WEIGHTED-BOOTSTRAP ALIGNMENT OF EXPLANATORY VARIABLES

Peter Hall[1,2], Xiaoyan Leng[3,*] and Hans-Georg Müller[2]

March 2007

Abstract

Adjustment for covariates is a time-honored tool in statistical analysis and is often implemented by including the covariates that one intends to adjust as additional predictors in a model. This adjustment often does not work well when the underlying model is misspecified. We consider here the situation where we compare a response between two groups. This response may depend on a covariate for which the distribution differs between the two groups one intends to compare. This creates the potential that observed differences are due to differences in covariate levels rather than "genuine" population differences that cannot be explained by covariate differences. We propose a bootstrap based adjustment method. Bootstrap weights are constructed with the aim of aligning bootstrap-weighted empirical distributions of the covariate between the two groups. Generally, the proposed weighted-bootstrap algorithm can be used to align or match the values of an explanatory variable as closely as desired to those of a given target distribution. We illustrate the proposed bootstrap adjustment method in simulations and in the analysis of data on the fecundity of historical cohorts of French-Canadian women.

Key words: Adjustment, Distribution function, French-Canadian women, Group comparison, Matching, Observational study, Odds ratio.

---

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Department of Statistics, University of California, Davis, CA 95616–8705, USA

[3] Department of Biostatistical Sciences, Wake Forest University School of Medicine, Winston-Salem, NC 27157–1063, USA

[*] Corresponding author

1 Introduction

When analyzing the effect of treatments or observed predictor categories on an observed response, it is well known that observed differences in the response may be caused by differences in covariates that are associated with the observed categories or treatment groups. Differences in response caused by such imbalances in covariates are however often not of much interest and therefore need to be adjusted, if we aim at the real effect of interest, which is the effect that would be observed under equal covariate distributions.

Methods that have been devised to avoid or adjust for such imbalances include various adjustment techniques such as propensity score or regressing the response on covariates. An example where the need for such adjustment arises is provided by studies that report differences of brain size between women and men. Such differences are highly significant if just brain sizes are compared. But this comparison per se is not necessarily meaningful, as it fails to adjust for general body size differences (Kimura, 1987; Ankney, 1992). While this exemplifies a situation where significant differences may arise due to a predictor variable which is omitted or is inappropriately adjusted for, the reverse also occurs, i.e., the detection of significant differences sometimes requires proper adjustment for covariates, as pointed out by Yu and Gastwirth (2003).

In section 5 below we discuss another example regarding the dependence of number of children on immigration status for French-Canadian women in a historical cohort. In this example, the age-at-death distribution differs between the immigrant and native-born cohorts, and at the same time age-at-death (during the fertile period) is strongly related to the number of children a woman has. These issues are related to the so-called "grandmother hypothesis" of human reproduction (Lahdenperä et al., 2004). The question then arises as to how to adjust for the differences in age-at-death between these two cohorts. We demonstrate that the proposed weighted-bootstrap provides a viable option for achieving this goal.

If a covariate is known to have a linear effect on the response, adjustment works by including this covariate in a linear regression model. This commonly used adjustment technique is not suitable when the effect of the covariate on the response is nonlinear (Hauck et al., 1998; Pocock et al., 2002). If an underlying regression relation is assumed to be linear, but in reality the regression relation is nonlinear and the fitted linear regression then must be interpreted as a projection of the underlying nonlinear regression relation on a linear approximation, issues of the effectiveness of such misspecified adjustments arise. Tests for linearity have low power against

detecting the presence of a nonlinear component in such settings (Dette and Munk, 1998; Eubank and Hart, 1992; Eubank and Spiegelman, 1990).

Hall et al. (1997) proposed a covariate-matched one-sided test based on bootstrap methods, which implicitly corrects for covariate differences. This approach is appropriate when treatment effects have means that are continuous functions of the covariates. It increases power over techniques that do not adjust for the differences in the covariate levels. Here we consider a more general notion of adjustment via weighted bootstrap. Bhattacharya and Gastwirth (1999) proposed a bandwidth-matched version of the Mantel-Haenszel estimator to construct odds-ratio estimates when the response variable is binary and there are differences in the distributions of a covariate. They showed that their estimator is consistent for the treatment main effect.

In this article, we propose a bootstrap method for aligning an explanatory variable in one population with that of another population or to any pre-specified distribution, so that the marginal distributions of the two treatment groups become similar. The proposed method is based on weighted-bootstrap alignment. Section 2 describes the proposed methodology in detail. Implementation of the weighted-bootstrap alignment is the theme of section 3. Simulations reported in section 4 illustrate the approach. We discuss the case of logistic regression in more detail. In section 5 we apply the weighted-bootstrap algorithm to a set of historical French-Canadian data on the fertility of female cohorts, with the goal of adjusting for age-at-death.

