Optimal Binary Prediction for Group Decision Making

Robert P. Lieli               Augusto Nieto Barthaburu
Department of Economics       Department of Economics
University of Texas, Austin   East Carolina University
September 29, 2008

Abstract

In this paper we address the problem of optimally forecasting a binary variable for a heterogeneous group of decision makers facing various (binary) decision problems that are only tied together by the unknown outcome. A typical example is a weather forecaster who needs to estimate the probability of rain tomorrow and then report it to the public. Given a conditional probability model for the outcome of interest (e.g., logit or probit), we introduce the idea of maximum welfare estimation and derive conditions under which traditional estimators such as maximum likelihood or (non-linear) least squares are asymptotically socially optimal even when the underlying model is misspecified.

JEL codes: C51, C25, C53, C44.

1 Introduction

Typically, the reason why one attempts to forecast the future value of a random variable $Y$ is because knowledge of this variable would enable a decision maker, or a group of decision makers, to choose the best action from a set of alternatives. The value of a forecast is thus the improvement in decisions under uncertainty afforded by it. Although most of traditional forecasting theory is based on expected loss minimization, the loss function is rarely derived explicitly form an underlying decision problem and the decision-supporting aspect of forecasting is generally neglected.
In recent years this traditional approach has come under criticism. In particular, as Granger and Pesaran (2000a,b) and Elliott and Lieli (2007) argue, when forecasts are constructed to aid in a decision making process, it is important to take into account the nature of the decision problem and the preferences of the decision maker (henceforth DM) at all stages of the forecasting procedure. This includes the specification of a relevant loss function and estimation.

The literature on decision-based forecasting is fairly recent, and most of the previous work in the area has focused on forecast evaluation. Papers discussing decision-based forecast evaluation include McCulloch and Rossi (1990), West, Edison and Cho (1993), Diebold, Gunther and Tay (1997), Pesaran and Skouras (2001), and Bond and Satchell (2004). Treatments in which decision based forecasting methods are proposed are found in Elliott and Lieli (2007) and Crone, Lessmann and Stahlbock (2005).

In this paper we concentrate on two-action, two-state decision/forecasting problems. The action to be taken by each DM is denoted by \( a \in \{-1, 1\} \), and \( Y \in \{-1, 1\} \) denotes the state of the world, to be observed only after the decision has been made. Information about the future outcome \( Y \) is summarized by a \( k \)-dimensional vector \( X \) of covariates, observed prior to making a decision. The following examples illustrate the setting:

**Example 1** A loan officer observes the credit history, income, household size, etc. of an applicant \((X)\). The possible outcomes are default \((Y = 1)\) or no default \((Y = -1)\). The possible actions are grant the loan \((a = 1)\) or not \((a = -1)\).

**Example 2** A weather forecaster observes today’s weather conditions \((X)\) in a certain geographic region. The possible states of the world are rain tomorrow \((Y = 1)\) or no rain tomorrow \((Y = -1)\). Farmers in the region have to decide whether to irrigate \((a = 1)\) or not \((a = -1)\); commuters have to decide whether to take along an umbrella \((a = 1)\) or not \((a = -1)\), etc.

We will make an explicit distinction between the forecast user (i.e. the DM) and the forecaster. We further distinguish between two kinds of forecasting activities, which we will refer to as “private consulting” and “public service”. In the consultant’s problem, illustrated
by Example 1, there is a single forecast user with well-defined preferences facing a concrete decision problem. Given data on $X$ and $Y$, and the DM’s utility function, the forecaster’s job is to predict $Y$ in a way that will lead the particular DM in question to their expected utility maximizing action. On the other hand, the public service forecasting problem (Example 2) is characterized by the presence of a heterogeneous population of decision makers facing potentially different binary decision problems, which are only tied together by a binary outcome $Y$. Given data on $X$ and $Y$, the forecaster provides a social service by constructing a single public forecast of $Y$; this information is then used as an input in many separate decision making processes.

The consultant’s problem was studied most recently by Elliott and Lieli (2007). They show that traditional binary forecasting methods, such as logit or probit estimated by maximum likelihood (ML), may easily contradict a given DM’s decision-theoretic objective. The reason for this is that traditional estimators of potentially misspecified parametric probability models, such as (quasi) ML, provide a global asymptotic approximation to the true conditional probability function $p(x) = P(Y = 1 | X = x)$ over the support of $X$. A global approximation may miss important features of a given decision problem; in particular, it may misrepresent those points $x$ at which the optimal decision changes. It is precisely at these points where it is important to approximate $p(x)$ locally well—at other points relatively small estimation errors are inconsequential for decision making.

In contrast, in this paper we concentrate on the public service forecasting problem, i.e. the multiple DM setting (Example 2). We argue that in constructing the forecast, the forecaster will generally need to strike a balance between the objectives of various decision makers. Different DMs have different points at which their optimal decision changes, and with a given parametric specification it might not be possible to fit the model through all these points. In this case the forecaster will need to provide an approximation to $p(x)$ that leads the population of DMs to “adequately good decisions” (this will be made more precise below). It will be shown that global approximations to $p(x)$ might be optimal in the sense of maximizing some social welfare objective set forth by the forecaster.

The contributions of this paper are threefold. First, we introduce the idea of a public
service forecasting problem, in which the forecaster maximizes a weighted sum of individual (expected) utilities. Thus we extend the literature on decision-based forecasting, which currently only deals with the single DM case, i.e. setups similar to the consultant’s problem. Second, we introduce the “maximum welfare” (MW) estimator through the forecaster’s maximization problem, and explore some of its properties. Third, we spell out conditions under which traditional binary prediction methods such as logit or probit regressions estimated by ML or (nonlinear) least squares (LS) can be interpreted, asymptotically, as socially optimal. In particular, we identify some special populations of decision makers and social welfare objectives for which these methods are optimal even when the estimated model is misspecified.

The paper will be organized as follows: In Section 2 we present an intuitive discussion of the public forecasting problem. In Section 3 we formalize the problem, while introducing the notion of social welfare estimation. In Section 4 we provide conditions on individual preferences and the social welfare objective under which ML and LS estimation can be interpreted as socially optimal. In Section 5 we conduct a Monte Carlo exercise to compare the performance of the MW estimator to other estimators. In Section 6 we apply the proposed MW estimator to data on Coronary Heart Disease (CHD) incidence. Section 7 summarizes and concludes. Proofs of all theoretical results are collected in the Appendix.

2 An informal discussion of the forecasting problem

In the public service forecasting problem each individual DM is assumed to have a utility function of the form

$$U(a,Y) = u_{a,Y}, \quad a \in \{-1, 1\}, Y \in \{-1, 1\}.$$ 

The utility function $U(a,Y)$ can be conveniently represented by the two-by-two matrix given in Table 2. (More generally, the DM’s utility function could also depend on the vector $X$ of covariates, but to avoid complications beyond the focus of this paper, we rule this possibility out.) We introduce the following assumption about the utility function, which ensures that the decision problem is non-trivial:
Table 1: The DM’s preferences

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = -1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a = 1$</td>
<td>$u_{1,1}$</td>
<td>$u_{1,-1}$</td>
</tr>
<tr>
<td>$a = -1$</td>
<td>$u_{-1,1}$</td>
<td>$u_{-1,-1}$</td>
</tr>
</tbody>
</table>

Assumption 1 $u_{1,1} > u_{1,-1}$ and $u_{-1,-1} > u_{-1,1}$.

In this simple setting the conditional distribution of the outcome $Y$ given the covariates $X$ is completely characterized by the conditional probability $p(X) = P(Y = 1 \mid X)$. As will be shown in Section 3, the expected utility maximizing decision rule of an individual DM is of the form $a^*(X) = \text{sign}[p(X) - c]$, where the optimal cutoff $c$ is determined by the utilities in Table 2 and the function $\text{sign}(x)$ is defined as 1 if $x > 0$ and $-1$ if $x \leq 0$. In less cryptic terms, the optimal decision rule states “take action $a = 1$ if $p(X) > c$ and take action $a = -1$ if $p(X) \leq c$”.

Let $\mathcal{X} \subset \mathbb{R}^k$ denote the support of $X$, and let $m(x, \theta)$ be a potentially misspecified parametric model of the conditional probability function $x \mapsto p(x)$, $x \in \mathcal{X}$. (Logit or probit are the most common examples.) As pointed out by Elliott and Lieli (2007), if there is only one decision maker, expected utility maximization dictates that the model should be fit accurately at those points $x$ in $\mathcal{X}$ where the cutoff intersects $p(x)$. The magnitude of errors is inconsequential away from these locations as long as the fitted model is on the “correct” side of the cutoff.

