11-2012

A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-Continuous Exposure

Wei Wang
University of Pennsylvania

Dylan S. Small
University of Pennsylvania

Recommended Citation

Wang, W., & Small, D. S. (2012). A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-Continuous Exposure. *The International Journal of Biostatistics, 8* (1), http://dx.doi.org/10.1515/1557-4679.1389

A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-Continuous Exposure

Abstract

The attributable fraction of a disease due to an exposure is the fraction of disease cases in a population that can be attributed to that exposure. We consider the attributable fraction for a semi-continuous exposure, that is an exposure for which a clump of people have zero exposure and the rest of the people have a continuously distributed positive exposure. Estimation of the attributable fraction involves estimating the conditional probability of having the disease given the exposure. Three main approaches to estimating the probability function are (1) a classical method based on sample averages; (2) parametric regression methods such as logistic regression models and power models; and (3) nonparametric regression methods including local linear smoothing and isotonic regression. We compare performance of these methods in estimating the attributable fraction for a semi-continuous exposure in a simulation study and in an example.

Keywords

attributable risk, monotonicity, nonparametric regression, power model

Disciplines

Biostatistics

The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 32

A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-continuous Exposure

Wei Wang, *Center for Outcomes Research, Children's Hospital of Philadelphia*
Dylan Small, *Department of Statistics, the Wharton School, University of Pennsylvania*

Recommended Citation:
Wang, Wei and Small, Dylan (2012) "A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-continuous Exposure," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 32.
DOI: 10.1515/1557-4679.1389

A Comparative Study of Parametric and Nonparametric Estimates of the Attributable Fraction for a Semi-continuous Exposure

Wei Wang and Dylan Small

Abstract

The attributable fraction of a disease due to an exposure is the fraction of disease cases in a population that can be attributed to that exposure. We consider the attributable fraction for a semi-continuous exposure, that is an exposure for which a clump of people have zero exposure and the rest of the people have a continuously distributed positive exposure. Estimation of the attributable fraction involves estimating the conditional probability of having the disease given the exposure. Three main approaches to estimating the probability function are (1) a classical method based on sample averages; (2) parametric regression methods such as logistic regression models and power models; and (3) nonparametric regression methods including local linear smoothing and isotonic regression. We compare performance of these methods in estimating the attributable fraction for a semi-continuous exposure in a simulation study and in an example.

KEYWORDS: attributable risk, monotonicity, nonparametric regression, power model

Author Notes: The authors would like to thank the referees for very helpful comments that have improved our paper.

1 Introduction

The attributable fraction of a disease due to an exposure is the proportion of disease cases which would be eliminated if everybody's exposure was set to zero. The AF is an important measure of the public health impact of the exposure on disease burden (Deubner, Wilkinson, Helms, Tyroler, and Hames, 1980, Rothman and Greenland, 1998, Benichou, 2001). In this paper, we compare the performance of different methods of estimating the AF when the exposure is *semi-continuous*. An exposure is semi-continuous when a clump of people have zero exposures and the rest of the people have continuously distributed positive exposures. For example, malaria parasites in children is a semi-continuous exposure; we will estimate the AF for fever due to malaria parasites in a malaria endemic area in Section 3.

If there are no confounders of the exposure-disease relationship (or if we are considering the AF within a stratum of confounders), then the AF is the following (Benichou, 2005, Chen, 2008):

$$AF = \frac{P(\text{Disease}) - P(\text{Disease}|\text{Exposure} = 0)}{P(\text{Disease})}. \tag{1}$$

The classical estimate of AF is given by plugging sample proportions of $P(\text{Disease})$ and $P(\text{Disease}|\text{Exposure=0})$ into (1) (Benichou, 2005, Chen, 2008). Smith, Schellenberg, and Hayes (1994) pointed out that when the proportion of subjects with zero exposure is small, it will be hard to estimate $P(\text{Disease}|\text{Exposure} = 0)$ and the classical estimate of AF will have wide confidence limits. For the exposure of malaria parasites in a malaria endemic area, parasite prevalence in young children may exceed 80 percent so that the proportion of children with zero exposure is small. When the proportion of people with zero exposure is small, one way to improve precision is to assume that low exposure is equivalent to zero exposure and estimate $P(\text{Disease}|\text{Exposure} = 0)$ by the sample proportion of people with disease with zero or low exposure. However, the resulting estimate may strongly rely on the definition of low exposure; see the example provided by Smith et al. (1994) for estimating the attributable fraction of fever due to malaria parasites. If the disease probability is increasing as the exposure increases even at low exposures, the AF will be underestimated by grouping low exposure with zero exposure.

To borrow strength in estimating $P(\text{Disease}|\text{Exposure} = 0)$ without assuming that $P(\text{Disease}|\text{Exposure=0}) = P(\text{Disease}|\text{Exposure is 0 or low})$, regression models for $P(\text{Disease}|\text{Exposure})$ can be used to estimate $P(\text{Disease}|\text{Exposure=0})$. Logistic regression is a frequently applied regression method to estimate the AF (Bruzzi, Green, Byar, Brinton, and Schairer, 1985, Greenland and Drescher, 1993, Drescher and Schill, 1991, Smith et al., 1994). In practice, the mechanism of the exposure on the disease is usually unknown and can be very complicated, and so the true

model is not necessarily in a logistic form. Power models extend logistic regression by considering transformations of the exposure variable such as logarithm and fractional polynomials (Royston, Ambler, and Sauerbrei, 1999, Royston, Sauerbrei, and Becher, 2010, Smith et al., 1994).

An alternative to parametric regression for estimating the AF is nonparametric modeling of $P(\text{Disease}\,|\,\text{Exposure})$, particularly when no prior knowledge about the shape of the true curve is available. To our knowledge, nonparametric regression has not been used to estimate the AF previously. However, nonparametric methods have often been applied to analyze medical or health-related data, for example, to estimate the effective dose level of dose-response curves (Bhattacharya and Kong, 2007, Dette, Neumeyer, and Pilz, 2005, Müller and Schmitt, 1988, Park and Park, 2006), to estimate the relative risk functions in case-control studies (Zhao, Kristal, and White, 1996) and to study the relationship between biomarker and disease risk (Ghosh, 2007). In nonparametric regression, it improves efficiency to incorporate any known shape constraints on the regression function.

Under certain circumstances, it is often thought that $P(\text{Disease}\,|\,\text{Exposure})$ is a monotone increasing function of the exposure level. For instance, the probability of suicide ideation is thought to be an increasing function of the level of hopelessness and depression (Wetzel, 1976). The probability of developing lung cancer is thought to be an increasing function of cigarettes smoked per day (Morabia and Wynder, 1991). The probability of death is assumed to be a monotone function of the severity of burn injury (Wolfe, Roi, and Margosches, 1981). A dose response curve is often assumed to be non-decreasing (Bhattacharya and Kong, 2007, Dette et al., 2005, Müller and Schmitt, 1988, Park and Park, 2006). Higher levels of a biomarker are often assumed to be associated with monotone increasing disease risk (Ghosh, 2007). For estimating the AF in a logistic regression framework accounting for interactions, incorporating the monotonicity constraint on the exposure has been found improving the accuracy substantially (Traskin*, Wang*, Have, and Small, first published online June 21, 2012).

