Partial Identification and Inference in Binary Choice and Duration Panel Data Models

Jason R. Blevins

Duke University

December 4, 2009

INTRODUCTION

- Fixed effects panel data binary choice model: for all $t = 0, \ldots, T-1$,

$$y_t = 1\{x_t'\beta + c + u_t \geq 0\}.$$

- No distributional assumption on u_t. For all x and c, $F_{u_t|xc}$ satisfies the following:
  a. $F_{u_t|xc} = F_{u_0|xc}$ for all t.
  b. The support of $F_{u_0|xc}$ is $\mathbb{R}$.
- If the first component of $x_1 - x_0$ has support everywhere on $\mathbb{R}$ conditional on almost every value of the remaining components, then β is point identified.
- What can we learn about β when x_t is discrete or bounded?
- Extensions: lagged dependent variable models and panel data duration models.

Limited Support Regressors

Assumption (Discrete Regressors)

x_t is a discrete random vector with finite support. That is, $|\mathcal{X}| < \infty$, where $|\mathcal{X}|$ denotes the cardinality of the set $\mathcal{X}$.

Assumption (Bounded Regressors)

The first component of $x_1 - x_0$, has positive density everywhere on $\mathcal{W}_1 \subseteq \mathbb{R}$ for almost every value of the remaining components.

Examples are common in practice: age, education, income, marital status, race, gender, number of children, etc.

CONTRIBUTION

- Consider semiparametric fixed effects panel data models and panel data duration models with discrete or continuous regressors:
    - Characterizations of the identified set Θ_I.
    - Consistent estimators $\hat{\Theta}_n$ of Θ_I.
    - Rates of convergence of $\hat{\Theta}_n$ to Θ_I.
    - Construction of confidence regions for Θ_I.
- Develop general theorems for establishing the above in a class of models with similar properties.
- Conditions for obtaining rates of convergence in partially identified irregular models.

Related Literature

- Criterion function based inference: Manski and Tamer (2002), Chernozhukov, Hong, and Tamer (2007), Romano and Shaikh (2008, 2009), and Bugni (2008).
- Semiparametric models with limited support regressors: Bierens and Hartog (1988), Horowitz (1998), Manski and Tamer (2002), Magnac and Maurin (2008), Honoré and Lleras-Muney (2006), and Komarova (2008).
- Semiparametric fixed effects panel data and transformation models: Manski (1987), Honoré and Kyriazidou (2000), Ridder (1990), Horowitz (1996), Abrevaya (2000), and Chen (2002).
- Partial identification in panel data models: Honoré and Tamer (2006), Chernozhukov, Fernández-Val, Hahn, and Newey (2009), and Rosen (2009).

IDENTIFICATION

- Models with a finite vector of parameters θ in some parameter space Θ.
- Semiparametric models with unknown infinite-dimensional components.
- The distribution of observables P_{θ_0} is induced by some $\theta_0 \in \Theta$.
- A model is *point identified* if only θ_0 is consistent with P_{θ_0}.
- A model is *partially identified* if there exist multiple values of θ that are observationally equivalent to θ_0 (that is, such that $P_\theta = P_{\theta_0}$).
- The set of all such θ is the *identified set* and is denoted Θ_I.
- See Manski (2003) and Tamer (2009) for surveys.

Fixed Effects Model: Identification

Theorem
In the fixed effects model the identified set can be written

$$\Theta_I = \{\theta : \operatorname{sgn}\left(P(y_1 = 1 \mid x) - P(y_0 = 1 \mid x)\right) \\ = \operatorname{sgn}\left((x_1 - x_0)'\beta\right)\ F_x - a.s.\} .$$

Criterion Function Based Estimation

- Suppose there exists a function Q which is maximized exactly on Θ_I.
- Let Q_n denote the sample analog.
- For a sequence $\tau_n \xrightarrow{p} 0$, define:

$$\hat{\Theta}_n(\tau_n) \equiv \left\{ \theta \in \Theta : Q_n(\theta) \geq \sup_{\Theta} Q_n - \tau_n \right\}.$$

- Intuition: if $Q_n \to Q$ then we might expect $\hat{\Theta}_n \to \Theta_I$.
- We now have sets converging to sets so we need to define consistency.