Figure 1 illustrates this point. In the figure the solid line represents the function $p(x)$, obtained by “pasting” parts of two logistic c.d.f.’s. We drew a large number of independent observations on $X$ from the uniform $[-3,6]$ distribution, and generated the corresponding $Y$ outcomes according to $p(X)$. The dark dashed line, a misspecified linear logit estimated by ML, represents a globally good fit. The light dashed line, the same model fitted by the “maximum utility” estimator in Elliott and Lieli (2007), represents a locally good fit. We see that the global fit is closer to $p(x)$ than the local fit for most $x \in \mathcal{X}$. However, a locally good fit leads to optimal decisions for all values of $X$, as it is always on the same side of the
DM’s cutoff as $p(x)$. On the other hand, the global fit misses the intersection point between $p(x)$ and the cutoff line, and therefore it leads to suboptimal decisions in a sizable portion of $X$. Hence, although closer to the truth on average, the global fit is inferior to the local fit for individual decision making.

In contrast to Elliott and Lieli (2007), in this paper we concentrate on the social service framework (multiple DMs), in which the forecaster’s problem is more complex. In this setting, each individual DM $j$ has an optimal decision rule of the form $a_j^*(X) = \text{sign}[p(X) - c_j]$, where the optimal cutoff $c_j$ may vary with $j$, i.e. from individual to individual. Figure 2 provides an illustration. In the figure we depict a relatively simple problem with only three DMs who possess different optimal cutoffs. A logit specification based on a linear index cannot possibly run through all the intersection points between $p(x)$ and the individual cutoffs. Therefore, given the model specification, it is not possible to ensure correct decisions for all DMs. Fitting the model will necessarily entail implicit or explicit interpersonal utility comparisons among individuals.

As Figures 1 and 2 suggest, a globally good fit of a conditional probability model may not be consistent with maximizing the expected utility of a given DM. However, it may well be consistent with maximizing the social welfare of a group of decision makers (the public forecaster’s problem). A natural question to ask is the following: Under what conditions

Figure 1: Globally good fit vs. expected utility maximizing (locally good) fit.
do traditional estimators (LS, ML) of potentially misspecified parametric models lead to empirical decision rules that maximize the social welfare of a population of DMs (asymptotically)? In Sections 3 and 4 we provide a formal answer to this question. In doing so, we give an economically relevant interpretation of traditional estimation methods in the context of binary prediction models; specifically, our results show that these methods possess asymptotic optimality properties for certain—rather special—populations of DMs and social welfare objectives even under model misspecification.

3 Formal statement of the forecasting problem

Conditional on the observed value of $X$, an individual DM’s expected utility maximization problem is given by

$$
\max_{a \in \{-1,1\}} E[u(a,Y) \mid X] = \max_{a \in \{-1,1\}} \{u_{a,1}p(X) + u_{a,-1}[1 - p(X)]\}.
$$

It is implicit in the formulation of problem (1) that there is no feedback from the action taken to the (conditional) probability of the outcome $Y$; otherwise one should write $p_a(x)$ in place of $p(x)$. We do not allow for such feedback from the action to the outcome of interest.

It is easy to see that (1) leads to the optimal decision rule $a^*(X) = sign[p(X) - c]$, where
the optimal cutoff \( c \) is given by

\[
c = (u_{-1,-1} - u_{1,-1})/b \quad \text{with} \quad b = u_{1,1} - u_{-1,1} + u_{-1,-1} - u_{1,-1}.
\]  

(2)

Under Assumption 1, \( b > 0 \) and the optimal cutoff \( c \) lies strictly between 0 and 1. The numerator of \( c \) measures the net benefit from taking the correct action (i.e. \( a = -1 \)) when \( Y = -1 \). The denominator \( b \) is the absolute net benefit, totaled across states, from taking correct action. Hence, \( c \) can be interpreted as the relative net benefit of taking action \(-1\) versus action \(1\). The higher this value, the higher the probability of \( Y = 1 \) has to be to make it worthwhile for the DM to take action \( 1 \) instead of \(-1\).

As shown by Elliott and Lieli (2007), problem (1) can be reformulated as the unconditional maximization problem

\[
\max_{a(\cdot)} E[u(a(X), Y)] = \frac{b}{4} \max_{a(\cdot)} E[(Y + 1 - 2c)a(X)] + \frac{b}{4} E(Y + 1 - 2c) + E[u(-1, Y)],
\]

(3)

where maximization is over all measurable functions \( a(\cdot) \) that map \( X \) into \( \{-1, 1\} \). Problems (1) and (3) are equivalent in the following sense: Suppose that the action \( a^*(x) \in \{-1, 1\} \) solves (1) for each \( x \in X \). Then the function \( x \mapsto a^*(x) \) solves (3). Conversely, if the function \( a^\dagger : X \to \{-1, 1\} \) solves (3), then for \( x \in X \) the action \( a^\dagger(x) \in \{-1, 1\} \) solves (1) except maybe on a set \( E \subset X \) with \( P(X \in E) = 0 \). Thus the individually optimal decision rule \( a^*(x) = \text{sign}[p(x) - c] \) can equivalently be characterized as a solution to (3).

We assume that there is a population of DMs, indexed by \( j = 1, \ldots, J \). (We could easily use a general index set, but a finite population allows for a particularly simple and transparent exposition.) For each DM there is a random draw \((Y_j, X_j)\) from a distribution \( F_{YX} \) on \( \{-1, 1\} \times X \). Thus, \((Y_j, X_j)\) are assumed to be identically distributed random variables, but not necessarily independent. In fact, an important case is when \((Y_j, X_j) = (Y, X)\) for all \( j \), i.e. all DMs observe the same realization of \( X \) before taking action, and eventually experience the same outcome \( Y \) as in Example 2.

Each DM has a utility function \( u_j(a_j, y_j) \) and seeks to maximize their own expected utility conditional on \( X_j \). Suppose further that the forecaster adopts the parametric model
\( m(x, \theta) \) for \( p(x) \), where \( \theta \in \Theta \subset \mathbb{R}^d \), and estimates the unknown parameter vector to be equal to \( \hat{\theta} \). Given that the forecaster reports the estimated conditional probability function \( m(x, \hat{\theta}) \) to the public, decision makers will be assumed to mimic their optimal decision rule simply by plugging this estimate in place of \( p(x) \), i.e. taking the action

\[
\hat{a}_j(X_j) = \text{sign}[m(X_j, \hat{\theta}) - c_j], \tag{4}
\]

where \( c_j \) is individual \( j \)'s optimal cutoff (2) derived from their utility function \( u_j \).

Conditional on the model specification \( m(\cdot, \theta) \), any given DM \( j \) would like the forecaster to choose \( \hat{\theta} \in \Theta \) so as to maximize their own expected utility. By equation (3), this could be accomplished by solving

\[
\max_{\hat{\theta}} E\{(Y + 1 - 2c_j)\text{sign}[m(X, \hat{\theta}) - c_j]\}. \tag{5}
\]

However, the forecaster is concerned with providing a single forecast to be used by all DMs. In this paper we assume that the forecaster’s goal is to maximize a weighted sum of the individual objective functions (5). Since individual preferences enter (5) only through \( c_j \), we can consolidate decision makers sharing the same optimal cutoff value into “types”. In particular, suppose that individual cutoffs take on \( T \) different values \( c_1, \ldots, c_T \). The forecaster’s objective is to solve

\[
\max_{\theta} W(\theta) = \max_{\theta} \sum_{t=1}^{T} \lambda_t E\{(Y + 1 - 2c_t)\text{sign}[m(X, \theta) - c_t]\}, \tag{6}
\]

where \( \lambda_t \geq 0 \) and \( \sum_{t=1}^{T} \lambda_t = 1 \).

The coefficients \( \lambda_t, t = 1, \ldots, T \), represent the total weight given by the social welfare function to DMs of type \( t \) (i.e. those with cutoff \( c_t \)). The relative magnitudes of the type welfare weights reflect two different factors. First, given an equal number of DMs of each type, the \( \lambda \)'s can differ because (some) individuals in one group are weighted more heavily than individuals in another group. Second, even if each individual DM \( j \) is weighted equally, some types may have more individuals than others.

We formulated the objective function \( W(\theta) \) under the presumption that individual DMs will act in accordance with (4). This assumption abstracts from the possibility of strategic
behavior on the part of individual agents: If DMs know the forecaster’s objective function, in principle they could try to make adjustments to the forecast provided so that it fits their preferences more closely. Considering this potential feedback in the forecaster’s objective would introduce a complicated fixed point problem. Instead we simply make the behavioral assumption that agents act as if they fully believed the reported forecast.