2 Weighted Bootstrap Alignment

In this section we develop the basic methodology. Its implementation to the adjustment problem will be discussed in the next section. Suppose data sets $\mathcal{X}^{(c)} = \{(X_i^{(c)}, Y_i^{(c)}),\ 1 \le i \le n_c\}$ are available for $c = 1, 2$, where the Xs are predictors or explanatory variables and the Ys are responses. We wish to devise an adjustment procedure by aligning the explanatory variables so that they have similar marginal populations. Attention is confined to that part of the supports of the two explanatory variables that is common to both samples, and we assume that the supports of the distributions of the explanatory variables are overlapping throughout for both populations.

Let $p^{(c)} = (p_1^{(c)}, \ldots, p_{n_c}^{(c)})$ denote a probability distribution on the set $\{1, \ldots, n_c\}$, and define

$$F^{(c)}\big(x|p^{(c)}\big) = \sum_{i=1}^{n_c} p_i^{(c)}\, I\big(X_i^{(c)} \le x\big)\,,$$

the weighted-bootstrap distribution function corresponding to explanatory variables in $\mathcal{X}^{(c)}$. Put

$$T\big(p^{(1)}, p^{(2)}\big) = \int \big\{F^{(1)}\big(x|p^{(1)}\big) - F^{(2)}\big(x|p^{(2)}\big)\big\}^2 \, w(x)\, dx\,, \tag{1}$$

denoting a measure of the distance between $F^{(1)}(\cdot|p^{(1)})$ and $F^{(2)}(\cdot|p^{(2)})$. Here, w is a bounded, nonnegative weight function, for example the density of a probability distribution with location and scale similar to those of the data. The default choice is a constant. Let $D^{(c)}(p^{(c)})$ be a measure of distance between the uniform probability distribution, $p^{(c)}_{\text{unif}} = (1/n_c, \ldots, 1/n_c)$, on $\{1, \ldots, n_c\}$, and the more general multinomial distribution, $p^{(c)}$. We wish to choose $p^{(1)}$ and $p^{(2)}$ so as to reduce the distance between $F^{(1)}(\cdot|p^{(1)})$ and $F^{(2)}(\cdot|p^{(2)})$, relative to that in the case $p^{(1)} = p^{(1)}_{\text{unif}}$ and $p^{(2)} = p^{(2)}_{\text{unif}}$.

There are several ways of achieving this end. In particular, (a) we could keep $F^{(2)}(\cdot|p^{(2)})$ fixed and move $F^{(1)}(\cdot|p^{(1)})$ towards it (or vice versa); or (b) we could simultaneously move each of $F^{(1)}(\cdot|p^{(1)})$ and $F^{(2)}(\cdot|p^{(2)})$ towards the other. These methods are described more closely as follows: (a) given a candidate value t of $T(p^{(1)}, p^{(2)}_{\text{unif}})$, choose $p^{(1)}$ to minimize $D^{(1)}(p^{(1)})$ subject to $T(p^{(1)}, p^{(2)}_{\text{unif}}) = t$; or (b) given a candidate value t of $T(p^{(1)}, p^{(2)})$, choose $p^{(1)}$ and $p^{(2)}$ simultaneously to minimize $D^{(1)}(p^{(1)}) + \alpha\, D^{(2)}(p^{(2)})$ subject to $T(p^{(1)}, p^{(2)}) = t$. In the latter case, α is a constant selected by the data analyst, perhaps taken equal to 1 or to n_2/n_1.

Many distance measures $D^{(c)}$ are possible, but for definiteness we consider only those in the power divergence class (Cressie and Read, 1984), more precisely the following subset of this class:

$$\begin{aligned}
D^{(c)}\big(p^{(c)}\big) &= n_c - \textstyle\sum_{i=1}^{n_c} \big(n_c p_i^{(c)}\big)^\rho \quad \text{where } 0 < \rho < 1, \text{ or} \\
D^{(c)}\big(p^{(c)}\big) &= -\textstyle\sum_{i=1}^{n_c} \log\big(n_c p_i^{(c)}\big) \quad \text{where } \rho = 0, \text{ or} \\
D^{(c)}\big(p^{(c)}\big) &= \textstyle\sum_{i=1}^{n_c} p_i^{(c)} \log\big(n_c p_i^{(c)}\big) \quad \text{where } \rho = 1.
\end{aligned} \tag{2}$$

Each distance function vanishes if $p^{(c)} = p^{(c)}_{\text{unif}}$ and is strictly positive otherwise.

We could choose the measure of distance between distributions of explanatory variables to be different from (1). We also note the similarity of the third divergence measure in (2) with Theil's dissimilarity index (Theil, 1967). The main advantage of measures (1) and (2) is that their derivatives are linear in elementary functions of the $p_i^{(c)}$s. Further details on the solution of the constrained optimization problem

$$D^{(c)}\big(p^{(c)}\big) = \min \quad \text{subject to} \quad T\big(p^{(1)}, p^{(2)}\big) = t \tag{3}$$

are provided in Appendix A1. Note that the equation determining the solution (8) yields a set of n_1 equations for $p_1^{(1)}, \ldots, p_{n_1}^{(1)}$, given t.

