CEMMAP Lectures on

# PUBLIC POLICY IN AN UNCERTAIN WORLD: Analysis and Decisions

Harvard University Press, 2013

**Charles F. Manski**

Department of Economics and Institute for Policy Research
Northwestern University

March 25-26, 2013

# Broad Themes

Society needs to face up to the uncertainties that attend policy formation.

The current practice of policy analysis hides uncertainty.

Researchers use untenable assumptions to make exact predictions of policy outcomes.

Credible policy analysis would explicitly express the limits to knowledge.

I consider how policy makers can reasonably make decisions in an uncertain world.

*Book Contents*

*Lecture Sequence and Sources*

## 1. Policy Analysis with Incredible Certitude

PPUW Chapter 1

Manski, C., "Credible Interval Estimates for Official Statistics with Survey Nonresponse," March 2013

## 2. Predicting Policy Outcomes

PPUW Chapter 2

Manski, C. and J. Pepper, "Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections," *Journal of Quantitative Criminology*, Vol. 29, 2013, No. 1, pp. 123-141.

## 3. Predicting Behavior

PPUW Chapter 3

Manski, C. "Identification of Income-Leisure Preferences and Evaluation of Income Tax Policy," January 2013.

## 4. Planning with Partial Knowledge, Diversified Treatment

PPUW Chapters 4 and 5

Manski, C. "Diversified Treatment under Ambiguity," *International Economic Review*, Vol. 50, No. 4, 2009, pp. 1013-1041.

## 5. Two Problems in Medical Decision Making

PPUW Chapter 4

Manski, C. "Vaccination with Partial Knowledge of External Effectiveness," *Proceedings of the National Academy of Sciences*, Vol. 107, No. 9, 2010, pp. 3953-3960.

Manski, C. "Diagnostic Testing and Treatment under Ambiguity: Using Decision Analysis to Inform Clinical Practice," *Proceedings of the National Academy of Sciences*, Vol. 110, No. 6, 2013, pp. 2064-2069.

## 6. Policy Analysis and Decisions

PPUW Chapter 6

# POLICY ANALYSIS
# WITH INCREDIBLE CERTITUDE

## The Logic and Credibility of Empirical Research

The logic of inference is summarized by the relationship:

**assumptions + data ⇒ conclusions.**

Stronger assumptions yield stronger conclusions.

There is a tension between the strength of assumptions and their credibility.

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

*Credibility* is a primitive concept that defies deep definition.

The Oxford English Dictionary (OED) defines *credibility* as "the quality of being credible."

The OED defines *credible* as "capable of being believed; believable."

It defines *believable* as "able to be believed; credible."

And so we come full circle.

Whatever credibility may be, it is subjective.

Analysts should agree on the logic of inference, but they often disagree about the credibility of assumptions.

Disagreement can persist when multiple contradictory assumptions are consistent with the available data.

Such assumptions are *nonrefutable*.

An analyst can pose a nonrefutable assumption and displace the burden of proof, stating

“I will maintain this assumption until it is proved wrong.”

## Incentives for Certitude

A researcher can resolve the tension between the credibility and power of assumptions by posing assumptions of varying strength and determining the conclusions that follow.

In practice, policy analysis tends to sacrifice credibility in return for strong conclusions. Why so?

Analysts respond to incentives.

* The scientific community rewards strong novel findings.

* The public wants unequivocal policy recommendations.

These incentives make it tempting for researchers to maintain assumptions far stronger than they can persuasively defend, in order to draw strong conclusions.

A story circulates about an economist's attempt to describe his uncertainty about a forecast to U. S. President Lyndon B. Johnson.

The economist presented the forecast as a likely range of values for the quantity under discussion. Johnson is said to have replied

> "Ranges are for cattle. Give me a number."

An econometrics colleague who frequently acts as a consultant stated the incentive argument this way:

> "You can't give the client a bound. The client needs a point."

I have found a common perception that policy makers are either psychologically unwilling or cognitively unable to cope with uncertainty.

Consultants argue that pragmatism dictates point prediction, even though it may not be credible.

Making point predictions also has been advocated in philosophy of science.

When there are multiple explanations for available data, philosophers recommend using a criterion such as "simplicity" to choose one of them.

**Some Manifestations of Incredible Certitude**

***conventional certitudes***

Predictions that are generally accepted as true, but that are not necessarily true.

(Examples: CBO scoring, reporting of official statistics)

***dueling certitudes***

Contradictory predictions based on alternative non-refutable assumptions.

(Example: RAND and IDA studies of drug policy)

***conflating science and advocacy***

Reversal of the direction of the logical relationship

$$\text{assumptions} + \text{data} \Rightarrow \text{conclusions},$$

seeking assumptions that imply specified conclusions.

(Example: Friedman advocacy of school vouchers)

***wishful extrapolation***

The use of untenable assumptions to extrapolate. (Example: FDA drug approval)

***illogical certitude*s**

Deductive errors, particularly non sequiturs. (Example: heritability research)

***media overreach***

Premature or exaggerated reporting of policy analysis. (Example: *NYT* on "$320,000 Kindergarten Teachers")

# CBO Scoring of Legislation

*Conventional certitude* is exemplified by Congressional Budget Office (CBO) scoring of federal legislation.

The CBO was established in the Congressional Budget Act of 1974. The Act has been interpreted as mandating the CBO to provide point predictions (*scores*) of the budgetary impact of legislation.

CBO scores are conveyed in letters that the Director writes to leaders of Congress.

They are not accompanied by measures of uncertainty.

CBO scores have achieved broad acceptance within American society.

They are used by both Democratic and Republican Members of Congress.

Media reports largely take them at face value.

*The Patient Protection and Affordable Care Act of 2010*

In March 2010 the CBO scored the combined consequences of the Patient Protection and Affordable Care Act and the Reconciliation Act of 2010.

Director Douglas Elmendorf wrote to Nancy Pelosi:

"CBO and JCT estimate that enacting both pieces of legislation . . . . would produce a net reduction of changes in federal deficits of $138 billion over the 2010–2019 period as a result of changes in direct spending and revenue."

Media reports largely accepted the CBO scores as fact.

A rare commentator who rejected the CBO score was Douglas Holtz-Eakin, a former CBO director. He wrote

"In reality, if you strip out all the gimmicks and budgetary games and rework the calculus, a wholly different picture emerges: The health care reform legislation would raise, not lower, federal deficits, by $562 billion."

The CBO and Holtz-Eakin scores differed by $700 billion. Yet they shared the common feature of certitude. Both were presented as exact, with no expression of uncertainty.

This provides an example of *dueling certitudes*.

*Interval Scoring*

The CBO has established an admirable reputation for impartiality.

Perhaps it is best to leave well enough alone and have the CBO express certitude when it scores legislation, even if the certitude is conventional rather than credible.

I worry that the existing social contract to take CBO scores at face value will eventually break down.

I think it better for the CBO to act to protect its reputation than to have some disgruntled group in Congress or the media declare that the emperor has no clothes.

A simple approach would be to provide interval forecasts of the budgetary impacts of legislation.

The CBO would produce two scores for a bill, a low score and a high score, and report both.

If the CBO must provide a point prediction for official purposes, it can continue to do so, with some convention used to locate the point within the interval forecast.

*Can Congress Cope with Uncertainty?*

I have received disparate reactions when I have suggested interval scoring to economists and policy analysts.

Academics react positively, but persons who have worked in the federal government tend to be skeptical.

Some assert that members of Congress are psychologically or cognitively unable to deal with uncertainty.

Some assert that Congressional decision making is a noncooperative game in which expression of uncertainty may yield inferior outcomes.

*British Norms*

The norms for government forecasting in the United Kingdom differ from those in the United States.

The Bank of England publishes probabilistic inflation forecasts presented visually as a *fan chart*.

The government requires an *Impact Assessment* for legislation submitted to Parliament.

The originating agency must state lower and upper bounds for the net benefits of the proposal, as well as a point estimate.

# Credible Interval Estimates for Official Statistics with Survey Nonresponse

Government agencies commonly report official statistics based on survey data as point estimates, without accompanying measures of error.

In the absence of agency guidance, users of the statistics can only conjecture the error magnitudes.

Agencies could mitigate misinterpretation of official statistics if they were to measure errors and report them.

Agencies could report sampling error using established statistical principles.

It is more challenging to report nonsampling errors because there are many sources and there has been no consensus about how to measure them.

This paper considers error due to survey nonresponse.

I summarize research deriving interval estimates that make no assumptions about the values of missing data.

In the absence of assumptions, one can obtain computable bounds on the population parameters that official statistics intend to measure.

To illustrate, I present interval estimates of median household income, the family poverty rate, and the unemployment rate in the United States.

I also explore the middle ground between interval estimation making no assumptions and traditional point estimation assuming that nonresponse is random.

*My Concern*

Reporting official statistics as point estimates without error measures encourages the public to believe that errors are small and inconsequential.

Some official statistics have become *conventional certitudes*—estimates that are generally accepted as true but that are not necessarily true.

In the absence of agency guidance, persons who understand that official statistics are subject to error must fend for themselves and conjecture the error magnitudes.

Thus, users of official statistics may misinterpret the information that the statistics provide.

*Why Study nonresponse error?*

1. Nonresponse is common.

Unit and item nonresponse often make data missing for substantial fractions of the persons sampled.

For example, the statistics on household income reported by the U. S. Census Bureau are based on the Annual Social and Economic (ASEC) Supplement to the Current Population Survey (CPS).

During 2002-2012, 7 to 9 percent of the sampled households yielded no income data due to unit nonresponse and 41 to 47 percent of the interviewed households yielded incomplete data due to item nonresponse.

2. Statistical agencies have formed point estimates by assuming, without justification, that nonresponse is random conditional on specified observed covariates.

These assumptions have been implemented as weights for unit nonresponse and imputations for item nonresponse, despite the fact that agencies do not know the consequences.

A Census Bureau document describing the American Housing Survey is revealing. The document states

"Some people refuse the interview or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones . . . . For most missing answers, an answer from a similar household is copied. The Census Bureau does not know how close the imputed values are to the actual values."

3. Methodological research has shown how to form interval estimates that make no assumptions about nonresponse.

See Manski (1989, 1994, 2003, 2007); Horowitz and Manski (1998, 2000).

We also know how to form confidence intervals that jointly measure sampling error and nonresponse error.

See Horowitz and Manski (2000), Imbens and Manski (2004), Stoye (2009).

Thus, we know how to report credible interval estimates for official statistics with survey nonresponse.

# Illustrative Reporting of U.S. Official Statistics

*BLS Reporting of Employment Statistics*

Each month, the BLS issues *The Employment Situation*.

The BLS reported on October 5, 2012:

"The unemployment rate decreased to 7.8 percent in September, and total nonfarm payroll employment rose by 114,000."

The unemployment-rate statistic is based on data on households sampled in the CPS.

The one on nonfarm employment is based on data collected from employer establishments sampled in the Current Employment Statistics survey.

The BLS monthly release reports employment statistics as point estimates, without measures of potential error.

A Technical Note issued with the release contains a section on *Reliability of the estimates* that acknowledges the possible presence of errors, beginning with the statement:

"Statistics based on the household and establishment surveys are subject to both sampling and nonsampling error."

The section describes the conventional use of standard errors and confidence intervals to measure sampling error, providing a few numerical illustrations.

The Technical Note then turns to nonsampling errors, stating that they

"can occur for many reasons, including the failure to sample a segment of the population, inability to obtain information for all respondents in the sample, inability or unwillingness of respondents to provide correct information on a timely basis, mistakes made by respondents, and errors made in the collection or processing of the data."

The Note does not indicate the magnitudes of the nonsampling errors that may be present in the employment statistics.

*Census Reporting of Income Statistics*

Each year Census reports statistics on the income distribution based on data in the ASEC supplement to CPS.

In a news release of September 12, 2012, Census declared:

"The nation's official poverty rate in 2011 was 15.0 percent, with 46.2 million people in poverty. After three consecutive years of increases, neither the poverty rate nor the number of people in poverty were statistically different from the 2010 estimates."

Thus, the Census release provided point estimates, acknowledged but did not quantify sampling error, and did not mention nonsampling error.

A Census publication gives this explanation for the decision not to report standard errors for income statistics:

"While it is possible to compute and present an estimate of the standard error based on the survey data for each estimate in a report, there are a number of reasons why this is not done. A presentation of the individual standard errors would be of limited use, since one could not possibly predict all of the combinations of results that may be of interest to data users. Additionally, data users have access to CPS microdata files, and it is impossible to compute in advance the standard error for every estimate one might obtain from those data sets."

This explains why Census cannot measure sampling error for every logically possible application of the CPS data. It does not explain why the Bureau does not report sampling error for the income statistics highlighted in news releases.

# Interval Estimation Without Assumptions on Nonresponse

Many official statistics aim to measure parameters of a probability distribution $P(y|x)$ of an outcome y conditional on covariates x.

The parameter of interest, say $\theta[P(y|x)]$, may be the conditional mean or a quantile of y.

Surveys draw stratified random samples of population units, ask sample members to report (y, x), and use the responses to estimate official statistics.

Estimation of statistics with no assumptions about nonresponse is basically a matter of contemplating all the values that the missing data might take.

*Identification Analysis*

To study inference, it is productive to first consider identification and then statistical inference.

In identification analysis, one supposes that all population units are sample members.

Hence, one knows the distributions of y and x for units who report them.

These are $P(y|z_y = 1)$, $P(x|z_x = 1)$, and $P(y, x|z_y = z_x = 1)$.

Here $z_y = 1$(or 0) if a population unit would (or not) report y, and $z_x = 1$(or 0) if the unit would (or not) report x.

One also knows the distribution $P(z_y, z_x)$ of response.

Given knowledge of these observable distributions, one may determine what this implies about $\theta[P(y|x)]$.

The generic finding is that $\theta[P(y|x)]$ lies in a set of values called its *identification region* or *identified set*.

The parameter is *point-identified* if the identification region contains just one point.

It is *partially identified* if the region contains multiple values but is a proper subset of the space of all logically possible values of the parameter.

*Statistical Inference*

Suppose that one draws a random sample of N units.

A natural way to estimate $\theta[P(y|x)]$ is to use observed empirical distributions to estimate their population counterparts.

This yields an estimate of the identification region of the parameter.

The Strong Law of Large Numbers implies that this set-valued estimate is consistent.