We consider various nonparametric estimators that incorporate the monotonicity constraint in this article. The purpose of this paper is to compare the performance of various parametric and nonparametric estimates of the AF for semi-continuous exposures via simulation studies and in an example. The rest of the article is organized as follows. Section 2 introduces some notation and lists the competing estimators. In Section 3, we apply these estimators to estimate the proportion of fever cases attributable to malaria parasites. Simulations studies are presented in Section 4. A summary is given in Section 5.

2 Estimators

Let Y denote presence (1) or absence (0) of the disease and let X denote the exposure. The conditional probability of disease at exposure level x is $P(Y=1|X=x)$. We assume that there are no confounders of the disease-exposure relationship or that we are considering the AF within a stratum of confounders. Benichou (2001) discusses ways to estimate the AF for the entire population based on estimates of the AF within each stratum of confounders; we provide further discussion in Section 5.

Under the assumption of no confounders, the AF is

$$\frac{P(Y=1)-P(Y=1|X=0)}{P(Y=1)} = P(X>0|\text{Y=1})\left(1-\frac{1}{R}\right), \tag{2}$$

where R is the relative risk of disease with exposure greater than zero compared to zero exposure. The classical estimate of the AF is to plug the sample proportions $\hat{P}(Y=1)$ and $\hat{P}(Y=1|X=0)$ into the left hand side of (2), or equivalently the sample proportion $\hat{P}(X>0|Y=1)$ and $\hat{R}=\frac{\hat{P}(Y=1|X>0)}{\hat{P}(Y=1|X=0)}$ into the right hand side of (2) . We denote this classical estimate based on sample averages by S. S is a nonparametric estimate of the AF.

To estimate the AF using regression methods, we note that the AF (2) can be written as

$$AF = \int \left[1-\frac{P(Y=1|X=0)}{P(Y=1|X=x)}\right] dF(x|Y=1), \tag{3}$$

where $F(x|Y=1)$ is the conditional distribution of the exposure in the subpopulation of people with disease (Benichou and Gail, 1990, Deubner et al., 1980, Greenland and Drescher, 1993). Based on (3), from a random sample of size N from the population, one can estimate the AF by

$$\widehat{AF} = \frac{1}{\sum_{i=1}^{N} I_{\{Y_i=1\}}} \sum_{i=1,Y_i=1}^{N} \left[1-\frac{\hat{P}(Y_i=1|X_i=0)}{\hat{P}(Y_i=1|X_i=x_i)}\right], \tag{4}$$

where $\hat{P}(Y_i=1|X_i=x_i)$, $\hat{P}(Y_i=1|X_i=0)$, $i=1,\ldots,N$, are estimates from a regression model of the conditional probability of disease at exposure levels. Using a regression model, we are not limited to categorical exposures. If the exposure is continuous, one can use it directly or categorize the exposure.

Let $p_{x_i}=P(Y_i=1|X_i=x_i)$. Assume $\hat{p}_{x_i}$ is equal to the truth p_{x_i}. Under regularity conditions, an informal justification of $\widehat{AF}$ converging to the AF on the left hand side of (2) is the following.

$$\widehat{AF} \;=\; \frac{1}{\sum_{i=1}^{N} I_{\{Y_i=1\}}} \sum_{i=1,Y_i=1}^{N} \left(1-\frac{p_0}{p_{x_i}}\right)$$

$$
\begin{aligned}
&= \frac{1}{\sum_{i=1}^{N} I_{\{Y_i=1\}}} \sum_{i=1}^{N} \left[\left(1 - \frac{p_0}{p_{x_i}} \right) I_{\{Y_i=1\}} \right] \\
&= \left(\frac{N}{\sum_{i=1}^{N} I_{\{Y_i=1\}}} \right) \frac{1}{N} \sum_{i=1}^{N} \left[\left(1 - \frac{p_0}{p_{x_i}} \right) I_{\{Y_i=1\}} \right] \\
&\to \frac{1}{\mathrm{E} I_{\{Y_i=1\}}} \mathrm{E} \left[\left(1 - \frac{p_0}{p_{x_i}} \right) I_{\{Y_i=1\}} \right] \\
&= \frac{1}{P(Y_i=1)} \left[P(Y_i=1) - \mathrm{E} \left(\frac{p_0}{p_{x_i}} I_{\{Y_i=1\}} \right) \right] \\
&= \frac{1}{P(Y_i=1)} \left[P(Y_i=1) - p_0 \right],
\end{aligned}
$$

where the last equality follows from expressing $\mathrm{E} I_{\{Y_i=1\}}$ as $\mathrm{E}(p_{x_i})$ by conditional expectation. Thus, a good estimate of the regression model may be of importance for estimating AF.

We will consider the following regression estimators of AF based on plugging into (4) the estimates of the following different regression models: logistic regression (Lg), power model (P), local linear smoothing (L), isotonic regression (I), local linear smoothing followed by isotonic regression (LI), and isotonic regression model followed by local linear smoothing (IL). The logistic regression model is linear in the exposure x, i.e., $\mathrm{logit}[P(Y=1|X=x)] = \alpha + \beta x$. The power model is $\mathrm{logit}[P(Y=1|X=x)] = \alpha + \beta (x)^{\tau}$ (Royston et al., 1999, 2010, Smith et al., 1994). We consider equally spaced grid points of τ from -10 to 10 with $\log x$ being used in place of x^0. We choose the increment for the grid as 0.1 to ease interpretation and for $\tau < 0$, we use the shifted exposure $1 + x_i$ to avoid a 0 denominator. We choose the τ that minimizes the deviance.

Local linear smoothing is a popular nonparametric regression technique. In our setting, local linear smoothing borrows information on the disease cases from a neighborhood of a given exposure to estimate the conditional probability at the given exposure. In local linear smoothing, a bandwidth parameter controls the smoothness of the fit. The asymptotically optimal bandwidth involves unknown quantities and practically it is not directly applicable. A common strategy for choosing the bandwidth is to use leave-one-out cross validation, but it is also computationally intensive. Here we apply the rule of thumb proposed by Rice (1984) (see also Dette et al. (2005), Dette and Scheder (2010), Müller and Schmitt (1988)). Let $\hat{\sigma}^2 = \sum_{i=1}^{N-1} (Y_{i+1} - Y_i)^2 / [2(N-1)]$. The bandwidth is chosen via a grid search of equally spaced points at which the following quantity is minimized $\sum_{i=1}^{N} \left(y_i - \hat{P}_h(Y_i = 1 | X_i = x_i) \right)^2 / N + 3\hat{\sigma}^2 / (2Nh)$. Following Dette et al. (2005), we use the Epanechnikov kernel. It is possible that the local linear estimator may go

beyond 1 or below 0. In such case, truncation is then applied (Aragaki and Altman, 1997, Signorini and Jones, 2004).

