Statistica Sinica **10**(2000), 517-544

THE DERIVATION OF A LATENT THRESHOLD INSTRUMENTAL VARIABLES MODEL

Mark E. Glickman and Sharon-Lise T. Normand

Boston University and Harvard Medical School

Abstract: Measuring the causal effect of a treatment from observational data is often difficult because the treatment status of a subject may be confounded with non-randomized factors, such as those that affect a subject's choice of treatment. An approach to remedying this problem is through the use of instrumental variables. We extend the instrumental variables framework proposed by Angrist, Imbens and Rubin (1996) by introducing a latent "threshold to receive treatment" parameter for each unit in the study. Incorporation of latent thresholds in the model permits the inclusion of discrete or continuous instruments, covariate information, and flexible choices of distributions. We apply our methodology to examine the effect of cardiac catheterization on short-term survival of a cohort of elderly heart attack patients.

Key words and phrases: Acute myocardial infarction, causal inference, iterative simulation, observational studies, treatment effect.

1. Introduction

A major difficulty in measuring treatment effects in non-randomized observational studies is that a subject's treatment status may depend on confounding factors that can also affect the subject's response. An approach to address this problem involves the use of instrumental variables techniques. The main difference between instrumental variable methods and covariate adjustment approaches is that the former assumes that the treatment assignment is non-ignorable, and requires the presence of an "instrument" in order to reduce or remove confounding bias. An instrumental variable is designed to balance unobserved covariates across treatment groups so that the mechanism by which units obtain values of an instrument may be considered ignorable. In addition to balancing unobserved covariates, an instrumental variable is assumed to covary with treatment status. Identifying a variable satisfying these conditions can often be difficult. In some instances, an instrumental variable can be found that is explicitly randomized to units, in which case the instrument clearly balances unobserved covariates. For example, randomized studies confounded with treatment non-compliance may be analyzed by employing treatment assignment as an instrument (Rubin (1998),

Imbens and Rubin (1997a)). However, most applications of instrumental variable techniques must be argued on a case-by-case basis.

Exploration of the instrumental variables approach within the statistics community has been attracting more attention. Nagelkerke, Fidler and Buwalda (1988) use an instrumental variable approach to make inferences about disease statuses of patients from diagnostic tests. Angrist and Krueger (1992) use instrumental variables to measure the effect of age at school entry on educational attainment. Stefanski and Buzas (1995) examine an instrumental variable approach to binary response regression. Angrist, Imbens and Rubin (1996) and Imbens and Rubin (1997a) present a more foundational approach and lay out assumptions to develop models with binary instrumental variables for inferring causal effects.

This article develops a general framework for Bayesian inference for a particular class of instrumental variable models. Our model, which is founded on the assumptions laid out in Angrist, Imbens and Rubin (1996), specifies the effect of the instrumental variable through a conditional structure via unobserved latent variables. As we subsequently describe, these latent variables have the natural interpretation of being the individuals' "thresholds" to receive treatment, and the inclusion of these latent variables extends the work of Angrist, Imbens and Rubin to permit greater modeling flexibility while still retaining their basic assumptions. For example, our framework allows for multi-valued instrumental variables, a variety of distributional assumptions for the data, and incorporation of covariate data. We note, however, that with greater flexibility in modeling, there is greater room for model misspecification and overfitting. Thus model diagnostics and selection are particularly important components in our framework.

We present our modeling framework for inferring causal effects with instrumental variables in Section 2. We show how the assumptions of Angrist, Imbens and Rubin can be combined into a parametric model by the inclusion of a latent threshold parameter for each unit. Innovations in computational methods for model fitting through Markov chain Monte Carlo simulation, and in particular the Gibbs sampler (e.g., Gelfand and Smith (1990)), permit straightforward inferential procedures. We apply our approach in Section 3 to measure the effect of cardiac catheterization on a cohort of elderly heart-attack patients. We discuss limitations and extensions of our model in Section 4.

2. A Latent Threshold Parameter Model

Our interest centers on measuring the causal effect of a dichotomous treatment, denoted D, on a response, denoted Y, from an observational study involving n subjects. For the development of our model, three variables (and possibly

covariates) are observed for each subject: the binary treatment status D_i for subject i, with $D_i = 0$ when subject i is exposed to control and $D_i = 1$ when subject i is exposed to treatment; an instrumental variable Z_i for subject i; and one of two potential responses, Y_{i1} and Y_{i0}, corresponding respectively to the response when subject i is exposed to treatment or to control. Under our framework, at most one of Y_{i1} or Y_{i0} can be observed, so the unobserved response (or, more precisely, averages of unobserved responses) must be inferred in the model fitting process. Both sets of outcome variables, Y_{i1} and Y_{i0}, and the instrumental variable, Z_i, can be discrete, continuous, or mixtures of discrete and continuous. The difficulty in measuring the causal effect of treatment from observational data is that a subject's treatment status, D_i, may depend on factors that are related to the subject's potential responses Y_{i1} and Y_{i0}, so that an analysis which ignores these confounding factors will result in incorrect inferences. Intuitively, the instrumental variable, Z_i, can often be thought of as taking on randomly assigned values (as if the Z_i themselves were randomized), but having a strong relationship to treatment status. Instrumental variable techniques remove or reduce confounding by projecting the Y_{i1}, Y_{i0}, D_i and the covariates into the space spanned by the Z_i.

We make four assumptions commonly used in instrumental variable models. The first is the Stable Unit Treatment Value Assumption (SUTVA), as described by Rubin (1978, 1980, 1990). Under this assumption, observed or potentially observed values for a subject are unaffected by those of any other subject. The second assumption, often termed "exclusion restriction", is that potential responses Y_{i0} and Y_{i1} do not depend on the value of the instrumental variable Z_i. This assumption is discussed in econometric literature on instrumental variables, and more recently in Angrist, Imbens and Rubin (1996) Viewing treatment status, $D_i(Z_i)$, as a function of Z_i, the third assumption is the monotonicity of D_i with respect to Z_i. We assume that for any two potentially observed values of Z_i, $z_1^* < z_2^*$, that $D_i(z_1^*) \leq D_i(z_2^*)$ for all i. This assumption asserts a particular functional relationship between Z_i and D_i that ensures the identifiability of treatment effects. Other assumptions can be used in place of monotonicity (e.g., constant treatment effect) as required. Finally, we assume that the mechanism generating the Z_i is ignorable (Rubin (1978)). Thus, the Z_i are assumed to be independent of unobserved information.

The four underlying assumptions can be unified into a single framework by making use of the following key idea. For monotonicity to hold, a subject will be exposed to "treatment" when the subject's value of Z_i is high, and will be exposed to "control" when Z_i is low, assuming a subject is willing to belong to either group. At some point between these extremes, a subject must possess

a "threshold" value that partitions the values of Z into those that result in the subject's exposure to the treatment and those that result in the subject's exposure to the control. In this formulation of the problem, the exposure status depends exclusively on whether Z_i is greater or less than the subject's threshold. A threshold for one subject can clearly be different from another, and potential responses Y_{i0} and Y_{i1} will typically depend on a subject's threshold. It is the introduction of a subject's threshold parameter into the modeling framework that permits a characterization of a large and flexible class of models for causal effects using instrumental variables.

More formally, let Γ_i be the latent threshold parameter for subject i that determines treatment status given the value of Z_i. Conditional on Γ_i, the treatment status of subject i is given by

$$D_i = \begin{cases} 1, & \text{if } Z_i \geq \Gamma_i \\ 0, & \text{if } Z_i < \Gamma_i. \end{cases}, \tag{1}$$

The model in (1) is equivalent to the definition of monotonicity described earlier under a mild regularity condition. To prove equivalence, it is straightforward to see (1) implies that for any $z_1^* < z_2^*$ and conditional on any fixed $\Gamma_i = \gamma$, $D_i(z_1^*) \leq D_i(z_2^*)$. The converse can be shown by construction, making the mild assumption that D_i is right-continuous with respect to z. We can then choose $\Gamma_i = \inf\{z : D_i = 1\}$, which satisfies the conditions of our definition. If unit i would never be exposed to treatment, then we set $\Gamma_i = \infty$; if unit i would always be exposed to treatment, then we set $\Gamma_i = -\infty$. Note that in many cases Γ_i need not be uniquely defined, particularly when Z is a discrete variable, but this poses no difficulties in the development of the model.

