of heart disease in the population can tighten the bounds. A simple way to do this is to extend Propositions 6.1 and 6.2 to cases in which $P(y = 1)$ is permitted to vary over a restricted range rather than over the interval [0, 1].

King and Zeng (2002) extend the propositions in this manner and reconsider the numerical example above when $P(y = 1)$ is permitted to vary over the range [0.05, 0.15]. Under the assumption that $P(y = 1)$ lies in this interval, they find that RR $\in$ [1.46, 1.53] and AR $\in$ [0.021, 0.056].

Sampling from One Response Stratum

Suppose that one has a random sample of ill persons but not of healthy ones, so $P(x \mid y = 1)$ is known but not $P(x \mid y = 0)$. Consider Propositions 6.3 and 6.4 in this setting.

First, suppose that the outcome distribution is known. By Proposition 6.3, $P(y = 1 \mid x = k) \in [0.06, 1]$, $P(y = 1 \mid x = j) \in [0.04, 1]$, RR $\in$ [0.06, 23.5], and AR $\in$ [−0.94, 0.96]. Thus, knowing the marginal probability of heart disease has little identifying power in this example.

Now, suppose that the covariate distribution is known. The parameters of the example give $c = 0.83$, so Proposition 6.4 yields $P(y = 1 \mid x = k) \in [0, 1]$ and $P(y = 1 \mid x = j) \in [0, 0.67]$. The quantity $d = 0.4$, so AR $\in$ [0, 0.33]. Thus, knowing the prevalence of smoking in the population reveals little about the magnitudes of the response probabilities but reveals a fair bit about attributable risk and point-identifies relative risk.

Endnotes

Sources and Historical Notes

The analysis in Sections 6.4 and 6.5 originally appeared in Manski (1995, 2001). In particular, Propositions 6.1 and 6.2 are based on Manski (1995, Chapter 4), and Propositions 6.3 and 6.4 on Manski (2001, Propositions 3 and 4). Manski and Lerman (1977) recommended collection of auxiliary outcome data to learn $P(y)$. Hsieh, Manski, and McFadden (1985) showed that auxiliary covariate data can enable deduction of $P(y)$.

Cornfield (1951) showed that the rare-disease assumption point-identifies relative risk. The lack of relevance of relative risk to public health has long been criticized; see, for example, Berkson (1958), Fleiss (1981, Section 6.3), and Hsieh, Manski, and McFadden (1985).

The term *reverse regression* has been used in the literature on mean regression to denote the conditional expectation $E(x|y)$, when the conditional expectation of interest is $E(y|x)$; see Goldberger (1984).

7

Analysis of Treatment Response

7.1. Anatomy of the Problem

The four remaining chapters of this book study a pervasive and distinctive problem of missing outcomes. The problem is the non-observability of *counterfactual outcomes* in empirical analysis of treatment response.

Studies of treatment response aim to predict the outcomes that would occur if different treatment rules were applied to a population. Treatments are mutually exclusive, so one cannot observe the outcomes that a person would experience under all treatments. At most, one can observe the outcome that a person experiences under the treatment actually received. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.

For example, suppose that patients ill with a specified disease can be treated by drugs or by surgery. The relevant outcome might be life span. One may want to predict the life spans that would occur if patients with specified risk factors were to receive each treatment. The available data may be observations of the actual life spans of these patients, some of whom were treated by drugs and the rest by surgery.

Predicting Outcomes Under Conjectural Treatment Rules

To formalize the inferential problem, let each member j of population J have covariates $x_j \in X$ and a *response function* $y_j(\cdot)$: $T \to Y$ mapping the mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Let $z_j \in T$ denote the treatment that person j receives and $y_j \equiv y_j(z_j)$ be the outcome that he experiences. Then $y_j(t)$, $t \neq z_j$ are counterfactual outcomes.

Let $y(\cdot)$: $J \to Y^{|T|}$ be the random variable mapping the population into their response functions. Let z: $J \to T$ be the *status quo treatment rule* mapping the members of J into the treatments that they actually receive. Response functions are not observable, but covariates, realized treatments, and realized outcomes may be observable. If so, random sampling from J reveals the status quo (outcome, treatment) distributions $P(y, z|x)$ as well as the covariate distribution $P(x)$.

