LECTURES ON PARTIAL IDENTIFICATION
University of Stavanger, August 30–September 3, 2010

Charles F. Manski
Department of Economics, Northwestern University

August 30. Lectures 1 and 2: Prediction with Incomplete Data

August 31. Lectures 3 and 4: Analysis of Treatment Response

September 1. Lectures 5 and 6: Planning under Ambiguity

September 2.
  Lecture 7: Partial Identification of Counterfactual Choice Probabilities
  Lecture 8: Rounding Probabilistic Expectations in Surveys

September 3. Lectures 9 and 10: Identification of Treatment Response with Social Interactions

Sources by Lecture

Lectures 1 and 2: IPD Chapters 1 through 4

Lectures 3 and 4: IPD Chapters 7 through 10

Lectures 5 and 6: IPD Chapter 11

Lecture 7: IPD Chapter 13

Lecture 8

Lectures 9 and 10
Statistical inference uses sample data to draw conclusions about a population probability distribution of interest.

Data alone do not suffice. Informative inference always requires assumptions about the sampling process. It often requires assumptions about the population.
Identification and Statistical Inference


“In our discussion we have used the phrase “a parameter that can be determined from a sufficient number of observations.” We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from “a sufficiently large number of observations” we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is suspect.”
It has been common to think of identification as a binary event—a parameter is either identified or it is not.

It has been traditional to combine available data with assumptions strong enough to yield point identification. However, these assumptions often are not well motivated, and researchers often debate their validity.

Empirical researchers should be concerned with the credibility of inference.

Credibility is a subjective matter, yet I take there to be wide agreement on a principle that I have called

*The Law of Decreasing Credibility:* The credibility of inference decreases with the strength of the assumptions maintained.

This principle implies that empirical researchers face a dilemma as they decide what assumptions to maintain. Stronger assumptions yield inferences that may be tighter but less credible.

Methodological research cannot resolve the dilemma but can clarify its nature.
Studies of Partial Identification

Consider inference when a specified sampling process generates data from a specified population.

First study inference on the population probability distribution when no assumptions are placed on this distribution. The usual finding is a set-valued *identification region* (or *identified set*). When this region is smaller than the set of all logically possible distributions, the distribution is *partially identified*.

Then ask how the identification region shrinks if specified assumptions are imposed. This quantifies the *identifying power* of the assumptions.

Researchers often want to learn particular parameters of a probability distribution. Study of identification of the distribution yields findings on the identification of parameters.
I recommend first considering weaker, more credible assumptions and then stronger, less credible ones.

Conservative analysis enables researchers to learn from available data without imposing untenable assumptions. It enables establishment of a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate. It makes plain the limitations of the available data.
Identification for Prediction and Decisions

The basic prediction problem is to learn the probability distribution $P(y|x)$ of an outcome $y$ conditional on covariates $x$.

In many decision problems, the relative merits of alternative actions depend on an outcome distribution. I study how a decision maker might reasonably choose an action when the available data and credible assumptions only partially identify this distribution. This is decision making under ambiguity.

Note: Some research is not concerned with prediction or decision making. Scientists sometimes motivate research as an effort to “understand” a phenomenon or to determine “causality.” I do not do this.
MISSING OUTCOMES

Inference with missing outcomes is a matter of contemplating all values that the missing data might take. The set of feasible outcome distributions is determined by considering all logically possible configurations of the missing data.

Let each member of the population be characterized by a triple \((y, z, x)\). Here \(y \in X\) is the outcome to be predicted and \(x \in X\) are observable covariates. The indicator \(z = 1\) if \(y\) is observable and \(z = 0\) otherwise.

A sampling process draws persons at random from the population. For each \(i = 1, \ldots, \infty\), outcome \(y_i\) is observed if \(z_i = 1\) and missing if \(z_i = 0\).

The objective is to use the available data to learn about \(P(y|x)\).
Application: Nonresponse in Survey Research

Nonresponse is a perennial concern in survey research. Some persons are not interviewed and some who are interviewed do not answer some questions. Longitudinal surveys experience attrition.

Application: Non-Observable Counterfactual Outcomes

Analysis of treatment response aims to predict the outcomes that would occur if alternative treatment rules were applied to a population. A fundamental problem is that one cannot observe the outcomes a person would experience under all treatments. At most one can observe the outcome that a person experiences under the treatment he actually receives. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.
Anatomy of the Problem

By the Law of Total Probability

\[ P(y \mid x) = P(y \mid x, z = 1)P(z = 1 \mid x) + P(y \mid x, z = 0)P(z = 0 \mid x). \]

The sampling process reveals \( P(y \mid x, z = 1)P(z = 1 \mid x) \) and \( P(z = 0 \mid x) \).

The sampling process is uninformative regarding \( P(y \mid x, z = 0) \). Hence, \( P(y \mid x) \) lies in the identification region

\[ \tilde{H}[P(y \mid x)] = [P(y \mid x, z = 1)P(z = 1 \mid x) + \gamma P(z = 0 \mid x), \gamma \in \Gamma_Y], \]

where \( \Gamma_Y \) denotes the set of all probability distributions on the set \( Y \).

The identification region is informative when \( P(z = 0 \mid x) < 1 \) and is the single distribution \( P(y \mid x, z = 1) \) when \( P(z = 0 \mid x) = 0 \). Hence, \( P(y \mid x) \) is partially identified when \( 0 < P(z = 0 \mid x) < 1 \) and is point-identified when \( P(z = 0 \mid x) = 0 \).
Identification of Parameters

The above concerns identification of the entire outcome distribution. A common objective is to infer a parameter of this distribution. For example, one may want to learn the conditional mean \( \operatorname{E}(y|x) \).

Let \( \theta(\cdot) \) be a function mapping probability distributions on \( Y \) into the real line. Consider inference on the parameter \( \theta[P(y|x)] \). The identification region is the set of all values \( \theta \) can take when \( P(y|x) \) ranges over its feasible values. Thus,

\[
\mathcal{H}\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in \mathcal{H}[P(y|x)]\}.
\]
By the Law of Total Probability,

\[ P(y \in B \mid x) = P(y \in B \mid x, z = 1)P(z = 1 \mid x) + P(y \in B \mid x, z = 0)P(z = 0 \mid x). \]

The sampling process reveals \( P(z \mid x) \) and \( P(y \in B \mid x, z = 1) \), but is uninformative about \( P(y \in B \mid x, z = 0) \). This quantity lies between zero and one. This yields the “worst-case” bound on \( P(y \in B \mid x) \):

\[
P(y \in B \mid x, z = 1)P(z = 1 \mid x) \leq P(y \in B \mid x) \leq P(y \in B \mid x, z = 1)P(z = 1 \mid x) + P(z = 0 \mid x).
\]

The bound is sharp. That is, the lower and upper bounds are the smallest and largest feasible values of \( P(y \in B \mid x) \). The width is \( P(z = 0 \mid x) \).

The identification region is the interval between the lower and upper bounds.
The Distribution Function

Let $y$ be real-valued. Let $B = (-\infty, t]$ for a specified $t$. Then the bound is

$$P(y \leq t \mid x, z = 1)P(z = 1 \mid x) \leq P(y \leq t \mid x)$$

$$\leq P(y \leq t \mid x, z = 1)P(z = 1 \mid x) + P(z = 0 \mid x).$$

The feasible distribution functions are all increasing functions of $t$ that lie within the bound for all values of $t$.

Quantiles

The bound on the distribution function can be inverted to bound quantiles of $P(y \mid x)$.

This bound is informative on both sides if $P(z = 0 \mid x) < \min(\alpha, 1 - \alpha)$. Then

$$[\alpha - P(z = 0 \mid x)]/P(z = 1 \mid x)$$-quantile of $P(y \mid x, z = 1)$

$$\leq \alpha$$-quantile of $P(y \mid x) \leq$$

$$[\alpha/P(z = 1 \mid x)]$$-quantile of $P(y \mid x, z = 1)$. 
Means of Functions of y

Consider $E[g(y)]$, where $g(\cdot)$ has range $[g_0, g_1]$.

The Law of Iterated Expectations gives

$$E[g(y) \mid x] = E[g(y) \mid x, z = 1] P(z = 1 \mid x) + E[g(y) \mid x, z = 0] P(z = 0 \mid x).$$

The sampling process reveals $E[g(y) \mid x, z = 1] P(z = 1 \mid x)$. The data are uninformative about $E[g(y) \mid x, z = 0]$, which can take any value in the interval $[g_0, g_1]$. Hence, the identification region for $E[g(y) \mid x]$ is

$$\mathcal{H}\{E[g(y) \mid x]\} = [E[g(y) \mid x, z = 1] P(z = 1 \mid x) + g_0 P(z = 0 \mid x),
E[g(y) \mid x, z = 1] P(z = 1 \mid x) + g_1 P(z = 0 \mid x)].$$

This interval has width $(g_1 - g_0) P(z = 0 \mid x)$.

If $g(\cdot)$ is a bounded function, the severity of the identification problem varies directly with the probability of missing data. If $g(\cdot)$ is unbounded, the sampling process per se is uninformative about $E[g(y) \mid x]$.

Contrast this result with that for $M[g(y) \mid x]$. There the sampling process is informative when $P(z = 0 \mid x) < \frac{1}{2}$, regardless of whether $g(\cdot)$ is bounded.
Estimation of Identification Regions

The above identification regions are functions of \( P(y \mid x, z = 1) \) and \( P(z \mid x) \). Given sample data, they can be consistently estimated by the sample analogs \( P_N(y \mid x, z = 1) \) and \( P_N(z \mid x) \). Thus, \( \mathcal{H}[P(y \mid x)] \) can be consistently estimated by

\[
H_N[P(y \mid x)] = [P_N(y \mid x, z = 1)P_N(z = 1 \mid x) + \gamma P_N(z = 0 \mid x), \gamma \in \Gamma_Y].
\]

Correspondingly, \( \mathcal{H}\{E[g(y) \mid x]\} \) can be consistently estimated by

\[
H_N\{E[g(y) \mid x]\} = [E_N[g(y) \mid x, z = 1]P_N(z = 1 \mid x) + g_0P_N(z = 0 \mid x),
E_N[g(y) \mid x, z = 1]P_N(z = 1 \mid x) + g_1P_N(z = 0 \mid x)].
\]

Note: The sample analog of an identification region is a function of the sample data and, hence, is a random set.
Confidence Sets

The statistics literature on point estimation of parameters has used confidence sets to measure sampling imprecision. The standard definition of a confidence set applies to parameters that are partially identified. One can also define confidence sets for identification regions.

Consider $E[g(y)|x]$. A confidence set can be constructed by widening the estimate given above to

$$[E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_0P_N(z = 0|x) - \delta_{0N},$$
$$E_N[g(y)|x, z = 1]P_N(z = 1|x) + g_1P_N(z = 0|x) + \delta_{1N}].$$

$\delta_{0N} > 0$ and $\delta_{1N} > 0$ are suitably chosen data-dependent numbers.
DISTRIBUTIONAL ASSUMPTIONS

Inference with no distributional assumptions provides a natural starting point for empirical analysis but ordinarily will not be the ending point. Having determined what can be learned without assumptions, one should then ask what more can be learned if plausible assumptions are imposed.
Missingness at Random

A common practice has been to assume that data are missing at random. Formally, the assumption is

\[ P(y|x, z = 0) = P(y|x, z = 1). \]

It follows that \( P(y|x) = P(y|x, z = 1). \)

This assumption is not *refutable* (or *testable*). The available data are uninformative about \( P(y|x, z = 0). \) Hence, it is logically possible that this distribution is the same as \( P(y|x, z = 1). \)
Refutable and Non-Refutable Assumptions

Assume that $P(y|x, z = 0) \in \Gamma_{0Y}$ for a specified $\Gamma_{0Y} \subset \Gamma_Y$. This assumption is non-refutable. The identification region for $P(y|x)$ is

$$
\mathcal{H}_0[P(y|x)] = [P(y|x z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{0Y}].
$$

Assume that $P(y|x) \in \Gamma_{1Y}$ for a specified $\Gamma_{1Y} \subset \Gamma_Y$. This assumption may be refutable. The data alone reveal that $P(y|x) \in \mathcal{H}[P(y|x)]$. Hence, the identification region with the assumption is

$$
\mathcal{H}_1[P(y|x)] = \mathcal{H}[P(y|x)] \cap \Gamma_{1Y}.
$$

The assumption is refutable if there exist possible values for $P(y|x z = 1)$ and $P(z|x)$ such that $\mathcal{H}_1[P(y|x)]$ empty. If this set is empty, $P(y|x)$ cannot lie in $\Gamma_{1Y}$. 
Refutability and Credibility

It is important not to confuse refutability with credibility.

Refutability is a matter of logic. Credibility is a subjective matter.

Refutability is a property of an assumption and the empirical evidence. An assumption is refutable if it is inconsistent with some possible configuration of the empirical evidence. It is non-refutable otherwise.

Credibility is a property of an assumption and the person contemplating it. An assumption is credible to the degree that someone thinks it so.
Wage Regressions

A major problem of missing outcome data in labor economics occurs in efforts to estimate wage regressions, which measure how market wages vary with schooling, work experience, and demographic background.

Surveys provide covariate data for each respondent and wage data for those who work. Surveys do not ordinarily provide wage data for respondents who do not work. Economists consider these wages to be well-defined but unobserved. They are the counterfactual wages that non-workers would earn if they were to work.
The Reservation Wage Model of Labor Supply

Assume each person knows the wage $y$ he would receive if he were to work. The person chooses to work if $y > R$, called the reservation wage, and chooses not to work if $y < R$.

Wages are observed when $y > R$ and are missing when $y < R$. Assume that $P(y = R) = 0$. Then the reservation-wage model implies that

$$
P(y | x, z = 1) = P(y | x, y > R),$$
$$
P(y | x, z = 0) = P(y | x, y < R),$$

The reservation-wage model per se has no identifying power. The model places no assumptions on the distribution $P(y | x, y < R)$. To have identifying power, the model must be augmented by assumptions on the distribution $P(y, R | x)$ of market and reservation wages.
Wages Missing at Random

Wage data are missing at random if \( P(y \mid x, y < R) = P(y \mid x, y > R) \).

This is logically possible. In particular, it occurs if \( R = y + u \), where \( u \perp y \).

The question is whether this distributional assumption is credible.
Homogeneous Reservation Wages

Assume that all persons with covariates x have the same reservation wage \( R(x) \). This homogeneous-reservation-wage assumption partially identifies \( P(y|x) \).

Let \( y^*(x) \) denote the smallest observed wage for persons with covariates x. Then \( y^*(x) > R(x) \). Hence, for all \( t \geq y^*(x) \),

\[
P(y \leq t | x) = P(y \leq t | x, y < R(x)) = 1.
\]

It follows that for all such \( t \),

\[
P(y \leq t | x) = P(y \leq t | x, z = 1)P(z = 1 | x) + P(y \leq t | x, z = 0)P(z = 0 | x)
= P(y \leq t | x, z = 1)P(z = 1 | x) + P(z = 0 | x).
\]

Thus, \( P(y \leq t | x) \) is point-identified for \( t \geq y^*(x) \).

The \( \alpha \)-quantile of \( P(y|x) \) is point-identified when \( \alpha > P(z = 0 | x) \). In particular, \( Q_\alpha(y|x) = Q_\alpha(y|x, z = 1) \), where \( a \equiv [\alpha - P(z = 0 | x)]/P(z = 1 | x) \).

The question is whether this distributional assumption is credible.
The Normal-Linear Model of Market and Reservation Wages

A common practice has been to restrict \( P(y, R \mid x) \) to a parametric family of distributions. The normal-linear model has received considerable attention. This assumes that

\[
P(\log y, \log R \mid x) \sim N[(x_1, x_2), \Sigma].
\]

This assumption reduces inference on \( P(y \mid x) \) to inference on \( (\beta_1, \beta_2, \Sigma) \).

If the assumption is correct and \( x \) has full rank, there exists one value of \( (\beta_1, \beta_2, \Sigma) \) that implies the observed \( P(y \mid x, z = 1) \) and \( P(z \mid x) \). Hence, point identification is the norm.

The assumption is refutable. If it is incorrect, typically no value of the parameters generates the observed \( P(y \mid x, z = 1) \) and \( P(z \mid x) \).

The question is whether this distributional assumption is credible.
**Selection Models**

The normal-linear model of market and reservation wages exemplifies a class of parametric selection models that describe how outcomes and missingness vary with observed and unobserved covariates. The expression “selection model” is used because these models aim to explain missingness as the result of a choice, such as the choice to work in the reservation wage model.

The selection models used in empirical research usually have the form

\[ y = xb + \delta, \]

\[ z = 1[xc + \epsilon > 0], \]

\[ P(\delta, \epsilon \mid x) \sim N(0, \Sigma). \]
Parametric Mean Regression with Missing Outcomes

Drop all the assumptions of the parametric selection model except for the mean regression assumption

$$E(y|x) = f(x, b).$$

If $y$ is bounded, this parametric model partially identifies $b$.

Consider the joint identification region for $[E(y|x = \xi), \xi \in X]$ using the empirical evidence alone. A parameter value $b \in B$ is feasible if and only if the implied values of $[f(\xi, b), \xi \in X]$ lie within this region; that is, if

$$[f(\xi, b), \xi \in X] \in H[E(y|x = \xi), \xi \in X],$$

where

$$H[E(y|x = \xi), \xi \in X] = \times_{\xi \in X} H[E(y|x = \xi)],$$

and

$$H[E(y|x = \xi)] = [E(y|x = \xi, z = 1)P(z = 1|x = \xi) + y_0 P(z = 0|x = \xi),$$

$$E(y|x = \xi, z = 1)P(z = 1|x = \xi) + y_1 P(z = 0|x = \xi)].$$

Here $y_0$ and $y_1$ are the smallest and largest logically possible values of $y$. 
Thus, $b$ is feasible if and only if

$$E(y \mid x = \xi, z = 1)P(z = 1 \mid x = \xi) + y_0P(z = 0 \mid x = \xi) \leq f(\xi, b)$$

$$\leq E(y \mid x = \xi, z = 1)P(z = 1 \mid x = \xi) + y_1P(z = 0 \mid x = \xi)$$

for all $\xi \in X$.

