Google

Search phrase	Hits in mio.
Generalised least squares	16.7
Matching estimation	3.4
Ordinary leas squares	1.8
Logit	1.1
Tobit	1.0
Cox regression	0.9
GARCH	0.7
Poisson regression	0.4
Propensity score matching	0.4
Kernel regression	0.3
EGARCH	0.03

Matching estimation

Michael Lechner
Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen

Roma, April 2011

Swiss Institute for Empirical Economic Research
University of St.Gallen

Plan for today's lecture

a) Some additional background on matching methods in general

b) Paper on comparing different matching estimators

Please ask questions whenever something is unclear!!!

Papers distributed (incl. slides)

Huber, Lechner, Wunsch (2010),
Frölich (2007)

Frölich, Lechner (2010) [+FL 11]

Lechner (2010)

What is matching?

> Matching is a statistical tool that allows
  - to compare features of the distribution of *Y*
  - in different subsamples defined by an integer *D*
  - such that on a hypothetical distribution of *X* is the same in all subsamples
> After matching, (adjusted) *X* and *D* are independent !
> If mean is target, matching provides weights for a weighted mean of *Y*
  - If weights are applied to *X*, weighted *X* will have same mean (marginal distribution) in all subsamples
> Allows comparison of *Corr(Y,D)* that is free from spurious correlation coming from different distributions of X in the different subsamples

What is propensity score matching?

> *X* may be high dimensional

  - course of dimensionality may become a problem

> Useful property introduced by Rosenbaum and Rubin (1983)

  - assume binary D (can be generalized)

$$p(x) = P(D = 1 \mid X = x) \quad \Rightarrow \quad X \coprod D \mid p(x)$$

> Adjusting distribution of p(x) are enough !

> If model for p(x) is known ➔ dimension reduction!

Where is it used?

Average treatment effect on the treated (ATET) based on CIA

$$Y^1, Y^0 \coprod D \mid X = x \quad \forall x \in \chi$$

$$Y = DY^1 + (1-D)Y^0$$

$$ATET = \theta = E(Y^1 - Y^0 \mid D=1) = E(Y \mid D=1) - \underbrace{E_{X|D=1}[E(Y \mid X = x, D=0) \mid D=1]}_{E_{p(X)|D=1}[E(Y|p(X)=p(x),D=0)|D=1]}$$

> Estimand for ATET:

$$\theta = E[Y \mid D=1] - \int_0^1 E[Y \mid D=0, p(X)=\rho] f_{p(X)|D=1}(\rho) d\rho$$

Where is it used?

IV estimation with need to control for confounders

- IV estimation with X (LATE)
  - Frölich (2007, Journal of Econometrics)

$$p_Z(x) := P(Z = 1 \mid X = x)$$

$$E(Y^1 - Y^0 \mid T = c) = \frac{\int [E(Y \mid p_z(x), Z = 1) - E(Y \mid p_z(x), Z = 0)] f_{p_Z(x)}(x) dx}{\int [E(D \mid p_z(x), Z = 1) - E(D \mid p_z(x), Z = 0)] f_{p_Z(x)}(x) dx}$$

$$E(Y^1 - Y^0 \mid T = c) = \frac{Pscore\ Matching\ estimator\ Z \to Y \quad (ATE)}{Pscore\ Matching\ estimator\ Z \to D \quad (ATE)}$$

  - Application in Frölich and Lechner (2010, JASA)
  - Also applies to continuous instrument

Where is it used?