As the joint distribution of \( Y \) and \( X \) is unknown, one cannot directly solve (6) in practice. Given a sample of observations \( \{(X_i, Y_i)\}, i = 1, 2, \ldots, n \) available to the forecaster, it is natural to replace the population objective function (6) with its sample average \( W_n(\theta) \), defined as

\[
W_n(\theta) = n^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \lambda_t \left\{ (Y_i + 1 - 2c_t) \text{sign}[m(X_i, \theta) - c_t] \right\}.
\]

The maximum welfare (MW) estimator is then defined as a maximizer of \( W_n(\theta) \). We note that the MW estimator can be regarded as a generalization of Manski’s (1975, 1985) maximum score estimator. We introduce the following assumptions:

**Assumption 2**

(a) \( \{(Y_i, X_i')\}_{i=1}^{\infty} \) is a (strictly) stationary, ergodic sequence of observations.

(b) \( \Theta \subset \mathbb{R}^d \) is compact.

(c) \( x \mapsto m(x, \theta) \) is Borel-measurable for all \( \theta \in \Theta \) and \( 0 \leq m(x, \theta) \leq 1 \) for all \( x \in X, \theta \in \Theta \).

(d) \( \theta \mapsto m(x, \theta) \) is continuous on \( \Theta \) for all \( x \in X \).

(e) Either all components of \( X \) are discrete or \( P[m(X, \theta) = k] = 0 \) for all \( \theta \in \Theta \) and \( k \in [0, 1] \).

It is easy to adopt the arguments in Elliott and Lieli (2007) to show that under Assumption 2, \( W_n(\theta) \) converges to \( W(\theta) \), uniformly in \( \theta \), and that for any sequence \( \hat{\theta}_n^{MW} \in \arg \max W_n(\theta) \), the distance between \( \hat{\theta}_n^{MW} \) and \( \arg \max W(\theta) \) goes to zero almost surely as \( n \to \infty \) even when \( \arg \max W(\theta) \) is a set. In short, Assumption 2 allows one to think of the maximum welfare estimator as asymptotically maximizing \( W(\theta) \).

The theoretical question we now plan to address is the following. Is there a population of decision makers and a set of welfare weights under which familiar estimators of the model \( m(x, \theta) \), such as maximum likelihood and (non-linear) least squares, are asymptotically socially optimal? In other words, are there cutoffs \( c_t \) and corresponding welfare weights \( \lambda_t \), \( t = 1, \ldots, T \), such that solving (6) is equivalent to maximizing the asymptotic likelihood function or the asymptotic least squares objective? Is there a set of cutoffs and weights that
“works” regardless of the data generating process and the extent of potential model misspecification? The answer is yes; Section 4 is devoted to the development of these results.

4 The asymptotic social optimality of various parametric estimators

For analytical convenience, in our theoretical derivations we will work with a continuum of decision makers with optimal cutoffs in the interval $(0, 1)$. Let $\lambda(c)$ be a Lebesgue density defined on $(0, 1)$. As before, the density $\lambda(c)$ may simply reflect the concentration of decision makers around the cutoff $c$, but it might also incorporate the forecaster’s value judgements concerning decision makers of this type. The forecaster chooses $\theta$ to maximize the continuous analog of the social welfare function:

$$\max_{\theta} W(\theta) = \max_{\theta} \int_0^1 \lambda(c) E\left\{ (Y + 1 - 2c) \text{sign}(m(X, \theta) - c) \right\} dc.$$ 

Results derived for continuous populations can also be interpreted as applicable to discrete populations with a large number of types ($T \to \infty$).

4.1 The correctly specified case

As discussed in Section 3, the function $p(x)$ leads each decision maker in the population to an individually optimal decision. Therefore, reporting $p(x)$ (the truth) as a public forecast must be socially optimal. We formalize this intuition in the following proposition.

**Proposition 1** Let $\mathcal{G}$ denote the set of real valued measurable functions defined on the support of $X$. Then the function $p(x) = P(Y = 1 \mid X = x)$ solves the social welfare maximization problem

$$\max_{g \in \mathcal{G}} \int_0^1 \lambda(c) E\left\{ (Y + 1 - 2c) \text{sign}(g(X) - c) \right\} dc$$

for any weight function (density) $\lambda(c)$ defined on $(0, 1)$. 

11
A formal proof can be found in Appendix A. It is immediate from Proposition 1 that if there exists (unique) \( \theta^* \in \Theta \) such that \( m(x, \theta^*) = p(x) \) for all \( x \in \mathcal{X} \), i.e. the parametric model \( m(x, \theta) \) is correctly specified for \( p(x) \), then the true parameter value \( \theta^* \) solves \( \max_{\theta \in \Theta} W(\theta) \) for any weight function \( \lambda(c) \). Thus, under correct specification any consistent estimator of \( \theta \) is asymptotically socially optimal for any set of welfare weights.

4.2 The misspecified case

More generally (and realistically), let \( m(x, \theta) \) be a potentially misspecified parametric model for \( p(x) \), and let \( \hat{\theta} \) denote an estimator of \( \theta \) obtained by maximizing a sample objective \( Q_n(\theta) \). Suppose that under certain conditions \( \hat{\theta} \) can be shown to converge in probability to the unique maximizer \( \theta^* \) of a non-stochastic objective function \( Q(\theta) \), where \( \theta^* = \theta^o \) whenever the model is correctly specified. Suppose further that there exists a density \( \lambda_Q(c) \), independent of \( p(x) \), the distribution of \( X \) and the model specification \( m(\cdot, \theta) \), such that

\[
\int_0^1 \lambda_Q(c) E\{(Y + 1 - 2c)\text{sign}[m(X, \theta) - c]\} dc = AQ(\theta) + B \quad \text{for all } \theta \in \Theta, \tag{7}
\]

where \( A > 0 \) and \( B \) are constants (as a function of \( \theta \)). Clearly, under such \( \lambda_Q(c) \) maximization of \( W(\theta) \) is equivalent to that of \( Q(\theta) \). We will therefore say that \( \hat{\theta} \) is, asymptotically, a social welfare maximizing estimator under \( \lambda_Q = \lambda_Q(c) \) or, more precisely, that the estimation method with the asymptotic objective function \( Q(\theta) \) is socially optimal under \( \lambda_Q \). The density \( \lambda_Q \) will be referred to as the robust welfare weight function implicit in \( Q(\theta) \).

The critical aspect of the definition of \( \lambda_Q \) is that (7) is supposed to hold, with the same \( \lambda_Q \), for all conditional probability functions \( p(x) \), all distributions of \( X \), and all model specifications \( m(\cdot, \theta) \) for which \( Q(\theta) \) is well-defined and \( \hat{\theta} \) is consistent for \( \theta^* = \arg \max Q(\theta) \). Hence, the estimation method that relies, asymptotically, on maximizing \( Q(\theta) \) will be socially optimal under \( \lambda_Q \) even if the model being estimated is severely misspecified for \( p(x) \). (This is why the qualifier “robust” is appropriate.) The independence requirements built into the definition of a robust welfare weight function ensure that if (7) is satisfied by some \( \lambda_Q \), then, under mild conditions, there does not exist another \( (\lambda_Q)' \) that makes the LHS of (7) identically equal to some other increasing function of \( Q(\theta) \). This uniqueness proposition is formally stated and proven in Appendix B.
So what does an estimator’s robust welfare weight function reveal about the estimator? First recall that a given weight function is open to many interpretations. A “canonical” interpretation can be obtained by assuming that the distribution of decision makers is uniform across all possible types on \((0, 1)\); any deviation of \(\lambda^Q\) from uniformity is then attributable to value judgements implicit in the criterion function \(Q(\theta)\). Alternatively, one can assume that the forecaster wants to treat every decision maker equally so that \(\lambda^Q\) shows the distribution of types for which maximization of \(Q(\theta)\) is socially optimal.

When the model \(m(x, \theta)\) is misspecified, it may not be possible to choose \(\theta\) in a way so that all decision makers are lead to their individually optimal decision rules. No matter what estimation method is used to fit the model to the data, there will be (implicit) tradeoffs between the utilities of various types of decision makers. The robust weight function \(\lambda^Q\) associated with a given estimation method makes these tradeoffs explicit; it reveals which types of decision makers are favored relative to others whenever misspecification forces such tradeoffs in fitting the model. In particular, if the weight assigned to a given type increases, then the optimal decision rule for that type will be reproduced more closely by the asymptotic fit of the model.

