5B: Selection

© A. Colin Cameron
U. of Calif. - Davis

.

OeNB Summer School 2010
Microeconometrics
Oesterreichische Nationalbank (OeNB), Vienna, Austria

Based on
A. Colin Cameron and Pravin K. Trivedi,
Microeconometrics: Methods and Applications (MMA), ch.14
Microeconometrics using Stata (MUS), ch.14.
Data examples are from MUS.

Aug 30 - Sep 3, 2010

1. Introduction

- Analysis of censored data had same process determining the censored and uncensored data
  - selection models relax to allow different models for participation and outcome
- Models include
  - two-part model (with independent processes)
  - sample selection model (two-part model with correlated processes)
  - inverse-probability weighted estimators
  - Roy model where y depends in part on a binary outcome
  - the Roy model is related to the treatment evaluation literature
- Some methods assumes selection on observables only $(\mathbf{x})$
  - others additionally allow for selection on unobservables (u).

Outline

1. Introduction
2. Selection: Heckman sample selection model
3. Selection: Roy model
4. Selection: Mixed discrete / continuous structural economic models
5. Selection: Simultaneous equations models
6. Selection: Semiparametric estimation
7. Selection: Inverse-probability weighting
8. Selection: Treatment evaluation literature

2. Heckman sample selection model: summary

- Selection on observables and unobservables.
- Participation: We observe whether y_1^* is positive or negative
  - $y_1^* = \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1$
  - $y_1 = 1[y_1^* > 0]$
- Outcome: Only positive values of y_2^* are observed
  - $y_2^* = \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2$
  - $y_2 = y_2^*$ if $y_1 = 1$.
- Heckman's two-step procedure:
  - **1.** Estimate $\boldsymbol{\beta}_1$ by probit for $y_1^* > 0$ or $y_1^* < 0$ with regressors $\mathbf{x}_{1i}$.
  - Calculate $\widehat{\lambda}_i = \lambda(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1) = \phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)/\Phi(\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1)$.
  - **2.** For observed y_2 estimate $\boldsymbol{\beta}_2$ and δ in the OLS regression

$$y_{2i} = \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + \delta\widehat{\lambda}_i + w_i.$$

- If $\delta = 0$ then reduces to two part model.

3. Roy Model: Overview

- Selection on observables and unobservables.
- Suppose y is always observed, but only in one of two states.
  - e.g. observe wages if union job or if not union job
  - e.g. observe wages if get training or if do not get training
  - e.g. observe health expenditures if have health insurance or if do not.
- Control for self-selection on unobservables (not just observables).
  - e.g. people select into health insurance if they think they are likely to have high health expenditures, and we do not have data to control for this.

Roy Model: Definition

- We observe state $y_1 = 1$ or $y_1 = 0$ according to

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0. \end{cases}$$

- The consequent outcome is

$$y = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ y_3^* & \text{if } y_1^* \leq 0. \end{cases}$$

- Usual model

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \varepsilon_2 \\ y_3^* &= \mathbf{x}_3'\boldsymbol{\beta}_3 + \varepsilon_3. \end{aligned}$$

where errors are joint normal with means 0 and normalization $\sigma_1^2 = 1$.

Roy Model: Estimation

- Can estimate by ML.
- More common to use Heckman two-step estimator using

$$\begin{aligned} \mathrm{E}[y|\mathbf{x}, y_1^* > 0] &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda(\mathbf{x}_1'\boldsymbol{\beta}_1) \\ \mathrm{E}[y|\mathbf{x}, y_1^* \leq 0] &= \mathbf{x}_3'\boldsymbol{\beta}_3 - \sigma_{13}\lambda(-\mathbf{x}_1'\boldsymbol{\beta}_1), \end{aligned}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ and we have used $\sigma_1^2 = 1$.

  - First-stage probit of $y_1^* > 0$ yields $\widehat{\boldsymbol{\beta}}_1$ and hence $\lambda(\mathbf{x}_1'\widehat{\boldsymbol{\beta}}_1)$.
  - Two separate OLS regressions then lead to direct estimates of $\boldsymbol{\beta}_2, \sigma_{12}$ and $\boldsymbol{\beta}_3, \sigma_{13}$.
  - Estimates of σ_2^2 and σ_3^2 can then be obtained using the squared residuals from the regressions.

