FOUNDATIONS OF ECONOMIC SURVEY RESEARCH

Lecture I. Sampling Theory
Lecture II. Survey Design and Response Models

Daniel McFadden
Econometrics Laboratory
University of California, Berkeley

Gorman Lectures, 2004

Lecture II.

1. Subject behavior in economic surveys
2. Response errors
3. Example: Kahneman-Tversky
4. Example: Kahneman-McFadden
5. Example: Hurd-McFadden
5. Models for detection and correction
6. Experiments in surveys to identify and correct response errors

Surveys

- Surveys are “structured conversations between strangers”, subject to most of the communication problems that arise in ordinary conversations
  - Inattention
  - Misunderstanding
  - Strategic motives
  - Posturing and projection
- Cognitive tasks are required that may be misinterpreted or processed incorrectly
- Retrieval of Memories and Facts may be incomplete and inaccurate
  - Analogy to test-taking

Survey Response Process

- Comprehension
  - Attend to question, instructions, identify focus, translate concepts and logic
- Retrieval
  - Plan retrieval process, retrieve generic and specific memories, reconstruct details
- Judgment
  - Evaluate reconstructed memories, draw inferences, Integrate retrieved material, make inferences, estimates
- Response
  - Map estimate to response category, edit response

  - From Tourangeau *et al* The Psychology of Survey Response

Comprehension

- Attend to question, instructions, identify focus, translate concepts and logic
  - Attention to instruction and the terms and qualifications in the question
  - Inattention, misunderstanding, misinterpretation
  - Identifying question focus
  - Translating concepts and logic into personal system

Example: “How much have you spent on food away from home in the past six months?”

Parsing the question –

Restaurants only? Fast food? Snacks? Drinks? Food/entertainment packages? Inclusive holidays? Purchases for others? Take-out food consumed at home? Groceries?

Significant event or date to demark six months?

Why are they asking?

- To see if I am a fast food junkie?
- To see if I am over-indulgent?
- To see if I am normal?
- To see if I have a full life?

Retrieval

- Recall relevant information from long-term memory
- Retrieval plan: Concentrate on events, budget, typical day or week or whole period? Top down or bottom up?
- Retrieve specific bits – distinctive events, remembered quantities and prices, or total outlays
- Reconstruct details
- Influenced by conceptual match with memory organization, question focus and cues

Food away from home example continued

- Recall of specific food purchase events
- Reconstruction of typical purchase patterns
- Recall of benchmarks – total income over the period, typical total food expenditures per day.

Judgment

- The processes respondents use to integrate retrieved information
  - Judge completeness and accuracy of retrieved memories
  - Inferences based on process of retrieval
  - Inferences to fill in gaps
- Date, duration, frequency judgments
  - Telescoping
  - Duration neglect
- Overall estimate
- Adjustment for retrieval omissions

Food away from home example continued

- Recall significant events and estimate their costs, reconstructing memories of such events as necessary, then estimate the cumulative contribution of insignificant events
- Compare for reasonableness with total income, specific event memories

Reporting

- Map answer onto appropriate scale
- Understanding and interpetation of scale categories; e.g., interpretation of “seldom” or “often”
- Classification of answer
- Editing of response for acceptability, consistency
- Give truthful answer, a misrepresention, an evasive or non-informative one, or a non-response?

Response Errors

- Misreporting of economic facts can arise from each stage of the response process
- Survey design can influence errors, perhaps differentially at various stages of the response process
- Food in restaurants last week may be answered by enumeration, and may be reported more accurately than food away from home in the past six months
- Known cognitive effects can be influenced by survey design

Cognitive Anomalies	
Retrieval	
Availability	Memory reconstruction is tilted toward most available information
Primacy/Recency	Initial and final events are the most available
Regression	Attribution of causal structure to observations, failure to anticipate regression to the mean
Representativeness	Frequency neglect in exemplars
Saliency	Dimensions judged most salient are over-emphasized

Cognitive Anomalies	
Judgment	
Anchoring	Numerical cues in questions are most available
Context	Environment of task influences how it is interpreted, what is salient
Framing,Reference Point, Status Quo	Form influences saliency, “The devil you know …”
Superstition	Non-consequentialist reasoning
Temporal	Telescoping, duration neglect