We will now derive the robust welfare weight function implicit in the nonlinear least squares and (quasi) maximum likelihood estimators.

4.2.1 The least squares estimator

Let \(m(x, \theta), \theta \in \Theta\), be a potentially misspecified parametric model for \(p(x)\). The least squares estimator of \(\theta\) is defined as the solution \(\hat{\theta}_{n}^{LS}\) to the maximization problem

\[
\max_{\theta \in \Theta} Q_n(\theta) = \max_{\theta \in \Theta} -n^{-1} \sum_{i=1}^{n} [\tilde{Y}_i - m(X_i, \theta)]^2,
\]

where \(\tilde{Y}_i = \frac{1}{2}(Y_i + 1)\). Suppose that Assumption 2 parts (a) through (d) are satisfied. Then by the strong law of large numbers (ergodic theorem) \(Q_n(\theta)\) converges almost surely to

\[
Q(\theta) \equiv -E\{[\tilde{Y} - m(X, \theta)]^2\}
\]

for any fixed value of \(\theta\) as \(n \to \infty\). Of course, pointwise convergence is not enough to guarantee that \(\hat{\theta}_{n}^{LS}\) converges to a maximizer of \(Q(\theta)\)—the key requirement for this is uni-
form convergence (in prob.) of $Q_n(\theta)$ to $Q(\theta)$; see, e.g., Amemiya (1985) or White (1996). The assumptions above are actually sufficient for the strong law of large numbers to hold uniformly in $\theta$; c.f. White (1996, Thm. A.2.2). Additional regularity conditions on $m(x, \theta)$ and the distribution of $X$ can be used to ensure that $Q(\theta)$ has a unique global maximum. We will henceforth think of the least squares estimator as asymptotically maximizing (8).

The robust welfare weight function associated with the least squares estimator is presented in the following proposition:

**Proposition 2** Let $\lambda^{LS}(c) = 1_{(0,1)}(c)$, where $1_A(x)$ denotes the indicator function on the set $A \subset \mathbb{R}$. Let $W^{LS}(\theta)$ denote the social welfare objective under $\lambda^{LS}$. If Assumption 2 part (c) holds, then $W^{LS}(\theta) = AQ(\theta) + B$ with $A = 2$ and $B = 1$.

The proof of Proposition 2 can be found in Appendix C. The only role Assumption 2(c) plays there is to ensure that $Q(\theta)$ and $W(\theta)$ are well-defined. The rest of Assumption 2 is not needed at all in proving Proposition 2—those conditions only help in interpreting it. As explained above, they ensure that the LS estimator indeed asymptotically maximizes $Q(\theta)$, and the MW estimator indeed asymptotically maximizes $W(\theta)$.

Proposition 2 can be interpreted in the following way: Consider estimating the model $m(x, \theta)$ by least squares using a large sample of observations. If the model is misspecified, then $m(x, \hat{\theta}_n^{LS})$ will not be a perfect fit for $p(x)$ even asymptotically. Proposition 2 states that if the distribution of decision makers is uniform across all possible types in $(0, 1)$, then using least squares corresponds to treating all decision makers equally in balancing estimation errors due to misspecification.

### 4.2.2 The (quasi) maximum likelihood estimator

Let $m(x, \theta), \theta \in \Theta$, be a potentially misspecified parametric model for $p(x)$. The maximum likelihood estimator of $\theta$ is defined as a maximizer $\hat{\theta}_n^{ML}$ of the likelihood function

$$L_n(\theta) = n^{-1} \sum_{i=1}^{n} \frac{1}{2} \left\{ (1 + Y_i) \log[m(X_i, \theta)] + (1 - Y_i) \log[1 - m(X_i, \theta)] \right\}.$$ 

Suppose that, in addition to Assumption 2 parts (a) through (d), the following condition is satisfied:
Assumption 3 $E \sup_{\theta \in \Theta} |\log[m(X, \theta)]| < \infty$ and $E \sup_{\theta \in \Theta} |\log[1 - m(X, \theta)]| < \infty$.

Given the maintained assumptions, $L_n(\theta)$ converges almost surely to the asymptotic log-likelihood function

$$L(\theta) = \frac{1}{2} E \left\{ (1 + Y) \log[m(X, \theta)] + (1 - Y) \log[1 - m(X, \theta)] \right\}$$

uniformly in $\theta$ (c.f. White 1996, Thm. A.2.2). Additional regularity conditions on $m(x, \theta)$ and the distribution of $X$ can be used to ensure that $L(\theta)$ has a unique global maximum. We will henceforth think of the maximum likelihood estimator as asymptotically maximizing the function $L(\theta)$.

As shown in Appendix D (see the proof of Lemma 1), if the model $m(x, \theta)$ is not bounded away from zero and one as $(x, \theta)$ varies over $X \times \Theta$, then it is not possible to find $\lambda(c)$ that makes $W(\theta)$ an exact positive affine transformation of $L(\theta)$. Nevertheless, it is still possible to find $\lambda(c)$ for which this is approximately true; further, the approximation error can be made arbitrarily small, uniformly in $\theta$. The following proposition presents the approximate robust welfare weights associated with the maximum likelihood estimator:

**Proposition 3** Given (small) $\epsilon > 0$, define

$$\lambda_{\epsilon}^{ML}(c) = \frac{1(\epsilon, 1 - \epsilon)(c)}{2[\log(1 - \epsilon) - \log(\epsilon)]c(1 - c)}$$

and

$$W_{\epsilon}^{ML}(\theta) = \int_{0}^{1} \lambda_{\epsilon}^{ML}(c) E \{ (Y + 1 - 2c) \text{sign}[m(X, \theta) - c] \} dc.$$

If Assumption 2(c) and Assumption 3 are satisfied, then

$$\sup_{\theta \in \Theta} \left| W_{\epsilon}^{ML}(\theta) - (A_{\epsilon}L(\theta) + B_{\epsilon}) \right| \to 0 \text{ as } \epsilon \to 0,$$

where $A_{\epsilon} = 2[\log(1 - \epsilon) - \log(\epsilon)]^{-1}$ and $B_{\epsilon} = -\frac{\log(\epsilon) + \log(1 - \epsilon)}{\log(\epsilon) - \log(1 - \epsilon)}$.

The proof of Proposition 3 can be found in Appendix D. Assumption 2(c) and 3 ensure that the asymptotic log-likelihood is well-defined; in addition, Assumption 3 plays an important role in showing that the approximation error $W_{\epsilon}^{ML}(\theta) - A_{\epsilon}L(\theta) - B_{\epsilon}$ is uniformly small.
in $\theta$ for small $\epsilon$, i.e. maximizing $W_{\epsilon}^{ML}(\theta)$ is practically equivalent to maximizing $L(\theta)$. The rest of Assumption 2 is not used in the proof. A discrete analog of this result was developed earlier in Lieli and Nieto (2006), the working paper version of this article.

The robust weight function $\lambda_{\epsilon}^{ML}(c)$ is supported on the interval $(\epsilon, 1 - \epsilon)$, is symmetric around $c = 1/2$ and has a $U$-shape. If the distribution of forecast users is uniform across the types $(\epsilon, 1 - \epsilon)$ then using ML implies that DMs with cutoffs close to zero or one get the most consideration. Alternatively, if the forecaster’s intention is to treat all forecast users equally, the ML estimator is asymptotically optimal provided that the distribution of types is known to be heavily concentrated around 0 and 1.

5 Monte Carlo experiments

We now present some Monte Carlo exercises that are designed to give insight into the workings of the maximum welfare estimator. The data generating process is given by

$$p(X) = \Lambda(\theta_0 + \theta_1 X + \theta_2 X^2),$$

where $\theta_0 = -10$, $\theta_1 = 1$, $\theta_2 = -0.026$, and $\Lambda(\cdot)$ is the c.d.f. of the logistic distribution. The distribution of $X$ is uniform over the interval $[4, 24]$. The conditional probability function $p(x)$ is represented by the black solid line in Figure 3. The asymptotic approximation to $p(x)$ provided by the ML estimate of a misspecified linear logit model is shown in Figure 3 as the dash-dot line. (The estimate is based on a sample of 15,000 observations.)

