
Patrick Kline    Andres Santos
UC Berkeley      UC San Diego

September 21, 2012
The Problem

Missing data is ubiquitous in modern economic research

- Roughly one quarter of earnings observations in CPS and Census.
- Problem can be worse in proprietary surveys and experiments.
The Problem

**Missing data** is ubiquitous in modern economic research

- Roughly one quarter of earnings observations in CPS and Census.
- Problem can be worse in proprietary surveys and experiments.

Equally ubiquitous solutions: **Missing at Random (MAR)**

- Justification for imputation procedures in CPS and Census.
- And for ignoring missingness altogether...

Patrick Kline UC Berkeley
The Problem

**Missing data** is ubiquitous in modern economic research

- Roughly one quarter of earnings observations in CPS and Census.
- Problem can be worse in proprietary surveys and experiments.

Equally ubiquitous solutions: **Missing at Random (MAR)**

- Justification for imputation procedures in CPS and Census.
- And for ignoring missingness altogether...

**Question:** How can we evaluate sensitivity of conclusions to MAR?

- Want to consider plausible deviations from MAR without presuming much about selection mechanism.
- And to enable study of sensitivity at different points in conditional distribution (tails likely more sensitive).
Consider a triplet \((Y, X, D)\) with \(Y \in \mathbb{R}, X \in \mathbb{R}^l, D \in \{0, 1\}\).

\[
Y = X'\beta(\tau) + \epsilon \quad P(\epsilon \leq 0|X) = \tau
\]

and \(D = 1\) if \(Y\) is observable, and \(D = 0\) if \(Y\) is missing.
The Model

Consider a triplet \((Y, X, D)\) with \(Y \in \mathbb{R}, X \in \mathbb{R}^l, D \in \{0, 1\}\).

\[
Y = X' \beta(\tau) + \epsilon \quad P(\epsilon \leq 0|X) = \tau
\]

and \(D = 1\) if \(Y\) is observable, and \(D = 0\) if \(Y\) is missing.

Without Missing Data

- Quantile regression as a summary of conditional distribution.
- Under misspecification \(\Rightarrow\) best approximation to true model.
The Model

Consider a triplet \((Y, X, D)\) with \(Y \in \mathbb{R},\ X \in \mathbb{R}^l,\ D \in \{0, 1\}\).

\[
Y = X' \beta(\tau) + \epsilon \quad P(\epsilon \leq 0|X) = \tau
\]

and \(D = 1\) if \(Y\) is observable, and \(D = 0\) if \(Y\) is missing.

Without Missing Data

- Quantile regression as a summary of conditional distribution.
- Under misspecification \(\Rightarrow\) best approximation to true model.

Without MAR

- Point identification fails.
- Need to consider misspecification and partial identification.
Define the conditional distribution functions,

\[ F_{y|x}(c) \equiv P(Y \leq c | X = x) \]  
\[ F_{y|d,x}(c) \equiv P(Y \leq c | D = d, X = x) \]
Define the conditional distribution functions,

\[ F_{y|x}(c) \equiv P(Y \leq c | X = x) \quad F_{y|d,x}(c) \equiv P(Y \leq c | D = d, X = x) \]

**Missing at Random**

- Rubin (1974). Assume that \( F_{y|x}(c) = F_{y|1,x}(c) \) for all \( c \in \mathbb{R} \).
Define the conditional distribution functions,

\[ F_{y|x}(c) \equiv P(Y \leq c | X = x) \quad \text{and} \quad F_{y|d,x}(c) \equiv P(Y \leq c | D = d, X = x) \]

**Missing at Random**

- Rubin (1974). Assume that \( F_{y|x}(c) = F_{y|1,x}(c) \) for all \( c \in \mathbb{R} \).

**Nonparametric Bounds**

- Manski (1994). Exploit that \( 0 \leq F_{y|0,x}(c) \leq 1 \) for all \( c \in \mathbb{R} \).
- Often bounds are uninformative.
- And typically overly conservative.

Between these extremes lie a continuum of selection mechanisms...
Deviation from MAR

Characterize selection as distance between $F_{y|1,x}$ and $F_{y|0,x}$:

$$d(F_{y|1,x}, F_{y|0,x}) \leq k$$

- A way of nonparametrically indexing set of selection mechanisms:
  - Missing at random corresponds to imposing $k = 0$.
  - Manski Bounds corresponds to imposing $k = \infty$.

- Allows study of sensitivity to deviations from MAR (e.g. what level of $k$ is necessary to overturn conclusions regarding $\beta(\tau)$?)

- And in some cases $k$ may be estimated using validation data.
Nominal Identified Set

- Find possible quantiles under restriction $d(F_y|1,x, F_y|0,x) \leq k$.
- Bound $\beta(\tau)$ as a function of $\tau, k$ allowing for misspecification.
General Outline

Nominal Identified Set

- Find possible quantiles under restriction $d(F_y|1,x, F_y|0,x) \leq k$.
- Bound $\beta(\tau)$ as a function of $\tau, k$ allowing for misspecification.

Inference

- Obtain distribution of estimates of boundary of nominal identified set.
- Exploit distribution as a function of $(\tau, k)$ for sensitivity analysis.
General Outline

Nominal Identified Set

• Find possible quantiles under restriction $d(F_y|1,x, F_y|0,x) \leq k$.
• Bound $\beta(\tau)$ as a function of $\tau, k$ allowing for misspecification.