The distinctive problem of the analysis of treatment response is to predict the outcomes that would occur under alternatives to the status quo treatment rule. Let $\tau: J \to T$ be a *conjectural treatment rule*, the outcomes of which one would like to predict. Thus, person j's outcome under rule τ would be $y_j(\tau_j)$. This outcome is counterfactual whenever $\tau_j \neq z_j$. Hence, the sampling process does not reveal the conjectural outcome distributions $P[y(\tau)|x]$. The problem is to combine empirical knowledge of $P(y, z|x)$ with credible prior information to learn about $P[y(\tau)|x]$.

To simplify the presentation, the analysis in this chapter supposes that the covariate space X is finite and that $P(x = x, z = t) > 0$, $(t, x) \in T \times X$. These regularity conditions are maintained without further reference.

The Selection Problem

Researchers studying treatment response often want to predict the outcomes that would occur under conjectural treatment rules in which all persons with the same covariates receive the same treatment. Consider, for example, the medical setting described earlier. Let the relevant covariate be age. Then one treatment rule might mandate that all patients receive drugs, another that all patients receive surgery, and yet another that older patients receive drugs and younger ones receive surgery.

By definition, $P[y(t)|x = x]$ is the distribution of outcomes that would occur if all persons with covariate value x were to receive a specified treatment t. Hence prediction of outcomes under rules mandating uniform treatment conditional on covariates requires inference on the distributions $\{P[y(t)|x], t \in T\}$. The problem of identification of these distributions from knowledge of $P(y, z|x)$ is commonly called the *selection problem.*

The selection problem has the same structure as the missing-outcomes problem of Chapter 1 and Section 3.2. To see this, write

$$P[y(t)|x = x]$$

$$= P[y(t)|x = x, z = t]P(z = t|x = x) + P[y(t)|x = x, z \neq t]P(z \neq t|x = x)$$

$$= P(y|x = x, z = t)P(z = t|x = x) + P[y(t)|x = x, z \neq t]P(z \neq t|x = x). \quad (7.1)$$

The first equality is the Law of Total Probability. The second holds because $y(t)$ is the outcome experienced by persons who receive treatment t. The sampling process reveals $P(y|x = x, z = t)$, $P(z = t|x = x)$, and $P(z \neq t|x = x)$, but it is uninformative about $P[y(t)|x = x, z \neq t]$. Hence, the identification region for $P[y(t)|x = x]$ using the empirical evidence alone is

$$H\{P[y(t)|x = x]\} =$$

$$\{P(y|x = x, z = t)P(z = t|x = x) + \gamma P(z \neq t|x = x), \gamma \in \Gamma_Y\}. \quad (7.2)$$

Now, consider the collection of conjectural outcome distributions $\{P[y(t)|x], t \in T\}$. The sampling process is jointly uninformative about the counterfactual outcome distributions $\{P[y(t)|x = x, z \neq t], t \in T, x \in X\}$, which can take any value in $\times_{(t,x) \in T \times X} \Gamma_Y$. Hence the identification region for $\{P[y(t)|x], t \in T\}$ using the empirical evidence alone is the Cartesian product

$$H\{P[y(t)|x], t \in T\} = \times_{(t,x) \in T \times X} H\{P[y(t)|x = x]\}. \quad (7.3)$$

Observe that the empirical evidence cannot refute the hypothesis that all treatments have the same outcome distribution, conditional on x. Consider the hypothesis: $P[y(t)|x] = P[y(t')|x]$, all $(t, t') \in T \times T$. Identification region (7.3) necessarily contains distributions that satisfy this hypothesis. The easiest way to see this is to observe that the empirical evidence cannot refute the much stronger hypothesis $\{y_j(t) = y_j$, all $(t, j) \in T \times J\}$; that is, each person's counterfactual outcomes could in principle be the same as the outcome that he actually experiences.

Random Treatment Selection

A familiar "solution" to the selection problem is to assume that the status quo treatment rule makes realized treatments statistically independent of response functions, conditional on x; that is,

$$P[y(\cdot)|x] = P[y(\cdot)|x, z]. \quad (7.4)$$

This assumption implies that, for each $t \in T$ and $x \in X$,

$$P[y(t)|x = x] = P[y(t)|x = x, z = t] = P(y|x = x, z = t). \quad (7.5)$$

The sampling process reveals $P(y|x = x, z = t)$. Hence, assumption (7.4) point-identifies $P[y(t)|x = x]$.