One can also form confidence sets to measure the uncertainty created by sampling variation.

## Outcome Nonresponse

Derivation of identification regions is particularly simple when only y has nonresponse, x always being observed.

The Law of Total Probability gives

$$P(y|x) = P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x),$$

where $z \equiv z_y$.

The empirical evidence reveals $P(z|x)$ and $P(y|x, z = 1)$.

The evidence is uninformative regarding $P(y|x, z = 0)$.

Hence, the identification region for P(y|x) is

$$H[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y],$$

where $\Gamma_Y$ is the set of all probability distributions on Y.

The identification region for $\theta[P(y|x)]$ is

$$H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}.$$

*Event Probabilities*

Consider $P(y \in B \mid x)$. 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 evidence reveals $P(z \mid x)$ and $P(y \in B \mid x, z = 1)$, but not $P(y \in B \mid x, z = 0)$.

This yields the following sharp 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 Distribution Function*

Suppose that y is real-valued. Consider the distribution function $P(y \leq t|x)$, $t \in R$.

Application of the event-probability bound to $P(y \leq t|x)$ gives

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

The feasible distribution functions are all increasing functions of t that take values between the lower and upper bounds for all values of t.

*Quantiles*

Consider the α-quantile $Q_\alpha(y|x)$.

Let $y_0$ and $y_1$ denote the smallest and largest logically possible values of y.

Manski (1994) shows that the sharp lower and upper bounds on $Q_\alpha(y|x)$ are $r(\alpha, x)$ and $s(\alpha, x)$, where

$$r(\alpha, x) \equiv [\alpha - P(z = 0 \mid x)]/P(z = 1|x) \text{ quantile of } P(y|x, z = 1)$$
$$\text{if } P(z = 0|x) < \alpha,$$
$$\equiv y_0 \text{ otherwise.}$$

$$s(\alpha, x) \equiv \alpha/P(z = 1|x) \text{ quantile of } P(y|x, z = 1)$$
$$\text{if } P(z = 0|x) \leq 1 - \alpha,$$
$$\equiv y_1 \text{ otherwise.}$$

*Means of Functions of the Outcome*

Consider the conditional mean E[g(y)|x]. Here g(·) has bounded range, with lower and upper bounds $g_0$ and $g_1$.

The Law of Iterated Expectations gives

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

The evidence reveals E[g(y)|x, z = 1] and P(z|x). E[g(y)|x, z = 0] can take any value in the interval $[g_0, g_1]$.

Hence,

$$H\{E[g(y)|x]\} = [E[g(y)|x, z = 1]P(z = 1|x) + g_0P(z = 0|x), E[g(y)|x, z = 1]P(z = 1|x) + g_1P(z = 0|x)].$$

*Estimation of the Bounds*

The sharp bounds on quantiles and means may be obtained by the same simple argument.

Wherever outcome data are missing, insert the smallest and largest values of y to obtain the bounds.

Estimation of the bounds with sample data is easy.

To estimate the lower (upper) bound, one supposes that $y_i = y_0$ ($y_1$) for every sample member i with missing data.

One then computes the usual point estimate of the parameter.

The present minimal and maximal imputations differ from *hot-deck* imputations applied to the CPS and other surveys.

Census describes the hot deck this way:

"This method assigns a missing value from a record with similar characteristics, which is the hot deck. Hot decks are defined by variables such as age, race, and sex. Other characteristics used in hot decks vary depending on the nature of the unanswered question. For instance, most labor force questions use age, race, sex, and occasionally another correlated labor force item such as full- or part-time status."

Thus, the agency staff select some sub-vector of covariates (say $x_0$) for which response is complete and determines the empirical distribution $P_N(y|x_0, z_y = 1)$ among sample members who have this value of $x_0$ and who report their outcomes.

An outcome is imputed to a sample member i with missing data by drawing a realization at random from $P_N(y|x_0 = x_{0i}, z_y = 1)$.

The CPS document offers no evidence that hot-deck imputation yields an outcome distribution for missing data that is close to the actual distribution of such outcomes.

# Interval Estimates for Median Household Income, Family Poverty, and Unemployment

Official statistics describing income and employment in the United States are based on data collected in the CPS.

The sampling unit is a household, which may include one or more families or unrelated individuals.

Data on annual income in the preceding year are collected in the ASEC Supplement, administered February–April.

Data on the current employment status of civilian adult household members are collected in the monthly administration of the CPS.

I use ASEC data collected in 2002-2012 to form interval estimates of median household income and the fraction of families with income below the official poverty threshold in the years 2001-2011.

I use monthly CPS data to form interval estimates of the unemployment rate in March 2012-2012.

I do not weight observations and do not seasonally adjust the unemployment rate.

I do not report sampling confidence intervals.

The sample sizes are so large that sampling uncertainty is a trivial consideration relative to the identification problem stemming from lack of knowledge of missing data.

TABLE 1: Interval Estimates Recognizing Item Nonresponse

| A. Annual Median Household Income | | | | |
|---|---|---|---|---|
| Year | Interviewed Households | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
| 2001 | 78265 | 0.461 | [32000, 102000] | 44200 |
| 2002 | 78310 | 0.466 | [31225, 104612] | 44231 |
| 2003 | 77149 | 0.464 | [32040, 106000] | 45100 |
| 2004 | 76447 | 0.463 | [33000, 106847] | 46324 |
| 2005 | 75939 | 0.440 | [35600, 101000] | 48078 |
| 2006 | 75477 | 0.423 | [36462, 99999] | 50002 |
| 2007 | 75872 | 0.428 | [38000, 101747] | 52000 |
| 2008 | 76185 | 0.417 | [39157, 100100] | 52004 |
| 2009 | 76260 | 0.416 | [37481, 99999] | 51157 |
| 2010 | 75188 | 0.414 | [36909, 100000] | 51016 |
| 2011 | 74838 | 0.408 | [38000, 100000] | 52000 |

| B. Annual Family Poverty Rate | | | | |
|---|---|---|---|---|
| Year | Interviewed Families | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
| 2001 | 89063 | 0.436 | [0.110, 0.315] | 0.146 |
| 2002 | 89098 | 0.440 | [0.112, 0.328] | 0.152 |
| 2003 | 87948 | 0.438 | [0.117, 0.331] | 0.158 |
| 2004 | 87149 | 0.437 | [0.118, 0.331] | 0.160 |
| 2005 | 86882 | 0.415 | [0.121, 0.319] | 0.161 |
| 2006 | 86222 | 0.400 | [0.120, 0.323] | 0.154 |
| 2007 | 86955 | 0.403 | [0.120, 0.324] | 0.154 |
| 2008 | 87562 | 0.392 | [0.126, 0.320] | 0.162 |
| 2009 | 88957 | 0.388 | [0.138, 0.338] | 0.176 |
| 2010 | 87076 | 0.390 | [0.138, 0.344] | 0.178 |
| 2011 | 86038 | 0.384 | [0.139, 0.339] | 0.176 |

| C. March Unemployment Rate | | | | |
|---|---|---|---|---|
| Year | Number of Interviewed Civilian Adults | Fraction with Missing Data | Interval Estimate | Point Estimate with Imputations |
| 2002 | 108901 | 0.0028 | [0.057, 0.062] | 0.058 |
| 2003 | 110375 | 0.0032 | [0.060, 0.065] | 0.061 |
| 2004 | 108221 | 0.0030 | [0.057, 0.062] | 0.057 |
| 2005 | 106913 | 0.0035 | [0.052, 0.057] | 0.052 |
| 2006 | 106234 | 0.0030 | [0.047, 0.051] | 0.047 |
| 2007 | 105392 | 0.0039 | [0.044, 0.050] | 0.044 |
| 2008 | 105643 | 0.0047 | [0.049, 0.057] | 0.050 |
| 2009 | 106923 | 0.0052 | [0.085, 0.093] | 0.086 |
| 2010 | 107582 | 0.0053 | [0.095, 0.103] | 0.096 |
| 2011 | 105774 | 0.0072 | [0.085, 0.096] | 0.086 |
| 2012 | 105314 | 0.0076 | [0.078, 0.090] | 0.079 |

TABLE 2: Interval Estimates Recognizing Item and Unit Nonresponse

| A. Annual Median Household Income | | | | |
|---|---|---|---|---|
| Year | Households in Sample | Fraction with Unit Non-response | Fraction with Item Non-response | Interval Estimate |
| 2001 | 84831 | 0.077 | 0.425 | [28200, 123643] |
| 2002 | 85092 | 0.080 | 0.429 | [27141, 104611] |
| 2003 | 84116 | 0.083 | 0.425 | [28000, 106000] |
| 2004 | 83932 | 0.089 | 0.422 | [28240, 106487] |
| 2005 | 83009 | 0.085 | 0.403 | [30652, 101000] |
| 2006 | 82554 | 0.086 | 0.387 | [31309, 99999] |
| 2007 | 82235 | 0.077 | 0.395 | [33200, 95008] |
| 2008 | 81904 | 0.070 | 0.388 | [35000, 100100] |
| 2009 | 81938 | 0.069 | 0.387 | [33004, 99999] |
| 2010 | 81737 | 0.080 | 0.381 | [31800, 100000] |
| 2011 | 81573 | 0.088 | 0.372 | [32132, 100000] |

| B. Annual Family Poverty Rate | | | | |
|---|---|---|---|---|
| Year | Families in Sample | Fraction with Unit Non-response | Fraction with Item Non-response | Interval Estimate |
| 2001 | 95629 | 0.069 | 0.406 | [0.102, 0.362] |
| 2002 | 95880 | 0.071 | 0.409 | [0.104, 0.376] |
| 2003 | 94915 | 0.073 | 0.406 | [0.108, 0.380] |
| 2004 | 94634 | 0.079 | 0.403 | [0.109, 0.384] |
| 2005 | 93952 | 0.075 | 0.384 | [0.112, 0.370] |
| 2006 | 93299 | 0.076 | 0.370 | [0.111, 0.374] |
| 2007 | 93318 | 0.068 | 0.376 | [0.112, 0.370] |
| 2008 | 93281 | 0.061 | 0.368 | [0.118, 0.362] |
| 2009 | 94635 | 0.060 | 0.365 | [0.130, 0.378] |
| 2010 | 93625 | 0.070 | 0.363 | [0.128, 0.390] |
| 2011 | 93228 | 0.077 | 0.354 | [0.128, 0.390] |

| C. March Unemployment Rate | | | | |
|---|---|---|---|---|
| Year | Civilian Adults in Sample | Fraction with Unit Non-response | Fraction with Item Non-response | Interval Estimate |
| 2002 | 113963 | 0.044 | 0.0027 | [0.053, 0.123] |
| 2003 | 115116 | 0.041 | 0.0031 | [0.056, 0.123] |
| 2004 | 113461 | 0.046 | 0.0028 | [0.054, 0.126] |
| 2005 | 112648 | 0.051 | 0.0033 | [0.048, 0.129] |
| 2006 | 111620 | 0.048 | 0.0029 | [0.043, 0.120] |
| 2007 | 111028 | 0.051 | 0.0037 | [0.040, 0.122] |
| 2008 | 110761 | 0.046 | 0.0045 | [0.046, 0.122] |
| 2009 | 111541 | 0.041 | 0.0050 | [0.073, 0.149] |
| 2010 | 112190 | 0.041 | 0.0051 | [0.089, 0.159] |
| 2011 | 111061 | 0.048 | 0.0069 | [0.078, 0.162] |
| 2012 | 111121 | 0.052 | 0.0072 | [0.071, 0.162] |

# Interval Estimation with Assumptions on Nonresponse

Interval estimates that place no assumptions on the values of missing data have maximal credibility.

Yet they may be excessively conservative if agency analysts have some understanding of the nature of nonresponse.

Traditional point estimates assuming random nonresponse have maximal precision.

Yet they suppose more understanding of nonresponse than agencies typically possess.

The middle ground obtains interval estimates based on assumptions that may include random nonresponse as one among various possibilities.

I restrict attention to outcome nonresponse and suppose that the outcome takes finitely many values.

I think it unlikely that any one middle-ground assumption about the nature of nonresponse will be appropriate in all settings.

Hence, I do not propose adoption of any particular assumption for reporting of official statistics.

I only pose some alternatives that statistical agencies may want to consider.

*Bounding the Distance Between the Distributions of Missing and Observed Outcomes*

Abstractly, one might assume that $P(y|x, z = 0) \in \Gamma_{cY}$, where $\Gamma_{cY}$ is a specified constrained set of outcome distributions.

The evidence combined with the assumption yields the identification region

$$H_c[P(y|x)] \equiv [P(y|x\, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{cY}].$$

More specifically, a middle-ground assumption might assert that $P(y|x, z = 0)$ lies in a neighborhood of $P(y|x, z = 1)$.

1. Assume that some fraction of nonresponse is random and that the remaining fraction arises from an unknown mechanism. This assumption asserts that

$$P(y|x, z = 0) = (1 - \delta)P(y|x, z = 1) + \delta\gamma,$$

where $\gamma$ is an unknown outcome distribution and $\delta$ is the fraction of nonresponse drawn from $\gamma$. Then

$$H_c[P(y|x)] = \{P(y|x, z = 1)[P(z = 1|x) + (1 - \delta)P(z = 0|x)] + \gamma[\delta P(z = 0|x)], \gamma \in \Gamma_Y\}.$$

The identification region has the same form as the region that would be obtained without assumptions, if the fraction of missing data were $\delta P(z = 0|x)$ rather than $P(z = 0|x)$.

2. Assume that the probability with which each missing outcome value occurs is not too different from the corresponding probability for observed outcomes. This assumption asserts that

$$\left|P(y = k|x, z = 0) - P(y = k|x, z = 1)\right| \leq \lambda_k, \quad \text{all } k \in Y.$$

Here $0 \leq \lambda_k \leq 1$ is the assumed maximum deviation between the probability that y = k conditional on the outcome being missing and observed. Hence,

$$\Gamma_{cY} = [\gamma \in \Gamma_Y: \left|\gamma(k) - P(y = k|x, z = 1)\right| \leq \lambda_k, \quad \text{all } k \in Y].$$

*Assumptions on Response Propensities*

The evidence reveals $P(z = 1|x)$ but only partially reveals $P(z = 1|x, y)$.

Assumptions that constrain the latter response propensities may have identifying power for $P(y|x)$.