One way of choosing t is to interpret $T(p^{(1)}, p^{(2)})$ as a goodness of fit statistic, and use Monte Carlo methods to compute a value of t equal to the 95% point (say) for the test. For example, consider pooling the samples $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ to obtain a new sample $\mathcal{X}$ of size $n_1 + n_2$. Using the uniform bootstrap, draw samples $\mathcal{X}_*^{(1)}$ and $\mathcal{X}_*^{(2)}$, of sizes n_1 and n_2 respectively, by sampling with replacement from $\mathcal{X}$; compute the corresponding version of $F^{(c)}(\cdot|p_{\rm unif}^{(c)})$, denoted by $F_*^{(c)}(\cdot|p_{\rm unif}^{(c)})$, for $c = 1, 2$, using the explanatory data in $\mathcal{X}_*^{(1)}$ and $\mathcal{X}_*^{(2)}$; and thence compute the bootstrap version $T_*(p_{\rm unif}^{(1)}, p_{\rm unif}^{(2)})$ of $T(p_{\rm unif}^{(1)}, p_{\rm unif}^{(2)})$. Using Monte Carlo simulation, calculate the 95% point of the distribution of $T_*(p_{\rm unif}^{(1)}, p_{\rm unif}^{(2)})$, conditional on $\mathcal{X}$, and take this as the t for implementing the algorithm suggested in the previous section.

There is no need to base these calculations entirely on the pooled sample. We might instead proceed iteratively, as follows. (a) Move the two empirical explanatory distributions closer together by a relatively small amount, less than that implied by the critical point t suggested in the previous paragraph, and leading to weights $p^{(1)}$ and $p^{(2)}$; (b) pool the resulting partially-aligned weighted empirical populations, reweighting them in the pooled population in proportion to the original sample sizes n_1 and n_2; (c) derive a new critical point t using bootstrap arguments once more, but this time with the resampling weights equal to $p^{(1)}$ and $p^{(2)}$ from (a), reweighted as suggested in (b); (d) move the weighted empirical explanatory distributions a little closer still, but not as far as suggested by the value t computed in step (c), and in the process deriving new versions of $p^{(1)}$ and $p^{(2)}$; (e) recalculate the critical point t, using the bootstrap methods from (c) but with $p^{(1)}$ and $p^{(2)}$ from (d), reweighted in the pooled population as suggested in (b); (f) move the distributions a little closer still, and so on.

Our ability to successfully implement algorithms such as this depends on whether we can actually reduce the size of $T(p^{(1)}, p^{(2)})$ to the level required by the Monte Carlo approach. In many situations we can, as further detailed in Appendix A2.

3 Implementing Weighted-Bootstrap Alignment

Assume data $\mathcal{X}^{(c)} = \{(X_i^{(c)}, Y_i^{(c)}),\ 1 \le i \le n_c\}$ are available for $c = 1, 2$, as described in section 2. Or equivalently in the form of X, Y, I, where X is an explanatory variable, Y a response and I an indicator, $I = 0, 1$, indicating from which population among two possible populations the

observation is coming from; we set $I = 0$ for population 1 and $I = 1$ for population 2. We also assume an underlying generalized regression relationship

$$E(Y|X, I) = g(\beta_0 + \beta_1 X + \beta_2 I + \beta_3 XI) = g(\eta(X, I)), \tag{4}$$

with a smooth link function g that is often nonlinear and a linear predictor $\eta(X, I) = \beta_0 + \beta_1 X + \beta_2 I + \beta_3 XI$. The link function g and the parameters are unknown. The marginal distributions of the predictor variable X, $F_1(x) = P(X \leq x|I = 0)$ and $F_2(x) = P(X \leq x|I = 1)$, are assumed to differ between the two populations. We also assume that the corresponding probability density functions (p.d.f.) f_1, f_2 exist.

Since the link function g is unknown, equation (4) cannot be used for adjustment. Instead, classic adjustment by linear regression modelling is frequently employed for the case where Y is continuous (perhaps even normal, or normal after a suitable data transformation). In this adjustment model, g is simply assumed to be the identity function, $g(x) = x$. One then fits the linear model,

$$E(Y|X, I) = \beta_0 + \beta_1 X + \beta_2 I + \beta_3 XI, \tag{5}$$

to the data. Testing for a difference between the two populations is equivalent to testing $\beta_2 = \beta_3 = 0$ in this model.

Another example is the case where a binomial regression is assumed, and the response Y is thought to be a Bernoulli variable. The canonical model for this case is logistic regression, where $g(x) = \exp(x)/(1 + \exp(x))$ in (4). An observed statistically significant odds ratio or log odds ratio for unmatched predictors may then simply be the result of differences in the distributions of explanatory variables; see Bhattacharya and Gastwirth (1999). The proposed bootstrap alignment algorithm addresses adjustment in various models with one unifying approach. The basic principle in applying the weighted-bootstrap adjustment is that virtually all estimates and other statistics that we compute can be expressed in terms of the empirical distribution function (edf) of the predictors. We then replace the edf by an alternative edf where the empirical weights $1/n_c$ have been replaced by the alternative weights $p_i^{(c)}$ that are computed in the weighted bootstrap algorithm.

