Efficient estimation of the attributable fraction when there are monotonicity constraints and interactions

MIKHAIL TRASKIN†
Amazon.com

WEI WANG∗†
Center for Outcomes Research, Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA
wangw@email.chop.edu

THOMAS R. TEN HAVE
Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA 19104, USA

DYLAN S. SMALL
Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA

SUMMARY
The PAF for an exposure is the fraction of disease cases in a population that can be attributed to that exposure. One method of estimating the PAF involves estimating the probability of having the disease given the exposure and confounding variables. In many settings, the exposure will interact with the confounders and the confounders will interact with each other. Also, in many settings, the probability of having the disease is thought, based on subject matter knowledge, to be a monotone increasing function of the exposure and possibly of some of the confounders. We develop an efficient approach for estimating logistic regression models with interactions and monotonicity constraints, and apply this approach to estimating the population attributable fraction (PAF). Our approach produces substantially more accurate estimates of the PAF in some settings than the usual approach which uses logistic regression without monotonicity constraints.

Keywords: Attributable fraction estimation; Logistic regression; Monotone regression with interactions.

1. INTRODUCTION
The population attributable fraction (PAF) for a given exposure is the fraction of disease cases that can be attributed to that exposure. The PAF is an important measure of the public health impact of the exposure on disease burden and is useful for deciding which public health interventions to give priority to (Deubner and others, 1980; Rothman and Greenland, 1998). As an example, we consider the fraction of

†These authors contributed equally to this work.
∗To whom correspondence should be addressed.
suicidal ideation that is attributable to the exposures of hopelessness and depression in Section 6; another example is the fraction of lung cancer that is attributable to smoking.

The maximum likelihood method is a commonly used method for estimating the PAF that requires a correct model for the probability of disease given the exposure and the confounders (in addition to the assumption that all confounders of the exposure–disease relationship are measured; see Section 2). Such a method is natural when the mechanisms that bring about the disease are well understood. When there is more understanding about the exposure mechanism, Sjölander (2010) proposed using a regression model for the exposure combined with inverse probability weighting to estimate the PAF and Sjölander and Vansteelandt (2011) proposed a doubly robust approach that is consistent if either a regression model for the outcome or a regression model for the exposure is correct.

For settings in which the mechanisms that bring about the disease are well understood and the maximum likelihood method is thought to be a reasonable method of estimation, the probability of disease given the exposure and the confounders is typically modeled using logistic regression (Deubner and others, 1980; Bruzzi and others, 1985; Greenland and Drescher, 1993). A correct model for the probability of disease will often need to allow for interactions between the exposure and the confounders, and between the confounders themselves. Another feature of the probability of disease given the exposure and confounders is that the probability is often thought, based on subject matter knowledge, to be a monotone function of the exposure and possibly of some of the confounders. For example the probability of suicide ideation is thought to be an increasing function of hopelessness and depression (Wetzel, 1976) and the probability of developing lung cancer is thought to be an increasing function of cigarettes smoked per day (Morabia and Wynder, 1991).

Monotonicity constraints on the exposure have not heretofore been made use of in logistic regression models for estimating the PAF. The monotonicity constraints in a logistic regression model with first-order interactions between the exposure and the confounders and within the confounders can be expressed as linear constraints, where there are on the order of \(2^q + 1\) constraints with \(q\) being the number of confounding variables, more details are given in Section 3. Geyer (1991) discussed how to calculate the maximum likelihood of a logistic regression model with one predictor variable with linear or smooth non-linear constraints using available optimization packages. The challenge in our setting is that for even a moderate number \(q = 30\) of confounding variables, there are already over a billion constraints, and it is practically infeasible to use available optimization packages with this number of constraints. Davidov and Peddada (2011) discussed a related problem of order-restricted inference for multivariate binary data in which the order restriction constraints can be described by a set of linear inequalities whose dimension increases rapidly with the number of responses in the multivariate vector; they find a clever way to reduce the dimensionality of the constraint set. In this paper, we solve our computational problem by showing that the large number of linear constraints can be reduced to a small number of non-smooth non-linear constraints and then maximize the logistic regression likelihood subject to these non-linear constraints. In simulation studies, we show that by maximizing the logistic regression likelihood subject to the monotonicity constraints, we estimate the PAF substantially more accurately in some settings than the usual method of estimating the PAF that is based on logistic regression without monotonicity constraints.

Another approach to regression with monotonicity constraints is the isotonic regression algorithm (Barlow and others, 1972). To apply an isotonic regression algorithm, we need to introduce an order (or partial order) on sample observations. However, in the case of continuous predictors, such an ordering can be imposed with positive probability only if the response surface increases monotonically in all predictor variables. A simple illustration of this issue is given by a problem with two continuous predictors, \(x_1\) and \(x_2\), where the response surface increases monotonically in \(x_1\) only. The sample is assumed to consist of \(n\) observations where the \(\{(x_1^i, x_2^i)\}_{i=1}^n\) in the sample are iid random vectors. For this setting, we can say that point \(x\) dominates point \(x'\) only if \(x_1' < x_1\) and \(x_2' = x_2\). With probability one, there will be no pairs of points in the sample for which we can say one point dominates the other, and thus, there is probability
Efficient estimation of the attributable fraction