Let $B_0$ denote the subset of $B$ for which these inequalities hold. Then $B_0$ is the identification region for $b$. It follows that $[f(\xi, b), \xi \in X]$, $b \in B_0$ is the identification region for $[E(y \mid x = \xi), \xi \in X]$.

The parametric regression model is refutable if it is logically possible for $B_0$ to be empty. In that event, no value of $b$ yields a function $f(x, b)$ that equals a feasible value of $E(y \mid x)$. Hence, the model is incorrect.

The set $B_0$ can be consistently estimated by the modified-minimum-distance method.
ANALYSIS OF TREATMENT RESPONSE

Studies of treatment response aim to predict the outcomes that would occur if alternative treatment rules were applied to a population.

One cannot observe the outcomes that a person would experience under all treatments. At most, one can observe a person’s realized outcome; that is, the one he experiences under the treatment he actually receives. The counterfactual outcomes that a person would have experienced under other treatments are logically unobservable.

Example: Suppose that patients ill with a specified disease can be treated by drugs or surgery. The relevant outcome might be life span. One may want to predict the life spans that would occur if all patients of a certain type were to be treated by drugs. The available data may be observations of the realized life spans of patients in a study population, some of whom were treated by drugs and the rest by surgery.
The Selection Problem

Let the set $T$ list all feasible treatments.

Let each member $j$ of a study population have covariates $x_j \in X$.

Let each $j$ have a response function $y_j(\cdot): T \rightarrow Y$ that maps the mutually exclusive and exhaustive treatments $t \in T$ into outcomes $y_j(t) \in Y$. Thus, $y_j(t)$ is the outcome that person $j$ would realize if he were to receive treatment $t$. $y_j(t)$ is a potential, latent, or conjectural outcome.

Let $z_j \in T$ denote the treatment received by person $j$. Then $y_j \equiv y_j(z_j)$ is the realized outcome. The outcomes $[y_j(t), t \neq z_j]$ he would have experienced under other treatments are counterfactual.

Observation may reveal the distribution $P(y, z | x)$ of realized outcomes and treatments for persons with covariates $x$. Observation cannot reveal the distribution of counterfactual outcomes.
Consider prediction of the outcomes that would occur if all persons with the same observed covariates were to receive the same treatment.

Prediction of outcomes under a policy mandating treatment \( t \) for persons with covariates \( x \) requires inference on \( P[y(t) \mid x] \).

The problem of identification of \( P[y(t) \mid x] \) from knowledge of \( P(y, z \mid x) \) is called the *selection problem*. This expression refers to the fact that treatment selection determines which potential outcome is observable.
Prediction Using the Empirical Evidence Alone

The selection problem has the same structure as the missing-outcomes problem.

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

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

The first equality is the Law of Total Probability. The second holds because \(y(t)\) is the outcome realized by persons who receive treatment \(t\).

The identification region for \(P(y(t)|x)\) using the data alone is

\[
\mathcal{H}\{P(y(t)|x)\} = \{P(y|x, z = t)P(z = t|x) + \gamma P(z \neq t|x), \gamma \in \Gamma_y\}.
\]

This has the same form as the identification region for \(P(y|x)\) when outcome data are missing. The outcome there was \(y\) and \(\{z = 1\}\) indicated observability of \(y\). The outcome here is \(y(t)\) and \(\{z = t\}\) indicates observability of \(y(t)\).
Average Treatment Effects

Researchers often compare policies mandating alternative treatments, say \( t \) and \( t' \). It is common to use data on realized treatments and outcomes to infer the *average treatment effect* \( E[y(t) \mid x] - E[y(t') \mid x] \).

Let \( Y \) have smallest and largest elements \( y_0 \) and \( y_1 \) respectively. The identification region for the average treatment effect is the interval

\[
\{E[y(t) \mid x] - E[y(t') \mid x]\} =
\]

\[
[E(y \mid x, z = t)P(z = t \mid x) + y_0P(z \neq t \mid x) - E(y \mid x, z = t')P(z = t' \mid x) - y_1P(z \neq t' \mid x),
\]

\[
E(y \mid x, z = t)P(z = t \mid x) + y_1P(z \neq t \mid x) - E(y \mid x, z = t')P(z = t' \mid x) - y_0P(z \neq t' \mid x)].
\]

This interval necessarily contains the value zero. Its width is

\[
(y_1 - y_0)[P(z \neq t \mid x) + P(z \neq t' \mid x)] = (y_1 - y_0)[2 - P(z = t \mid x) - P(z = t' \mid x)].
\]

Let \( t \) and \( t' \) be the only feasible treatments in the study population. Then \( P(z = t \mid x) + P(z = t' \mid x) = 1 \). Hence, the interval has width \( (y_1 - y_0) \).
Treatment at Random

The analog to the assumption of outcomes missing at random is

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

This is credible in classical randomized experiments, where an explicit randomization mechanism has been used to assign treatments and all persons comply with their treatment assignments. Its credibility in other settings is almost invariably a matter of controversy.
Illustration: Sentencing and Recidivism

Consider how the sentencing of offenders may affect recidivism.

Data are available on the outcomes experienced by offenders given the sentences that they receive. However, researchers have long debated the counterfactual outcomes that offenders would experience if they were to receive other sentences. Moreover, the sentencing rules that judges actually use are largely unknown. Thus, predicting the response of criminality to sentencing might reasonably be studied using the empirical evidence alone.

Manski and Nagin (1998) analyzed data on the sentencing and recidivism of males in the state of Utah who were born from 1970 through 1974 and who were convicted of offenses before they reached age 16. We compared recidivism under the two main sentencing options available to judges: confinement in residential facilities (t = b) and sentences that do not involve residential confinement (t = a). The outcome of interest was taken to be a binary measure of recidivism, with \( y = 1 \) if an offender is not convicted of a subsequent crime in the two-year period following sentencing, and \( y = 0 \) if the offender is convicted of a subsequent crime.
The data reveal that

Probability of residential treatment: $P(z = b) = 0.11$

Recidivism probability in sub-population receiving residential treatment:

$P(y = 0 \mid z = b) = 0.77$

Recidivism probability in sub-population receiving nonresidential treatment: $P(y = 0 \mid z = a) = 0.59$.

Consider two policies, one mandating residential treatment for all offenders and the other mandating non-residential treatment. The recidivism probabilities under these policies are $P[y(b) = 0]$ and $P[y(a) = 0]$ respectively.

Assuming treatment at random,

$P[y(b) = 0] = P(y = 0 \mid z = b) = 0.77$

$P[y(a) = 0] = P(y = 0 \mid z = a) = 0.59$.

Using the data alone,

$H\{P[y(b) = 0]\} = [0.08, 0.97] \quad H\{P[y(a) = 0]\} = [0.53, 0.64]$.

The identification region for the average treatment effect is

$H\{P[y(b) = 0] - P[y(a) = 0]\} = [-0.56, 0.44]$. 
Assumptions Linking Outcomes Across Treatments

Thus far, analysis of treatment response is simply an instance of missing outcomes. The problem takes a distinctive character when one poses assumptions that link outcomes across treatments. Then observation of the realized outcome $y_j$ can be informative about the counterfactual outcomes $y_j(t), t \neq z_j$.

The *perfect-foresight optimization* model assumes $y_j(t) \leq y_j$ for all $t$ and $j$.

*Monotone treatment response* assumes that treatments are ordered and $y_j(\cdot)$ is monotone in the treatment. Hence,

\[
    t > z \implies y_j(t) \geq y_j. \quad t < z \implies y_j(t) \leq y_j.
\]
**Homogeneous Linear Response**

Let treatments be real-valued. Assume that each response function is linear in $t$, with slope that is homogeneous across the population. Thus,

$$y_j(t) = \beta t + \epsilon_j.$$ 

The homogeneous-linear-response assumption per se reveals nothing about the slope parameter $\beta$. For each person $j$, only $(y_j, z_j)$ are observable. This observable pair satisfies the equation

$$y_j = \beta z_j + \epsilon_j.$$ 

Given any conjectured value for $\beta$, this equation is satisfied by setting the unobserved value of $\epsilon_j$ equal to $y_j - \beta z_j$. Hence, the assumption and the empirical evidence are uninformative about $\beta$. 

“The” Instrumental Variable Estimator

Let \( v \) be a real covariate, called an *instrumental variable*. Assume that

\[
\text{Cov}(v, \epsilon) = 0, \quad \text{and} \quad \text{Cov}(v, z) \neq 0.
\]

Then

\[
0 = \text{Cov}(v, \epsilon) = \text{Cov}(v, y - \beta z) \\
= E[v(y - \beta z)] - E(v)E(y - \beta z) \\
= E(vy) - \beta E(vz) - E(v)E(y) + \beta E(v)E(z) = \text{Cov}(v, y) - \beta \text{Cov}(v, z).
\]

Hence, \( \beta \) is point-identified, with

\[
\beta = \frac{\text{Cov}(v, y)}{\text{Cov}(v, z)}.
\]

Researchers sometimes call the sample analog of \( \frac{\text{Cov}(v, y)}{\text{Cov}(v, z)} \) “the” instrumental variables estimator. This designation has historical foundation but no compelling scientific basis. There are many distributional assumptions using instrumental variables. Each such assumption may be combined with a variety of restrictions on the shapes of response functions.
Randomized Experiments

Many researchers argue that the assumption of treatment-at-random should have primacy among all of the assumptions that might be brought to bear in analysis of treatment response. That is,

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

The rationale for giving this assumption special status is the strong credibility that it enjoys in classical randomized experiments.
The classical argument for experiments with randomly assigned treatments is generally attributed to Fisher (1935) and can be phrased as follows:

Let random samples of persons be drawn from the population of interest and formed into treatment groups. Let all members of a treatment group be assigned the same treatment and suppose that each subject complies with the assigned treatment. Then the distribution of outcomes experienced by the members of a treatment group will be the same (up to random sampling error) as would be observed if the treatment in question were received by all members of the population.

The argument applies both to controlled experiments, in which a researcher purposefully randomizes treatment assignments, and to so-called natural experiments, in which randomization is a consequence of some process external to the research project. The randomization mechanism is irrelevant. What matters is that randomization makes it credible to assume that $z$ is statistically independent of $y(\cdot)$. 
Experiments in Practice

Classical randomized experiments have clear appeal, but they typically cannot be performed in practice.

(1) The classical argument supposes that subjects are drawn at random from the population of interest. Yet participation in experiments ordinarily cannot be mandated in democracies. Hence, experiments in practice usually draw subjects at random from a pool of persons who volunteer to participate. So one learns about treatment response within the population of volunteers rather than within the population of interest.

(2) The argument supposes that all participants in the experiment comply with their assigned treatments. In practice, subjects often do not comply.

(3) The argument supposes that one observes the realized treatments, outcomes, and covariates of all participants in the experiment. In practice, experiments may have missing data. A particularly common problem is missing outcome data when researchers lose contact with participants before their outcomes can be recorded.

Hence, identification problems typically occur in experiments in practice.
The Mixing Problem

The mixing problem arises when data from a randomized experiment are used to predict the outcomes of a policy that offers but does not mandate treatments.

The mixing problem is the converse of the selection problem. The selection problem arises when treatment selection in the study population is decentralized, but treatment would be mandated under the policy of interest.

Conversely, the mixing problem arises when treatment selection in the study population is mandated (as in a randomized experiment), but treatment would be decentralized under the policy of interest.
Consider choice among alternative preschool policies. Experiments with preschool interventions have sought to learn the outcomes that occur when members of treatment and control groups respectively enroll and do not enroll in a proposed program. Observation of the treatment/control group reveals outcomes if the program is mandated/unavailable.

Suppose that the policy under consideration would make the program available, but not mandated. The experiment does not reveal outcomes under this policy. Outcomes depend on program participation and on the joint distribution of treatment response, quantities that the experiment does not reveal.
**The Mixing Problem with Data from the Perry Preschool Project**

---

### Experimental Evidence on High School Graduation

\[ P[y(b) = 1] = 0.67 \quad P[y(a) = 1] = 0.49 \]

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Graduation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>experimental evidence alone</td>
<td>[0.16, 1]</td>
</tr>
<tr>
<td>5/10 population receives treatment b</td>
<td>[0.17, 0.99]</td>
</tr>
<tr>
<td>statistically independent outcomes</td>
<td>[0.33, 0.83]</td>
</tr>
<tr>
<td>monotone treatment response</td>
<td>[0.49, 0.67]</td>
</tr>
<tr>
<td>outcome maximization</td>
<td>[0.67, 1]</td>
</tr>
<tr>
<td>+ statistically independent outcomes</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Econometric analyses of treatment response commonly use instrumental variable (IV) assumptions to identify treatment effects. Yet the credibility of IV assumptions is often a matter of considerable disagreement. There is therefore good reason to consider weaker but more credible assumptions. To this end, we introduce monotone instrumental variable (MIV) assumptions and the important special case of monotone treatment selection (MTS). We study the identifying power of MIV assumptions alone and combined with the assumption of monotone treatment response (MTR). We present an empirical application using the MTS and MTR assumptions to place upper bounds on the returns to schooling.
The Returns to Schooling

Labor economists studying schooling as a treatment suppose that each individual $j$ has a log(wage) function $y_j(t)$, giving the log(wage) that $j$ would receive were he to obtain $t$ years of schooling. Observing realized covariates, schooling, and wages in the population, labor economists may seek to learn the mean returns to completing $t$ years of schooling relative to $s$ years. Some researchers assume that treatment selection is exogenous. Some use family or environmental attributes as instrumental variables.

The *monotone treatment selection* (MTS) assumption asserts that persons who select higher levels of schooling have weakly higher mean wage functions than do those who select lower levels of schooling.

The *monotone treatment response* (MTR) assumption asserts that, ceteris paribus, each person’s wage increases weakly as a function of conjectured years of schooling.

Combining these assumptions yields informative upper bounds on the returns to schooling.
Basic Ideas in the Analysis of Treatment Response

Each member $j$ of a population $J$ has a *response function* $y_j(\cdot): T \rightarrow Y$ mapping the mutually exclusive and exhaustive *treatments* $t$ into *outcomes* $y_j(t)$.

Let $z_j$ denote the treatment that $j$ actually receives. Then $y_j = y_j(z_j)$ is $j$'s *realized outcome* and $y_j(t), t \neq z_j$ are *latent outcomes*.

$P[y(\cdot), z, y]$ gives the population distribution of response functions, realized treatments, and realized outcomes.

Response functions are not observable but realized treatments and outcomes are observable. So the analyst can empirically learn $P(z, y)$ by collecting data from a random sample of individuals. The analyst's problem is to combine this empirical evidence with prior information in order to learn about $P[y(\cdot)]$. 
Some Assumptions On The Distribution of Outcomes

**Exogenous Treatment Selection**: \( E[y(t)|z] = E[y(t)]. \)

**Monotone Treatment Selection**: \( u_2 > u_1 \Rightarrow E[y(t)|z = u_2] \geq E[y(t)|z = u_1]. \)

Let \( v \) be an observed covariate.

**Instrumental Variable**: \( E[y(t)|v] = E[y(t)]. \)

**Monotone Instrumental Variable**: \( u_2 > u_1 \Rightarrow E[y(t)|v = u_2] \geq E[y(t)|v = u_1]. \)

Some Assumptions on Response Functions

**Linear Treatment Response**: \( y_j(t) = \beta t + \epsilon_j. \)

**Monotone Treatment Response**: \( t_2 > t_1 \Rightarrow y_j(t_2) \geq y_j(t_1). \)
We begin from the fact that

\[ E[y(t) \mid v = u] = E(y \mid v = u, z = t)P(z = t \mid v = u) + E[y(t) \mid v = u, z \neq t]P(z \neq t \mid v = u). \]

Let \([K_0, K_1]\) be the range of \(y\). Each quantity on the right hand side is identified except for \(E[y(t) \mid v = u, z \neq t]\), which may take any value in \([K_0, K_1]\). Hence we have the \textbf{no-assumptions bound} (Manski, 1989)

\[
E(y \mid v = u, z = t)P(z = t \mid v = u) + K_0P(z \neq t \mid v = u) \\
\leq E[y(t) \mid v = u] \leq \\
E(y \mid v = u, z = t)P(z = t \mid v = u) + K_1P(z \neq t \mid v = u).
\]
An MIV assumption implies the inequality

$$u_1 \leq u \leq u_2 \Rightarrow E[y(t)|v = u_1] \leq E[y(t)|v = u] \leq E[y(t)|v = u_2].$$

Hence $E[y(t)|v = u]$ is no smaller than the lower bound on $E[y(t)|v = u_1]$ and no larger than the upper bound on $E[y(t)|v = u_2]$. Hence

**Proposition 1:** Let the MIV Assumption hold. Then

$$\sup_{u_1 \leq u} [E(y|v = u_1, z = t)\cdot P(z = t|v = u_1) + K_0\cdot P(z \neq t|v = u_1)]$$

$$\leq E[y(t)|v = u] \leq$$

$$\inf_{u_2 \geq u} [E(y|v = u_2, z = t)\cdot P(z = t|v = u_2) + K_1\cdot P(z \neq t|v = u_2)].$$

In the absence of other information, this bound is sharp. ■
**Corollary:** Let the MTS Assumption hold. Then

\[
\begin{align*}
\text{u} < \text{t} \quad &\Rightarrow \quad K_0 \leq \mathbb{E}[y(t) \mid z = u] \leq \mathbb{E}(y \mid z = t) \\
\text{u} = \text{t} \quad &\Rightarrow \quad \mathbb{E}[y(t) \mid z = u] = \mathbb{E}(y \mid z = t) \\
\text{u} > \text{t} \quad &\Rightarrow \quad \mathbb{E}(y \mid z = t) \leq \mathbb{E}[y(t) \mid z = u] \leq K_1.
\end{align*}
\]

and

\[
K_0 P(z < t) + \mathbb{E}(y \mid z = t) P(z \geq t) \leq \mathbb{E}[y(t)] \leq K_1 P(z > t) + \mathbb{E}(y \mid z = t) P(z \leq t).
\]