4. Mixed Discrete / Continuous Structural Economic Models

- Sample selection and Roy models have been obtained from utility maximization.
- Leading examples are
  - Heckman (1974) for labor supply
    - whether to work (participation) and amount worked (outcome).
  - Dubin and McFadden (1984) for appliance choice and energy consumption
    - whether has or electric appliances (discrete) and energy consumed given appliance choice
  - Hanemann (1984)
    - brand choice (discrete) and amount consumed given brand choice.

5. Further Topics: Simultaneous Equations Tobit Models

- A general bivariate example with endogenous regressors is

$$\begin{aligned} y_1^* &= \mathbf{x}_1'\boldsymbol{\beta}_1 + \alpha_1 y_2^* + \gamma_1 y_2 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2'\boldsymbol{\beta}_2 + \alpha_2 y_1^* + \gamma_2 y_1 + \varepsilon_2 \end{aligned}$$

- Here both y_2^* or y_2 appear in the first equation (and similarly y_1 or y_1^* in the second equation).
  - Identification conditions require some to be dropped (coherency conditions).
- Simplest to have r.h.s. endogenous variables be the latent variables y_2^* or y_1^*.
  - Then obtain a reduced form for y_1^* and y_2^*, in exactly the same way as regular linear simultaneous equations
  - Do Tobit estimation on this reduced form.
- More difficult when r.h.s. endogenous variables are the observed variables y_2 or y_1.

6. Semiparametric methods

- Consistency of preceding estimators requires correct specification of the error distribution.
  - Any misspecification leads to inconsistency
    - e.g. failure of normality or homoskedasticity.
  - so should generally treat Tobit estimates with skepticism
    - exception may be top-coded income if believe income is lognormal.
- Semiparametric estimators do not require specification of distribution of the error distribution.

Semiparametric methods: Tobit model

- Tobit MLE is very fragile to distributional misspecification
  - inconsistent if errors are nonnormal
  - inconsistent even if errors are normal but heteroskedastic
  - So need methods with fewer assumptions.
- Censored least absolute deviations (CLAD) for left-censoring at zero
  - $\widehat{\boldsymbol{\beta}}_{CLAD}$ minimizes $Q(\boldsymbol{\beta}) = \sum_i |y_i - \max(0, \mathbf{x}_i'\boldsymbol{\beta})|$.
  - Intuition is that censoring changes the mean of the data but not the median (if less than 50% of data is censored).
  - Least absolute deviations is regression analog of median.
  - Consistency requires that $\varepsilon|\mathbf{x}$ has median zero (e.g. errors are i.i.d.).

Semiparametric methods: Selection models

- For sample selection less has been done. Recall:

$$\begin{aligned}\mathsf{E}[y_2|y_1^*>0] &= \mathbf{x}_2'\boldsymbol{\beta}_2+\delta\times\mathsf{E}[\varepsilon_1|\varepsilon_1>-\mathbf{x}_1'\boldsymbol{\beta}_1] \\ &= \mathbf{x}_2'\boldsymbol{\beta}_2+g(\mathbf{x}_1'\boldsymbol{\beta}_1)\end{aligned}$$

- So we need $g(\mathbf{x}_1'\boldsymbol{\beta}_1)$ without specifying functional form of $g(\cdot)$
  - Heckman 2-step where at second step include a polynomial in $\mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1$ or $\widehat{\lambda}_i$.
  - e.g. regress noncensored y_{2i} on $\mathbf{x}_{2i}$, $\widehat{\lambda}_i$, $\widehat{\lambda}_i^2$ and $\widehat{\lambda}_i^3$
  - But need good discriminating model between $\mathbf{x}_1$ and $\mathbf{x}_2$ and still based on strong assumptions.