The Γ_i can be viewed not just as parameters in a latent probability model, but as potentially observable quantities. If, in an experimental setting, one could observe for subject i many values of the instrumental variable along with the resulting treatment status, then Γ_i could be determined by noting the value of the instrumental variable at which treatment status switches. Thus the Γ_i are interpretable as quantities that could be known if appropriate data were collected on individuals. In typical studies, where only a single value of Z_i is observed for each subject, the Γ_i would need to be inferred in the model fitting process. Specifically, once Z_i and D_i are observed for unit i, (1) constrains the value of Γ_i to be less than or equal to Z_i (if $D_i = 1$ was observed), or greater than Z_i (if $D_i = 0$ was observed).

Our basic model makes the following three general distributional assumptions.

$$(Y_{i0}, Y_{i1}) | D_i, \Gamma_i, Z_i, x_i, \pi \sim G(y_0, y_1 | D_i, \Gamma_i, x_i, \pi) \tag{2}$$

$$D_i = \begin{cases} 1, & \text{if } Z_i \geq \Gamma_i, \\ 0, & \text{if } Z_i < \Gamma_i. \end{cases} \tag{3}$$

$$\Gamma_i | Z_i, u_i, \pi \sim p(\gamma | Z_i, u_i, \pi), \tag{4}$$

where x_i is covariate data in the response model for subject i, u_i is covariate data (possibly overlapping with x_i) in the threshold model for subject i, π is a vector of model parameters, and G and p are assumed probability distributions. We also assume the Z_i are generated by an ignorable data mechanism,

$$Z_i | u_i, \eta \sim q(z | u_i, \eta), \tag{5}$$

where q is an assumed probability distribution, and the parameters η may or may not be known in advance. In our modeling framework, we condition on the Z_i so that inferences about η (if η is unknown) are irrelevant.

The (joint) distribution in (2) relates the potential responses to both observed variables (treatment status, covariate data) and latent variables (threshold, other model parameters). The exclusion restriction assumption implies that this distribution does not depend on any of the Z_i. Because at least one of Y_{i1} or Y_{i0} will not be observed, its distribution can be integrated out. An important feature of the response model is that differences among subjects' treatment effects not already explained by observed covariate data x_i can be incorporated through the latent threshold parameter Γ_i. The role of Γ_i in the response model is therefore to mitigate the bias associated with unobserved variables. Our framework also permits the inclusion of observed covariate information in (4), the model for the threshold parameters. An important benefit of having the flexibility to model the threshold parameters is that covariates might provide information that allows the thresholds to be inferred with greater precision. The stronger the relationship, the more precisely the latent thresholds can be inferred. This in turn results in less uncertainty about the distribution for the potential responses, so more precise causal effects can be inferred. We explore threshold models with covariates in Section 3.2.

The model framework in (2), (3) and (4) explicitly addresses the SUTVA, the exclusion restriction of Z_i given D_i, and the monotonicity of D_i with respect to Z_i. With ignorability assumed for Z_i, our model can therefore be seen as an extension of the modeling framework of Angrist, Imbens and Rubin that allows for arbitrary distributional assumptions both at the data and threshold levels of the model, inclusion of covariates, arbitrary functional relationships between covariates and response, and so on. Some common examples of instrumental variable models for causal effects can be seen as special cases of our framework.

Example 1. Econometric program evaluation models

A class of econometric models that posit a latent variable has been considered by Heckman and Robb (1985) and Heckman and Hotz (1989), and discussed by Imbens and Angrist (1994). Such models have been used, for example, to determine the effect of manpower training on productivity or earnings. Letting Y_i denote the observed response, a linear response model assumes

$$Y_i = \beta_0 + D_i\beta_1 + X_i\beta_2 + \epsilon_i. \tag{6}$$

Here X_i is a matrix of covariates, ϵ_i is an error term centered at 0, and

$$D_i = \begin{cases} 1, & \text{if } D_i^* \geq 0, \\ 0, & \text{if } D_i^* < 0, \end{cases} \tag{7}$$

with

$$D_i^* = \alpha_0 + X_i\alpha_1 + Z_i\alpha_2 + \nu_i, \tag{8}$$

where ν_i is an error term centered at 0, Z_i is a scalar variable, and the Z_i are independent of ν_i and ϵ_i. In this model, Y_i might be the observed productivity for subject i, D_i would be an indicator of whether subject i was selected for a productivity training program, and X_i might be socio-demographic information about subject i. Letting

$$\Gamma_i = \frac{-\alpha_0 - X_i\alpha_1 - \nu_i}{\alpha_2}$$

and setting

$$D_i = \begin{cases} 1, & \text{if } Z_i \geq \Gamma_i, \\ 0, & \text{if } Z_i < \Gamma_i, \end{cases}$$

a special case of our model results. Compared to these linear latent variable models, our framework has several advantages. First, the latent variable model described above assumes that, conditional on the covariates, the treatment effect β_1 is constant. In contrast, our model does not assume a constant treatment effect. Differences among subjects due to unobserved variables as they relate to the response may be difficult to incorporate into (6) in a meaningful way. Secondly, the latent parameters in the more conventional models lack a natural interpretation which is retained by our framework. Finally, our framework does not restrict the models to particular functional forms and distributional assumptions which may be problematic to incorporate in latent variable models similar to that in (6), (7), and (8).

Example 2. Models with a binary instrument

Binary instrument models are commonly used to identify causal effects in many applications. Angrist (1990) examines the effect of Vietnam veteran status on income using draft lottery numbers (high/low) as an instrument. Imbens and Rubin (1997b) reanalyze data from a study by Angrist and Krueger (1991) on the effect of education on earnings, using season of birth (first or fourth quarters) as the instrument. The treatment non-compliance problem, in which the random assignment to treatment group may be considered an instrument, has been examined by Efron and Feldman (1991) who, from a randomized clinical trial with non-compliance, measure the effectiveness of a drug for lowering cholesterol levels. Imbens and Rubin (1997a) present a foundational development of using instrumental variables in randomized studies involving non-compliance. For concreteness, we assume that our binary instrument application involves a randomized study with non-compliance, though our discussion easily extends to other situations with binary instruments.

Assume the binary instrument, Z_i, is the treatment assignment variable where $Z_i = 0$ is assignment to the control group and $Z_i = 1$ is assignment to the treatment group. Following the development of Imbens and Rubin (1997a), the monotonicity assumption for a model with an instrument Z_i that takes on values 0 or 1 implies the existence of three types of subjects:

1. those for whom $D_i = 0$ always,
2. those for whom $D_i = c$ when $Z_i = c$ for $c = 0, 1$, and
3. those for whom $D_i = 1$ always.

The first type of subject never takes the treatment regardless of treatment assignment (a "never-taker"), the second type complies with treatment assignment (a "complier"), and the third type always takes the treatment regardless of treatment assignment (an "always-taker"). Incorporating the monotonicity restriction rules out the fourth type of individual who would do the opposite of the treatment assigned. This setup allows for four types of potential responses; Y_{i1} for always-takers, Y_{i0} for never-takers, Y_{i1} for compliers, and Y_{i0} for compliers. Causal inference can only be defined for compliers because only compliers can potentially be observed to take either treatment or control. It would not be meaningful to consider a causal effect for a noncomplier who could only ever be exposed to one treatment. It is for the compliers that an average causal effect is of interest. Imbens and Angrist (1994) refer to the causal effect for compliers as the "local average treatment effect" (LATE).

To map the problem into our framework, we assume that Γ_i can take on one of three values in the set $\{-0.5, 0.5, 1.5\}$. When $\Gamma_i = -0.5$, $\Gamma_i < Z_i$ for both possible values of Z_i, so subject i would be an always-taker. When $\Gamma_i = 1.5$,

$\Gamma_i > Z_i$ for both possible values of Z_i, so subject i would be a never-taker. Lastly, for $\Gamma_i = 0.5$, $D_i = 1$ when $Z_i = 1$, and $D_i = 0$ when $Z_i = 0$. This corresponds to subject i being a complier.