Hausdorff Distance

- We work in the Hausdorff metric $(\mathcal{P}(\Theta), d_H)$.
- Let (Θ, d) be a metric space where d is the Euclidean distance.
- Let $d(\theta, A) \equiv \inf_{\theta' \in A} d(\theta, \theta')$ be the distance between θ and a set A.
- For a pair of subsets $A, B \subset \Theta$, the *Hausdorff distance* between A and B is

$$d_H(A, B) = \max\left\{\sup_{\theta \in B} d(\theta, A), \sup_{\theta \in A} d(\theta, B)\right\}.$$

- $d_H(A, B) = 0$ if and only if $A = B$.
- Now, we say that $\hat{\Theta}_n$ is consistent for Θ_I if $d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.

HAUSDORFF DISTANCE

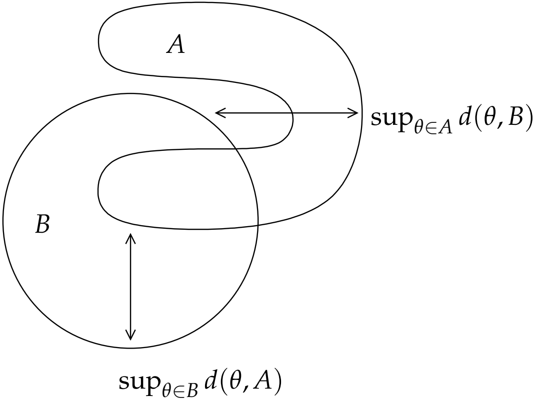


Figure: Hausdorff distance example.

CONSISTENCY IN GENERAL MODELS

Theorem (Consistency in General Models)

Suppose the following:

a. $\Theta \subset \mathbb{R}^k$ *is nonempty and compact with respect to the Euclidean metric.*

b. *There exists a function* $Q : \Theta \to \mathbb{R}$ *such that* $\arg\max_\Theta Q = \Theta_I$ *and for all* $\varepsilon > 0$ *there exists a* $\delta_\epsilon > 0$ *such that* $\sup_{\Theta \setminus \Theta_I^\varepsilon} Q \leq \sup_\Theta Q - \delta_\varepsilon$.

c. $Q_n(\theta)$ *is jointly measurable in* θ *and* $(x_1, \ldots, x_T, y_1, \ldots, y_T)$.

d. $\sup_\Theta |Q_n - Q| = O_p(1/b_n)$ *for some sequence* $b_n \to \infty$.

Under the conditions above:

1. *If* $\tau_n \xrightarrow{p} 0$, *then* $\sup_{\theta \in \hat{\Theta}_n} d(\theta, \Theta_I) \xrightarrow{p} 0$.
2. *If* $\tau_n \xrightarrow{p} 0$ *and* $\tau_n b_n \xrightarrow{p} \infty$, *then* $\lim_{n\to\infty} P(\Theta_I \subseteq \hat{\Theta}_n) = 1$.

Thus, if $\tau_n \xrightarrow{p} 0$ *and* $\tau_n b_n \xrightarrow{p} \infty$ *then* $d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.

Fixed Effects Model: Consistent Estimation

Objective function:

$$Q(\theta) = \mathrm{E}\left[(y_1 - y_0)\,\mathrm{sgn}\left((x_1 - x_0)'\beta\right)\right].$$

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - y_{i0})\,\mathrm{sgn}\left((x_{i1} - x_{i0})'\beta\right)$$

We show the following:

- $\arg\max_\theta Q = \Theta_I$.
- $\sup_\Theta |Q_n - Q| = O_p(n^{-1/2})$, that is, $b_n = n^{1/2}$.

Thus, $\hat{\Theta}_n$ is consistent for Θ_I.