Cognitive Anomalies	
Reporting	
Focal	Artificial or “rounded-off” response
Projection	Response edited fo enhance image
Strategic	Deliberate misrepresentation for strategic purposes

Example: Kahneman-Tversky

Experiment 1 (N = 152)	Choice	Experiment 2 (N – 155)	Choice
A: 200 people saved	72%	C: 400 people die	22%
B: 600 saved with probability 1/3, 0 saved with probability 2/3	28%	D: 0 die with probability 1/3 600 die with probability 2/3	78%

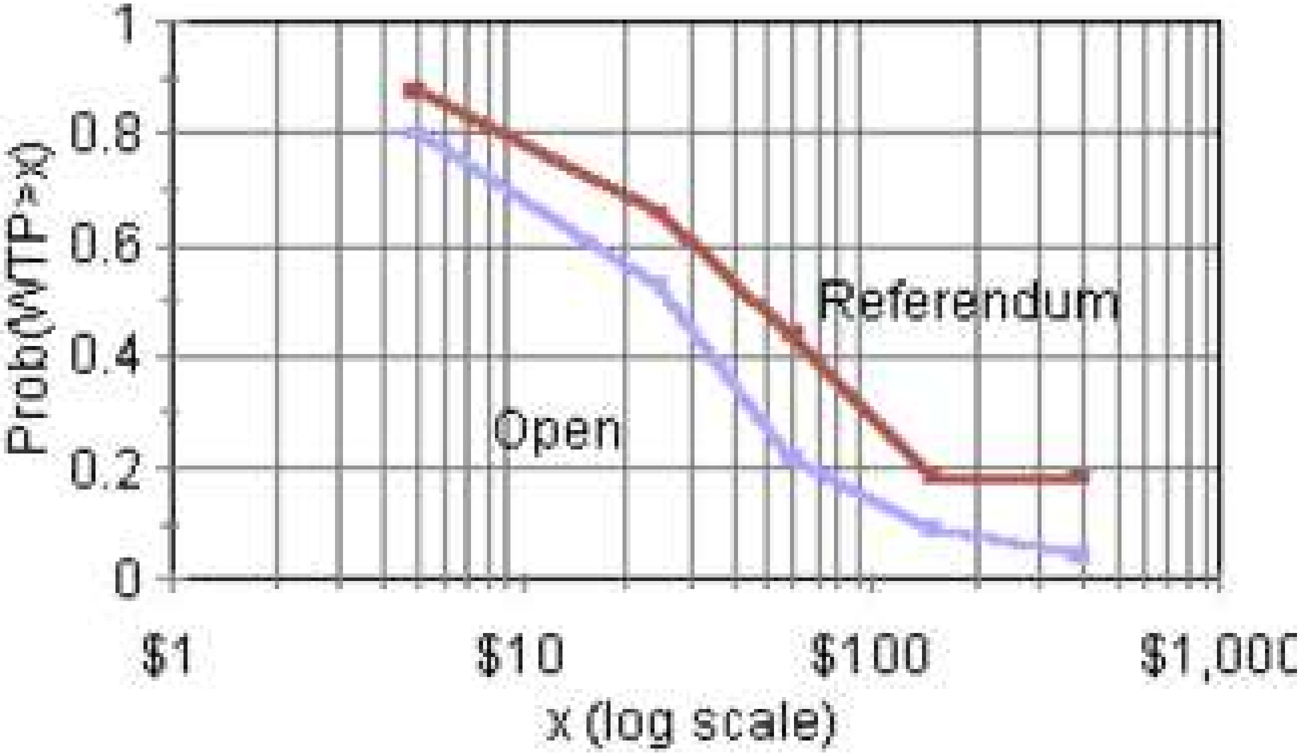
Anchoring in economic questions

- A bracket question (e.g., “Did you spend more than $800 in the past six months for food away from home?”) induces a response that is pulled toward the numerical cue, more so when the quantity is not easily retrieved from memory
- Example (Kahneman-McFadden) Willingness to pay for seabirds