We shall compare the proposed weighted-bootstrap alignment with the standard adjustment provided by applying (5). Assume we have n_1 samples for population 1, denoted by $\mathcal{X}^{(1)} = \{(X_i^{(1)}, Y_i^{(1)}), 1 \leq i \leq n_1\}$ and similarly n_2 samples from population 2, denoted by $\mathcal{X}^{(2)} =$

$\{(X_i^{(2)}, Y_i^{(2)}), 1 \le i \le n_2\}$.

Let

$$F^{(c)}\big(x\big|p^{(c)}\big) = \sum_{i=1}^{n_c} p_i^{(c)}\, I\big(X_i^{(c)} \le x\big)\,, \quad c = 1, 2,$$

be the weighted bootstrap distribution function corresponding to $\mathcal{X}^{(c)}$, as previously defined in section 2, and put

$$t = T\big(p^{(1)}, p^{(2)}\big) = \int \{F^{(1)}\big(x\big|p^{(1)}\big) - F^{(2)}\big(x\big|p^{(2)}\big)\}^2\, \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx, \tag{6}$$

by taking $w(x)$ to be the standard normal p.d.f., where $p^{(2)} = p^{(2)}_{\text{unif}}$. As mentioned in the previous section, one has several options for implementing weighted-bootstrap alignment. One may use the weighted-bootstrap algorithm to align the predictor distribution as closely as possible to various possible target distributions. Targeting one of the population distributions in this process can be understood as analogous to matching subjects to those with the target distribution.

In the following implementation, the target is chosen as the distribution of the predictor in population 2, towards which we align the predictor distribution in population 1, using distance measure $D^{(c)}$, defined in (2) with $\rho = 1$. Then we find

$$p_i^{(1)} = \exp\left\{a_1 + a_2\left(\sum_{j=1}^{n_1} p_j^{(1)}\, a_{ij}^{(11)} - n_2^{-1} \sum_{j=1}^{n_2} a_{ij}^{(12)}\right)\right\}, \quad i = 1, \ldots, n_1, \tag{7}$$

where the unknown constants a_1, a_2 are determined by the constraints

$$T\big(p^{(1)}, p^{(2)}\big) = t \quad \text{and} \quad \sum_{i=1}^{n_1} p_i^{(1)} = 1.$$

The latter can be guaranteed by a normalization step. The $a_{ij}^{(k\ell)}$ appearing above are defined in section 2, and in particular

$$a_{ij}^{(11)} = 1 - \Phi(X_i^{(1)} \vee X_j^{(1)}), \quad 1 \le i, j \le n_1,$$

$$a_{ij}^{(12)} = 1 - \Phi(X_i^{(1)} \vee X_j^{(2)}), \quad 1 \le i \le n_1, 1 \le j \le n_2,$$

where Φ is the Gaussian c.d.f..

We choose t by treating $T\big(p^{(1)}, p^{(2)}\big)$ as a goodness of fit statistic, and use uniform bootstrap with replacement on the pooled $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ to draw samples $\mathcal{X}_*^{(1)}$ and $\mathcal{X}_*^{(2)}$ of size n_1 and n_2 to obtain the bootstrap version $T_*\big(p^{(1)}_{\text{unif}}, p^{(2)}_{\text{unif}}\big)$ of $T\big(p^{(1)}_{\text{unif}}, p^{(2)}_{\text{unif}}\big)$. We then compute a value of t equal to the 95% point for the test through Monte Carlo simulations, i.e., $P_*(T_* > t) = 0.95$

for the bootstrap probability P_*. Generally, the choice of t should be as small as feasible. The considerations in Appendix A2 demonstrate the important fact that achieving small values is realistic. The choice of the 95% point is practically feasible, yet does not lead to unrealistically small choices of t, as one might get when aiming at larger rejection levels.

Once t has been selected, we solve the nonlinear equations

$$\begin{cases} p_i^{(1)} = \exp\left\{a_1 + a_2 \left(\sum_{j=1}^{n_1} p_j^{(1)} a_{ij}^{(11)} - n_2^{-1} \sum_{j=1}^{n_2} a_{ij}^{(12)} \right)\right\} \\ T\big(p^{(1)}, p^{(2)}_{\text{unif}}\big) = t \\ \sum_{i=1}^{n_1} p_i^{(1)} = 1 \end{cases}$$

for $p^{(1)}$ and a_1, a_2 numerically. The resulting $p^{(1)} = \{p_1^{(1)}, \ldots, p_{n_1}^{(1)}\}$ then provide the sought adjustment of the distribution F_1 towards the target distribution F_2. The adjusted distribution is given by

$$F^{(1)}(x|p^{(1)}) = \sum_{i=1}^{n_c} p_i^{(1)} I\{X_i^{(1)} \le x\}.$$

To obtain parameter estimates under this weighted-bootstrap adjusted distribution within the classic linear adjustment model (5), we use weighted least squares to estimate the parameters, with the components of $p^{(1)}$ serving as case weights for observations from population 1, and the components of $p^{(2)} = p^{(2)}_{\text{unif}}$ serving as case weights for observations from population 2. For weighted linear least squares, we obtain parameter estimates $\hat{\beta} = (X'VX)^{-1}(X'VY)$, where V is a $(n_1 + n_2) \times (n_1 + n_2)$ diagonal weight matrix for which the diagonal consists of the elements $p^{(c)}$, arranged sequentially. To use weighted-bootstrap adjustment in the framework of a binomial regression model, the $p^{(c)}$ are similarly used as case weights in the weight matrix in the weighted iterated least squares steps that correspond to the scoring method to solve the score equation of the generalized linear model; see Nelder and Wedderburn (1972) for details.