Assumption 7.4 is credible in a classical randomized experiment, in which an explicit randomization mechanism has been used to assign treatments and all persons comply with their treatment assignments. The credibility of the assumption in other applied settings almost invariably is a matter of controversy.

The Task Ahead

This opening section introduced the general problem of predicting outcomes under conjectural treatment rules and then examined basic elements of the selection problem. The remainder of this chapter uses a social-planning problem to motivate rules mandating uniform treatment conditional on covariates and to show how the selection problem affects treatment choice. Chapters 8 and 9 study the identifying power of some monotonicity assumptions that may be credible and useful in the analysis of treatment response. Chapter 10 studies the *mixing problem*; that is, the problem of predicting outcomes under conjectural treatment rules that do not mandate uniform treatment conditional on covariates.

7.2. Treatment Choice in Heterogeneous Populations

An important practical objective of empirical studies of treatment response is to provide decision makers with information useful in choosing treatments. Often the decision maker is a *planner* who must choose treatments for a heterogeneous *treatment population*. The planner may want to choose treatments whose outcomes maximize the welfare of the treatment population.

For example, consider a physician choosing medical treatments for a population of patients. The physician may observe each patient's demographic attributes, medical history, and the results of diagnostic tests. He may then choose a treatment rule that makes treatment a function of these covariates. If the physician acts on behalf of his patients, the outcome of interest may be a measure of patient health status and welfare may be this measure of health status minus the cost of treatment, measured in comparable units.

As another example, consider a judge choosing sentences for a population of convicted offenders. The judge may observe each offender's past criminal record, demeanor in court, and other attributes. Subject to legislated sentencing guidelines, she may consider these covariates when choosing sentences. If the judge acts on behalf of society, the outcome of interest may be a measure of recidivism, and welfare may be this measure

of recidivism minus the cost of carrying out a sentence.

I present here a simple formulation of the planner's problem that motivates inference on the outcome distributions $\{P[y(t)|x],\ t \in T\}$ in general and the conditional mean outcomes $\{E[y(t)|x],\ t \in T\}$ in particular. I first specify the planner's choice set and objective function. I then derive the optimal treatment rule.

The Choice Set

Suppose that there is a finite set T of mutually exclusive and exhaustive treatments. A planner must choose a treatment for each member of the treatment population, denoted J^*. Each member j of population J^* has observable covariates $x_j \in X$ and an unobservable response function $y_j(\cdot)$: $T \to Y$ mapping treatments into real-valued outcomes.

The treatment population J^* is *identical in distribution* to the study population J, in which treatments have already been selected and outcomes have been realized. Thus (J^*, Ω, P) is a probability space whose probability measure P coincides with that of (J, Ω, P). The only difference between J and J^* is that the status quo treatment rule z has been applied in the former population, whereas a treatment rule has yet to be chosen in the latter.

There are no budgetary or other constraints that make it infeasible to choose some treatment rules. However, the planner cannot distinguish among persons with the same observed covariates and so cannot implement treatment rules that systematically differentiate among these persons. Hence, the feasible non-randomized rules are functions mapping the observed covariates into treatments.[1] Thus, uniform treatment conditional on covariates emerges naturally out of the planner's problem.

Formally, let Z(X) be the space of all functions mapping X into T. Let $z(\cdot) \in Z(X)$. Then feasible treatment rules have the form $\tau_j = z(x_j),\ j \in J^*$.

The Objective Function

Suppose that the planner wants to maximize population mean welfare. Let the welfare obtained from assigning treatment t to person j have the additive form $y_j(t) + c(t, x_j)$. Here $c(\cdot, \cdot)$: $T \times X \to R$ is a real-valued *cost function* known to the planner at the time of treatment choice. For each $z(\cdot) \in Z(X)$, let $E\{y[z(x)] + c[z(x), x]\}$ denote the mean welfare that would occur if the planner were to choose treatment rule $z(\cdot)$. Then the planner wants to solve the problem

$$\max_{z(\cdot) \in Z(X)} \quad E\{y[z(x)] + c[z(x), x]\}. \tag{7.6}$$

In the case of a physician, for example, $y_j(t)$ may measure the health status of patient j following receipt of treatment t, and $c(t, x_j)$ may be the (negative-valued) cost of treatment. The physician may know the costs of alternative treatments but not their health outcomes. Similarly, in the case of a judge, $y_j(t)$ may measure the rate of recidivism of offender j following receipt of sentence t, and $c(t, x_j)$ may be the cost of carrying out the sentence. Again, the judge may know the costs of alternative sentences but not their recidivism outcomes.