The set of DMs we will consider consists of four types having cutoffs $c_1 = 0.07$, $c_2 = 0.22$, $c_3 = 0.27$, $c_4 = 0.32$ with the same weight assigned to each. As shown in Figure 3, these cutoffs were chosen so that the large sample ML estimate of the linear logit model leads to a suboptimal decision for a relatively large set of $X$-values for most types of DMs. Specifically, the entry entitled “ML (logit, lin.)” in the “Asymptotic” column of Table 2 shows that this model only achieves 90.6% of the optimal value of the forecaster’s objective that would obtain if $p(x)$ were known. However, if the same model is fitted by MW, the theoretical optimum is almost achieved in the limit (98.9%). As Figure 3 shows, the MW estimator succeeds in approximating the optimal decision rule for each type very closely despite the misspecified
The observations above illustrate the defining property of the MW estimator—conditional on model specification, it is designed to make the limiting value of forecaster’s welfare objective as large as any other method. It might be argued, however, that this property is of limited practical importance. After all, likelihood-based specification tests should asymptotically reveal the inadequacy of the linear model against the quadratic alternative with probability one. As this model is also consistently estimated by ML, the forecaster’s unconditional welfare optimum is asymptotically achieved—see the line “ML (w/test)” in Table 2. Another alternative is the use of a flexible nonparametric estimator that can recover $p(x)$ in large samples under general conditions. For example, as a benchmark, we implement the Nadaraya-Watson estimator of $E(Y|X = x) = 2p(x) - 1$ with normal kernels and bandwidths chosen by least squares cross validation. As the line entitled “Nonparametric” in Table 2 shows, $p(x)$ is again recovered in the limit and the welfare optimum is achieved.

Therefore, we also investigate the performance of the MW estimator in samples small enough so that misspecification becomes hard to detect. Furthermore, in a small sample the flexibility of nonparametric methods may well be offset by large bias and variance. The small sample exercise consists of 500 Monte Carlo repetitions. In each cycle we generate an estimation sample of size $N = 80$. With 80 observations a size 5% Lagrange multiplier test
Table 2: Monte Carlo simulation results

<table>
<thead>
<tr>
<th>Method</th>
<th>Asymptotic</th>
<th>Small sample (n = 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In-sample</td>
<td>Out-of-sample</td>
</tr>
<tr>
<td>( p(x) ) known</td>
<td>0.302=100.0</td>
<td>0.305=100.0</td>
</tr>
<tr>
<td>Nonparametric</td>
<td>0.301=99.7</td>
<td>0.342=112.1</td>
</tr>
<tr>
<td>ML (logit, linear)</td>
<td>0.274=90.6</td>
<td>0.290=95.1</td>
</tr>
<tr>
<td>ML (logit, w/test)</td>
<td>0.302=100.0</td>
<td>0.321=105.1</td>
</tr>
<tr>
<td>MW (logit, linear)</td>
<td>0.299=98.9</td>
<td>0.350=114.6</td>
</tr>
</tbody>
</table>

\( c_1 = 0.07, c_2 = 0.22, c_3 = 0.27, c_4 = 0.32, \) equal weights

Note: The figures give the value of the forecaster’s welfare objective. Asymptotic results are based on a single sample of size 15,000. Small sample results are averages over 500 Monte Carlo cycles.

(in effect, a \( t \)-test) rejects \( H_0 : \theta_2 = 0 \) only about 40% of the time. The method entitled “ML (logit, w/test)” takes specification testing into account: If in a given sample the linear logit specification is rejected against the quadratic alternative, then the ML estimator is applied to the correct (quadratic) model. Conversely, if the linear model is not rejected, then the ML estimator is applied to the linear model.

For each estimation sample we also generate an independent evaluation sample of size \( M = 5000 \). Given an estimate of \( p(x) \) computed from the estimation sample using some method, \( W_n(\cdot) \) is calculated in-sample (i.e., over the estimation sample with \( n = N \)) as well as out of sample (i.e., over the evaluation sample with \( n = M \)). As the evaluation sample is large, \( W_M(\cdot) \approx W(\cdot) \) in each cycle.

The small sample simulation results are also reported in Table 2. In estimating \( p(x) \) using the misspecified linear logit model, MW outperforms ML by a fairly large margin in terms of in-sample welfare. Specifically, ML achieves 95.1% in comparison with the case in which \( p(x) \) is assumed to be known, while MW achieves 114.6%. That MW outperforms ML in-sample for a given parametric specification is a necessary consequence of the definition of the MW estimator. More importantly, the effect prevails out of sample as well, though the difference is quite a bit smaller. (Whether 86.6% vs. 90.7% is an economically relevant difference cannot really be judged outside the context of a concrete application.) In any case, the extent to which the MW estimator of the linear model asymptotically outperforms the ML
estimator appears to be large enough for the effect to hold up with just 80 observations for estimation. In fact, even if the possibility of specification testing is taken into account, the MW estimate of the linear model still outperforms ML out of sample, though the difference between the two methods is smaller still (88.3% vs. 90.7%).

Note that MW estimator of the linear model overfits the data in that it delivers higher in-sample welfare than \( p(x) \) by more than 14%. The nonparametric estimator also overfits at 112.1%. Of course, out of sample both methods are inferior to the knowledge of \( p(x) \), but the MW estimator still outperforms the nonparametric method by a small margin (87.2% vs. 90.7%). (The nonparametric estimator is roughly as good as ML with testing.) One reason why the MW estimator of the linear model is capable of outperforming the nonparametric estimator in this example is because the limit of the estimator closely reproduces the optimal decision rule for all types despite the misspecified functional form (again, see Figure 3). If misspecification is severe enough so that this is no longer possible, then the nonparametric estimator will do better asymptotically, but in small samples the MW estimator might still come out on top.

6 Empirical illustration

Data and model specification In this section we use data on Coronary Heart Disease (CHD) incidence from the Framingham Offspring Heart Study to study the MW estimator. This data set has been widely employed in various heart disease risk assessment studies; see, e.g., Wilson et al. (1998). The data set consists of records of five medical exams, conducted on a large cohort of individuals (approx. 2500 males and 2500 females) over the span of twenty years. At each exam date a number of variables were recorded for participating individuals, including major risk factors for the development of CHD. We will use data on risk factors from Exam 1; our dependent variable is an indicator of whether an individual developed CHD within 16 years of Exam 1.

A set of standard risk factors used to predict CHD incidence consists of age, gender, smoking habits, presence of diabetes, systolic blood pressure (SBP), and cholesterol level. Cholesterol level entails two measurements: LDL (low density lipoprotein or "bad") choles-
terol is conducive to the development of CHD, while HDL (high density lipoprotein or “good”) cholesterol is an inhibitor. We will conduct our analysis conditional on gender and smoking habits, i.e. we estimate separate models for male smokers, male non-smokers, female smokers and female non-smokers. (The group “smokers” consists of current or former smokers at the time of Exam 1; “non-smokers” had never smoked.) We exclude diabetes from the analysis, as there are few individuals with this condition in the data set.

For each group we estimate predictive models of the form

\[ P(CHD = 1 \mid X) = F(\theta_0 + \theta_1 AGE + \theta_2 \frac{HDL}{LDL} + \theta_3 SBP), \]

where \( F(\cdot) \) is a c.d.f. and \( X = (AGE, LDL, HDL, SBP) \). We will consider two choices for \( F \): the logistic c.d.f. (logit model) and the Cauchy c.d.f. (“cauchit” model). While logit is by far the more common choice, there are no a priori reasons, other than convenience, for excluding other c.d.f.’s. Koenker and Yoon (2007) demonstrate that the choice of \( F \) is not inconsequential, and while logit and probit are often close, this does not mean that other c.d.f.’s do not potentially fit the data better. Another uncertain aspect of the model specification is the functional form through which HDL and LDL enter (10). We use the ratio of HDL to LDL as in Kinosian et al. (1994), while some researchers, e.g. Wilson et al. (1998), use (categorized) levels.

The public forecasting problem In order to implement the MW estimator, we must specify cutoffs and corresponding weights. While these are primitives for the MW estimator, we will not simply employ an arbitrary set of cutoffs/weights; rather we will motivate their choice by a “story” that casts CHD risk assessment as a public forecasting problem. We consider a government agency that maintains a website designed to help people decide whether to buy CHD insurance. Suppose in particular that people who visit the site can enter their risk factors into a heart disease risk calculator and, based on a version of model (10), the system returns the probability of a CHD event over the next 16 years. Individuals are then assumed to use this information in their purchase decision. (There are many actual examples of such web-based calculators hosted by various institutions—they are easily found by a Google search of the phrase “heart disease risk calculator”.)
Individual DMs are endowed with CRRA utility functions $v(z) = z^{1-\rho}/(1-\rho)$, defined over yearly income $z$, and with risk aversion coefficient $\rho = 2$. If someone suffers a heart attack in a given year, the cost of treatment is assumed to be $30,000. The premium for heart disease coverage is $1,000 per year. The possible consequences of buying/not buying insurance for a given year are summarized in Table 3. Individual $j$ with income $z_j$ will buy insurance if and only if the probability of a CHD event in that year is greater than the cutoff

$$\tilde{c}_j = \frac{v(z) - v(z - $1,000)}{v(z) - v(z - $30,000)}.$$ 

However, the reported probability is that of having a CHD event over a 16 year time horizon. Decision makers are assumed to convert their one-year probability cutoffs $\tilde{c}_j$ into 16-year cutoffs $c_j$ using the formula $c_j = 1 - (1 - \tilde{c}_j)^{16}$. This formula assumes that the reported 16-year CHD probability is distributed evenly over that time horizon with CHD occurring in a given year independently of other years.