Inference

• Obtain distribution of estimates of boundary of nominal identified set.
• Exploit distribution as a function of $(\tau, k)$ for sensitivity analysis.

Changes in Wage Structure

• Examine changes in wage structure across Decennial Censuses.
• Measure departures from MAR in matched CPS-SSA.
Literature Review

Missing Data:

Sensitivity Analysis:
Altonji, Elder, and Taber (2005); Rosenbaum and Rubin (1983); Rosenbaum (1987, 2002).

Misspecification

Misspecification and Partial Identification
1 Nominal Identified Set

2 Parametric Approximation

3 Inference

4 Changes in Wage Structure

5 CPS-SSA Analysis
Bounds on Conditional Quantiles

Define true conditional quantile \( q(\tau|x) \) and non-missing probability \( p(x) \):

\[
F_{y|x}(q(\tau|x)) = \tau \quad p(x) \equiv P(D = 1|X = x)
\]

Goal: Obtain identified set for \( q(\tau|x) \) under hypothetical \( d(F_{y|1,x}, F_{y|0,x}) \leq k \).

For the distance metric we use Kolmogorov-Smirnov, which is given by:

\[
S(F) \equiv \sup_{x \in \mathcal{X}} KS(F_{y|1,x}, F_{y|0,x}) = \sup_{x \in \mathcal{X}} \sup_{c \in \mathbb{R}} |F_{y|1,x}(c) - F_{y|0,x}(c)|
\]

Comments

- KS provides control over maximal distance between \( F_{y|1,x} \) and \( F_{y|0,x} \).
- Nests a wide nonparametric class of potential selection mechanisms.
Choice of Metric

Information necessarily lost with scalar index of selection, but ...

- Not ruling out selection mechanisms as done in parametric approaches.
- Different levels of selection can be considered at each quantile.
- Easy extension to different weights on covariate realizations.
- Scalar metric well suited to sensitivity analysis.

\[ S(F) \]

Example: Suppose the data generating process is given by:

\[(Y, v) \sim N(0, \rho_1 \rho_1) \]

\[ D = I \{ \mu + v > 0 \} \]

where \( \mu \) is chosen so that the missing probability is 25% to match data.

<table>
<thead>
<tr>
<th>Density</th>
<th>Value 1</th>
<th>Value 2</th>
<th>Value 3</th>
<th>Value 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0672</td>
<td>0.40</td>
<td>0.2778</td>
<td>0.70</td>
</tr>
<tr>
<td>0.20</td>
<td>0.1355</td>
<td>0.50</td>
<td>0.3520</td>
<td>0.80</td>
</tr>
<tr>
<td>0.30</td>
<td>0.2069</td>
<td>0.60</td>
<td>0.4304</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Patrick Kline UC Berkeley
Choice of Metric

Information necessarily lost with scalar index of selection, but ...

- Not ruling out selection mechanisms as done in parametric approaches.
- Different levels of selection can be considered at each quantile.
- Easy extension to different weights on covariate realizations.
- Scalar metric well suited to sensitivity analysis.

What is a big $S(F)$?
Choice of Metric

Information necessarily lost with scalar index of selection, but ...
- Not ruling out selection mechanisms as done in parametric approaches.
- Different levels of selection can be considered at each quantile.
- Easy extension to different weights on covariate realizations.
- Scalar metric well suited to sensitivity analysis.

What is a big $S(F)$?

Example: Suppose the data generating process is given by:

\[(Y, v) \sim N(0, (\frac{1}{\rho} \rho)) \quad D = 1\{\mu + v > 0\} .\]

where $\mu$ is chosen so that the missing probability is 25% to match data.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$S(F)$</th>
<th>$\rho$</th>
<th>$S(F)$</th>
<th>$\rho$</th>
<th>$S(F)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0672</td>
<td>0.40</td>
<td>0.2778</td>
<td>0.70</td>
<td>0.5165</td>
</tr>
<tr>
<td>0.20</td>
<td>0.1355</td>
<td>0.50</td>
<td>0.3520</td>
<td>0.80</td>
<td>0.6158</td>
</tr>
<tr>
<td>0.30</td>
<td>0.2069</td>
<td>0.60</td>
<td>0.4304</td>
<td>0.90</td>
<td>0.7377</td>
</tr>
</tbody>
</table>
Figure: Missing and Observed Outcome CDFs

Patrick Kline UC Berkeley
Figure: Distance Between Missing and Observed Outcome CDFs
Suppose a fraction $k$ of the missing population is distributed according to an arbitrary CDF $\tilde{F}_{y|x}$, while the remaining fraction $1-k$ of that population are missing at random in the sense that they are distributed according to $F_{y|1,x}$. Then:

$$F_{y|0,x}(c) = (1-k)F_{y|1,x}(c) + k\tilde{F}_{y|x}(c),$$

where $\tilde{F}_{y|x}$ is unknown, and the above holds for all $x \in \mathcal{X}$ and any $c \in \mathbb{R}$. Now:

$$S(F) = \sup_{x \in \mathcal{X}} \sup_{c \in \mathbb{R}} |F_{y|1,x}(c) - k\tilde{F}_{y|x}(c) - (1-k)F_{y|1,x}(c)| $$

$$= k \times \sup_{x \in \mathcal{X}} \sup_{c \in \mathbb{R}} |F_{y|1,x}(c) - \tilde{F}_{y|x}(c)|.$$

**Worst Case:** $S(F) = k$. Thus, $k$ gives bound on the fraction of the missing sample that is not well represented by the observed data distribution.
Nominal Identified set for $q(\tau|\cdot)$

Assumption (A)
(i) $X \in \mathbb{R}^l$ has finite support $\mathcal{X}$.
(ii) $F_{y|d,x}(c)$ is continuous, strictly increasing $\forall c$ with $0 < F_{y|d,x}(c) < 1$.
(iii) $D$ equals one if $Y$ is observable and zero otherwise.