In the absence of other information, these bounds are sharp. ■
Identification Using MTR and MIV Assumptions

The MTR assumption implies (Manski, 1997, Corollary M1.2)

\[
E(y \mid v = u, t \geq z) \cdot P(t \geq z \mid v = u) + K_0 \cdot P(t < z \mid v = u)
\]

\[
\leq E[y(t) \mid v = u] \leq
\]

\[
E(y \mid v = u, t \leq z) \cdot P(t \leq z \mid v = u) + K_1 \cdot P(t > z \mid v = u).
\]

**Proposition 3:** Let the MTR and MIV Assumptions hold. Then

\[
\sup_{u_i \leq u} [E(y \mid v = u_i, t \geq z) \cdot P(t \geq z \mid v = u_i) + K_0 \cdot P(t < z \mid v = u_i)]
\]

\[
\leq E[y(t) \mid v = u] \leq
\]

\[
\inf_{u_2 \geq u} [E(y \mid v = u_2, t \leq z) \cdot P(t \leq z \mid v = u_2) + K_1 \cdot P(t > z \mid v = u_2)].
\]

In the absence of other information, this bound is sharp. ■
Corollary: Let the MTR and MTS Assumptions hold. Then

\[ u < t \quad \Rightarrow \quad E(y|z = u) \leq E[y(t)|z = u] \leq E(y|z = t) \]
\[ u = t \quad \Rightarrow \quad E[y(t)|z = u] = E(y|z = t) \]
\[ u > t \quad \Rightarrow \quad E(y|z = t) \leq E[y(t)|z = u] \leq E(y|z = u) \]

and

\[ \sum_{u < t} E(y|z = u) \cdot P(z = u) + E(y|z = t) \cdot P(z \geq t) \leq E[y(t)] \]
\[ \leq \sum_{u > t} E(y|z = u) \cdot P(z = u) + E(y|z = t) \cdot P(z \leq t). \]

In the absence of other information, these bounds are sharp. ■
Empirical Analysis of The Returns to Schooling

We use data on 1,257 randomly sampled NLSY white male respondents who, in 1994, reported that they were full-time year-round workers with positive wages. We exclude the self-employed. We observe each respondent’s 1994 hourly log-wage \((y)\) and years of schooling \((z)\). We seek to learn

\[
\Delta(s, t) = \text{E}[y(t)] - \text{E}[y(s)].
\]

Under the MTR and MTS assumptions, the sharp upper bound on \(\Delta(s, t)\) is the upper bound on \(\text{E}[y(t)]\) minus the lower bound on \(\text{E}[y(s)]\).
Table 1: NLSY Empirical Mean log(wages) and Distribution of Years of Schooling

| z  | E(y|z) | P(z)  | Sample Size |
|----|-------|-------|-------------|
| 8  | 2.249 | 0.014 | 18          |
| 9  | 2.302 | 0.018 | 22          |
| 10 | 2.195 | 0.018 | 23          |
| 11 | 2.346 | 0.025 | 32          |
| 12 | 2.496 | 0.413 | 519         |
| 13 | 2.658 | 0.074 | 93          |
| 14 | 2.639 | 0.083 | 104         |
| 15 | 2.693 | 0.035 | 44          |
| 16 | 2.870 | 0.189 | 238         |
| 17 | 2.775 | 0.038 | 48          |
| 18 | 3.006 | 0.051 | 64          |
| 19 | 3.009 | 0.020 | 25          |
| 20 | 2.936 | 0.021 | 27          |
| Total | 1 | 1257 |
Table 2: MTR-MTS Upper Bounds on Returns to Schooling

<table>
<thead>
<tr>
<th>s</th>
<th>t</th>
<th>Estimate</th>
<th>0.95 Bootstrap Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>0.390</td>
<td>0.531</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>0.334</td>
<td>0.408</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>0.445</td>
<td>0.525</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>0.313</td>
<td>0.416</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>0.253</td>
<td>0.307</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>0.159</td>
<td>0.226</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>0.202</td>
<td>0.288</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>0.304</td>
<td>0.369</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>0.165</td>
<td>0.256</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>0.386</td>
<td>0.485</td>
</tr>
<tr>
<td>18</td>
<td>19</td>
<td>0.368</td>
<td>0.539</td>
</tr>
<tr>
<td>19</td>
<td>20</td>
<td>0.296</td>
<td>0.486</td>
</tr>
<tr>
<td>----</td>
<td>----</td>
<td>----------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>0.397</td>
<td>0.450</td>
</tr>
</tbody>
</table>
It is common to have partial knowledge of policy impacts.

Ambiguity occurs when one policy is superior in some states of nature and the other is superior in other states.

Ambiguity is the result of many identification problems that are prevalent in empirical research.

unobervability of counterfactual policy outcomes: One may observe the outcomes of realized policies, but the outcomes of other policies are unobservable. Determination of an optimal policy requires comparison of all feasible policies.
How might a planner cope with ambiguity?

The Bayesian prescription asserts a subjective distribution on the states of nature and maximizes subjective expected welfare.

A subjective distribution is a form of knowledge, and the planner may have no credible basis for asserting one.

I have studied the *minimax-regret* (MR) criterion and the *adaptive minimax-regret* (AMR) criterion, a simple dynamic extension.
Fractional Treatment Allocations

Suppose that a planner can treat persons differentially.

Examples: Medical treatments, sentencing of offenders, active labor-market programs, tax schedules.

He may make a singleton allocation, assigning all observationally identical persons to the same treatment.

He could choose a fractional allocation, randomly assigning positive fractions of these persons to both treatments.

Portfolio choice has long been framed as a choice among fractional allocations, but social planning has commonly been viewed as a choice between singleton allocations.
Fractional allocations cope with ambiguity through diversification.

Diversification enables a decision maker to balance two types of potential error.

A Type A error occurs when treatment a is chosen but is actually inferior to b, and a Type B error occurs when b is chosen but is inferior to a.

The singleton allocation assigning everyone to treatment a entirely avoids type B errors but may yield Type A errors, and vice versa for singleton assignment to treatment b.

Fractional allocations make both types of errors but reduce their potential magnitudes.
Choosing Treatments for X-Pox
(Manski, C. Identification for Prediction and Decision, 2007, Chapter 11)

Suppose that a new viral disease, x-pox, is sweeping the world.

Medical researchers have proposed two mutually exclusive treatments, a and b, which reflect alternative hypotheses, $H_a$ and $H_b$, about the nature of the virus.

If $H_t$ is correct, all persons who receive treatment $t$ survive and all others die.

It is known that one of the two hypotheses is correct, but it is not known which. Thus, there are two states of natures, $\gamma = H_a$ and $\gamma = H_b$.

The objective is to maximize the survival rate of the population.
There are two singleton rules, one giving treatment a to the entire population and the other giving b.

Consider the rule in which a fraction \( \delta \in [0, 1] \) of the population receives treatment b and the remaining \( 1 - \delta \) receives treatment a. The fraction who survive is

\[
\delta \cdot 1[\gamma = H_b] + (1 - \delta) \cdot 1[\gamma = H_a].
\]

A Bayes rule generically sets \( \delta = 0 \) or 1.

The maximin and minimax-regret rules set \( \delta = \frac{1}{2} \).
Organization of the Paper

One-Period Planning with Individualistic Treatment and Linear Welfare (Review of Manski, 2007, Chapter 11)

Nonlinear Welfare

Interacting Treatments

Dynamic Planning Problems

Two-Planner Games
One-Period Problems with Individualistic Treatment and Linear Welfare

There are two treatments, labeled a and b. Let $T = \{a, b\}$.

Each member $j$ of population $J$ has a response function $y_j(\cdot)$: mapping treatments $t$ into outcomes $y_j(t)$.

$P[y(\cdot)]$ is the population distribution of treatment response. The population is large, with $P(j) = 0$ for all $j \in J$. 
The task is to allocate the population to treatments. An allocation \( \delta \in [0, 1] \) randomly assigns a fraction \( \delta \) of the population to treatment \( b \) and the remaining \( 1 - \delta \) to treatment \( a \).

The planner wants to maximize mean welfare.

Let \( u_j(t) = u_j[y(t), t] \) be the net contribution to welfare when person \( j \) receives treatment \( t \) and realizes outcome \( y_j(t) \).

Let \( \alpha = \mathbb{E}[u(a)] \) and \( \beta = \mathbb{E}[u(b)] \).

Mean welfare with allocation \( \delta \) is

\[
W(\delta) = \alpha(1 - \delta) + \beta \delta = \alpha + (\beta - \alpha)\delta.
\]
Treatment Choice Under Ambiguity

\[ \delta = 1 \text{ is optimal if } \beta \geq \alpha \text{ and } \delta = 0 \text{ if } \beta \leq \alpha. \] The problem is treatment choice when \((\alpha, \beta)\) is partially known.

Let \(S\) index the feasible states of nature. The planner knows that \((\alpha, \beta)\) lies in the set \([(\alpha_s, \beta_s), s \in S]\). Let this set be bounded. Let

\[
\alpha_L = \min_{s \in S} \alpha_s, \quad \beta_L = \min_{s \in S} \beta_s,
\]

\[
\alpha_U = \max_{s \in S} \alpha_s, \quad \beta_U = \max_{s \in S} \beta_s.
\]

The planner faces ambiguity if \(\alpha_s > \beta_s\) for some values of \(s\) and \(\alpha_s < \beta_s\) for other values.
Bayes Rules

A Bayesian planner places a subjective distribution \( \pi \) on \( S \) and solves

\[
\max_{\delta \in [0, 1]} E_{\pi}(\alpha) + [E_{\pi}(\beta) - E_{\pi}(\alpha)]\delta,
\]

where \( E_{\pi}(\alpha) = \int \alpha_s d\pi \) and \( E_{\pi}(\beta) = \int \beta_s d\pi \).

Chooses \( \delta = 0 \) if \( E_{\pi}(\beta) < E_{\pi}(\alpha) \) and \( \delta = 1 \) if \( E_{\pi}(\beta) > E_{\pi}(\alpha) \).

The Maximin Criterion

A maximin planner solves

\[
\max_{\delta \in [0, 1]} \min_{s \in S} \alpha_s + (\beta_s - \alpha_s)\delta.
\]

Let \((\alpha_L, \beta_L)\) be feasible. Then the decision is
\( \delta = 0 \) if \( \beta_L < \alpha_L \) and \( \delta = 1 \) if \( \beta_L > \alpha_L \).
The Minimax-Regret Criterion

The regret of allocation $\delta$ in state of nature $s$ is the difference between the maximum achievable welfare and the welfare achieved with allocation $\delta$.

Maximum welfare in state of nature $s$ is $\max (\alpha_s, \beta_s)$.

The minimax-regret criterion is

$$\min_{\delta \in [0, 1]} \max_s \max (\alpha_s, \beta_s) - [\alpha_s + (\beta_s - \alpha_s)\delta].$$
The Minimax-Regret Allocation
(Identification for Prediction and Decision, Complement 11A)

Let $S(a) = \{ s \in S: \alpha_s > \beta_s \}$ and $S(b) = \{ s \in S: \beta_s > \alpha_s \}$.

Let $M(a) = \max_{s \in S(a)} (\alpha_s - \beta_s)$; $M(b) = \max_{s \in S(b)} (\beta_s - \alpha_s)$.
Then

$$\delta_{MR} = \frac{M(b)}{M(a) + M(b)}.$$

If $(\alpha_L, \beta_U)$ and $(\alpha_U, \beta_L)$ are feasible, then

$$\delta_{MR} = \frac{\beta_U - \alpha_L}{(\alpha_U - \beta_L) + (\beta_U - \alpha_L)}.$$
Proof:

The maximum regret of $\delta$ is max $[R(\delta, a), R(\delta, b)]$, where

$$R(\delta, a) \equiv \max_{s \in S(a)} \alpha_s - [(1 - \delta)\alpha_s + \delta\beta_s]$$

$$= \max_{s \in S(a)} \delta(\alpha_s - \beta_s) = \delta M(a),$$

$$R(\delta, b) \equiv \max_{s \in S(b)} \beta_s - [(1 - \delta)\alpha_s + \delta\beta_s]$$

$$= \max_{s \in S(b)} (1 - \delta)(\beta_s - \alpha_s) = (1 - \delta)M(b),$$

are maximum regret on $S(a)$ and $S(b)$. 
Both treatments are undominated, so \( R(1, a) = M(a) > 0 \) and \( R(0, b) = M(b) > 0 \).

As \( \delta \) increases from 0 to 1, \( R(\cdot, a) \) increases linearly from 0 to \( M(a) \) and \( R(\cdot, b) \) decreases linearly from \( M(b) \) to 0.

Hence, the MR rule is the unique \( \delta \in (0, 1) \) such that
\[
R(\delta, a) = R(\delta, b).
\]

This yields the result.
Planning with Observable Covariates

A planner may systematically differentiate among persons with different observed covariates \( \xi \in X \).

He may segment persons by \( \xi \) and treat each group as the population. This works when the objective function is separable in covariates but not otherwise.

The Bayes objective function is always separable.

Maximin is separable if \((\alpha_{\xi_L}, \beta_{\xi_L}), \xi \in X\) is feasible.

Minimax-regret is separable if \((\alpha_{\xi_L}, \beta_{\xi_U}), \xi \in X\) and \((\alpha_{\xi_U}, \beta_{\xi_L}), \xi \in X\) are feasible.

Nonseparability occurs if the planner has information that relates the values of \((\alpha_{\xi_s}, \beta_{\xi_s})\) across \( \xi \in X \).
Planning with Multiple Treatments

The MR allocation is not always fractional when a planner allocates the population among more than two treatments.

Stoye (2007) has studied a class of such problems and has found that the MR allocations are subtle to characterize. They often are fractional, but he gives an example in which there exists a unique singleton allocation.
Nonlinear Welfare

Monotone Transformations of the Welfare Function

Let $W(\delta) = f[\alpha + (\beta - \alpha)\delta]$, where $f(\cdot)$ is strictly increasing.

The Bayes decision is generically singleton if $f(\cdot)$ is convex, but it may be fractional if $f(\cdot)$ has concave segments.

The shape of $f(\cdot)$ does not affect the maximin decision.

The minimax-regret allocation is fractional whenever $f(\cdot)$ is continuous and the optimal choice is ambiguous.

If $f(\cdot) = \log(\cdot)$ and $\{(\alpha_L, \beta_U), (\alpha_U, \beta_L)\}$ are feasible, then

$$
\delta_{MR} = \frac{\alpha_U(\beta_U - \alpha_L)}{\alpha_U(\beta_U - \alpha_L) + \beta_U(\alpha_U - \beta_L)}.
$$
Allocation of an Endowment Between a Safe and a Risky Asset

Consider an investor who must allocate an endowment between a safe and a risky asset. The safe asset is treatment $a$, with known return $\alpha$. The risky asset is $b$, with return known to lie in $[\beta_L, \beta_U]$. Assume that $\alpha \in [\beta_L, \beta_U]$.

A Bayesian investor sets $\delta = 0$ if $E_\pi(\beta) < \alpha$. He diversifies if $E_\pi(\beta) > \alpha$ and $\int f(\beta)d\pi < f(\alpha)$. He sets $\delta > 0$ if $\int f(\beta)d\pi \geq f(\alpha)$.

A maximin investor sets $\delta = 0$. A minimax-regret investor always diversifies, the specific allocation depending on the shape of $f(\cdot)$.
Fixed Costs

Let treatments a and b have known fixed costs $C(a)$ and $C(b)$. Let welfare be linear. The MR allocation is fractional if the fixed costs are small, but singleton if they are large.

If $C \equiv C(a) = C(b)$, then

$$\delta_{\text{FMR}} = 0 \quad \text{if} \quad M(b) \leq \min \{M(a), \delta_{\text{MR}}M(a) + C\}$$

$$\delta_{\text{FMR}} = \delta_{\text{MR}} \quad \text{if} \quad \delta_{\text{MR}}M(a) + C \leq \min \{M(a), M(b)\}$$

$$\delta_{\text{FMR}} = 1 \quad \text{if} \quad M(a) \leq \min \{M(b), \delta_{\text{MR}}M(a) + C\}.$$
Deontological Welfare Functions

Deontological ethics supposes that choices may have intrinsic value, apart from their consequences.

Fixed costs can be interpreted as expressing the deontological idea that any use of treatment a or b is bad.

Equal Treatment of Equals is a deontological principle

Fractional allocations adhere to the principle in the \emph{ex ante} sense that all persons have equal probabilities of receiving particular treatments.

Fractional allocations are inconsistent with equal treatment in the \emph{ex post} sense that all persons do not actually receive the same treatment.
From the ex ante perspective, all treatment allocations are deontologically equivalent.

From the ex post perspective, singleton allocations are advantageous relative to fractional ones.

Placing value C on equal ex post treatment does not alter the MR allocation if $C < \min \{M(a), M(b)\} - \delta_{MR}M(a)$. The MR allocation if singleton otherwise.
Interacting Treatments

Response function $y_j(\cdot, \cdot): T \times [0, 1] \to Y$ maps treatments $t$ and allocations $\delta$ into outcomes $y_j(t, \delta)$.

Let $u_j(t, \delta) = u_j[y(t, \delta), t, \delta]$. Let $\alpha(\delta) \equiv E[u(a, \delta)]$ and $\beta(\delta) \equiv E[u(b, \delta)]$.

Then welfare is $W(\delta) = \alpha(\delta)(1 - \delta) + \beta(\delta)\delta$.

The optimal allocation may be singleton or fractional, depending on how $\alpha(\cdot)$ and $\beta(\cdot)$ vary with $\delta$.

Ambiguity with interacting treatments is more severe than with individualistic treatment. Optimization previously required knowing only whether $\alpha$ is larger or smaller than $\beta$. It now requires knowledge of the functions $\alpha(\cdot)$ and $\beta(\cdot)$. 
Choosing Medical Treatments for an Infectious Disease

A population is susceptible to an infectious disease. Treatment a is therapy after infection and b is vaccination before infection.

Assume: The infection rate of unvaccinated persons decreases with the fraction of the population who are vaccinated. Thus, $\alpha(\delta)$ weakly increases with $\delta$.