The tools used to exploit such assumptions are Bayes Theorem and the fact that probabilities of mutually exclusive and exhaustive events sum to one.

For each $k \in Y$, Bayes Theorem gives

$$P(y = k|x) = P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k).$$

The evidence reveals $P(y = k|x, z = 1)$ and $P(z = 1|x)$. Hence, constraints on $P(z = 1|x, y = k)$ restrict $P(y = k|x)$.

Summing $P(y = k|x)$ across $k \in Y$ gives

$$1 = \sum_{k \in Y} P(y = k|x)$$

$$= \sum_{k \in Y} P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k).$$

Further constraints may be posed as assumptions asserting that $[P(z = 1|x, y = k), k \in Y]$ lies in a specified set of $|Y|$-dimensional vectors, say F.

Random nonresponse assumes $P(z = 1|x, y) = P(z = 1|x)$.

A middle-ground assumption might assert that each component of $[P(z = 1|x, y = k), k \in Y]$ lies in some neighborhood of $P(z = 1|x)$.

In particular, one might think it credible to assume that

$$\alpha_k P(z = 1|x) \leq P(z = 1|x, y = k) \leq \beta_k P(z = 1|x), \quad k \in K$$

for specified constants $0 \leq \alpha_k \leq 1 \leq \beta_k$, $k \in Y$.

Then $F = \times_{k \in Y} [\alpha_k P(z = 1|x), \beta_k P(z = 1|x)]$.

*Comparison of Assumptions Constraining Missing Outcomes and Response Propensities*

Both classes of assumptions have identifying power.

However, they may differ in refutability, mathematical transparency, and subjective transparency.

## Refutability

Assumptions that directly constrain the distribution of missing outcomes are nonrefutable: the available data imply no constraints on the missing data.

Assumptions that constrain response propensities may be refutable.

One may find that there exists no feasible vector [$P(z = 1|x, y = k), k \in Y$]. Then one should conclude that the vector does not lie in F.

Mathematical Transparency

The identification region for P(y|x) with partially random nonresponse has the same simple form as the one obtained without assumptions.

The region implied by assumptions on response propensities must be determined indirectly.

## Substantive Transparency

Nonresponse to surveys is a behavioral phenomenon, wherein persons choose to be interviewed and to respond to the questions asked.

An analyst may find it natural to pose conjectures about how response behavior varies with personal characteristics.

Contrariwise, an analyst may find it artificial to pose conjectures about the distribution of missing data, which is determined jointly by response behavior and by the population distribution of (y, x).

*Assumptions Restricting Temporal Variation in Unknowns*

Suppose that data have been collected in periods t = 1,. . ,T.

Index probability distributions by the period to which they pertain.

Suppose that one wants to use the data to learn the time-series $\theta[P_t(y|x)]$, t = 1, . . . , T.

Identification analysis depends on whether one maintains assumptions relating unknowns across time periods.

## 1. Identification without Temporal Assumptions

The period-specific findings extend immediately to the time series.

The joint identification region for the time-series vector of statistics is the Cartesian product of the period-specific regions.

$$H\{\theta[P_t(y|x)],\ t = 1, \ldots, T\} = \underset{t = 1, \ldots, T}{\times} H_t\{\theta[P_t(y|x)]\}.$$

This provides the basis for determination of the identification region for any function of the time-series vector of statistics.

## 2. Identification with Temporal Assumptions

Given assumptions relating unknowns across time periods, the joint identification region for the time-series vector may be a proper subset of the region obtained without these assumptions.

The joint region generally has no simple explicit form, but it can be determined numerically.

Many temporal assumptions may be reasonable to conjecture.

Among them, ones supposing that unknown quantities do not change too rapidly with time may be particularly credible.

There are multiple ways to formalize the idea.

# Conclusion

The norm in reporting official statistics is to acknowledge nonresponse error verbally but not quantitatively.

Statistical agencies neither specify nor attempt to justify the assumptions of conditionally random nonresponse used to generate imputations and weights.

Agencies could better inform the public if they were to measure potential errors and report them.

This paper has shown how to form interval estimates that face up to nonresponse.

I presented maximally credible estimates that make no assumptions about the nature of nonresponse.

I showed how assumptions imply narrower intervals.

I did not recommend particular assumptions.

I do recommend that analysts at statistical agencies should consider carefully the types of assumptions they deem credible enough to maintain, determine their identifying power, and report interval estimates accordingly.

# PREDICTING POLICY OUTCOMES

Ideally, persons who are not expert in research methodology would be able to trust the conclusions of policy analysis.

They would be able to believe predicted policy outcomes without concern about the process used to produce them.

Unfortunately, consumers of policy analysis cannot safely trust the experts.

They need to understand, at least in broad terms, how predictions depend on maintained assumptions and available data.

# Analysis of Treatment Response

Prediction of the outcomes of alternative policies is often called *analysis of treatment response*.

One contemplates a policy that has been or may be applied to a population.

The outcome that a person (or other unit) would experience under a potential treatment is his *treatment response*.

A common simplifying assumption is that response is *individualistic*: A person's outcome depends only on his own treatment.

Policy analysis uses data on a *study population* to predict policy outcomes in a *population of interest*.

A researcher observes the treatments and outcomes realized under a status quo policy in the study population.

He combines the data with assumptions to predict the outcomes that would occur if a specified policy were to be implemented in the population of interest.

Example: Sentencing and Recidivism

Possible sentences of convicted offenders are

A. non-confinement B. confinement.

The status quo policy is judicial discretion in sentencing.

A proposed policy may mandate confinement.

The problem is to predict recidivism under this policy.

*Inference Problems: Statistical Imprecision & Identification*

*Statistical inference* uses data on a sample of the study population to predict outcomes in the entire population, under the policy actually implemented there.

*Identification analysis* studies extrapolation from the study population. One may want to predict

* the outcomes of new policies.
* different outcomes than those observed.
* outcomes in other populations.

**Increasing sample size improves statistical precision, but it does not diminish identification problems.**

Identification is often the main difficulty.

A basic identification problem stems from the unobservability of *counterfactual outcomes*.

Example: Judicial Sentencing of Convicted Offenders

One can observe the recidivism of an offender under the sentence he received.

One cannot observe the recidivism the offender would have experienced under other sentences.

**The unobservability of counterfactual outcomes is a matter of logic. It cannot be addressed solely by collecting further data.**

**Prediction of counterfactual outcomes requires assumptions.**

**Weak assumptions yield interval predictions of outcomes.**

**Stronger assumptions tighten the intervals.**

**Sufficiently strong assumptions yield point predictions.**

# Predicting Outcomes of Policies Mandating Treatment

Suppose that some members of a study population receive treatment A and others receive B.

One observes the realized treatments and outcomes.

Consider prediction of outcomes in the study population under a policy mandating treatment B.

Three common assumptions are

1. *Treatment response is individualistic*.
2. 1 + *Identical Treatment Units*
3. 1 + *Identical Treatment Groups*.

*Individualistic Treatment Response*

Persons receiving B in the status quo would realize the same outcomes under the new policy.

Persons receiving A in the status quo have counterfactual outcomes under the new policy.

Mean outcome under policy mandating B

=

Observed mean outcome of persons receiving B
x
fraction of persons receiving B

+

**Unobserved mean outcome of persons receiving A**
x
fraction of persons receiving A

Hence, we obtain an interval prediction.

*Identical Treatment Units*

Assume that different persons (units) respond to treatment identically. That is, if different persons were to receive the same treatment, they would experience the same outcome.

Suppose one wants to predict the outcome of a policy in which person j would receive treatment B.

Suppose one can find an identical person in the study population, say k, that actually received treatment B.

One concludes that if j were to receive B, he would experience the outcome realized by k.

*Controlled experiments* use this reasoning.

The researcher prepares specimens j and k, intending them to be identical in every relevant respect. He applies treatment B to k and observes the outcome.

If the specimens are identical, the observed outcome of k is the outcome that j would experience if it were to receive B.

When studying persons, analysts cannot prepare identical specimens. Persons inevitably are *heterogeneous*.

Nevertheless, analysts sometimes mimic the reasoning of controlled experiments. They match persons who appear similar and assume that they respond identically to treatment.

A common practice is to compare outcomes at two points in time, before and after an event occurs.

The person's environment before the event is treatment A and the environment following the event is B.

A *before-and-after study* assumes that, except for occurrence of the event, the person is identical at the earlier and later points in time.

*Identical Treatment Groups*

Assume that treatment groups A and B have identical response distributions.

A *treatment group* is a collection of units in a study population who receive the same treatment under the status quo policy. Those receiving treatment A are one treatment group and those receiving B are another group.

The members of a treatment group may be heterogeneous.

The *response distribution* of a group describes this heterogeneity, giving the frequencies of different response patterns within the group.

Suppose that one wants to predict the outcomes of a policy mandating treatment B.

Assume that groups A and B are identical.

If group A were to receive treatment B, their outcome distribution would be the same as the one actually realized by group B.

Hence, the outcome distribution of group A under a policy mandating treatment B would be the same as the outcome distribution realized by group B under the status quo policy.

*Experiments with Random Assignment of Treatments*

The assumption of identical groups is often suspect in observational studies, where an analyst observes the outcomes of treatments selected in a decentralized manner.

The assumption has high credibility when the status quo is a classical randomized experiment.

In a randomized experiment, two random samples of persons are drawn from a study population. The members of one sample are assigned to treatment A and the members of the other are assigned to B.

The experiment is "classical" if treatment response is individualistic and all subjects comply with their assigned treatments.

Random sampling implies that both treatment groups are likely to have distributions of treatment response that are similar to the population response distribution.

The degree of similarity increases as the samples grow.

Hence, it is credible to assume that the treatment groups are identical, or at least very similar.

*Randomized Experiments in Practice*

Researchers often say that randomized experiments are the "gold standard" for evidence on treatment response.

They have in mind an ideal case.

Experiments may deviate substantially from this ideal.

There are multiple reasons, each of which generates an identification problem.

# Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections

Researchers have long used data on homicide rates to examine the deterrent effect of capital punishment.

A large body of work addresses the question, yet the literature has failed to achieve consensus.

Ehrlich, *AER* (1975):

"In fact, the empirical analysis suggests that on the average the tradeoff between the execution of an offender and the lives of potential victims it might have saved was of the order of 1 for 8 for the period 1933–1967 in the United States."

National Research Council (1978):

"The current evidence on the deterrent effect of capital punishment is inadequate for drawing any substantive conclusion."

National Research Council (2012):

"The committee concludes that research to date on the effect of capital punishment on homicide is not informative about whether capital punishment decreases, increases, or has no effect on homicide rates.

A fundamental difficulty is that the outcomes of counterfactual policies are unobservable.

Data alone cannot reveal what the homicide rate in a state without (with) a death penalty would have been had the state (not) adopted a death penalty statute.

Data must be combined with assumptions about treatment selection and/or treatment response.

It is tempting to impose assumptions strong enough to yield a definitive finding.

However, the assumptions may be inaccurate, yielding flawed and conflicting conclusions.

We study inference under weaker assumptions that partially identify deterrent effects.

We consider the problem of drawing inferences on the deterrent effects of a death penalty statute using data from repeated cross sections of states.

We focus on the years following the 1972 Supreme Court case *Furman vs. Georgia*, which resulted in a moratorium on the application of the death penalty, and the 1976 case *Gregg vs. Georgia*, which ruled that the death penalty could be applied subject to certain criteria.

We examine the effect of death penalty statutes on murder rates in 1975 and 1977. In 1975 the death penalty was illegal throughout the country. In 1977 thirty-two states had death penalty statutes.

Table 1: Homicide Rates per 100,000 Residents by Year and Treatment Status in 1977

| | Group | | Total |
|---|---|---|---|
| Year | Untreated | Treated | |
| 1975 | 8 | 10.3 | 9.6 |
| 1977 | 6.9 | 9.7 | 8.8 |
| Total | 7.5 | 10 | 9.2 |

The treated states legalized the death penalty after the *Gregg* decision and the untreated ones did not.

Assuming random assignment in 1977 gives 2.8 (9.7 – 6.9).

A before-and-after comparison of the treated gives −0.6 (9.7 – 10.3).

The *difference-in-difference* (DID) estimate is 0.5 [(9.7 – 10.3) – (6.9 – 8.0)].

# Average Treatment Effects and the Selection Problem

We consider inference on

$$ATE_{dX} = E[Y_d(1)|X] - E[Y_d(0)|X].$$

t = 1 is a sanctions regime with a death penalty statute.

t = 0 is one without such a statute.

$Y_d(t)$ is the homicide rate at date d with treatment t.

X are observed covariates.

d indicates 1975 or 1977 (= 0 if 1975, = 1 if 1977).

$Z_{jd}$ = 1 if state j has a statute in year d. $Z_{jd}$ = 0 otherwise.

The observed murder rate is $Y_{jd} = Y_{jd}(1)Z_{jd} + Y_{jd}(0)(1 - Z_{jd})$.

The Law of Iterated Expectations gives

$E[Y_d(1)] =$

$$E[Y_d(1)|Z_d = 1]P(Z_d = 1) + E[Y_d(1)|Z_d = 0]P(Z_d = 0).$$

The data reveal that

$P(Z_1 = 1) = 0.70.$ $E[Y_1(1)|Z_1 = 1] = 9.7.$

The data do not reveal $E[Y_d(1)|Z_d = 0]$.

*Linear Homogeneous Treatment Response*

Let $Y_{jd}(t) = \alpha_j + \beta \cdot d + \gamma \cdot t + \delta_{jd}$.

Then $Y_{jd} = \alpha_j + \beta \cdot d + \gamma \cdot Z_{jd} + \delta_{jd}$.

Let $X_j = 1$ or 0 if j is treated ($Z_{j1} = 1$) or untreated ($Z_{j1} = 0$).

Assume that $E(\delta_d | X, Z_d) = 0$ and $E(\alpha | X, Z_d) = E(\alpha | X)$.

Then $E(Y_d | X, Z_d) = E(\alpha | X) + \beta \cdot d + \gamma \cdot Z_d$.