Difference in difference estimation with confounders

See Lechner (2010) survey

$$E_{X|T=1,D=1}E\ \ Y_t \mid X = x, D = d\ \ =$$

$$= E_{p(X,t,d)|TD=1}E\ \ Y_t \mid p(X,t,D') = p(x,t,d), D = d\ \ ;$$

$$p(X,t,d) := P(TD = 1 \mid X = x, (T,D) \in \{(t,d),(1,1)\}].$$

$$ATET = E_{X|TD=1}\theta(x)$$

$$= E_{p(X,1,1)|TD=1}E\ \ Y \mid p(X,1,1) = p(x,1,1), T = 1, D = 1$$

$$-E_{p(X,0,1)|TD=1}E\ \ Y \mid p(X,0,1) = p(x,0,1), T = 0, D = 1$$

$$-E_{p(X,1,0)|TD=1}E\ \ Y \mid p(X,1,0) = p(x,1,0), T = 1, D = 0$$

$$+E_{p(X,0,0)|TD=1}E\ \ Y \mid p(X,0,0) = p(x,0,0), T = 0, D = 0\ .$$

How to control for many covariates? Reliable estimators for propensity score matching

Coauthors: Martin Huber & Conny Wunsch

HLW 10

Swiss Institute for Empirical Economic Research
University of St.Gallen

Research question of LMW 10 in a nutshell

Paper: Which is the best estimator for balancing covariate distributions in a typical (labour market) policy evaluation in terms of finite sample properties?

Introduction

Matching estimators

> Estimate effect of binary / discrete / continuous interventions (*treatments*)

> Make groups comparable w.r.t. the distribution of observed covariates *X* and compare outcome in the adjusted groups (weighted mean)

> *Selection on observables*: Identify causal effect

> *Selection on unobservables*: Tool to remove observable differences

- DiD(*X*), IV -LATE(*X*): Brief discussion at the end

Introduction

Propensity score matching estimators (1)

> Intuition of *propensity score matching estimator*:
'Make observations comparable' w.r.t. their probability of being observed in 1 of the 2 groups (cond. on X → propensity score)

> Rosenbaum & Rubin (1983) show that in this case observations in the different groups have the same joint distribution of covariates

> Matching estimators are usually implemented semiparametrically

- Propensity score ($P(D=1|X)$) is specified parametrically
- Relation of covariates (X) and treatment (D) to outcomes is left unrestricted (nonparametric)

Introduction

Propensity score matching estimators (2)

> Propensity score matching estimators are an important tool in labour economics (now spreading to other fields as well)

- Effect heterogeneity
- Robustness without curse of dimensionality (because of semiparametric modelling)

Introduction

Propensity score matching estimators (3)

> Many practical properties of estimators not yet clear

- Asymptotics may be a poor guide in semiparametric econometrics
- Finite sample results established by unrealistic Monte Carlo studies
- Many estimators not considered in analytical and MC studies
- (How to perform reliable inference? *different paper*)

Introduction

Key idea of HLW paper

> Our research question:

Which is the best propensity score matching estimator in practice?

(practice := labour market evaluations and beyond)

> Key elements used to obtain the answer

- A large real data set as basis of the Monte Carlo simulations
- Variety of relevant scenarios based on this real data
- Basic & 'optimised' versions of the estimators
- Same trimming rules for all (!) estimators
  (trimming idea: limit impact of single observations on estimator)
- Regressions to summarize the huge amount of information in the MC

Introduction

Key findings

> 'Naïve' versions of all estimators are pretty bad

> Optimising estimators is important

- Trimming is very important for all estimators
- Tuning the matching estimators helps dramatically

> Among the optimized estimators, differences are not large

> Hard to see why the favourites of the Monte Carlos studies conducted so far should be preferred
(Frölich, 2004, Busse, DiNardo, McCrary, 2009)

Introduction

Limitations

> Only parametric propensity scores

- Only feasible option with many regressors

> Monte Carlo design specific for evaluation studies

- Design has many feature that are common in other applications as well
  - binary and (semi-) continuous outcomes and regressors
  - different degrees of selection (share and strength)
  - different sample sizes
  - different types of effects

> Only ATET (computation time)

- Most frequently estimated parameter in practice
- Results for ATENT will be similar
- Unlikely that results will change much for ATE (weighted sum of ATET and ATENT)