Assume: Vaccination of a person always prevents his infection. Thus, $\beta(\delta) = \beta$ does not vary with $\delta$. 
**Linear Treatment Interactions**

Assume that $\alpha(\cdot)$ has the known linear form

$$\alpha(\delta) = \alpha(0)(1 - \delta) + \alpha(1)\delta.$$  

Then the optimization problem is

$$\max_{\delta \in [0, 1]} \alpha(0)(1 - \delta)^2 + \alpha(1)\delta(1 - \delta) + \beta \delta.$$  

The optimal allocation is

$$\delta^* = \frac{1}{2} + \frac{1}{2}[\beta - \alpha(0)]/[\alpha(1) - \alpha(0)]$$

if $\delta^* \in [0, 1]$. The optimum is to vaccinate no one if $\delta^* < 0$ and to vaccinate everyone if $\delta^* > 1$.  

Treatment under Ambiguity

Assume: $\alpha(\cdot)$ is bounded and weakly increasing, the bounds being real numbers $L$ and $U$ such that $L \leq \alpha(0) \leq \alpha(1) \leq U$. $\beta$ is known.

The minimax-regret allocation is

$$\delta_{MR} = \begin{cases} 1 & \text{if } \beta > U, \\ \frac{U - L}{(U - L) + (U - \beta)} & \text{otherwise.} \end{cases}$$

$\delta_{MR}$ is always positive, no matter how small $\beta$ is.
Dynamic Planning Problems

In each period $n = 0, \ldots, N$, a planner must choose treatments for the current cohort of a population.

Learning is possible, with observation of the outcomes experienced by earlier cohorts informing treatment choice for later cohorts.

Fractional treatment allocations generate randomized experiments yielding outcome data on both treatments.

Sampling variation is not an issue when cohorts are large. All fractional allocations yield the same information.
The Adaptive Minimax-Reget Criterion

In each period, the \textit{adaptive minimax-regret (AMR)} criterion applies the static minimax-regret criterion using the information available at the time.

The AMR criterion is an appealing myopic rule. It treats each cohort as well as possible, in the MR sense, given the available knowledge. It does not ask the members of one cohort to sacrifice for the benefit of future cohorts.

Unless fixed costs or deontological considerations make the AMR allocation singleton, it maximizes learning about treatment response.
Treating a Life-Threatening Disease

Close approximations to the AMR rule could be implemented in centralized health care systems where government or private agencies directly assign treatments.

Let \( y(t) \) be the number of years that a patient lives during the five years following receipt of treatment \( t \).

The outcome gradually becomes observable as time passes. Initially, \( y_j(t) \in [0, 1, 2, 3, 4, 5] \). A year later, one knows whether \( y_j(t) = 0 \) or \( y_j(t) \geq 1 \). And so on until year five.

Assume that \( u(t) = y(t) \). Assume no initial knowledge of \( \beta \).
<table>
<thead>
<tr>
<th>cohort or year (n or k)</th>
<th>death rate in k\textsuperscript{th} year after treatment</th>
<th>bound on $\beta$ for cohort n</th>
<th>AMR allocation for cohort n</th>
<th>maximum regret of AMR allocation for cohort n</th>
<th>mean life span achieved by cohort n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Status Quo</td>
<td>[0, 5]</td>
<td>0.30</td>
<td>1.05</td>
<td>3.74</td>
</tr>
<tr>
<td>1</td>
<td>0.20</td>
<td>[0.90, 4.50]</td>
<td>0.28</td>
<td>0.72</td>
<td>3.72</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>[1.78, 4.42]</td>
<td>0.35</td>
<td>0.60</td>
<td>3.78</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>[2.64, 4.36]</td>
<td>0.50</td>
<td>0.43</td>
<td>3.90</td>
</tr>
<tr>
<td>4</td>
<td>0.05</td>
<td>[3.48, 4.32]</td>
<td>0.98</td>
<td>0.02</td>
<td>4.28</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>[4.30, 4.30]</td>
<td>1</td>
<td>0</td>
<td>4.30</td>
</tr>
</tbody>
</table>
Randomized clinical trials (RCTs) are used to learn about medical innovations. The allocations produced by the AMR criterion differ from the practice of RCTs in many ways.

**Fraction of the Population Receiving the Innovation**

The AMR allocation can take any value in [0, 1]. The sample receiving the innovation in RCTs is typically a very small fraction of the population, with sample size determined by conventional calculations of statistical power.

**Group Subject to Randomization**

Under the AMR criterion, the persons receiving the innovation are randomly drawn from the full patient population. Clinical trials randomly draw subjects from pools of persons who volunteer to participate.
Measurement of Outcomes

Under the AMR criterion, one observes the health outcomes of interest as they unfold over time. RCTs typically have short durations of two to three years. Hence, medical researchers often measure surrogate outcomes rather than outcomes of real interest.

Blinding of Treatment Assignment

Under the AMR criterion, assigned treatments are known to patients and their physicians. Blinded treatment assignment has been the norm in clinical trials of new drugs.

Use of Empirical Evidence in Decision Making

Choosing a treatment allocation to minimize maximum regret is remote from the way that RCTs are used in decision making. The conventional approach is to perform a hypothesis test. The null hypothesis is that the innovation is no better than the status quo.
Two-Planner Games

Consider a two-planner setting where the planners may have different welfare functions and beliefs. Any departure from the status quo require agreement of the two planners.

When policy choice is framed as a binary decision, let the status quo be chosen if either planner prefers \( \delta = 0 \) to \( \delta = 1 \).

An innovation replaces the status quo if both prefer \( \delta = 1 \) to \( \delta = 0 \).

Thus, the innovation bears the *burden of proof*.

When policy choice is framed as selection of a treatment allocation, there may exist fractional allocations that both planners prefer.
Noncooperative Application of the MR Criterion

Let the two planners be $m = 1$ and $m = 2$. Let each have a welfare function that is monotone in some mean outcome. Suppose that both planners use the MR criterion.

Let $\delta_{mMR}$ be the allocation that $m$ would choose if he were able to dictate policy. Let $\delta_{1MR} \leq \delta_{2MR}$.

The set of pareto efficient allocations is $[\delta_{1MR}, \delta_{2MR}]$.

Consider a decision process where each planner announces his preferred allocation, after which the smaller of the reported values is chosen. Truthful revelation is the dominant strategy. Thus, $\delta_{1MR}$ is the chosen allocation.
Conclusion

Analysis of choice with incomplete information has usually presumed that the decision maker places a subjective distribution on the feasible states of nature and maximizes expected utility. A subjective distribution is a form of knowledge. There may be no credible basis to assert one.

Diversification is commonplace in private decisions with incomplete information. Yet discussions of social planning commonly presume that observationally identical persons should receive the same treatment. A planner need not limit consideration to singleton allocations.
I have focused attention on the minimax-regret criterion, which has received remarkably little attention from economists in the long period since it was proposed by Savage (1951).

The MR criterion diversifies treatment in a large class of planning problems, including ones with nonlinear welfare functions, interacting treatments, dynamics with learning, and non-cooperative aspects.

There are important problems in which the MR allocation is not fractional. It is easy to see why the MR allocation is singleton when fixed costs or deontological considerations loom large. It is harder to intuit the subtlety of the allocation in settings with more than two treatments.
Vaccination with Partial Knowledge of External Effectiveness

The problem of choosing an optimal vaccination policy for a population susceptible to infectious disease has drawn considerable attention.

Researchers have typically assumed the planner knows how vaccination affects illness rates.
There are two reasons why a planner may have only partial knowledge of the effect of vaccination on illness.

He may only partially know the *internal* effectiveness of vaccination in generating an immune response that prevents a vaccinated person from becoming ill or infectious.

He may only partially know the *external* effectiveness of vaccination in preventing transmission of disease to members of the population who are unvaccinated or unsuccessfully vaccinated.

A randomized clinical trial reveals the internal effectiveness of vaccination, but it does not reveal the external effect of applying different vaccination rates to the population.
Researchers have used epidemiological models to forecast outcomes with counterfactual vaccination policies.

They typically do not assess the accuracy of their assumptions about individual behavior, social interactions, and disease transmission.
I study choice of vaccination rate when a planner has partial knowledge of the external effectiveness of vaccination.

The objective is to minimize the social cost of illness and vaccination.

The planner observes the illness rate of a study population whose vaccination rate has been chosen previously.

He assumes that the illness rate of unvaccinated persons weakly decreases as the vaccination rate increases, but he does not know the magnitude of the preventive effect of vaccination.
Optimal Vaccination: An Illustration

Suppose that the planner must choose the vaccination rate for a large population of observationally identical persons.

Assume that vaccination always prevents a vaccinated person from becoming ill.

Let $p(t)$ be the *external-response function*, giving the fraction of unvaccinated persons who become ill when the vaccination rate is $t$.

The fraction of the population who become ill is $p(t)(1 - t)$. 
The planner wants to minimize a social cost function with two components, the harm caused by illness and the cost of vaccination.

Let $a = 1$ denote the mean social harm caused by illness and let $c > 0$ denote the mean social cost per vaccination, measured in commensurate units. The social cost of vaccination rate $t$ is

$$K(t) = p(t)(1 - t) + ct.$$  

The planner wants to solve the problem $\min_{t \in [0, 1]} K(t)$. 

Let \( p(t) = \rho(1 - t) \) and \( 0 < \rho \leq 1 \).

The optimal vaccination rate is

\[
    t^* = \arg\min_{t \in [0, 1]} \rho(1 - t)^2 + ct.
\]

The optimal rate is

\[
    t^* = \begin{cases} 0 & \text{if } 2\rho < c, \\ 1 - c/(2\rho) & \text{if } 2\rho \geq c. \end{cases}
\]
Partial Knowledge of External Effectiveness

The planner observes the vaccination and illness rates of a study population, whose vaccination rate has been chosen to be some value less than one.

He assumes that the study population and the treatment population have the same external-response function.

He assumes that the illness rate of unvaccinated persons weakly decreases as the vaccination rate increases.

He makes no assumption about the magnitude of the external effect of vaccination.
Let $r < 1$ denote the observed vaccination rate and $q(1 - r)$ denote the observed realized illness rate. The two maintained assumptions are

**Assumption 1 (Study Population):** The planner observes $r$ and $q(1 - r)$. He knows that $q = p(r)$.

**Assumption 2 (Vaccination Weakly Prevents Illness):** The planner knows that $p(t)$ is weakly decreasing in $t$.

These assumptions imply that

\[ t \leq r \implies p(t) \geq q, \]
\[ t \geq r \implies p(t) \leq q. \]
Dominance

A candidate vaccination rate $t$ is strictly dominated if any of these conditions hold:

(a) Let $c < q$. Then $t$ is strictly dominated if $t < r$.

(b) Let $c > q$. Then $t$ is strictly dominated if

$$t > r + q(1 - r)/c.$$ 

(c) Let $c > 1$. Then $t$ is strictly dominated if

$$(1 - q)/(c - q) < t \leq r$$
or if $t > \max (r, 1/c)$.
The Minimax Rate

\[ t^m = 0 \quad \text{if} \quad c > 1 \text{ and } 1 \leq q(1-r) + cr, \]

\[ = r \quad \text{if} \quad c > 1 \text{ and } 1 \geq q(1-r) + cr \]

\[ \quad \text{or if} \quad q < c < 1, \]

\[ = \text{all } t \in [0, 1] \quad \text{if} \quad c = q \text{ and } q = 1, \]

\[ = \text{all } t \in [r, 1] \quad \text{if} \quad c = q \text{ and } q < 1, \]

\[ = 1 \quad \text{if} \quad c < q. \]
Minimax-Regret Rate

(a) Let \( c \leq q \). Then the minimax-regret vaccination rate is

\[
\begin{align*}
t^\text{mr} & = \frac{(q + cr)}{(q + c)}. \\
\end{align*}
\]

(b) Let \( c > q \). Then the minimax-regret vaccination rate is

\[
\begin{align*}
t^\text{mr} & = \arg\min_{t \in [0, 1]} 1[t < r] \cdot \max \left\{ (1 - q)(1 - t), (1 - t) + c(t - r), (c - q)t \right\} \\
& + 1[t \geq r] \cdot \max \left\{ q(1 - t), c(t - r), (c - q)t \right\}. \\
\end{align*}
\]
Related Planning Problems

The analysis extends to settings where vaccination has imperfect but known internal effectiveness.

Population members may have observable covariates.

A planner who cannot mandate vaccination may provide incentives for private vaccination.

In dynamic planning problems, a planner vaccinates a sequence of cohorts, using observation of past outcomes to inform present decisions.
Looking beyond vaccination, the analysis demonstrates how one may address a class of choice problems where a planner observes the outcome of a status-quo policy and feels able to partially extrapolate to counterfactual policies.

Manski (2006) studied the criminal-justice problem of choosing a rate of search for evidence of crime, when a planner has partial knowledge of the deterrent effect of search on the rate of crime commission. I considered a planner who wants to minimize the social cost of crime, search, and punishment. The planner observes the crime rate under a status-quo search rate and assumes that the crime rate falls as the search rate rises.

The formal structure of this planning problem is similar to that of the vaccination problem, the substantive difference between the two notwithstanding.
PARTIAL IDENTIFICATION OF COUNTERFACTUAL CHOICE PROBABILITIES

Economists have long used random utility models (RUMs) that point identify counterfactual choice probabilities.

Strong predictive power requires strong assumptions. Researchers should be willing to give up some predictive power in return for enhanced credibility.

This leads to the study of behavioral models that partially identify counterfactual choice probabilities.

Outline

Choice as Treatment Response

Prediction with a Finite Universe of Alternatives
  Anatomy of the Problem
  Random Utility Models with Strict Preferences
  Linear Utility Models with Random Coefficients
  Models Assuming a Positive Fraction of Persons with Strict Preferences

Finite-Sample Confidence Sets
  Observed Binary Choice Sets
  Observed Multinomial Choice Sets
Choice as Treatment Response

General Setup

There is a population $J$ and a set of treatments $T$. Each $j \in J$ has a response function $y_j(\cdot) : T \rightarrow Y$ mapping $t \in T$ into $y_j(t) \in Y$.

The population is a probability space. $P(y(\cdot))$ describes treatment response across the population. $P(j) = 0, j \in J$.

Person $j$ receives treatment $z_j \in T$ and realizes outcome $y_j = y_j(z_j)$. The outcomes $[y_j(t), t \neq z_j]$ are counterfactual. Observation reveals $P(z, y)$.

The problem is to predict the outcomes that would occur if members of the population were to receive unrealized treatments.

Note: One may condition on any observable covariates.
Discrete Choice

There is a population $J$ and a universe of alternatives $A$. The treatments $T$ are the non-empty finite subsets of $A$, each being a possible choice set.

Response function $y_j(\cdot)$ is a *choice function* giving the alternative that person $j$ would choose when facing any choice set. Thus, $y_j(C)$ is the choice that $j$ would make if he were to face set $C$. The realized choice is $y_j = y_j(z_j)$, where $z_j$ is the choice set that $j$ actually faces.

The problem is to predict the counterfactual choices that would occur if members of the population were to face unrealized choice sets.

This paper considers prediction of behavior in a counterfactual scenario where all persons face the same set $C$. The objective is to predict the fraction of persons who would choose $c \in C$. Thus, the problem is to learn $P[y(C) = c]$. 
Exogenous Choice Set Assumption

Realized choice sets $z$ and choice functions $y(\cdot)$ are statistically independent.

$$P[y(\cdot)] = P[y(\cdot)|z]$$

Assumptions on the Distribution of Types

A type is defined by a choice function. Persons with the same choice function are of the same type.

In the econometric literature, assumptions restricting the set of types appear through the basic RUM assumption that all persons behave rationally and the common additional assumption that utility functions are linear in parameters.

Assumptions on the shape of the distribution of types appear when researchers suppose that the distribution of preference parameters has a particular form.
Prediction with a Finite Universe of Alternatives

Setup

The relevant universe for counterfactual prediction comprises all alternatives that appear in a realized choice set or the counterfactual choice set.

Let \((D_m, m \in M)\) be the realized choice sets. The objective is to predict choice from \(C\). Hence, the universe is \(A = C \cup (D_m, m \in M)\).

Let \(M\) be finite. Then there exist finitely many choice functions and, hence, finitely many types.

Let \(y_k(\cdot)\) be the choice function for type \(k\). Let \(K\) denote the set of all types.

Let \(\pi_k\) be the fraction of the population who are type \(k\). Then \(\pi = (\pi_k, k \in K)\) is the multinomial distribution of types.

A behavioral model is an assumption that \(\pi\) lies in some set \(\Pi\) of distributions.
The Identification Regions for \( \pi \) and \( P[y(C) = c] \)

\( P[y(C) = c] \) is related to \( \pi \) by the linear equation

\[
(4) \quad P[y(C) = c] = \sum_{k \in K} 1[y_k(C) = c] \cdot \pi_k.
\]

Choice data reveal the choice probabilities \( P[y(D_m) = d], d \in D_m, m \in M \). These are related to \( \pi \) through the linear equations and inequalities

\[
(5) \quad P[y(D_m) = d] = \sum_{k \in K} 1[y_k(D_m) = d] \cdot \pi_k, \quad d \in D_m, \quad m \in M.
\]

\[
(6) \quad \sum_{k \in K} \pi_k = 1; \quad \pi_k \geq 0, \quad k \in K.
\]

The behavioral model assumes that

\[
(7) \quad \pi \in \Pi.
\]

The identification region is the set \( H(\pi) \) that solve (5), (6), and (7). If the model is correct, \( H(\pi) \) is non-empty. If \( H(\pi) \) is empty, the model is incorrect.
The identification region for $P[y(C) = c]$ is the set

$$
\{P[y(C) = c] \} = \{ 1[y_k(C) = c] \cdot \lambda_k, \lambda \in H(\pi) \} 
$$

obtained by applying equation (4) with all feasible values of $\pi$. 
Linear Behavioral Models

A model is *linear* if $\Pi$ is a set of distributions that satisfies specified linear equalities or inequalities. For example, assume that the population does not contain some types $K_0$. This linear model sets $\pi_k = 0$, $k \in K_0$.

The identification regions resulting from linear models have simple structures. $H(\pi)$ is convex and $H\{P[y(C) = c]\}$ is an interval.