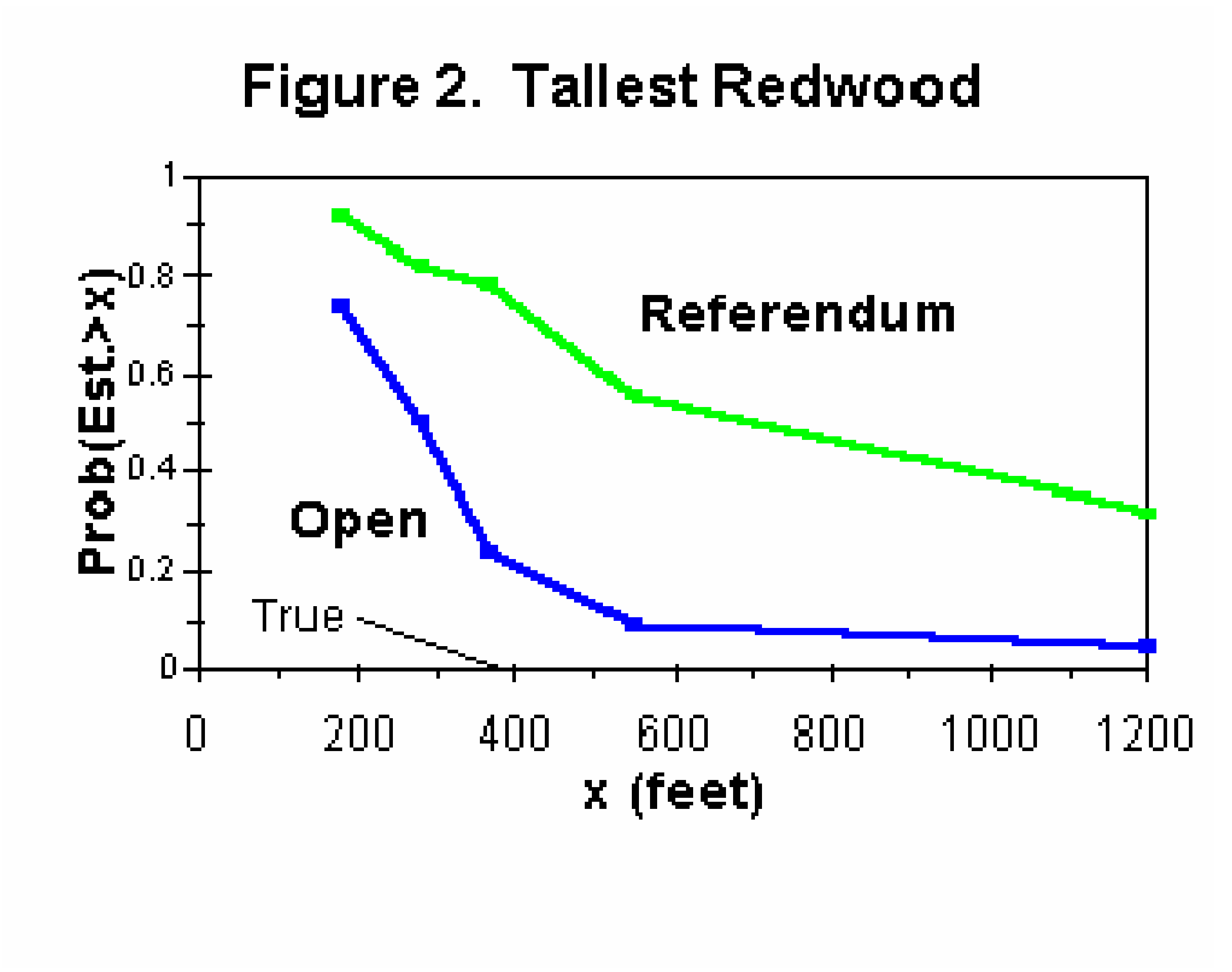
- "There is a population of several million seabirds living off the Pacific coast, from San Diego to Seattle. The birds spend most of their time many miles away from shore and few people see them. It is estimated that small oil spills kill more than 50,000 seabirds per year, far from shore. Scientists have discussed methods to prevent seabird deaths from oil, but the solutions are expensive and extra funds will be required to implement them. It is usually not possible to identify the tankers than cause small spills and to force the companies to pay. Until this situation changes, public money would have to be spent each year to save the birds. We are interested in the value your household would place on saving about 50,000 seabirds each year from the effects of offshore oil spills.