For obtaining a monotone estimate of the regression function $P(Y = 1|X = x)$, a classical method is isotonic regression (Barlow, Bartholomew, Bremner, and Brunk, 1972, Bhattacharya and Kong, 2007). The obtained estimate behaves like a step function, being flat in certain regions and then having jumps. Individuals in the same flat region share a common estimated probability value and thus the same relative risk.

The local linear estimator may not be monotone and the isotonic regression estimate may not be smooth. A hybrid approach is to combine local linear smoothing and isotonic regression to obtain a smooth, monotonic estimate. In this approach, two estimators can be constructed: local linear smoothing an isotonic estimate or isotonizing a local linear estimator, where the monotonicity constraint is preserved for the latter. For constructing the IL and LI estimators, we use the same grid points as the local linear estimator L. Sometimes the estimate L is monotone itself in which case we take LI as L without the additional isotonization step.

Besides the hybrid approach, other approaches have also been proposed for monotone smoothing (Dette et al., 2005, Müller and Schmitt, 1988, Park and Park, 2006). Dette et al. (2005) also show that their estimator has exactly the same first-order asymptotic properties as that of Müller and Schmitt (1988). The authors also compare their estimator with the LI estimator and found no clear ordering between the LI estimator and their estimator. Dette and Scheder (2010) conduct a detailed numerical comparison to estimate the effective dose in quantal bioassay for estimators of Dette et al. (2005), Müller and Schmitt (1988), Park and Park (2006). The authors consider repeated and non-repeated measurement designs, and find in both cases the comparison of the estimates yields a similar picture. We will only consider the LI and IL approaches to monotone smoothing henceforth.

3 Example: Attributable Fraction of Fever Due to Malaria

Malaria is an infectious disease caused by a parasite that is a major public health problem in many countries. Fever is the most characteristic clinical feature of malaria. However, fevers caused by malaria parasites often cannot be distinguished on the basis of clinical features from fevers caused by other common childhood infections such as the common cold, pneumonia, influenza, viral hepatitis or typhoid fever (Hommel, 2002, Koram and Molyneux, 2007). One aid to deciding whether a fever is caused by malaria parasites is to measure the density of malaria parasites

in the child's blood. But even if a child has fever and has a high parasite density, the fever might still be caused by another infection. Estimation of the proportion of fever attributable to malaria parasite is important for understanding the burden of the disease and changes in the burden.

We consider data from repeated cross-sectional surveys of parasitaemia and fever among 408 children up to 1 year old in a village in the Kilombero district in Tanzania. The data were described in (Kitua, Smith, Alonso, Masanja, Urassa, Menendez, Kimario, and Tanner, 1996) and analyzed by Vounatsou, Smith, and Smith (1998). For each sampled child, the child's axillary temperatures was measured and fever was defined as an axillary temperature of 37.5° C (99.5° F) or higher. Also, a finger prick blood sample of the child's blood was taken and, after being dried and stained, was examined under a light microscope for malaria parasites. The malaria parasite density per cubic milliliter (μl) was assessed by counting how many parasites were found for the first approximately 200 white blood cells counted, and then multiplying by (8000/number of white blood cells counted), under the assumption that there are 8000 white blood cells per μl.

Let Y be the response of having a fever ($Y = 1$) or not ($Y = 0$) and let X represent the parasite density. Many of the parasite densities are clustered around 0 while the rest range continuously up to the maximum $399,952$. To investigate the relationship between fever and parasite, we first summarize the data by calculating the averages of Y within intervals of X in Table 1. Due to the sparsity of high parasite densities, the intervals are constructed such that the lengths are almost doubled and each interval contains at least 30 observations. In Table 1, we see that the fever rates are increasing as the parasite density increases except for the last interval. For parasite densities less than or equal to $25,781$, there were 161 fever cases were observed among 304 children. Among the total 104 children with parasite densities greater than $25,781$, there were 103 fevers cases and a single non-fever case at $X = 138,677$.

In Figure 1, we plot the fever rates and 95% confidence intervals against the right end point of the intervals on a logarithm base 10 scale in Table 1. The confidence intervals are computed by the "Wilson" method for binomial probabilities (Agresti and Coull, 1998). The plot suggests the fever rate increases sharply for low exposures (the original scale of x) and then approach a constant around 1 for high exposures. Overall, the plot suggests a monotone pattern of the conditional probability $P(Y = 1|X = x)$.

Figure 2 shows $\hat{P}(Y = 1|X = x)$ on a logarithm base 10 scale of X from estimators I, IL, L, LI, Lg, P. For the power model estimator P, the power τ is estimated to be 0.7. Grid points of h for estimators L, IL and LI are equally spaced from 1 to the maximum of x on a grid of length 1000. Bandwidth chosen for the three estimators are respectively $23,221$, $23,622$ and $23,221$. In the plot, we see

Table 1: Distribution of malaria parasite densities and fever rates in children sampled in Kilombero District, Tanzania in 1993-1994.

Parasite density (parasites/μl)	Number of observations	Fever rate
0	116	0.457
1-3000	77	0.468
3001-7000	32	0.594
7001-15,000	40	0.650
15,001-30,000	42	0.714
30,001-60,000	35	1.000
60,001-120,000	36	1.000
120,001-399,952	30	0.967

that all the estimators suggest a similar pattern of a sharply increasing rate of the probability that approaches 1. The local linear estimator *L* gives an estimated value of 1 of the probability at 4.69, but then dips below 1 starting around 5.07. The fit of *L* is clearly affected by the outlier at $\log_{10} x = \log_{10} 138{,}677 \approx 5.14$ where we see a valley of the fit occurs. Estimator *I* gives estimated probability of 1 right after the outlier. Fitted probabilities given by estimators *Lg* and *P* are very close to each other, and so are those given by estimators *IL* and *LI*. Estimator *IL* replaces the sudden jump of estimator *I* at 5.14 by a smooth increasing step. On the other hand, estimator *LI* seems to be a compromise between *I* and *L*: filling up the valley and then increasing smoothly to 1.