The lower and upper bounds of this interval solve linear programming problems. The lower bound solves

\[
(9) \quad \min_{\lambda} \sum_{k \in K} 1[y_k(C) = c] \cdot \lambda_k
\]

subject to

\[
P[y(D_m) = d] = \sum_{k \in K} 1[y_k(D_m) = d] \cdot \lambda_k, \quad d \in D_m, \ m \in M,
\]

\[
\sum_{k \in K} \lambda_k = 1; \quad \lambda_k \geq 0, \ k \in K; \quad \lambda \in \Pi.
\]

The upper bound solves the analogous problem.
Random Utility Models with Strict Preferences

Assume that all persons have strict preference orderings on A. There are $|A|!$ strict preference orderings, hence this many feasible types.

The simplest non-trivial case occurs when A has three elements, (c, d, e). There are six feasible types, with preference orderings:

(c > d > e), (c > e > d), (d > c > e), (d > e > c), (e > c > d), (e > d > c).

There are four non-trivial choice sets: (c, d, e), (c, d), (c, e), and (d, e).
The relationship between choice probabilities and the preference distribution is

(10a) \( P[y(c, d, e) = c] = P(c > d > e) + P(c > e > d) \),
(10b) \( P[y(c, d, e) = d] = P(d > c > e) + P(d > e > c) \),
(10c) \( P[y(c, d, e) = e] = P(e > c > d) + P(e > d > c) \).

(11a) \( P[y(c, d) = c] = P(c > d > e) + P(c > e > d) + P(e > c > d) \),
(11b) \( P[y(c, d) = d] = P(d > c > e) + P(d > e > c) + P(e > d > c) \).

(12a) \( P[y(c, e) = c] = P(c > d > e) + P(c > e > d) + P(d > c > e) \),
(12b) \( P[y(c, e) = e] = P(e > d > c) + P(e > c > d) + P(d > e > c) \).

(13a) \( P[y(d, e) = d] = P(c > d > e) + P(d > e > c) + P(d > c > e) \),
(13b) \( P[y(d, e) = e] = P(c > e > d) + P(e > d > c) + P(e > c > d) \).
All Persons Face \((c, d, e)\) and the Objective is to Predict Choice from \((c, d)\)

\[
(15) \quad P[y(c, d, e) = c] \leq P[y(c, d) = c] \leq P[y(c, d, e) = c] + P[y(c, d, e) = e].
\]

Some Persons Face \((c, d)\), Some Face \((d, e)\), and the Objective is to Predict Choice from \((c, e)\)

\[
(17) \quad \max\{0, P[y(c, d) = c] - P[y(d, e) = e]\} \leq P[y(c, e) = c]
\leq \min\{1, P[y(c, d) = c] + P[y(d, e) = d]\}.
\]

Letting \(S = P[y(c, d) = c] + P[y(d, e) = d]\), bound (17) can be rewritten as

\[
(17') \quad \max(0, S - 1) \leq P[y(c, e) = c] \leq \min(1, S).
\]

This is the triangle inequality of Marschak (1960). Marschak presumed that one observes the three binary choice probabilities \(P[y(c, d) = c]\), \(P[y(d, e) = d]\), and \(P[y(c, e) = c]\). His objective was to test their consistency with a strict-preference RUM. They are consistent if and only if \(P[y(c, e) = c]\) lies within the bound.
Now characterize alternatives as attribute bundles and assume that persons have linear utility functions with random coefficients.

Econometricians generally assume that the distribution of coefficients is continuous. When the universe of alternatives is finite, all persons with sufficiently similar preference parameters have the same choice function. Hence, the set of types is finite.

Let each \( d \in A \) have a distinct observable attribute vector \( w_d \). The utility of \( d \) to person \( j \) is \( u_j(d) = w_d \theta_j \), where \( \theta_j \) is the person’s preference parameters.

Assume that almost all members of the population have strict preference orderings on \( A \). A sufficient condition is continuity of the distribution of \( \theta \).
Feasible preference orderings

Let \( A_\ell, \ell \in L \) denote the \(|A|! \) distinct ordered representations of \( A \). Let \( w_{\ell n} \) be the attribute bundle of the \( n^{th} \) element of the ordered set \( A_\ell \). Let

\[
\Theta_\ell = [\theta: w_{\ell 1} \theta > w_{\ell 2} \theta > \ldots > w_{\ell |A|} \theta].
\]

A person has preference ordering \( \ell \) if his preference parameter lies in \( \Theta_\ell \).

Set \( \Theta_\ell \) is empty if \( A_\ell \) contains an element \( n \) such that \( w_{\ell n} \) is a convex combination of other attribute bundles that follow or precede it in preference order.

Finite-Support Representation of the Preference Distribution

Let \( A_\ell, \ell \in L^* \) denote the feasible ordered representations of \( A \). All \( \theta \in \Theta_\ell \) yield the same choice function. Hence, we can select any \( \theta_\ell \in \Theta_\ell \) and proceed as if all persons with \( \theta \in \Theta_\ell \) have parameter \( \theta_\ell \).

Thus, the usual assumption that the distribution of \( \theta \) is continuous is observationally equivalent to assuming that it has finite support \((\theta_\ell, \ell \in L^*)\).
Counterfactual Prediction

Equation (4) takes the form

\begin{equation}
\Pr[y(C) = c] = \sum_{\ell \in L^*} \mathbb{1}[w_c \theta_\ell \geq w_d \theta_\ell, \ d \in C] \cdot \pi_\ell.
\end{equation}

Equation (5) takes the form

\begin{equation}
\Pr[y(D_m) = d] = \sum_{\ell \in L^*} \mathbb{1}[w_d \theta_\ell \geq w_e \theta_\ell, \ e \in D_m] \cdot \pi_\ell, \ d \in D_m, \ m \in M.
\end{equation}
**Numerical Illustration**

Consider a setting with two attributes, the first being quality and the second being price. The first preference parameter is normalized to 1 for the entire population. The second is uniformly distributed on the interval \([-4, -0.25]\).

There are five alternatives (a, b, c, d, e), with

- \(w(a) = (0, 0)\)
- \(w(b) = (1, 1)\)
- \(w(c) = (0.25, 0.5)\)
- \(w(d) = (0.75, 1.25)\)
- \(w(e) = (1.5, 2)\).

The observed choice sets are (a, b), (a, c), (a, d), (b, c). The counterfactual choice set is (a, e).

The result is \(P[y(a, e) = a] \in [0.8, 0.9066]\).
Finite-Sample Confidence Sets

Suppose one observes the choices of a random sample of the population. I develop confidence sets for \( P[y(C) = c] \) that are valid for all sample sizes.

**Observed Binary Choice Sets**

Let all observed choice sets be binary, with \( D_m = (d_m, e_m) \). Then (5) becomes

\[
\begin{align*}
(26) \quad P[y(D_m) = d_m] &= \sum_{k \in K} 1[y_k(D_m) = d_m] \pi_k, \quad m \in M.
\end{align*}
\]

The analogous equation for \( P[y(D_m) = e_m] \) is redundant.

Let \( p_m \equiv P[y(D_m) = d_m] \) and \( p \equiv (p_m, m \in M) \). Let \( N_m \) be the number of sample members facing \( D_m \). Let \( \bar{y}_m \) be the frequency of choice of \( d_m \).

I use \( [(\bar{y}_m, N_m), m \in M] \) to construct a confidence set for \( p \). This yields a confidence set for \( P[y(C) = c] \).
Confidence set for $p$

The components of $(\bar{y}_m, \ m \in M)$ are statistically independent, conditional on sample size. Hence, we may consider each $m$ separately and use $(\bar{y}_m, N_m)$ to construct a confidence set for $p_m$. The coverage probability for $p$ of the Cartesian product of these sets is the product of their coverage probabilities.

Let $\alpha_m \in (0, 1)$ be a desired coverage probability for $p_m$. We want a procedure whose *confidence coefficient* (infimum of the coverage probabilities across all feasible values of $p_m$) is at least $\alpha_m$. Clopper and Pearson (1934) developed such a confidence interval. Also see Blyth and Still (1983), Casella (1986), and Brown *et al.* (2001).

Let $Q(\bar{y}_m, N_m, \alpha_m)$ be any confidence interval for $p_m$ that has confidence coefficient $\alpha_m$ or greater. Consider the $|M|$-dimensional rectangle

\begin{equation}
Q = \bigotimes_{m \in M} Q(\bar{y}_m, N_m, \alpha_m).
\end{equation}

Conditional on $(N_m, \ m \in M)$, the coverage probability for $p$ of set $Q$ is at least $\alpha = \prod_{m \in M} \alpha_m$.
Confidence Set for $P[y(C) = c]$ 

Let $q \in Q$ and let $H_q(\pi)$ be the set of type distributions that satisfy the conditions

\begin{equation}
q_m = \sum_{k \in K} 1[y_k(D_m) = d_m] \cdot \pi_k, \ m \in M.
\end{equation}

\begin{align*}
\sum_{k \in K} \pi_k &= 1; \quad \pi_k \geq 0, \ k \in K; \quad \pi \in \Pi.
\end{align*}

Then $H_Q(\pi) = \bigcup_{q \in Q} H_q(\pi)$ is a confidence set for $\pi$ with coverage probability at least $\alpha$.

Moreover,

\begin{equation}
H_Q\{P[y(C) = c]\} = \{ \sum_{k \in K} 1[y_k(C) = c] \cdot \lambda_k, \ \lambda \in H_Q(\pi) \}
\end{equation}

is a confidence set for $P[y(C) = c]$ with coverage probability at least $\alpha$. 
Construction of $H_Q \{P[y(C) = c]\}$ is simple when the behavioral model is linear.

Let $Q(\bar{y}_m, N_m, \alpha_m)$ be the interval $[q_0(\bar{y}_m, N_m, \alpha_m), q_1(\bar{y}_m, N_m, \alpha_m)]$. Then $H_Q \{P[y(C) = c]\}$ is an interval whose lower bound solves the linear programming problem

\begin{equation}
\min_{\lambda} \sum_{k \in K} 1[y_k(C) = c] \cdot \lambda_k \tag{30}
\end{equation}

subject to

$$q_0(\bar{y}_m, N_m, \alpha_m) \leq \sum_{k \in K} 1[y_k(D_m) = d] \cdot \lambda_k \leq q_1(\bar{y}_m, N_m, \alpha_m), \quad m \in M,$$

$$\sum_{k \in K} \lambda_k = 1; \quad \lambda_k \geq 0, \quad k \in K; \quad \lambda \in \Pi.$$ 

The upper bound solves the analogous problem, with a max replacing the min.
**Observed Multinomial Choice Sets**

Let some observed choice sets have three or more alternatives. Construction of confidence sets is similar in principle but differs in an important detail.

Let \( p_{md} \equiv P[y(D_m) = d] \), \( p_m \equiv \{P[y(D_m) = d], d \in D_m\} \), and \( p \equiv (p_m, m \in M) \). Let \( N_m \) be the number of sample members facing \( D_m \). Let \( \bar{y}_{md} \) be the empirical frequency of choice of \( d \), and let \( \bar{y}_m = (\bar{y}_{md}, d \in D_m) \).

The components of \((\bar{y}_m, m \in M)\) are statistically independent conditional on sample size, so we may consider each \( m \) separately and use \((\bar{y}_m, N_m)\) to construct a confidence set for \( p_m \). The coverage probability for \( p \) of the Cartesian product of these sets is the product of their coverage probabilities.

The difference in detail concerns construction of the confidence set for \( p_m \). When a choice set is multinomial, \( p_m \) has multiple non-redundant elements and one faces the problem of joint inference on a vector of multinomial probabilities. Research on this subject has a long history but proposed procedures have apparently been studied only from an asymptotic perspective.
In the apparent absence of multinomial confidence sets with known finite-sample coverage, I give a conservative suggestion. 

Let \( e_m \) be the redundant element of \( D_m \). The problem is to use \( (\bar{y}_m, N_m) \) to construct a set \( Q(\bar{y}_m, N_m, \alpha_m) \) that covers \( (p_{md}, d \in D_m/e_m) \) with probability \( \alpha_m \).

For each \( d \in D_m/e_m \) and \( \gamma_{md} \in (0, 1) \), one can use the Clopper-Pearson or another procedure to construct an interval \( Q_d(\bar{y}_m, N_m, \gamma_{md}) \) that covers \( p_{md} \) with probability \( \gamma_{md} \) or greater.

Let \( Q(\bar{y}_m, N_m, \alpha_m) \) be the \( \{|D_m|-1\} \)-dimensional rectangle

\[
(31) \quad Q(\bar{y}_m, N_m, \alpha_m) = \prod_{d \in D_m/e_m} Q_d(\bar{y}_m, N_m, \gamma_{md}).
\]

The Bonferroni inequality shows that this confidence set for \( (p_{md}, d \in D_m/e_m) \) has coverage probability at least equal to

\[
(32) \quad \gamma_m = 1 - \sum_{d \in D_m/e_m} (1 - \gamma_{md}).
\]

Hence, coverage of \( \alpha_m \) is achieved by choosing \( (\gamma_{md}, d \in D_m/e_m) \) so that \( \gamma_m \geq \alpha_m \).
Rounding Probabilistic Expectations in Surveys

Charles F. Manski
Department of Economics and Institute for Policy Research, Northwestern University

Francesca Molinari
Department of Economics, Cornell University
Rounding is the familiar practice of reporting one value whenever a real number lies in an interval.

Consider meteorological reports of wind direction.

Reports issued to the public commonly delineate 8 directions. Those to aircraft pilots delineate 36 directions.

The extent of rounding is common knowledge.
Recipients of numerical data are unsure of the extent of rounding in other settings.

Consider responses to the question: What time is it?

If someone says 4:01 PM, one might infer that the person is rounding to the nearest minute.

If someone says 4 PM, one may be uncertain whether the person is rounding to the nearest minute, quarter hour, or half hour.

One may also be uncertain whether a person rounds to simplify communication or to convey partial knowledge.
Uncertainty about rounding is common when researchers analyze survey responses to numerical questions.

Questionnaires generally do not request that respondents round to a specified degree.

There are no firm conventions for rounding responses.

Hence, researchers cannot be sure how much rounding there may be in survey data.
The common practice has been to take numerical responses at face value.

When researchers show concern about data accuracy, they typically assume the classical errors-in-variables model.

The structure of the errors produced by rounding differs from the errors-in-variables model.
We study the rounding of responses to questions that ask persons to state the percent-chance that a future event will occur.

Dominitz and Manski (1997) observed that respondents tend to report values at one-percent intervals at the extremes (i.e., 0, 1, 2 and 98, 99, 100) and at five-percent intervals elsewhere (i.e., 5, 10, . . . , 90, 95). Responses are more bunched at 50 percent than at adjacent round values (40, 45, 55, 60).

This finding has been corroborated repeatedly in subsequent studies.
It seems evident that respondents to subjective probability questions round their responses, but to what extent?

When someone states 30 percent, the person may be rounding to the nearest one, five, or ten percent.

Even more uncertain is how to interpret responses of 0, 50, and 100 percent. They may be sharp expressions of beliefs, or indicate gross rounding.
Although survey data do not directly reveal the extent of rounding in observed responses, analysis of patterns of responses can be informative.

We analyze responses to the expectations module in the 2006 administration of the HRS.

We find that respondents differ considerably in their response tendencies. A small fraction only report (0, 50, 100). Most make fuller use of the 0-100 scale.
We use each person's response pattern across questions to infer the extent to which he rounds his responses to particular questions.

When person j answers question k with the value $v_{jk}$, we infer from his response pattern that the quantity of interest actually lies in an interval $[v_{jkl}, v_{jku}]$ that contains $v_{jk}$.

We bring to bear our work on analysis of interval data to characterize the consequences of rounding of outcomes for best prediction (Manski and Tamer, 2002; Beresteanu and Molinari, 2008).

We use the HRS data to illustrate.
Finally, we propose enrichment of surveys by probing to learn the extent and reasons for rounding.

The only way to strengthen inference without weakening credibility is to collect richer data on expectations.

We report an exploratory study, in which we administered a sequence of questions that follows the usual percent-chance item with further questions that probe to learn the extent and reasons for rounding.
Response Patterns to Probabilistic Expectations Questions in the HRS

In 2006, the HRS administered a module of 38 probabilistic questions to 17,191 respondents.

Table 1 presents the response patterns for 15 questions.

The respondent numbers vary across questions because the HRS makes extensive use of skip sequencing.