The assumption that the planner wants to maximize population mean welfare has normative, analytical, and practical appeal. This criterion function is standard in the public economics literature on social planning, which assumes that the objective is to maximize a utilitarian social welfare function. Linearity of the expectation operator yields substantial analytical simplifications, particularly through use of the law of iterated expectations. The practical appeal is that, as shown below, a planner choosing treatments to maximize mean welfare will want to learn mean treatment response, the main statistic reported in empirical studies of treatment response.

The Optimal Treatment Rule

The solution to optimization problem (7.6) is to assign to each member of the population a treatment that maximizes mean welfare conditional on the person's observed covariates. To show this, let $1[\cdot]$ be the indicator function taking the value one if the logical condition in the brackets holds and the value zero otherwise. For each $z(\cdot) \in Z(X)$, use the Law of Iterated Expectations to write

$$E\{y[z(x)] + c[z(x), x]\} = E\{E\{y[z(x)] + c[z(x), x] \mid x\}\}$$

$$= \sum_{x \in X} P(x = x)\{\sum_{t \in T} \{E[y(t) \mid x] + c(t, x)\} \cdot 1[z(x) = t]\}. \qquad (7.7)$$

For each $x \in X$, the quantity $\sum_{t \in T} \{E[y(t) \mid x = x] + c(t, x)\} \cdot 1[z(x) = t]$ is maximized by choosing $z(x)$ to maximize $E[y(t) \mid x] + c(t, x)$ on $t \in T$. Hence, a treatment rule $z^*(\cdot)$ is optimal if, for each $x \in X$, $z^*(x)$ solves the problem

$$\max_{t \in T} E[y(t) \mid x = x] + c(t, x). \qquad (7.8)$$

A planner who knows the conditional mean outcomes $E[y(\cdot) \mid x] \equiv \{E[y(t) \mid x = x], t \in T, x \in X\}$ can implement the optimal treatment rule. However, the selection problem and other identification problems limit the information that studies of treatment response provide. Sections 7.3 and 7.4

show how the selection problem affects treatment choice. Complements 7A and 7C consider the implications of other identification problems.

7.3. The Selection Problem and Treatment Choice

Let Y contain its lower and upper bounds $y_0 \equiv \inf_{y \in Y}$ and $y_1 \equiv \sup_{y \in Y}$. For $t \in T$ and $x \in X$, use the Law of Iterated Expectations to write

$$E[y(t) \mid x = x] =$$

$$E(y \mid x = x, z = t) \cdot P(z = t \mid x = x) + E[y(t) \mid x = x, z \neq t] \cdot P(z \neq t \mid x = x). \qquad (7.9)$$

The empirical evidence reveals $E(y \mid x = x, z = t)$ and $P(z \mid x = x)$, but it is uninformative about $E[y(t) \mid x = x, z \neq t]$. Hence, the identification region for $E[y(t) \mid x = x]$ using the empirical evidence alone is the closed interval

$$H\{E[y(t) \mid x = x]\}$$

$$= [E(y \mid x = x, z = t) \cdot P(z = t \mid x = x) + y_0 \cdot P(z \neq t \mid x = x),$$
$$E(y \mid x = x, z = t) \cdot P(z = t \mid x = x) + y_1 \cdot P(z \neq t \mid x = x)]. \qquad (7.10)$$

This result is a direct application of Proposition 1.1.

The object of interest is the collection of conditional mean outcomes $E[y(\cdot) \mid x]$. Its identification region $H\{E[y(\cdot) \mid x]\}$ is the rectangle

$$H\{E[y(\cdot) \mid x]\}$$

$$= \times_{(t, x) \in T \times X} [E(y \mid x = x, z = t) \cdot P(z = t \mid x = x) + y_0 \cdot P(z \neq t \mid x = x),$$
$$E(y \mid x = x, z = t) \cdot P(z = t \mid x = x) + y_1 \cdot P(z \neq t \mid x = x)]. \qquad (7.11)$$

This is the identification region because the empirical evidence is uninformative about the set of counterfactual means $\{E[y(t) \mid x = x, z \neq t], (t, x) \in T \times X\}$.

