline non-parametric models

For a sample s selected from the population U according to a sample design $p(\cdot)$, we denote by $\pi_i > 0$ the probability of unit i to be selected in the sample and $\pi_{ij} > 0$ the joint probability of units i and j to be selected in the sample with $\pi_{ii} = \pi_i$. We look for an estimator of $\boldsymbol{\beta}$ and of OR taking the auxiliary variable Z, with values $z_1, \ldots, z_N$, into account.

The regression coefficient $\boldsymbol{\beta}$ is a nonlinear finite population function of totals defined by the implicit equation (2). The functional method by Deville [1999], extended to the nonparametric case by Goga and Ruiz-Gazen [2014], is used to build a nonparametric estimator of $\boldsymbol{\beta}$. Let $M = \sum_{i \in U} \delta_{y_i}$ be the finite measure assigning the unit mass to each y_i, $i \in U$, and zero elsewhere, where δ_{y_i} is the Dirac function at y_i, $\delta_{y_i}(y) = 1$ for $y = y_i$ and zero elsewhere. Consider also the functional T defined by

$$T(M; \boldsymbol{\beta}) = \sum_{i \in U} \mathbf{x}_i (y_i - \mu(\mathbf{x}_i' \boldsymbol{\beta})) = \sum_{i \in U} \mathbf{t}_i(\boldsymbol{\beta}). \tag{3}$$

Then, the regression coefficient $\boldsymbol{\beta}$ is the solution of the implicit equation

$$T(M; \boldsymbol{\beta}) = 0. \tag{4}$$

The measure M may be estimated by using the Horvitz-Thompson weights $d_i = 1/\pi_i$ or the linear calibration weights [Deville, 1999]. The functional method allows us to use nonparametric weights for estimating the logistic regression coefficient. Remark that the method is general and may be applied for any parameter $\boldsymbol{\beta}$ defined as a solution of estimating equations.
Goga [2005] suggests using nonparametric weights based on B-spline regression to estimate totals for variables which are related nonlinearly to the auxiliary information and Goga and Ruiz-Gazen [2014] suggest penalized B-spline regression to estimate totals or nonlinear parameters such as a Gini index. The B-splines functions [Dierckx, 1995] are known for their flexibility to model nonlinear trend in the data and by their numerical stability and ease of implementation. Let $B_1, \ldots, B_q$, where $q = m + K$ denote the B-spline functions of degree m and with K interior knots. Then, the B-spline nonparametric weights [Goga, 2005] are given by:

$$w_{is}^b = d_i \left(\sum_{k \in U} \mathbf{b}(z_k) \right)' \left(\sum_{k \in s} d_k \mathbf{b}(z_k) \mathbf{b}'(z_k) \right)^{-1} \mathbf{b}(z_i), \tag{5}$$

where $\mathbf{b}(z_i) = (B_1(z_i), \ldots, B_q(z_i))'$. The weights w_{is}^b depend only on the auxiliary variable and are similar to calibration weights [Deville and Särndal, 1992]. They allow to estimate exactly the population size N, $\sum_{i \in s} w_{is}^b = N$, and the total of the auxiliary variable Z, $\sum_{i \in s} w_{is}^b z_i = \sum_{i \in U} z_i$. We use here w_{is}^b to estimate the logistic regression coefficient and the odds ratio efficiently. More exactly, we estimate M by $\widehat{M} = \sum_{i \in s} w_{is}^b \delta_{y_i}$. Plugging $\widehat{M}$ into the

functional expression of $\boldsymbol{\beta}$ given by (4) yields the B-spline nonparametric estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$:

$$T(\widehat{M};\widehat{\boldsymbol{\beta}}) = 0, \tag{6}$$

which means that $\widehat{\boldsymbol{\beta}}$ is the solution of the implicit equation $\sum_{i\in s} w_{is}^b \mathbf{x}_i(y_i - \mu(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})) = 0$.