It follows that $\gamma$ has the DID form

$$\gamma = [E(Y_1 | X = 1, Z_1 = 1) - E(Y_0 | X = 1, Z_0 = 0)] - [E(Y_1 | X = 0, Z_1 = 0) - E(Y_0 | X = 0, Z_0 = 0)].$$

*Partial Identification Assuming Bounded Outcomes*

Recall that

$$E[Y_d(1)] =$$

$$E[Y_d(1)|Z_d = 1]\cdot P(Z_d = 1) + E[Y_d(1)|Z_d = 0]\cdot P(Z_d = 0).$$

Across all states and years, the observed homicide rate always was in the range [0.8, 32.8].

Assume that $E[Y_1(1)|\, Z_1 = 0] \in [0, 35]$. Then

$$E[Y_d(1)] \in \{E[Y_d(1)|Z_d = 1]\cdot P(Z_d = 1) + 0\cdot P(Z_d = 0),$$
$$E[Y_d(1)|Z_d = 1]\cdot P(Z_d = 1) + 35\cdot P(Z_d = 0)\}.$$

Table 2: Partial Identification of the ATE Under the Bounded Outcomes Assumption

| | 1975 | 1977 |
|---|---|---|
| Probability of Death Penalty Statute: $P(Z_d = 1)$ | 0 | 0.7 |
| Murder Rate with Statute: $E[Y_d(1)\|Z_d = 1]$ | N. A. | 9.7 |
| Murder Rate without Statute: $E[Y_d(0)\|Z_d = 0]$ | 9.6 | 6.9 |
| **Bounds:** | | |
| $E[Y_d(1)]$ | [0, 35] | [6.8 , 17.3] |
| $E[Y_d(0)]$ | 9.6 | [2.1 , 26.6] |
| $ATE_d$ | [-9.6, 25.4] | [-19.8 ,15.2] |

## Middle-Ground Assumptions

We consider assumptions that presume some commonality across states or time, but not homogeneity.

We do not endorse any particular assumption. We demonstrate how the conclusions drawn depends on the assumptions imposed.

This provides a spectrum of possibilities to readers of research on deterrence.

*Date-Invariant Treatment Effects*

Assume that $ATE_1 = ATE_0$.

Then the date-invariant ATE must lie in the intersection of the two date-specific intervals shown in Table 1, these being [-9.6, 25.4] and [-19.8, 15.2]. The result is [−9.6, 15.2].

Here is a longer derivation that will have payoff later.

Assume that mean treatment response at date d has the form

$$E[Y_1(t)] = E[Y_0(t)] + \beta.$$

Then $E[Y_1(1)] - E[Y_1(0)] = E[Y_0(1)] - E[Y_0(0)]$.

Let $E_t \equiv E[Y_0(t)]$. Then $E[Y_1(t)] = E_t + \beta$.

These equations reduce the number of unknown mean potential outcomes by one.

Without the assumption, we do not know the four quantities $E[Y_d(t)]$, $d = 0, 1$; $t = 0, 1$.

With it, we do not know the three quantities $(E_0, E_1, \beta)$.

To obtain the identifying power of the assumption, first consider each pair (d, t) and obtain the identification region for $E[Y_d(t)]$ using only the assumption of bounded outcomes. Let this interval be called $[L_d(t), U_d(t)]$.

Combining this with date invariance yields

$$L_0(t) \le E_t \le U_0(t), \qquad t = 0, 1;$$
$$L_1(t) \le E_t + \beta \le U_1(t), \qquad t = 0, 1.$$

The feasible values of $(\beta, E_0, E_1)$ satisfy the four inequalities.

Then $\beta \in [-7.5, 17.0]$, $E_o = 9.6$, $E_1 \in [0, 24.8]$.

Hence, ATE $\in [-9.6, 15.2]$.

*Bounding Time-Series Variation in Mean Response Levels*

The above analysis did not restrict time-series variation in treatment response levels.

The assumption and data implied that $\beta \in [-7.5, 17.0]$.

One might not think it credible that large variations in potential homicide rates could have occurred over a short time period. Assuming that $\beta$ lies in a narrower interval may imply a narrower bound on the ATE.

Figure 1 displays the bound on the ATE as a function of $\beta$.

$\beta \leq -3 \Rightarrow \text{ATE} > 0.$

$\beta = 0 \Rightarrow \text{ATE} \in [-2.8, 7.7]$

$\beta \geq 8 \Rightarrow \text{ATE} < 0.$

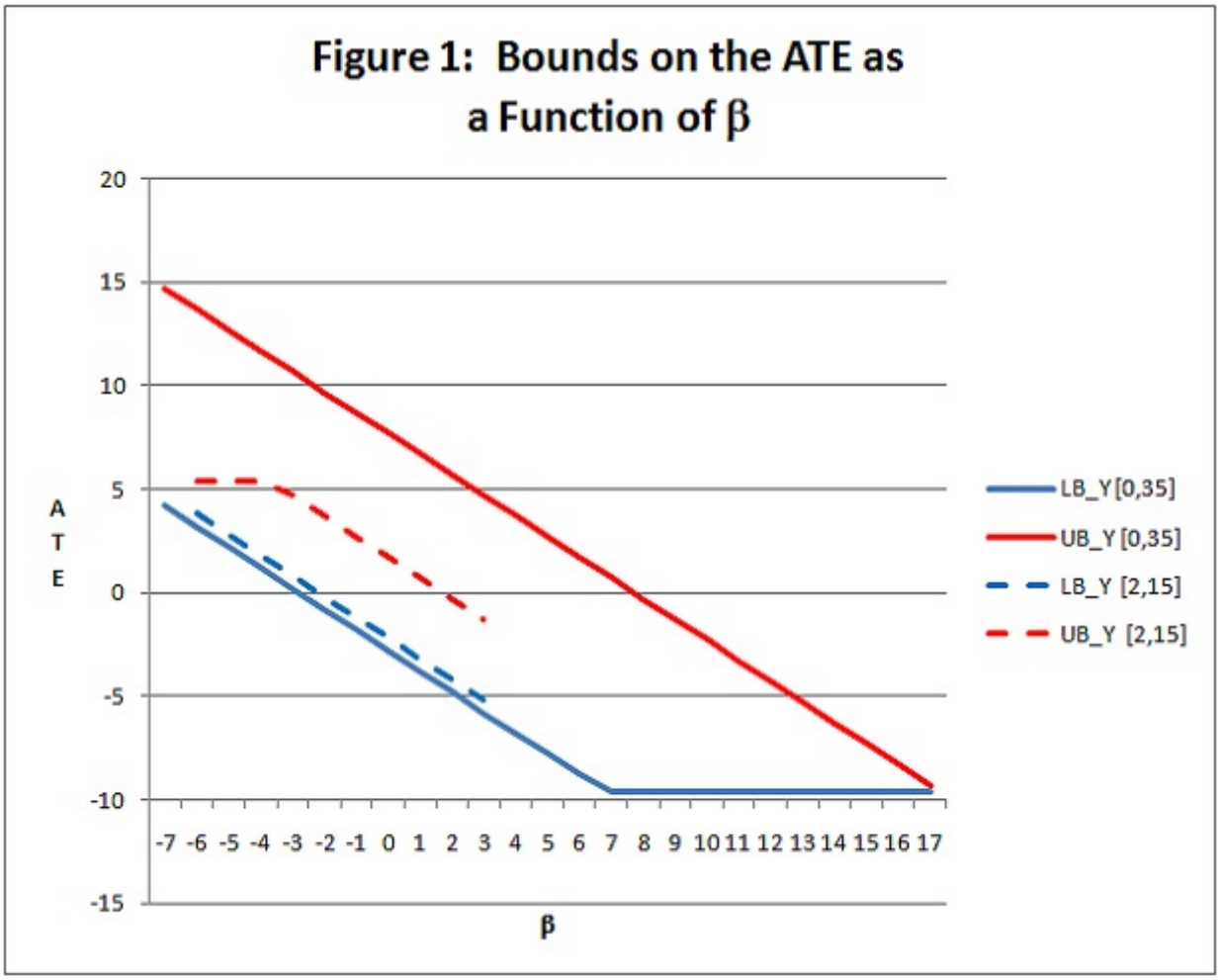


Figure 1: Bounds on the ATE as a Function of β
20
15
10
5
0
-5
-10
-15
ATE
-7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
β
LB_Y [0,35]
UB_Y [0,35]
LB_Y [2,15]
UB_Y [2,15]

*Tighter Bounds on Counterfactual Mean Response Levels*

We have so far assumed only that $0 \leq E[Y_d(t) | Z_d \neq t] \leq 35$.

Of the 102 state-specific homicide rates in 1975 and 1977, the central ninety percent fall in the interval [2, 15].

Assume that $2 \leq E[Y_d(t) | Z_d \neq t] \leq 15$. Then

$\in [-6.1, 3.0]$, ATE $\in [-5.2, 5.4]$.

Figure 1 displays how the bound on the ATE varies with .

$\beta \leq -2 \Rightarrow$ ATE $> 0$.

$\beta = 0 \Rightarrow$ ATE $\in [-2.2, 1.7]$

$\beta \geq 2 \Rightarrow$ ATE $< 0$.

*Date-Invariant Treatment Effects Combined with Covariate-Invariant Date Intercepts*

Let X separate states into K groups.

Let $E_{t|X} \equiv E[Y_0(t)|X]$. Assume that $E[Y_1(t)|X] = E_{t|X} + \beta$.

Let $[L_d(t|x), U_d(t|x)]$ be the bound on $E[Y_d(t)|X]$ obtained using only the assumption that outcomes lie in [0, 35]. Then

$$L_0(t|X) \le E_{t|X} \le U_0(t|X), \qquad t = 0, 1; \text{ all } X$$
$$L_1(t|X) \le E_{t|X} + \beta \le U_1(t|X), \qquad t = 0, 1; \text{ all } X.$$

The feasible values of the (2K + 1) unknowns ($\beta$, $E_{0|X}$, $E_{1|X}$, all X) satisfy these 4K inequalities.

We evaluate the ATE with two definitions of X.

First, X indicates whether a state has a death penalty statute in 1977.

Second, we let X indicate the location of a state in one of four census regions.

*Treatment Group as the Covariate*

Let $X_j = 1$ if $Z_{j1} = 1$ and $X_j = 0$ if $Z_{j1} = 0$. Then the eight inequalities are

$$\begin{aligned} E_{0|0} &= E(Y_0|X=0), & E_{0|0}+\beta &= E(Y_1|X=0),\\ 0 &\le E_{1|0} \le 35, & 0 &\le E_{1|0}+\beta \le 35,\\ E_{0|1} &= E(Y_0|X=1), & 0 &\le E_{0|1}+\beta \le 35,\\ 0 &\le E_{1|1} \le 35, & E_{1|1}+\beta &= E(Y_1|X=1). \end{aligned}$$

The equalities in the first row point-identify the date intercept

$$\beta = E(Y_1|X=0) - E(Y_0|X=0).$$

We find $\beta = -1.1$.

With knowledge of $\beta$,

$$E_{0|0} = E(Y_0 | X = 0).$$
$$E_{0|1} = E(Y_0 | X = 1).$$
$$E_{1|0} \in [\max(0, -\beta), \min(35, 35 - \beta)].$$
$$E_{1|1} = E(Y_1 | X = 1) - \beta.$$

Then

$$ETT \equiv E_{1|1} - E0|1$$
$$= [E(Y_1|X=1) - E(Y_0|X=1)] - [E(Y_1|X=0) - E(Y_0|X=0)].$$

$$ETU \equiv E_{1|0} - E_{0|0}$$
$$\in [\max(0, -\beta) - E(Y_0|X=0), \min(35, 35 - \beta) - E(Y_0|X=0)].$$

| Table 3: Treatment Effects with Date-Invariant Treatment Effects, with and without Covariate-Invariant Date Intercepts | | | |
|---|---|---|---|
| Assumption | ETT | ETU | ATE |
| Linear Response | 0.5 | 0.5 | 0.5 |
| Bounded Outcomes, 1977 | [−25.3, 9.7] | [−6.9, 28.1] | [−19.8,15.2] |
| Bounded Outcomes, 1975 | [−35.0, 35.0] | [−9.6, 25.4] | [−9.6, 25.4] |
| Date-Invariant Treatment Effects | | | |
| No Covariate | | | [−9.6, 15.2] |
| Region as Covariate | | | [−9.0, 10.1] |
| Treatment Group as Covariate | 0.5 | [−6.9, 27.0] | [−1.9 , 8.3] |

*Bounded Instrumental Variables*

Traditional instrumental variables (IVs) assume that specified groups of treatment units have the same mean treatment response or the same average treatment effects.

It often is difficult to motivate IV assumptions, but it may be easier to motivate weaker ones asserting that mean response or average treatment effects do not differ too much across groups.

Such assumptions uses *bounded instrumental variables*.

To demonstrate, consider identification of the ATE when the researcher selects a $\Delta \geq 0$ and assumes that

$$| ATE_{dx} - ATE_{dx'} | \leq \qquad \text{for all x and x}'.$$

$= 0$ gives a traditional IV assumption asserting that groups of states with different covariates have the same ATE.

$\Delta > 0$ supposes that the ATE may differ across groups by no more than $\Delta$.

Figure 2 maintains the earlier assumptions, adds the new one, and displays the bound on the ATE as a function of $\Delta$.

Figure 3 sets $\Delta=0$ and displays the bound as a function of $\beta$.