The four distributions of potential responses in our framework are

$$
\begin{aligned}
Y_{i1}|\Gamma_i = -0.5 &\sim G_1(y|D = 1, \Gamma_i = -0.5, \pi),\\
Y_{i0}|\Gamma_i = 0.5 &\sim G_0(y|D = 0, \Gamma_i = 0.5, \pi),\\
Y_{i1}|\Gamma_i = 0.5 &\sim G_1(y|D = 1, \Gamma_i = 0.5, \pi),\\
Y_{i0}|\Gamma_i = 1.5 &\sim G_0(y|D = 0, \Gamma_i = 1.5, \pi).
\end{aligned}
$$

Details for performing likelihood-based inference for such binary instrument models are described in Angrist, Imbens and Rubin (1996).

Inference for specific cases of our general framework can be performed using the method of moments, which is often implemented as a two-stage least squares procedure. Typical applications of the method of moments for instrumental variable estimators can be found in standard textbooks in econometrics, such as Bowden and Turkington (1984). Imbens and Angrist (1994) describe conditions under which a method of moments instrumental variable estimator can be equated to a weighted sum of local average treatment effects. Small sample properties of two-stage least squares estimators can be found, for example, in Buse (1992) and Phillips (1983).

Likelihood-based approaches, including Bayesian methods, have only been investigated in the context of simple models such as those described in Angrist, Imbens and Rubin (1996) with binary instruments, and in constant treatment effect models as in Heckman and Robb (1985). For our latent threshold framework, iterative simulation via the Gibbs sampler provides an important tool for model fitting. We demonstrate this approach in the following section, where we infer the causal effect of undergoing cardiac catheterization on mortality on elderly heart attack patients.

3. Effect of Cardiac Catheterization on Short-term Mortality

McClellan, McNeil and Newhouse (1994), hereafter MMN, and McClellan and Newhouse (1993) study the benefits for elderly patients of a diagnostic cardiac catheterization during the first or "index" admission for acute myocardial infarction (AMI). The data from their study included over 200,000 elderly Medicare patients discharged from a hospital with a principal diagnosis of AMI (all ICD-9-CM codes 410 except those with a 2 in the fifth position) in 1987. Their work used a generalized method of moments (GMM) instrumental variable approach (Chamberlain (1987)) through a two-stage least squares procedure to estimate a treatment effect from their data. We reexamine this question in the

context of our framework using a data set consisting of 3667 elderly Alabama residents admitted to an Alabama hospital in 1990 with a principal diagnosis of AMI.

3.1. Data description

For each patient in the data set, we have recorded whether a patient died within 30 days of admission to the hospital (Y), and whether a patient had undergone a cardiac catheterization during the index admission for the AMI (D). We also consider the following patient covariate information for modeling 30-day mortality: gender, race (black/non-black), age, and urban/rural status of a patient's residence (whether a patient's residence belongs to a Metropolitan Statistical Area). These data were obtained retrospectively from Medicare utilization claims data. We discuss other covariate information to include in the threshold model in Section 3.2. Table 1 shows the distribution of patient covariates and their relationship with 30-day mortality and frequency of undergoing cardiac catheterization.

Table 1. Distribution of patient covariates and their relationship to 30-day mortality and frequency of undergoing catheterization. Data were obtained from a cohort of 3667 elderly Medicare beneficiaries residing in Alabama discharged with a principal diagnosis of AMI in 1990.

Stratum	Sample Total	Percent Undergoing Catheterization	Percent Dead
Male	1791	37.2	23.6
Female	1876	26.5	26.5
Non-black	3165	33.1	25.1
Black	502	23.1	25.1
Age 65–69	792	52.1	16.4
Age 70–74	863	45.1	17.3
Age 75–79	832	30.0	22.7
Age 80–84	657	14.0	34.1
Age 85+	523	4.0	43.6
Urban	2535	36.4	23.6
Rural	1132	21.3	28.5
Total	3667	31.8	25.1

Approximately 32% of the sample underwent a cardiac catheterization, with younger patients having more frequent use of the procedure. Males, non-black and urban patients tend to undergo cardiac catheterizations more often than females, black and rural patients. The overall mortality rate for the sample was

25%, with older patients experiencing 30-day mortality substantially more often than younger patients.

We also have recorded the distance between a patient's residence and the nearest hospital equipped to perform cardiac catheterizations, and the distance between a patient's residence and the nearest hospital not equipped to perform cardiac catheterizations regularly (we treat a hospital as "equipped" to perform a cardiac catheterization if the hospital performed at least 5 catheterizations on AMI patients in 1990; this definition is also used by MMN). Distances are measured as the number of miles between the centroids of zip code areas. These distance measures will be used to construct the instrumental variable used in our analysis, which we describe below.

The goal of our analysis is to infer the expected difference in 30-day mortality when patients undergo catheterization versus when patients do not undergo catheterization. Clearly, causal inference is hindered by unobservable biases because the data were not the result of a randomized design. For example, patient severity on admission is an unobserved factor that may relate to both the decision to "treat" (catheterization presents greater risk to patients in more serious condition) and survival probability. To account for these confounding biases, we use an instrument similar to that described by MMN, and McClellan and Newhouse (1993). We define "differential distance" to be the distance from a patient's residence to the nearest hospital equipped to perform a catheterization subtracted from the distance from a patient's residence to the nearest hospital not equipped to perform a catheterization. Intuitively, the larger the differential distance, the more accessible the hospital with a catheterization facility compared to one without a catheterization facility, so the greater the chance a randomly selected patient would undergo a cardiac catheterization if the patient's differential distance is large than if it is small (or negative). Figure 1 displays the distribution of differential distances in our data set. The differential distances are between -68 and 22.4 miles, with a median differential distance at 0 miles.

MMN argue differential distance as an instrument both intuitively and empirically. They show that the distribution of virtually all available patient covariates stratified by differential distance groupings are similar. This adds credibility to the belief that using differential distance may substantially reduce the impact of hidden biases. As suggested above, monotonicity can be argued on the grounds that a patient not having undergone a catheterization would not have elected to undergo a catheterization if he or she lived even farther from a hospital equipped to perform a catheterization. The exclusion restriction assumption in the context of our example states that whether a patient would die, given both covariate information and whether a patient underwent a catheterization, is not affected by a patient's differential distance. The SUTVA most likely holds for our data

because the decision for one patient to undergo a catheterization is probably not directly affected by another patient's decision.

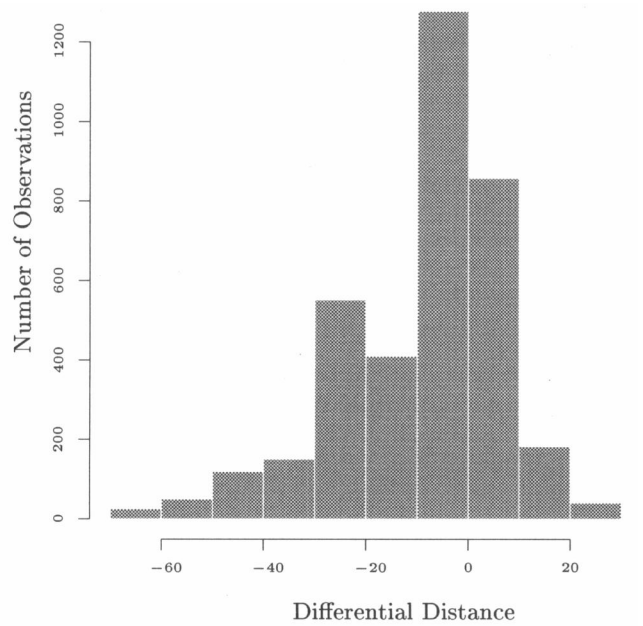

Figure 1. Empirical distribution of differential distances. Differential distance is defined as the difference between the distance a patient would need to travel to the nearest hospital equipped to perform a cardiac catheterization and the distance a patient would need to travel to the nearest hospital not equipped to perform a cardiac catheterization. Positive differential distances correspond to patients living closer to hospitals equipped to perform catheterizations. The lowest differential distance is −68 miles, and the largest is 22.4 miles. The median differential distance is 0 miles.