Let \( Y \) denote presence (1) or absence (0) of the disease, \( E \) the exposure, \( X \) measured pre-exposure covariates. Let \( Y_0 \) denote what the presence or absence would be if the exposure were to be eliminated, \( P(Y = 1) \) the probability of disease and \( P(Y_0 = 1) \) the “hypothetical probability of disease in the same population but with all exposure eliminated” (Benichou, 2001; Sjölander and Vansteelandt, 2011). The PAF for \( E \) is the proportion of disease that would be eliminated if all exposure were eliminated, i.e.

\[
\text{PAF} = \frac{P(Y = 1) - P(Y_0 = 1)}{P(Y = 1)} = \frac{P(Y = 1) - \int P(Y_0 = 1|X = x) \, dF(x)}{P(Y = 1)}.
\]

In order to identify the PAF based on the observed data, we need to make some assumption that connects the observed data to the counterfactual \( Y_0 \). We make the assumption that all confounders of the disease–exposure relationship are measured and contained in \( X \), and that consequently, ignorable treatment assignment (Rosenbaum and Rubin, 1983) holds.

**Assumption 1 (Ignorable Treatment Assignment)**

\[
P(Y_0 = 1|X = x) = P(Y_0 = 1|X = x, E = 0) = P(Y = 1|X = x, E = 0).
\]

Under Assumption 1, the PAF can be written as

\[
\text{PAF} = 1 - \frac{\int P(Y = 1|E = 0, X = x) \, dF(x)}{P(Y = 1)}, \tag{2.1}
\]

which is equal to the expression (2.2) for the PAF in (Sjölander and Vansteelandt, 2011). The PAF can be equivalently written as

\[
\text{PAF} = \int \left[ 1 - \frac{P(Y = 1|E = 0, X = x)}{P(Y = 1|E = e', X = x)} \right] \, dF(e, x|Y = 1), \tag{2.2}
\]

where \( F(e, x|Y = 1) \) is the conditional distribution of \((E, X)\) in the subpopulation of people with disease (Deubner and others, 1980; Benichou and Gail, 1990; Greenland and Drescher, 1993).

From a random sample of the population of size \( n \), indexed by superscript \( i = 1, \ldots, n \), one can estimate the PAF by

\[
\hat{\text{PAF}} = \frac{1}{\sum_{i=1}^{n} I_{\{Y_i=1\}}} \sum_{\{l: Y_l=1, 1 \leq l \leq n\}} \left[ 1 - \frac{\hat{P}(Y_l = 1|E_l = 0, X_l' = x_l')}{\hat{P}(Y_l = 1|E_l = e', X_l' = x_l')} \right], \tag{2.3}
\]

where \( I_{\{Y_i=1\}} \) is an indicator function. The PAF in (2.2) is defined through an integral. The integrand, which is called the individual attributable fraction (IAF), is also often of interest. The IAF represents the chance that an individual disease case with confounders \( x \) is attributable to an exposure. A
setting in which the IAF is useful is, for example, in making clinical decisions about whether to treat a child with fever with an antimalarial treatment based on the density of malaria parasites in the child’s blood (Rougemont and others, 1991; Smith and others, 1994).

3. Method

To save notation and for generality, we do not write out explicitly the exposure variable $E$ in this section and in the supplementary material available at Biostatistics online. Instead, we just assume $E$ is one of the components of $x$. We use lower indices to index the components of the predictor vector $x$. We assume that we have a model with $p$ predictors $x_i \in \mathcal{X}_i$ and a response $y \in \{0, 1\}$. In what follows, we shall assume that $\mathcal{X}_i = [x_{il}, x_{iu}]$, where $x_{il}$ and $x_{iu}$ are the lower and upper bounds for the range of $x_i$. For notational convenience, we also add $x_0 \in \mathcal{X}_0 = \{1\}$ to the list of the predictors, so that we can write any model up to a full quadratic model in a uniform way. We shall work with the full quadratic logistic regression model, where we assume

$$P(Y = 1 | X = x) = \Lambda \left( \sum_{i=0}^{p} \sum_{j=i}^{p} \beta_{ij} x_i x_j \right),$$

where

$$\Lambda(x) = \frac{e^x}{1 + e^x}$$

is the logistic distribution function. To understand the notation just introduced, notice that a linear model (on the logistic link scale) with an intercept assumes that all coefficients $\beta_{ij}$ with index $i > 0$ are equal to 0, thus giving

$$P(Y = 1 | X = x) = \Lambda \left( \sum_{j=0}^{p} \beta_{0j} x_0 x_j \right) = \Lambda \left( \sum_{j=0}^{p} \beta_{0j} x_j \right),$$

since we assumed $x_0 = 1$. If we want to work with a full interactions model (without quadratic terms), then in (3.1) we have to set $\beta_{ij} = 0$ for $i > 0$.

Without loss of generality, we assume that $P(Y = 1 | X = x)$ is monotone in the first $1 \leq d \leq p$ predictors. Note that our model allows for interaction between any subset of variables, not just “monotone” and “non-monotone” variables. Because the function $\Lambda(\cdot)$ is monotonically increasing in its argument, the monotonicity assumption on $P(Y = 1 | \cdot)$ is equivalent to the assumption that the linear predictor

$$\eta(x) = \sum_{i=0}^{p} \sum_{j=i}^{p} \beta_{ij} x_i x_j \equiv \sum_{i \leq j} \beta_{ij} x_i x_j$$

is monotone in the corresponding predictor variables. Therefore, the monotonicity assumption (that is, $P(Y = 1 | \cdot)$ is monotonically increasing in $x_i$, $i = 1, \ldots, d$) amounts to the following constraint