Table 2 contains the estimated AF using these different methods. The estimated AFs range from 0.2691 to 0.3309. Estimator I gives the biggest $\widehat{AF}$ and estimator Lg gives the smallest estimate. Bootstrap percentile confidence intervals are also reported. These confidence intervals were formed by resampling the original data 1000 times with replacement and then looking at the lower 2.5%-th and upper 97.5%-th quantiles of the $\widehat{AF}$'s. The length of the bootstrap confidence interval is reported in the bottom row of the table. The logistic regression estimator Lg has the smallest confidence interval. However, Lg is a parametric estimator that can be biased if the logistic regression model is not true. As we shall see in Section 4, Lg performs the worst in terms of mean squared error in our simulation study. The nonparametric regression estimators L, IL and LI have somewhat smaller confidence intervals than the classical nonparametric estimator S.

Table 2: $\widehat{AF}$ and the 95% bootstrap confidence intervals for different estimators.

Estimator	S	I	Lg	P	L	IL	LI
$\widehat{AF}$	0.2939	0.3309	0.2691	0.3116	0.2732	0.2847	0.2731
Lower CI	0.1772	0.2485	0.2003	0.1928	0.1306	0.1305	0.1306
Upper CI	0.4221	0.4408	0.3474	0.4104	0.3404	0.3582	0.3404
Length of CI	0.245	0.192	0.147	0.218	0.210	0.228	0.210

4 Simulation Study

In our simulation study, we will consider the effects of the sample size, the proportion of subjects with zero exposure and the disease-exposure relationship $P(Y = 1|X = x)$. We consider sample sizes, $n = 30$, 100, 500 and 1000; the relative performance of the estimators for $n = 1000$ was very similar to that of $n = 500$ so we only report results for $n = 30, 100$ and 500. We simulate the exposure variable X as being zero with probability q and being uniform on $[0,1]$ with probability $1-q$. For the proportion of zero exposures q, we consider values 0.1, 0.3 and 0.5.

The binary response Y is simulated according to the probability models $P(Y = 1|X = x) = a + (1-a)f(x)$. For $f(x)$, we consider the seven models:

$$\begin{aligned} f_1 &= 1-\exp(-\sqrt{12x}), \\ f_2 &= \sqrt{x}, \\ f_3 &= x+\frac{\sin(2\pi x)}{2\pi}, \\ f_4 &= x, \\ f_5 &= \frac{1}{1+\exp(5-10x)}, \\ f_6 &= 0.9x^2+0.1x, \\ f_7 &= x^7. \end{aligned}$$

Figure 3 shows the shapes of these function for $a = 0$. The models include a Weibull model (f_1), a straight line (f_4) and a logit model (f_5). Models f_1 and f_5 are also used in (Dette and Scheder, 2010, Park and Park, 2006). These models can be roughly classified into two groups according to their instantaneous rates of change at $x = 0$: (1) $f'(0)$ is greater than 1 for f_1, f_2 and f_3; (2) $f'(0)$ is less than or equal to 1 for f_4, f_5, f_6 and f_7. The instantaneous rates of change as $P(Y = 1|X = x)$ approaches 1 for these models can also be roughly classified into two groups: one group with f_1, f_2 and f_5, and another group with f_3, f_4, f_6 and f_7. The parameter a controls the probability of disease at zero exposure, $P(Y = 1|X = 0)$. The models in Figure 3

have $P(Y=1|X=0)=0$ but in practice there is likely to be some probability of the disease due to causes other than the exposure so that $P(Y=1|X=0)>0$. So to allow for $P(Y=1|X=0)=a>0$, we linearly transform the probability models f in Figure 3 by the "intercept" a and a "slope" $1-a$ with $a=0.1$, 0.3, 0.5 ,0.7 respectively. For instance for function f_4 with $a=0.1$, the transformed model is $P(Y=1|X=x)=0.9x+0.1$. Note that a higher value of a also has the indirect effect of decreasing the rate of change around $x=0$ of $P(Y=1|X=x)$.

The true AF value is calculated as

$$1-\frac{P(Y=1|X=0)}{qP(Y=1|X=0)+(1-q)\int_{x>0}P(Y=1|X=x)dx}.$$

The grid points of h for estimators L, LI and IL are from 0.001 to 1 with increment 0.001. At every combination (n,a,q,f), we run 1000 simulations and calculate the MSE for each estimator. The MSE of the estimator S is selected as a baseline and the relative efficiencies (RE) of the other estimators are then obtained as the ratio of their MSE's to the baseline. To get an overall evaluation of the performance of these estimators, we look at the averaged relative efficiency (ARE) at each (n,f), (n,a) and (n,q). Specifically, at each (n,f), the ARE for an estimator is calculated as an averaged value of the RE's over the 12 combinations of (a,q). The ARE's at each (n,a) and (n,q) are calculated similarly and we summarize these ARE's in Tables 3-5 respectively. In each column of the table, the smallest ARE is in bold font and the largest is italicized. Finally, we compare the ARE's over all combinations of (n,a,q,f) and list them in Table 6.

We first consider the effects of the shape of the disease-exposure relationship and sample size in Table 3. In general, for fixed sample size, the regression estimators do better compared to S for functions f_4, f_5, f_6 and f_7, for which $f'(0)\leq 1$, than for functions f_1, f_2 and f_3, for which $f'(0)>1$. The larger $f'(0)$ is, i.e., the steeper the rate of change of the probability of disease given exposure is at zero exposure, the less information there is for the regression estimators to gain over S from borrowing information around the neighborhood of zero exposure to estimate the disease probability at zero exposure and the more potential there is for bias from attempting to borrow information. Among all models, f_1 has the steepest increase in disease rate at zero exposure. For f_1, most estimators except estimator I are worse than S (i.e., have ARE greater than 1) for most sample sizes. The nonparametric regression estimators L, LI and IL, which borrow information on Y values from the neighborhood of $x=0$ to estimate $P(Y=1|X=0)$, tend to overestimate $P(Y=1|X=0)$ for small sample sizes and thus underestimate the AF. For $n=30$, estimator L has the largest ARE. However, for the larger sample size of $n=500$, bias in L, LI and IL are less of a problem; L and LI are only 9% worse than S and IL is 4% better than S. The estimator I is a little better than S for all sample sizes

for f_1. The power model is comparable to S for all sample sizes for f_1. Logistic regression performs poorly for larger sample sizes for f_1; for $n = 500$, logistic regression has a MSE more than 10 times that of S. There are similar patterns in comparisons among estimators for f_2 and f_3 as for f_1. However, for f_2 and f_3, all of the regression estimators perform better than S for the small sample size of $n = 30$, and for f_3, most are better than S for $n = 100$. Estimator I continues to perform well and be better than S for all sample sizes for f_2 and f_3. IL is better than I for sample sizes $n = 30$ and $n = 100$ for f_2 and f_3 but a little worse for $n = 500$. The logistic regression estimator does not perform as badly for f_2 and f_3 as for f_1 but still has an ARE compared to S of around 2 for $n = 500$.