An iterative Newton-Raphson method is used to compute $\widehat{\boldsymbol{\beta}}$. Consider for that the derivative of the functional T given in (3) with respect to $\boldsymbol{\beta}$:

$$\frac{\partial T}{\partial \boldsymbol{\beta}} = -\sum_{i\in U} \nu(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' = \mathbf{X}'\boldsymbol{\Lambda}(\boldsymbol{\beta})\mathbf{X} := \mathbf{J}(\boldsymbol{\beta}), \tag{7}$$

with $\mathbf{X} = (\mathbf{x}_i')_{i\in U}$ and $\boldsymbol{\Lambda}(\boldsymbol{\beta}) = -\text{diag}(\nu(\mathbf{x}_i'\boldsymbol{\beta}))$ with $\nu(\mathbf{x}_i'\boldsymbol{\beta}) = \mu(\mathbf{x}_i'\boldsymbol{\beta})(1 - \mu(\mathbf{x}_i'\boldsymbol{\beta}))$. The 2×2 matrix $\mathbf{X}'\boldsymbol{\Lambda}(\boldsymbol{\beta})\mathbf{X}$ is invertible and $\mathbf{J}(\boldsymbol{\beta})$ is definite negative. From (7), the matrix $\mathbf{J}(\boldsymbol{\beta})$ is unknown and may be estimated by using the nonparametric weights w_{is}^b:

$$\widehat{\mathbf{J}}_w(\boldsymbol{\beta}) = -\sum_{i\in s} w_{is}^b \nu(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i\mathbf{x}_i' = \mathbf{X}_s'\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta})\mathbf{X}_s, \tag{8}$$

where $\widehat{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) = -\text{diag}(w_{is}^b\nu(\mathbf{x}_i'\boldsymbol{\beta}))_{i\in s}$ and $\mathbf{X}_s = (\mathbf{x}_i')_{i\in s}$. Then, the r-th step of the Newton-Raphson algorithm is:

$$\widehat{\boldsymbol{\beta}}_r = \widehat{\boldsymbol{\beta}}_{r-1} - \widehat{\mathbf{J}}_w(\widehat{\boldsymbol{\beta}}_{r-1})T(\widehat{M};\widehat{\boldsymbol{\beta}}_{r-1}), \tag{9}$$

where $\widehat{\boldsymbol{\beta}}_{r-1}$ is the value of $\widehat{\boldsymbol{\beta}}$ obtained at the $(r-1)$-th step. $\widehat{\mathbf{J}}_w(\widehat{\boldsymbol{\beta}}_{r-1})$ is the value of $\widehat{\mathbf{J}}_w(\boldsymbol{\beta})$ and $T(\widehat{M};\widehat{\boldsymbol{\beta}}_{r-1})$ the value of $T(\widehat{M};\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{r-1}$. Iterating to convergence produces the nonparametric estimator $\widehat{\boldsymbol{\beta}}$ and the estimated Jacobian matrix $\widehat{\mathbf{J}}_w(\widehat{\boldsymbol{\beta}})$. The odds ratio is estimated by $\widehat{\text{OR}} = \exp(\hat{\beta}_1)$ and $\widehat{\mathbf{J}}_w(\widehat{\boldsymbol{\beta}})$ is used in Section 3 to estimate the variance of $\hat{\boldsymbol{\beta}}$.

3 Variance estimation and confidence intervals

3.1 Asymptotic variance of the B-spline estimator of OR

The coefficient $\boldsymbol{\beta}$ of the logistic regression defined in (2) is a nonlinear function of totals and the nonparametric weights w_{is}^b add even more nonlinearity.