- *Non-Decisive, Decoupled Payment Vehicle:*
- "We want to know if you support an operation that would be sure to save 50,000 seabirds each year, and would be paid for with extra federal or state taxes. The extra taxes to your household if the operation takes place would be your household's share of the actual cost, and would not depend your answer on this survey. The operation will stop when ways are found to prevent oil spills, or to identify the tankers that cause them and make their owners pay for the operation.
- *Open-Ended Elicitation:*
- "What is the MOST you would be willing to pay in extra federal or state taxes per year at which you would vote for this operation?
  $__________ per year.
- *Referendum Elicitation (with Open-Ended Followup):*
- "Would you vote for this operation if it cost your household $___ per year in extra federal or state taxes? Yes ___ No ___. What is the MOST you would be willing to pay per year at which you would vote for this operation? $__________ per year.

Willingness to Pay for Seabirds



Recall/reconstruction of a fact



Hurd-McFadden AHEAD study

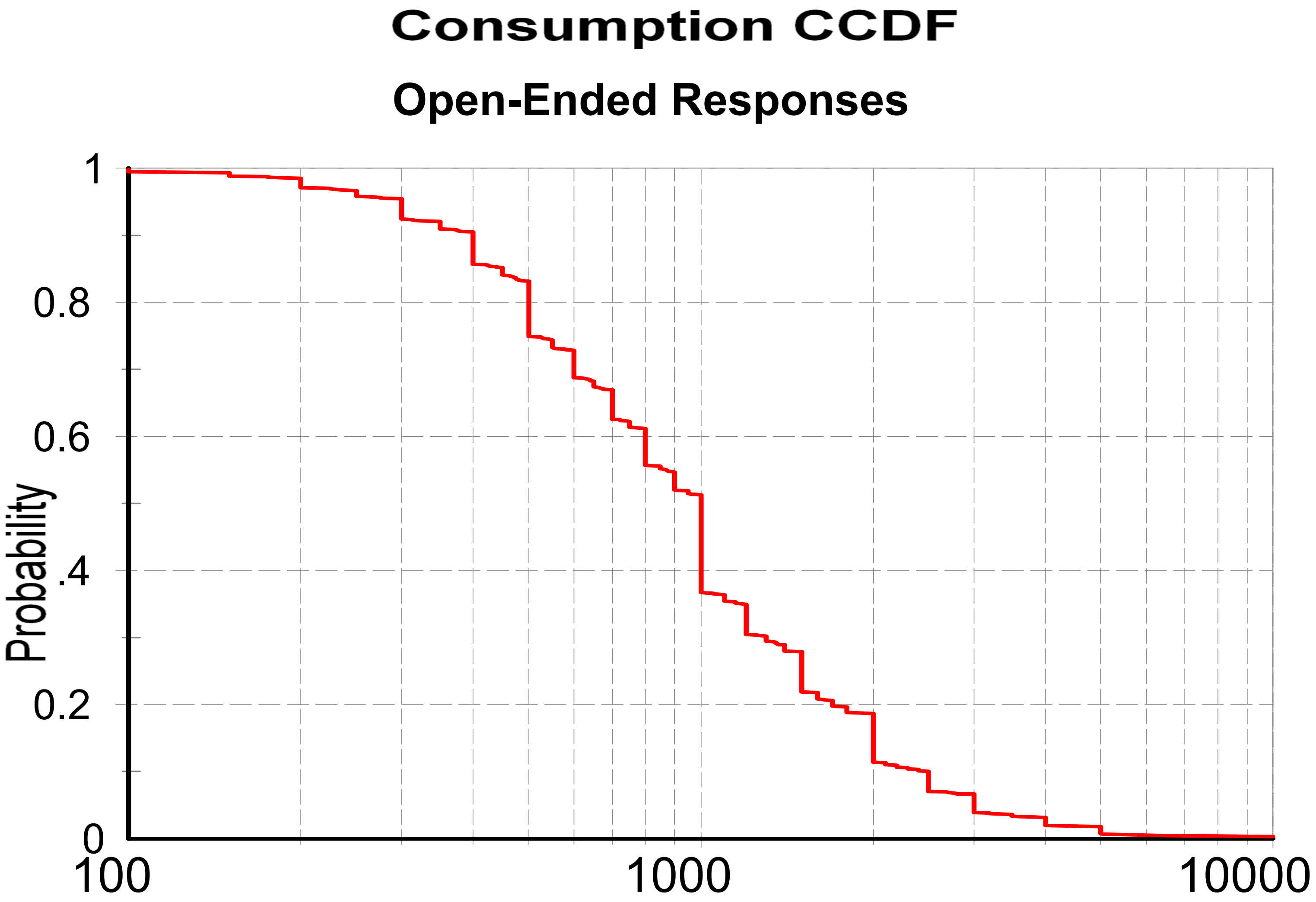
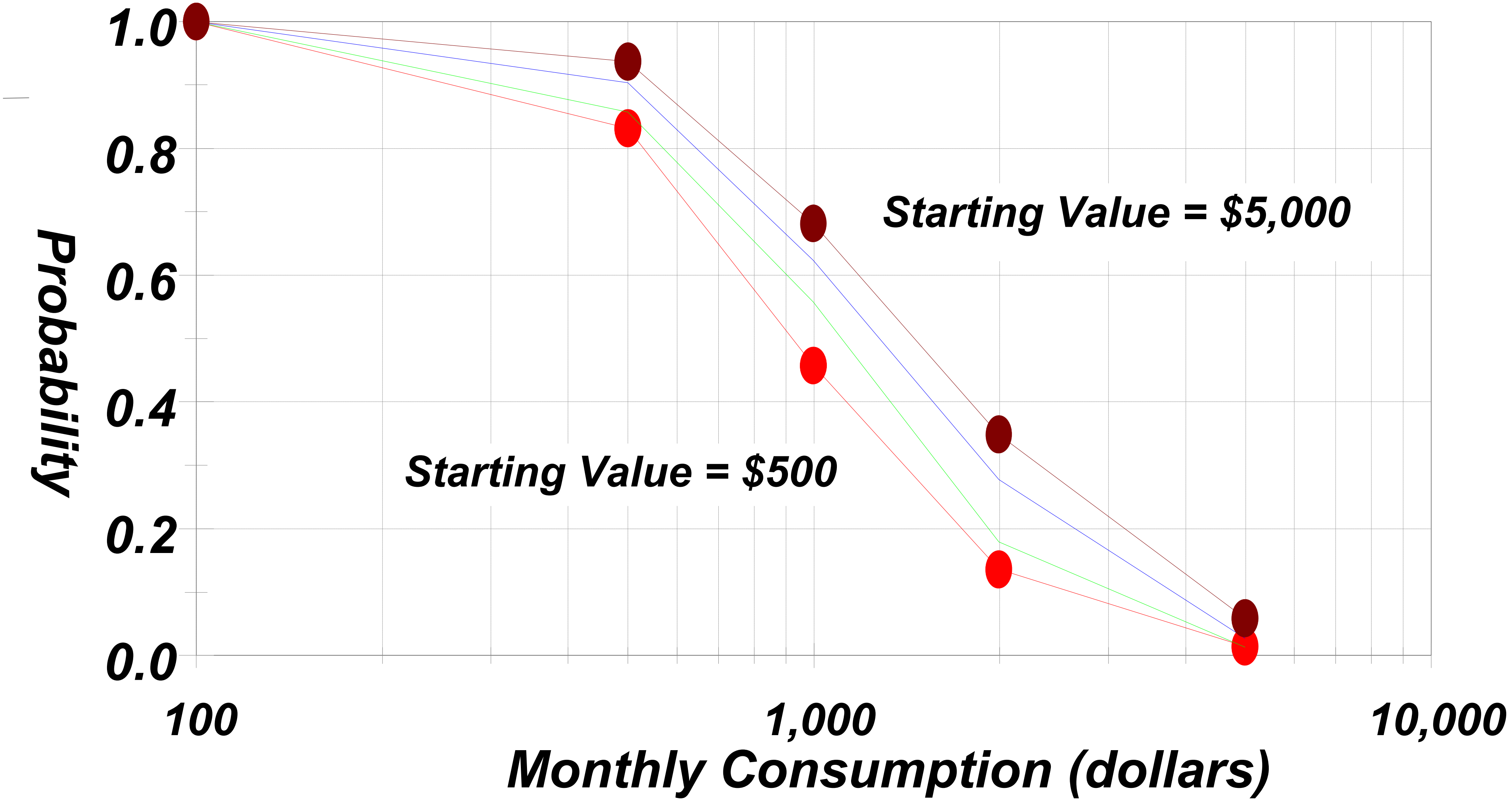