Table 3: The averaged relative efficiency (ARE) of the regression estimators compared to S at each (n, f).

n		f_1	f_2	f_3	f_4	f_5	f_6	f_7
30	I	**0.8412**	0.7607	0.7246	*0.7167*	*0.7311*	*0.7039*	*0.6733*
	IL	1.1306	**0.5570**	**0.4297**	0.4597	0.5190	0.4846	0.4712
	L	*1.3691*	0.7111	0.5723	0.4878	0.4671	0.4417	0.3978
	LI	1.3406	0.6740	0.5204	**0.4472**	**0.4302**	**0.3903**	**0.2993**
	Lg	1.1224	0.6323	0.5425	0.4708	0.5353	0.4582	0.4385
	P	0.9497	*0.8474*	*0.8021*	0.6753	0.5415	0.5798	0.6274
100	I	**0.9219**	0.8572	0.8447	*0.8513*	*0.8498*	*0.8408*	*0.8108*
	IL	1.4721	0.8472	**0.5440**	0.5378	0.6849	0.6384	0.6544
	L	1.9139	*1.0461*	0.7460	0.5302	0.5280	0.4986	0.4609
	LI	1.9046	1.0252	0.7226	0.5121	**0.5045**	**0.4742**	**0.3833**
	Lg	*2.3381*	**0.8434**	0.7217	**0.5094**	0.7913	0.5671	0.5492
	P	1.0069	1.0405	*1.0855*	0.8200	0.5164	0.5884	0.4323
500	I	0.9721	**0.9472**	**0.9370**	*0.9345*	0.9501	0.9362	0.9300
	IL	**0.9631**	1.0979	0.9961	0.6191	0.9902	0.9925	1.1734
	L	1.0860	1.2869	1.0978	0.6050	0.6933	0.6902	0.6822
	LI	1.0857	1.2809	1.0929	**0.5975**	0.6865	0.6850	0.6581
	Lg	*10.1670*	*2.2590*	*1.8206*	0.6529	*2.2242*	*1.1847*	*1.3129*
	P	1.0025	1.1281	1.2886	0.9129	**0.5262**	**0.6389**	**0.2153**

For the functions f_4, f_5, f_6 and f_7, we would expect that the regression estimators would perform relatively better compared to S than for f_1, f_2 and f_3 because the more slowly increasing rate of $P(Y = 1|X = x)$ in x means there is more to gain from borrowing information around the neighborhood of $x = 0$ to estimate $P(Y = 1|X = 0)$. This is indeed the case, and for the functions f_4, f_5, f_6 and f_7, the regression estimators generally perform better than S. The estimator I continues

to always perform better than S. However, the estimator I, which performed well compared to other regression estimators for f_1, f_2 and f_3, is generally the worst among the regression estimators for f_4, f_5, f_6 and f_7. The estimator LI is generally the best among the nonparametric regression estimators, followed by L and IL. Logistic regression performs better than S for the smaller sample sizes of $n = 30$ and $n = 100$ for f_4, f_5, f_6 and f_7 and for $n = 500$ for f_4 but worse for $n = 500$ for f_5, f_6 and f_7. The power model estimator P generally does well and is the best estimator for $n = 500$ for f_5, f_6 and f_7.

We now examine the effect of the parameter a, the probability of disease given zero exposure, in Table 4. As a increases, the rate of change of the probability of disease at zero exposure decreases. We expect that as the rate of change of the probability at zero exposure decreases, (i.e., a increases), there is more to gain from borrowing information from the neighborhood of zero to estimate the probability of disease at zero exposure and that the regression estimators will do better compared S. This is indeed the case in Table 4 for all of the regression estimators except I, which does worse compared to S as a increases.

Next, we examine the effect of the parameter q, the probability of zero exposure, in Table 5. In general, as q increases, the efficiency of the regression estimators compared to S decreases (i.e., the ARE increases). When q is larger, the sample average estimate of $P(Y = 1|X = 0)$ that S uses is more accurate and there is less to gain from borrowing information from the neighborhood around zero exposure.

Our simulation study indicates that performance of the estimators usually depends on f, a, q and n. Table 6 presents the overall average relative efficiency over all the settings in the simulation study. The estimator LI performs the best. IL, L and P are a little worse, but considerably better than S. I is better than S but somewhat worse than LI, IL, L and P. Lg is considerably worse than S. The estimator LI also had the smallest ARE the most times over the different settings. We also note that LI was better than L for all f, a, q and n, although the gain was often small, indicating that incorporating monotonicity provides a consistent but small gain.

5 Summary

Estimation of the attributable fraction essentially depends on estimation of the underlying conditional probability of disease given the exposure. We have studied the performance of several different estimators when the exposure is semi-continuous, in particular the estimators S (based on sample averages), I (isotonic regression), IL (isotonic regression followed by local linear smoothing), L (local linear smoothing),

LI (local linear smoothing followed by isotonic regression), Lg (logistic regression) and P (power model).

The comparison among estimators depends on the sample size, the true conditional probability model, the value of the conditional probability at zero exposure, and the proportion of zero exposures. Based on our simulation study, we make the following recommendations:

- The classical estimate S can be improved upon by regression estimators, in particular when the change in the probability of disease given exposure at zero exposure is not steep.
- However, the logistic regression estimator Lg, where the logit of the probability is linear in the exposure, should be avoided. It is not robust to deviations from the logistic regression model being true and was worse than the classical estimate S in our study.
- The power model that was proposed by Smith et al. (1994) for estimating AFs generally works well and did considerably better than the classical estimate S in our study.
- Nonparametric regression estimators work well and improve considerably on the classical estimate S when the change in the probability of disease given exposure at zero exposure is not steep. When the change in the probability of disease given exposure at zero exposure is steep, the nonparametric regression estimators are sometimes a little worse and sometimes a little better than S. The nonparametric regression estimators performed similarly overall to the power model. The LI estimator was slightly better overall than the power model.
- There is a small, but consistent across different settings, gain to incorporating the constraint that the probability of disease is monotonically increasing in exposure into the nonparametric regression estimators, assuming this constraint is believed to be true.

6 Discussion

We have focused on estimating the AF when there are no confounders or when we are interested in estimating the AF within a stratum of confounders. When there are J strata of confounders, the overall AF can be estimated by combining the estimates of the AF within strata $1,\ldots,J$ using the weighted sum method (Benichou, 2001, Whittemore, 1982): $\hat{AF} = \sum_{j=1}^{J} w_j \hat{AF}_j$ where w_j is the proportion of diseased individuals in stratum j and $\hat{AF}_j$ is the AF estimate for stratum j. Properties

of the combined estimate need investigation and the argument may be based on the informal justification below (4).