$H\{E[y(\cdot) \mid x]\}$ is a bounded, proper subset of $(Y^{|T|} \times X)$ if Y has bounded range. Suppose that this is so. Then, without loss of generality, outcomes may be scaled to take values in the unit interval. Setting $y_0 = 0$ and $y_1 = 1$ gives $H\{E[y(\cdot) \mid x]\}$ the simpler form

$$H\{E[y(\cdot)\,|\,x]\} = \times_{(t,x)\in T\times X}\,[E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x),\; E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + P(z\neq t\,|\,x=x)]. \tag{7.12}$$

The analysis in the remainder of this chapter assumes that Y has bounded range and that outcomes have been scaled to lie in the unit interval. Thus (7.12) henceforth gives the identification region for $E[y(\cdot)\,|\,x]$ using the empirical evidence alone.

Dominated Treatment Rules

In general, the empirical evidence alone is not sufficiently informative about $E[y(\cdot)\,|\,x]$ to enable solution of optimization problem (7.8). What should the planner do?

Clearly, the planner should not choose a *dominated* treatment rule. (A treatment rule $z(\cdot)$ is dominated if there exists another feasible rule, say $z'(\cdot)$, that necessarily yields at least the mean welfare of $z(\cdot)$ and that performs strictly better than $z(\cdot)$ in some possible state of nature.) The rectangular form of $H\{E[y(\cdot)\,|\,x]\}$ makes it easy to determine what treatment rules are dominated.

Let $(t, x) \in T \times X$, and consider any rule that assigns treatment t to persons with covariate value x. By (7.12), the mean welfare yielded by this treatment choice can take any value in the interval

$$[E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + c(t, x),\; E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + P(z\neq t\,|\,x=x) + c(t, x)].$$

The mean welfare of another treatment choice, say t′, can take any value in the interval

$$[E(y\,|\,x=x, z=t')\cdot P(z=t'\,|\,x=x) + c(t', x),\; E(y\,|\,x=x, z=t')\cdot P(z=t'\,|\,x=x) + P(z\neq t'\,|\,x=x) + c(t', x)].$$

Treatment t is definitely inferior to t′ if the upper bound of the former interval is no larger than the lower bound of the latter one. This gives Proposition 7.1.

Proposition 7.1: Let $(t, x) \in T \times X$. Using the empirical evidence alone, a treatment rule that assigns treatment t to persons with covariates x is dominated if and only if there exists a treatment $t' \in T$ such that

$$E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + P(z\neq t\,|\,x=x) + c(t, x) \le E(y\,|\,x=x, z=t')\cdot P(z=t'\,|\,x=x) + c(t', x). \tag{7.13}$$

□

The special case in which all treatments have the same cost is noteworthy. Then there generically are no dominated treatments. To see this, let $c(t, x) = c(t', x)$ for all treatments t and t′. Then inequality (7.13) reduces to

$$E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + P(z\neq t\,|\,x=x) \le E(y\,|\,x=x, z=t')\cdot P(z=t'\,|\,x=x).$$

Observe that $P(z\neq t\,|\,x=x) \ge P(z=t'\,|\,x=x)$, $E(y\,|\,x=x, z=t) \in [0, 1]$, and $E(y\,|\,x=x, z=t') \in [0, 1]$. Hence, this inequality can never hold strictly. Moreover, it holds weakly only when $P(z\neq t\,|\,x=x) = P(z=t'\,|\,x=x)$, $E(y\,|\,x=x, z=t) = 0$, and $E(y\,|\,x=x, z=t') = 1$.

Choice Among Undominated Treatments

The fact that the empirical evidence does not enable determination of the optimal treatment rule does not imply that a planner should be paralyzed, unwilling and unable to choose a rule. It implies only that the planner cannot assert optimality for whatever rule he does choose.

The planner could, for example, reasonably apply the maximin rule, which calls for persons with covariates x to be assigned the treatment that maximizes the lower bound on $E[y(\cdot)\,|\,x=x]$. By (7.12), the maximin rule solves the problem

$$\max_{t\in T} E(y\,|\,x=x, z=t)\cdot P(z=t\,|\,x=x) + c(t, x). \tag{7.14}$$

This rule is simple to apply and to comprehend. From the maximin perspective, the desirability of treatment t increases with $E(y\,|\,x=x, z=t)$, the mean outcome experienced by persons who received this treatment, and with $P(z=t\,|\,x=x)$, the fraction of persons who received treatment t. The second factor gives form to the conservatism of the maximin rule—the more prevalent a treatment was in the study population, the more expedient it is to choose this treatment in the treatment population.