Figure 3. Consumption CCDF

By Starting Value, Complete Bracket Responses



Detection, control, and compensation for response errors

- **stand-alone or in-stream experimental treatments**
  - **Example: ask for health conditions using different question treatments**
  - **Audits and validation procedures**

Examples of Variable Types

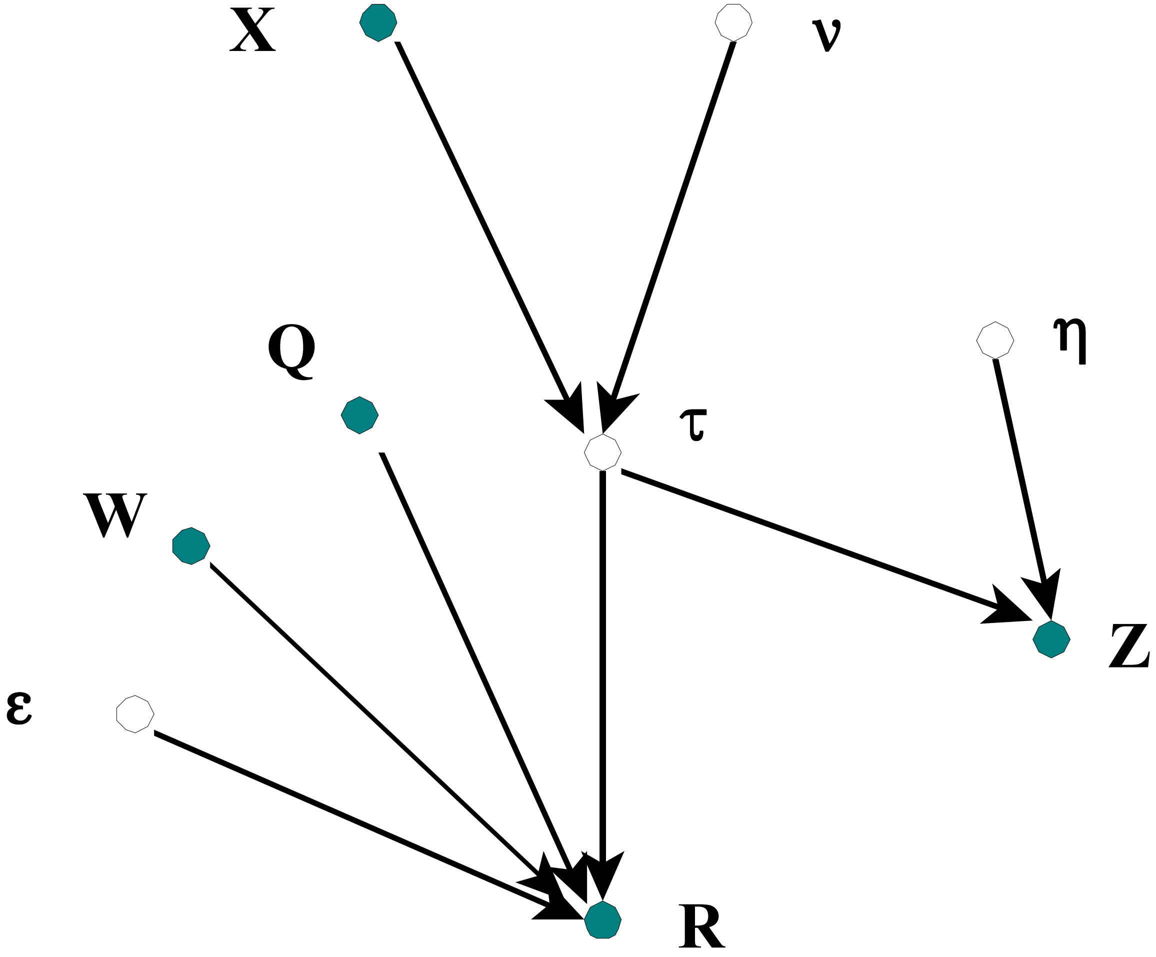
- **objective, verifiable**
  - **last month’s phone bill**
  - **individual audits, population distribution**
- **subjective but externally scalable**
  - **subjective mortality hazard rate**
  - **distribution from external life tables or observed mortality experience**
- **self-rated health status on a five-point scale**
  - **identification through axiomatic restrictions and/or indirect indicators**
- **health limitations**
  - **vignette anchoring**