We have considered cohort studies in this paper. The odds ratio from a case-control study approximates the relative risk from a corresponding cohort study when the disease is rare (Greenland and Drescher, 1993, Drescher and Schill, 1991), and consequently the attributable fraction can be estimated using (3) by plugging in regression model estimates of the odds ratio $\text{Odds}(Y_i = 1|X_i = 0)/\text{Odds}(Y_i = 1|X_i = x_i)$ for the case-control study for the relative risk $P(Y_i = 1|X_i = 0)/P(Y_i = 1|X_i = x_i)$ in (3) (Benichou, 2001, Bruzzi et al., 1985, Drescher and Schill, 1991). It would be useful to study in future work how the parametric and nonparametric estimates of the attributable fraction for a semi-continuous exposure considered in this paper for cohort studies perform in case-control studies.

References

Agresti, A. and B. A. Coull (1998): "Approximate is better than "exact" for interval estimation of binomial proportions," *American Statistician*, 52, 119–126.

Aragaki, A. and N. S. Altman (1997): *Local polynomial regression for binary response computer science and statistics: proceedings of the 29th symposium on the interface, ed. Scott D. W., Fairfax Station*, VA: Interface foundation of North America.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972): *Statistical inference under order restrictions: the theory and application of isotonic regression*, Wiley, New York.

Benichou, J. (2001): "A review of adjusted estimators of attributable risk," *Statistical Methods in Medical Research*, 10, 195–216.

Benichou, J. (2005): *Attributable risk. In: Armitage P, Colton T (editors),* Encyclopedia of Biostatistics, John Wiley & Sons: Chichester, UK, 2nd edition.

Benichou, J. and M. H. Gail (1990): "Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models," *Biometrics*, 46, 991–1003.

Bhattacharya, R. and M. Kong (2007): "Consistency and asymptotic normality of the estimated effective doses in bioassay," *Journal of Statistical Planning and Inference*, 137, 643658.

Bruzzi, P., S. B. Green, D. P. Byar, L. S. Brinton, and C. Schairer (1985): "Estimating the population attributable risk for multiple risk factors using case-control data," *American Journal of Psychiatry*, 122, 904–914.

Chen, Y. Q. (2008): *Attributable risk.* Encyclopedia of Medical Decision Making *(Ed. Michael Kattan)*, Thousand Oakes: SAGE.

Dette, H., N. Neumeyer, and K. F. Pilz (2005): "A note on nonparametric estimation of the effective dose in quantal bioassay," *Journal of American Statistical Association*, 100, 503–510.

Dette, H. and R. Scheder (2010): "A finite sample comparison of nonparametric estimates of the effective dose in quantal bioassay," *Journal of Statistical Computation and Simulation*, 80, 527544.

Deubner, D. C., W. E. Wilkinson, M. J. Helms, H. A. Tyroler, and C. G. Hames (1980): "Logistic model estimation of death attributable to risk factors for cardiovascular disease in evans county, georgia," *American Journal of Epidemiology*, 112, 135–143.

Drescher, K. and W. Schill (1991): "Attributable risk estimation from case-control data via logistic regression," *Biometrics*, 4, 1247–1256.

Ghosh, D. (2007): "Incorporating monotonicity into the evaluation of a biomarker," *Biostatistics*, 8, 402–413.

Greenland, S. and K. Drescher (1993): "Maximum likelihood estimation of the attributable fraction from logistic models," *Biometrics*, 49, 865–872.

Hommel, M. (2002): *Diagnostic methods in malaria.* Essential Malariology*, Warrell D. A., Gilles H. M. eds.*, Arnold; London, UK.

Kitua, A. Y., T. Smith, P. L. Alonso, H. Masanja, H. Urassa, C. Menendez, J. Kimario, and M. Tanner (1996): "*Plasmodium falciparum* malaria in the first year of life in an area of intense and perennial transmission," *Tropical Medicine and International Health*, 1, 475–484.

Koram, K. A. and M. E. Molyneux (2007): "When is "malaria" malaria? the different burdens of malaria infection, malaria disease and malaria-like illnesses"," *American Journal of Tropical Medicine and Hygiene*, 77, 1–5.

Morabia, A. and E. L. Wynder (1991): "Cigarette smoking and lung cancer cell types," *Cancer*, 68, 2074–2078.

Müller, H.-G. and T. Schmitt (1988): "Kernel and probit estimates in quantal bioassay," *Journal of American Statistical Association*, 83, 750–759.

Park, D. and S. Park (2006): "Parametric and nonparametric estimators of ed100α," *Journal of Statistical Computation and Simulation*, 76, 661672.

Rice, J. (1984): "Bandwidth choice for nonparametric kernel regression," *Annals of Statistics*, 12, 1215–1230.

Rothman, K. J. and S. Greenland (1998): *Causation and causal inference.* Modern Epidemiology, Lippincott Williams & Wilkins: New York.

Royston, P., G. Ambler, and W. Sauerbrei (1999): "The use of fractional polynomials to model continuous risk variables in epidemiology," *International Epidemiological Association*, 28, 964–974.

Royston, P., W. Sauerbrei, and H. Becher (2010): "Modeling continuous exposures with a 'spike' at zero: a new procedure based on fractional polynomials," *Statis-*

tics in Medicine, 29, 1219–1227.

Signorini, D. F. and M. C. Jones (2004): "Kernel estimators for univariate binary regression," *Journal of American Statistical Association*, 465, 119–126.

Smith, T., J. A. Schellenberg, and R. Hayes (1994): "Attributable fraction estimates and case definitions for malaria in endemic areas," *Statistics in Medicine*, 13, 2345–2358.

Traskin*, M., W. Wang*, T. R. T. Have, and D. S. Small (first published online June 21, 2012): "Efficient estimation of the attributable fraction when there are monotonicity constraints and interactions," *Biostatistics*, 10.1093/biostatistics/kxs019, *: joint first author.

Vounatsou, P., T. Smith, and A. F. M. Smith (1998): "Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions," *AppS*, 47, 575–587.

Wetzel, R. D. (1976): "Hopelessness, depression, and suicide intent," *Archives of General Psychiatry*, 33, 1069–1073.

Whittemore, A. S. (1982): "Statistical methods for estimating attributable risk from retrospective data," *Statistics in Medicine*, 1, 229–243.

Wolfe, R. A., L. D. Roi, and E. Margosches (1981): "Monotonic dichotomous regression: A burn care example," *Biometrics*, 37, 157–167.

Zhao, L. P., A. R. Kristal, and E. White (1996): "Estimating relative risk functions in case-control studies using a nonparametric logistic regression," *American Journal of Epidemiology*, 6, 598–609.