Literature

Asymptotic properties of estimators (more details later)

> Some results available

- Based on the (estimated) ***p-score***, IPW estimators are efficient
- The other estimators probably not

Econometrica, Vol. 66, No. 2 (March, 1998), 315–331

ON THE ROLE OF THE PROPENSITY SCORE IN EFFICIENT SEMIPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS

BY JINYONG HAHN[1]

Econometrica, Vol. 71, No. 4 (July, 2003), 1161–1189

EFFICIENT ESTIMATION OF AVERAGE TREATMENT EFFECTS USING THE ESTIMATED PROPENSITY SCORE

BY KEISUKE HIRANO, GUIDO W. IMBENS, AND GEERT RIDDER[1]

Econometrica, Vol. 74, No. 1 (January, 2006), 235–267

LARGE SAMPLE PROPERTIES OF MATCHING ESTIMATORS FOR AVERAGE TREATMENT EFFECTS

BY ALBERTO ABADIE AND GUIDO W. IMBENS[1]

Review of Economic Studies (1998) **65**, 261–294 0034-6527/98/00120261$02.00
© 1998 The Review of Economic Studies Limited

Matching As An Econometric Evaluation Estimator

JAMES J. HECKMAN
University of Chicago

HIDEHIKO ICHIMURA
University of Pittsburgh

and

PETRA TODD
University of Pennsylvania

Literature

Asymptotic properties of estimators (more details later)

> Several estimators used in practise not covered at all

- Radius matching (data dependent radius)
- Matching with parametric adjustment step
  - but: new discussion paper by Abadie & Imbens (NBER, 2009)

> Not clear for what sample sizes and dim(confounders) the asymptotic results are good guides for finite samples

Literature

Finite sample properties: General

> Almost all estimators depend on tuning parameters

- Very little guidance on how to set them in finite samples

> Trimming that affects only the small sample distribution (almost) not considered in the literature other than for IPW

- Usually considered as asymptotic problem
  - Common support problem (Crump, Hotz, Imbens, Mitnik, *2010,* Biometrika; Busso, DiNardo, McCrary, 2009)
  - A infinite-support-of-*X* issue (Khan & Tamer, 2009, mimeo)
- But: The idea already appeared in the survey by Imbens (2004) as a way to ensure common support (but this is a different problem)

Literature

Finite sample properties: Results from Monte Carlo studies

> Disagreement about the best estimator in general

- Local linear ridge kernel matching: Frölich (2004, *REStat)*
- (Normalized) IPW: Busse, DiNardo, McCrady (2009a,b, JBES & mimeo)

Literature

Finite sample properties: Shortcomings of MC studies so far

> Restrictive and artificial DGP's

> Many estimators that are commonly used are not considered (in particular for matching)

- Radius matching
- Matching with (parametric) bias-adjustment step
- Different values of tuning parameters
- Incomplete set of classes of estimators
- Trimming

Literature

Our intended contributions

- Small, medium & larger sample behaviour of many different versions of PS-score-matching-type estimators
- Design of a more realistic Monte Carlo study
- Small sample trimming
- Guidance for applied researchers on estimator choice

- Computer code *will be* made available in Gauss & Stata in the not so distant future

Monte Carlo study

General design of our *real* Mt. Carlo study

> How to construct a realistic design?

> The 5 components of our solution:

- Use very large sample of relevant data as population
  - Here: German data base for training evaluation (>115.000 controls, 3200 treated)
- Understand real selection process by estimating realistic selection equation
  - Here: Use probit with the usual covariates (41) of such an evaluation study
- Generate Mt Carlo sample from nontreated (effect known … =0!) by using estimated conditional selection probabilities to simulate treated
  - Common support ensured by construction
- Use the real outcome variables (and their real unspecified dependence on *X*)
  - Semi-continuous (earnings) and discrete outcomes

Monte Carlo study

General design of our *real* Mt. Carlo study

> How to construct a realistic design?