RATES OF CONVERGENCE IN SMOOTH MODELS

Theorem

Suppose that the assumptions for consistency are satisfied and that there exist positive constants $(\delta, \kappa, \gamma_1, \gamma_2)$ *with* $\gamma_1 \geq \gamma_2$ *such that for any* $\varepsilon \in (0, 1)$ *there are* $(\kappa_\varepsilon, n_\varepsilon)$ *such that for all* $n \geq n_\varepsilon$*,*

$$Q_n(\theta) \leq \sup_{\Theta} Q_n - \kappa \cdot (d(\theta, \Theta_I) \wedge \delta)^{\gamma_1}$$

uniformly on $\{\theta \in \Theta : d(\theta, \Theta_I) \geq (\kappa_\varepsilon / b_n)^{1/\gamma_2}\}$ *with probability at least* $1 - \varepsilon$*. If* $\tau_n \xrightarrow{p} 0$ *and* $\tau_n b_n \xrightarrow{p} \infty$*, then* $d_H(\hat{\Theta}_n, \Theta_I) = O_p(\tau_n^{1/\gamma_2})$*.*

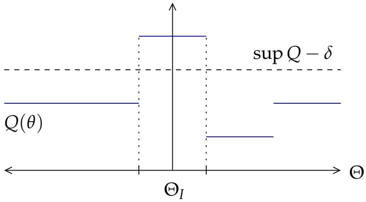
Rates of Convergence in Discrete Models

Theorem

Suppose the assumptions for consistency are satisfied and that there exists a positive constant δ such that

$$Q(\theta) \leq \sup_{\Theta} Q - \delta$$

for all $\theta \in \Theta \setminus \Theta_I$. If $\tau_n \xrightarrow{p} 0$ and $\tau_n b_n \xrightarrow{p} \infty$, then for any sequence r_n, $r_n d_H(\hat{\Theta}_n, \Theta_I) \xrightarrow{p} 0$.



FIXED EFFECTS MODEL: RATES OF CONVERGENCE

When x is discrete:

- Q is a step function and satisfies the constant majorant condition.
- $\hat{\Theta}_n$ converges to Θ_I arbitrarily fast.

When at least one component of $x_1 - x_0$ is continuous:

- Polynomial majorant condition: Q_n is quadratic outside of $b_n^{3/2}$ neighborhoods of Θ_I, corresponding to $\gamma_1 = 2$ and $\gamma_2 = 3/2$.
- The general rate of convergence is τ_n^{1/γ_2}.
- τ_n can be chosen arbitrarily close to $n^{-1/2}$.
- Therefore, $\hat{\Theta}_n$ can achieve rates arbitrarily close to, but slower than, $n^{-1/3}$.

Confidence Regions in General Models

- Conditions for discrete models.
- Let $C_n(\kappa_n)$ denote the set

$$C_n(\kappa_n) = \left\{ \theta \in \Theta : b_n Q_n(\theta) \geq \sup_{\Theta} b_n Q_n - \kappa_n \right\}.$$

- $\kappa_n = b_n \tau_n$ yields consistent estimates.
- We also want a sequence $\hat{c}_n$ which yields a $1 - \alpha$ confidence region:

$$\lim_{n \to \infty} P(\Theta_I \subseteq C_n(\hat{c}_n)) \geq 1 - \alpha.$$

Confidence Regions: Convergence of $\mathcal{Q}_n$

Inference is based on the following relationship:

$$
\begin{aligned}
P\left(\Theta_I \subseteq C_n(\hat{c}_n)\right) &= P\left(\inf_{\Theta_I} b_n Q_n \geq \sup_{\Theta} b_n Q_n - \hat{c}_n\right) \\
&= P\left(\mathcal{Q}_n \leq \hat{c}_n\right)
\end{aligned}
$$

where $\mathcal{Q}_n$ is our inferential statistic defined as:

$$\mathcal{Q}_n \equiv \sup_{\Theta} b_n Q_n - \inf_{\Theta_I} b_n Q_n$$

Assumption (Convergence of $\mathcal{Q}_n$)

Suppose that $P\{\mathcal{Q}_n \leq c\} \to P\{\mathcal{Q} \leq c\}$ *for each* $c \in \mathbb{R}$.