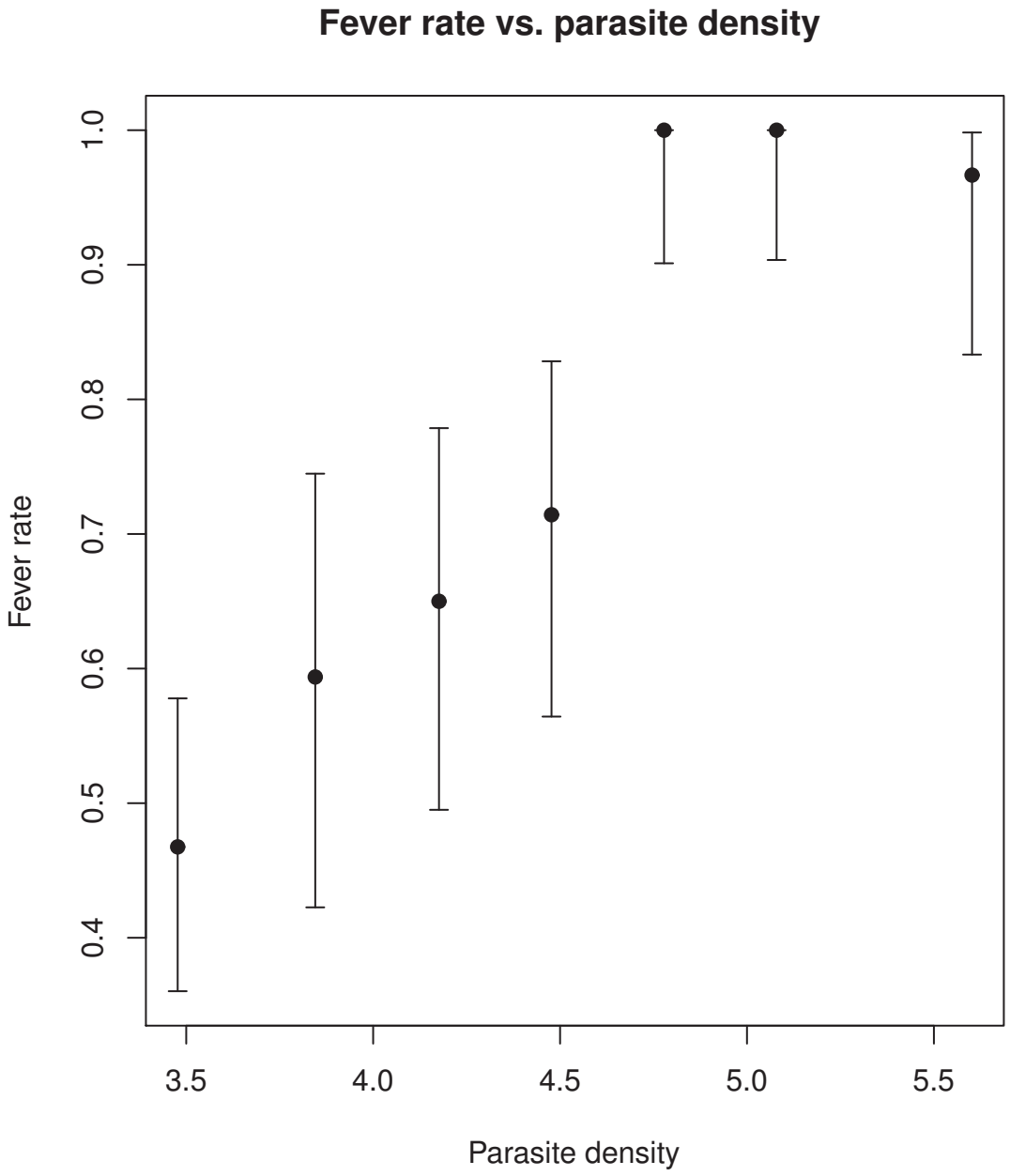

Figure 1: Fever rates for the intervals in Table 1 and the associated 95% confidence intervals. The fever rate is plotted versus the right end point of the interval in Table 1. A logarithmic (base 10) scale is used for the x (parasite density) axis.

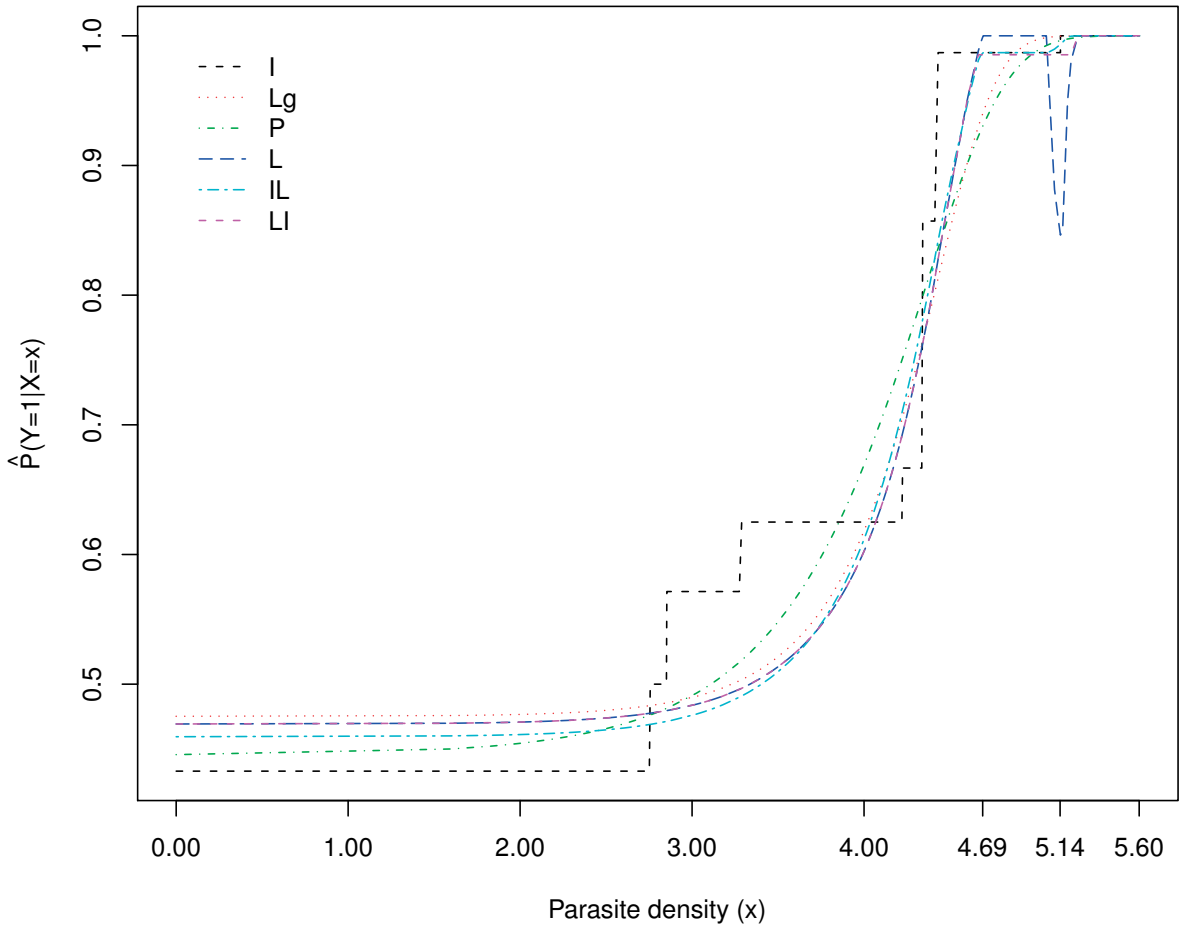

Figure 2: Estimated probability of having a fever from different regression estimators: I, IL, L, LI, Lg and P. A logarithmic (base 10) scale is used for the x (parasite density) axis.