VARIABLES

- **X observed exogenous variables (“multiple causes” such as family size, age) that influence latent true variable**
- **W observed exogenous variables that directly influence observed response (e.g., time delay influencing memory, measure of cognitive ability)**
- **Z observed indicators for latent true variable (e.g., self-reported reliability bounds, look-up value)**
- **Q question context/format (e.g., location of range card brackets, content of question instructions, question order), the treatment variable**
- **τ latent true variable (e.g., true phone bill)**
- **η,ν,ε unobserved disturbances**

- **The exogenous variables (X,Q,W) have a covariance matrix of full rank.**

DAG for causal paths



EQUATIONS

- **$\tau = t(X,v)$ multiple cause equation for true response**
- **$Z = h(\tau,\eta)$ (multiple) indicators for true response**
- **$R = m(\tau,Q,W,\varepsilon)$ model for the determination of observed response**

Objectives

1. **Recover (or bound) the conditional distribution F of τ given X, and/or its generalized conditional moments.**
2. **Recover (or bound) the function m(τ,Q,W,ε)**
3. **Test the hypothesis that m(τ,Q,W, ε) = τ for some question treatment Q_0. (This may be a maintained assumption for identification in some cases.)**
4. **Test the hypothesis that m(τ,Q_1,W, ε) = m(τ,Q_2,W, ε); i.e., that two question formats are equivalent.**
5. **Predict τ for an individual, given X,Z,Q,W,R, or alternately given X,Q,W,R**

Linear MIMC version

The dimensions of variables are 1×1 for τ and R, 1×k for X, 1×m for Z, 1×n for W, 1×q for Q.

$$\tau = t(X,v) = \kappa + X\alpha + \sigma v \sim N(\kappa + X\alpha,\sigma^2)$$

$$Z = h(\tau,\eta) = \gamma + \tau\delta + \eta K \sim N(\gamma + \tau\delta,K'K)$$

$$R = m(\tau,Q,W,\varepsilon) = \theta + \tau\beta + Q\lambda + W\pi + \rho\ \varepsilon$$
$$\sim N(\theta + \tau\beta + Q\lambda + W\tau,\rho^2)$$

The parameters of this model are κ, α (k×1), σ, γ (1×m), δ (1×m), K (m×m, upper triangular), θ, β, λ (q×1), π (n×1), ρ.

- Eliminating τ using the first equation, the observed dependent variables Z and R satisfy

- $Z = \gamma + \delta\kappa + X\alpha\delta + \eta K + \delta\sigma\nu$
- $R = \theta + \beta\kappa + X\alpha\beta + Q\lambda + W\pi + \beta\sigma\nu + \rho\varepsilon.$

- Then, they are distributed with conditional moments

- $E(Z|\ X,Q,W) = \gamma + \delta\kappa + X\alpha\delta$
- $E(R|X,Q,W) = \theta + \beta\kappa + X\alpha\beta + Q\lambda + W\pi$
- $Var(Z|X,Q,W) = \kappa N\kappa+\sigma^2\delta'\delta$
- $Var(R|X,Q,W) = \rho^2+\theta^2\sigma^2$
- $Cov(Z,R|X,Q,W) = \delta\alpha\beta$

Identification and estimation of this system

- Regressions of Z on constants and X, and of R on a constant, X, Q, and W return consistent asymptotically normal estimates of $\gamma + \delta\kappa$, $\alpha\delta$, $\theta + \beta\kappa$, $\alpha\beta$, λ, π and the conditional covariance matrix.
- These estimates are sufficient for some purposes, such as testing whether question format/context influences response (i.e., H_0: $\lambda = 0$) and adjusting responses to homogenize question effect (i.e., $R_{adj} = R + (Q_0\text{-}Q)\lambda$ produces an adjusted response that would be produced by common question format Q_0).