To set up the agency’s social welfare objective, we generate 10 cutoffs corresponding to the midpoints of the bins of a discretized version of the U.S. income distribution conditional on having annual household income greater than $30,000. (See Table 5 in Appendix E.) For a given cutoff, the weight $\lambda$ is chosen to be proportional to \([relative\ frequency\ of\ bin] \times [v(m) - v(m - 30,000)]\), where $m$ is the midpoint of the bin. Hence the forecaster is assumed to weigh those individuals more heavily for whom the total net utility gain resulting from correct classification is higher. As the CRRA utility function is strictly concave, this puts more weight on low income DMs.

**Estimation exercise** We perform the following exercise. We take the four groups (male smokers/non-smokers, female smokers/non-smokers), and select a random estimation sample
and evaluation sample from each. (The two subsamples are disjoint.) We then estimate model (10) six different ways: logit/cauchit estimated by ML, LS, and MW, respectively. As a benchmark, we also estimate $P(CHD = 1 \mid X)$ using a Nadaraya-Watson nonparametric regression with normal kernels and bandwidths chosen by least squares cross validation. The entire exercise is repeated 100 times.

We report the estimation results for male and female smokers in Table 4. (Results for the two other groups are similar.) The estimation and evaluation samples consisted of 800 and 500 observations, respectively. The reported numbers give the average value of the forecaster’s social welfare, both in-sample (i.e., over the estimation sample) and out-of-sample (i.e., over the evaluation sample).

Looking at the in-sample results, we see that the MW estimator delivers the highest welfare among all parametric estimators in both groups. Of course, this simply reflects the definition of the MW estimator. Note that the MW estimates of the cauchit and logit models achieve almost the same welfare, while there is a considerable gap between the ML/LS estimates of these two specifications. Specifically, ML and LS are both closer to MW for logit than for cauchit. This suggests that the logit functional form is a better approximation to $p(x)$ than cauchit, and also shows the relative insensitivity of the MW estimator towards misspecification.

While the nonparametric estimator yields higher in-sample welfare than any of the parametric ones, comparing with out-of-sample figures shows that this is mainly due to in-sample overfitting. The out-of-sample drop in welfare is by far the highest for the nonparametric estimator, followed by MW. For male smokers the nonparametric estimator ends up at the same level of welfare as MW, while for female smokers MW is clearly better. MW continues to outperform the ML/LS estimates of the cauchit model for both groups out-of-sample, offering further proof that this specification of $F$ is not fully appropriate. On the other hand, the ML/LS estimate of logit outperforms MW for male smokers, and is roughly as good as MW for female smokers. These out-of-sample results offer further evidence that model (10) with the logistic c.d.f. must be a reasonably good overall approximation to $p(X) = P(CHD = 1 \mid X)$ for these groups. Note, however, that for the MW estimator logit
Table 4: Estimation results for model (10)

<table>
<thead>
<tr>
<th>Method</th>
<th>Logit</th>
<th>Cauchit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NP</td>
<td>ML</td>
</tr>
<tr>
<td>In-sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male smokers</td>
<td>0.217</td>
<td>0.187</td>
</tr>
<tr>
<td>Female smokers</td>
<td>0.209</td>
<td>0.190</td>
</tr>
<tr>
<td>Out-of-sample</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male smokers</td>
<td>0.173</td>
<td>0.180</td>
</tr>
<tr>
<td>Female smokers</td>
<td>0.175</td>
<td>0.182</td>
</tr>
</tbody>
</table>

Note: The figures give the average value of the forecaster’s welfare objective.

versus cauchit makes virtually no difference.

These estimation results demonstrate the delicate interaction between model specification, method of estimation, and the goal of maximizing an objective of the form (6). One way of ensuring that (6) is (asymptotically) maximized is to ensure that the proposed model of \( p(x) \) is correctly specified. This underscores the importance of specification testing. However, in practice the number of alternative models against which the “null specification” could be tested may be very large. Furthermore, specification tests may not be powerful in certain directions, and there may not even exist a readily available test against reasonable alternatives.

So if a forecaster with a given welfare objective arrives at a candidate model for \( p(x) \) after theoretical considerations and/or conducting tests, it is still worthwhile to compare traditional estimators of the model to the MW estimator using an exercise similar to the one presented here. Doing so is in fact tantamount to a specification test. While this test is informal in that the relevant asymptotic distribution theory has not been worked out, it has the advantage that it tests precisely for those aspects of the model specification that affect the maximization of the given social welfare objective.
7 Conclusion

In this paper we defined and formalized the notion of public forecasting in an environment where there are two possible states of the world and many potential forecast users facing binary decisions. Motivated by the forecaster’s objective in such problems, we introduced the maximum welfare (MW) estimator for parametric conditional probability models. We derived conditions under which (non-linear) LS and ML are asymptotically MW estimators. Somewhat trivially, correct model specification is a sufficient condition for this. More interestingly, LS coincides with MW, regardless of possible model misspecification, if and only if uniform weight is given to all cutoff values in the (0, 1) interval. The same is true for ML if cutoffs at the extremes of this interval are weighted more heavily.

The paper has a positive as well as several normative interpretations. On the positive side, we showed that if a researcher writes down a parametric conditional probability model and estimates it using ML or LS, then, unless the model is correctly specified, the choice of the estimator makes implicit assumptions about the composition of forecast users and/or imposes specific value judgements concerning different types of decision makers. The positive contribution is making these implicit features of parametric estimators explicit.

On the normative side, there are a number of points that flow from the results in the paper. The first point concerns ML vs. LS estimation. Our results show that the LS estimator is easier to interpret than the ML estimator. The LS estimator is theoretically attractive in situations where (i) there is model uncertainty; (ii) the forecaster wants to treat all forecast users equally, and (iii) has an “ignorance prior” about the distribution of cutoffs. It is harder to rationalize the use of the ML estimator. Second, if the forecaster wants to use specific information about the distribution of cutoffs or wants to treat different types of forecast users differently, then under model uncertainty the MW estimator is a useful tool for implementing these goals. The third point concerns the importance of specification testing. If the conditional probability model is indeed correctly specified up to a finite dimensional vector of parameters, then any consistent estimator is socially optimal for any set of welfare weights. In large samples, when specification tests have good power, the advantage of the MW estimator might vanish. However, in finite samples the MW estimator might still do
better than ML or LS if there is a non-trivial amount of undetected misspecification.

Acknowledgements

Comments by the Associate Editor and two anonymous referees led to substantial improvements in the paper. We thank Ross Starr for numerous helpful discussions. We further benefitted from comments by Graham Elliott, Andra Ghent, Clive Granger, David Kaplan, Ulrich Mueller, Marco Ottaviani, Tridib Sharma, Joel Sobel and Chris Woodruff. We thank Majid Ezzati for providing the CHD data, and Jeff Racine for his code “N” for nonparametric estimation. All remaining errors are our responsibility.