4 Simulation Study

We draw the responses from a simulated logistic regression model with a quadratic predictor component. Responses were generated as independent Bernoulli variables with probabilities $P = E(Y|X, I) = g(\eta)$ (for the definition of η, see (4)). Here the link function is the *expit* function $g(x) = \text{expit}(x) = \exp(x)/(1 + \exp(x))$. The linear predictor is $\eta = \beta_0 + \beta_1 X + \beta_2 X^2$, so there is no influence of the population indicator I. However, the distributions of X in the

two populations are assumed to differ, with choices of F_1 and F_2 both normal, with means $-1/3$ and $1/3$ respectively, and common variance $1/3$. We chose sample sizes $n_1 = n_2 = 100$, and the parameter values $\beta_0 = -0.5, \beta_1 = 5$ and $\beta_2 = 10$. The bootstrap target value for t (see eq. (6)) was between 0.0053 and 0.0090, each estimated from 1000 bootstrap samples. Simulation results are based on $M = 1000$ Monte Carlo runs.

We applied the proposed weighted-bootstrap adjustment algorithm to the covariate X, aligning distribution F_1 towards distribution F_2, as described in the previous section. The results of this alignment are shown in Figure **??** for one typical Monte Carlo sample. The top panel demonstrates the adjustment in terms of aligning empirical distribution functions, and the bottom panel in terms of aligning density functions. To visualize the densities, they were smoothed with a kernel estimator. We find that the adjustment works well for this sample, as $\hat{F}_2$ and the weighted-bootstrap version $\hat{F}_{1\,\mathrm{adj}}$ of $\hat{F}_1$ are seen to be in reasonably close agreement. A consequence is that group comparisons under the almost identical covariate distributions F_2 and $F_{1\,\mathrm{adj}}$ will not be subject to distorting effects that might be caused by differences in covariate distributions, while such an effect may arise when comparisons are made under the clearly distinct distributions F_1 and F_2.

To compare the proposed bootstrap procedure with standard adjustments, we fitted four models which contain the population indicator I and the covariate X as predictor:
Model I (correctly specified): $E(Y|X,I) = \mathrm{expit}(\beta_{10} + \gamma_1 I + \beta_{11} X + \beta_{12} X^2)$ (where in truth $\gamma_1 = 0$);
Model II (misspecified adjustment): $E(Y|X,I) = \mathrm{expit}(\beta_{20} + \gamma_2 I + \beta_{21} X)$;
Model III (misspecified adjustment): $E(Y|I) = \mathrm{expit}(\beta_{30} + \gamma_3 I)$;
Model IV (bootstrap adjustment): $E(Y_{adj}|I) = \mathrm{expit}(\beta_{40} + \gamma_4 I)$.
For fitting Model IV, we use the $p_i^{(1)}$ (equation (7)) as case weights in the iterative weighted least squares algorithm. In practice, the weights can be implemented as case dispersion parameter or case weights when using SAS *PROC GENMOD*.

We compare these four models in terms of estimated log odds ratios between populations 2 and 1. The true log odds ratio is 0 as there is in fact no effect of group membership on the response. The simulation results for mean and mean-squared error (MSE), as well as the empirical coverage rate of maximum likelihood based asymptotic 95% confidence intervals, are given in Table 1.

Table 1: Comparison of log odds ratios for true model and three adjustment models in simulation of logistic regression model, based on 1000 Monte Carlo runs

Model	Mean	MSE	CI covers 1 (%)
I (true)	0.0166	0.1802	93.8%
II (misspecified)	0.1552	0.1686	92.5%
III (misspecified)	0.5924	0.4530	53.1%
IV (bootstrap adjusted)	-0.0259	0.1105	93.2%

We find that the weighted-bootstrap adjustment in Model IV leads to reasonably good estimates. For estimating the log-odds ratio under the null hypothesis, it performs better in this specific example than the maximum likelihood estimator in terms of MSE, but less well in terms of the coverage of the confidence interval for the odds ratio. As this is an observation from a simulation study with limited scope, this finding merely illustrates that the proposed method works remarkably well in some circumstances, and should not be interpreted too broadly.

Further discussing inference derived from 95% confidence intervals, we find that the bootstrap-adjusted model rejects the null hypothesis that the true log odds ratio is 1 in 6.8% of all Monte Carlo runs. This result compares well with the result obtained when fitting the true underlying model, where the null hypothesis is rejected in 6.2% of all runs. Adjustments obtained through the misspecified models work considerably less well, as seen from the values of the mean squared errors, particularly from the size of the biases. In particular, misspecified Model III is not at all useful for testing and inference as seen from the low empirical coverage of the confidence interval.