Lemma Under (A), if $S(F) \leq k$, then the identified set for $q(\tau|\cdot)$ is:

$$C(\tau, k) \equiv \{\theta : \mathcal{X} \rightarrow \mathbb{R} : q_L(\tau, k|x) \leq \theta(x) \leq q_U(\tau, k|x)\}$$

where the bounds $q_L(\tau, k|x)$ and $q_U(\tau, k|x)$ are given by:

$$q_L(\tau, k|x) \equiv F_{y|1,x}^{-1}\left(\frac{\tau - \min\{\tau + kp(x), 1\}\{1 - p(x)\}}{p(x)}\right)$$

$$q_U(\tau, k|x) \equiv F_{y|1,x}^{-1}\left(\frac{\tau - \max\{\tau - kp(x), 0\}\{1 - p(x)\}}{p(x)}\right)$$
Example – No Covariates and $p(x) = \frac{2}{3}$
Suppose \( X \) is binary so that \( X \in \{0, 1\} \), and write:

\[
Y = q(\tau | X) + \epsilon \quad P(\epsilon \leq 0 | X) = \tau
\]

Suppose that under MAR we have \( q(\tau_0 | X = 1) \neq q(\tau_0 | X = 0) \) for some \( \tau_0 \).

We can evaluate sensitivity of this conclusion to MAR by defining:

\[
k_0 \equiv \inf k : q_L(\tau_0, k | X=1) - q_U(\tau_0, k | X=0) \leq 0 \leq q_U(\tau_0, k | X=1) - q_L(\tau_0, k | X=0)
\]

**Comment**

- \( k_0 \) is the minimal level for overturning \( q(\tau_0 | X = 1) \neq q(\tau_0 | X = 0) \).
- Large \( k_0 \) indicates robust conclusion.
Example 2 (Distributional Analysis)

We want to know if $F_{y|x=1}$ first order stochastically dominates $F_{y|x=0}$.

Suppose that under MAR we find $q(\tau|X = 1) > q(\tau|X = 0)$ at all $\tau$.

We evaluate sensitivity of FOSD conclusion by examining:

$$k_0 \equiv \inf k : q_L(\tau, k|X = 1) \leq q_U(\tau, k|X = 0) \text{ for some } \tau \in (0, 1)$$

Comment

- $k_0$ is the minimal level of selection under which the conclusion of FOSD may be undermined.
Example 3 (Breakdown Analysis)

\[ Y = q(\tau|X) + \epsilon \quad \text{and} \quad P(\epsilon \leq 0|X) = \tau \]

Suppose that under MAR we have \( q(\tau|X = 1) \neq q(\tau|X = 0) \) for multiple \( \tau \).

More nuanced analysis can consider the quantile specific critical level:

\[ \kappa_0(\tau) \equiv \inf k : q_L(\tau, k|X=1) - q_U(\tau, k|X=0) \leq 0 \leq q_U(\tau, k|X=1) - q_L(\tau, k|X=0) \]

**Comment**

- Changes in \( \tau \) map out a “breakdown function” \( \tau \mapsto \kappa_0(\tau) \).
- Reveals differential sensitivity of the entire conditional distribution.
1 Nominal Identified Set

2 Parametric Approximation

3 Inference

4 Changes in Wage Structure

5 CPS-SSA Analysis
Adding Parametric Structure

With lots of covariates, convenient to assume a linear parametric model:

\[ q(\tau|X) = X'\beta(\tau) \]

Identified set for \( \beta(\tau) \) is intersection of \( C(\tau, k) \) with parametric models:

\[
\left\{ \beta(\tau) \in \mathbb{R}^l : q_L(\tau, k|X) \leq X'\beta(\tau) \leq q_U(\tau, k|X) \right\}
\]

Comments:

- Set of functions in identified set may be severely restricted.
- Inadvertently rewards misspecification.

Identification by misspecification
Figure: Linear Conditional Quantile Functions as a Subset of the Identified Set

- **Quantile Bounds**
- **Lowest Slope Line**
- **Highest Slope Line**
Adding Parametric Structure

Instead allow for misspecification in the linear quantile model

\[ Y = X' \beta(\tau) + \eta \]

- If identified, misspecification as “pseudo true” approximation

\[ \beta(\tau) \equiv \arg \min_{\gamma \in \mathbb{R}^l} \int (q(\tau|x) - x'\gamma)^2 dS(x) \]

- If partially identified, each \( \theta \in C(\tau, k) \) implies a pseudo true vector \( \beta(\tau) \)

\[ \mathcal{P}(\tau, k) \equiv \left\{ \beta \in \mathbb{R}^l : \beta = \arg \min_{\gamma \in \mathbb{R}^l} \int (\theta(x) - x'\gamma)^2 dS(x) \text{ for some } \theta \in C(\tau, k) \right\} \]

\Rightarrow \text{i.e. consider } \beta \in \mathbb{R}^l \text{ that are best approximation to some } \theta \in C(\tau, k).
Figure: Conditional Quantile and its Pseudo-True Approximation
Misspecification

Choice of quadratic loss allows for simple characterization of $\mathcal{P}(\tau, k)$.