References


Appendix

A. Proof of Proposition 1  By the law of iterated expectations and the fact that \(E(Y \mid X) = 2p(X) - 1\), we can write
\[
E\{(Y + 1 - 2c)\text{sign}[g(X) - c]\} = 2E\{[p(X) - c]\text{sign}[g(X) - c]\}.
\]
Further, the inequality
\[
|p(x) - c| = |p(x) - c|\text{sign}[p(x) - c] \geq |p(x) - c|\text{sign}[g(x) - c].
\]
is valid for all \(c \in (0, 1)\), \(x \in \mathcal{X}\) and \(g \in \mathcal{G}\). Integrating first with respect to the distribution of \(X\) and then with respect to the density \(\lambda(c)\) preserves this inequality. □

B. Proposition: Uniqueness of \(\lambda^\parallel\)  Suppose that there exist densities \(\lambda_i, i = 1, 2\) defined on \((0, 1)\), and strictly increasing functions \(g_i, i = 1, 2\) such that
\[
(i) \int_0^1 \lambda_i(c)E\{(Y + 1 - 2c)\text{sign}[m(X, \theta) - c]\}dc = g_i(Q(\theta)) \text{ for all } \theta \in \Theta \text{ whenever } Q(\theta) \text{ is well-defined over } \Theta;
\]
\[
(ii) \lambda_i \text{ and } g_i \text{ are independent of the model } \{m(\cdot, \theta) : \theta \in \Theta\} \text{ and the distribution of } (Y, X);
\]
\[
(iii) g_1 \text{ is invertible.}
\]
Suppose further that
\[
(iv) \text{ for the 'constant model' } m(x, \theta) \equiv \theta, \theta \in [0, 1], \text{ the objective } Q(\theta) \text{ is well-defined over } [0, 1] \text{ for any joint distribution of } (Y, X).
\]
Then \(\lambda_1 = \lambda_2 \text{ a.e.} \) (and so \(g_1 = g_2\)).

Proof: Let \(W_i(\theta) = \int_0^1 \lambda_i(c)E\{(Y + 1 - 2c)\text{sign}[m(X, \theta) - c]\}dc, i = 1, 2\), and write \(Q(\theta) = g_1^{-1}(W_1(\theta))\).

By the definition of \(\lambda_1\) and \(\lambda_2\), the equation
\[
W_2(\theta) = g_2(g_1^{-1}(W_1(\theta)))
\]
holds as an identity in \(\theta\) for any model specification and data generating process for which \(Q(\cdot)\) exists.

By assumption \((iv)\), we can consider the model specification \(m(x, \theta) = \theta, \theta \in [0, 1]\). Noting that \(\text{sign}[\theta - c] = 1_{[0, \theta]}(c) - 1_{[\theta, 1]}(c)\), it is easy to verify that \(W_i\) reduces to
\[
W_i(\theta) = 2p \left[\int_0^\theta \lambda_i(c)dc - \int_0^1 \lambda_i(c)dc\right] - 2 \left[\int_0^\theta c\lambda_i(c)dc - \int_0^1 c\lambda_i(c)dc\right],
\]
where \(p = P(Y = 1)\). In particular,
\[
W_i(0) = -2p + 2M_i \text{ and } W_i(1) = 2p - 2M_i,
\]
where \(M_i \in (0, 1)\) is the mean of \(\lambda_i\). Let \(h = g_2 \circ g_1^{-1}\). Substituting \((13)\) into \((11)\) yields
\[
-2p + 2M_2 = h(-2p + 2M_1) \quad \text{and} \quad 2p - 2M_2 = h(2p - 2M_1).
\]
By assumption \((iv)\), these equations must hold regardless of the data generating process; in particular, they must hold for all \(p \in [0, 1]\). Letting \(x = -2p + 2M_1\), we have, for \(x \in [-2 + 2M_1, 2M_1]\),

\[
h(x) = x - 2M_1 + 2M_2 \quad \text{and} \quad h(-x) = -x + 2M_1 - 2M_2.
\]

Substituting \(x = 0\) shows \(-2M_1 + 2M_2 = 2M_1 - 2M_2 = 0\), i.e. \(M_1 = M_2 := M\). Hence,

\[
h(x) = x \quad \text{for} \quad x \in [-2 + 2M, 2M],
\]

so that by (11) and the definition of \(h\),

\[
W_2(\theta) = W_1(\theta) \quad \text{for all} \quad \theta \quad \text{for which} \quad W_1(\theta) \in [-2 + 2M, 2M].
\]

Suppose that \(M \leq 1/2\), and let \(p = 0\). It is clear from (12) that in this case \(W_1(\theta)\) is decreasing in \(\theta\) with \(W_1(0) = 2M\) and \(W_1(1) = -2M \geq -2 + 2M\). Therefore \(W_1(\theta) \in [-2 + 2M, 2M]\) for all \(\theta \in [0, 1]\), which implies \(W_1(\theta) = W_2(\theta)\) for all \(\theta \in [0, 1]\). This is only possible if \(\lambda_1 = \lambda_2\) on \((0, 1)\) almost everywhere. Now suppose that \(M > 1/2\), and let \(p = 1\). It is clear from (12) that in this case \(W_1(\theta)\) is increasing in \(\theta\) with \(W_1(0) = -2 + 2M\) and \(W_1(1) = 2 - 2M < 2M\). Therefore \(W_1(\theta) \in [-2 + 2M, 2M]\) for all \(\theta \in [0, 1]\), which implies \(W_1(\theta) = W_2(\theta)\) for all \(\theta \in [0, 1]\). This is only possible if \(\lambda_1 = \lambda_2\) on \((0, 1)\) almost everywhere. Thus, we must have \(\lambda_1 = \lambda_2\) on \((0, 1)\) almost everywhere regardless of the value of \(M\). □

C. Proof of Proposition 2

Under Assumption 2(c), \(Q(\theta)\) and \(W(\theta)\) are well-defined for all \(\theta\). Recall that \(P(Y = 1 \mid X) = p(X)\) and \(P(Y = -1 \mid X) = 1 - p(X)\). We rewrite \(W(\theta)\) in the following way:

\[
W(\theta) = \int_0^1 \lambda(c) E_X \left( E \left[ \left| Y + 1 - 2c \right| \text{sign}(m(X, \theta) - c) \right] \right) dc
\]

\[
= 2 \int_0^1 \lambda(c) E_X \left\{ [p(X) - c] \text{sign}(m(X, \theta) - c) \right\} dc
\]

\[
= 2E_X \int_0^{m(X, \theta)} [p(X) - c] \lambda(c) dc - 2E_X \int_{m(X, \theta)}^1 [p(X) - c] \lambda(c) dc
\]

\[
= E_X \left\{ 2 \left[ \int_0^{m(X, \theta)} \lambda(c) dc - \int_{m(X, \theta)}^1 c \lambda(c) dc \right] p(X) \right\} - E_X \left\{ 2 \left[ \int_0^{m(X, \theta)} c \lambda(c) dc - \int_{m(X, \theta)}^1 c \lambda(c) dc \right] \right\}.
\]

The interchanging of the two integrals on the second line is justified by Fubini’s theorem. We can rewrite \(Q(\theta)\) in an analogous way:

\[
Q(\theta) = -E \left[ (\hat{Y} - m(X, \theta))^2 \right] = E \left\{ [2m(X, \theta) - 1] p(X) \right\} - E \left\{ [m(X, \theta)]^2 \right\}.
\]

Setting \(\lambda(c) = 1_{(0,1)}(c)\) gives

\[
\int_0^{m(X, \theta)} \lambda(c) dc - \int_{m(X, \theta)}^1 \lambda(c) dc = 2m(X, \theta) - 1 \quad \text{and}
\]

\[
\int_0^{m} c \lambda(c) dc - \int_{m}^1 c \lambda(c) dc = [m(X, \theta)]^2 - 1/2.
\]
Substituting (16) and (17) into (14) and (15) shows that $W(\theta) = 2Q(\theta) + 1$. □

D. Proof of Proposition 3 The proof is given in two lemmas. First we introduce notation. Given $A \subset \mathbb{R}$ define

$$W(\theta)_{m \in A} = \int_0^1 \lambda(c)E \left\{ (Y + 1 - 2c) \text{sign}(m(X, \theta) - c)|1_A(m(X, \theta)) \right\}dc \quad \text{and}$$
$$L(\theta)_{m \in A} = \frac{1}{2}E \left\{ \left[ (1 + Y) \log[m(X, \theta)] + (1 - Y) \log[1 - m(X, \theta)] \right]|1_A(m(X, \theta)) \right\}.$$

**Lemma 1** Let $\epsilon > 0$ and $I_\epsilon = [\epsilon, 1 - \epsilon]$. For $A_\epsilon$, $B_\epsilon$ and $W_{\epsilon ML}(\theta)$ given in Proposition 3,

$$W_{\epsilon ML}(\theta)_{m \in I_\epsilon} = A_\epsilon L(\theta)_{m \in I_\epsilon} + B_\epsilon.$$

**Proof:** Under Assumption 2(c) and 3, $W(\theta)$ and $L(\theta)$ are well-defined for all $\theta$. Consider the decomposition of $W(\theta)$ given in (14) and (15). Using the law of iterated expectations and the fact that $E(Y | X) = 2p(X) - 1$, we rewrite $L(\theta)$ in an analogous way:

$$L(\theta) = E \left\{ \log[m(X, \theta)] - \log[1 - m(X, \theta)] \right\}p(X) \quad \text{(18)}$$
$$+ E \left\{ \log[1 - m(X, \theta)] \right\}. \quad \text{(19)}$$