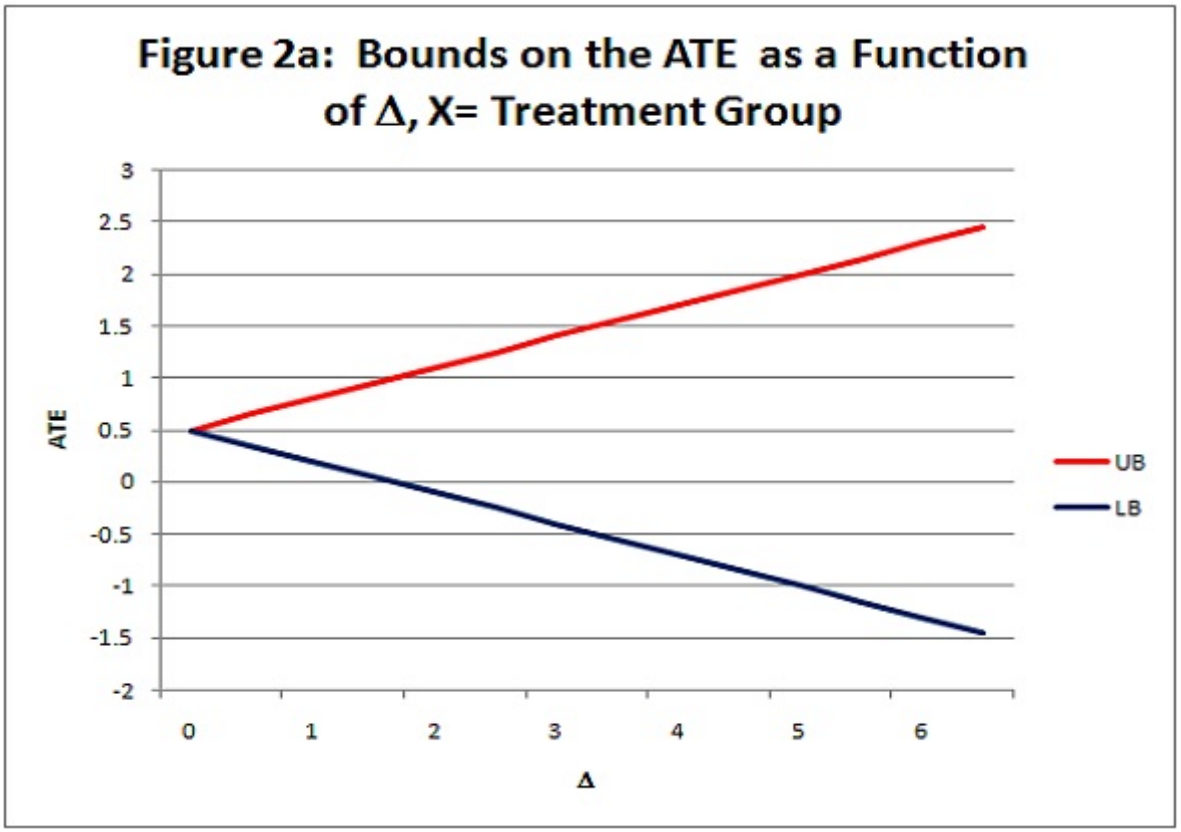


Figure 2a: Bounds on the ATE as a Function of Δ, X= Treatment Group
ATE
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
1
2
3
4
5
6
Δ
UB
LB

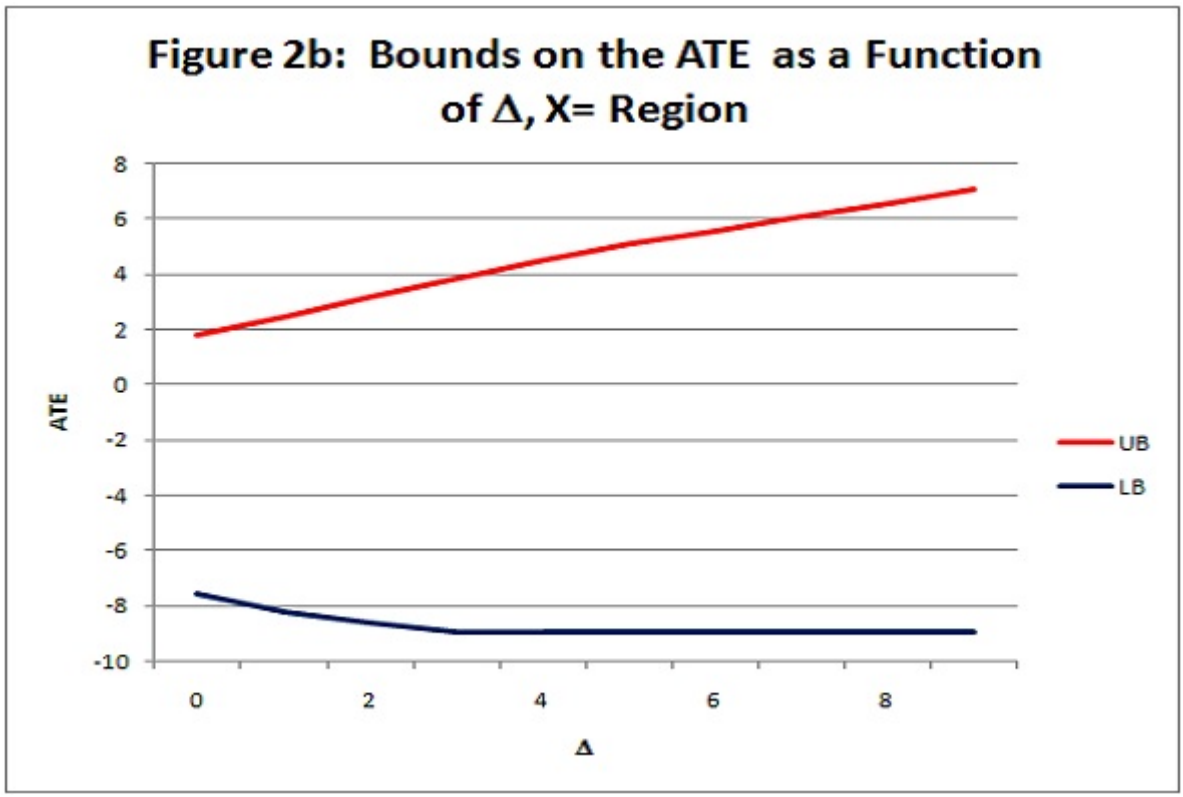


Figure 2b: Bounds on the ATE as a Function of Δ, X= Region
ATE
8
6
4
2
0
-2
-4
-6
-8
-10
0
2
4
6
8
Δ
UB
LB

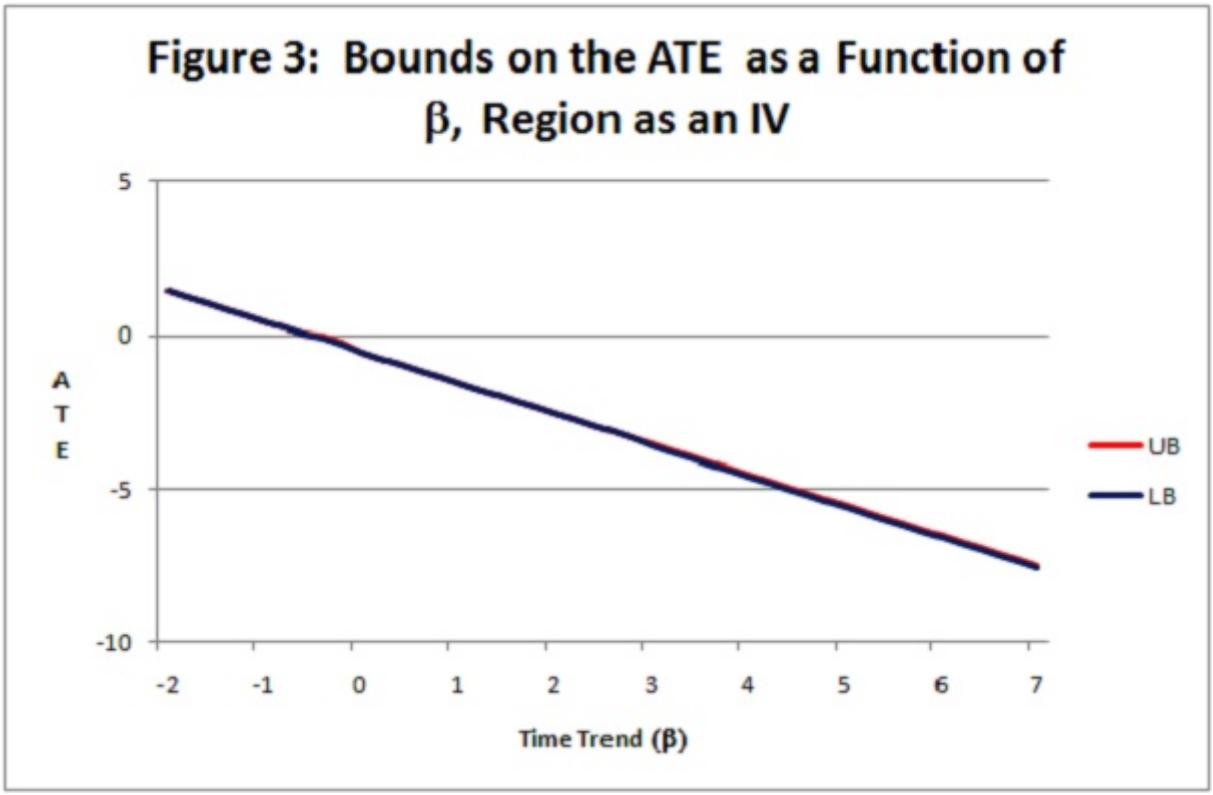


Figure 3: Bounds on the ATE as a Function of β, Region as an IV
5
0
-5
-10
ATE
-2
-1
0
1
2
3
4
5
6
7
Time Trend (β)
UB
LB

# Conclusion

Researchers have had persistent problems providing credible inference on the deterrent effect of the death penalty.

The 1978 NRC report warned of the shortcomings of the data and methods, and questioned whether empirical research could provide useful information at all.

Nevertheless, researchers have continued to examine the same or more recent data using similar methods.

Research continues to combine data with untenable assumptions. The results have been highly sensitive to these assumptions and no consensus has emerged.

We demonstrate what can be learned under relatively weak assumptions.

We study the identifying power of assumptions restricting variation in treatment response across places and time.

The results bound the deterrent effect of capital punishment.

By successively adding stronger identifying assumptions, the analysis makes transparent how assumptions shape inferences.

If one assumes only that outcomes are bounded, one cannot identify the sign of the ATE and one can only draw weak conclusions about its magnitude.

Those who find it credible to make further assumptions can obtain more informative findings.

Society at large can draw strong conclusions only if there is a consensus favoring particular assumptions.

Without such a consensus, data on sanctions and murder rates cannot settle the debate about deterrence.

However, data combined with weak assumptions can bound and focus the debate.

# PREDICTING BEHAVIOR

Suppose that everyone in the study population received A and no one received B. Thus, A is the *status quo* and B is an *innovation*.

The assumptions examined so far (individualistic response, identical units, identical groups) have no power to predict outcomes under an innovation.

How might one proceed?

*Modeling Treatment Response*

A broad idea is to model treatment response, imposing assumptions that relate the outcomes a person would experience under treatments A and B.

Such assumptions, combined with observation of the study population, can yield conclusions about the outcomes that would occur if persons were to receive the innovation.

*Revealed Preference Analysis*

Economists use this idea to predict behavioral response to new policy.

A treatment is the set of alternatives available to a person, called the *choice set*.

One observes the choice that persons make when facing status quo choice sets.

The problem is to predict the choices that persons would make when facing new choice sets.

Economists assume that a person prefers his choice to all other available alternatives.

Hence, observation of a choice partially reveals the person's preferences.

This partial knowledge of preferences, combined with assumptions about preferences, may enable prediction of the choice the person would make if facing a new choice set.

Assumptions that relate the preferences of different persons may also have predictive power.

# Identification of Income-leisure Preferences and Evaluation of Income Tax Policy


Summary

The merits of alternative income tax policies depend on preferences for income, leisure, and public goods.

Standard theory, which supposes that persons want more income and more leisure, does not predict how they resolve the tension between these desires.

Empirical studies of labor supply have not shed much light on the matter.

Researchers have imposed strong preference assumptions that lack foundation.

This paper examines the problem of inference on preferences.

I perform a basic revealed-preference analysis that imposes no assumptions on preferences beyond that persons prefer more income and leisure.

I find that observation of a person's labor supply under a status quo tax policy may bound his labor supply under a proposed policy or may have no implications, depending on the shapes of the two tax schedules and the location of status quo labor supply.

I next explore the identifying power of two assumptions restricting the distribution of income-leisure preferences.

One assumes that groups of persons who face different choice sets have the same distribution of preferences. The other adds restrictions on the shape of this distribution.

I then address utilitarian policy comparison with partial knowledge of preferences.

Partial knowledge of preferences implies partial knowledge of the welfare function. Hence, it may not be possible to rank policies.

# Background

Among his simplifying assumptions, Mirrlees (1971) wrote:

"The State is supposed to have perfect information about the individuals in the economy, their utilities and, consequently, their actions."

In his conclusion, he wrote:

"The examples discussed confirm, as one would expect, that the shape of the optimum earned-income tax schedule is rather sensitive to the distribution of skills within the population, and to the income-leisure preferences postulated. Neither is easy to estimate for real economies."

Economic theory is silent on the response of labor supply to income taxation.

The consensus in the empirical literature is that increasing tax rates usually reduces work effort. Researchers differ on the magnitude of the elasticities but not the sign.

The consensus on the sign of elasticities may be an artifact of model specification.

Researchers generally assume that labor supply varies monotonically with net wages.

They assume that the response of labor supply to net wage is homogeneous within broad groups.

Neither assumption has a foundation in theory or evidence.

# Basic Analysis of Revealed Preference

I assume only that utility is an increasing function of net income and leisure. In short, more is better.

Predicting population labor supply under a proposed policy simply requires aggregation of individual predictions. Hence, I focus on one person.

Tax Policy and Labor Supply

Person j has wage $w_j$ and unearned income $z_j$. He allocates one unit of time between leisure (L) and work (1 − L).

The status quo tax policy, S, subtracts tax revenue $R_{jS}(L)$ from gross income, leaving j with net income

$$Y_{jS}(L) \equiv w_j(1 - L) + z_j - R_{jS}(L).$$

The $R_{jS}(\cdot)$ notation allows the tax schedule to be specific to person j. Taxes may take positive or negative values.

j chooses L from choice set $\Lambda_j \subset [0, 1]$.

$U_j(Y, L)$ is the utility of (net income, leisure) pair (Y, L). Utility is strictly increasing in both arguments.