The weighted-bootstrap adjustment clearly improves upon the wrongly estimated log odds ratios that are observed for the misspecified models. It does not require to specify any particular model for the adjustment to be effective. Its success is the more remarkable as this algorithm is universally employable for covariate adjustment and is not particularly tuned towards this application.

5 Analyzing Fecundity of Historical Cohorts of French-Canadian Women

We apply the proposed weighted-bootstrap alignment algorithm to data on fertility and mortality of a historical (17th to 18th century) cohort of French-Canadian women. The cohort included 7246 women, of whom 915 were classified as immigrants and 6331 as native-born (non-immigrant). For this analysis, we only included the women who had at least one child, resulting in data on 748 immigrants and 5547 non-immigrants. The relationship between the number of children a woman bears and age-at-death is of interest for theories of cost of reproduction and the evolution of lifespan and menopause (Westendorp and Kirkwood, 1998; Doblhammer and Oeppen, 2003). These data have been analyzed previously (Nault et al. 1990; Le Bourg et al. 1993; Müller et al. 2002). One particular question of interest is whether the relationships are different for immigrants and native-born women. Immigration status is likely associated with a specific age-at-death distribution for these women. Therefore, adjustment for age when making the comparison of the number of children at death is expected to be of relevance.

The upper panel of Figure 2 shows density estimates for age-at-death for both immigrant and native-born women, confirming that there are differences between these groups. For native-born women, there is a "shoulder" in the age-at-death density between about 35 and 50 years, which is absent from the density for immigrant women. One guess would be that relatively more native-born women gave birth at age below 40, which at that time would have increased risk of early death (due to the frequent complications associated with child birth in the 17/18th century). This effect could give rise to the "shoulder" in the density of age-at-death for native-born women.

The relationship of the number of children and age-at-death is displayed in the lower panel of Figure 2. Interestingly, almost at each age-at-death, native-born women had more children than immigrant women, especially for the women in the later child-bearing ages after 35. During that period, the average number of children for native-born women kept steadily increasing, while that for immigrant women started to flatten.

We would expect that if we were to adjust the distribution of age-at-death of immigrant women toward that of native-born women, the resulting average number of children for immigrant women would be adjusted further downwards. This was confirmed by empirically integrating each of the two curves of average number of children by age-at-death over the respective

age-at-death density: If we integrate the observed cumulative number of children for the immigrant women over the density of age-at-death for native-born women we obtain an even smaller average number of children per immigrant woman, as compared to the observed average. The observed average corresponds to integrating over the density of age-at-death for immigrant women themselves. Adjustment for the disparity in age-at-death will thus have the effect of increasing the differences in the observed number of children between these two populations.

Table 2: Poisson regression of number of children born per woman in response to immigrant status as predictor (0= native-born, 1=immigrant). The estimate is for the parameter of the immigrant status indicator.

Model	Parameter	Estimate	Standard Error	p-value
I (unadjusted)	immigrant status	-0.0675	0.0145	3.26×10^{-6}
II (bootstrap adjusted)	immigrant status	-0.0929	0.0070	7.24×10^{-40}

We use the proposed bootstrap method to adjust the age-at-death distribution of immigrant women towards the age-at-death distribution of native-born women. The overall model is chosen as Poisson regression with canonical (log) link function, the number of children as outcome variable, and immigration status as predictor. Table 2 provides estimates and standard errors for the immigration status indicator variable. We first fit a Poisson regression model with only one predictor, the indicator for immigrant status (coded as 1 for immigrant women and as 0 for native-born women), and ignoring the differences in the age-at-death distribution (Model I). The difference in the number of children between the two populations is found to be significant, with a p-value in the order of 10^{-6}.

Applying the weighted-bootstrap alignment to the covariate age-at-death within the same model (Model II) leads to a much larger difference. In this approach we find an increase in the estimated difference in the number of children between the two groups, in accordance with the above heuristic consideration. The parameter estimate for immigrant status decreases from -0.0675 to -0.0929 and the p-value after adjustment is of order of 10^{-40} (see Table 2). This would translate in a reduction in the number of children of immigrant women as compared to native-born women by 6.5% according to Model I and by 8.9% according to the bootstrap-adjusted Model II.

This finding and the noticeable absence of later births for the immigrant women suggest clear differences in the fertility of native-born and immigrant women. One can only speculate about the cause for this difference. A contributing factor may have been the likely lower chance of the presence of a grandmother for the children of immigrant women, who might therefore have been available as child care givers to a larger extent for the native-born as compared to the immigrant women. This would be in accordance with the "grandmother hypothesis" of reproduction (Lahdenperä et al., 2004), which attempts to provide an evolution-based explanation for the presence of menopause in women. According to the hypothesis, menopause opens up resources from older women who do not need to care anymore for their own children and instead are available to care for the children of their own (and, in particular, daughter's) off-spring. Our analysis suggests that this hypothesis is consistent with the observed differences in child-bearing between immigrant and native-born women.