Table 2 describes response patterns across all the questions in the module. The sample members are the 16,674 respondents whose age is 50 or above.
<table>
<thead>
<tr>
<th>Question</th>
<th>N</th>
<th>NR</th>
<th>0</th>
<th>1-4</th>
<th>50</th>
<th>96-99</th>
<th>100</th>
<th>M10</th>
<th>M5</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3: rain or snow tomorrow</td>
<td>17191</td>
<td>0.029</td>
<td>0.218</td>
<td>0.015</td>
<td>0.150</td>
<td>0.004</td>
<td>0.047</td>
<td>0.468</td>
<td>0.066</td>
<td>0.003</td>
</tr>
<tr>
<td>P4: income keep up with cost of living</td>
<td>17191</td>
<td>0.069</td>
<td>0.173</td>
<td>0.012</td>
<td>0.182</td>
<td>0.002</td>
<td>0.063</td>
<td>0.387</td>
<td>0.108</td>
<td>0.003</td>
</tr>
<tr>
<td>P5: leave inheritance &gt;= $10,000</td>
<td>17191</td>
<td>0.053</td>
<td>0.159</td>
<td>0.004</td>
<td>0.067</td>
<td>0.008</td>
<td>0.447</td>
<td>0.209</td>
<td>0.052</td>
<td>0.001</td>
</tr>
<tr>
<td>P14: lose job during next year</td>
<td>4797</td>
<td>0.020</td>
<td>0.461</td>
<td>0.026</td>
<td>0.107</td>
<td>0.001</td>
<td>0.018</td>
<td>0.274</td>
<td>0.090</td>
<td>0.003</td>
</tr>
<tr>
<td>P15: find equally good job</td>
<td>4797</td>
<td>0.017</td>
<td>0.173</td>
<td>0.014</td>
<td>0.152</td>
<td>0.004</td>
<td>0.143</td>
<td>0.383</td>
<td>0.112</td>
<td>0.003</td>
</tr>
<tr>
<td>P18: work full time after age 65</td>
<td>5148</td>
<td>0.016</td>
<td>0.276</td>
<td>0.020</td>
<td>0.126</td>
<td>0.003</td>
<td>0.095</td>
<td>0.348</td>
<td>0.114</td>
<td>0.002</td>
</tr>
<tr>
<td>P28: live to be 75 or more</td>
<td>6713</td>
<td>0.040</td>
<td>0.053</td>
<td>0.004</td>
<td>0.222</td>
<td>0.005</td>
<td>0.152</td>
<td>0.375</td>
<td>0.144</td>
<td>0.004</td>
</tr>
<tr>
<td>P103: live independently at 75</td>
<td>2558</td>
<td>0.015</td>
<td>0.012</td>
<td>0.004</td>
<td>0.214</td>
<td>0.004</td>
<td>0.136</td>
<td>0.433</td>
<td>0.180</td>
<td>0.002</td>
</tr>
<tr>
<td>P70: medical expenses use savings</td>
<td>16754</td>
<td>0.064</td>
<td>0.254</td>
<td>0.011</td>
<td>0.137</td>
<td>0.001</td>
<td>0.056</td>
<td>0.357</td>
<td>0.118</td>
<td>0.002</td>
</tr>
<tr>
<td>P30: give help of $5000 or more</td>
<td>16754</td>
<td>0.027</td>
<td>0.382</td>
<td>0.008</td>
<td>0.114</td>
<td>0.002</td>
<td>0.118</td>
<td>0.263</td>
<td>0.084</td>
<td>0.001</td>
</tr>
<tr>
<td>P31: receive help of $5000 or more</td>
<td>16754</td>
<td>0.027</td>
<td>0.646</td>
<td>0.020</td>
<td>0.044</td>
<td>0.000</td>
<td>0.016</td>
<td>0.173</td>
<td>0.072</td>
<td>0.001</td>
</tr>
<tr>
<td>P32: move to nursing home in 5 years</td>
<td>10044</td>
<td>0.075</td>
<td>0.463</td>
<td>0.021</td>
<td>0.101</td>
<td>0.000</td>
<td>0.007</td>
<td>0.231</td>
<td>0.100</td>
<td>0.002</td>
</tr>
<tr>
<td>P34: U.S. have economic depression</td>
<td>16754</td>
<td>0.078</td>
<td>0.066</td>
<td>0.006</td>
<td>0.238</td>
<td>0.002</td>
<td>0.060</td>
<td>0.404</td>
<td>0.142</td>
<td>0.004</td>
</tr>
<tr>
<td>P110: Social Sec. will be less generous</td>
<td>16754</td>
<td>0.065</td>
<td>0.048</td>
<td>0.003</td>
<td>0.231</td>
<td>0.005</td>
<td>0.120</td>
<td>0.387</td>
<td>0.139</td>
<td>0.002</td>
</tr>
<tr>
<td>P47: mutual fund increase in value</td>
<td>16754</td>
<td>0.240</td>
<td>0.042</td>
<td>0.003</td>
<td>0.231</td>
<td>0.001</td>
<td>0.036</td>
<td>0.339</td>
<td>0.106</td>
<td>0.003</td>
</tr>
<tr>
<td>P114: mutual fund increase in real terms</td>
<td>16680</td>
<td>0.281</td>
<td>0.068</td>
<td>0.003</td>
<td>0.182</td>
<td>0.000</td>
<td>0.028</td>
<td>0.334</td>
<td>0.099</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Notes: N = sample size, NR = nonresponse, M10 = multiple of 10 but not (0, 50, 100), M5 = multiple of 5 but not of 10
<table>
<thead>
<tr>
<th></th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample size</td>
<td>mean items asked per person</td>
<td>mean items responded per person</td>
</tr>
<tr>
<td>all</td>
<td>6,774</td>
<td>9,900</td>
</tr>
<tr>
<td>age 50-54</td>
<td>594</td>
<td>988</td>
</tr>
<tr>
<td>age 55-59</td>
<td>966</td>
<td>1,443</td>
</tr>
<tr>
<td>age 60-64</td>
<td>859</td>
<td>1,434</td>
</tr>
<tr>
<td>age 65-69</td>
<td>1,396</td>
<td>1,833</td>
</tr>
<tr>
<td>age 70-74</td>
<td>1,147</td>
<td>1,511</td>
</tr>
<tr>
<td>age 75-79</td>
<td>823</td>
<td>1,069</td>
</tr>
<tr>
<td>age 80-84</td>
<td>590</td>
<td>817</td>
</tr>
<tr>
<td>age 85+</td>
<td>399</td>
<td>805</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>response pattern</th>
<th>all NR</th>
<th>all 0 or 100</th>
<th>all 0, 50 or 100</th>
<th>some M10</th>
<th>some M5</th>
<th>some 1-4 or 96-99</th>
<th>some other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>0.0227</td>
<td>0.0136</td>
<td>0.0195</td>
<td>0.2523</td>
<td>0.5322</td>
<td>0.1182</td>
<td>0.0415</td>
</tr>
<tr>
<td>age 50-54</td>
<td>0.0101</td>
<td>0.0034</td>
<td>0.0101</td>
<td>0.2239</td>
<td>0.5572</td>
<td>0.1431</td>
<td>0.0522</td>
</tr>
<tr>
<td>age 55-59</td>
<td>0.0145</td>
<td>0.0155</td>
<td>0.0135</td>
<td>0.2340</td>
<td>0.5507</td>
<td>0.1304</td>
<td>0.0414</td>
</tr>
<tr>
<td>age 60-64</td>
<td>0.0116</td>
<td>0.0116</td>
<td>0.0198</td>
<td>0.2305</td>
<td>0.5588</td>
<td>0.1246</td>
<td>0.0431</td>
</tr>
<tr>
<td>age 65-69</td>
<td>0.0193</td>
<td>0.0115</td>
<td>0.0179</td>
<td>0.2679</td>
<td>0.5186</td>
<td>0.1246</td>
<td>0.0401</td>
</tr>
<tr>
<td>age 70-74</td>
<td>0.0201</td>
<td>0.0131</td>
<td>0.0192</td>
<td>0.2720</td>
<td>0.5362</td>
<td>0.1037</td>
<td>0.0357</td>
</tr>
<tr>
<td>age 75-79</td>
<td>0.0243</td>
<td>0.0134</td>
<td>0.0122</td>
<td>0.2734</td>
<td>0.5237</td>
<td>0.1106</td>
<td>0.0425</td>
</tr>
<tr>
<td>age 80-84</td>
<td>0.0339</td>
<td>0.0153</td>
<td>0.0424</td>
<td>0.2322</td>
<td>0.5407</td>
<td>0.1051</td>
<td>0.0305</td>
</tr>
<tr>
<td>age 85+</td>
<td>0.0852</td>
<td>0.0351</td>
<td>0.0351</td>
<td>0.2607</td>
<td>0.4336</td>
<td>0.0927</td>
<td>0.0576</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>response pattern</th>
<th>all NR</th>
<th>all 0 or 100</th>
<th>all 0, 50 or 100</th>
<th>some M10</th>
<th>some M5</th>
<th>some 1-4 or 96-99</th>
<th>some other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>0.0276</td>
<td>0.0234</td>
<td>0.0261</td>
<td>0.2701</td>
<td>0.5014</td>
<td>0.1152</td>
<td>0.0363</td>
</tr>
<tr>
<td>age 50-54</td>
<td>0.0132</td>
<td>0.0142</td>
<td>0.0132</td>
<td>0.2713</td>
<td>0.5010</td>
<td>0.1498</td>
<td>0.0374</td>
</tr>
<tr>
<td>age 55-59</td>
<td>0.0152</td>
<td>0.0187</td>
<td>0.0146</td>
<td>0.2460</td>
<td>0.5253</td>
<td>0.1414</td>
<td>0.0388</td>
</tr>
<tr>
<td>age 60-64</td>
<td>0.0098</td>
<td>0.0160</td>
<td>0.0105</td>
<td>0.2301</td>
<td>0.5495</td>
<td>0.1360</td>
<td>0.0481</td>
</tr>
<tr>
<td>age 65-69</td>
<td>0.0262</td>
<td>0.0169</td>
<td>0.0278</td>
<td>0.2913</td>
<td>0.4817</td>
<td>0.1244</td>
<td>0.0316</td>
</tr>
<tr>
<td>age 70-74</td>
<td>0.0258</td>
<td>0.0192</td>
<td>0.0251</td>
<td>0.2727</td>
<td>0.5169</td>
<td>0.1039</td>
<td>0.0364</td>
</tr>
<tr>
<td>age 75-79</td>
<td>0.0196</td>
<td>0.0281</td>
<td>0.0271</td>
<td>0.2947</td>
<td>0.5108</td>
<td>0.0870</td>
<td>0.0327</td>
</tr>
<tr>
<td>age 80-84</td>
<td>0.0477</td>
<td>0.0355</td>
<td>0.0453</td>
<td>0.2852</td>
<td>0.4761</td>
<td>0.0759</td>
<td>0.0343</td>
</tr>
<tr>
<td>age 85+</td>
<td>0.0957</td>
<td>0.0609</td>
<td>0.0671</td>
<td>0.2820</td>
<td>0.4025</td>
<td>0.0658</td>
<td>0.0261</td>
</tr>
</tbody>
</table>

Notes: NR = nonresponse, M10 = multiple of 10 but not (0, 50, 100), M5 = multiple of 5 but not of 10
Empirical Analysis with Potentially Rounded Responses

Formation of Interval Data from Survey Responses

The idea is to replace a reported value $v_{jk}$ with an interval $[v_{jkL}, v_{jkU}]$ that contains $v_{jk}$.

We give an algorithm that balances the tension between wanting the interval to be narrower, hence more informative, and wider, hence more credible.

The algorithm uses only the set of responses that a person gives to a specified class of questions.
If a person does not respond to a question, the interval is [0, 100].

If a person only uses the values (0, 100) when replying to all questions in the class, we treat his data as if he is rounding grossly, with 0 implying [0, 50] and 100 implying [50, 100].

If a person only uses (0, 50, 100), we treat his data as if 0 implies [0, 25], the response 50 implies [25, 75], and 100 implies [75, 100].

If all responses are multiples of 10 and at least one is not (0, 50, 100), we treat his data as if he is rounding to the nearest 10.

And so on.
The algorithm is subject to two forms of misclassification, with asymmetric consequences for inference.

First, a given response may be less rounded than the interval yielded by the algorithm. Then our use of the data is correct, but it is less sharp than if we knew the true degree of rounding of the response.

Second, the actual rounding interval may not be a subset of the algorithm's interval. Then our use of the data is not correct.
Inference on Best Predictors with Interval Outcome Data

Empirical analysis with interval data is simply a matter of considering all points in the relevant interval to be feasible values of the quantity of interest.

Implementation can be easy or difficult, depending on the analysis.

We consider inference on best predictors under square loss with interval outcome data.
Assume that persons round to simplify communication rather than to express ambiguity. Hence, the latent subjective probability $v$ is well-defined.

Assume that the constructed intervals always contain the latent expectation.

Best nonparametric predictor

The identification region for $E(v \mid x)$ is (Manski and Tamer, 2002)

$$H[E(v \mid x)] = [E(v_L \mid x), E(v_U \mid x)].$$
**Best linear predictor**

The identification region for the parameter vector $\beta$ is (Beresteanu and Molinari, 2008)

$$H[\beta] = \left\{ \mathbf{b} \in \mathbb{R}^k : \mathbf{b} = \left[ E\left(x'x\right) \right]^{-1} E\left(x'\tilde{v}\right), \tilde{v} \in [v_L, v_U] \right\} \text{ with probability 1}. $$

Closed form bounds are available for the projection of $H(\beta)$ onto each of its components and for the best predictor itself, whose identification region is

$$H[\text{BLP}(v \mid x)] = \{xb, b \in H(\beta)\}.$$
Best nonparametric predictor assuming linear regression

The identification region for $\beta$ is (Manski and Tamer, 2002)

$$[b \in \mathbb{R}^k : xb \in H \left[ E(v|x) \right] \text{ for } x-a.s.]$$
Statistical Inference

The bounds for best nonparametric prediction can be estimated by standard nonparametric procedures.

Beresteanu and Molinari (2008) show that $H(\beta)$ can be estimated as a Minkowski average of set-valued random variables.

When the BLP is assumed best nonparametric, $H(\beta)$ can be estimated using methods in Manski and Tamer (2002) and Chernozhukov, Hong, and Tamer (2007).

Our empirical illustration reports confidence sets that (asymptotically) cover each point in the identification region with probability at least 95%.
Illustrative Application

We consider the subjective expectations of HRS respondents for survival to age 75.

We analyze data from the 2006 wave of the HRS, where 6,713 respondents below age 65 were asked:

What is the percent chance that you will live to be 75 or more?

We examine the variation of response with age and gender.
Table 3 reports the sample frequencies of the width of the intervals \([v_{jkL}, v_{jkU}]\) when the intervals are constructed using two versions of the algorithm.

The left/right panel assumes each respondent uses a common rounding rule when answering the health/all questions.

Table 4 reports estimates of \(H(\beta)\).

Figure 1 reports estimates for the identification regions of the best predictors.
Table 3: Distribution of \( v_u - v_L \) in the 2006 HRS Data

<table>
<thead>
<tr>
<th></th>
<th>based on questions about personal health</th>
<th>based on all expectations questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>163</td>
<td>2.43</td>
</tr>
<tr>
<td>2.5</td>
<td>306</td>
<td>4.56</td>
</tr>
<tr>
<td>5</td>
<td>2,663</td>
<td>39.67</td>
</tr>
<tr>
<td>10</td>
<td>2,496</td>
<td>37.18</td>
</tr>
<tr>
<td>25</td>
<td>203</td>
<td>3.02</td>
</tr>
<tr>
<td>50</td>
<td>611</td>
<td>9.1</td>
</tr>
<tr>
<td>100</td>
<td>271</td>
<td>4.04</td>
</tr>
<tr>
<td>N</td>
<td>6713</td>
<td></td>
</tr>
</tbody>
</table>

Notes:

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>(B)</th>
<th>(C)</th>
<th>(D)</th>
<th>(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>-0.9751</td>
<td>1.3752</td>
<td>-0.546</td>
<td>0.9841</td>
<td>-0.2239</td>
</tr>
<tr>
<td>UB</td>
<td>-1.0867</td>
<td>1.4902</td>
<td>-0.6632</td>
<td>1.0993</td>
<td>-0.3397</td>
</tr>
<tr>
<td></td>
<td>(-6.9854</td>
<td>-20.9382</td>
<td>-15.9322</td>
<td>5.6428</td>
<td>(-12.7083</td>
</tr>
<tr>
<td>Const.</td>
<td>50.9797</td>
<td>-13.8319</td>
<td>119.3794</td>
<td>8.7446</td>
<td>95.3106</td>
</tr>
<tr>
<td></td>
<td>(43.0622</td>
<td>58.8971</td>
<td>(2.1368</td>
<td>(101.9776)</td>
<td>(19.4765</td>
</tr>
<tr>
<td>N</td>
<td>6442</td>
<td>6713</td>
<td>6713</td>
<td>6442</td>
<td>6385</td>
</tr>
</tbody>
</table>

Notes: * 95-percent confidence intervals (in parentheses) based on the normal approximation; ** 95-percent Imbens-Manski confidence sets (in parentheses) based on the bootstrap; (A) \( v_L \) and \( v_U \) constructed using only questions about personal health; (B) \( v_L \) and \( v_U \) constructed using all questions; (C) \( v_L \) and \( v_U \) constructed using all questions and nonrespondents discarded; (D) \( v_L \) and \( v_U \) constructed using all questions, nonrespondents and respondents with all 0 or 100 discarded; (E) \( v_L \) and \( v_U \) constructed using all questions, nonrespondents, respondents with all 0 or 100, and respondents with all 0, 50 or 100 discarded.
Figure 1: Set Estimates (Light Grey) and 95% Confidence Sets (Dark Grey) For Males. Predictions of Levels: Nonparametric (Cell Means) on the Left, Parametric on the Right. Data Source: HRS2006.
We have performed exploratory data analysis to obtain a sense of the source of the wide intervals obtained when rounding is taken into account.

This analysis suggests that the wide intervals result primarily from the respondents in the (all NR), (all 0, 100), and (all 0, 50, 100) response categories.
Probing Beneath the Reported Expectations

The only way to strengthen inference without weakening credibility is to collect richer data.

A fruitful way to enrich the data is to follow a standard probabilistic expectations question with further questions that probe to learn the extent and reasons for rounding.
One can follow a standard percent-chance question with these questions:

Q1. When you said [X percent] just now, did you mean this as an exact number or were you rounding or approximating?

If a person answers “rounding or approximating,” one might ask

Q2. What number or range of numbers did you have in mind when you said [X percent]?

When the response to Q1 is an exact number, one might conclude that rounding was minimal.

When the response is rounding or approximating, one can use the response to Q2 to interpret the data.
We posed the survival question to 552 respondents to the American Life Panel (ALP), an internet survey of American adults administered by RAND, and followed it by Q1 and Q2.

Table 5 describes the findings.

We used the ALP data to estimate the BLP for the probability of survival to age 75 given age and gender.