$$\frac{\partial \eta(x)}{\partial x_k} = \sum_{i=0}^{k} \beta_{ik} x_i + \sum_{j=k}^{p} \beta_{kj} x_j \geq 0, \quad \forall x \in \mathcal{X}, k = 1, \ldots, d,$$

where $\mathcal{X} = \mathcal{X}_0 \times \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$. 
We estimate the parameters of the logistic regression model using the constrained maximum likelihood estimation. For a given sample \(S_n = (x^l, y^l)^n_{l=1}\), we solve the following convex optimization problem

\[
\max_{\beta} \ l(\beta | S_n),
\]

\[\text{s.t. } \forall x \in \mathcal{X} \quad \sum_{i=0}^{k} \beta_{lk} x^i + \sum_{j=k}^{p} \beta_{kj} x^j \geq 0, \quad k = 1, \ldots, d, \tag{3.2}\]

where \(l(\beta | S_n)\) is the log-likelihood function for the logistic regression model

\[
l(\beta | S_n) = \sum_{l=1}^{n} \ln \left\{ \Lambda \left( \sum_{i \leq j} \beta_{ij} x^i x^j \right)^{y^l} \left[ 1 - \Lambda \left( \sum_{i \leq j} \beta_{ij} x^i x^j \right) \right]^{1-y^l} \right\}.
\]

Notice that constraints (3.2) are variational, that is they must hold at any point \(x \in \mathcal{X}\). Next we discuss how these variational constraints can be efficiently verified.

### 3.1 Efficient constraint check

Maximum likelihood estimation for logistic regression is usually done by an iterative optimization algorithm. This means that at every iteration we need to verify that a given estimate satisfies the constraint. Because the constraints are linear, a brute-force approach could be used to verify all the \(d2^p\) resulting constraints, when we consider all the corners of the \(\mathcal{X}\) space. Clearly, for sufficiently large \(p\), this approach is not feasible. To deal with this issue we adopt the approach of Ben-Tal and others (2009), which allows us to convert a large number of linear constraints into a small set of non-linear constraints. Let

\[
x_{j0} = \frac{x_{jl} + x_{ju}}{2}
\]

and

\[
x_{jr} = \frac{x_{ju} - x_{jl}}{2}.
\]

We represent a point in \(\mathcal{X}_j\) as \(x_j = x_{j0} + \xi x_{jr}, \xi \in [-1, 1]\). Then the \(k\)th monotonicity constraint can be rewritten as

\[
\min_{-1 \leq \xi \leq 1, i \in \{1, \ldots, p\}} \left[ \sum_{i=0}^{k} \beta_{ik} (x_{j0} + \xi x_{jr}) + \sum_{j=k}^{p} \beta_{kj} (x_{j0} + \xi x_{jr}) \right] \geq 0
\]

or

\[
\sum_{i=0}^{k} \beta_{ik} x_{j0} + \sum_{j=k}^{p} \beta_{kj} x_{j0} + \min_{-1 \leq \xi \leq 1, i \in \{1, \ldots, p\}} \left[ \sum_{i=0}^{k} \beta_{ik} \xi x_{jr} + \sum_{j=k}^{p} \beta_{kj} \xi x_{jr} \right] \geq 0
\]

It is clear that

\[
\min_{-1 \leq \xi \leq 1, i \in \{1, \ldots, p\}} \left[ \sum_{i=0}^{k} \beta_{ik} \xi x_{jr} \right] = -\sum_{i=0}^{k} |\beta_{ik}| x_{jr}.
\]
Hence the $k$th variational constraint reduces to a simple inequality

$$\sum_{i=0}^{k} [\beta_{ik}x_{i0} - |\beta_{ik}|x_{ir}] + \sum_{j=k}^{p} [\beta_{kj}x_{j0} - |\beta_{kj}|x_{jr}] \geq 0. \quad (3.3)$$

This constraint is a cone centered at the origin. As an example, Figure 1 demonstrates the feasible region of the type I model considered in Section 4. This type of constraint introduces a number of algorithmic issues that are discussed in greater detail in the supplementary material available at Biostatistics online, where we solve this constrained optimization problem using the coordinate ascent algorithm, in the spirit of Friedman and others (2007) and Friedman and others (2010).

### 3.2 Bounds on the variable range

It may seem desirable to remove or loosen the assumption that $x_j \in [x_{jl}, x_{ju}]$ for a certain $x_j$ or all of the $x_j$'s. One scenario when this may seem desirable is for variables that are naturally bounded from below by 0, while remaining unbounded, at least, formally, from above.

If we do not have an upper bound on a variable, then the constrained optimization problem is often easier. For illustration, consider a simple example where the conditional probability $P(Y = 1|\cdot)$ is monotone only in predictor $x_1$. Assume $x_{jl} = 0$, $j \neq 1$, then for any $x_{ju}$ we have $x_{j0} = x_{jr}$ and $x_{ju} = 2x_{jr}$. Following (3.3), we obtain

$$\sum_{i=0}^{1} (\beta_{i1} - |\beta_{i1}|)x_{ir} + \sum_{j=1}^{p} (\beta_{1j} - |\beta_{1j}|)x_{j0} \geq 0.$$
Consider $\beta_{11}$ in the above expression. For the negative values of it, we have $x_1 (\beta_{11} - |\beta_{11}|) < 0$ and this term tends to $-\infty$ as $x_1 \to \infty$. Therefore, if we remove an upper bound on $x_1$ and keep only the lower bound of 0, then the problem reduces to the search over positive values of the associated coefficient $\beta_{11}$. The same holds for all other predictors.