Our solution:

- Use very large sample of relevant data as population
  - Here: German data base for training evaluation

Monte Carlo study

DGP: The data base (1)

- 2% random sample drawn from the population of all German employees subject to social insurance since 1990 (IEB)
- Participants of vocational class room training
- Nonparticipants do not enter any programme in first 12 months of UE spell
- About 115.000 nonparticipants; 3.200 participants

Monte Carlo study

DGP: The data base (2)

> Outcome variables

- Average earnings in 3 years after programme start for periods of employment (0 if not employed in this period)
- Some non-subsidised employment in 3 years after programme st.

> Confounders (41 variables)

- Socio-demographics (gender, nationality, kids, education, health)
- Information about last employment
- Information about last 10 years before programme
- Regional information

Monte Carlo study

General design of our *real* Mt. Carlo study

> How to construct a realistic design?

Our solution:

- Use very large sample of relevant data as population
  - Here: German data base for training evaluation
- Understand real selection process by estimating a selection equation and use the estimates then for simulation
  - Here: Use probit with the usual covariates of such an evaluation study

Monte Carlo study

DGP: Confounders and the estimated selection model (1)

Descriptive statistics of 'population'

Variable	Treated		Control		Standardized difference	Probit estimation	
	mean	std.	mean	std.	in %	coef.	std. error
Employed	.63	0.56	.48	0.50	9	-	-
Earnings in EUR	1193	1041	1115	1152	9	-	-
Constant term	-	-	-	-	-	-4.90	0.22
Age / 10	3.6	3.5	.84	1.1	8	1.60	0.11
… squared / 1000	1.4	1.4	.63	.85	3	2.01	-0.13
20 - 25	.21	binary	.41		22	0.19	0.04
Women	.57	.46	.50	.50	15	-1.16	0.24
Not German	.11 -	binary	.31		16	-0.13	0.03
Secondary degree	.32	binary	.47		15	0.21	0.02
University entrance qualification	.29	binary	.45		15	0.19	0.02
No vocational degree	.18	binary	.39		26	-0.07	0.03
At least one child in household	.42	binary	.49		22	-0.04	0.03
Last occupation: Non-skilled worker	.14	binary	.35		13	0.07	0.03
Last occupation: Salaried worker	.40	binary	.49		29	0.33	0.03
Last occupation: Part time	.22	binary	.42		12	0.36	0.05
UI benefits: 0	.33	binary	.47		16	-0.14	0.02
> 650 EUR per month	.26	binary	.44		7	0.13	0.03
Last 10 years before UE:	.49	.46	.34	.35	8	-0.30	0.04
share employed							
share unemployed	.06	.05	.11	.11	1	-0.55	0.10
share in programme	.01	.01	.04	.03	9	1.12	0.25
Last years before UE: share minor employment	.07	.03	.23	.14	15	-0.21	0.17
share part time	.16	.11	.33	.29	10	-0.22	0.05
share out-of-the labour force (OLF)	.28	.37	.40	.44	14	-0.30	0.04

Monte Carlo study

DGP: Confounders and the estimated selection model (2)