Suppose for the moment that we could find a density $\lambda(c)$ so that for any fixed number $m \in (0, 1)$ it is true that

$$2 \left[ \int_0^m \lambda(c)dc - \int_0^1 \lambda(c)dc \right] = A[\log(m) - \log(1 - m)] + B_1 \quad \text{and} \quad \text{(20)}$$
$$2 \left[ \int_0^m c\lambda(c)dc - \int_0^1 c\lambda(c)dc \right] = -A \log(1 - m) + B_2, \quad \text{(21)}$$

where $A > 0$, $B_1$ and $B_2$ are constants independent of $m$. These relationships would then imply that $W(\theta) = AL(\theta) + B$ holds with $B = B_1E[p(X)] - B_2$. However, note that the LHS of (20) converges to $-2$ as $m \to 0$, while the RHS converges to $-\infty$. A similar problem arises when $m \to 1$. Equation (21) also suffers from the same defect. We work around this problem by requiring that the choice of $\lambda(c)$ make (20) and (21) true only for $m \in [\epsilon, 1 - \epsilon] = I_\epsilon$, where $\epsilon$ is a pre-specified small positive number. In other words, we only require $W(\theta)_{m \in I_\epsilon} = AL(\theta)_{m \in I_\epsilon} + B$. In choosing $\lambda(c)$ we can then restrict attention to densities supported on $[\epsilon, 1 - \epsilon]$ without loss of generality.

Taking the derivative of (20) with respect to $m$ yields

$$4\lambda(m) = \frac{A}{m(1 - m)}, \quad m \in (\epsilon, 1 - \epsilon).$$

Given that $\lambda(m)$ is a density with support $(\epsilon, 1 - \epsilon)$, substituting $m = \epsilon$ and $m = 1 - \epsilon$ into (20) yields the following two equations:

$$A[\log(\epsilon) - \log(1 - \epsilon)] + B_1 = -2$$
$$A[\log(1 - \epsilon) - \log(\epsilon)] + B_1 = 2.$$
Solving for $A$ and $B_1$ gives $A_\epsilon = 2[(\log(1-\epsilon) - \log(\epsilon))^{-1}$ and $B_{1\epsilon} = 0$. Hence, the desired solution to equation (20) is given by the density
\[
\lambda^{ML}_\epsilon(c) = \frac{1_{(\epsilon,1-\epsilon)}(c)}{2[(\log(1-\epsilon) - \log(\epsilon))\epsilon(1-\epsilon)].}
\]
Of course, we still need to verify that $\lambda^{ML}_\epsilon(c)$ and $A_\epsilon$ make equation (21) identically true for $m \in [\epsilon, 1-\epsilon]$ and some choice of $B_2$ independent of $m$. Substituting $A_\epsilon$ and $\lambda^{ML}_\epsilon(c)$ directly into (21) shows that if one chooses
\[
B_{2\epsilon} = \frac{\log(\epsilon) + \log(1-\epsilon)}{\log(\epsilon) - \log(1-\epsilon)},
\]
then (21) indeed becomes an identity in $m$ over $[\epsilon, 1-\epsilon]$. This establishes $W^{ML}_\epsilon(m)_{m \in \ell_{\epsilon}} = A_{\epsilon}L(\theta)_{m \in \ell_{\epsilon}} + B_{\epsilon}$ with $B_{\epsilon} = -B_{2\epsilon}$. □

**Lemma 2** For $A_\epsilon$, $B_\epsilon$ and $W^{ML}_\epsilon(\theta)$ given in Proposition 3,
\[
\sup_{\theta \in \Theta} \left| W^{ML}_\epsilon(\theta) - (A_{\epsilon}L(\theta) + B_{\epsilon}) \right| \to 0 \text{ as } \epsilon \to 0.
\]

**Proof:** Clearly, we have
\[
W^{ML}_\epsilon(\theta) = W^{ML}_\epsilon(\theta)_{m \in \ell_{\epsilon}} + W^{ML}_\epsilon(\theta)_{m \in \ell'_{\epsilon}} \quad \text{and} \quad L(\theta) = L(\theta)_{m \in \ell_{\epsilon}} + L(\theta)_{m \in \ell'_{\epsilon}},
\]
where $I_{\epsilon} = [\epsilon, 1-\epsilon]$ and the superscript $c$ denotes complementation. By Lemma 1,
\[
W^{ML}_\epsilon(\theta)_{m \in \ell_{\epsilon}} = A_{\epsilon}L(\theta)_{m \in \ell_{\epsilon}} + B_{\epsilon}.
\]
Combining this with the decomposition of $W^{ML}_\epsilon(\theta)$ and $L(\theta)$ yields
\[
W^{ML}_\epsilon(\theta) = (A_{\epsilon}L(\theta) + B_{\epsilon}) = W^{ML}_\epsilon(\theta)_{m \in \ell'_{\epsilon}} - A_{\epsilon}L(\theta)_{m \in \ell'_{\epsilon}}
\]
so that
\[
\sup_{\theta} \left| W^{ML}_\epsilon(\theta) - (A_{\epsilon}L(\theta) + B_{\epsilon}) \right| \leq \sup_{\theta} \left| W^{ML}_\epsilon(\theta)_{m \in \ell'_{\epsilon}} \right| + A_{\epsilon} \sup_{\theta} \left| L(\theta)_{m \in \ell'_{\epsilon}} \right|.
\]
Assumption 3 implies $\left| L(\theta)_{m \in \ell'_{\epsilon}} \right| \to 0$ as $\epsilon \to 0$, uniformly in $\theta$. In addition, $A_{\epsilon} \to 0$ as $\epsilon \to 0$. Finally,
\[
\sup_{\theta} \left| W^{ML}_\epsilon(\theta)_{m \in \ell'_{\epsilon}} \right| \leq 2 \sup_{\theta} P[\epsilon \leq m(X, \theta) \leq 1-\epsilon],
\]
which also goes to zero as $\epsilon \to 0$ by Assumption 3. □

**E. Cutoffs and weights derived from the U.S. income distribution** Table 5 presents the cutoffs/weights we used in implementing the MW estimator in Section 6.

The underlying income distribution is based on the 2007 Current Population Survey. The source data can be found at the URL
\[
http://pubdb3.census.gov/macro/032007/hhinc/new06_000.htm
\]
We took the income distribution for all races, conditioned on having annual household income of at least $30,000$, and consolidated bins to have length $10,000$. We censored the distribution at $200,000$.  

30
Table 5: Cutoffs and weights used in Section 6

<table>
<thead>
<tr>
<th>Annual inc. (Thousands)</th>
<th>Rel. freq.</th>
<th>1 yr. cutoff</th>
<th>16 yr. cutoff</th>
<th>λ</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-40</td>
<td>0.1474</td>
<td>0.0049</td>
<td>0.0756</td>
<td>0.6824</td>
</tr>
<tr>
<td>40-50</td>
<td>0.1318</td>
<td>0.0114</td>
<td>0.1671</td>
<td>0.1582</td>
</tr>
<tr>
<td>50-60</td>
<td>0.1175</td>
<td>0.0154</td>
<td>0.2203</td>
<td>0.0692</td>
</tr>
<tr>
<td>60-70</td>
<td>0.0983</td>
<td>0.0182</td>
<td>0.2550</td>
<td>0.0350</td>
</tr>
<tr>
<td>70-80</td>
<td>0.0860</td>
<td>0.0203</td>
<td>0.2794</td>
<td>0.0206</td>
</tr>
<tr>
<td>80-90</td>
<td>0.0711</td>
<td>0.0218</td>
<td>0.2975</td>
<td>0.0123</td>
</tr>
<tr>
<td>90-100</td>
<td>0.0582</td>
<td>0.0230</td>
<td>0.3114</td>
<td>0.0076</td>
</tr>
<tr>
<td>100-150</td>
<td>0.1745</td>
<td>0.0255</td>
<td>0.3389</td>
<td>0.0119</td>
</tr>
<tr>
<td>150-200</td>
<td>0.0628</td>
<td>0.0278</td>
<td>0.3628</td>
<td>0.0020</td>
</tr>
<tr>
<td>200-</td>
<td>0.0523</td>
<td>0.0295</td>
<td>0.3802</td>
<td>0.0008</td>
</tr>
</tbody>
</table>

*Note: Cutoffs are calculated using bin midpoints. For the last bin cutoff is calculated using $250,000. This is roughly the median of the last bin.*