However, in practice, there are usually upper bounds on a variable. For example, it is unlikely that we will ever see a human who is taller than 9 feet. For the hopelessness and depression measures that we will analyze in Section 6, there is an upper bound on the measures because they are constructed using scales that have an upper bound. For variables that do have practical bounds, it is valuable to impose these bounds. The reason is that the monotonic logistic regression model is only an approximation to the truth. If we ask for this model to hold outside the range of plausible $\mathbf{x}$, then we may unnecessarily constrict the model’s ability to approximate the true $P(Y = 1|\mathbf{x})$ for the range of plausible $\mathbf{x}$. Consequently, it is valuable to only force the model to be monotonic over plausible ranges of $\mathbf{x}$.

4. Simulations

In this section, we compare our approach utilizing the monotonicity assumption with the conventional logistic regression approach for estimating the PAF and the IAF. We compare the two approaches by the following criteria: the absolute value of the bias (|Bias|), the square root of the mean squared error (RMSE), and the mean absolute deviation (MAD) of estimating the PAF, and also the square root of the Winsorized integrated mean squared error (RWIMSE) of estimating the IAF. The reason why we apply Winsorization is to reduce the influence of the outliers on the conventional logistic estimate. In our simulation studies, we see that fitting the model with the monotonicity constraint often improves results substantially.

We simulate the covariates from independent uniform $[0, 1]$ distributions and use six logistic models of two types in the simulation. The type I model has two covariates: $x_1$ and the exposure variable $e$ with the probability of disease being assumed to be monotonic in $e$ over $[0, 1]$. The type II model has five covariates with the probability of disease being assumed to be monotonic in $x_4$ and $e$ over $[0, 1]$. Specifically, the logit link function of the models are:

$Ia : -2 - 1.5x_1 + 1.02e - 1.01x_1 e$;
$Ib : -3 - 2x_1 + 2e + 3x_1 e$;
$Ic : -2 - 3x_1 + 2.01e - 2x_1 e$;
$IIa : -0.8 - 0.4x_1 - 0.6x_2 - 0.5x_3 + 0.7x_4 + 0.8e - 0.8x_1 x_2 - 0.7x_1 x_3 - 0.2x_1 x_4 - 0.3x_1 e$
   $- 0.6x_2 x_3 - 0.1x_2 x_4 - 0.2x_2 e - 0.3x_3 x_4 - 0.2x_3 e + 0.2x_4 e$;
$IIb : -1 - 2x_1 + x_2 - 2x_3 + 2x_4 + 2e - 3x_1 x_2 - 4x_1 x_3 - x_1 x_4 + x_1 e$
   $+ x_2 x_3 + 2x_2 x_4 + x_2 e - x_3 x_4 - 0.9x_3 e + 0.5x_4 e$;
$IIc : -3 - x_1 - 3x_2 - 3x_3 + 2x_4 + 2e - 4x_1 x_2 - 5x_1 x_3 - 0.8x_1 x_4 - 0.9x_1 e$
   $- 2x_2 x_3 + 2x_2 x_4 + 2x_2 e - 0.8x_3 x_4 - x_3 e + x_4 e$.

The true PAF values of these models are, respectively: 0.2604, 0.5059, 0.8412, 0.2142, 0.5154, and 0.7820. They are calculated from (2.1) using Monte Carlo methods by drawing $5 \times 10^6$ independent samples of covariates from the uniform distribution over $[0, 1]^d$, where $d$ is the number of covariates. The unconditional probability of disease $P(Y = 1)$ is calculated as $P(Y = 1) = \int P(Y = 1|E = e, \mathbf{X} = \mathbf{x}) dF(e, \mathbf{X})$, where $F(e, \mathbf{X})$ is the joint distribution of the exposure and confounders in the population. The probability in this integrand and also the probability in $\int P(Y = 1|E = 0, \mathbf{X} = \mathbf{x}) dF(x)$ are calculated for each sample and the integrals are computed as the sample averages of the probabilities. The standard error of this Monte Carlo estimate of the PAFs are at most 0.0004.
We fit the type I models (Ia, Ib, and Ic) with the intercept term and the main and interaction effects. To fit the type II models (IIa, IIb, and IIc), we use the same model selection procedure that we use in our data analysis (see Section 6) for deciding which interactions to include. We first consider a logistic regression model (without constraints) with intercept, main covariates and all first-order interactions, and then keep those interaction terms with p-values < 0.10. We then fit the monotonic logistic regression with these same chosen interactions. We ran 1000 simulations with sample sizes \( n = 50, 100, 200 \). The simulation results are summarized in Table 1 for estimating the PAF and in Table 2 for estimating the IAF.

As shown in Table 1, our estimator always has a smaller RMSE and MAD than the conventional logistic estimator, often much smaller. Our estimator with sample size \( n = 50 \) performs about as well or better than the conventional estimator with twice the sample size in terms of both the RMSE and MAD criterion. A similar pattern is also observed for |Bias|, except for model IIa when \( n = 200 \), the |Bias| of our estimator is a little bigger. The conventional logistic regression estimator is sometimes unstable resulting in huge bias and RMSE/MAD, whereas the monotonic estimator never has huge bias or RMSE/MAD. The conventional logistic regression estimator is unstable for the type I models when \( n = 50 \) for all settings considered and is also unstable for the type II models when \( n = 50 \) for all settings considered as well as for models IIb and IIc when \( n = 100 \), and for model IIc when \( n = 200 \). Histograms of the estimated PAFs are shown in Section 4 of the supplementary material available at Biostatistics online. To reduce the influence of the outliers on the conventional logistic estimated PAF, we apply Winsorization to set the estimated PAFs below a certain percentile to be that percentile. For instance, a 1% Winsorization would set all data below the first percentile to the 1st percentile. Tables S3 and S4 in Section 4 of the supplementary material (available at Biostatistics online) show the 1% and 5% Winsorized |Bias|, RMSE, and MAD for estimating the PAF in each model. In the tables, we see that our estimator always has a smaller RMSE and MAD, and