Table 6 presents the parameter estimates and Figure 2 reports the BLP estimates.
Table 5: Exact Answers versus Rounding and Approximating in ALP Data

<table>
<thead>
<tr>
<th>N</th>
<th>Exact Answers</th>
<th>Rounding or Approximating</th>
<th>Total *</th>
<th>Exact Number in Mind</th>
<th>Range and Exact Number in Mind</th>
<th>Range in Mind</th>
<th>Average Width of the Range ***</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>1-4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>50</td>
<td>104</td>
<td>60</td>
<td>44</td>
<td>14</td>
<td>(5)</td>
<td>34</td>
<td>18.44</td>
</tr>
<tr>
<td>96-99</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>4.00</td>
</tr>
<tr>
<td>100</td>
<td>32</td>
<td>27</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>37.33</td>
</tr>
<tr>
<td>M10</td>
<td>236</td>
<td>100</td>
<td>136</td>
<td>32</td>
<td>(18)</td>
<td>122</td>
<td>17.53</td>
</tr>
<tr>
<td>M5</td>
<td>145</td>
<td>56</td>
<td>89</td>
<td>19</td>
<td>(7)</td>
<td>77</td>
<td>16.55</td>
</tr>
<tr>
<td>other</td>
<td>13</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>(1)</td>
<td>8</td>
<td>22.88</td>
</tr>
<tr>
<td>Total</td>
<td>552</td>
<td>264</td>
<td>288</td>
<td>70</td>
<td>(31)</td>
<td>248</td>
<td>17.60</td>
</tr>
</tbody>
</table>

Notes: N = sample size, NR = nonresponse, M10 = multiple of 10 but not (0, 50, 100), M5 = multiple of 5 but not of 10. * One respondent who reported a probability of surviving to age 75 and older equal to 50 and reported to be approximating/rounding refused to answer the subsequent questions. ** Some respondents who declared they were approximating/rounding reported both a range and an exact number. *** One respondent reported a range with lower bound greater than upper bound, and one respondent did not report an upper bound. These two respondents were dropped from the average width calculation (these respondents were in the M10 group and in the 96-99 group).

Table 6: Point Estimates vs Set Estimates Conditioning on Age and Gender In ALP data

<table>
<thead>
<tr>
<th>Point Estimates*</th>
<th>Set Estimates**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LB</td>
</tr>
<tr>
<td>age</td>
<td>0.0794</td>
</tr>
<tr>
<td></td>
<td>(-0.0802 0.2390)</td>
</tr>
<tr>
<td></td>
<td>(-8.1421 -0.3866)</td>
</tr>
<tr>
<td>const.</td>
<td>68.1507</td>
</tr>
<tr>
<td></td>
<td>(60.5507 75.7508)</td>
</tr>
</tbody>
</table>

Notes: * 95-percent confidence intervals (in parentheses) based on the normal approximation; ** 95-percent Imbens-Manski confidence sets (in parentheses) based on the bootstrap.
Figure 3: Set Estimates (Light Grey) and 95% Confidence Sets (Dark Grey) Using ALP Data

Parametric Predictions of Levels for Males

Parametric Predictions of Levels for Females
We caution that our estimates using the ALP data accept at face value what respondents reported about their rounding practices.

Among the 264 persons who reported that their response was an exact number, almost a quarter (60) reported that their survival probability is precisely 50 percent.

Perhaps this is what they really believed, but it could be many of these respondents actually were rounding.
Conclusion

The ideas developed here should have broad application.

Consider a question asking respondents to state the number of hours they worked in the past week.

Many respondents may round their responses, with the extent of rounding differing across persons.

Examination of a person's response pattern across different numerical questions may provide a credible way to infer his rounding practice.

It may then be credible to interpret reported numerical values as intervals.

Finally, probing to ascertain the extent and reasons for rounding should be applicable broadly.
IDENTIFICATION OF TREATMENT RESPONSE WITH SOCIAL INTERACTIONS

Charles F. Manski
Northwestern University

http://faculty.wcas.northwestern.edu/~cfm754/
treatment_with_social_interactions.pdf
Overview

This paper develops a formal language for study of treatment response with social interactions, where each person’s outcome may depend on the entire vector of treatments received by the population.

I use this language to study identification when shape restrictions and distributional assumptions are placed on response functions.

The paper extends my earlier study of identification with individualistic response, emphasizing assumptions that may be credible in applications.

Hence, I primarily report findings of partial identification.
An early key result is that the traditional assumption of individualistic treatment response (ITR) is a polar case within the class of constant treatment response (CTR) assumptions.

Assumption CTR states that a person’s outcome remains constant when the population treatment vector varies in specified ways.

Leading cases are interactions within reference groups and distributional interactions.

I first study identification with assumption CTR alone.

I then strengthen it to semi-monotone treatment response (SMTR). This assumes that outcomes vary monotonically across ordered pairs of treatment vectors. Important cases are reinforcing and opposing interactions.
I next study distributional assumptions using instrumental variables. Findings obtained under assumption ITR extend when assumptions of statistical independence (SI) are posed in settings with interactions. However,

* Assumption SI has no power to identify counterfactual outcome distributions when interactions are unrestricted.

* The extended assumption SI may not be credible even if treatments are randomly assigned.

* Random assignment of realized treatments has strong identifying power when reference groups are small, limited power when distributional interactions occur in large groups, and no power in settings with strong dependence.
Finally, I consider models of endogenous interactions. Such models pose *structural equations* that take the outcome of each person to be a function of the treatment vector and of the outcomes of other members of the population. Response functions are the *reduced form*.

Identification of structural equations differs from identification of outcome distributions under potential treatment vectors. From the latter perspective, structural equations are useful if they have identifying power for the reduced form.

I compare the identifying power of complete and incomplete models.

Analysis of linear models illustrates that point identification of a complete model is not necessary for point identification of potential outcome distributions.
This paper does not

study treatments that assign persons to groups.

study dynamic treatment regimes.
Basic Concepts and Notation

*Individualistic Treatment Response*

Member \( j \) of population \( J \) has covariates \( x_j \) and response function \( y_j(\cdot) \) mapping treatments \( t \in T \) into outcomes \( y_j(t) \). Person \( j \) has realized treatment \( z_j \) and outcome \( y_j = y_j(z) \).

Observation of \([x_j, y_j, z_j], j \in J\] reveals \( P(x, y, z) \).

A common objective is to learn the outcome distribution \( P[y(t)] \) that would occur if all persons were to receive a specified treatment \( t \).
Extend the domain of $y_j(\cdot)$ from $T$ to $T^J \equiv \times_{k \in J} T$.

Now $y_j(t^j)$ is the outcome $j$ would experience if the population were to receive treatment vector $t^j = (t_k, k \in J)$.

Person $j$ realizes treatment $z_j$ and outcome $y_j = y_j(z^j)$. Observation of $[(x_j, y_j, z_j), j \in J]$ reveals $P(x, y, z)$.

The objective is inference on the outcome distribution $P[y(t^j)]$ for a specified treatment vector $t^j$.

I let $t$ denote the random variable generated by $(t_j, j \in J)$. Then $P(x, y, z, t)$ is the distribution of $[(x_j, y_j, z_j, t_j), j \in J]$.

I let $\tau$ denote a specific element of $T$. 

*Treatment with Social Interactions*
**Constant Treatment Response (Assumption CTR)**

These assumptions have limited identifying power, but often have high credibility. They provide a foundation on which further assumptions may be placed.

Consider person $j$. Let $c_j(\cdot): T^j \to C_j$ be a known function mapping treatment vectors onto a set $C_j$. A constant-response assumption asserts that

\[
(1) \quad c_j(t^j) = c_j(s^j) \Rightarrow y_j(t^j) = y_j(s^j).
\]

Thus, $j$ experiences the same outcome for all treatment vectors that form a level set of $c_j(\cdot)$.

I refer to $C_j$ as the set of *effective treatments* for person $j$. 
Consider inference on $y_j(t^j)$.

The researcher knows $y_j(t^j)$ if and only if $c_j(z^j) = c_j(t^j)$. When this event occurs, $z^j$ and $t^j$ are effectively the same treatment from the perspective of person $j$, yielding the same outcome $y_j(t^j) = y_j(z^j) = y_j$.

When $c_j(z^j) \neq c_j(t^j)$, assumption CTR and observation of $y_j$ do not reveal $y_j(t^j)$.

Hence, the identification region for $P[y(t^j)]$ is

$$
H\{P[y(t^j)]\} = P[y | c(z^j) = c(t^j)] \cdot P[c(z^j) = c(t^j)] + \delta \cdot P[c(z^j) \neq c(t^j)], \delta \in \Delta_Y.$$
It is common to assume that each member of the population has a known reference group, with interactions occurring within but not across groups. An effective treatment for person j is the sub-vector of treatments in his group.

Let \( G(j) \subset J \) denote the reference group of person j. Let \( T^{G(j)} \equiv \times_{k \in G(j)} T \). Let \( t^{G(j)} \equiv [t_k, k \in G(j)] \) be the group treatments. For \( j \in J \) and \( t^J \in T^J \), let \( C_j = T^{G(j)} \) and \( c_j(t^J) = t^{G(j)} \).

As defined here, reference groups are person-specific, treatment-invariant, and non-manipulable.
The identification region for $P[y(t')]$ is

$$H\{P[y(t')]\} = \left[ P(y|z^G = t^G) \cdot P(z^G = t^G) + \delta \cdot P(z^G \neq t^G), \delta \in \Delta_y \right].$$

One polar case is unrestricted interactions, where $G(j) = J$ for all $j \in J$. Then

$$H\{P[y(t')]\} = \left[ P(y|z^j = t^j) \cdot P(z^j = t^j) + \delta \cdot P(z^j \neq t^j), \delta \in \Delta_y \right].$$

Another is Assumption ITR, where $G(j) = j$ for all $j \in J$. Then

$$H\{P[y(t')]\} = \left[ P(y|z = t) \cdot P(z = t) + \delta \cdot P(z \neq t), \delta \in \Delta_y \right].$$
Distributional Interactions

A distributional interaction assumes that the outcome of person $j$ may vary with his own treatment and with the distribution of treatments among other members of the reference group.

The outcome does not vary with the size of the group or with permutations of the treatments received by other members of the group.

The assumption is empty when a group contains one or two persons, but is meaningful when the group is larger.
Let $C_j = T \times \Delta_T$.

For $j \in J$ with $\lvert G(j) \rvert > 1$, let $G(j)/j$ denote the reference group exclusive of person $j$ himself.

For $t^j \in T^j$, let $c_j(t^j) = [t_j, Q(t^{G(j)/j})]$, where $Q(t^{G(j)/j})$ is the within-group distribution of the treatments in $t^{G(j)/j}$.

For $j \in J$ with $\lvert G(j) \rvert = 1$, let $Q(t^{G(j)/j}) = \emptyset$.

Then

$$H\{P[y(t^j)]\} = \left\{ [P(y \mid z = t, Q(z^G) = Q(t^G)) \cdot P[z = t, Q(z^G) = Q(t^G)]ight. \\
+ \delta \cdot P(z \neq t \text{ or } Q(z^G) \neq Q(t^G), \delta \in \Delta_Y} \right\}.$$
Functional Interactions

A functional interaction assumes that $Q(t^{G(j)/j})$ affects outcomes solely through a functional, say $F(t^{G(j)/j})$.

In a *mean interaction*, treatments are real-valued and $F(t^{G(j)/j}) = E(t^{G(j)/j})$, the within-group mean of the treatments in $t^{G(j)/j}$.

In a *supremum interaction*, treatments are ordered and $F_Q(t^{G(j)/j}) = \sup(t^{G(j)/j})$.

Let $C_j = T \times \Phi$, where $\Phi$ is the range space for $F$. Let $c_j(t^j) = [t_j, F(t^{G(j)/j})]$. Then

$$H\{P[y(t^j)]\} = \{[P(y|z = t, F(z^{G}) = F(t^{G})] \cdot P[z = t, F(z^{G}) = F(t^{G})]$$

$$+ \delta \cdot P(z \neq t \text{ or } F(z^{G}) \neq F(t^{G}), \delta \in \Delta_Y\}.$$
Semi-Monotone Treatment Response (Assumption SMTR)

Let a constant-response assumption be imposed.

Considering person $j$, let $C_j$ be partially ordered. Thus, for distinct values $(c, c') \in C_j \times C_j$, either $c < c'$, $c > c'$ or $c \circ c'$.

Let $Y$ be a subset of the real line. Let $t^j$ and $s^j$ be two treatment vectors. Assumption SMTR asserts that

$$c_j(t^j) \geq c_j(s^j) \Rightarrow y_j(t^j) \geq y_j(s^j).$$
Consider the outcome of person \( j \) for treatment vector \( t^j \).
Let \( y_0 = \inf Y \) and \( y_1 = \sup Y \) be the logical lower and upper bounds on outcomes. Combining the empirical evidence with assumption SMTR yields this sharp bound on \( y_j(t^j) \):

\[
\begin{align*}
    &c_j(t^j) < c_j(z^j) \Rightarrow y_0 \leq y_j(t^j) \leq y_j \\
    &c_j(t^j) = c_j(z^j) \Rightarrow y_j(t^j) = y_j \\
    &c_j(t^j) > c_j(z^j) \Rightarrow y_j \leq y_j(t^j) \leq y_1 \\
    &c_j(t^j) \odot c_j(z^j) \Rightarrow y_0 \leq y_j(t^j) \leq y_1.
\end{align*}
\]

Let \( y_{jL}(t^j) \) and \( y_{jU}(t^j) \) denote the lower and upper bounds on \( y_j(t^j) \). The distribution of \( y_{jU}(t^j) \) stochastically dominates that of \( y(t^j) \), which dominates that of \( y_{jL}(t^j) \). Hence,

\[
H\{P[y(t^j)]\} = \{\delta \in \Delta_Y : P[y_U(t^j)] \succeq_{sd} \delta \succeq_{sd} P[y_L(t^j)]\},
\]

where \( \succeq_{sd} \) denotes weak stochastic dominance.
Reinforcing Interactions

Let T be partially ordered. Let j have reference group G(j) and let $T^{G(j)}$ inherit the partial ordering on T. A reinforcing interaction occurs when the treatments received by others in the group reinforce a person’s own treatment. Thus,

$$[t_k \geq s_k, \text{all } k \in G(j)] \Rightarrow y_j(t^j) \geq y_j(s^j).$$

A reinforcing distributional interaction occurs when

$$[t_j \geq s_j, Q(t^{G(j)/j}) \geq_{sd} Q(s^{G(j)/j})] \Rightarrow y_j(t^j) \geq y_j(s^j).$$

Example: Vaccination of person j against an infectious disease may reduce his chance of illness, and vaccination of other persons in contact with j may also reduce his chance of illness, reinforcing the effect of own vaccination.
Opposing Interactions

An opposing interaction occurs when the treatments received by others act in opposition to a person’s own treatment. Thus,

\[ t_j \geq s_j, \{t_k \leq s_k, k \in G(j)/j\} \Rightarrow y_j(t^j) \geq y_j(s^j). \]

An opposing distributional interaction occurs when

\[ t_j \geq s_j, Q(s^{G(j)/j}) \geq_{sd} Q(t^{G(j)/j}) \Rightarrow y_j(t^j) \geq y_j(s^j). \]

Example: Training person j in an occupation may increase his chance of employment in the occupation, but training other persons increases the supply of trained labor and, hence, may lower the chance that j finds employment.
Estimation of Identification Regions with Data on a Random Sample

Recall that, under assumptions CTR and SMTR,

\[
H\{P[y(t^j)]\} = \{P[y|c(z^j) = c(t^j)] \cdot P[c(z^j) = c(t^j)] \\
+ \delta \cdot P[c(z^j) \neq c(t^j)], \delta \in \Delta_Y\},
\]

\[
H\{P[y(t^j)]\} = \{\delta \in \Delta_Y: P[y_u(t^j)] \geq_{sd} \delta \geq_{sd} P[y_L(t^j)]\}.
\]

Suppose one draws a random sample of N persons, say \(J_N\), and observes \([c_j(z^j), c_j(t^j), y_j; j \in J_N]\). Then one may estimate these regions by their sample analogs

\[
H\{P_N[y(t^j)]\} = \{P_N[y|c(z^j) = c(t^j)] \cdot P_N[c(z^j) = c(t^j)] \\
+ \delta \cdot P_N[c(z^j) \neq c(t^j)], \delta \in \Delta_Y\},
\]

\[
H\{P_N[y(t^j)]\} = \{\delta \in \Delta_Y: P_N[y_u(t^j)] \geq_{sd} \delta \geq_{sd} P_N[y_L(t^j)]\}.
\]

Each \(H\{P_N[y(t^j)]\}\) converges to \(H\{P[y(t^j)]\}\) as \(N \to \infty\).
This argument requires observation of sample members’ realized effective treatments \([c_j(z^l), j \in J_N]\), not just their realized own treatments \((z_j, j \in J_N)\).

Observation of effective treatments is realistic in some applied settings. Realized treatments for the entire population may be set by known regulations, may be observable prices, or may be recorded in accessible administrative databases.

When population treatment data are not available, a survey researcher might ask sample members to report the treatments received by their reference groups.
Distributional Assumptions using Instrumental Variables

When studying individualistic response, an IV is a function $v = v(x, z)$ of covariates $x$ and realized treatments $z$. It has been common to assume statistical independence (SI). Thus, $P[y(\tau) | v] = P[y(\tau)]$.

When studying treatment with interactions, let $v = v(x^I, z^I)$. Let Assumption SI assert that $P[y(t^I) | v] = P[y(t^I)]$.

For $v \in V$, let $H\{P[y(t^I) | v = v]\}$ be the identification region for $P[y(t^I) | v = v]$ using shape restrictions but without use of Assumption SI.

Using Assumption SI,

$$H\{P[y(t^I)]\} = \cap_{v \in V} H\{P[y(t^I) | v = v]\}.$$
Using Realized Treatments as Instrumental Variables

In research with assumption ITR, it is common to let \( v = z \) and assume that \( P[y(\tau)] = P[y(\tau)|z] \).

Assumption ITR implies that \( P[y(\tau)|z = \tau] = P(y|z = \tau) \).

Observation of realized treatments and outcomes reveals \( P(y|z = \tau) \) if and only if \( P(z = \tau) > 0 \).

Hence, taking \( z \) to be the IV point-identifies \( P[y(\tau)] \) if and only if \( P(z = \tau) > 0 \).
An extension of this result holds under assumption CTR. Now the IV is the realized effective treatment $c(z^l)$.

Assume that $P[y(t^l)|c(t^l)] = P[y(t^l)|c(t^l), c(z^l)]$.