We approximate $\widehat{\boldsymbol{\beta}}$ in (6) by a linear estimator in two steps: we first treat the nonlinearity due to $\boldsymbol{\beta}$, and second the nonlinearity due to the nonparametric estimation. This procedure is different from Deville [1999]. From the implicit function theorem, there exists a unique functional $\widetilde{T}$ such that

$$\widetilde{T}(M) = \boldsymbol{\beta} \quad \text{and} \quad \widetilde{T}(\widehat{M}) = \widehat{\boldsymbol{\beta}}. \tag{10}$$

The functional $\widetilde{T}$ is Fréchet differentiable with respect to M. The derivative of $\widetilde{T}$ with respect to M, called the influence function, is defined by

$$I\widetilde{T}(M, \xi) = \lim_{\lambda \to 0} \frac{\widetilde{T}(M + \lambda\delta_\xi) - \widetilde{T}(M)}{\lambda},$$

where δ_ξ is the Dirac function at ξ. Under the assumptions given in Goga and Ruiz-Gazen [2014], we obtain the following first-order expansion:

$$\widetilde{T}(\widehat{M}) = \widetilde{T}(M) + \sum_{i \in U} (w_{is}^b - 1) I\widetilde{T}\,(M, y_i) + o_p(n^{-1/2}). \tag{11}$$

For $i \in U$, $I\widetilde{T}\,(M, y_i) = \mathbf{u}_i$ is called the linearized variable of $\widetilde{T}(M) = \boldsymbol{\beta}$ and equals:

$$\begin{aligned} \mathbf{u}_i &= -\left(\frac{\partial T}{\partial \boldsymbol{\beta}}\right)^{-1} IT(M, y_i; \boldsymbol{\beta}) = -\left(\mathbf{X}'\boldsymbol{\Lambda}(\boldsymbol{\beta})\mathbf{X}\right)^{-1} \mathbf{x}_i (y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta})) \\ &= -\mathbf{J}^{-1}(\boldsymbol{\beta}) \cdot \mathbf{t}_i(\boldsymbol{\beta}). \end{aligned} \tag{12}$$

The linearized variable $\mathbf{u}_i = (u_{i,0}, u_{i,1})'$ is a two-dimensional vector depending on the unknown parameter $\boldsymbol{\beta}$ and on totals contained in the matrix $\mathrm{J}(\boldsymbol{\beta})$. The second component $u_{i,1}$ of $\mathbf{u}_i$ is the linearized variable of β_1. Note that with a binary variable X, the odds ratio is given by $\mathrm{OR} = (N_{00}N_{11})/(N_{01}N_{10})$ where N_{00}, N_{01}, N_{10}, and N_{11} are the population counts associated with the contingency table. In this case, the linearized variable of β_1 has the expression:

$$u_{i,1} = \frac{1_{\{x_i=0, y_i=0\}}}{N_{00}} + \frac{1_{\{x_i=1, y_i=1\}}}{N_{11}} - \frac{1_{\{x_i=1, y_i=0\}}}{N_{10}} - \frac{1_{\{x_i=0, y_i=1\}}}{N_{01}} \tag{13}$$

and the same expression is obtained from (12) after some algebra.
Relation (11) may be written as:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \simeq \sum_{i \in s} w_{is}^b \mathbf{u}_i - \sum_{i \in U} \mathbf{u}_i, \tag{14}$$

namely, the B-spline nonparametric regression estimator $\hat{\boldsymbol{\beta}}$ is approximated by the weighted estimator $\sum_{i\in s} w_{is}^b \mathbf{u}_i$ of the finite population total of the linearized variable $\mathbf{u}_i$. In the following, the aim is to derive the asymptotic variance of $\hat{\boldsymbol{\beta}}$.

Using the weights d_i instead of w_{is}^b in (14) implies that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is:

$$\begin{aligned} \mathrm{AV}(\hat{\boldsymbol{\beta}}) = \mathrm{Var}\left(\sum_{i\in s} d_i \mathbf{u}_i\right) &= \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j) d_i d_j \mathbf{u}_i \mathbf{u}_j' \\ &= \mathbf{J}^{-1}(\boldsymbol{\beta})\,\mathrm{Var}(\hat{\mathbf{t}}_d(\boldsymbol{\beta}))\,\mathbf{J}^{-1}(\boldsymbol{\beta}), \end{aligned} \tag{15}$$

where $\mathrm{Var}(\hat{\mathbf{t}}_d(\boldsymbol{\beta}))$ is the variance of $\hat{\mathbf{t}}_d(\boldsymbol{\beta}) = \sum_{i\in s} d_i \mathbf{t}_i(\boldsymbol{\beta})$ with $\mathbf{t}_i(\boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))$:

$$\mathrm{Var}(\hat{\mathbf{t}}_d(\boldsymbol{\beta})) = \mathrm{Var}\left(\sum_{i\in s} d_i \mathbf{t}_i(\boldsymbol{\beta})\right) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j) d_i d_j \mathbf{t}_i(\boldsymbol{\beta})\mathbf{t}_j'(\boldsymbol{\beta}). \tag{16}$$

Note that Binder [1983] gives the same asymptotic expression for the variance.