Variable	Treated		Control		Standardized difference	Probit estimation	
	mean	std.	mean	std.	in %	coef.	std. error
Entering UE in 2000	.26	binary	.44		13	0.29	0.02
2001	.29	binary	.46		5	0.18	0.02
2003	.20	binary	.40		12	0.004	0.03
Share of population close to big city	.76	.73	.35	.37	6	0.09	0.02
Health restrictions	.09	binary	.29		13	-0.15	0.03
Never OLF	.14	binary	.34		6	0.12	0.03
Part time in last 10 years	.35	binary	.48		9	-0.12	0.03
Never employed	.11	binary	.31		17	-0.27	0.05
Duration of last employment > 1 year	.41	binary	.49		4	-0.13	0.02
Average earnings last 10 years when employed / 1000	.59	.52	.41	.40	13	-0.09	0.04
Women x age / 10	2.1	1.7	1.9	1.9	17	0.57	0.13
x squared / 1000	.83	.64	.85	.90	15	-0.57	0.17
x no vocational degree	.09	binary	.28	.36	15	-0.23	0.04
x At least one child in household	.32	binary	.47	.37	25	0.17	0.04
x share minor employment last year	.06	.02	.22	.13	16	0.71	0.18
x share OLF last year	.19	.18	.36	.35	3	0.22	0.05
x average earnings last 10 years when employed	.26	.19	.34	.30	16	-0.23	0.06
x entering UE in 2003	.10	binary	.30		6	-0.14	0.04
$X\hat{\beta}$	-1.7	-2.1	.42	.42	68	-	-
$\Phi(X\hat{\beta})$	.06	0.03	.05	.03	59	-	-
Number of obs., R^2 in %	3266		114349			3.6	

Note:

Monte Carlo study

General design of our *real* Mt. Carlo study

> How to construct a realistic design?

Our solution:

- Use very large sample of relevant data as population
  - Here: German data base for training evaluation
- Understand real selection process by estimating selection equation
  - Here: Use probit with the usual covariates of such an evaluation study
- Generate Mt Carlo sample from nontreated (effect known … =0!) by using estimated conditional selection probabilities
  - common support ensured by construction

Monte Carlo study

General design of our *real* Mt. Carlo study: Summary statistics

Table 3.2: Summary statistic of DGP's

Magnitude of selection	Share of treated in %	Normalized difference of p-score	*Pseudo-R^2* of probit in %	*Sample size considered*
Random	10	0	0	1200, 4800
	50	0	0	300, 1200, 4800
	90	0	0	1200, 4800
Observed	10	0.5	6	1200, 4800
	50	0.4	10	300, 1200, 4800
	90	0.5	6	1200, 4800
Strong	10	1.1	27	1200, 4800
	50	0.8	36	300, 1200, 4800
	90	0.8	27	1200, 4800

9 DGP's

Monte Carlo study

Data generating processes: Sample sizes and replications

> Number of replications is sample size dependent, because smaller samples lead to noisier estimates that are more difficult to pin down

Sample size	Replications	Computation time
300	16000	5-7 hours
1200	4000	1-3 days
4800	1000	2-4 weeks

> Do not use those estimators in the *large sample* that are clearly (!) inefficient in the medium sized sample

Monte Carlo study

Data generating processes: Experimental benchmarks (note)

> RMCS approach is close to 'checking' estimators based on how they can reproduce the results of an experimental control group with an observational group (effect is known in experiments as well, but with noise)

- Lalonde, (1986, AER), Heckman, Ichimura, Smith, Todd (1998, Emetrica), Deheija, Wahba (1999 JASA), Smith & Todd (2005, JEmetrics), etc. etc.

> Advantages of our approach are that

- We mimic the properties of repeated sampling inference (and still retain the plausibility of the DGP)
- The exact truth is known, instead of having a noisy experimental estimate
- CIA holds by construction

> Disadvantage

- True selection process may be different from the specified selection model in a way that would lead to a different performance of the estimators

Monte Carlo study

General design … : Related approaches in the literature

> Bertrand, Duflo, and Mullainathan (2004, QJE)

- DiD: Generate Placebo Laws using US state-year aggregated CPS data

> Diamond and Sekhon (2008, mimeo)

- Gen. matching: Use marginal distribution of *X*, specification of *p(x)*, and conditional expectation of outcome as estimated in LaLonde (1986, AER) data

> Lee and Whang (2009, mimeo)

- Draw samples from National Supported Work (NSW) data to study performance of test for zero conditional treatment effects (two specifications)

Specifications

Propensity score

> Two versions of propensity scores used for all estimators

- Correctly specified (with variables used in simulation of treated)
- Incorrectly specified
  - Omit interaction terms with sex and age nonlinearities
    - 10 significant variables omitted in estimation (misspecification)
  - CIA still valid, but functional misspecification