Let $L_{jS} \in \Lambda_j$ denote chosen leisure under policy S. Utility maximization implies

$$U_j[Y_{jS}(L_{jS}), L_{jS}] \geq U_j[Y_{jS}(L), L], \text{ all } L \in \Lambda_j.$$

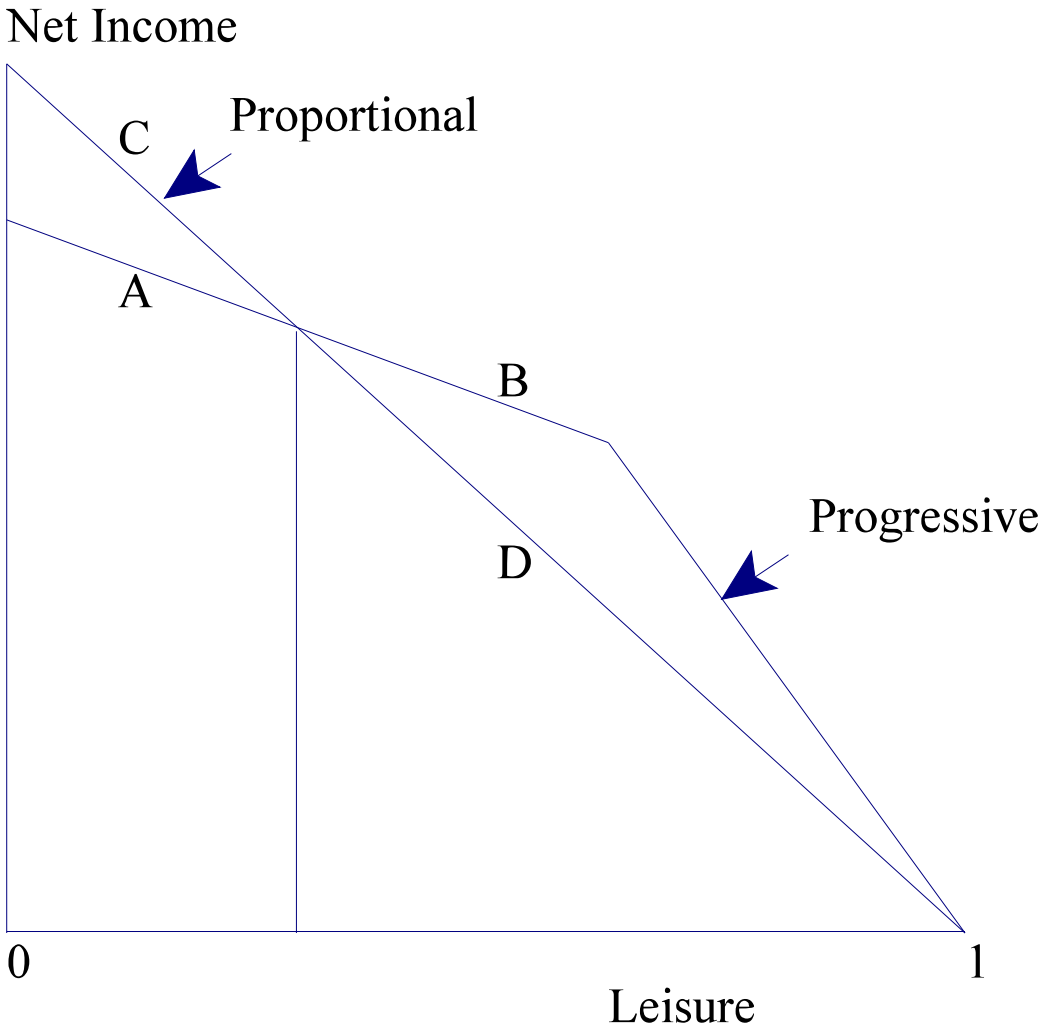
Note: Using $U_j(\cdot, \cdot)$ to express preferences suppresses the dependence of preferences on the public goods produced with tax revenue.

*Predicting Labor Supply under a Proposed Tax Schedule*

Let $R_{jT}(\cdot)$ denote the tax schedule under proposed policy T.

Suppose that one observes $L_{jS,}$ $Y_{jS}(\cdot)$, and $Y_{jT}(\cdot)$.

What can one predict about time allocation under T?


Net Income
Proportional
C
A
B
Progressive
D
0
1
Leisure

Let $L_{jS} < L_j^*$ and consider any $L > L_j^*$.

More is better $\Rightarrow$ $[Y_{jT}(L_{jS}), L_{jS}] \succ [Y_{jS}(L_{jS}), L_{jS}]$.

$[Y_{jS}(L), L] \succ [Y_{jT}(L), L]$.

Revealed preference $\Rightarrow$ $[Y_{jS}(L_{jS}), L_{jS}] \succeq [Y_{jS}(L), L]$.

Hence, $[Y_{jT}(L_{jS}), L_{jS}] \succ [Y_{jT}(L), L]$.

Thus, under policy T, person j would not choose any $L > L_j^*$.

*General Analysis*

Let

$\Lambda_{j<} \equiv \{L_< \in \Lambda_j: [Y_{jT}(L_<), L_<] < [Y_{jS}(L), L] \text{ for some } L \in \Lambda_j\}$.

$\Lambda_{j>} \equiv \{L_> \in \Lambda_j: [Y_{jT}(L_>), L_>] > [Y_{jS}(L_{jS}), L_{jS}]\}$.

Define $\Lambda_{j\leq}$ and $\Lambda_{j\geq}$ analogously.

*Proposition 1*:

Let $\Lambda_{j<}$ and $\Lambda_{j\geq}$ be non-empty. If j were to face T, he would not choose any $L \in \Lambda_{j<}$.

Let $\Lambda_{j\leq}$ and $\Lambda_{j>}$ be non-empty. If j were to face T, he would not choose any $L \in \Lambda_{j\leq}$.

*Welfare Implications*

If $\Lambda_{j<} = \Lambda_j$, replacement of S with T strictly decreases utility.

If $\Lambda_{j>}$ is non-empty, replacement of S with T strictly increases utility.

Basic analysis yields no welfare conclusions otherwise.

*Tax Schedules that Cross Once*

Let $Y_{jS}(\cdot)$ and $Y_{jT}(\cdot)$ be downward sloping.

Basic analysis can have predictive power for labor supply only if the two net-income functions cross at least once.

Let $Y_{jT}(\cdot)$ cross $Y_{jS}(\cdot)$ from above. Then

$\Lambda_{j<} = [L \in \Lambda_j: L > L_j^*]$.

$\Lambda_{j>}$ is non-empty if and only if $L_{jS} < L_j^*$.

*Aggregate Labor Supply*

Let $(J, \Omega, P)$ denote a population.

Let net income under T cross S from above.

Consider the sub-population whose net incomes cross at point $\ell$. Then

$$P(L_S < L^* \,|\, L^* = \ell) \;\leq\; P(L_T < L^* \,|\, L^* = \ell).$$

This is the only limited sense in which progressivity necessarily decreases labor supply.

## Restrictions on the Preference Distribution

To explore the middle ground between basic analysis and current practice, I use the discrete-choice framework developed in Manski (*IER*, 2007).

Let S and T generate a finite universe of potential (Y, L) values. That is,

$$A \equiv \{[Y_{jS}(L), L], [Y_{jT}(L), L], L \in \Lambda_j, j \in J\}$$

has finite cardinality.

Let almost all persons have strict preference orderings. There are $|A|!$ strict orderings on A. The number of feasible orderings is smaller because the assumption that more-is-better excludes some orderings.

*Numerical Illustration*

Let S tax at 15% up to \$50,000 and 25% above \$50,000. Let T tax at 20%. Let $\Lambda_j = \{0, ½\}$ for all $j \in J$.

Let persons have w = \$150,000 or \$500,000. Then

| | | Status Quo Schedule | | Proposed Schedule |
|---|---|---|---|---|
| (w = \$150,000) | | | | |
| Full-time work | a | [117,500, 0] | e | [120,000, 0] |
| Half-time work | b | [ 61,250, ½] | f | [ 60,000, ½] |
| (w = \$500,000) | | | | |
| Full-time work | c | [380,000, 0] | g | [400,000, 0] |
| Half-time work | d | [192,500, ½] | h | [200,000, ½]. |

The number of strict preference orderings is 8! = 40,320.

More-is-better implies that $g \succ c \succ e \succ a$, $h \succ d \succ b \succ f$, and $d \succ e$. Only 53 orderings satisfy these inequalities.

Identification Analysis

Let $A_k$, $k \in K$ denote the orderings satisfying more-is-better.

The preference distribution is multinomial with at most $|K|$ mass points.

Assumptions about preferences are restrictions on this multinomial distribution.

I consider these assumptions.

1. Groups who face different choice sets have the same distribution of preferences.

2. shape restrictions on the preference distribution.

Divide J into ($J_m$, $m \in M$) mutually exclusive and exhaustive groups according to the choice sets they face under S and T.

All members of group m face a common choice set $C_m$ of (Y, L) pairs under S and a common set $D_m$ under T. Groups m and m′ are distinct if $C_m \neq C_{m'}$ or $D_m \neq D_{m'}$.

Also divide J into a finite collection X of groups of persons with common observed covariates.

Define $J_{mx}$ accordingly.

Let $\pi_{mx} \equiv (\pi_{mxk}, k \in K)$ denote the distribution of preferences in group (m, x).

*Assumption 1*: For a specified class of groups $N \subset M \times X$, there exists a single distribution $\pi_N \equiv (\pi_{Nk}, k \in K)$ such that $\pi_{mx} = \pi_N$, all $(m, x) \in N$.

*Assumption 2*: $\pi_N$ lies in a specified set $\Pi_N$ of multinomial distributions.

Let $c_k(B)$ denote the (Y, L) pair that a person with preference ordering k would choose if he were to face choice set B.

Let c(B) denote the choice of a person randomly drawn from J.

Let $P[c(C_m) = (Y, L) \mid J_{mx}]$ give the observed fraction of sub-population $J_{mx}$ who make choice (Y, L) when facing their status quo choice set $C_m$.

Let $P[c(D_m) = (Y, L) \mid J_{mx}]$ give the unobserved fraction of persons in $J_{mx}$ who would choose (Y, L) if they were to face choice set $D_m$.

Assumption 1 yields

$$P[c(C_m) = (Y, L) \mid J_{mx}] = \sum_{k \in K} 1[c_k(C_m) = (Y, L)] \cdot \pi_{Nk},$$

$$(Y, L) \in C_m, \quad (m, x) \in N.$$

Assumption 2 states that $\pi_N \in \Pi_N$.

Also $\pi_N$ satisfies $\sum_{k \in K} \pi_{Nk} = 1, \quad \pi_{Nk} \geq 0, \quad k \in K.$

*Proposition 2*: Given Assumptions 1 and 2, the identification region for $\pi_N$ is the set $H(\pi_N)$ of multinomial distributions that satisfy these conditions.

When $\Pi_N$ places linear restrictions on $\pi_N$, $H(\pi_N)$ is the convex set of vectors in $R^{|K|}$ that solve a set of linear equalities and inequalities.

$H(\pi_N)$ is non-empty if the assumptions are correct. If $H(\pi_N)$ is empty, some assumption is incorrect.

When $H(\pi_N)$ is non-empty, $H(\pi_N)$ yields identification regions for choice probabilities and for tax revenue under T.

## Discussion

The analysis is mathematically simple. When $\Pi_N$ places linear restrictions on $\pi_N$, computation is tractable.

One would like to be able to characterize the identifying power of alternative assumptions. When $\Pi_N$ places linear restrictions on $\pi_N$, one can count the number of linear equations and inequalities. However, the structure of $H(\pi_N)$ depends delicately on the choice sets that the various groups face under S.

There does not seem to be an effective shortcut to characterize $H(\pi_N)$ or derived quantities. It appears necessary to impose assumptions and compute the identification regions.

A substantive issue is that we lack a credible basis for placing informative restrictions on preference distributions.

Standard theory only suggests that more-is-better. It is not clear how to credibly go beyond this.

The preference assumptions used in empirical research on labor supply have been motivated by a desire to obtain tractable point estimates, not by theory or evidence.

A Computational Experiment

Given data on labor supply under S, the problem is to predict tax revenue per capita under T.

I show the predictive power of a sequence of assumptions:
(1) more is better
(2) + persons in specified wage groups have the same distribution of preferences
(3) + preferences have the CES form
(4) + all CES utility functions in a wage group have the same elasticity of substitution.

*Tax Policies, Wages, Choice Sets, and Preferences*

Let S tax at 20% up to $100,000 and 30% above $100,000. Let T tax at 25%. The schedules cross when gross income equals $200,000.

The population has persons whose wages are ${50, 100, 150, 200, 250, 300, 350, 400} x 1000. The distribution is

$P(w = 50) = 0.70$, $P(w = 100) = 0.20$, $P(w = 150) = 0.05$,
$P(w = 200) = 0.02$, $P(w = 250) = 0.01$, $P(w = 300) = 0.0075$,
$P(w = 350) = 0.0075$, $P(w = 400) = 0.005$.

$\Lambda_j = \{0, ¼, ½, 1\}$ for all $j \in J$.

CES utility functions have the form

$$U_j(Y, L) = [\alpha_j(Y/400{,}000)^{\rho_j} + (1 - \alpha_j)L^{\rho_j}]^{1/\rho_j},$$

$\alpha \in \{0, 0.01, 0.02, \ldots, 0.99, 1\}$

$\rho \in \{-100, -90, \ldots, -20, -10, -1, -0.99, .0, \ldots, 0.99, 1\}.$

Hence, there are $101 \times 211 = 21{,}311$ CES preferences.

The population actually contains persons with 20 distinct CES preferences:

$(\alpha, \rho) \in \{0.25, 0.5, 0.65, 0.75\} \times \{-100, -0.5, 0, 0.5, 1\}.$

The distribution of $(\alpha, \rho)$ is uniform conditional on wage.

*Findings*

Actual tax revenues under S and T are $7,652 and $9,211.

Knowing only the wage distribution but not labor supply under S, revenue under T lies in the interval [0, $18969].

With data on status quo labor supply, assuming more-is-better yields the bound [$593, $18969].

Combine more-is-better with a version of Assumption 1: Persons with $w \leq \$200{,}000$ have the same preference distribution. Those with $w > \$200{,}000$ have the same distribution. Then revenue $\in$ [$3744, $14149].

Next suppose that all persons have CES utility functions. This yields the bound [$6883, $10444].

Finally, assume that the persons in each wage group have the same elasticity of substitution $\rho$. This yields an empty identification region for the preference distribution within the second wage group. Thus, the assumption of homogeneous $\rho$ within each wage group is rejected.

*Computation of the Bounds*

Computation ranges from easy to challenging.

The bound assuming more-is-better is simple to compute.