Lemma

If the above assumption holds, then for any
$\hat{c}_n \xrightarrow{p} c(1-\alpha) \equiv \inf\{c : P\{\mathcal{Q} \leq c\} \geq 1-\alpha\}$ *for some* $\alpha \in (0,1)$,

$$P\{\Theta_I \subseteq C_n(\hat{c}_n)\} \geq (1-\alpha) + o(1).$$

Confidence Regions: Algorithm

To construct a sequence $\hat{c}_n$ yielding conservative confidence regions $C_n(\hat{c}_n)$ with asymptotic coverage probability of at least $1 - \alpha$:

1. Choose a subsample size $m < n$ such that $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. Let M_n denote the number of subsets of size m and let κ_n be any sequence such that such that $C_n(\kappa_n)$ is a consistent estimator of Θ_I (e.g., $\kappa_n \propto \sqrt{\ln n}$).
2. Compute $\hat{c}_n$ as the $1 - \alpha$ quantile of the values $\{\hat{\mathcal{Q}}_{n,m,j}\}_{j=1}^{M_n}$ where

$$\hat{\mathcal{Q}}_{n,m,j} \equiv \sup_{\theta \in \Theta} b_m Q_{n,m,j}(\theta) - \inf_{\theta \in C_n(\kappa_n)} b_m Q_{n,m,j}(\theta)$$

   and $Q_{n,m,j}$ denotes the sample objective function constructed using the j-th subsample of size m.
3. Report $C_n(\kappa_n)$ as a consistent estimate of Θ_I and $C_n(\hat{c}_n)$ as a conservative confidence region.

Confidence Regions: Validity of Subsampling

Theorem

Suppose that:

- *the conditions for consistency hold,*
- *the constant majorant condition is satisfied,*
- $\mathcal{Q}_n$ *has a limiting distribution, and*
- $m \to \infty$*, and* $m/n \to 0$ *as* $n \to \infty$*.*

Let $1 - \alpha$ *denote the desired coverage level, where the distribution of* $\mathcal{Q}$ *is continuous at* $c(1-\alpha)$*. Then,* $\hat{c}_n \xrightarrow{p} c(1-\alpha)$*.*

FIXED EFFECTS MODEL: CONFIDENCE REGIONS

- Convergence of $\mathcal{Q}_n$:

$$\mathcal{Q}_n \xrightarrow{\mathrm{d}} \inf_{\theta \in \Theta_I} \mathbb{G}(\theta) \equiv \mathcal{Q}$$

  where $\mathbb{G}$ is a Gaussian process on Θ.
- Approximability of $\mathcal{Q}_n$: This holds easily when x is discrete since $\hat{\Theta}_n$ converges arbitrarily fast.
- A similar condition holds when x is bounded and estimators have polynomial rates of convergence.

LAGGED DEPENDENT VARIABLE MODEL

Suppose that for $t = 0$,

$$P(y_0 = 0 \mid x, c) = p_0(x, c),$$

and for $t = 1, \ldots, T-1$,

$$y_t = 1\{x_t'\beta + \gamma y_{t-1} + c + u_t \geq 0\},$$

where

- x_t is a random variable with support $\mathcal{X} \subset \mathbb{R}^k$,
- c is a real-valued random variable,
- $\theta = (\beta, \gamma) \in \Theta \subset \mathbb{R}^{k+1}$ are the parameters of interest,
- the disturbances u_t are serially independent with support $\mathbb{R}$.

Panel Data Duration Models

For all t,

$$\Lambda(y_t) = x_t'\beta + c + u_t$$

where

- Λ is a strictly increasing function,
- x_t is a random vector with support $\mathcal{X} \subset \mathbb{R}^k$,
- c is a real-valued random variable,
- $\theta = (\beta) \in \Theta \subseteq \mathbb{R}^k$ is the parameter of interest,
- u_t is stationary conditional on (x, c).

Objective function for estimating β (when $T = 2$):

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \operatorname{sgn}(y_1 - y_0) \cdot \operatorname{sgn}\left((x_1 - x_0)'\beta\right)$$

PANEL DATA DURATION MODELS: BOUNDING Λ

- Normalize $\Lambda(\bar{y}_0) = 0$.
- Suppose we know θ_0, then we can estimate $\Lambda(\bar{y})$:

$$\Gamma_n(\bar{y}, \lambda, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} (1\{y_1 > \bar{y}\} - 1\{y_0 > \bar{y}_0\}) \, 1\{(x_0 - x_1)' \beta_0 \leq \lambda\}.$$

- Thus, given an estimated set $\hat{\Theta}_n$, this suggests bounding $\Lambda(\bar{y})$:

$$\hat{\Lambda}_n(\bar{y}) = \left\{ \lambda : \lambda = \arg\max \Gamma_n(\bar{y}, \lambda, \hat{\theta}_n) \text{ for some } \theta_n \in \hat{\Theta}_n \right\}$$