For B-spline basis functions formed by step functions on intervals between knots ($m = 1$), the weights w_{is}^b yield the post-stratified estimator of $\boldsymbol{\beta}$ [Rao et al., 2002]. Linear calibration weights lead to the case treated by Deville [1999]. Consider now the general case of nonparametric weights w_{is}^b given in (5), then the right hand side of (14) is a nonparametric estimator for the total of the linearized variable $\mathbf{u}_i$ and a supplementary linearization step is needed. It can be written as a generalized regression estimator (GREG):

$$\sum_{i\in s} w_{is}^b \mathbf{u}_i - \sum_{i\in U} \mathbf{u}_i = \sum_{i\in s} d_i(\mathbf{u}_i - \widehat{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)) - \sum_{i\in U}(\mathbf{u}_i - \widehat{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)),$$

where $\widehat{\boldsymbol{\theta}}_u = (\sum_{i\in s} d_i \mathbf{b}(z_i)\mathbf{b}'(z_i))^{-1}(\sum_{i\in s} d_i \mathbf{b}(z_i)\mathbf{u}_i')$. In order to derive the asymptotic variance of the nonparametric estimator of $\boldsymbol{\beta}$, we assume that $||\mathbf{x}_i|| < C$ for all $i \in U$ with C a positive constant independent of i and N, and $||\cdot||$ is the Euclidian norm. Then, the linearized variable verifies $N||\mathbf{u}_i|| = O(1)$ uniformly in i, because

$$N||\mathbf{u}_i|| \leq ||N\mathbf{J}^{-1}(\boldsymbol{\beta})||_2\, ||\mathbf{x}_i||\, |y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))| = O(1).$$

where the matrix norm $||\cdot||_2$ is defined by $||\mathbf{A}||_2^2 = \mathrm{tr}(\mathbf{A}'\mathbf{A})$.

Under the assumptions of Theorem 7 in Goga and Ruiz-Gazen [2014] on the B-splines functions and the sampling design, the nonparametric estimator $\sum_{i\in s} w_{is}^b \mathbf{u}_i$ is asymptotically equivalent to

$$\sum_{i\in s} w_{is}^b \mathbf{u}_i - \sum_{i\in U} \mathbf{u}_i \simeq \sum_{i\in s} d_i(\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)) - \sum_{i\in U}(\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)), \tag{17}$$

where $\widetilde{\boldsymbol{\theta}}_u = (\sum_{i\in U} \mathbf{b}(z_i)\mathbf{b}'(z_i))^{-1} \sum_{i\in U} \mathbf{b}(z_i)\mathbf{u}_i'$. This states that the B-spline nonparametric estimator of $\sum_{i\in U} \mathbf{u}_i$ is asymptotically equivalent to the generalized difference estimator. We interpret this result as fitting a nonparametric model on the linearized variable $\mathbf{u}_i$ taking into account the auxiliary information z_i. Nonparametric models are a good choice when the linearized variable obtained from the first linearization step does not depend linearly on z_i, as it is the case in the logistic regression, which implies a second linearization step.

Putting together (14) and (17), we can approximate the variance of $\hat{\boldsymbol{\beta}}$ by the Horvitz-Thompson variance of the residuals $\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)$,

$$\mathrm{AV}(\hat{\boldsymbol{\beta}}) = \sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j) d_i d_j \left(\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)\right)\left(\mathbf{u}_j - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_j)\right)'. \tag{18}$$

The B-spline nonparametric fitting allows large flexibility and implies that the residuals $\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i)$ have a smaller dispersion than with a linear fitting regression.