6 Discussion

The proposed weighted-bootstrap alignment procedure for covariate adjustment requires that the supports of the observed marginal distributions of the covariates for the two groups overlap. In some cirucmstances, this may require to restrict the analysis to ranges of the covariate where this is actually the case. As omitting data from non-overlapping regions will reduce efficiency, a possible strategy is to sample more data, assuming the underlying true marginal distributions are overlapping which is usually not a restrictive assumption. If the sample size from one population is restricted, for example if this sample consists of subjects with a rare characteristic, but it is easy to increase the sample size for a comparison population, then it may be a feasible strategy to obtain a very large sample from the second population and to retain those subjects whose covariate range is similar to that of the restricted population.

Weighted-bootstrap alignment has the advantage that it avoids the need to find exact matches as required for case-control studies. This is a consequence of the possibility to reduce the value of $T\big(p^{(1)}, p^{(2)}\big)$ without further restrictions except the overlap of the marginal distributions. As a consequence, the proposed method can be used for adjustment in situations where exact or close matches cannot be found. It can be implemented to mimic matching by choosing the target distribution for the alignment to be the distribution of one of the underlying populations.

The weighted bootstrap alignment provides an adjustment method that is quite general and can be adapted to any situation where estimators or tests can be represented as functionals of the empirical distribution of the data. The method does not require specification of a particular statistical model, in contrast to classical model-specific adjustment methods. This model-free feature is particularly useful when a suitable model is not known a priori. For example, the exploratory phase of choosing a suitable model for the data can then be conducted after the data have been aligned, which reduces the complexity of this task. In situations where one deals with reasonably large samples, it would be possible to set a fraction of the data aside for alignment, followed by model selection. The selected model would then be applied to the remainder of the data. Further research about the behavior of the method under alternative hypotheses and for a wider range of situations will be of interest.

Acknowledgments

We wish to thank two referees for very helpful comments that led to an improved version. In particular, the discussion section reflects the suggestions of one referee. This work was supported in part by National Science Foundation grants DMS03-54448 and DMS05-05537.

Appendix

A1. Solution of problem (3)

We begin by noting that for $c = 1$,

$$\tfrac{1}{2}\,\frac{\partial T}{\partial p_i^{(1)}} = \sum_{j=1}^{n_1} p_j^{(1)}\, a_{ij}^{(11)} - \sum_{j=1}^{n_2} p_j^{(2)}\, a_{ij}^{(12)}\,,$$

where

$$a_{ij}^{(k\ell)} = W(\infty) - W(X_i^{(k)} \vee X_j^{(\ell)}), \quad W(x) = \int_{u<x} w(u)\, du.$$

The derivative of $D^{(c)}(p^{(c)})$ with respect to $p_i^{(c)}$ is proportional to $-(p_i^{(c)})^{\rho-1}$ if in (2) we take $0 \le \rho < 1$, and to $\log p_i^{(c)}$ plus a constant if we take $\rho = 1$. In consequence, if we solve the constrained optimization problem using Lagrange multipliers we find, when $F^{(2)}(\cdot|p^{(2)})$ is kept fixed at $F^{(2)}(\cdot|p^{(2)}_{\text{unif}})$ and $F^{(1)}(\cdot|p^{(1)})$ is moved towards it, that $p_i^{(1)} = \xi_\rho\{\eta_i(a_1, a_2)\}$, where

$$\eta_i(a_1, a_2) = a_1 + a_2 \left(\sum_{j=1}^{n_1} p_j^{(1)}\, a_{ij}^{(11)} - n_2^{-1} \sum_{j=1}^{n_2} a_{ij}^{(12)} \right), \tag{8}$$

$\xi_\rho(u) = u^{-1/(1-\rho)}$ for $0 \le \rho < 1$, $\xi_1(u) = e^u$, and a_1, a_2 are determined by the constraints $\sum_i p_i^{(1)} = 1$ and $T(p^{(1)}, p^{(2)}_{\rm unif}) = t$. Analogous formulas hold for the case where one would simultaneously move $F^{(1)}(\cdot|p^{(1)})$ and $F^{(2)}(\cdot|p^{(2)})$ towards each other.

A2. Size of T in (1)

Note that the value of t computed in either of the two Monte Carlo algorithms described on Section 2 will be of size n^{-1}, where $n = n_1 \wedge n_2$. (In the first algorithm this follows from the fact that the difference between distribution functions in the integrand of the definition of the bootstrap statistic $T_*(p^{(1)}_{\rm unif}, p^{(2)}_{\rm unif})$ is of size $n^{-1/2}$.) However, we claim that the infimum of $T(p^{(1)}, p^{(2)})$ is often of order $O(n^{-2})$, and so a level of n^{-1} is generally achievable.

Our claim is easiest to verify when we simultaneously move $F^{(1)}(\cdot|p^{(1)})$ and $F^{(2)}(\cdot|p^{(2)})$ towards each other. There, the least possible value of $T(p^{(1)}, p^{(2)})$ equals

$$\inf_{1 \le i_1 \le n_1,\, 1 \le i_2 \le n_2} \int \big\{ I\big(X^{(1)}_{i_1} \le x\big) - I\big(X^{(2)}_{i_2} \le x\big) \big\}^2 \, w(x)\, dx \,.$$

Using moment properties of spacings of order statistics as described for example in Chapter 21 of Shorack and Wellner (1986), we may show that, if the two explanatory data sets are sourced from distributions whose supports overlap in a region where their respective densities are bounded away from zero, and if (i_{10}, i_{20}) represents the value of (i_1, i_2) at the infimum, then $E(X^{(1)}_{i_{10}} - X^{(2)}_{i_{20}})^2 = O(n^{-2})$. (Here we assume that $\int w < \infty$.) It follows that $E\{\inf T(p^{(1)}, p^{(2)})\} = O(n^{-2})$.