**Lemma:** Under (A), if $S(F) \leq k$ and $\int xx' dS(x)$ is invertible:

$$\mathcal{P}(\tau, k) = \left\{ \beta = \left[ \int xx' dS(x) \right]^{-1} \int x\theta(x) dS(x) : q(\tau, k|x) \leq \theta(x) \leq q(\tau, k|x) \right\}$$

**Note:** It follows that $\mathcal{P}(\tau, k)$ is convex.

**One more assumption:** We will assume the measure $S$ is known.

- Analogous to having a known loss function.
Parameter of Interest

Inference on parameters of the form $\lambda' \beta(\tau)$ for some $\lambda \in \mathbb{R}^l$.

**Corollary:** The identified set for $\lambda' \beta(\tau)$ is $[\pi_L(\tau, k), \pi_U(\tau, k)]$, where:

\[
\pi_L(\tau, k) \equiv \inf_{\theta} \lambda' \left[ \int xx'dS \right]^{-1} \int x\theta(x)dS(x) \quad \text{s.t.} \quad q_L(\tau, k|x) \leq \theta(x) \leq q_U(\tau, k|x)
\]

\[
\pi_U(\tau, k) \equiv \sup_{\theta} \lambda' \left[ \int xx'dS \right]^{-1} \int x\theta(x)dS(x) \quad \text{s.t.} \quad q_L(\tau, k|x) \leq \theta(x) \leq q_U(\tau, k|x)
\]

**Comments:**

- Examples: individual coefficients and fitted values.
- Bounds sharp for fixed $\tau$ and $k$.
- Bounds become wider with $k$ and change across $\tau$. 
**Bounds on the Process**

Previous corollary implies that if $S(F) \leq k$, then $\lambda' \beta(\cdot)$ belongs to:

$$
G(k) \equiv \left\{ g : [0, 1] \rightarrow \mathbb{R} : \pi_L(\tau, k) \leq g(\tau) \leq \pi_U(\tau, k) \text{ for all } \tau \right\}
$$

Unfortunately, $G(k)$ is not a sharp identified set of the process $\lambda' \beta(\cdot)$.

However ...

- The bounds $\pi_L(\cdot, k)$ and $\pi_U(\cdot, k)$ are in identified set where finite.
- The bounds of $G(k)$ are sharp at every point of evaluation $\tau$.
- If $\theta \notin G(k)$, then the function $\theta(\cdot)$ cannot equal $\lambda' \beta(\cdot)$.
- Ease of analysis and graphical representation.
Example 1 (cont)

\[ Y = \alpha(\tau) + X' \beta(\tau) + \eta \]

Suppose that under MAR we have \( \beta(\tau_0) \neq 0 \) for some specific quantile \( \tau_0 \).

We can evaluate sensitivity of this conclusion to MAR by defining:

\[ k_0 \equiv \inf k : \pi_L(\tau_0, k) \leq 0 \leq \pi_U(\tau_0, k) \]

Comment

- \( k_0 \) is the minimal level of selection necessary to overturn \( \beta(\tau_0) \neq 0 \).
Example 2 (cont)

\[ Y = \alpha(\tau) + X'\beta(\tau) + \eta \]

Suppose that under MAR we have \( \beta(\tau) > 0 \) for multiple \( \tau \).

We evaluate sensitivity to conclusion of \( F_{y|x} \) being increasing at some \( \tau \):

\[ k_0 \equiv \inf k : \pi_L(\tau, k) \leq 0 \text{ for all } \tau \in [0, 1] \]

Comment

- \( k_0 \) is the minimal level of selection that overturns \( \beta(\tau) > 0 \) for some \( \tau \).
- \( \pi_L(\cdot, k_0) \) is in identified set for \( \beta(\cdot) \) under \( S(F') \leq k \).
Example 3 (cont)

\[ Y = \alpha(\tau) + X'\beta(\tau) + \eta \]

Suppose that under MAR we have \( \beta(\tau) \neq 0 \) for multiple \( \tau \).

More nuanced analysis can consider the quantile specific critical level:

\[ \kappa_0(\tau) \equiv \inf k : \pi_L(\tau, k) \leq 0 \leq \pi_U(\tau, k) \]

Comment

- Changing \( \tau \) maps out a “breakdown function” \( \tau \mapsto \kappa_0(\tau) \).
- Reveals differential sensitivity of the entire conditional distribution.
1 Nominal Identified Set
2 Parametric Approximation
3 Inference
4 Changes in Wage Structure
5 CPS-SSA Analysis
Estimating Bounds

- Study estimators for bound functions $\pi_L(\tau, k)$, $\pi_U(\tau, k)$ given by:

\[
\hat{\pi}_L(\tau, k) \equiv \inf_{\theta} \lambda' \left[ \int x x' dS(x) \right]^{-1} \int x \theta(x) dS(x) \quad \text{s.t.} \quad \hat{q}_L(\tau, k|x) \leq \theta(x) \leq \hat{q}_U(\tau, k|x)
\]

\[
\hat{\pi}_U(\tau, k) \equiv \sup_{\theta} \lambda' \left[ \int x x' dS(x) \right]^{-1} \int x \theta(x) dS(x) \quad \text{s.t.} \quad \hat{q}_L(\tau, k|x) \leq \theta(x) \leq \hat{q}_U(\tau, k|x)
\]

- Need distribution as processes on $L^\infty(B)$, where for $0 < 2\epsilon < \inf_x p(x)$:

\[
B \equiv \left\{ (\tau, k) : \begin{array}{ll}
(i) & kp(x)(1 - p(x)) + 2\epsilon \leq \tau p(x) \\
(ii) & kp(x)(1 - p(x)) + 2\epsilon \leq (1 - \tau)p(x) \\
(iii) & k \leq \tau \\
(iv) & k \leq 1 - \tau
\end{array} \right\}
\]

Comments:

- The bounds $\pi_L(\tau, k)$ and $\pi_U(\tau, k)$ are finite everywhere on $B$.
- Large or small values of $\tau$ must be accompanied by small values of $k$. 