7. Inverse probability weighting overview

- Alternative way to handle selection
  - selection is we lose some data due to sample selection, attrition, ...
  - and do analysis only on the selected sample.
- Inverse probability weighting
  - assume selection is on observables only
  - then do weighted estimation
  - the weights are the inverse of the probability of selection
  - this downweights "oversampled" observations (like a weighted mean)
  - the weights may be known (e.g. from stratified sampling)
  - or the weights may be estimated from the data.
- Can be applied to wide range of methods
  - But assumes selection is on observables only.

Selected sample

- We want $\widetilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$
  - where $\widetilde{\boldsymbol{\theta}}$ is the estimator based on selected sample and
  - $\boldsymbol{\theta}_0$ is the limit of the estimator $\widehat{\boldsymbol{\theta}}$ if we had the complete nonselected sample.
- First, define $\boldsymbol{\theta}_0$.
  - $\widehat{\boldsymbol{\theta}}$ maximizes $\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$ based on complete nonselected sample
    - ★ where e.g. $q(\mathbf{w}_i, \boldsymbol{\theta}) = -(y_i - \mathbf{x}_i' \boldsymbol{\theta})^2$ or $q(\mathbf{w}_i, \boldsymbol{\theta}) = \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta})$.
  - $\boldsymbol{\theta}_0$ maximizes $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$.
- Then define the selection process
  - $s_i = 1$ if data $\mathbf{w}_i$ are observed and $s_i = 0$ otherwise.

Weighted m-estimator

- Naive estimation with selection sample: $\widetilde{\boldsymbol{\theta}}$ maximizes

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{N} s_i q(\mathbf{w}_i, \boldsymbol{\theta}).$$

  - Inconsistent if selection is on endogenous y!
  - Though may be consistent for selection on exogenous regressors x if we add an assumption about correct model specification.
  - Formally, need $\boldsymbol{\theta}_0$ that maximizes $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$ also maximizes $\mathrm{E}[s \times q(\mathbf{w}, \boldsymbol{\theta})]$.
- Instead weighted m-estimator with known selection probabilities: $\widehat{\boldsymbol{\theta}}_w$ maximizes

$$\sum_{i=1}^{N} \frac{s_i}{p(\mathbf{v}_i)} q(\mathbf{w}_i, \boldsymbol{\theta}),$$

  - where we know $p(\mathbf{v}_i) = \Pr[s_i = 1 | \mathbf{v}_i]$ and $\mathbf{v}_i$ contains $\mathbf{w}_i$
  - $p(\mathbf{v}_i)$ is known e.g. from stratified sampling or variable probability sampling
  - note that if $\mathbf{w}_i = (y_i, \mathbf{x}_i)$ then $\mathbf{v}_i$ includes y_i so $p(\mathbf{v}_i)$ controls for selection on y as well as on $\mathbf{x}$.

Estimated weights

- If we don't know the weights then estimate them and control for estimation error.
- Assumptions
  - there are extra variables $\mathbf{z}_i$ that are **always observed**
  - once we condition on $\mathbf{z}_i$ the selection probability no longer depends on $\mathbf{w}_i$ (the outcome of interest observed only if $s_i = 1$)
    - ★ $\Pr[s_i = 1|\mathbf{z}_i,\mathbf{w}_i] = \Pr[s_i = 1|\mathbf{z}_i] = p(\mathbf{z}_i)$
  - We have a valid parametric model $p(\mathbf{z}_i, \gamma)$ for $p(\mathbf{z}_i)$.
- Then the weighted m-estimator with estimated selection probabilities: $\widehat{\boldsymbol{\theta}}_w$ maximizes

$$\sum\nolimits_{i=1}^N \frac{s_i}{p(\mathbf{z}_i, \widehat{\gamma})} q(\mathbf{w}_i, \boldsymbol{\theta})$$

  - where $\widehat{\gamma}$ first maximizes $\sum_{i=1}^N \{s_i \ln p(\mathbf{z}_i, \gamma) + (1 + s_i) \ln(1 - p(\mathbf{z}_i, \gamma))\}$.