7.4. Instrumental Variables

Section 7.3 considered a planner who confronts the selection problem using the empirical evidence alone. Credible distributional assumptions may enable the planner to shrink the identification region for $E[y(\cdot)|x]$ and hence shrink the set of undominated treatment rules. Many distributional assumptions make use of instrumental variables.

Treatment-Specific Assumptions

The selection problem is a matter of missing outcome data, so all of the analysis of Chapter 2 may be brought to bear. Thus, suppose that person j is characterized by an observable covariate v_j in a finite space V. Let $P(y, z, x, v)$ denote the joint distribution of (y, z, x, v). The covariate v serving as an instrumental variable need not be distinct from the covariate x used to make treatment choices, but it simplifies analysis if v contains information not conveyed by x. Hence, the presentation here assumes that $P(v = v, z = t|x) > 0$ for all $(v, t) \in V \times T$.

Let $t \in T$ and $x \in X$. To help identify $E[y(t)|x = x]$, a planner could impose any of the distributional assumptions studied in Chapter 2. This section shows how Assumptions MAR, SI, MMAR, and MI may be applied to the analysis of treatment response. Assumptions MM and MMM will be considered separately in Chapter 9.

In the context of treatment response, Assumptions MAR, SI, MMAR, and MI are as follows.

Outcomes Missing-at-Random (Assumption MAR):

$$P[y(t)|x = x, v] = P[y(t)|x = x, v, z = t] = P[y(t)|x = x, v, z \neq t]. \quad (7.15)$$

Statistical Independence of Outcomes and Instruments (Assumption SI):

$$P[y(t)|x = x, v] = P[y(t)|x = x]. \quad (7.16)$$

Means Missing-at-Random (Assumption MMAR):

$$E[y(t)|x = x, v] = E[y(t)|x = x, v, z = t] = E[y(t)|x = x, v, z \neq t]. \quad (7.17)$$

Mean Independence of Outcomes and Instruments (Assumption MI):

$$E[y(t)|x = x, v] = E[y(t)|x = x]. \quad (7.18)$$

These assumptions restrict the distribution of outcomes under a specified treatment t for persons with specified covariates x. A planner can consider each value of (t, x) in turn and decide what assumption to assert.

Assumptions MAR and MMAR point-identify $E[y(t)|x = x]$. The other assumptions generally do not yield point identification but do shrink the identification region. Propositions 7.2 through 7.5 give the results. These propositions are immediate extensions of corresponding ones in Chapter 2, so proofs are omitted.

Proposition 7.2: Let assumption MAR hold. Then $P[y(t)|x = x]$ is point-identified with

$$P[y(t)|x = x] = \sum_{v \in V} P(y|x = x, v = v, z = t)P(v = v|x = x). \quad (7.19)$$

□

Proposition 7.3: Let assumption SI hold. Then the identification region for $P[y(t)|x = x]$ is

$$H_{SI}\{P[y(t)|x = x]\} = \bigcap_{v \in V} \{P(y|x = x, v = v, z = t)P(z = t|x = x, v = v) + \gamma_v \cdot P(z \neq t|x = x, v = v),\ \gamma_v \in \Gamma_Y\}. \quad (7.20)$$

□

Proposition 7.4: Let assumption MMAR hold. Then $E[y(t)|x = x]$ is point-identified with

$$E[y(t)|x = x] = \sum_{v \in V} E(y|x = x, v = v, z = t)P(v = v|x = x). \quad (7.21)$$

□

Proposition 7.5: Let assumption MI hold. Then the identification region for $E[y(t)|x = x]$ is the closed interval

$$H_{MI}\{E[y(t)|x = x]\} = [\max_{v \in V} E\{y \cdot 1[z = t]|x = x, v = v\},\ \min_{v \in V} E\{y \cdot 1[z = t] + 1[z \neq t]|x = x, v = v\}]. \quad (7.22)$$

□

Statistical Independence of Response Functions and Instruments

Whereas the assumptions considered above are treatment-specific, one could instead have information that restricts the joint distribution of the response function $y(\cdot)$. An especially prominent assumption is

Statistical Independence of Response Functions and Instruments (Assumption SI-RF):

$$P[y(\cdot)|x = x, v] = P[y(\cdot)|\, x = x]. \tag{7.23}$$

Assumption SI-RF strengthens assumption SI. The latter assumption, when applied to all treatments, asserts that each component $[y(t), t \in T]$ of the outcome vector $y(\cdot)$ is statistically independent of v. Assumption SI-RF asserts that $[y(t), t \in T]$ are jointly independent of v.