We write the asymptotic variance in (18) in a matrix form similar to (15). We have

$$\mathbf{u}_i - \widetilde{\boldsymbol{\theta}}_u' \mathbf{b}(z_i) = -\mathbf{J}^{-1}(\boldsymbol{\beta})\left(\mathbf{t}_i(\boldsymbol{\beta}) - \widetilde{\boldsymbol{\theta}}_t' \mathbf{b}(z_i)\right)$$

with

$$\widetilde{\boldsymbol{\theta}}_t = (\sum_{i\in U} \mathbf{b}(z_i)\mathbf{b}'(z_i))^{-1} \sum_{i\in U} \mathbf{b}(z_i)\mathbf{t}_i'(\boldsymbol{\beta})$$

and $\mathbf{t}_i$ the score functions. Then, the asymptotic variance of $\hat{\boldsymbol{\beta}}$ becomes:

$$\mathrm{AV}(\hat{\boldsymbol{\beta}}) = \mathbf{J}^{-1}(\boldsymbol{\beta})\,\mathrm{Var}(\hat{\mathbf{e}}_d(\boldsymbol{\beta}))\,\mathbf{J}^{-1}(\boldsymbol{\beta}) \tag{19}$$

where $\hat{\mathbf{e}}_d(\boldsymbol{\beta}) = \sum_{i\in s} d_i \mathbf{e}_i(\boldsymbol{\beta})$ is the Horvitz-Thompson estimator of the residual $\mathbf{e}_i(\boldsymbol{\beta}) = \mathbf{t}_i(\boldsymbol{\beta}) - \widetilde{\boldsymbol{\theta}}_t' \mathbf{b}(z_i)$ of $\mathbf{t}_i(\boldsymbol{\beta})$ using B-spline nonparametric estimation and $\mathrm{Var}(\hat{\mathbf{e}}_d(\boldsymbol{\beta}))$ is obtained as in (16). Result given in (19) shows that improving the estimation of $\boldsymbol{\beta}$ is equivalent to improving the estimation of the score functions $\mathbf{t}_i = \mathbf{x}_i(y_i - \mu(\mathbf{x}_i'\boldsymbol{\beta}))$.

3.2 Variance estimation and confidence interval for the odds ratio

The linearized variable $\mathbf{u}_i$ is unknown and is estimated by:

$$\hat{\mathbf{u}}_i \quad = \quad -\widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}})\,\mathbf{x}_i(y_i - \mu(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})) = -\widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}})\,\hat{\mathbf{t}}_i$$

where the matrix $\widehat{\mathbf{J}}_w$ is computed according to (8) and $\hat{\mathbf{t}}_i$ is the estimation of $\mathbf{t}_i(\boldsymbol{\beta})$ with $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Assuming that all $\pi_{ij} > 0$, the asymptotic variance $\mathrm{AV}(\widehat{\boldsymbol{\beta}})$ given in (18) or (19) is estimated by the Horvitz-Thompson variance estimator with $\mathbf{u}_i$ replaced by $\hat{\mathbf{u}}_i$:

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = \sum_{i\in s}\sum_{j\in s}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}d_i d_j \hat{\mathbf{u}}_i\hat{\mathbf{u}}_j' \quad = \quad \widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}})\,\widehat{\mathrm{V}}_{\mathrm{HT}}(\hat{\mathbf{e}}_d(\widehat{\boldsymbol{\beta}}))\,\widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}}) \quad (20)$$

where $\widehat{\mathrm{V}}_{\mathrm{HT}}(\hat{\mathbf{e}}_d)$ is the Horvitz-Thompson variance estimator of $\hat{\mathbf{e}}_d(\hat{\boldsymbol{\beta}}) = \sum_{i\in s} d_i\hat{\mathbf{e}}_i(\hat{\boldsymbol{\beta}})$ with $\hat{\mathbf{e}}_i(\hat{\boldsymbol{\beta}}) = \hat{\mathbf{t}}_i - \widehat{\boldsymbol{\theta}}_{\hat{t}}'\mathbf{b}(z_i)$ and $\widehat{\boldsymbol{\theta}}_{\hat{t}} = (\sum_{i\in s} d_i\mathbf{b}(z_i)\mathbf{b}'(z_i))^{-1}\sum_{i\in s} d_i\mathbf{b}(z_i)\hat{\mathbf{t}}_i'$.