Implementation

- Do the following
  - 1. Do a flexible logit or probit of s_i on $\mathbf{z}_i$ for the full selected sample.
  - 2. Then do weighted estimation of y_i on $\mathbf{x}_i$ with selected sample only $(s_i = 1)$
  - 3. Conservative inference uses the robust standard errors from this regression
- Stata example:
  - `logit s z`
  - `poisson y x if s==1 [pweight 1/p]`
- Improvement
  - Inference ignored first-step estimation of the selection probabilities
  - Intuitively this should lead to larger standard errors for $\widehat{\boldsymbol{\theta}}_w$
  - But in fact it leads to smaller standard errors for $\widehat{\boldsymbol{\theta}}_w$
  - So conservative inference (report smaller t-statistics than truth)
  - Wooldridge (2002) shows how to get the correct smaller standard errors which is desirable but not necessary.

8. Treatment Effects Estimation

- Example is treatment effect of training $(d = 1)$ on earnings (y).
- We observe
  - continuous outcome y_i
  - binary treatment d_i $(= 1$ if treated and $= 0$ if not treated).
- For each person there are two potential outcomes
  - $y_{0i} = y_i$ if $d_i = 0$
  - $y_{1i} = y_i$ if $d_i = 1$.
- The evaluation problem is: we only observe

$$\begin{aligned} y_i &= d_i y_{1i} + (1 - d_i) y_{0i} \\ &= y_{0i} + d_i (y_{1i} - y_{0i}) \end{aligned}$$

  but we want to compute the treatment effect

$$\Delta_i = y_{1i} - y_{0i}.$$

- We are concerned that the treated are self-selected, so selection problem.

- Selection on observables methods
  - control function
  - matching
  - propensity score matching
  - regression discontinuity design (sharp)
- Selection additionally on unobservables methods
  - parametric (Roy model)
  - instrumental variables and LATE
  - panel data
  - differences in differences

Random experiment

- Classic example is experiment where randomly assign d_i.
- Then average treatment effect (ATE)

$$\text{ATE} = \bar{y}_1 - \bar{y}_0$$

- This is OLS estimate of α in regression

$$y_i = \alpha d_i + u_i.$$

- It can be more efficient to use OLS estimate of α in regression

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha d_i + u_i.$$

  - Reason: Regressors may reduce σ_u^2 so smaller standard errors.

Selection on observables only: Control function approach

- OLS estimate of α in regression

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha d_i + u_i.$$

- Consistency requires $\mathrm{E}[u_i|d_i] = \mathrm{E}[y_{ii} - \mathbf{x}_i'\boldsymbol{\beta} - \alpha d_i|d_i] = 0$
  - this assumes no selection on unobservables.
- Equivalently need $y_{0i}, y_{1i} \perp d_i|\mathbf{x}_i$
  - versus $y_{0i}, y_{1i} \perp d_i$ under random assignment.
- Best to have many regressors and flexible model.
- Restricts the treatment effect α to be equal for all individuals
  - matching relaxes this.

Selection on observables only: Matching

- Matching approach
  - More flexible as treatment effect can vary over individuals.
  - For each distinct value of $\mathbf{x}_i$ the average treatment effect is the difference in average value of y for the treated and untreated

$$\text{ATE}|\mathbf{x}_i = (\bar{y}_1|\mathbf{x}_i) - (\bar{y}_0|\mathbf{x}_i)$$

  - Then average the ATE over the distinct values of $\mathbf{x}_i$.
- Problem is may have too many distinct values of $\mathbf{x}_i$.
  - Instead do propensity score matching.

Selection on observables only: Propensity score matching

- For each individual calculate the probability of treatment, called the propensity score
$$p_i = \Pr[d_i = 1|\mathbf{x}_i]$$
  - Use a flexible logit model or kernel regression.
- Compare y_1 and y_0 for those with similar $\widehat{p}$.
  - Several ways to do this.
- For example, interval or stratification matching
  - For observations with similar $\widehat{p}_i$, say $\widehat{p}_i \in A_j$, the average treatment effect is
$$\text{ATE}|\widehat{p}_i \in A_j = (\overline{y}_1|\widehat{p}_i \in A_j) - (\overline{y}_0|\widehat{p}_i \in A_j)$$
  - Then average the ATE over A_j.
- Again consistency requires $y_{0i}, y_{1i} \perp d_i|\mathbf{x}_i$
  - Called ignorability assumption or unconfoundedness or conditional independence.
  - Also need overlap assumption that for each p_i there are both treated and untreated.