<table>
<thead>
<tr>
<th>Model</th>
<th>PAF</th>
<th>Criterion</th>
<th>( n = 50 )</th>
<th>( n = 100 )</th>
<th>( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Logit</td>
<td>Mon</td>
<td>Logit</td>
<td>Mon</td>
</tr>
<tr>
<td>Ia</td>
<td>0.2604</td>
<td>1.5012 \times 10^{12}</td>
<td>0.1518</td>
<td>0.1336</td>
<td>0.1063</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.7472 \times 10^{13}</td>
<td>0.3737</td>
<td>0.5636</td>
<td>0.3001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5012 \times 10^{12}</td>
<td>0.3196</td>
<td>0.4123</td>
<td>0.2525</td>
</tr>
<tr>
<td>Ib</td>
<td>0.5059</td>
<td>9.007 \times 10^{3}</td>
<td>0.0074</td>
<td>0.1464</td>
<td>0.0041</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.7762 \times 10^{5}</td>
<td>0.3514</td>
<td>0.5201</td>
<td>0.2911</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.0077 \times 10^{3}</td>
<td>0.3067</td>
<td>0.3601</td>
<td>0.2448</td>
</tr>
<tr>
<td>Ic</td>
<td>0.8412</td>
<td>4.5037 \times 10^{3}</td>
<td>0.0553</td>
<td>0.0655</td>
<td>0.0342</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.4242 \times 10^{5}</td>
<td>0.2079</td>
<td>0.1988</td>
<td>0.1393</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5038 \times 10^{3}</td>
<td>0.1502</td>
<td>0.1378</td>
<td>0.1079</td>
</tr>
<tr>
<td>IIa</td>
<td>0.2142</td>
<td>1.5673 \times 10^{13}</td>
<td>0.1176</td>
<td>0.0872</td>
<td>0.0834</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.3182 \times 10^{14}</td>
<td>0.3121</td>
<td>0.3912</td>
<td>0.2408</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.5673 \times 10^{13}</td>
<td>0.2592</td>
<td>0.2851</td>
<td>0.2019</td>
</tr>
<tr>
<td>IIb</td>
<td>0.5154</td>
<td>3.1659 \times 10^{12}</td>
<td>0.0065</td>
<td>0.0655</td>
<td>0.0342</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0913 \times 10^{13}</td>
<td>0.2413</td>
<td>2.9670 \times 10^{4}</td>
<td>0.1637</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.1659 \times 10^{12}</td>
<td>0.1984</td>
<td>938.3955</td>
<td>0.1306</td>
</tr>
<tr>
<td>IIc</td>
<td>0.7820</td>
<td>2.2522 \times 10^{3}</td>
<td>0.3326</td>
<td>4.4714 \times 10^{12}</td>
<td>0.1641</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.1208 \times 10^{4}</td>
<td>0.5853</td>
<td>7.5750 \times 10^{13}</td>
<td>0.4719</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.2524 \times 10^{5}</td>
<td>0.5111</td>
<td>4.4714 \times 10^{12}</td>
<td>0.3870</td>
</tr>
</tbody>
</table>
Efficient estimation of the attributable fraction

Table 2. Comparison of the RWIMSE of estimating the IAF between the conventional logistic estimate (logit) and the monotonicity constrained estimate (mon)

<table>
<thead>
<tr>
<th>Model</th>
<th>PAF</th>
<th>Winsorization (%)</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Logit</td>
<td>Mon</td>
<td>Logit</td>
</tr>
<tr>
<td>Ia</td>
<td>0.2604</td>
<td>100</td>
<td>5.9245 × 10^7</td>
<td>0.3825</td>
<td>2.8611 × 10^6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>4.6366</td>
<td>0.3375</td>
<td>1.1671</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.7552</td>
<td>0.2374</td>
<td>0.5310</td>
</tr>
<tr>
<td>Ib</td>
<td>0.5059</td>
<td>100</td>
<td>9.0743 × 10^7</td>
<td>0.3805</td>
<td>1.8323 × 10^7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>6.5039</td>
<td>0.3522</td>
<td>1.4493</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.5764</td>
<td>0.3343</td>
<td>0.4825</td>
</tr>
<tr>
<td>Ic</td>
<td>0.8412</td>
<td>100</td>
<td>1.1780 × 10^7</td>
<td>0.2425</td>
<td>0.9369</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>0.2823</td>
<td>0.1949</td>
<td>0.1682</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.1896</td>
<td>0.1658</td>
<td>0.1292</td>
</tr>
<tr>
<td>Ila</td>
<td>0.2142</td>
<td>100</td>
<td>6.7172 × 10^7</td>
<td>0.3235</td>
<td>1.8101 × 10^6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>1.1236</td>
<td>0.2710</td>
<td>1.3459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.4176</td>
<td>0.1996</td>
<td>0.4018</td>
</tr>
<tr>
<td>IIb</td>
<td>0.5154</td>
<td>100</td>
<td>4.2835 × 10^7</td>
<td>0.2684</td>
<td>3.1919 × 10^6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>0.2423</td>
<td>0.2170</td>
<td>0.2789</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.1852</td>
<td>0.1796</td>
<td>0.1802</td>
</tr>
<tr>
<td>IIc</td>
<td>0.7820</td>
<td>100</td>
<td>7.6152 × 10^7</td>
<td>0.6437</td>
<td>6.5146 × 10^7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75</td>
<td>0.6392</td>
<td>0.6437</td>
<td>0.5991</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50</td>
<td>0.6392</td>
<td>0.6422</td>
<td>0.5913</td>
</tr>
</tbody>
</table>