3.2. Modeling catheterization effect

In this section, we construct four instrumental variable models for measuring the effect of cardiac catheterization. Our models are four different variations of the framework established in (2), (3), and (4). For $i = 1, \ldots, 3667$, let

$$Y_i = \begin{cases} 1 & \text{if patient } i \text{ dies within 30 days of hospital admission,} \\ 0 & \text{otherwise,} \end{cases}$$

$$D_i = \begin{cases} 1 & \text{if patient } i \text{ underwent a catheterization,} \\ 0 & \text{otherwise,} \end{cases}$$

Z_i = Differential distance stratum for patient i,

x_i = Vector of mortality covariates for patient i,

Γ_i = Threshold distance for patient i.

Also, denote Γ to be the entire vector of Γ_i, D to be the vector of D_i, Y to be the vector of Y_i, Z to be the vector of Z_i, and X to be the matrix of x_i.

Model 1:

For the first model, we assume that the latent threshold variable, Γ_i, for patient i has a discrete distribution, taking on one of only seven values. The potential values of Γ_i are displayed in Table 2. These values were chosen so that the distribution of differential distance were relatively equally balanced between neighboring pairs of thresholds distances. The two outermost thresholds, corresponding to -80 miles and 30+ miles, represent always-takers and never-takers, respectively. For example, suppose a patient lives 20 miles closer to a catheterization facility than a non-catheterization facility and does not undergo a catheterization. Then this patient's threshold to receive treatment must be greater than 20. Among the possible threshold values, $\gamma_i = 30.1$ is the only possibility (this conclusion applies to the model where the γ_i are discrete variables). Because this is the highest threshold value, we would conclude that this patient would not undergo a catheterization at any differential distance. Such patients are likely to be so unhealthy that the risk operative mortality would be dangerously high.

Table 2. Threshold distances for Model 1. The latent variable Γ_i is assumed to take on one of only seven values listed above.

Threshold	Threshold Distance
γ_1	−80 miles
γ_2	−29.9 miles
γ_3	−14.9 miles
γ_4	−0.9 miles
γ_5	1.1 miles
γ_6	10.1 miles
γ_7	30.1 miles

The model we assume for mortality is

$$\text{pr}(Y_i = 1|D, \Gamma, Z, X, \lambda, \delta, \alpha, p) = \text{logit}^{-1}(W_i'(\lambda + D_i\delta) + x_i'\alpha), \tag{9}$$

where W_i is a vector of 7 components with jth component $W_{ij} = 1$ if $\Gamma_i = \gamma_j$ and $W_{ij} = 0$ otherwise, λ is the vector of effects associated with the threshold groups,

δ is the vector of interaction effects of threshold group and treatment status, and α are the covariate effects on 30-day mortality. The parameters p are discussed below. The term $W_i'(\lambda + D_i\delta)$ assigns a different treatment effect (on the logit scale) to each threshold group. This recognizes that the treatment effect may differ depending on the latent threshold group, Γ_i, to which a patient is inferred to belong. Note that the first and last components of δ are not identifiable because patients for whom $\Gamma_i = \gamma_1$ are always treated and patients for whom $\Gamma_i = \gamma_7$ are never treated, so these two parameters are set to 0.

The model for treatment status can be written as

$$D_i = \begin{cases} 1, & \text{if } Z_i \geq \Gamma_i, \\ 0, & \text{if } Z_i < \Gamma_i. \end{cases} \tag{10}$$

As before, this condition ensures both monotonicity of treatment status with the instrument, and in conjunction with the model for mortality, ensures the exclusion restriction of the instrument given treatment status.

We assume a general discrete model for threshold groups,

$$W_i | Z, X, \lambda, \delta, \alpha, p \sim \text{Multinomial}(1, p), \tag{11}$$

where $p = (p_1, \ldots, p_7)$ is a vector of 7 probabilities summing to 1. This model can also be written as $\text{pr}(\Gamma_i = \gamma_j) = p_j$ for all $j = 1, \ldots, 7$, noting the correspondence in representation between Γ_i and W_i. We do not incorporate covariate information at this level of the model as we assume that the distribution of W_i is completely specified given the p_j.

We choose a noninformative proper prior distribution on the the parameters to reflect our initial uncertainty,

$$\begin{aligned} \lambda, \delta, \alpha &\sim \text{N}(\mathbf{0}, 100 \cdot I), \\ p = (p_1, \ldots, p_7) &\sim \text{Dirichlet}(0.5, \ldots, 0.5), \end{aligned} \tag{12}$$

where I is the identity matrix. These prior parameter values were chosen to allow the data to dominate inferences.

Model 2:

The second model is identical to the Model 1 with one exception. Rather than positing seven values for Γ_i, Model 2 posits only three. Thus, there are fewer parameters in the model for mortality (because W_i is a vector of only 3 values), and the vector p of multinomial probabilities has three elements. The model with seven threshold levels may be over-parametrized, and a model that assumes only three threshold levels may sufficiently describe the variability in thresholds across units. Table 3 shows the three threshold distances assumed for

Model 2. For Model 2, a patient with $\Gamma_i = \gamma_1$ is a never-taker, and a patient with $\Gamma_i = \gamma_3$ is an always-taker. Only when patient i is inferred to have $\Gamma_i = \gamma_2$ is a treatment effect defined.

Table 3. Threshold distances for Model 2. The latent variable Γ_i is assumed to take on one of only three values listed above.

Threshold	Threshold Distance
γ_1	−80 miles
γ_2	0 miles
γ_3	30 miles

Model 3:

In the third model, the thresholds are assumed to be continuously distributed. The model we assume for mortality is given by

$$\text{pr}(Y_i = 1|D, \Gamma, Z, X, \delta, \alpha) = \text{logit}^{-1}(D_i\delta_D + \Gamma_i\delta_\Gamma + D_i\Gamma_i\delta_{D\Gamma} + x_i'\alpha). \quad (13)$$

The parameters δ_D, δ_Γ and $\delta_{D\Gamma}$ are the effects of treatment, latent threshold, and their interaction, respectively. We let δ denote the collection of these three parameters. This component of Model 3 is analogous to Models 1 and 2 where each treatment/threshold combination corresponded to different effects on mortality. The current model is more restrictive in that the log-linearity of the threshold effect on mortality is assumed.

The model for treatment status is assumed to be

$$D_i = \begin{cases} 1, & \text{if } Z_i \geq \Gamma_i, \\ 0, & \text{if } Z_i < \Gamma_i. \end{cases} \quad (14)$$

The model in (14) is identical to the model for treatment status in Models 1 and 2.

Finally, the model for thresholds is assumed to be normally distributed, that is

$$\Gamma_i|Z, X, \delta, \alpha, \mu, \sigma^2 \sim \text{N}(\mu, \sigma^2). \quad (15)$$

This model assumes that the Γ_i come from a single normal distribution.

A noninformative prior distribution is assumed for all parameters, that is,

$$\begin{aligned} \delta, \alpha &\sim \text{N}(\mathbf{0}, 100 \cdot I), \\ \mu &\sim \text{N}(0, 10000), \\ p(\sigma^2) &\propto 1/\sigma^2. \end{aligned} \quad (16)$$

These distributions reflect the initial uncertainty in the parameter values.

Model 4:

Model 4 is similar to Model 3 in that the threshold is assumed to be continuous. The difference is that the distribution of the thresholds are assumed to depend on covariates. Thus, the model for Γ_i is a linear regression with

$$\Gamma_i | Z, X, \delta, \alpha, u_i, \beta, \sigma^2 \sim \mathrm{N}(u_i'\beta, \sigma^2), \tag{17}$$

where u_i is a vector of covariates for patient i that relate to the value of the threshold parameter, and β is a vector of the effects of these covariates. The covariates used in this stage of the model, all categorical, are displayed in Table 4. Apart from age information, all of the covariates at this stage of the model are indicators of comorbid conditions (i.e., conditions of poor health not directly related to the severity of the patient's AMI). At the physicians' discretion, patients who are unhealthy by virtue of having several comorbid conditions do not typically undergo catheterizations because they lead to revascularization procedures such as coronary angioplasty or bypass surgery. Although there is no significant surgical risk associated with catheterization, the same can not be said of the risk associated with revascularization procedures. In such a situation, a patient would likely have a large value of Γ_i, and including the comorbid information would aid in inferring the Γ_i more precisely. This would in turn result in greater precision in inferences about the local average treatment effect.