The result is identical if we keep one of $p^{(1)}$ and $p^{(2)}$ fixed and adjust the other. Now, however, it is necessary to assume that the supports of the distributions from which the two explanatory data sets came are identical. We shall suppose in addition that each sampled density is bounded away from zero on its support, although weaker conditions are possible. Choose $p^{(1)} = p^{(1)}_\#$, say, to maximize $F^{(1)}(x|p^{(1)})$ at each x, subject to $F^{(1)}(x|p^{(1)}) \le F^{(2)}(x|p^{(2)}_{\rm unif})$ for all values x that are strictly less than $\min(\max_i X^{(1)}_i, \max_i X^{(2)}_i)$. This defines $p^{(1)}_\#$ uniquely, and of course, the infimum of $T(p^{(1)}, p^{(2)}_{\rm unif})$ over $p^{(1)}$ does not exceed $T(p^{(1)}_\#, p^{(2)}_{\rm unif})$. It may be proved that

$$E\big\{F^{(1)}\big(x|p^{(1)}_\#\big) - F^{(2)}\big(x|p^{(2)}_{\rm unif}\big)\big\}^2 = O\big(n^{-2}\big)$$

uniformly in x. Since the argument of the expectation vanishes when x lies outside the support of the two sampled distributions, which by assumption is bounded, then $E\{T(p^{(1)}_\#, p^{(2)}_{\rm unif})\} = O(n^{-2})$ and hence $E\{\inf T(p^{(1)}, p^{(2)}_{\rm unif})\} = O(n^{-2})$.

References

Ankney, C.D. (1992). Sex-differences in relative brain size. *Intelligence* **16**, 329-336.

Bhattacharya, P.K. and Gastwirth, J.L. (1999) Estimation of the odds-ratio in an observational study using bandwidth matching. *Journal of Nonparametric Statistics* **11**, 1-12.

Cressie, N.A.C. and Read, T.R.C. (1984) Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society Series B* **46**, 440-464.

Dette, H. and Munk, A. (1998) Validation of linear regression models. *Annals of Statistics* **26**, 778-800.

Doblhammer, G. and Oeppen, J. (2003) Reproduction and longevity among the British peerage: the effect of frailty and health selection. *Proceedings of the Royal Society B* **270**, 1541-1547.

Eubank, R.L. and Hart, J.D. (1992) Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* **20**, 1412-1425.

Eubank, R.L. and Spiegelman, C.H. (1990) Testing the goodness of fit of a linear-model via nonparametric regression techniques. *Journal of the American Statistical Association* **85**, 387-392.

Gail, M. and Simon, R. (1985) Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361-372.

Hall, P., Huber, C. and Speckman, P.L. (1997) Covariate-matched one-sided tests for the difference between functional means. *Journal of the American Statistical Association* **92**, 1074-1083.

Hauck, W.W., Anderson, S. and Marcus, S.M. (1998) Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Controlled Clinical Trials* **19**, 249-256.

Kimura, D. (1987) Are men's and women's brains really different? *Canadian Psychology* **28**, 133-147.

Lahdenperä M., Lummaa V., Helle S., Tremblay M. and Russell A.F. (2004) Fitness benefits of prolonged post-reproductive lifespan in women. *Nature* **428**, 178 - 181.

Le Bourg, E., Legare, J., Desjardins, B. and Charbonneau, H. (1993) Reproductive life of French-Canadians in the 17th-18th centuries - A search for a trade-off between early fecundity and longevity. *Experimental Gerontology* **28**, 217-232.

Müller, H.G., Chiou, J.M., Carey, J.R. and Wang, J.L. (2002) Fertility and lifespan: Late children enhance female longevity. *Journal of Gerontology A* **57**, B202-B206.

Nault, F., Desjardins, B. and Legare, J. (1990) Effects of reproductive behavior on infant mortality of French-Canadians during the seventeenth and eighteenth centuries. *Population Studies* **44**, 273-285.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society A* **135**, 370-384.

Pocock, S.J., Assmann, S.E., Enos, L.E. and Kasten, L.E. (2002) Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* **21**, 2917-2930.

Shorack, G.R., Wellner, J.A. (1986). Empirical Processes with Applications to Statistics. New York: Wiley.

Theil, H. (1967) Economics and Information Theory. Chicago, IL: Rand McNally.

Westendorp, R.G.J. and Kirkwood, T.B.L. (1998) Human longevity at the cost of reproductive success. *Nature* **396**, 743-746.

Yu, B. and Gastwirth, J.L. (2003). The 'reverse' Cornfield inequality and its use in the analysis of epidemiologic data. *Statistics in Medicine* **22**, 3383-3401.