generally smaller |Bias| except for model Ila. The conventional logistic estimate is still influenced by the outliers. For example, for model IIc when \(n = 200\), the RMSE is still of a large scale \(10^5\) under a 5% Winsorization.

To obtain the RWIMSE for estimating the IAF, we first calculate the integrated squared error (ISE) for the simulated samples. For each \(k, k = 1, \ldots, 1000\), the error is

\[
\int \left[ \frac{\hat{P}_k(Y = 1|E=0, X=x)}{P(Y = 1|E=0, X=x)} - \frac{P(Y = 1|E=0, X=x)}{P(Y = 1|E=e, X=x)} \right]^2 dF(e, x),
\]

where \(F\) is the joint distribution function of \((E, X)\) and \(\hat{P}_k\) is the estimate from the \(k\)th simulated sample. The integral is approximately computed as

\[
\frac{1}{5 \times 10^6} \sum_{B=1}^{5 \times 10^6} \left[ \frac{\hat{P}_k(Y = 1|E=0, X=x_{B})}{\hat{P}_k(Y = 1|E=e, X=x_{B})} - \frac{P(Y = 1|E=0, X=x_{B})}{P(Y = 1|E=e, X=x_{B})} \right]^2,
\]

where \((e, x_{B})\) is a random sample of covariates drawn from the uniform \([0, 1]\) distributions. Usual logistic regression occasionally produces very inaccurate estimates of the IAF while monotonic logistic regression never produced highly inaccurate estimates. To compare the performance of the estimators without these occasional large errors dominating, we apply Winsorization to set the ISEs above a certain percentile to be that percentile. For instance, a 75% Winsorization would set all data above the 75th percentile to the 75th percentile while a 100% Winsorization is equivalent to no Winsorization. The RWIMSE is then calculated as the square root of the sample average of these Winsorized ISEs. Table 2 contains the RWIMSE.
where 100%, 75%, and 50% Winsorization are applied. The table shows that without Winsorization the conventional logistic regression estimate usually has a very large square root mean integrated squared error (RIMSE), while our estimator has a much smaller RIMSE. In particular, our estimator is much more stable for the sample sizes $n = 50$ and $n = 100$. Although Winsorization reduces the RIMSE for both estimators, the square root Winsorized IMSE (RWIMSE) of the conventional logistic estimator is still substantially larger than that of our estimator. For example, in Table 2 for model Ib when $n = 200$, even if a 50% Winsorization is applied, the RWIMSE of the conventional logistic estimate is more than twice as large as that of our approach. We also notice that in the table, the RWIMSE of our estimate with sample size $n = 50$ or $n = 100$ is usually comparable with or smaller than that of the conventional estimate with double the sample size.

5. Inference

Parameter estimates of the logistic regression are usually assumed to have a normal distribution, even though this is true only asymptotically, and confidence intervals or statistical tests are constructed under the assumption of normality. Since in the case of monotone regression we are dealing with a constrained optimization problem, the assumption of normality is no longer true even asymptotically, unless the true parameter value is within the feasible set. For finite sample size we can expect concentration of probability mass on the boundary of the feasible set. Similar issues arise when considering the distribution of the PAF estimates.

For finding confidence intervals for monotone logistic regression, we propose an approach in which the width of the confidence interval is the same as this of logistic regression without monotonicity constraints. We expect this approach to be conservative because we expect the monotone logistic regression to have less variance than the logistic regression without monotonicity constraints. Specifically, our procedure for computing confidence intervals for parameters of the monotone logistic regression is as follows:

1. Estimate parameters using the monotone regression and obtain estimate $\hat{\beta}_{\text{mon}}$.
2. Estimate parameters using the logistic regression, this also gives SE estimates.
3. Use SEs from the previous step to compute CI of the form $\hat{\beta}_{\text{mon}} \pm t_{\alpha/2} \text{SE}$.

Simulation results for the above procedure are given in supplementary material available at Biostatistics online. In our simulations for models I and II, we observed that the actual coverage of these confidence intervals tends to be higher than the nominal CI level, suggesting that this procedure is conservative, i.e. it might be possible to develop narrower intervals that maintain the nominal CI level. Hwang and Peddada (1994) discuss confidence intervals for the means of elliptically symmetric that take advantage of order restrictions, and some ideas from this work may be useful for future research in our setting.

As for the PAF confidence intervals, analogous to the above procedure for computing CIs for the coefficients, one may use the following algorithm, based on non-parametric bootstrap:

1. Compute $\hat{\text{PAF}}_{\text{mon}}$—PAF estimate based on the monotone regression.
2. Compute $\hat{\text{PAF}}$—PAF estimate based on the usual logistic regression.
3. Run non-parametric bootstrap and for every bootstrap sample, estimate the PAF using the standard logistic regression model. Determine $q_{l,\alpha} = -\alpha/2$ and $q_{h,\alpha} = 1 - \alpha/2$ quantiles of the bootstrap distribution of PAF estimates.
4. In the spirit of bias-corrected confidence intervals (Davison and Hinkley, 1997), compute the confidence interval as $(\hat{\text{PAF}}_{\text{mon}} + \hat{\text{PAF}} - q_{h,\alpha}, \hat{\text{PAF}}_{\text{mon}} + \hat{\text{PAF}} - q_{l,\alpha})$. 