Let $C \equiv \bigcup_{j \in J} C_j$. Let $C(t^l) \equiv \{\gamma \in C: P[c(t^l) = \gamma] > 0\}$. Then

$$H\{P[y(t^l)]\} = \{ \sum_{\gamma \in C_1(t^l)} P[y|c(t^l) = \gamma, c(z^l) = \gamma] \cdot P[c(t^l) = \gamma] + \delta \cdot P[c(t^l) \in C_0(t^l)], \; \delta \in \Delta_{\gamma}\}.$$  

Here $C_1(t^l) \equiv \{\gamma \in C(t^l): P[c(z^l) = \gamma|c(t^l) = \gamma] > 0\}$,

$C_0(t^l) \equiv \{\gamma \in C(t^l): P[c(z^l) = \gamma|c(t^l) = \gamma] = 0\}$. 
P[y(t^i)] is point-identified if

\[ P[c(t^i) = \gamma] > 0 \Rightarrow P[c(z^i) = \gamma \mid c(t^i) = \gamma] > 0. \]

Then \( C_0(t^i) \) is empty, \( C_1(t^i) = C(t^i) \), and

\[ P[y(t^i)] = \sum_{\gamma \in C(t^i)} P[y \mid c(t^i) = \gamma, c(z^i) = \gamma] \cdot P[c(t^i) = \gamma]. \]
In the polar case of Assumption ITR,

\[
H\{P[y(t^I)]\} = \{ \sum_{\tau \in T_1(t^I)} P(y|t = \tau, z = \tau) \cdot P(t = \tau) + \delta \cdot P[t \in T_0(t^I)], \ \delta \in \Delta_y \}.
\]

Here \( T_1(t^I) \equiv [\tau \in T(t^I): P(z = \tau|t = \tau) > 0] \)

\( T_0(t^I) \equiv [\tau \in T(t^I): P(z = \tau|t = \tau) = 0]. \)

In the polar case of an unrestricted interaction, Assumption SI is tautological, and so has no identifying power. Hence,

\[
H\{P[y(t^I)]\} = P(y) \text{ if } z^I = t^I \text{ and } \Delta_y \text{ otherwise.}
\]
Comparison with Assumption CTR Alone

When only Assumption CTR is maintained, observation of $y_j$ and $z^j$ reveals $y_j(t^j)$ if and only if $c_j(z^j) = c_j(t^j)$. Hence, the size of $\mathbb{H}\{P[y(t^j)]\}$ grows directly with $P[c(z^j) \neq c(t^j)]$.

When assumptions CTR and SI are combined, the evidence reveals $P[y(t^j) \mid c(t^j) = \gamma]$ if and only if $P[c(z^j) = \gamma \mid c(t^j) = \gamma] > 0$. The size of $\mathbb{H}\{P[y(t^j)]\}$ grows with $P[c(t^j) \in C_0(t^j)]$.

Potential-outcome distribution $P[y(t^j)]$ may be point-identified even if $P[c(z^j) \neq c(t^j)] = 1$. 
Random Assignment of Realized Treatments

Let J be a large population (a continuum), and let the realized treatments of all members of J be assigned randomly.

Let the ex ante probability distribution on T produced by the randomization mechanism be denoted q. Thus, q(\tau) is the probability that a person is assigned to treatment \tau.

*Note:* It has been common to perform experiments in which a finite sample of a large population are randomly assigned to treatment. I do not consider such experiments here.
Random assignment has different identifying power when reference groups are small and large.

It has strong identifying power if the population partitions into many small groups, whatever the nature of the within-group interaction.

It generically has no identifying power in large-group settings with strong dependence, where the treatments assigned to a small set of persons may affect the population outcome distribution.

It has limited identifying power in large-group settings with distributional interactions.
Partition of the Population into Many Small Groups

When J is a continuum and treatments are assigned at random to the entire population, realized treatments $z$ are statistically independent of $y(\cdot)$ with ex ante probability one. Also, $P(z = \tau) = q(\tau)$ with ex ante probability one.

Recall that, under Assumption ITR,

$$P[y(\tau)] = P[y(\tau)|z = \tau] = P(y|z = \tau) \text{ for all } \tau \in T.$$  

Random assignment point-identifies $P[y(\tau)]$ when the randomization mechanism has $q(\tau) > 0$. 
This reasoning extends to situations in which the population partitions into a continuum of reference groups, each of equal finite size.

Then assumption ITR holds when the population is defined to be the collection of groups rather than persons.

It further extends to cases in which the population partitions into finitely many sub-collections of groups, each sub-collection being composed of a continuum of groups of a specific finite size.
Interactions with Strong Dependence

Let there exist a \( J_0 \subset J \) such that \( P(J_0) = 0 \), but \( P[y(t^j)] \) varies with \( (t_j, j \in J_0) \). Then treatment response has **strong dependence**. I borrow this term from analysis of stochastic processes.

If strong dependence is feasible, random assignment generically does not have identifying power.

The reason is that random assignment does not pin down the treatments realized by particular persons. It only places an ex ante probability distribution on their treatments.

Yet strong dependence permits the outcome distribution to vary with the treatments realized by particular persons.
To illustrate, consider a population with a set of leaders, role models, or central nodes.

I say that a set \( J_0 \) of persons may have *pervasive influence* on population \( J \) if, for each \( j \in J \), \( c_j(t^j) = c_j(s^j) \) only if \( t_k = s_k, k \in J_0 \).

Random assignment almost never has identifying power when the cardinality of \( J_0 \) is infinite, and it has no identifying power with positive ex ante probability when \( |J_0| \) is finite.

This is apparent from inspection of the identification region under Assumption SI.
This region is \( H\{P[y(t^j)]\} = \Delta_y \) if \( P[c(t^j) \in C_0(t^j)] = 1 \).

When \( J_0 \) may have pervasive influence, \( P[c(t^j) \in C_0(t^j)] = 1 \) if there exists any \( k \in J_0 \) such that \( z_k \neq t_k \).

Random assignment does not pin down the realized treatments \((z_k, k \in J_0)\). It only establishes an ex ante probability distribution for them.

The ex ante probability that \((z_k = t_k, k \in J_0)\) is \( \prod_{k \in J_0} q(t_k) \).

This probability is generically less than one when \( |J_0| \) is finite and zero when \( |J_0| = \infty \).

Hence, random assignment always yields positive ex ante probability that \( H\{P[y(t^j)]\} = \Delta_y \), and probability one when \( |J_0| = \infty \).
Strong dependence cannot occur in large groups with distributional interactions.

Recall that a distributional interaction is assumption CTR with $c_j(t^j) = [t_j, Q(t^{G(j)/j})]$, where $Q(t^{G(j)/j})$ is the within-group distribution of the treatments in $t^{G(j)/j}$.

If the group is large, in the sense that $P[G(j)] > 0$, variation in the treatments received by a small set of persons does not affect $Q(t^{G(j)/j})$. Hence, strong dependence cannot occur.
In a large group, \(Q(t^{G(j)}) = Q(t^{G(j)})\). Random assignment implies that \(Q(z^{G(j)}) = q\), the ex ante probability distribution on \(T\) produced by the randomization mechanism. Hence, assumption SI takes the form

\[
P[y(t^j)\mid t, Q(t^G)] = P[y(t^j)\mid (t, Q(t^G)), (z, q)].
\]

Random assignment point-identifies the outcome distribution in groups where \(Q(t^G) = q\), and has no identifying power in groups where \(Q(t^G) \neq q\).

Hence, the identification region for \(P[y(t^j)]\) is

\[
H\{P[y(t^j)]\} = \left\{ \sum_{\tau \in T} P[y\mid t = \tau, Q(t^G) = q, z = \tau] \cdot P[t = \tau, Q(t^G) = q] + \delta \cdot P[Q(t^G) \neq q], \ \delta \in \Delta_{\gamma} \right\}.
\]
Models of Endogenous Social Interactions

Models of endogenous interactions assume that there exist *structural functions* $f^j = [f_j(\cdot), j \in J]$ such that the outcomes $y^j(t^j) = [y_j(t^j), j \in J]$ solve the *structural equations*

$$y_j(t^j) = f_j[t_j, t^{jj}, y^{jj}(t^j)], \quad j \in J.$$  

Here $t^{jj} = (t_k, k \in J, k \neq j)$ and $y^{jj}(t^j) = [y_k(t^j), k \in J, k \neq j]$ are the treatment and outcome vectors for the population exclusive of person $j$.

Function $f_j(\cdot)$ permits $y_j(t^j)$ to be determined by $j$’s own treatment as well as by the treatments and outcomes of other members of the population.

The term *exogenous* interaction describes $t^{jj}$ as an argument of $f_j(\cdot)$, while *endogenous* interaction describes $y^{jj}(t^j)$. 
An outcome vector $y^J(t^J)$ that solves the structural equations is a *reduced form*. A model is *complete* if the equations have a unique solution for all feasible structural functions $f^J$. A model is *incomplete* if the equations may have multiple solutions or no solutions.

Research in econometrics has long been concerned with identification of structural functions. Observation of $P(z, y)$ reveals that $y_j = f_j(z_j, z_j^{\perp}, y_j^{\perp}), j \in J$. Partial or point identification of $f^J$ may become feasible when the empirical evidence is combined with shape restrictions or distributional assumptions on $f^J$. 
Our objective is identification of potential outcome distributions. Thus, we are concerned with the use of endogenous-interactions models to identify the reduced forms of structural equations, not with identification of structural equations per se.

Let $t^j$ be a counterfactual treatment vector. A model has identifying power for $P[y(t^j)]$ if the empirical evidence and maintained assumptions on $t^j$ imply restrictions on $P[y(t^j)]$. 
Econometrics as practiced by Arthur (Art) Goldberger demonstrates extraordinary sensitivity to issues of measurement and model specification, and unusual care and caution in interpretation of results, as well as a thorough and comprehensive mastery of econometric theory. His landmark 1964 book, *Econometric Theory*, set a new standard of rigor in econometrics, and at the same time treated the important problems posed by limited and qualitative dependent variables years before any other text. Art Goldberger’s work ranges from early contributions to macro modeling through demand analysis, multivariate modeling with latent variables, and models for sample selectivity, to his highly regarded work on important social issues of heritability of IQ, effectiveness of public versus private schools, and measurement of salary discrimination. Goldberger’s influential work, especially on modeling latent or unobservable variables, is widely known and applied in sociology and psychology as well as in economics. Art has been at Wisconsin for many years, and this association is an important reason for Wisconsin’s continuing reputation as a leading center for quantitative social sciences.

The quality and influence of Art Goldberger’s work has earned him many professional honors. He is a Fellow of the Econometric Society, the American Statistical Association, the American Academy of Arts and Sciences, and the American Association for the Advancement of Science, and has
Well, that's one position, that the entire content in a structural model is simply in the restrictions, if any, that it implies on the reduced form—that's true. That gives priority to the reduced form. There is another notion, that if one structural parameter changes, a lot of parameters will change in the reduced form. But the rest of the structure won't change. That's a distinct idea from the restrictions. And, of course, the concept of autonomy is not confined to simultaneous equation models.

I do think there's some tendency to introduce simultaneity when there's no reason for it, when a multivariate regression model would suffice. You have exogenous variables and you have endogenous variables and that's the way it ought to be, but there's some effort to invent a simultaneous structure. That is to say, you feel obligated to have the endogenous variables depending
Complete Models

Suppose the empirical evidence and assumptions imply that the structural functions lies in a set \( \Phi \).

Suppose the structural equations have a unique solution for each \( f^d \in \Phi \), denoted \( y^J(t^J, f^d) \).

Let \( P[y(t^J, f^d)] \) denote the outcome distribution when \( y^J(t^J, f^d) \) is the population outcome vector. Then the set of feasible potential outcome distributions is \( \{ P[y(t^J, f^d)], f^d \in \Phi \} \).

This shows that point identification of \( f^d \) implies point identification of \( P[y(t^J)] \). However, point identification of \( f^d \) is not necessary for point identification of \( P[y(t^J)] \).
Incomplete Models with Solutions in all States of Nature

Consider incomplete models having at least one solution to the structural equations for every feasible value of $f^d$ and multiple solutions for some values.

For each $f^d \in \Phi$, let $\Omega(t, f^d)$ denote the set of solutions. Then the set of feasible outcome distributions is $\{P[y(t^j, f^d)], y^j(t^j, f^d) \in Y(t^j, f^d), f^d \in \Phi\}$.

Now $P[y(t^j)]$ is generically partially identified. The outcome distribution is point identified only when all solutions are permutations of one another, in the sense that all feasible outcome vectors $[y^j(t^j, f^d) \in Y(t^j, f^d), f^d \in \Phi]$ yield the same outcome distribution. This is a special circumstance even when $f^d$ is point identified.
Researchers sometimes transform incomplete models into complete ones by combining the structural equations with “equilibrium selection” assumptions predicting the solution that will occur among the set of possible solutions. The credibility of such assumptions must be assessed on a case by case basis.
Incomplete Models with No Solution in Some States of Nature

Consider incomplete models having no solution for some feasible value of \( f^d \). There are two ways to interpret non-existence of a solution.

One might interpret non-existence to mean that the structural vector under consideration is not feasible. One then eliminates the vector from \( \Phi \). Then non-existence logically cannot occur.

One might interpret non-existence to mean that the endogenous-interactions model is silent on the population outcome vector. Then the model has no identifying power. This interpretation is common in game theory, where a finding that no equilibrium exists is taken to mean that the specified equilibrium concept makes no prediction about the actions chosen by players.
Models of Distributional Exogenous and Endogenous Interactions in Large Groups

Let the population partition into large groups characterized by values for a covariate $x$. Thus, all persons with the same value of $x$ belong to the same group.

Suppose that distributional exogenous and endogenous interactions may occur within each group. Thus, for $t^j \in T^j$ and $x \in X$, $y_j(t^j)$ solves

$$y_j(t^j) = f_j\{t_j, P(t^j|x_j), P[y(t^j)|x_j]\}.$$ 

Here $P(t|x_j)$ and $P[y(t^j)|x_j]$ are the distributions of treatments and outcomes in the group with covariate $x_j$. 
Without further assumptions, this model restricts the response functions \([y_j(\cdot), j \in J]\) only by requiring that interactions occur within the groups defined by values of \(x\).

The model does not thus far imply that interactions are distributional. The reason is that each person’s own treatment may affect his own outcome in a distinct manner. Hence, \(P[y(t^I) | x_j]\) may vary with the group treatment vector \(t^{G(j)}\), not just with the group treatment distribution \(P(t | x_j)\).
Empirical researchers often assume the linear model

\[ y_j(t^j) = \alpha + \beta_1 t_j + \beta_2 E(t \mid x_j) + \gamma E[y(t^j) \mid x_j] + u_j. \]

\((\alpha, \beta_1, \beta_2, \gamma)\) are parameters and \(u_j\) is a person-specific variable.

Taking expectations of both sides, conditional on \(x_j\), yields the equilibrium condition

\[ E[y(t^j) \mid x_j] = \alpha + (\beta_1 + \beta_2)E(t \mid x_j) + \gamma E[y(t^j) \mid x_j] + E(u \mid x_j). \]

Unless \(\gamma = 1\), the unique equilibrium value of \(E[y(t^j) \mid x_j]\) is

\[ E[y(t^j) \mid x_j] = \frac{\alpha}{1 - \gamma} + \frac{\beta_1 + \beta_2}{1 - \gamma} E(t \mid x_j) + \frac{E(u \mid x_j)}{1 - \gamma}. \]
This implies the reduced form

$$y_j(t^j) = \frac{\alpha}{1 - \gamma} + \beta_1 t_j + \frac{\gamma \beta_1 + \beta_2}{1 - \gamma} E(t|x_j) + \frac{\gamma}{1 - \gamma} E(u|x_j) + u_j.$$ 

Thus, response functions are linear in treatments, the slope parameters for own treatments and group-mean treatments being the same for all members of the population.

The model does not yet enable prediction of outcomes under potential treatment vectors. The reason is that it does not yet restrict the person-specific variables ($u_j, j \in J$).
Manski (1993) studied identification of the structural parameters. However, the present objective is identification of potential outcome distributions, not structural parameters.

Consider the assumption $E(u \mid z, x) = 0$. Then

\[
E(y \mid z, x) = \frac{\alpha}{1 - \gamma} + \beta_1 z + \frac{\gamma \beta_1 + \beta_2}{1 - \gamma} E(z \mid x)
\]

\[
= \varphi_0 + \varphi_1 z + \varphi_2 E(z \mid x).
\]

The empirical evidence reveals $E(y \mid z, x)$ on the support of $(z, x)$. Hence, $\varphi$ is point-identified if the support of $[1, z, E(z \mid x)]$ is not contained in a linear subspace of $\mathbb{R}^3$. 
Knowledge of $\varphi$, combined with observation of $P(z, y, x)$, implies knowledge of $(u_j, j \in J)$. This implies knowledge of all of the response functions $[y_j(\cdot), j \in J]$. This yields knowledge of $P[y(t^l)]$ for all $t^l \in T^l$.

Point-identification of the reduced-form parameters $\varphi$ does not imply point-identification of the structural parameters $(\alpha, \beta_1, \beta_2, \gamma)$. Parameter $\beta_1$ is point-identified under the maintained assumptions, but $(\alpha, \beta_2, \gamma)$ are not.

This illustrates that point identification of a complete endogenous-interactions model is not necessary for point identification of potential outcome distributions.
Conclusion

This paper has developed a formal language for study of treatment response with social interactions, and has used it to derive new findings on identification.

Findings on identification obtained under assumption ITR were extended to assumption CTR. These include identification with assumption CTR alone and when this shape restriction is strengthened to semi-monotone response.
I studied distributional assumptions using instrumental variables. Findings obtained under assumption ITR extend when Assumption SI is invoked in settings with social interactions.

The extended version of assumption SI has no power to identify counterfactual outcome distributions when social interactions are unrestricted.

Random assignment of realized treatments has strong identifying power when reference groups are small, limited power when distributional interactions occur in large groups, and generically no power in settings with strong dependence.
The final part of the paper considered models of endogenous social interactions.

I emphasized that identification of structural equations differs from identification of potential outcome distributions.

I compared the identifying power of complete and incomplete models.

I used familiar linear models to illustrate that point identification of a complete model is not necessary for point identification of outcome distributions.