Patrick Kline 
UC Berkeley
Estimating Bounds

- Recall \( q_L(\tau, k|x) \) and \( q_U(\tau, k|x) \) were defined as quantiles of \( F_{y|1,x} \):

\[
q_L(\tau, k|x) = \arg \min_{c \in \mathbb{R}} Q_x(c|\tau, \tau+kp(x)) \quad q_U(\tau, k|x) = \arg \min_{c \in \mathbb{R}} Q_x(c|\tau, \tau-kp(x))
\]

where the family of criterion functions \( Q_x(c|\tau, b) \) is given by:

\[
Q_x(c|\tau, b) \equiv \left( P(Y \leq c, X = x, D = 1) + bP(D = 0, X = x) - \tau P(X = x) \right)^2
\]

- This suggests an extremum estimation approach given by:

\[
\hat{q}_L(\tau, k|x) = \arg \min_{c \in \mathbb{R}} Q_{x,n}(c|\tau, \tau+\hat{p}(x)) \quad \hat{q}_U(\tau, k|x) = \arg \min_{c \in \mathbb{R}} Q_{x,n}(c|\tau, \tau-\hat{p}(x))
\]

where the criterion function \( Q_{x,n}(c|\tau, b) \) is the immediate sample analogue.
Asymptotic Distribution

Assumptions (B)

(i) $F_{y|1,x}$ has a continuous bounded derivative $f_{y|1,x}$

(ii) $f_{y|1,x}$ has a continuous bounded derivative $f'_{y|1,x}$

(iii) The matrix $\int xx'dS(x)$ is invertible.

(iv) $f_{y|1,x}$ is positive “over relevant range”.

Theorem Under Assumptions (A) and (B), if \(\{Y_i, X_i, D_i\}_{i=1}^n\) is IID, then:

$$\sqrt{n}(\hat{\pi}_L - \pi_L) - \hat{\pi}_U - \pi_U \xrightarrow{\mathcal{L}} G,$$

where $G$ is a gaussian process on $L^\infty(\mathcal{B}) \times L^\infty(\mathcal{B})$. 
Proof Outline

**Step 1:** Study distribution of minimizers of $Q_{x,n}(c|\tau, b)$ as a function of $(\tau, b)$.

- Obtain uniform asymptotic expansions for the minimizers.
- $Q_{x,n}(c|\tau, b)$ has enough structure to establish equicontinuity.

**Step 2:** Find distribution $(\hat{q}_L, \hat{q}_U)$ in $L^\infty(\mathcal{B} \times \mathcal{X}) \times L^\infty(\mathcal{B} \times \mathcal{X})$.

- Simply a restriction of the process derived in Step 1.

**Step 3:** Establish the distribution of $(\hat{\pi}_L, \hat{\pi}_U)$ on $L^\infty(\mathcal{B})$.

- Straightforward due to linear program.
Example 1 (cont)

Suppose under MAR we find that \( \beta(\tau_0) \neq 0 \) for some specific quantile \( \tau_0 \). Minimal level of selection necessary to undermine this conclusion is:

\[
k_0 \equiv \inf k : \pi_L(\tau_0, k) \leq 0 \leq \pi_U(\tau_0, k)
\]

Let \( r_{1-\alpha}^{(i)}(k) \) be the \( 1 - \alpha \) quantile of \( G^{(i)}(\tau_0, k) \) and define:

\[
\hat{k}_0 \equiv \inf k : \hat{\pi}_L(\tau_0, k) - \frac{r_{1-\alpha}^{(1)}(k)}{\sqrt{n}} \leq 0 \leq \hat{\pi}_U(\tau_0, k) + \frac{r_{1-\alpha}^{(2)}(k)}{\sqrt{n}}
\]

Then \( k_0 \in [\hat{k}_0, 1] \) with asymptotic probability greater than or equal to \( 1 - \alpha \).

Comments

- One sided confidence interval (rather than two sided) is natural.
- Relevant critical value depends on transformation of \( G \).
Example 2 (cont)

Suppose under MAR we find that $\beta(\tau) > 0$ for multiple $\tau$ and recall that

$$k_0 \equiv \inf k : \pi_L(\tau, k) \leq 0 \quad \text{for all } \tau \in [0, 1]$$

Let $r_{1-\alpha}(k)$ be the $1 - \alpha$ quantile of $\sup_{\tau} G^{(1)}(\tau, k)/\omega_L(\tau, k)$ and define:

$$\hat{k}_0 \equiv \inf k : \sup_{\tau} \hat{\pi}_L(\tau, k) - \frac{r_{1-\alpha}(k)}{\sqrt{n}} \omega_L(\tau, k) \leq 0$$

Then $k_0 \in [\hat{k}_0, 1]$ with asymptotic probability greater than or equal to $1 - \alpha$.

Comments

- Weight function $\omega_L$ allows to adjust for different asymptotic variances.
- Result exploits uniformity in $\tau$ but not in $k$. 