This algorithm produced approximately valid CIs in our simulations. The actual coverage of these CIs was higher than the nominal level in some but not all cases. More detailed discussion of PAF CI computation together with simulation results on the coverage is available in the supplementary material available at Biostatistics online.

6. Data Analysis

We apply our approach to estimate the attributable fraction of hopelessness and depression for suicidal ideation among elderly depressed patients. Hopelessness is a system of negative expectations concerning oneself and one’s future life (Beck and others, 1975). Though hopelessness is associated with one’s self depression, it is oriented to the future as opposed to the present state (Beck and others, 1975). Studies have shown that both hopelessness and depression are important predictors of suicide and suicide ideation in psychiatric patients, and it is thought that the probability of suicide ideation is an increasing function of hopelessness and depression (Beck, 1967; Wetzel, 1976; Beck and others, 1985, 1990). Cheung and others (2006) used logistic regression to estimate the attributable fraction of hopelessness and depression for suicidal ideation among the general population in Hong Kong.

The population we consider is patients in primary care practices. The data we use come from the PROSPECT (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) study (Bruce and others, 2004). The disease we focus on is suicidal ideation 4 months after the beginning of the study, as indicated by whether the score for suicide ideation is greater than zero. This is the measure of suicidal ideation suggested by Bruce and others (2004). We are interested in how much of the suicidal ideation is attributable to hopelessness and depression at the beginning of the study. Following Cheung and others (2006), we measure hopelessness using the Beck Hopelessness Scale (BHS) (Beck and Steer, 1988), which ranges from 0 to 19 (higher value means more hopeless). Depression is measured by the Beck Depression Inventory (BDI) score (Beck and others, 1990), which ranges from 0 to 17 (higher value means more depressed).

Other variables measured for each patient include age, gender, marital status, race, and education years. We consider the 590 patients in the study with no missing data. Figure 2 shows the suicidal ideation rates and the associated 95% confidence interval at each BDI value. These are exact confidence intervals for the proportion parameter of a binomial distribution, assuming that the number of successes at each value of the predictor variable is an independent random variable having a binomial distribution (Agresti and Coull, 1998). The plot suggests a monotone relationship between suicidal ideation and the BDI. A similar monotone pattern is also observed for suicidal ideation rate and BHS (see Figure S1 in the supplementary material available at Biostatistics online). A logistic regression model with all variables in the model without interactions shows that BHS and BDI are important predictors for suicidal ideation with estimated coefficients 0.1687 and 0.0911, and associated p-values $6.56 \times 10^{-5}$ and 0.072, respectively.

The PAF for hopelessness is the proportion of suicide ideation that would be prevented if all patients’ hopelessness was reduced to zero on the BHS scale, while keeping all other variables (including depression) constant. The PAF for depression is the proportion of suicide ideation that would be prevented if all patients’ depression was reduced to 0 on the BDI scale, while keeping all other variables (including hopelessness) fixed. The estimated PAFs for hopelessness and depression under the logistic model with no interactions are 0.6885 and 0.2905, respectively, as shown in the first column of Table 3.

We are interested in seeing how the estimated PAF will change if interactions are included into the logistic regression model, in particular the interactions involving the monotonic variables BHS and BDI. We first consider a logistic regression model with all first-order interactions and then keep those interaction terms with p-values $< 0.10$. The p-value of interactions between BDI and education years, BDI and gender, and BHS and gender are small, 0.0667, 0.0681, and 0.0033, respectively, and the p-value of other interactions are large with the smallest being 0.1545. Hence, we consider the logistic regression model,
Fig. 2. Plot of the averaged suicidal ideation rate and the associated 95% confidence interval at each BDI value.

Table 3. Comparison of the estimated PAF and the associated 95% bootstrap confidence interval among the approaches: logistic regression without interactions, logistic regression with interactions (logit) and our approach controlling for interactions and using monotonicity assumption (mon)

<table>
<thead>
<tr>
<th></th>
<th>Logit (no interaction)</th>
<th>Logit (logit)</th>
<th>Mon (mon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BHS</td>
<td>0.69 (0.53, 0.97)</td>
<td>0.54 (0.25, 1)</td>
<td>0.60 (0.31, 1)</td>
</tr>
<tr>
<td>BDI</td>
<td>0.29 (0.03, 0.63)</td>
<td>0.21 (0, 0.62)</td>
<td>0.39 (0.11, 0.8)</td>
</tr>
</tbody>
</table>

treating the BHS and BDI relationships with ideation as monotonic in estimating the attributable fraction. We compare the differences in the PAF estimates between the conventional approach and our approach that treats the BHS and BDI relationships with suicide ideation as monotonic.

Table 3 shows the difference in the PAF estimates among the three approaches: logistic regression without interactions, logistic regression with interactions and our approach controlling for interactions and using the monotonicity constraints. The associated 95% non-parametric bootstrap confidence intervals for each estimate are also reported, where the intervals are constructed with 1999 bootstrap samples as described in Section 5. Since the PAF is between 0 and 1, truncation is applied if any lower bound is < 0 or any upper bound exceeds 1.