The prominence of assumption SI-RF derives from its credibility when the study population are subjects in a randomized experiment. In a randomized experiment, the instrumental variable v designates the treatment group in which each subject has been placed; thus $V = T$. Randomization implies that $y(\cdot)$ is statistically independent of the designated treatment v, so assumption SI-RF holds.

The classical theory of randomized experiments assumes that all subjects comply with their designated treatments; that is, $z = v$. In this special case, application of Proposition 7.3 to each treatment t shows that $P[y(t)|x = x]$ is point-identified, with

$$H_{SI}\{P[y(t)|x = x]\} = P(y|x = x, z = t). \tag{7.24}$$

This finding only uses assumption SI; it does not require the full force of assumption SI-RF.

When some subjects do not comply, randomization of designated treatments generally does not point-identify $P[y(t)|x = x]$. In this case, assumption SI-RF may have identifying power beyond that obtained when assumption SI is applied to all treatments. However, the form of identification regions under assumption SI-RF is largely an open question.[2]

Complement 7A. Identification and Ambiguity

The treatment-choice problem examined in this chapter is an instance of choice under *ambiguity*. In general, a decision maker with a known choice set who wants to maximize an unknown objective function is said to face a problem of choice under ambiguity. A common source of ambiguity is partial knowledge of a probability distribution describing a relevant population—the decision maker may know only that the distribution of interest is a member of some set of distributions. This is the generic situation of a decision maker who seeks to learn a population distribution empirically, but whose data and other information do not point identify the

distribution. Thus, identification problems in empirical analysis generate problems of choice under ambiguity.

The term *ambiguity* appears to originate in Ellsberg (1961), which posed thought experiments in which subjects were asked to draw a ball from either of two urns, one with a known distribution of colors and the other with an unknown distribution of colors. Keynes (1921) and Knight (1921) used the term *uncertainty*, but uncertainty has since come to be used to describe optimization problems in which the objective function depends on a known probability distribution.

Dominated Treatment Rules

Manski (2000) shows that the social planner of Section 7.2 faces a problem of treatment choice under ambiguity whenever identification problems prevent the planner from knowing enough about mean treatment response to be able to determine the optimal rule.

Considering the matter in abstraction, suppose that a planner learns from the available studies that $E[y(\cdot)|x]$ lies in some identification region $H\{E[y(\cdot)|x]\}$. This information may not suffice to solve problem (7.8) but may suffice to determine that some treatment rules are dominated.

A feasible treatment rule $z(\cdot)$ is dominated if there exists another feasible rule, say $z'(\cdot)$, that necessarily yields at least the social welfare of $z(\cdot)$ and that performs strictly better than $z(\cdot)$ in some possible state of nature. Thus $z(\cdot) \in Z(X)$ is dominated if there exists a $z'(\cdot) \in Z(X)$ such that

$$\sum_{x \in X} P(x = x)\{\sum_{t \in T} [\eta(t, x) + c(t, x)]\cdot 1[z(x) = t]\}$$
$$\leq \sum_{x \in X} P(x = x)\{\sum_{t \in T} [\eta(t, x) + c(t, x)]\cdot 1[z'(x) = t]\}$$

for all $\eta \in H\{E[y(\cdot)|x]\}$ and

$$\sum_{x \in X} P(x = x)\{\sum_{t \in T} [\eta(t, x) + c(t, x)]\cdot 1[z(x) = t]\}$$
$$< \sum_{x \in X} P(x = x)\{\sum_{t \in T} [\eta(t, x) + c(t, x)]\cdot 1[z'(x) = t]\}$$

for some $\eta \in H\{E[y(\cdot)|x]\}$, where $\eta(t, x)$ is a feasible value of $E[y(t)|x]$.

Choice Among Undominated Rules

The central difficulty of choice under ambiguity is that there is no clearly best way to choose among undominated actions. Two common suggestions are application of the *maximin rule* or a *Bayes decision rule*.