Selection on observables only: Regression discontinuity design

- Assume treatment occurs if variable s_i crosses a threshold

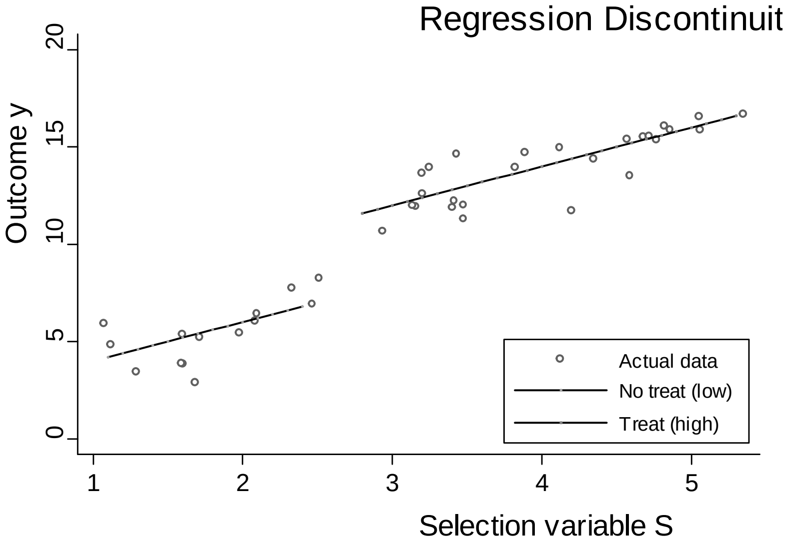
$$d_i = 1[s_i > s^*]$$

  - e.g. admitted to college if SAT score $>$ 1200.
- Then if s_i also determines the outcome y_i the treatment effect is
  - $(y$ for s just above $s^*)$ minus $(y$ for s just below $s^*)$
- The treatment effect is $\widehat{\alpha}$ from OLS estimation of

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + k(s_i) + \alpha d_i + u_i$$

  - where $k(s_i)$ is for example a cubic in s_i.

Regression discontinuity design



Selection on unobservables: IV, 2SLS and LATE

- With selection on unobservables (as well as observables) d_i is endogenous in
$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \alpha d_i + u_i.$$
- One solution is instrumental variables.
  - Assume there exists $\mathbf{z}_i$ such that $\mathrm{E}[u_i|\mathbf{z}_i] = 0$.
  - Estimate by IV (just-identified) or 2SLS (over-identified).
- The instrument may effect only a subset of population
  - e.g. earnings (y) and high school graduation (d)
  - instrument (z) is minimum school leaving age
  - effects only those likely to have $d = 0$.
- Local average treatment effect (LATE) covers this case
  - interpret 2SLS estimate as applying only to "compliers"
  - these are people subject to the treatment and
  - can explain why different instruments give different 2SLS estimates.

Selection on unobservables: panel data

- Binary treatment regressor is now d_{it} ($= 1$ if individual i receives treatment in period t)
- Assume a fixed effects model for outcome y_{it}

$$y_{it} = \phi d_{it} + \delta_t + \alpha_i + \varepsilon_{it},$$

  - where δ_t is a time-specific fixed effect
  - α_i is an individual-specific fixed effect possibly correlated with d_{it} (selection on unobservables).
- The individual effects α_i can be eliminated by first-differencing. Then

$$\Delta y_{it} = \phi \Delta d_{it} + (\delta_t - \delta_{t-1}) + \Delta \varepsilon_{it}.$$

  - The treatment effect ϕ can be consistently estimated by pooled OLS regression of Δy_{it} on Δd_{it} and a full set of time dummies.
- Essential assumption is d_{it} correlated only with time-invariant component α_i of the error.