Monte Carlo Experiments: Model

A representative fixed effects binary choice model:

$$y_{it} = 1\{x_{i1t} + \beta x_{i2t} + c_i + u_{it} \geq 0\},$$

where

$$\begin{aligned} x_{i1t} &\sim \text{Uniform}(\{-2,-1,0,1,2\}), \\ x_{i2t} &\sim \text{Uniform}(\{1,2,3,4,5,6,7,8,9,10\}), \\ c_i &= (x_{i11} + x_{i12} + x_{i21} + x_{i22})/4, \\ u_{it} &\sim \text{Normal}(0,1). \end{aligned}$$

Population parameter: $\theta_0 = \beta_0 = -0.15$.
Identified set: $\Theta_I = [-0.163, -0.148]$.

Monte Carlo Experiments: Objective Function

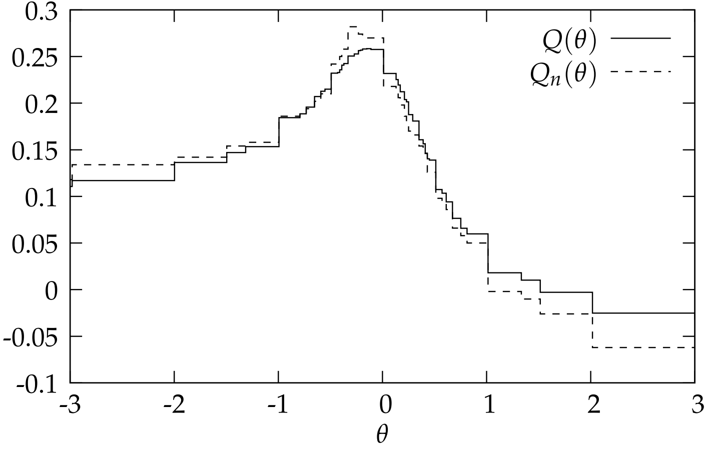


Figure: $Q(\theta)$ and one realization of $Q_n(\theta)$ for $n = 500$.

ESTIMATES WITH $\kappa_n \propto \sqrt{\ln n}$

κ_n	n	Mean $\hat{\Theta}_n$	St. Dev.	Coverage
$0.20\sqrt{\ln n}$	500	[-0.401, 0.044]	[0.096, 0.076]	0.98
	1000	[-0.365, 0.019]	[0.069, 0.051]	0.99
	2000	[-0.328, -0.001]	[0.045, 0.016]	0.99
	4000	[-0.305, -0.003]	[0.037, 0.003]	1.00
	8000	[-0.277, -0.003]	[0.032, 0.005]	1.00
	16000	[-0.256, -0.003]	[0.019, 0.000]	1.00
	32000	[-0.246, -0.003]	[0.010, 0.000]	1.00
	64000	[-0.239, -0.003]	[0.015, 0.000]	1.00
$0.05\sqrt{\ln n}$	500	[-0.258, -0.041]	[0.083, 0.085]	0.66
	1000	[-0.247, -0.033]	[0.061, 0.065]	0.76
	2000	[-0.231, -0.044]	[0.048, 0.064]	0.81
	4000	[-0.215, -0.050]	[0.038, 0.065]	0.84
	8000	[-0.203, -0.055]	[0.031, 0.064]	0.88
	16000	[-0.196, -0.064]	[0.027, 0.062]	0.95
	32000	[-0.194, -0.078]	[0.021, 0.060]	0.97
	64000	[-0.188, -0.096]	[0.017, 0.051]	0.99

CONFIDENCE REGIONS WITH $\kappa_n \propto \sqrt{\ln n}$

			Empirical Coverage			
κ_n	m	n	0.750	0.900	0.950	0.990
$0.20\sqrt{\ln n}$	$n^{3/5}$	500	0.934	0.978	0.994	0.993
		1000	0.910	0.976	0.984	0.996
		2000	0.936	0.989	0.991	0.997
		4000	0.978	0.989	0.990	0.995
		8000	0.985	0.991	0.994	0.995
		16000	0.986	0.994	0.997	0.997
		32000	0.997	1.000	1.000	1.000
		64000	1.000	1.000	1.000	1.000
$0.05\sqrt{\ln n}$	$n^{3/5}$	500	0.608	0.775	0.877	0.918
		1000	0.721	0.853	0.897	0.936
		2000	0.808	0.912	0.931	0.946
		4000	0.859	0.926	0.939	0.957
		8000	0.900	0.918	0.926	0.945
		16000	0.898	0.954	0.965	0.974
		32000	0.939	0.975	0.980	0.982
		64000	0.969	0.983	0.992	0.995