Table 4. Covariates for thresholds in Model 4. The covariates include age information and comorbid conditions. Less than 8–9% of the 3667 patients in the sample experience the comorbid conditions, with the exception of chronic pulmonary disease which affects 14.4% of the sample.

Covariate for modeling threshold parameters	Frequency in sample
Age 65–69	0.216
Age 70–74	0.235
Age 75–79	0.227
Age 80–84	0.179
Age 85+	0.143
Cancer	0.021
Connective tissue disorder	0.003
Dementia	0.006
Uncomplicated diabetes	0.093
Diabetes with end organ damage	0.082
Chronically debilitating neurological disorders	0.014
Paralysis	0.003
Chronic pulmonary disease	0.144
Chronic renal failure without dialysis	0.020

The prior distribution is the same as in Model 3, replacing the distribution for μ with

$$\beta \sim \mathrm{N}(0, 10000). \tag{18}$$

Summary of four models:

The key features of the four models can be summarized as follows:

- Model 1: Seven threshold levels modeled multinomially, no covariates in threshold model,
- Model 2: Three threshold levels modeled multinomially, no covariates in threshold model,
- Model 3: Continuous threshold levels assumed to follow a normal distribution, no covariates in threshold model, and
- Model 4: Continuous threshold levels assumed to follow a normal distribution conditional on covariates (i.e., linear regression).

3.3. Analysis via iterative simulation

The four models in the preceding section were fit using Markov chain Monte Carlo simulation via the Gibbs sampler. This involved iteratively simulating values from three sequences of conditional posterior distributions. The details of the Markov chain simulation can be found in Appendix A.

For each model, a single "pilot" Gibbs sampler with starting values at the prior means was run to determine regions of the parameter space with high posterior mass. Four parallel Gibbs samplers for each model were then run with overdispersed starting values relative to the draws of the parameter values from the pilot sampler. Each sampler was run for 20000 iterations and convergence was diagnosed by examining the potential scale reduction (Gelman and Rubin (1992)) of the parameters in the mortality model, and the parameters in the model for the thresholds. The potential scale reduction is an estimate of the factor by which the variance of the current distribution of draws in the Gibbs sampler will decrease with continued iterations. Values near 1 are indicative of convergence. After appropriately transforming variables, all the estimated potential scale reductions for all parameters based on samples beyond iteration 10000 were no more than 1.02, which appears close enough to 1 for practical purposes to assume the Gibbs sampler has reached its stationary distribution. We used the final 10000 posterior draws from each of the four sampler series in each model as the final sample upon which to base inferences. In addition to producing model parameter draws for each Gibbs sampler iteration beyond iteration 10000, we also produced draws of the overall local average treatment effect (LATE), and the LATE stratified by covariates.

3.4. Results

The posterior distribution of the LATE for each of the four models is shown in Figure 2. In each model, the LATE is calculated at every iteration of the Gibbs sampler for patients having a defined treatment effect. Patients inferred to have threshold values of γ_1 and γ_7 in Model 1, and γ_1 and γ_3 in Model 2, are excluded from the LATE calculation because a causal effect is not defined for these threshold values. In Models 3 and 4, threshold values inferred to be lower than the minimum differential distance or higher than the maximum differential distance are excluded from the LATE calculation. The posterior mean LATE for the four models range from -0.146 for Model 4 to -0.247 for Model 2. Models 1 and 2 show greater variability of the LATE. This is due to the larger proportion of observations that are excluded when constructing the average treatment effect.

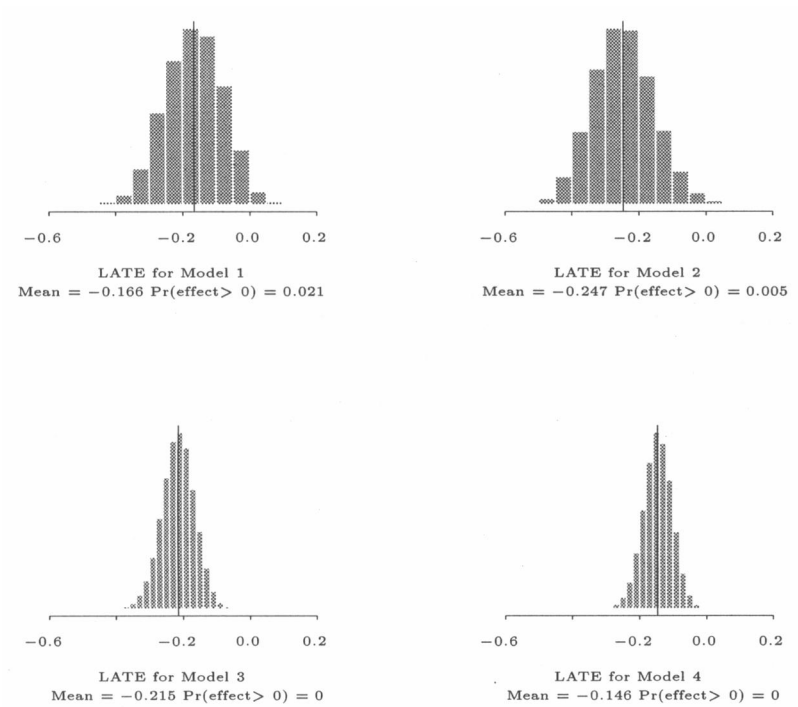

Figure 2. Posterior distribution of the local average treatment effects of catheterization on mortality from our four instrumental variable models. The local average treatment effect is defined as the average mortality when all patients undergo catheterization less the average mortality when all patients do not undergo catheterization for patients who are inferred to potentially receive both treatments. The solid vertical line corresponds to the posterior means.

Table 5. Posterior summaries of the local average treatment effect for each of the four fitted models. Negative values indicate that undergoing cardiac catheterization increases probability of 30-day survival. The treatment effect for individual strata and the overall treatment effect were computed from the model fit. For each stratum, the posterior means are displayed with the 95% central posterior intervals in parentheses below.

Stratum	Model 1	Model 2	Model 3	Model 4
Overall	−0.166	−0.247	−0.215	−0.146
	(−0.334, −0.005)	(−0.416, −0.063)	(−0.315, −0.120)	(−0.233, −0.064)
Male	−0.161	−0.247	−0.215	−0.149
	(−0.331, −0.007)	(−0.417, −0.063)	(−0.320, −0.120)	(−0.239, −0.066)
Female	−0.170	−0.246	−0.214	−0.143
	(−0.340, −0.001)	(−0.417, −0.062)	(−0.312, −0.120)	(−0.230, −0.060)
Non-black	−0.165	−0.249	−0.217	−0.148
	(−0.335, −0.006)	(−0.419, −0.063)	(−0.319, −0.121)	(−0.236, −0.065)
Black	−0.171	−0.237	−0.199	−0.132
	(−0.340, 0.016)	(−0.405, −0.058)	(−0.297, −0.109)	(−0.219, −0.052)
Age 65–69	−0.144	−0.233	−0.195	−0.141
	(−0.307, −0.004)	(−0.403, −0.057)	(−0.304, −0.102)	(−0.229, −0.064)
Age 70–74	−0.147	−0.229	−0.188	−0.135
	(−0.302, −0.003)	(−0.398, −0.055)	(−0.292, −0.100)	(−0.218, −0.060)
Age 75–79	−0.166	−0.242	−0.207	−0.145
	(−0.332, −0.003)	(−0.414, −0.060)	(−0.309, −0.114)	(−0.235, −0.061)
Age 80–84	−0.193	−0.266	−0.257	−0.176
	(−0.394, −0.003)	(−0.440, −0.070)	(−0.367, −0.144)	(−0.291, −0.064)
Age 85+	−0.205	−0.274	−0.291	−0.189
	(−0.435, −0.005)	(−0.449, −0.074)	(−0.411, −0.154)	(−0.329, −0.056)
Rural	−0.166	−0.248	−0.225	−0.153
	(−0.343, 0.007)	(−0.412, −0.065)	(−0.322, −0.130)	(−0.242, −0.067)
Urban	−0.165	−0.246	−0.210	−0.143
	(−0.337, −0.006)	(−0.420, −0.061)	(−0.314, −0.115)	(−0.231, −0.062)

Table 5 summarizes the posterior distribution of the overall LATE for all four models, and the distributions stratified by gender, race, age, and urbanicity. The table shows posterior mean estimates along with central 95% posterior intervals. All four posterior intervals for the overall local average treatment effects are below 0. This suggests evidence under the modeling assumptions that catheterization does indeed have a positive effect. The catheterization effect appears to be larger (a more negative difference) for older patients, though the posterior intervals reveal a fair amount of variability. The treatment effect stratified by gender, race or urbanicity does not differ substantially from the overall average treatment effect.