The bound assuming CES preferences is tractable because the number of distinct choice functions is fairly small. The set of 21,311 distinct values of $(\alpha, \rho)$ partitions into a small number of 108 or 99 equivalency classes, within which distinct $(\alpha, \rho)$ values yield the same choices under both tax schedules.

The most challenging task is to compute the bound assuming that persons within each wage group have the same distribution of preferences, but not assuming that preferences are CES.

The members of each wage group face 24 distinct (Y, L) pairs under S and T. Thus, the total number of strict preference orderings in each group is $24! \cong 6.2 \times 10^{23}$.

An algorithm that determines and excludes orderings inconsistent with the more-is-better assumption reduces the number of feasible orderings to 7, 654 and 4,000.

# Utilitarian Policy Evaluation

A familiar exercise in normative public economics poses a utilitarian social welfare function and ranks tax policies by the welfare they achieve.

This requires knowledge of preferences to (1) predict tax revenues and (2) compute welfare.

There are multiple difficulties.

1. Basic revealed preference analysis does not predict labor supply under policies that change the production of public goods that directly affect utility.

2. Computation of welfare requires knowledge of the preferences of the population, not just their labor supply.

3. Policies which fix public-good production before observation of labor supply generically yield budget surpluses or deficits, because the ability to predict labor supply is incomplete.

# Enriching the Data for Identification of Preferences

We lack the knowledge of preferences necessary to perform credible utilitarian evaluation of income tax policy.

The consensus that increasing tax rates reduces work effort is premature.

Knowledge of preferences for public goods is almost non-existent.

I do not expect that new theory will sharpen our knowledge of preferences.

The only way is to obtain richer data.

*Promoting Exogenous Variation in Choice Sets*

Economists have long appreciated the identifying power of exogenous variation in choice sets.

In this paper, exogenous variation is expressed through Assumption 1, which places restrictions on the preference distribution assumed to be invariant across groups facing different status-quo choice sets.

Natural variation in choice sets arises from wage heterogeneity and from the diversity of tax schedules across jurisdictions.

Governments may be able to enhance variation by decentralizing tax policy and by performing experiments that randomize persons into alternative tax schedules.

*Observation of Individual Behavior in Multiple Choice Settings*

My analysis has supposed that a researcher observes one status-quo time allocation per person.

The identifying power of revealed preference analysis grows if it is feasible to observe individual behavior in multiple choice settings.

## Identification

The stream of basic revealed preference analysis running from Samuelson (1938, 1948) through Afriat (1967), Varian (1982), and others recognizes that observation of neoclassical consumer demand has increasing power to predict counterfactual behavior as choices from more budget sets are observed.

I expect that basic revealed preference analysis of time allocation in settings with multiple choice observations would yield similar qualitative results.

That is, transitivity and the assumption that more is better should yield tighter bounds on counterfactual labor supply than when only one status-quo time allocation is observed.

Extension of the discrete choice analysis in this paper to settings with multiple choice observations per person is a straightforward generalization of Proposition 2.

The basic idea is that observation of multiple choices enlarges the set of linear restrictions that Assumption 1 places on the preference distribution. Hence, it tightens the identification region.

## Data Sources

### *Longitudinal observation of time allocation under varying wages or tax schedules*

An analytical issue is that interpretation of longitudinal data as in this paper rests on acceptance of the classical static model.

If labor supply is dynamic, repeated observations provide data on one sequential choice path rather than data on choice from multiple independent choice sets.

Interpretation of the data from a dynamic perspective requires assumptions about the information that persons possess when making choices, the way they form expectations for relevant future events, and the criteria they use to make decisions under uncertainty.

*Stated-choice analysis*

A researcher may pose multiple hypothetical choice settings to a person and ask the person to predict the choice he would make in each setting.

A practical advantage is that a researcher can elicit predictions of behavior in a wide spectrum of hypothetical choice settings.

An analytical issue is that interpretation of stated-choice data requires assumptions about the way that persons construe the scenarios posed and the cognitive processes they use when responding to questions.

Collecting longitudinal and stated-choice data will not "solve" the problem of identifying income-leisure preferences.

Yet I think that careful efforts to collect and study such data can add significantly to the little we now know.

# PLANNING WITH PARTIAL KNOWLEDGE

I have explained some of the immense difficulty of predicting policy outcomes.

The point predictions produced by analysts are achieved by imposing strong assumptions that rarely have foundation.

Analysis with more credible assumptions typically yields interval rather than point predictions.

How may policy making reasonably cope with this difficulty?

I use elementary decision theory to study policy choice by a *planner* — an actual or idealized solitary decision maker who acts on behalf of society.

Why study a planner, when policy in democracies emerges from the interaction of many persons and institutions?

**Policy choice in an uncertain world is subtle even when a society agrees on what it wants and what it believes.**

**Even in a cohesive society, there is no optimal decision, at most various reasonable ones.**

A planner personifies a cohesive society that forthrightly acknowledges and copes with uncertainty.

*Illustration: Treating X-Pox*

Suppose that a new disease called x-pox is sweeping a community. It is impossible to avoid infection. If untreated, infected persons always die.

Medical researchers propose treatments A and B.

Researchers know that one is effective, but they do not know which one. Administering both is fatal.

Thus, a person will survive if and only if she is administered the effective treatment alone.

There is no time to experiment.

A health agency must decide how to treat the community. The agency wants to maximize the survival rate.

It can select one treatment and administer it to everyone. Then the entire population will either live or die.

Or it can give one treatment to some fraction of the community and the other to the remaining fraction.

If half the community receives each treatment, the survival rate is certain to be fifty percent.

What should the agency do?

Decision theory can be used to motivate either policy.

Economists typically assume that the decision maker places a subjective probability distribution on unknown quantities and maximizes *expected utility*.

A health agency treating X-pox would give everyone the treatment with the higher subjective probability of success.

A decision maker lacking a subjective distribution faces a problem of choice under *ambiguity*.

A health agency applying the *maximin* or *minimax-regret* criterion would give half the population each treatment.

## Diversified Treatment under Ambiguity

In the x-pox example, giving half the population A and half B diversifies treatment.

An individual cannot not diversify her own treatment. Each person receives one treatment and either lives or dies.

Yet the community can diversify by having positive fractions of the population receive each treatment.

Thus, private diversification of treatment for x-pox is impossible, but communal diversification is possible.

*Financial Diversification*

Financial diversification is a familiar recommendation for portfolio allocation.

A portfolio is diversified if an investor allocates positive fractions of wealth to different investments.

An investor with full knowledge would not diversify. He would invest fully in the investment with the highest return.

**The rationale for diversification arises purely from incompleteness of knowledge.**

Diversification enables someone who is uncertain about the returns to investments to balance potential errors.

*Treatment Diversification*

Treatment choice is diversified if a planner allocates positive fractions of the population to each treatment.

Diversification enables a planner who is uncertain about treatment response to balance potential errors.

A *Type A error* occurs if a person receives treatment A but treatment B is better. A *Type B* error is analogous.

A planner who maximizes expected social welfare may diversify treatment if he is *risk averse*.

A planner who uses the minimax-regret criterion always diversifies. This criterion balances the potential welfare effects of Type A and Type B errors.

*Treatment Diversification Differs from Profiling*

Diversification calls for *randomly* different treatment of persons.

Profiling calls for *systematically* different treatment of persons who differ in observed attributes.

Profiling may be good policy when a planner knows how treatment response varies across persons.

Diversification may appeal when a planner does not know how treatment response varies across persons.

# One-Period Planning with Individualistic Treatment and Linear Welfare

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

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

Let $\alpha \equiv E[u(a)]$ and $\beta \equiv E[u(b)]$.

Mean welfare with allocation $\delta$ is

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

*Treatment Choice Under Ambiguity*

$\delta = 1$ is optimal if $\beta \geq \alpha$ and $\delta = 0$ 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

$$\alpha_L \equiv \min_{s \in S} \alpha_s, \quad \beta_L \equiv \min_{s \in S} \beta_s,$$
$$\alpha_U \equiv \max_{s \in S} \alpha_s, \quad \beta_U \equiv \max_{s \in S} \beta_s.$$

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

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

*The Minimax-Regret Allocation*

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

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

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

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

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

*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 assumptions relate $(\alpha_{\xi s}, \beta_{\xi s})$, $\xi \in X$.

*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

$$\delta_{FMR} = 0 \quad \text{if} \quad M(b) \leq \min \{M(a), \delta_{MR}M(a) + C\}$$
$$\delta_{FMR} = \delta_{MR} \quad \text{if} \quad \delta_{MR}M(a) + C \leq \min \{M(a), M(b)\}$$
$$\delta_{FMR} = 1 \quad \text{if} \quad M(a) \leq \min \{M(b), \delta_{MR}M(a) + C\}.$$

*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 *ex ante* sense that all persons have equal probabilities of receiving particular treatments.

Fractional allocations are inconsistent with equal treatment in the *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.

# Adaptive Diversification

Suppose that, in each period n = 0 , . . . , N, a planner must choose treatments for the current cohort of a population.

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

Diversification generates 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 *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 β.

| Table 1: Treating a Life-Threatening Disease | | | | | | |
|---|---|---|---|---|---|---|
| cohort or year (n or k) | death rate in $k^{th}$ year after treatment | | bound on β for cohort n | AMR allocation for cohort n | maximum regret of AMR allocation for cohort n | mean life span achieved by cohort n |
| | Status Quo | Innovation | | | | |
| 0 | | | [0, 5] | 0.30 | 1.05 | 3.74 |
| 1 | 0.20 | 0.10 | [0.90, 4.50] | 0.28 | 0.72 | 3.72 |
| 2 | 0.05 | 0.02 | [1.78, 4.42] | 0.35 | 0.60 | 3.78 |
| 3 | 0.05 | 0.02 | [2.64, 4.36] | 0.50 | 0.43 | 3.90 |
| 4 | 0.05 | 0.02 | [3.48, 4.32] | 0.98 | 0.02 | 4.28 |
| 5 | 0.05 | 0.02 | [4.30, 4.30] | 1 | 0 | 4.30 |

## *The AMR Criterion and Randomized Clinical Trials*

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.

*Adaptive Partial Drug Approval*

The modern drug approval process requires pharmaceutical firms to demonstrate that new drugs are safe and effective through performance of randomized clinical trials (RCTs).

In the U.S., a new drug goes through three phases of RCTs.

After phase 3, the firm files a New Drug Application. The Food and Drug Administration (FDA) approves or disapproves the drug after reviewing the trial findings.

FDA evaluation occurs with partial knowledge of treatment response.

A Type B error occurs when a new drug that is inferior to a status quo is approved because it appears superior when evaluated using the available information.

A Type A error occurs when a new drug that is superior to the status quo is disapproved because it appears inferior when evaluated using the available information.

Some Type B errors are eventually corrected through the FDA *post-market surveillance program*, which analyzes data on the outcomes experienced when the drug is used in clinical practice.

Type A errors often are permanent because, after a drug is disapproved, use of the drug ceases.

## Binary Versus Partial Approval

The FDA practice of framing approval as a yes/no decision between full approval and complete disapproval constrains the set of policy options.

The idea of adaptive diversification suggests establishment of an *adaptive partial drug approval* process, where the extent of the permitted use of a new drug would vary as evidence accumulates.

The stronger the evidence on outcomes of interest, the more that use of a new drug would be permitted.

If the FDA could assign medical treatments directly, it could implement adaptive diversification as described earlier.

The initial allocation of patients to the status quo and the new drug would reflect the initial knowledge available about the effectiveness of the innovation.

As evidence from trials accumulates, the FDA would revise the allocation accordingly.

Eventually, the FDA would learn which treatment is best. At this point a binary approval decision would be made.

The FDA does not have the power to mandate treatment.

It can only place an upper bound on use of a drug by approving its production and marketing.

In this legal setting, I suggest empowering the FDA to grant limited-term sales licenses while clinical trials are underway.

A license would permit a firm seeking approval of a new drug to sell no more than a specified quantity over a specified time period.

The permitted quantity would change over time as knowledge of treatment response accumulates.

# TWO PROBLEMS IN MEDICAL DECISION MAKING UNDER AMBIGUITY

1. Vaccination with Partial Knowledge of External Effectiveness

2. Diagnostic Testing and Treatment under Ambiguity

# 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 become 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

$$t^* = \underset{t \in [0, 1]}{\operatorname{argmin}} \rho(1 - t)^2 + ct.$$

The optimal rate is

$$t^* = 0 \quad \text{if } 2\rho < c.$$
$$= 1 - c/(2\rho) \quad \text{if } 2\rho \geq c.$$

*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

$$t \leq r \Rightarrow p(t) \geq q,$$
$$t \geq r \Rightarrow p(t) \leq q.$$

*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

$$(1 - q)/(c - q) < t \leq r \text{ or if } t > \max(r, 1/c).$$

*The Minimax Rate*

$$
\begin{aligned}
t^m &= 0 && \text{if} \quad c > 1 \text{ and } 1 \leq q(1-r) + cr, \\
&= r && \text{if} \quad c > 1 \text{ and } 1 \geq q(1-r) + cr \\
& && \text{or if} \quad q < c < 1, \\
&= \text{all } t \in [0, 1] && \text{if } c = q \text{ and } q = 1, \\
&= \text{all } t \in [r, 1] && \text{if } c = q \text{ and } q < 1, \\
&= 1 && \text{if} \quad c < q.
\end{aligned}
$$

*Minimax-Regret Rate*

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

$$t^{mr} = (q + cr)/(q + c).$$

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

$$t^{mr} = \underset{t \in [0, 1]}{\text{argmin}}\ 1[t < r]\cdot\{\max [(1 - q)(1 - t), (1 - t) + c(t - r), (c - q)t]\}$$

$$+ 1[t \geq r]\cdot\{\max [q(1 - t), c(t - r), (c - q)t]\}.$$

*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.

# Diagnostic Testing and Treatment under Ambiguity: Using Decision Analysis to Inform Clinical Practice

*Clinical Practice Guidelines (CPGs)*

A 2011 report by the Institute of Medicine (IOM) gave this definition for CPGs:

"Clinical practice guidelines are statements that include recommendations intended to optimize patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options."

The report calls for the development of *rigorous* CPGs. Yet the standards proposed by the IOM are vague.

The IOM report brought to bear no formal decision analysis. It discussed decision analysis only briefly, stating

"A frontier of evidence-based medicine is decision analytic modeling in health care alternatives' assessment. . . . . Although the field is currently fraught with controversy, the committee acknowledges it as exciting and potentially promising, however, decided the state of the art is not ready for direct comment."