ESTIMATES WITH $\kappa_n = 0$

n	Mean $\hat{\Theta}_n$	St. Dev.	Coverage
500	[-0.210, -0.096]	[0.080, 0.097]	0.34
1000	[-0.192, -0.109]	[0.058, 0.083]	0.29
2000	[-0.178, -0.122]	[0.047, 0.068]	0.31
4000	[-0.171, -0.126]	[0.039, 0.061]	0.30
8000	[-0.166, -0.132]	[0.031, 0.047]	0.35
16000	[-0.162, -0.135]	[0.027, 0.037]	0.41
32000	[-0.163, -0.143]	[0.022, 0.020]	0.50
64000	[-0.161, -0.143]	[0.017, 0.014]	0.61

Confidence Sets with $\kappa_n = 0$

	Empirical Coverage			
n	0.750	0.900	0.950	0.990
500	0.429	0.517	0.534	0.536
1000	0.377	0.461	0.495	0.501
2000	0.381	0.424	0.458	0.465
4000	0.372	0.416	0.430	0.447
8000	0.399	0.421	0.433	0.440
16000	0.442	0.457	0.459	0.461
32000	0.514	0.517	0.518	0.521
64000	0.622	0.622	0.622	0.622

Conclusion

- General results for a new class of models:
    - consistency
    - rates of convergence
    - inference
- Consider several specific models:
    - fixed effects panel data models
    - lagged dependent variable models
    - panel data duration models
- Future work:
    - Properties of $\hat{\Lambda}_n(\bar{y})$.
    - Inference for the true parameter θ_0.
    - Multinomial choice.

References I

Abrevaya, J. (2000). Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics 95*, 1–23.

Bierens, H. J. and J. Hartog (1988). Non-linear regression with discrete explanatory variables, with an application to the earnings function. *Journal of Econometrics 38*, 269–299.

Bugni, F. A. (2008). Bootstrap inference in partially identified models. Working paper, Northwestern University.

Chen, S. (2002). Rank estimation of transformation models. *Econometrica 70*, 1683–1697.

Chernozhukov, V., I. Fernández-Val, J. Hahn, and W. Newey (2009). Identification and estimation of marginal effects in nonlinear panel models. Unpublished manuscript, Massachusetts Institute of Technology.

Chernozhukov, V., H. Hong, and E. Tamer (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica 75*, 1243–1284.

Honoré, B. E. and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica 68*, 839–874.

Honoré, B. E. and A. Lleras-Muney (2006). Bounds in competing risks models and the war on cancer. *Econometrica 74*, 1675–1698.

Honoré, B. E. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica 74*, 611–629.

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica 64*, 103–137.

REFERENCES II

Horowitz, J. L. (1998). *Semiparametric Methods in Econometrics*. New York: Springer.

Komarova, T. (2008). Binary choice models with discrete regressors: Identification and misspecification. Unpublished manuscript, London School of Economics.

Magnac, T. and E. Maurin (2008). Partial identification in monotone binary models: Discrete regressors and interval data. *Review of Economic Studies 75*, 835–864.

Manski, C. F. (1987). Semiparametric analysis of random effects linear models from binary panel data. *Econometrica 55*, 357–362.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer-Verlag.

Manski, C. F. and E. Tamer (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica 70*, 519–546.

Ridder, G. (1990). The non-parametric identification of generalized accelerated failure-time models. *Review of Economic Studies 57*, 167–182.

Romano, J. P. and A. M. Shaikh (2008). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference 138*, 2786–2807.

Romano, J. P. and A. M. Shaikh (2009). Inference for the identified set in partially identified econometric models. *Econometrica*, forthcoming.

Rosen, A. (2009). Set identification via quantile restrictions in short panels. Working Paper CWP26/09, CEMMAP.

Tamer, E. (2009). Partial identification in econometrics. Unpublished manuscript, Northwestern University.