The variance estimator of $\hat{\beta}_1$ is obtained from (20) as:

$$\widehat{V}(\hat{\beta}_1) = \widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}})\,\widehat{\mathrm{V}}_{\mathrm{HT}}(\hat{e}_{d,2}(\widehat{\boldsymbol{\beta}}))\,\widehat{\mathbf{J}}_w^{-1}(\widehat{\boldsymbol{\beta}}),$$

where $\hat{e}_{d,2}(\widehat{\boldsymbol{\beta}})$ is the second component of $\hat{\mathbf{e}}_d(\widehat{\boldsymbol{\beta}})$ so that, under regularity conditions, the $(1-\alpha)\%$ normal interval for OR is:

$$\mathrm{CI}_{1-\alpha}(\mathrm{OR}) = \left[\exp\left(\hat{\beta}_1 - z_{\alpha/2}\left(\hat{V}(\hat{\beta}_1)\right)^{1/2}\right), \exp\left(\hat{\beta}_1 + z_{\alpha/2}\left(\hat{V}(\hat{\beta}_1)\right)^{1/2}\right)\right],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of a $\mathcal{N}(0,1)$ variable. It is not symmetric around the estimated odds ratio but provides more accurate coverage rates of the true population value for a specified α [Heeringa et al., 2017]).

4 Two small applications

We compare the asymptotic variance of different estimators of the odds ratio in the simple case of one binary risk variable using two data sets. As previously mentioned, in this context, the odds ratio is a simple function of four counts. We focus on the simple random sampling without replacement and compare three estimators. The first one is the Horvitz-Thompson estimator

which does not use the auxiliary variable and whose asymptotic variance is given by (15). The second estimator is the generalized regression estimator which takes the auxiliary variable into account through a linear model, fitting the linearized variable against the auxiliary variable. The third estimator is the B-spline calibration estimator with an asymptotic variance given by (19). In order to gain efficiency, the auxiliary variable has to be related to the linearized variable. In the context of one binary factor, the linearized variable is given by (13) and takes four different values, which depend on the values of the variables X and Y. In order to be related to the linearized variable, the auxiliary variable has to be related to the product of the two variables X and Y, which is a strong property. Moreover, because $u_{i,1}$, X, and Y are discrete, using auxiliary information does not necessarily lead to an important gain in efficiency as illustrated by the first health survey example. The gain in efficiency however is significant in some other cases. In the second example using labor survey data, the gain in using the B-splines calibration estimator compared to the Horvitz-Thompson estimator is significant because the auxiliary variable is related to the variable Y but also to the factor X; X and Y being related to one another, too.

Example from the California Health Interview Survey

The data set comes from the Center for Health Policy Research at the University of California. It was extracted from the adult survey data file of the California Health Interview Survey in 2009 and consists of 11074 adults. The response dummy variable equals one if the person is currently insured; the binary factor equals one if the person is currently a smoker. The auxiliary variable is age and we consider people who are less than 60 years old. The data are presented in detail in Lumley [2011].

We compare the Horvitz-Thompson, the generalized regression, and the B-splines calibration estimators in terms of asympotic variance. In order to calculate the B-splines functions, we use the SAS procedure *transreg* and take $K = 15$ knots and B-splines of degree $m = 3$. The gain in using the generalized regression estimator compared to the Horvitz-Thompson estimator is only 0.01%. It is 1.5% when using B-splines instead of the generalized regression. When changing the number of knots and the degree of the B-spline functions, the results remain similar and the gain remains under 2%. In this example, there is no gain in using auxiliary information even with flexible B-splines, because the auxiliary variable is not related enough to the linearized variable. The linearized variable takes negative values for smokers

without insurance and non smokers with insurance, positive values for smokers with insurance and non smokers without insurance. Age is not a good predictor for this variable, because we expect to find sufficient people of any age in each of the four categories (smokers/non smokers $\times$ insurance/no insurance). Incorporating this auxiliary information brings no gain.