In fact, the foundations of decision analysis were largely in place more than fifty years ago.

Applications within medicine have been promoted for over thirty years by the Society for Medical Decision Making.

If the IOM report were to embrace decision analysis, it would observe that rigorous decision making requires one to specify

(i) the choice set and the objective one wants to achieve

(ii) the knowledge one has of patient health status and treatment response

(iii) the decision criterion one uses when partial knowledge makes optimization infeasible.

The report would observe that medical research speaks only to the second factor. Research may help a clinician predict treatment response, but it cannot tell the clinician what objective he should want to achieve and what decision criterion he should use when optimization is infeasible.

*Testing and Treatment Decisions*

I study a common scenario regarding diagnostic testing and treatment.

A patient presents to a clinician, who obtains initial evidence on health status.

The clinician prescribes a treatment immediately or orders a test that may yield further evidence on health status.

In the latter case, he prescribes a treatment after observation of the test result.

*Example: Aggressive Treatment with Positive Testing*

Clinicians often decide between *aggressive treatment* of an illness and *active surveillance* (aka *watchful waiting*).

Before prescribing treatment, they may order a diagnostic test.

A common practice is to choose aggressive treatment if the test result is positive and active surveillance if it is negative or if the patient is not tested.

I call this *aggressive treatment with positive testing*.

*Optimization and Decision under Ambiguity*

Given a specified objective and sufficient knowledge of response to testing and treatment, one can optimize care.

Medical decision analysis of this type appears to have originated with Phelps and Mushlin (*MDM*, 1988).

They assumed that clinicians have rational expectations and maximize expected utility.

Then the usefulness of testing is expressed by the *expected value of information*, defined by Meltzer (*JHE*, 2001) as

"the change in expected utility with the collection of information."

My concern is with settings in which the clinician not only lacks rational expectations but is unable to credibly assert a subjective distribution for response to testing and treatment.

Then the clinician faces a problem of decision making under ambiguity.

## Optimal Testing and Treatment

Let the clinician's objective be to optimize care on average across the patients in his practice.

The patient population is predetermined and is large.

Patients always comply with the clinician's decisions.

x denotes the initially observed covariates of a patient.

t denotes a treatment. $t = A$ or $t = B$.

$s = 1$ or 0 indicates whether the clinician orders the test.

r is the test result: p (positive) or n (negative).

The actions that the clinician may choose and the

knowledge of patient covariates accompanying each action may be expressed as a decision tree.

The clinician chooses s = 0 or s = 1with knowledge of x.

If he chooses s = 0, he chooses t = A or t = B with knowledge of x.

If he chooses s = 1, he chooses t = A or t = B with knowledge of (x, r).

*Feasible Testing and Treatment Allocations*

The clinician can use x to profile.

He cannot profile within the group of patients having the same value of x, but he can randomly test some fraction and not the remainder.

$\delta_S(x)$ = fraction of tested patients with covariates x.

$\delta_{T0}(x)$ = fraction of untested patients with covariates x who receive treatment B.

$\delta_{T1}(x, r)$ = fraction of tested patients with covariates (x, r) who receive B.

*The Welfare Function*

The clinician aggregates the benefits and harms of making a particular testing and treatment decision for a given patient into a scalar welfare measure y.

y(s, t) summarizes the clinician's overall assessment of the benefits and harms that would occur if he were to make testing decision s and treatment decision t.

y(s, t) may vary across patients.

Mean welfare across the population of patients is determined by the fraction of those in each covariate group that the clinician assigns to each testing-treatment option.

Let P(x) denote the fraction of patients with covariate x.

Let f(r|x) denote the fraction of patients with covariates x who would have test result r if they were to be tested.

Let E[y(s, t)|x] be the mean welfare if all patients with covariates x were to receive (s, t).

Let E[y(s, t)|x, r] be the mean welfare if all patients with covariates x and test result r were to receive (s, t).

Let $\delta = [\delta_S(x), \delta_{T0}(x), \delta_{T1}(x, r), x \in X, r \in \{p, n\}]$ denote any testing-treatment allocation.

The mean welfare resulting from allocation $\delta$ is

$$
\begin{aligned}
W(\delta) \;=\; & \sum_{x \in X} P(x)\big[[1 - \delta_S(x)][1 - \delta_{T0}(x)]E[y(0, A)|x] \\
& + [1 - \delta_S(x)]\delta_{T0}(x)E[y(0, B)|x] \\
& + \sum_{r \in \{p, n\}} f(r|x)\{\delta_S(x)[1 - \delta_{T1}(x, r)]E[y(1, A)|x, r] \\
& + \delta_S(x)\delta_{T1}(x, r)E[y(1, B)|x, r]\}\big].
\end{aligned}
$$

An optimal allocation maximizes $W(\delta)$. An optimum is

$$\delta_S(x) = 1 \text{ if } \sum_{r \in \{p, n\}} f(r|x)[\max\{E[y(1, A)|x, r], E[y(1, B)|x, r]\}] \geq \max\{E[y(0, A)|x], E[y(0, B)|x]\},$$

$$= 0 \text{ otherwise.}$$

$$\delta_{T0}(x) = 1 \text{ if } E[y(0, B)|x] \geq E[y(0, A)|x],$$

$$= 0 \text{ otherwise.}$$

$$\delta_{T1}(x, p) = 1 \text{ if } E[y(1, B)|x, p] \geq E[y(1, A)|x, p],$$

$$= 0 \text{ otherwise.}$$

$$\delta_{T1}(x, n) = 1 \text{ if } E[y(1, B)|x, n] \geq E[y(1, A)|x, n],$$

$$= 0 \text{ otherwise.}$$

# Identification of Response to Testing and Treatment when ATPT is Standard Practice

Determination of the optimum requires sufficient knowledge of $f(r|x)$, $E[y(0, t)|x]$, and $E[y(1, t)|x, r]$ for $t \in \{A, B\}$, $r \in \{n, p\}$, and $x \in X$.

In principle, one might obtain this knowledge by performing a randomized trial.

An ideal trial with four arms, one for each value of (s, t), would reveal the distribution of test results and treatment response.

However, an ideal trial often is infeasible. Partial knowledge is common in practice.

I analyze identification of response to testing and treatment when one observes a study population where ATPT was standard clinical practice.

The identification problem arises from the unobservability of counterfactual testing and treatment outcomes.

To focus attention on this core difficulty, I abstract from others that may arise in practice.

One observes the entire study population rather than a sample.

The study population has the same composition as the population to be treated.

Outcomes are bounded on [0, 1].

*Basic Analysis with no Distributional Assumptions*

To optimize care, the clinician wants to learn {E[y(0, t)|x], E[y(1, t)|x, r], f(r|x)} for $t \in \{A, B\}$, $r \in \{n, p\}$, $x \in X$.

Given the ATPT practice, the evidence reveals nothing about E[y(0, B)|x], E[y(1, B)|x, n], and E[y(1, A)|x, p].

The evidence partially identify E[y(0, A)|x], E[y(1, A)|x, n], E[y(1, B)|x, p], and f(r|x).

Let z = 1 if a person in the study population was tested and z = 0 if he was not.

Let g(n, x) = f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x)

g(p, x) = f(r = p|x, z = 1)P(z = 1|x) + P(z = 0|x).

The results are these bounds:

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

$$E[y(1, A)|x, n, z = 1]f(r = n|x, z = 1)P(z = 1|x)/g(n, x) \leq E[y(1, A)|x, n] \leq \{E[y(1, A)|x, n, z = 1]f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x)\}/g(n, x).$$

$$E[y(1, B)|x, p, z = 1]f(r = p|x, z = 1)P(z = 1|x)/g(p, x) \leq E[y(1, B)|x, p] \leq \{E[y(1, B)|x, p, z = 1]f(r = p|x, z = 1)P(z = 1|x) + P(z = 0|x)\}/g(p, x).$$

$$f(r = n|x, z = 1)P(z = 1|x) \leq f(r = n|x) \leq f(r = n|x, z = 1)P(z = 1|x) + P(z = 0|x).$$

*Random Testing*

Each of the five quantities that is partially identified in the basic analysis becomes point identified if testing is random conditional on x.

Random testing may occur through performance of a randomized trial of testing, which is not prohibited by the ATPT practice.

Or it may occur without an explicit trial if clinicians caring for the study population make testing decisions that are statistically independent of test results and of response to testing and treatment.

Random testing implies these equalities:

$E[y(0, A)|x] = E[y(0, A)|x, z = 0]$.

$E[y(1, A)|x, n] = E[y(1, A)|x, n, z = 1]$.

$E[y(1, B)|x, p] = E[y(1, B)|x, p, z = 1]$.

$f(r = n|x) = f(r = n|x, z = 1)$.

The assumption of random testing does not help to identify $E[y(0, B)|x]$, $E[y(1, B)|x, n]$, and $E[y(1, A)|x, p]$.

*Test Result as a Monotone Instrumental Variable*

Patients with negative results on a diagnostic test are often thought to be healthier than ones with positive results.

Hence, a clinician may find it credible to predict that patients with negative test results have better future prospects, on average, than do patients with positive results.

Formally, the clinician may find it credible to assume

$$E[y(s, t)|x, n] \geq E[y(s, t)|x, p]$$

for specified values of (s, t).

Then test result is a *monotone instrumental variable* (MIV). See Manski and Pepper (*ECMA*, 2000).

The MIV assumption for (s, t) = (1, A) implies that E[y(1, A)|x, p] is no larger than the basic upper bound on E[y(1, A)|x, n].

If one also assumes random testing, then E[y(1, A)|x, p] is no larger than the known value of E[y(1, A)|x, n].

The MIV assumption for (s, t) = (1, B) implies that E[y(1, B)|x, n] is no smaller than the basic lower bound on E[y(1, B)|x, p].

If one also assumes random testing, then E[y(1, B)|x, n] is no smaller than the known value of E[y(1, B)|x, p].

*Monotone Response to Testing*

A clinician may believe that testing cannot directly improve welfare but may decrease it.

For example, he may think that testing has no therapeutic effect but may be invasive or costly.

Formally, the clinician may find it credible to assume the inequality

$$y(0, t) \geq y(1, t)$$

for specified values of t and for all patients. This is a *monotone-treatment-response* (MTR) assumption. See Manski (*ECMA*, 1997).

The MTR assumption for t = B yields a lower bound on E[y(0, B)|x], the bound depending on what other assumptions are imposed.

A simple finding emerges if one combines the MTR assumption with the MIV assumption for (s, t) = (1, B).

Then E[y(0, B)|x] is no smaller than the basic lower bound on E[y(1, B)|x, p].

If one also assumes random testing, then E[y(0, B)|x] is no smaller than the known value of E[y(1, B)|x, p].

# Patient Care under Ambiguity

Suppose that a clinician wants to choose a testing-treatment allocation that maximizes welfare.

He has only partial knowledge of the welfare function.

The clinician should not choose a dominated allocation.

How should he choose among undominated allocations?

The question has no unambiguously correct answer.

Hence, the IOM report should not have supposed that CPGs can "optimize patient care."

A clinician might maximize expected welfare.

He might use the maximin or minimax-regret criterion.

These criteria yield different prescriptions for decision making.

Each may be described as "reasonable," but none as "optimal."

# Developing Clinical Guidelines under Ambiguity

I think it important to separate two tasks for CPGs.

One is to characterize medical knowledge.

The other is to make recommendations for patient care.

Characterization of available medical knowledge has substantial potential to improve clinical practice.

It should draw on all available evidence, experimental and observational.

It should maintain assumptions that are sufficiently credible to be taken seriously.

It should combine the evidence and assumptions to draw logically valid conclusions.

I am skeptical whether CPGs should make recommendations for patient care.

Making recommendations for patient care asks the developers of CPGs to aggregate the benefits and harms of care into a scalar measure of welfare.

It requires them to specify a decision criterion to cope with partial knowledge.

These activities might be uncontroversial if there were consensus about how welfare should be measured and what decision criterion should be used.

However, care recommendations may be contentious if perspectives vary across clinicians, patients, and other relevant parties.

An alternative to having CPGs make care recommendations would be to bring specialists in decision analysis into the clinical team.

Modern clinical practice often has a group of professionals jointly contribute to patient care.

However, existing patient-care teams do not ordinarily draw on professionals having specific expertise in the framing and analysis of complex decision problems.

It may be that adding such professionals to clinical teams would be more beneficial to patient care than asking physicians to adhere to care recommendations made by distant organizations.

# POLICY ANALYSIS AND DECISIONS

I have described the practice of policy analysis and the inferential problems that researchers confront.

I have argued that credible analysis typically yields interval rather than point predictions of policy outcomes.

I have examined how a policy maker might reasonably choose policy with partial knowledge.

I conclude with some thoughts that connect policy analysis and decisions.

*Institutional Separation of Analysis and Decisions*

Modern societies have created an institutional separation between policy analysis and decision, with analysts reporting findings to policy makers.

Separation of analysis and decision making, the former aiming to inform the latter, appears advantageous from the perspective of division of labor.

However, the current practice of policy analysis does not serve policy makers well.

The consumers of policy analysis cannot trust the producers.

Everyone concerned with policy making should keep in mind several dangers of incredible certitude.

1. Policy making with incredible certitude seeks to maximize the social welfare that would prevail if untenable assumptions were to hold, not *actual* welfare.

2. Incredible certitude inhibits performance of new research aiming to learn about treatment response.

3. Incredible certitude does not recognize the value of treatment diversification as a means to cope with uncertainty and learn.

*Doing Better*

Rather than require the consumers of policy analysis to guess how to interpret incredible point predictions, researchers could provide credible interval predictions.

Some think this idea impractical. Policy makers may be psychologically unwilling or cognitively unable to cope with uncertainty.

I do not know for sure that analysis providing credible interval predictions will yield better policy decisions than prediction with incredible certitude.

To claim this would subject me to a charge of incredible certitude, which I certainly want to avoid.

What I do suggest is application of the lessons of this book to policy analysis itself.

Point prediction of policy outcomes is a status quo treatment.

Provision of credible interval predictions is an innovation.

An outcome of interest is the quality of policy decisions.

Society has little knowledge of the relative merits of the status quo and the innovation.

I suggest adaptive diversification.