Patrick Kline  
UC Berkeley
Example 3 (cont)

Suppose under MAR we find that $\beta(\tau) \neq 0$ for multiple $\tau$ and recall that:

$$\kappa_0(\tau) \equiv \inf k : \pi_L(\tau, k) \leq 0 \leq \pi_U(\tau, k)$$

For $(\omega_L, \omega_U)$ positive weight functions, let $r_{1-\alpha}$ be the $1 - \alpha$ quantile of:

$$\sup_{\tau, k} \max \left\{ \frac{|G^{(1)}(\tau, k)|}{\omega_L(\tau, k)}, \frac{|G^{(2)}(\tau, k)|}{\omega_U(\tau, k)} \right\}$$

Then with asymptotic probability at least $1 - \alpha$ for all $\tau$, $\kappa_0(\tau)$ lies between:

$$\hat{\kappa}_L(\tau) \equiv \inf k : \hat{\pi}_L(\tau, k) - \frac{r_{1-\alpha}}{\sqrt{n}} \omega_L(\tau, k) \leq 0 \text{ and } 0 \leq \hat{\pi}_U(\tau, k) + \frac{r_{1-\alpha}}{\sqrt{n}} \omega_U(\tau, k)$$

$$\hat{\kappa}_U(\tau) \equiv \sup k : \hat{\pi}_L(\tau, k) + \frac{r_{1-\alpha}}{\sqrt{n}} \omega_L(\tau, k) \geq 0 \text{ or } 0 \geq \hat{\pi}_U(\tau, k) - \frac{r_{1-\alpha}}{\sqrt{n}} \omega_U(\tau, k)$$
Weighted Bootstrap

**Question** How do we obtain a consistent estimator for $r_{1-\alpha}$?

**Answer** Perturb the objective function and recompute (weighted bootstrap).

In all examples, $r_{1-\alpha}$ is quantile of $L(G_\omega)$ where $L$ is Lipschitz and

$$G_\omega(\tau, k) = \begin{pmatrix} G^{(1)}(\tau, k) / \omega_L(\tau, k) \\ G^{(2)}(\tau, k) / \omega_U(\tau, k) \end{pmatrix}$$

**In Particular**

- In Example 1 $\theta \mapsto L(\theta)$ is $L(G_\omega) = G_\omega^{(i)}(\tau_0, k)$.
- In Example 2 $\theta \mapsto L(\theta)$ is $L(G_\omega) = \sup_\tau G_\omega^{(1)}(\tau, k)$.
- In Example 3 $\theta \mapsto L(\theta)$ is $L(G_\omega) = \sup_{\tau, k} \max\{|G_\omega^{(1)}(\tau, k)|, |G_\omega^{(2)}(\tau, k)|\}$.

**Goal** Construct a general bootstrap procedure for quantiles of $L(G_\omega)$. 

Patrick Kline UC Berkeley
Weighted Bootstrap

**Step 1** Generate a random sample of weights \( \{W_i\} \) and define the criterion:

\[
\tilde{Q}_{x,n}(c|\tau, b) \equiv \left( \frac{1}{n} \sum_{i=1}^{n} W_i \{1\{Y_i \leq c, X_i = x, D_i = 1\} + b1\{D_i = 0, X_i = x\} - \tau 1\{X_i = x\}\} \right)^2
\]

Using \( \tilde{Q}_{x,n} \) instead of \( Q_{x,n} \) obtain analogues to \( \hat{q}_L(\tau, k|x) \) and \( \hat{q}_U(\tau, k|x) \)

\[
\tilde{q}_L(\tau, k|x) = \arg \min_{c \in \mathbb{R}} \tilde{Q}_{x,n}(c|\tau, \tau + k\tilde{p}(x)) \quad \tilde{q}_U(\tau, k|x) = \arg \min_{c \in \mathbb{R}} \tilde{Q}_{x,n}(c|\tau, \tau - k\tilde{p}(x))
\]

where \( \tilde{p}(x) \equiv (\sum_i W_i 1\{D_i = 1, X_i = x\})/(\sum_i W_i 1\{X_i = x\}). \)

**Step 2** Using the bounds \( \tilde{q}_L(\tau, k|x) \) and \( \tilde{q}_U(\tau, k|x) \) from **Step 1**, obtain:

\[
\tilde{\pi}_L(\tau, k) \equiv \inf_{\theta} \lambda' \left[ \int x x' dS(x) \right]^{-1} \int x \theta(x) dS(x) \quad \text{s.t.} \quad \tilde{q}_L(\tau, k|x) \leq \theta(x) \leq \tilde{q}_U(\tau, k|x)
\]

\[
\tilde{\pi}_U(\tau, k) \equiv \sup_{\theta} \lambda' \left[ \int x x' dS(x) \right]^{-1} \int x \theta(x) dS(x) \quad \text{s.t.} \quad \tilde{q}_L(\tau, k|x) \leq \theta(x) \leq \tilde{q}_U(\tau, k|x)
\]
Weighted Bootstrap

**Step 3** Using the bounds $\tilde{\pi}_L(\tau, k)$ and $\tilde{\pi}_U(\tau, k)$ from Step 2, define:

$$\tilde{G}_\omega(\tau, k) = \sqrt{n} \left( \frac{(\tilde{\pi}_L - \hat{\pi}_L)/\hat{\omega}_L}{(\tilde{\pi}_U - \hat{\pi}_U)/\hat{\omega}_U} \right)$$

where $\hat{\omega}_L(\tau, k)$ and $\hat{\omega}_U(\tau, k)$ are estimators for $\omega_L(\tau, k)$ and $\omega_U(\tau, k)$.