Example from the French Labor Survey

We consider 14621 wage-earners under 50 years of age, from the French labour force survey. The initial data set consists of monthly wages in 2000 and 1999. A dummy variable W00 equals one if the monthly wage in 2000 exceeds 1500 euros and zero otherwise. The same for W99 in 1999. The population is divided in lower and upper education groups. The value of the categorical factor DIP equals one for people with a university degree and zero otherwise. W00 corresponds to the binary response variable Y while the diploma variable DIP corresponds to the risk variable X. The variable W99 is the auxiliary variable Z. In this context, the odds ratio is a simple function of four counts. We focus on the simple random sampling without replacement and compare three estimators. The first one is the Horvitz-Thompson estimator which does not use the auxiliary variable and whose asymptotic variance is given by (15). The second estimator is the generalized regression estimator which takes the auxiliary variable into account through a linear model, fitting the linearized variable against the auxiliary variable. The third estimator is the B-spline calibration estimator with an asymptotic variance given by (19).

To compare the Horvitz-Thompson estimator with the generalized regression estimator and the B-splines calibration estimator, we first calculate the gain in terms of asymptotic variance. We consider $K = 15$ knots and the degree $m = 3$. The gain in using the generalized estimator compared to the Horvitz-Thompson estimator is 20%. It is 33% when using B-splines. The result is almost independent of the number of knots and, of the degree of B-spline functions. When the total number of knots varies from 5 to 50 and the degree varies from 1 to 5, the gain is between 32% and 34%. The nonlinear link between the linearized variable of a complex parameter with the auxiliary variable explains the gain in using a nonparametric estimator compared to an estimator based on a linear model [Goga and Ruiz-Gazen, 2014]. For the odds ratio with one binary factor, the linearized variable is discrete and the linear model does not fit the data.

5 Discussion

In the presence of one auxiliary variable known for all the population units, a B-splines approach is easy to implement and can improve on the precision of the Horvitz-Thompson estimator for an odds-ratio parameter if the auxiliary variable is well related with the variable of interest. It is possible to take into account more than one auxiliary variable by using some generalized additive model and consider some B-splines estimator as proposed above for each of the additive components. The theory however needs further development.

Acknowledgement: we thank Benoît Riandey for drawing our attention to the odds ratio.

References

A. Agresti. Categorical data analysis. john wiley & sons. *Inc., Publication*, 2002.

D. A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 279–292, 1983.

J. C. Deville. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology*, 25(2):193–204, 1999.

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

P. Dierckx. *Curve and Surface Fitting with Splines*. Monographs on numerical analysis. Clarendon Press, 1995. ISBN 9780198534402. URL `https://books.google.fr/books?id=-RIQ3SR0sZMC`.

C. Goga. Réduction de la variance dans les sondages en présence d'information auxiliaire: Une approche non paramétrique par splines de régression. *Canadian Journal of Statistics*, 33(2):163–180, 2005.

C. Goga and A. Ruiz-Gazen. Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):113–140, 2014.

S. G. Heeringa, B. T. West, and P. A. Berglund. *Applied Survey Data Analysis*. Chapman and Hall/CRC, 2017.

E. Korn and B. Graubard. *Analysis of Health Surveys*. Wiley Series in Survey Methodology. Wiley, 1999. ISBN 9780471137733. URL https://books.google.fr/books?id=y1wZAQAAIAAJ.

T. Lumley. *Complex Surveys: A Guide to Analysis Using R*. Wiley Series in Survey Methodology. Wiley, 2011. ISBN 9781118210932. URL https://books.google.fr/books?id=L96ludyhFBsC.

J. Rao, W. Yung, and M. Hidiroglou. Estimating equations for the analysis of survey data using poststratification information. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 364–378, 2002.