In Table 3, we see that the two methods that control for interactions give smaller estimates of the PAF than the method that does not control for interactions. By looking at equation (2.2) that defines the PAF,
we see that it is important to use a correct model for $P(Y = 1|E, X)$ in estimating the PAF. Consequently, because there are significant interactions, it is important to control for these interactions in the model to obtain good estimates of the PAF.

7. Discussion

In this paper, we study how to estimate the attributable fraction when, based on subject matter knowledge, it is thought that the exposure variable and possibly some confounding variables have a monotone effect on the probability of disease, and there are also potentially interactions between the exposure variable and the confounding variables or interactions among the confounding variables. We develop an approach to estimating the probability of having a disease under the monotonicity constraint for logistic models. We have shown in simulation studies in Section 4 that our approach can provide substantially more accurate estimates of the PAF than the usual approach that is based on logistic regression without taking advantage of the monotonicity constraints. Our approach can be generalized to fit any other generalized linear model with interactions and monotonicity constraints by expressing the optimization problem as the convex optimization problem (3.2) and solving it by using the efficient constraint check of Section 3.1 and the coordinate ascent method in the supplementary material available at Biostatistics online. This would be useful for estimating the fraction of a count outcome that are attributable to an exposure, e.g. the fraction of seizures that are due to helminthic infections (Montano and others, 2005).

When there are a large number of variables in the model for the probability of disease given exposure, some sort of model selection or shrinkage estimation may be desired. For this setting, our approach can be used together with $l_1$ or $l_2$ regularization (Zou and Hastie, 2005). Regularization is achieved through introduction of a penalty function of the form

$$P(\beta, \alpha, \gamma) = \alpha \sum_{i,j \geq 0} \gamma_{ij}|\beta_{ij}| + \frac{1 - \alpha}{2} \sum_{i,j \geq 0} \gamma_{ij}\beta_{ij}^2,$$

where $\alpha \in [0, 1]$ is the desired trade-off between $l_1$ and $l_2$ regularization and $\gamma_{ij} \geq 0$, $i \geq 0$, $j \geq 1$ are individual parameter weights that allow for different weighting of individual parameters (they are usually set all equal). As a result, instead of maximizing the usual logistic log-likelihood function $l(\beta)$, we maximize a penalized likelihood function $l(\beta) - P(\beta, \alpha, \gamma)$ subject to the same inequality constraints as in Section 3. This is still a convex optimization problem that can be solved with the same coordinate ascent algorithm with only minor algorithmic complications resulting from non-differentiability of the penalty function for $\alpha > 0$ (that is for penalties that include $l_1$ regularization); see the supplementary material (available at Biostatistics online) for further discussion.

Instead of using the efficient estimate of $P(Y = 1|E = 0, X)$ that we have developed to directly estimate the PAF by (2.3), our estimate can be combined with an estimate of $P(E = 0|X)$ to estimate the PAF doubly robustly, using the double robust estimator developed by Sjölander and Vansteelandt (2011). Specifically, we can plug our estimate of $P(Y = 1|E = 0, X)$ into the estimator $\hat{P}_0(\hat{\alpha}, \hat{\beta})$ at the end of Section 3.2 of Sjölander and Vansteelandt (2011). It would be of interest to investigate the performance of this doubly robust estimator in future work.

We have developed a frequentist approach to incorporating monotonicity constraints in estimating the attributable fraction in this paper. Our approach enables prior information about monotonicity constraints to be incorporated into the analysis. It would be of interest to consider a Bayesian approach to incorporating the prior information about monotonicity constraints. For instance, if a covariate $x$ was believed to have a monotone increasing effect on the probability, the prior of the associated regression parameter $\beta$ would be such that $P(\beta < 0) = 0$. If one were not absolutely sure about the validity of the monotonicity constraint, the Bayesian approach would allow one to hedge one’s bets by using a prior that places some probability on $\beta > 0$. 
Besides their use in estimating attributable fractions, monotonicity constraints are often used in other contexts such as testing for sufficient cause interactions on the basis of contrasts between regression coefficients (VanderWeele, 2009). It would be of future research interest to consider whether our approach to efficient estimation of logistic regression when there are monotonicity constraints and interactions would be useful for these contexts.

We have seen in the simulation study that incorporating monotonicity constraints into PAF estimation can substantially increase efficiency. However, at the present time, we are only able to produce a conservative confidence interval for the PAF using the constrained estimator that is of the same width as the confidence interval based on the estimator without the monotonicity constraints. This has some analogues with $l_1$ regularized regression where the estimates have better mean squared error than usual regression, but it is challenging to provide valid confidence intervals that are not conservative (see Tibshirani, 1996; Osborne and others, 2000; Chatterjee and Lahiri, 2010). It will be valuable future research to produce shorter, less conservative confidence intervals based on the monotonicity constrained estimator of the PAF.

Implementation of the developed approach is programmed in the statistical software R. R code and instructions for using the developed approach are provided in the supplementary material available at Biostatistics online.

**Supplementary Material**

Supplementary material is available at http://biostatistics.oxfordjournals.org.

**Acknowledgments**

The authors would like to thank the referees and associate editors for very helpful comments that have improved our paper. *Conflict of Interest*: None declared.

**Funding**

This work was supported in part by NSF grant SES-0961971.

**References**


Efficient estimation of the attributable fraction


[Received January 15, 2012; revised May 9, 2012; accepted for publication May 14, 2012]