A planner using the maximin rule selects a treatment rule that maximizes the minimum mean welfare attainable under all possible states of nature. This means solution of the optimization problem

$$\max_{z(\cdot)\in Z^*(X)} \min_{\eta \in H\{E[y(\cdot)|x]\}} \sum_{x\in X} P(x=x)\{\sum_{t\in T} [\eta(t,x)+c(t,x)]\cdot 1[z(x)=t]\},$$

where $Z^*(X)$ denotes the set of undominated treatment rules.

Bayesian decision theorists recommend that a decision maker facing ambiguity place a subjective distribution on the states of nature and maximize expected welfare with respect to this distribution. In the treatment choice context, the planner would place a probability measure π on the set $H\{E[y(\cdot)|x]\}$, where π expresses the decision maker's personal beliefs about where $E[y(\cdot)|x]$ may lie within $H\{E[y(\cdot)|x]\}$. The planner would then solve the optimization problem

$$\max_{z(\cdot)\in Z^*(X)} \int \sum_{x\in X} P(x=x)\{\sum_{t\in T} [\eta(t,x)+c(t,x)]\cdot 1[z(x)=t]\}\, d\pi.$$

The maximin rule and Bayes rules are reasonable ways to make decisions under ambiguity, but there is no optimal way to behave in the absence of credible information on the location of $E[y(\cdot)|x]$ within $H\{E[y(\cdot)|x]\}$. Wald (1950), who proposed and studied the maximin rule, did not contend that the rule is optimal, only that it is reasonable. Considering the case in which the objective is to minimize rather than maximize an objective function, he wrote (Wald, 1950, p.18): "a minimax solution seems, in general, to be a reasonable solution of the decision problem."

Bayesians often present *procedural rationality* arguments for use of Bayes decision rules. Savage (1954) showed that a decision maker whose choices are consistent with a certain set of axioms can be interpreted as using a Bayes rule. Many decision theorists consider the Savage axioms, or other sets of axioms, to be a priori appealing. Acting in a manner that is consistent with these axioms does not, however, imply that chosen actions yield good outcomes. Berger (1985, p. 121) calls attention to this, stating: "A Bayesian analysis may be 'rational' in the weak axiomatic sense, yet be terrible in a practical sense if an inappropriate prior distribution is used."

Complement 7B: Sentencing and Recidivism

The question of how judges should sentence convicted juvenile offenders has long been of interest to policy makers and criminologists. Manski and

Nagin (1998) analyzed data on the sentencing of 13,197 juvenile offenders in Utah and their subsequent recidivism. We compared recidivism under the two main sentencing options available to judges: confinement in residential facilities ($t = 1$) and sentences that do not involve confinement ($t = 0$).

Let the outcome take the value $y = 1$ if an offender is convicted of a subsequent crime in the two-year period following sentencing, and the value $y = 0$ otherwise. The empirical distribution of treatments and outcomes among the observed offenders was found to be as follows:

residential treatment: $P(z = 1) = 0.11$,
recidivism given residential treatment: $P(y = 1 | z = 1) = 0.77$,
recidivism given nonresidential treatment: $P(y = 1 | z = 0) = 0.59$.

The problem is to use this empirical evidence to draw conclusions about the response probabilities $P[y(1) = 1]$ and $P[y(0) = 1]$.

The empirical evidence alone reveals that

$$P[y(1) = 1] \in [0.08,\ 0.97] \qquad P[y(0) = 1] \in [0.53, 0.64].$$

If one assumes that judges sentence offenders randomly, then

$$P[y(1) = 1] = 0.77 \qquad P[y(0) = 1] = 0.59.$$

Random sentencing did not seem a credible assumption, so we considered two alternative models of treatment selection. The *outcome optimization* model assumes that judges aim to minimize the chance of recidivism. The empirical evidence plus this assumption can be shown to imply that

$$P[y(1) = 1] \in [0.61,\ 0.97] \qquad P[y(0) = 1] \in [0.61, 0.64].$$

The *skimming* model assumes that judges classify offenders as "higher risk" or "lower risk," sentencing only the former to residential confinement. The empirical evidence plus this assumption can be shown to imply that

$$P[y(1) = 1] \in [0.08,\ 0.77] \qquad P[y(0) = 1] \in [0.59, 0.64].$$

Thus, conclusions about response to treatment depend critically on the assumptions imposed.