Estimators

… for average treatment effect on the treated (ATET)

> Estimand for ATET:

$$\theta = E[Y \mid D=1] - \int_0^1 E[Y \mid D=0, p(X)=\rho] f_{p(X)|D=1}(\rho) d\rho$$

> General structure of estimators

$$\hat{\theta} = \frac{1}{N_1} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_0} \sum_{j=1}^{N} (1-D_j) \hat{w}_j Y_j$$

> Estimators differ in how they weight the controls

Estimators

… for average treatment effect on the treated (ATET)

> Main classes of estimators

- Inverse probability weighting (IPW)
- Matching
- Kernel
- Parametric: Tobit, Probit, OLS, double robust

Inverse Probability Weighting (IPW)

Estimation principle

> Weight observations by their probability to be selected

> Inverse probability weighting estimators used for a long time (IPW, Horvitz, Thomson, 1952)

> Busse, DiNardo, McCrary (2009) point to relevance of normalizing the weights

$$w_j \equiv \frac{\frac{\hat{p}(X_j)}{1-\hat{p}(X_j)}}{\frac{1}{N_0}\sum_{l=1}^{N}\frac{1-d_l \quad \hat{p}(X_l)}{1-\hat{p}(X_l)}}; \quad \sum_{j=1}^{N} 1(d_i=0) w_j = 1$$

> Weights might become extreme! ➔ Trimming!!

Trimming (1): A (new) proposal

> Desirable principles (when treatment effects are heterogeneous)

- Do not trim much, as bias (or variance) might become a problem
- Trimming should disappear as sample size increases (no asympt. bias)
- Computationally easy & fast (computation time already excessive)

> Idea: Limit impact of observation with 'a too large weight' on estimate

- Remove observations that tend to 'dominate' the estimator (they lead to an increase in the variability), because they have to too large share of the weights
- Imbens (2004) suggested **T=5%** (without theory)
- Here, **T=6%, 4%** (because large samples are almost not subject to trimming)

$$w_i = \underline{1}\left(w_i / \sum_{j=1}^{N} \underline{1}\ d_i = 0\ \ w_j \leq T\% \right)$$

Swiss Institute for Empirical Economic Research
University of St.Gallen

Trimming (2)

> Implementation

- Use IPW formula to determine trimming level
  - all estimators are trimmed in exactly the same way
  - trimming depends only on the propensity score and the sample size, not on the particular estimator
- Trimming operates only in the subsample of controls (for ATET)
  - Common support adjustments trim in the sample of the treated!

Trimming (3)

Table C.1: Number of deleted non-treated observations for different levels of trimming and different DGP's

Data generating processes			Trimming levels						
Magnitude of selection	Share of treated in %	Sample size	4%	5%	6%	7%	8%	9%	10%
Correct specification of the propensity score									
Random	10	1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	50	300	0.48	0.21	0.10	0.06	0.03	0.02	0.01
		1200	-	-	-	-	-	-	-
		4800	-	-	-	-	-	-	-
	90	1200	0.01	0.00	0.00	0.00	0.00	-	-
		4800	-	-	-	-	-	-	-
Strong	10	1200	0.73	0.41	0.25	0.17	0.11	0.08	0.06
		4800	0.02	0.01	0.01	0.00	0.00	0.00	-
	50	300	3.99	2.86	2.17	1.71	1.40	1.16	0.98
		1200	1.45	0.97	0.68	0.51	0.39	0.30	0.24
		4800	0.34	0.21	0.14	0.10	0.08	0.06	0.05
	90	1200	4.30	3.26	2.57	2.08	1.74	1.48	1.27
		4800	2.34	1.69	1.31	1.04	0.86	0.74	0.62

Only few observations are trimmed
(but they matter a lot!)