Order condition for identification

- **Counting empirical moments and parameters, the conditional mean for Z determines m(k+1) quantities, the conditional mean of R determines 1+k+n+q quantities, and the covariance matrix determines (m+1)(m+2)/2 quantities. The system contains 5+k+n+q+m(m+5)/2 parameters. Then 3-mk normalizing restrictions are needed to identify the structural parameters from the first and second moments.**
- **Restrictions Needed for Identification**

m\k	**0**	**1**	**2**	**3+**
1	**3**	**2**	**1**	**0**
2	**3**	**1**	**0**	**0**
3+	**3**	**0**	**0**	**0**

Rank conditions

- **Without normalizations, the location κ and scale σ of τ are arbitrary, in the sense that γ, δ, θ, β can be adjusted commensurately to yield observationally equivalent equations for Z and R. Then, two normalizations on these 2(m+1) parameters are needed to fix τ. If k > 0, these two normalizations meet the necessary order conditions. However, if k = 0, so there are no observed causes of τ, an additional normalization is needed.**
- **The most common method of normalizing the location and scale of τ would be through an assumption that one component of Z is an unbiased estimate of τ; e.g., $\gamma_1 = 0$ and $\delta_1 = 1$, so that $E(Z_1 - \tau|\tau) = 0$. This is reasonable if Z_1 is an audited or look-up value for the latent variable, or has external validity for determining the location and scale of τ. These normalizations allow κ and α to be estimated consistently from the regression of Z_1 on X.**

- If k > 0, the parameters γ_i and δ_i for i = 2,...,m are estimated consistently from the regression of Z_i on a constant and the composite variable $\kappa + X\alpha$, and θ, β, λ, π are estimated consistently from the regression of R on a constant, the composite variable $\kappa + X\alpha$, Q, and W. This establishes identification, and also gives a consistent estimation method.
- If k = 0, then the parameters α are absent, the parameters γ_j and δ_i are not identified from the regression of Z_i on a constant, and the parameters θ and β are not identified from the regression of R on a constant, Q, and W. An additional normalizing assumption, such as $\beta = 1$, is needed to identify θ, and m - 1 normalizing assumptions are needed to identify the γ_i and δ_i. In many cases, these normalizations will have no good external justification. Thus, k > 0 is very helpful for identification. Note that the presence of W, even if it contains variables distinct from X, does not aid identification.

Estimating or bounding τ

Best linear unbiased predictors when Z is not observed,

$$\tau^e = (R - \theta - Q\lambda - W\pi)/\beta$$

with $E\tau^e = \tau$ **and** $E(\tau^e - \tau)^2 = \rho^2/\beta^2$

If Z is observed,

$$\tau^e = [(Z-\gamma)(K'K)^{-1}\delta'\rho^2$$
$$+ (R - \theta - Q\lambda - W\pi)\beta]/[\rho^2\delta(K'K)^{-1}\delta' + \beta^2],$$

with $E\tau^e = \tau$ **and** $E(\tau^e - \tau)^2 = \rho^2/[\rho^2\delta(K' K)^{-1}\delta' + \beta^2]$**.**

A NON-MIMC FORMULATION USING QUANTILE METHODS

Suppose the question treatments are indexed by Q = 0,...,q. Suppose Q = 0 denotes a "neutral" or "gold standard" treatment. Assume that $m(\tau,0,W,\varepsilon) \equiv \tau$. This assumption might be justified because this particular format is known to be exact, or because it is taken as the definition of τ. Consider the simple case where ε does not enter m. Assume that m is increasing in τ. Let $F_Q(R|W)$ be the conditional distribution of R given W and Q, and note that $F_0(R|W) = F_0(R)$. Then $F_0^{-1}(F_Q(R|W))$ recovers the value of τ associated with each R and question treatment. This is an elementary version of the use of quantile methods developed by Matzkin (1999). Conditional quantiles estimated using kernel methods will work, as might some "nearest neighbor" methods.

Conclusion: Experiments in Surveys to Detect and Correct Response Error

- Using the linear parametric MIMC model, or nonlinear, nonparametric generalizations as a template, identify data structures sufficient for identification
- Design experiments in surveys to provide the necessary data structures and variation
- Use the combined data and analysis to provide consistent estimates of population conditional distribution, and in some cases best predictions of unconfounded individual response
- Example: Hurd-McFadden analysis of models for correction of anchoring effects for consumption and savings in the AHEAD panel.
- Example: McFadden-Winter-Schwarz experiment in the Retirement Perspectives Survey of AARP members on order and range effects on reported purchase of nursing home insurance