Table 6. Posterior summaries of covariate effect on mortality for each of the four fitted models. The larger the covariate effect, the greater the impact on the probability of 30-day mortality. For each covariate, the posterior means are displayed with the 95% central posterior intervals in parentheses below. Higher parameter values indicate a greater probability of 30-day mortality.

Effect	Model 1	Model 2	Model 3	Model 4
Female	−0.235 (−0.704, 0.100)	−0.098 (−0.336, 0.135)	−0.097 (−0.326, 0.130)	−0.089 (−0.318, 0.142)
Non-white	−0.585 (−1.407, −0.013)	−0.181 (−0.512, 0.148)	−0.169 (−0.493, 0.147)	−0.182 (−0.507, 0.138)
Age 65–69	−1.933 (−2.932, −1.049)	−1.092, (−1.485, −0.715)	−0.993 (−1.457, −0.545)	−1.133 (−1.537, −0.739)
Age 70–74	−1.928 (−2.855, −1.100)	−1.140 (−1.513, −0.782)	−1.042 (−1.492, −0.600)	−1.204 (−1.605, −0.813)
Age 75–79	−1.641 (−2.473, −0.905)	−0.966 (−1.314, −0.627)	−0.886 (−1.326, −0.453)	−1.110 (−1.546, −0.684)
Age 80–84	−1.150 (−1.951, −0.456)	−0.557 (−0.895, −0.229)	−0.485 (−0.920, −0.050)	−0.802 (−1.308, −0.302)
Age 85+	−0.722 (−1.557, −0.059)	−0.268 (−0.622, 0.085)	−0.198 (−0.650, 0.254)	−0.670 (−1.341, −0.009)
Rural	−0.034 (−1.433, 1.316)	−0.040 (−0.364, 0.244)	0.070 (−0.181, 0.318)	0.107 (−0.135, 0.353)

Table 6 shows posterior summaries of the covariate effects, α, on mortality after controlling for the effect of catheterization and threshold to receive treatment. All four models seem relatively similar in their parameter summaries, which may be interpreted identically across models. According to the 95% posterior intervals, gender, race and urbanicity have little effect on mortality beyond the effect explained by catheterization and the threshold parameter. Younger patients, not surprisingly, tend to survive longer than older patients, as indicated by the positive relationship between the posterior parameter estimates and age. The posterior intervals for the components of α are fairly wide, particularly for Models 1 and 2. The reason for the posterior variability in Models 1 and 2 is related to the degeneracy that could occur during the iterative simulation. If during an iteration no patient is inferred to have a particular threshold value, then the conditional posterior distribution of the covariate effects, α, will be degenerate. Such a degeneracy is much less likely to occur in Models 3 and 4.

A deeper examination of the results of fitting Models 3 and 4 reveals that the inclusion of covariates in the threshold model does not seem to add substantially in explaining the variability of the thresholds. While age has a strong positive association with effect on a patient's threshold (older patients are inferred to

have larger values of Γ_i), and all the comorbid conditions except uncomplicated diabetes are associated with larger thresholds, the reduction in variance due to these additional covariates is small. In particular, from Model 3, a 95% central posterior interval for σ, the standard deviation of the Γ_i unconditional on covariates is (39.29, 50.28) with an estimated posterior mean of 44.38. From Model 4, where σ is fit conditional on the covariates, the 95% central posterior interval is (37.46, 49.67) with an estimated posterior mean of 42.96. This drop in σ could potentially be greater if more predictive covariates were available.

The large variability in the posterior distribution for the local average treatment effect, particularly in Models 1 and 2, can be explained by examining the posterior distribution of threshold probabilities. This is summarized in Figures 3 and 4. For Model 1, the medians of the posterior distributions for $\text{pr}(\Gamma_i = \gamma_1)$ and $\text{pr}(\Gamma_i = \gamma_7)$ are 0.111 and 0.400, respectively. This indicates that an average treatment effect is computed based on an average of only $1-0.111-0.400 = 48.9\%$ of the data. Similarly, for Model 2, the posterior medians of $\text{pr}(\Gamma_i = \gamma_1)$ and $\text{pr}(\Gamma_i = \gamma_3)$ are 0.279 and 0.556, respectively. Here, an average of only 16.6% of the data is used to calculate a local average treatment effect. This loss in efficiency is, in essence, the cost for accounting for the effects of confounding biases. It is also worth noting that the posterior distribution of threshold parameters, Γ_i, do not change smoothly, so the multinomial assumption on the Γ_i captures features of the underlying continuous distribution of Γ_i that might not otherwise be apparent. By contrast, Models 3 and 4 make a strong assumption about the functional relationship (i.e., log-linear) of Γ_i and survival probability.

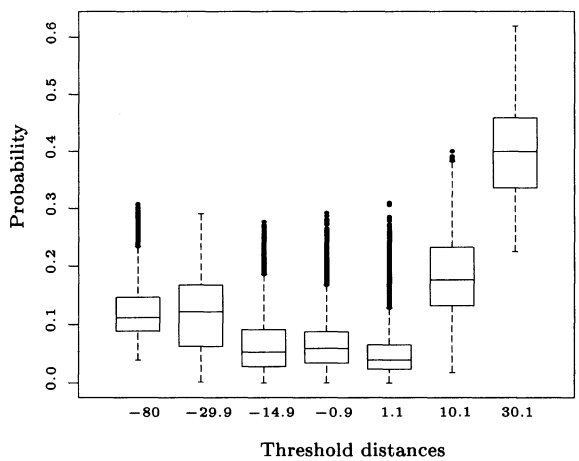

Figure 3. Posterior distribution for the probabilities of belonging to each latent threshold group in Model 1. The high probability of belonging to the group with the largest thresholds suggest that a substantial fraction of the patients in the sample would never undergo a catheterization given their covariates.

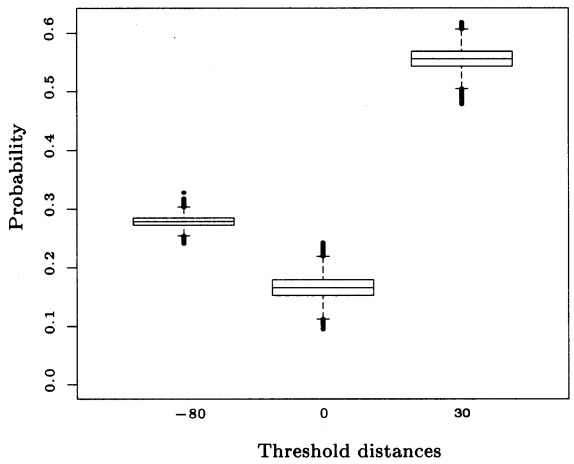

Figure 4. Posterior distribution for the probabilities of belonging to each latent threshold group in Model 2. The local average treatment effect is defined only for patients inferred to have a threshold distance of 0, so that effectively less than 20% of the sample is used to estimate the treatment effect.

3.5. Model comparison

To compare the models, we examine how well they each predict responses on a subset of the sample. We refit each of our models leaving out a random "validation" sample of 200 observations, the same 200 observations for each model. Each Gibbs sampler was run for a burn-in period of 3000 iterations, with starting values at the posterior means from the previous fits. For the next 1000 iterations, we calculated a measure of predictive fit; given the model parameters at an iteration, we calculated for out-of-sample patient i, $i = 1, \ldots, 200$, the probability of 30-day mortality, π_i. We then computed the average log-(predictive)-likelihood (ALPL) for these 200 cases,

$$\text{ALPL} = \frac{1}{200} \sum_{i=1}^{200} \left(y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \right),$$

where y_i is 1 if the patient died within 30 days, and 0 otherwise. Larger values of ALPL indicate better prediction to the validation sample. A model that predicted with only 50% accuracy (random guessing) would produce an ALPL of $\log 0.5 = -0.693$.