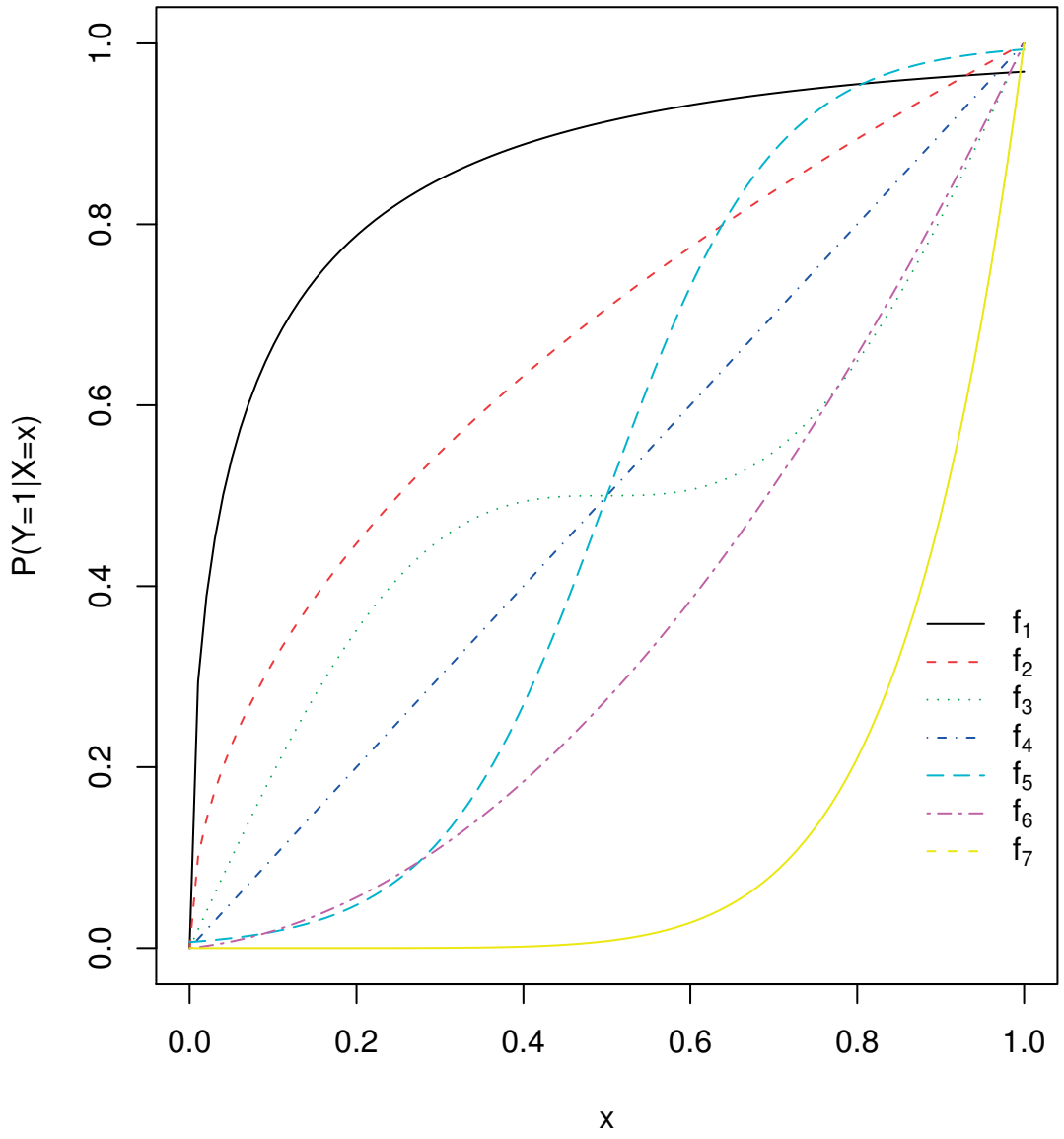

Figure 3: Models of $P(Y = 1|X = x)$ used in simulations.

Table 4: The averaged relative efficiency (ARE) of the regression estimators compared to S at each (n,a).

	30				100				500			
	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
I	**0.5663**	*0.7348*	*0.7966*	*0.8460*	**0.7573**	*0.8719*	*0.8840*	*0.9020*	0.9149	0.9352	0.9617	0.9637
IL	0.7245	0.6172	0.5269	0.4467	1.0165	**0.7944**	0.6644	0.5983	1.1972	1.0068	0.8998	0.8004
L	*0.9258*	0.6495	0.5255	0.4403	1.2276	0.8293	0.6571	0.5568	1.1604	0.8984	0.7781	0.6725
LI	0.8993	0.6139	**0.4711**	**0.3597**	1.2197	0.8113	**0.6253**	**0.5016**	1.1602	0.8953	**0.7695**	**0.6529**
Lg	0.7470	**0.5684**	0.5369	0.5477	*1.5501*	0.8118	0.6581	0.5915	*6.2955*	*2.3601*	*1.5153*	*1.0414*
P	0.7762	0.7097	0.6801	0.7045	0.8995	0.8003	0.7432	0.6941	**0.9007**	**0.8213**	0.7863	0.7559

Table 5: The averaged relative efficiency (ARE) of the regression estimators compared to S at each (n,q).

	30			100			500		
	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
I	*0.5729*	0.7845	0.8504	0.7649	0.8834	*0.9131*	0.9037	0.9510	0.9770
IL	**0.3826**	**0.6365**	0.7174	0.6765	0.7942	0.8346	0.8445	1.0203	1.0633
L	0.4747	0.6865	0.7445	0.8108	0.8137	0.8285	0.8087	0.8871	0.9362
LI	0.4144	0.6395	**0.7040**	0.7665	**0.7913**	**0.8106**	0.7938	0.8821	0.9326
Lg	0.3846	0.6482	0.7672	*0.8523*	*0.9778*	0.8785	*3.7163*	*2.8984*	*1.7945*
P	0.5139	*0.7859*	*0.8530*	**0.6678**	0.8201	0.8649	**0.7576**	**0.8276**	**0.8630**

Table 6: The overall averaged relative efficiency (ARE) of the estimators.

S	I	IL	L	LI	Lg	P
1.0000	0.8445	0.7744	0.7768	0.7483	1.4353	0.7726