Selection on unobservables: differences in differences

- Specialize preceding as follows
  - two time periods (1 and 2)
  - treatment occurs only in period 2 so
    - ★ $d_{i1} = 0$ for all individuals
    - ★ $d_{i2} = 1$ for treated and $d_{i2} = 0$ for the nontreated.
- The subscript t can be dropped and

$$\Delta y_i = \phi d_i + \delta + v_i,$$

  - where d_i is a binary treatment variable indicating whether or not the individual received treatment.
- The treatment effect can be estimated by OLS of Δy_i on an intercept and d_i.

- So do OLS of Δy_i on an intercept and d_i.
  - OLS reduces to $\widehat{\phi} = \Delta\bar{y}^{tr} - \Delta\bar{y}^{nt}$ where
    - ★ $\Delta\bar{y}^{tr}$ is sample average of Δy_i for the treated $(d_i = 1)$
    - ★ $\Delta\bar{y}^{nt}$ is sample average of Δy_i for nontreated $(D_i = 0)$.
    - ★ This estimator is called the differences-in-differences (DID) estimator.
- The DID estimator does not require panel data!
  - suppose two separate cross-sections are available for the two periods.
  - in the second period compute the averages $\bar{y}_2^{tr}$ and $\bar{y}_2^{nt}$ for the treated and untreated groups.
  - in the first pre-treatment period compute similar averages $\bar{y}_1^{tr}$ and $\bar{y}_1^{nt}$.
  - then compute $\widehat{\phi} = (\bar{y}_2^{tr} - \bar{y}_1^{tr}) - (\bar{y}_2^{nt} - \bar{y}_1^{nt})$.
- Example average annual earnings
  - for group eligible for treatment are $10,000$ before treatment and $13,000$ after treatment so $\bar{y}_2^{tr} - \bar{y}_1^{tr} = 3,000$
  - for group not eligible for treatment are $15,000$ before treatment and $17,000$ after treatment so $\bar{y}_2^{nt} - \bar{y}_1^{nt} = 2,000$.
  - The DID estimate $\widehat{\phi}$ is then $3,000 - 2,000 = 1,000$.

Selection on unobservables: parametric models (Roy model)

- The Roy model specifies a particular model
  - $y_1 = d$ is the treatment indicator and
  - we observe either $y = y_2$ if $d = 1$ or $y = y_3$ when $d = 0$.
- This allows for both
  - selection on observables (via regressors $\mathbf{x}$)
  - selection on unobservables (ε).

9. Some References

- These references are mainly ones that refer to the recent literature.
- Semiparametric estimation

  - Chen, S. (2010), "An integrated maximum score estimator for a generalized censored quantile regression model," *Journal of Econometrics*, 155, 90-98.
  - Blundell, R. and J.L. Powell (2007), "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, 141, 65–83.
  - Hong, H., and E. Tamer (2003), " Inference in censored models with endogenous regressors," *Econometrica*, 71, 905-932.

- Inverse probability weighting

  - Wooldridge, J.M. (2002), "Inverse probability weighted m-estimators for sample selection, attrition and stratification," *Portuguese Economic Journal*, 1, 117-139.

- Panel data
  - Semykina, A. and J.M. Wooldridge (2010), "Estimating panel data models in the presence of endogeneity and selection," *Journal of Econometrics*, forthcoming.
- Instrumental variables
  - Chernozhukov, V., and C. Hansen (2005), "An IV model for quantile treatment effects," *Econometrica*, 73, 245-261.
  - d'Haultfoeuille, X. (2010), "A New Instrumental method for dealing with endogenous selection," *Journal of Econometrics*, 154, 1-15.
- Treatment Evaluation
  - CT(2005) MMA chapter 25.
  - Angrist, J. D., and J.-S. Pischke (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.
  - Heckman, J.J., J.I. Tobias, and E. Vytlacil (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85, 748-755.
  - Imbens, G. W., and T. Lemieux (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics* 142 (2), 615–35.