**Step 4** Estimate $r_{1-\alpha}$, the $1 - \alpha$ quantile of $L(G_\omega)$ by $\tilde{r}_{1-\alpha}$ defined as:

$$\tilde{r}_{1-\alpha} \equiv \inf \left\{ r : P \left( L(\tilde{G}_\omega) \geq r \left| \{Y_i, X_i, D_i\}_{i=1}^n \right) \geq 1 - \alpha \right\}$$

**Comments**

- Notice probability is conditional on $\{Y_i, X_i, D_i\}_{i=1}^n$ but not on $\{W_i\}_{i=1}^n$.
- In practice $\tilde{r}_{1-\alpha}$ can be obtained through simulations.
Assumptions (C)

(i) $\omega_L$ and $\omega_U$ are strictly positive and continuous on $\mathcal{B}$.
(ii) $\hat{\omega}_L$ and $\hat{\omega}_U$ are uniformly consistent on $\mathcal{B}$.
(iii) $W$ is positive a.s. independent of $(Y, X, D)$.
(iv) $W$ satisfies $E[W] = 1$ and $Var(W) = 1$.
(v) The transformation $L$ is Lipschitz continuous.
(vi) The cdf of $L(G_\omega)$ is strictly increasing and continuous at $r_{1-\alpha}$.

Theorem Under Assumptions (A)-(C), if $\{Y_i, X_i, D_i, W_i\}_{i=1}^n$ are IID, then:

$$\tilde{r}_{1-\alpha} \overset{p}{\rightarrow} r_{1-\alpha}$$
1 Nominal Identified Set

2 Parametric Approximation

3 Inference

4 Changes in Wage Structure

5 CPS-SSA Analysis
Roadmap


- Assess sensitivity of results to deviations from MAR.
Roadmap


- Assess sensitivity of results to deviations from MAR.

Then... How worried should we be?

- Investigate nature of deviations from MAR in matched CPS-SSA data
- Test for and measure departures from ignorability using KS metric.
Quantile Specific Returns

Like Angrist, Chernozhukov and Fernandez-Val (2006) we estimate:

\[ Y_i = X_i' \gamma(\tau) + E_i \beta(\tau) + \epsilon_i \quad P(\epsilon_i \leq 0|X_i, E_i) = \tau \]

where \( Y_i \) is log average weekly earnings, \( E_i \) is years of schooling, and \( X_i \) consists of intercept, black dummy and quadratic in potential experience.

Sample Restrictions

- Black and white men age 40-49 with education \( \geq 6 \) years.
- \( Y_i \) treated as missing for all obs with allocated earnings or weeks worked.
Data quality is deteriorating

Table: Fraction of Observations in Estimation Sample with Missing Weekly Earnings

<table>
<thead>
<tr>
<th>Census Year</th>
<th>Total Number of Observations</th>
<th>Allocated Earnings</th>
<th>Allocated Weeks Worked</th>
<th>Fraction of Total Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>80,128</td>
<td>12,839</td>
<td>5,278</td>
<td>19.49%</td>
</tr>
<tr>
<td>1990</td>
<td>111,070</td>
<td>17,370</td>
<td>11,807</td>
<td>23.09%</td>
</tr>
<tr>
<td>2000</td>
<td>131,265</td>
<td>26,540</td>
<td>17,455</td>
<td>27.70%</td>
</tr>
<tr>
<td>Total</td>
<td>322,463</td>
<td>56,749</td>
<td>34,540</td>
<td>23.66%</td>
</tr>
</tbody>
</table>
Figure: Worst Case Nonparametric Bounds on 1990 Medians and Linear Model Fits for Two Experience Groups of White Men.
Figure: Nonparametric Bounds on 1990 Medians and Best Linear Approximations for Two Experience Groups of White Men Under $S(F) \leq 0.05$. 
Figure: Uniform Confidence Regions for Schooling Coefficients by Quantile and Year Under Missing at Random Assumption ($S(F) = 0$).
Figure: Uniform Confidence Regions for Schooling Coefficients by Quantile and Year Under $S(F) \leq 0.05$. 
**Figure:** Uniform Confidence Regions for Schooling Coefficients by Quantile and Year Under $S(F) \leq 0.175$ (1980 vs. 1990).

Patrick Kline UC Berkeley
Figure: Confidence Intervals for Fitted Values Under $S(F) \leq 0.05$. 

Patrick Kline UC Berkeley
Distributional Sensitivity

Found critical $k$ at which $\pi^{80}_U(\tau, k) \geq \pi^{90}_L(\tau, k)$ for all $\tau$.

... more informative find a $\tau$ specific critical $k$ for each $\tau$.

Define $\tau$-“breakdown” point $\kappa_0(\tau)$ as the smallest $k \in [0, 1]$ for which

$$\pi^{80}_U(\tau, k) - \pi^{90}_L(\tau, k) \geq 0$$

⇒ pointwise defines a function $\kappa_0$ which at each $\tau$ gives critical $k$.

Comments

• $\kappa_0$ function summarizes distributional sensitivity to MAR assumption.
• Use $(\tau, k)$ uniformity to build confidence interval for $\kappa_0(\tau)$ uniform in $\tau$. 
Intersection of Sample Bounds

Breakdown Function

\( dY/dE \)

\( k \)

\( \tau \)
Figure: Breakdown Curve (1980 vs 1990).
1. Nominal Identified Set
2. Parametric Approximation
3. Inference
4. Changes in Wage Structure
5. CPS-SSA Analysis
How worried should we be?