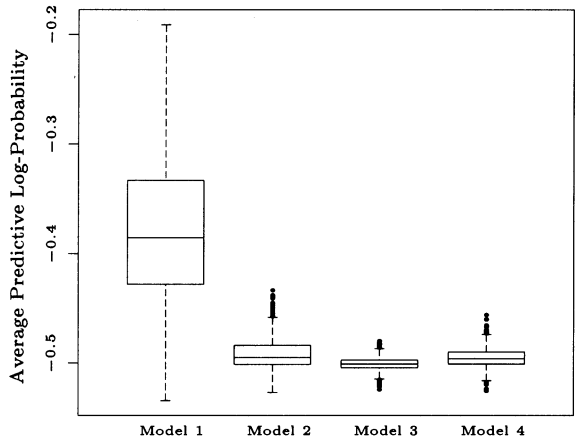

Figure 5. Posterior distributions for the average log-(predictive)-likelihood (ALPL) computed over the 200 observations left out of the model fit. Larger values of ALPL indicate better predictions. The plot demonstrates that, on average, Model 1 outpredicts the other three models.

Figure 5 shows the comparison of the posterior distributions of ALPL for each of the four models via boxplots. The figure indicates that all four models predict better than random, and, on average, Model 1 predicts the validation sample responses substantially better than the other three models. The large posterior variability of the ALPL for Model 1 may be related to the degeneracies induced by the model fit in which no patients are inferred to belong to a threshold group. A comparison of the ALPL for Models 1 and 2 reveals that using seven threshold levels per patient captures greater variability than using only three. The distribution of the ALPL for Models 3 and 4 demonstrate that using a continuous threshold parameter is too restrictive an assumption compared to a multinomial threshold with seven levels. It appears from the comparison of distributions for Models 3 and 4 that using covariate information at the threshold level improves predictability, though not substantially.

We also fit the model using a GMM approach via the two-stage least squares analysis used in MMN. From this analysis, the estimated overall treatment effect from our data is -0.1520, with an approximate standard error of 0.115. This result is comparable with those of our models, though the GMM standard error is slightly larger. We expect the results to be somewhat similar because the large sample size guarantees that the GMM estimator is well-behaved. Because our framework uses likelihood-based inference rather than moment method inference, the properties of our procedure are more easily understood, especially in small to moderate sized samples.

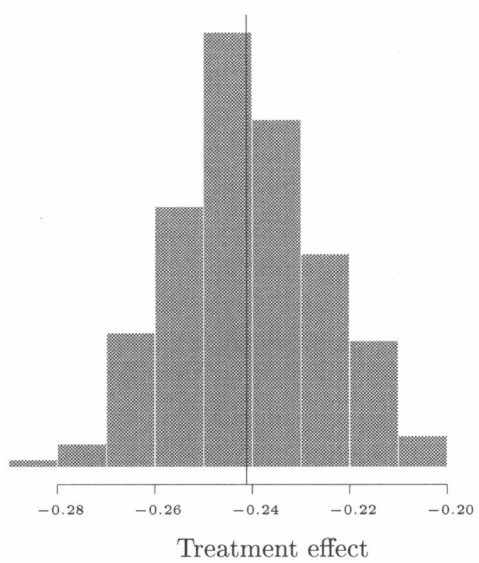

Figure 6. Posterior distribution of the average treatment effect of undergoing catheterization from a logistic regression model. The average treatment effect is defined as the average mortality when all patients undergo catheterization less the average mortality when all patients do not undergo catheterization. The solid vertical line corresponds to the posterior mean of -0.2412.

3.6. Unconfounded treatment assignment

For comparative purposes, we fit a model for mortality that ignores differential distance as an instrumental variable. This can be accomplished by fitting a logistic regression model using a uniform prior distribution on the regression coefficients to obtain (potentially biased) inferences about the effect of catheterization on mortality. We simulated average treatment effects from a normal approximation to the posterior distribution of regression coefficients. This was carried out by drawing individual samples from the normal distribution of coefficients, calculating the posterior probabilities of mortality for each patient conditional on undergoing and not undergoing catheterization, simulating two Bernoulli random variables (catheterization, no catheterization) for each patient, and then calculating the average difference across all patients. This process was performed 200 times to produce an approximate posterior distribution for the (biased) treatment effect. The distribution is summarized in Figure 6. The posterior mean from this model is -0.2412. The distribution is tightly centered at the mean, with a 95% central posterior interval of $(-0.2697, -0.2121)$. This wrongly suggests that undergoing a catheterization improves chances of 30-day survival with complete certainty. Naturally, this naive analysis does not account for the possible confounding biases.

4. Conclusions

The importance of using instrumental variables in non-randomized study designs can be seen from the catheterization example. Using only covariate adjustment results in model inferences that are incorrectly precise. Incorporating an instrumental variable into our model produces inferences with appropriately large variability that accounts for the selection biases in the study design. The increase in uncertainty in our particular model is a result of patients being classified into compliers and non-compliers, treatment effects not being defined for the latter group.

As with any modeling framework, an important issue in the use of our methodology is that the model assumptions apply to the data. This is especially important for assumptions such as the monotonicity of treatment choice with the instrument, which cannot be verified empirically, though has testable implications. The reason differential distance in our example can be argued to satisfy monotonicity relies on the nature of the index AMI hospital admission. Because of the emergent nature of an AMI, individuals who suffer an infarct are typically admitted to a nearby hospital. Whether a patient is taken to a hospital equipped to perform a catheterization versus one that does not is, to a large degree, a function of the distances to different candidate hospitals. By contrast, if a study were performed to assess the effect of the treatment on patient outcome of early-stage breast cancer via breast-conserving surgery versus mastectomy (breast removal), then differential distance to facilities equipped with radiation therapy facilities would not be an appropriate instrument. In this case, it is not likely that a patient would choose breast-conserving surgery if the differential distance to a hospital with radiation therapy facilities were smaller. Because the successful treatment of breast cancer does not depend so crucially on the prompt delivery of the patient to a hospital, the choice of the type of hospital is not as much a function of distances to various hospitals as it is of other factors. Factors such as choice of physician or recommendations to undergo treatment at a particular facility, make monotonicity an untenable relationship between differential distance and treatment choice.

The difficulty in using our framework is that greater flexibility in choice of models requires more care in selecting appropriate models. The approach we have taken with our mortality model involved examining several instrumental variable models and comparing predictive characteristics. Alternative approaches would involve learning the shapes of the functional relationships through flexible regression methods (e.g., non-parametric regression procedures) and then modeling the instrument and threshold parameters as continuous variables after properly choosing a functional form. This is an area for future work.

The use of latent threshold parameters in our instrumental variable approach not only permits the specification of commonly used assumptions in a conveniently parametrized model, but permits a flexible choice of assumptions about the probability models and the data structure. With the recent addition of computational techniques such as iterative simulation to perform inference, fitting the models we propose, or their extensions, is straightforward.

Acknowledgements

The authors wish to thank Guido Imbens, Ree Dawson, Barbara McNeil, David Hoaglin, Mark McClellan, and two anonymous referees for their helpful comments.

Appendix

Conditional Distributions for MCMC Sampling

Gibbs sampling for Models 1 and 2

Conditional posterior distribution of λ, δ, α

The conditional posterior distribution of the mortality effect parameters is proportional to a product of a binomial likelihood involving terms that only appear in (9) and a normal prior distribution. Generating parameter values from this distribution may be carried out through rejection sampling, as described by Zeger and Karim (1991). Following their approach, the product of the binomial likelihood and normal prior distribution is approximated by a normal distribution with the same mode and twice the variance. A single draw is obtained by simulation from the approximating normal distribution, and this is accepted with a probability proportional to the ratio of the actual posterior density to the approximating normal density; otherwise the draw is rejected. This process is repeated until a draw is accepted. The resulting draw is a sample from the desired conditional posterior distribution.

Conditional posterior distribution of p

The conditional posterior distribution of multinomial probabilities for threshold groupings is proportional to a product of a multinomial likelihood and a (conjugate) Dirichlet prior distribution. The resulting product is therefore a Dirichlet density. A sample from a Dirichlet distribution may be obtained, for example, by generating Gamma random variables with shape parameters equal to the Dirichlet parameters, and then computing the ratio of each Gamma draw to the sum of all the Gamma draws.