**Goal:** Employ 1973 CPS-SSA File to assess $S(F)$.

Data on SSA and IRS earnings for respondents to March CPS

**Sample Restrictions**

- Black and white men between ages of 25 and 55.
- More than 6 years of schooling.
- Must have reported working at least one week in past year.
- Drop self-employed and occupations likely to receive tips.
- Drop observations with IRS earnings $\leq 1000$ or $\geq 50000$.

**Comments**

- Roughly 7.2% of observations have unreported CPS earnings.
- Use IRS rather than SSA earnings due to topcoding.
Assessing $S(F)$

Define,

$$p_L(x, \tau) \equiv P(D = 1 | X = x, F_y|x(Y) \leq \tau) \quad p_U(x, \tau) \equiv P(D = 1 | X = x, F_y|x(Y) > \tau)$$

Leads to alternative expression for distance between $F_{y|1,x}$ and $F_{y|0,x}$

$$|F_{y|1,x}(q(\tau|x)) - F_{y|0,x}(q(\tau|x))| = \frac{\sqrt{(p_L(x, \tau) - p(x))(p_U(x, \tau) - p(x))\tau(1 - \tau)}}{p(x)(1 - p(x))}$$

Comments

- Emphasizes the effect of selection.
- Only need estimate of $P(D = 1 | X = x, F_y|x(Y) = \tau)$.
- Use earnings information on nonrespondents to estimate selection.
Three Logit Models

\[ P(D = 0|X = x, F_y|x(Y) = \tau) = \Lambda(\beta_1 \tau + \beta_2 \tau^2 + \delta_x) \]  
(1)

\[ P(D = 0|X = x, F_y|x(Y) = \tau) = \Lambda(\beta_1 \tau + \beta_2 \tau^2 + \gamma_1 \delta_x \tau + \gamma_2 \delta_x \tau^2 + \delta_x) \]  
(2)

\[ P(D = 0|X = x, F_y|x(Y) = \tau) = \Lambda(\beta_{1,x} \tau + \beta_{2,x} \tau^2 + \delta_x) \]  
(3)

**Comments**

- Five year age categories, Four schooling (< 12, 12, 13 − 15, 16).
- Drop small cells (< 50 obs).
- Only need estimate of \( P(D = 1|X = x, F_y|x(Y) = \tau) \).
- Model (2) substantially increases Likelihood over model (1).
- LR test cannot reject model (2) for model (3).
Table: Logit Estimates of \( P(D = 0|X = x, F_y|x(Y) = \tau) \) in 1973 CPS-IRS Sample

<table>
<thead>
<tr>
<th>( b_1 )</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_1 )</td>
<td>-1.06</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.43)</td>
<td>(5.44)</td>
<td></td>
</tr>
<tr>
<td>( b_2 )</td>
<td>1.09</td>
<td>3.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.41)</td>
<td>(4.08)</td>
<td></td>
</tr>
<tr>
<td>( \gamma_1 )</td>
<td>0.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \gamma_2 )</td>
<td>1.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.73)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Log-Likelihood | -3,802.91 | -3798.48 | -3759.97 |
| Parameters | 37 | 39 | 105 |
| Number of observations | 15,027 | 15,027 | 15,027 |
| Demographic Cells | 35 | 35 | 35 |

**Ages 25-55**

| Min KS Distance | 0.02 | 0.02 | 0.01 |
| Median KS Distance | 0.02 | 0.05 | 0.12 |
| Max KS Distance (\( S(F') \)) | 0.02 | 0.17 | 0.67 |

**Ages 40-49**

| Min KS Distance | 0.02 | 0.02 | 0.01 |
| Median KS Distance | 0.02 | 0.05 | 0.08 |
| Max KS Distance (\( S(F') \)) | 0.02 | 0.09 | 0.39 |

Note: Asymptotic standard errors in parentheses.
MAR clearly violated

- Very high and very low earning individuals mostly likely to have missing earnings on average.
- But missingness pattern appears to be heterogenous across demographic cells.
- Difficult to have guessed pattern a priori.

Degree of Heterogeneity Affects Bottom Line

- Model 1: $S(F) = 0.02$
- Model 2: $S(F) = 0.09$
- Model 3: $S(F) = 0.39$
Distance Between Missing and Nonmissing CDFs by Quantile of IRS Earnings and Demographic Cell

\( \tau \)

\( k \)

Patrick Kline UC Berkeley
Conclusion

**Theory:** When data are poor, useful to check sensitivity to violations of MAR.

- KS provides natural metric for assessing violations of MAR.
- Methods developed here enable study of parametric approximating models.
- And allow for assessment of distributional sensitivity to MAR assumption.
Conclusion

**Theory:** When data are poor, useful to check sensitivity to violations of MAR.
- KS provides natural metric for assessing violations of MAR.
- Methods developed here enable study of parametric approximating models.
- And allow for assessment of distributional sensitivity to MAR assumption.

**Empirics:** Reexamine the quantile specific returns to education.
- Measured changes in wage structure between 1980-1990 fairly robust (except at low end of distribution).
- But changes over 1990-2000 easily confounded by a bit of selection and deterioration in quality of Census data.
- 1973 CPS-SSA file provides evidence of selection and heterogeneity.