Conditional posterior distribution of Γ_i

Conditional on the data and the remaining parameters, the posterior distribution of the Γ_i are independent and may be drawn individually. The distribution of a single Γ_i is proportional to the product of three terms,

$$f(\Gamma_i|\lambda,\delta,\alpha,Y,D,X,Z,p)$$
$$\propto \left(Q_i^{Y_i}(1-Q_i)^{(1-Y_i)}\right)\cdot\left((\Delta[Z_i\geq\Gamma_i])^{D_i}(\Delta[Z_i<\Gamma_i])^{(1-D_i)}\right)\cdot\left(p_1^{W_{i1}}\cdots p_K^{W_{iK}}\right),$$

where

$$Q_i=\text{logit}^{-1}(W_i'(\lambda+D_i\delta)+x_i'\alpha).$$

and where $\Delta[\cdot]$ is 1 if the argument is true, and 0 otherwise. This product can be evaluated for each of the K values of Γ_i and then standardized to sum to 1, where $K=7$ for Model 1 and $K=3$ for Model 2. Note that the product evaluates to 0 when $Z_i<\Gamma_i$ and $D_i=1$, or when $Z_i\geq\Gamma_i$ and $D_i=0$, so that sampling may be performed more efficiently by excluding threshold groupings resulting in 0 probability.

Gibbs sampling for Models 3 and 4

Conditional posterior distribution of δ,α

Analogous to Models 1 and 2, the conditional posterior distribution of δ,α is proportional to a product of a binomial likelihood and a normal prior distribution. Values of δ and α are simulated via rejection sampling, as described above.

Conditional posterior distribution of μ or β

The conditional posterior density of μ in Model 3 is proportional to the product of a normal prior, centered at 0, and a normal likelihood where the Γ_i are the data, and σ^2 is the variance (treated as known). The resulting density is therefore normal, and values from this distribution may be simulated without difficulty.

For Model 4, the conditional posterior density of β is proportional to the product of a multivariate normal prior, centered at 0, and a normal likelihood which is the regression of the Γ_i on the covariates, u_i, given the variance is σ^2. The resulting density is multivariate normal, so that values of β may be simulated using standard methods.

Conditional posterior distribution of σ^2

The conditional posterior density of σ^2 is the product of the prior of $1/\sigma^2$ and a scaled reciprocal-χ^2 likelihood (product of normal densities). The resulting density is therefore also a scaled reciprocal-χ^2 density with the appropriate degrees of freedom ($n-2$ for Model 3, and $n-16$ for Model 4, where $n=3667$, the sample size). Drawing from the reciprocal-χ^2 distribution can be accomplished by drawing from the χ^2 distribution with the same degrees of freedom, and then taking the reciprocal of the result.

Conditional posterior distribution of Γ_i

As in Models 1 and 2, the conditional posterior distribution of the Γ_i are

independent, so they may be drawn separately. The conditional posterior distribution of Γ_i in Models 3 and 4 is proportional to the product of three terms,

$$p(\Gamma_i) \propto \left(Q_i(\Gamma_i)^{Y_i}(1 - Q_i(\Gamma_i))^{(1-Y_i)}\right) \cdot \left((\Delta[Z_i \geq \Gamma_i])^{D_i}(\Delta[Z_i < \Gamma_i])^{(1-D_i)}\right) \cdot \varphi(\Gamma_i|\mu_i, \sigma^2), \tag{A.1}$$

where

$$Q_i(\Gamma_i) = \text{logit}^{-1}(D_i\delta_D + \Gamma_i\delta_\Gamma + D_i\Gamma_i\delta_{D\Gamma} + x_i'\alpha)$$

and $\varphi(\cdot|\mu_i, \sigma^2)$ is a normal density with mean μ_i ($\mu_i = \mu$ in Model 3, and $\mu_i = u_i'\beta$ in Model 4) and variance σ^2.

Obtaining a direct draw from this density can be difficult. Instead, we apply the weighted bootstrap (Smith and Gelfand (1992)), which is closely related to the SIR algorithm (Rubin (1988)). This can be applied as follows. The product of the second and third factors in (A.1) correspond to an unnormalized truncated normal density. We simulate eight values at random from this truncated normal density. Denote these eight values $\gamma_1, \ldots, \gamma_8$. Now compute the eight unnormalized importance weights

$$Q_i(\gamma_j)^{Y_i}(1 - Q_i(\gamma_j))^{(1-Y_i)}$$

for $j = 1, \ldots, 8$, and normalize them to sum to 1. Now resample a single value from the eight with the computed importance weights. This procedure results in a value that is approximately drawn from the desired conditional posterior distribution.

References

Angrist, J. (1990). Lifetime earnings and the Vietnam era draft lottery: evidence from Social Security administrative records. *Amer. Econom. Rev.* **80**, 313-336.

Angrist, J., Imbens, G. and Rubin, D. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91**, 444-472.

Angrist, J. and Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings. *Quart. J. Econom.* **106**, 979-1014.

Angrist, J. and Krueger, A. (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *J. Amer. Statist. Assoc.* **87**, 328-336.

Bowden, R. and Turkington, D. (1984). *Instrumental Variables.* Cambridge University Press, Cambridge.

Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica* **60**, 173-180.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34**, 305-334.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable (with discussion). *J Amer. Statist. Assoc.* **86**, 9-26.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, 972-985.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7**, 457-511.

Heckman, J. and Hotz, J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. *J. Amer. Statist. Assoc.* **84**, 862-874.

Heckman, J. and Robb, R. (1985). Alternative methods for evaluating the impact of interventions. In *Longitudinal Analysis of Labor Market Data.* (Edited by J. Heckman and B. Singer). Cambridge University Press, New York.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467-475.

Imbens, G. and Rubin, D. (1997a). Bayesian causal inference in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305-327.

Imbens, G. and Rubin, D. (1997b). Estimating the outcome distributions for compliers in instrumental variables models. *Rev. Econom. Stud.* **64**, 555-574.

McClellan, M. and Newhouse, J. (1993). The marginal benefits of medical treatment intensity: acute myocardial infarction in the elderly. Technical Report, Harvard University and the National Bureau of Economic Research, Cambridge.

McClellan, M., McNeil, B. and Newhouse, J. (1994). Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. *J. Amer. Medical Assoc.* **272**, 859-866.

Nagelkerke, N., Fidler, V. and Buwalda, M. (1988). Instrumental variables in the evaluation of diagnostic test procedures when the true disease state is unknown. *Statist. Medicine* **7**, 739-744.

Phillips, P. (1983). Exact small sample theory in the simultaneous equations model. Chapter 8. In *The Handbook of Econometrics* **1** (Edited by Z. Griliches and M. Intriligator). North-Holland, Amsterdam.

Rubin, D. (1978). Bayesian inference for causal effects. *Ann. Statist.* **6**, 34-58.

Rubin, D. (1980). Discussion of "Randomization analysis of experimental data in the Fisher randomization test," by Basu. *J. Amer. Statist. Assoc.* **75**, 591-593.

Rubin, D. (1988). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics* **3** (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 395-402. Oxford University Press, Cambridge, MA.

Rubin, D. (1990). Comment on "Neyman (1923) and causal inference in experiments and observational studies." *Statist. Sci.* **5**, 472-480.

Rubin, D. (1998). More powerful randomization-based p-values in double-blind trials with noncompliance. *Statist. Medicine* **17**, 371-385.

Smith, A. F. M. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling-resampling perspective. *Amer. Statist.* **46**, 84-88.

Stefanski, L. and Buzas, J. (1995). Instrumental variable estimation in binary regression measurement error models. *J. Amer. Statist. Assoc.* **90**, 541-550.

Zeger, S. and Karim, M. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86**, 79-86.

Department of Mathematics, Boston University, 111 Cummington St., Boston MA 02215, U.S.A.

E-mail: mg@math.bu.edu

Department of Health Care Policy, Harvard Medical School, U.S.A.

E-mail: sharon@hcp.med.harvard.edu

(Received February 1996; accepted June 1999)