Inverse Probability Weighting (IPW)

Asymptotic properties

- Asymptotically efficient estimator based on series estimation (Hirano, Imbens, Ridder, 2003, Emetrica)
- Easy to compute distribution of IPW for parametric propensity score (2-stage GMM à la Newey, 1994, EconLetters)

Inverse Probability Weighting (IPW)

Finite sample properties: Estimators

> IPW best estimator in some simulation studies

(Busse, DiNardo, McCrady, 2009a,b, mimeo)

- Normalisation of weights important

> Other studies (e.g. Frölich, 2004) show that IPW may have very erratic small sample properties

- trimming of very large probabilities probably very important

Estimators

… for average treatment effect on the treated (ATET)

> Main classes of estimators

- IPW
- **Matching**
- Kernel
- Tobit, Probit, OLS
- Double robust

Matching estimators

Estimation principle

> Form explicit comparison group which has the same distribution of *X* as the treatment group

- Compute mean of *Y* in this 'matched comparison group'

> Do this pairing / matching by finding for every treated one or several non-treated with the same (or similar) values of *X*

> Here we consider only 'matching-with-replacement'

- Same control observations may be used many times
- Only possibility if treatment shares > 50%

> Benchmark: 1 : 1 matching $w_j \equiv \frac{N_0}{N_1} \sum_{i:D_i=1} 1 \min \| \hat{p}(X_j) - \hat{p}(X_i) \|$

Matching estimators

Asymptotic properties

> 1:M matching estimators on X (Abadie, Imbens, 2006, Emetrica)

- Consistent, inefficient, asymptotically normally distributed, asymptotically biased if dim(X) > 1
- Bias can be removed by nonparametric regression

> 1:M matching estimators based on parametrically estimated propensity score (Abadie, Imbens, 2009, NBER WP)

- Consistent, asymptotically normally distributed, inefficient
- More precise than using true propensity score

Matching estimators

'Real' matching estimators considered

> Issues

- Asymptotic bias ➔ regression adjustment
- Inefficient , too few controls ➔ radius matching (3 different radi)
- Inefficient (p-score, not X) ➔ additional covariates (sex, prev. earnings) with Mahalanobis matching

> Combine all these features

> Additionally, logit instead of linear regression for binary outcomes

> 48 (x2) estimators for employment, 32 (x2) for earnings

Matching estimators

Monte Carlo performance: Summary

> Radius matching better than nearest neighbour matching
  - In particular for the smaller samples
  - Low sensitivity to choice of radius (data driven algorithm not bad)

> Regression adjustments
  - Non-linear or linear regression reduces bias and adds variance compared to unadjusted radius matching with same radius (bias reduction dominates)
  - Logit adjustments (for employment) leads to better estimator in *smaller samples*
  - Linear adjustment works as good as logit in *larger samples*

> Additional variables in 1st step (Mahalanobis matching) reduce RMSE

> Using linear index instead of p-score leads to some RMSE reductions in smaller sample (no difference for large sample)

> Nearest neighbour m. has on average >50% larger RMSE than best methods

Estimators

… for average treatment effect on the treated (ATET)

> Main classes of estimators

- IPW
- Matching
- **Kernel**
- Tobit, Probit, OLS
- Double robust

Kernel matching

Estimation principle

> Estimate $E(Y|X=x, D=0)$ by nonparametric kernel methods

- could also be done for E(Y|X=x, D=1) (not considered here)

> Average this regression function with respect to the distribution of $X|D=1$

Kernel matching

Asymptotic properties

> Kernel estimator that are 'asymptotically linear with trimming' (dim(kernel) at least as large as dim(X))
  - consistent and asymptotically normally distributed (no asymptotic bias);
  - not necessarily efficient if p-score used
  - see Heckman, Ichimura, Todd (1998, REStud)

> Series estimators combined with IPW
  - consistent and efficient
  - Imbens, Newey, Ridder (2007)

> Hahn (1998, Emetrica) proposes alternative regressions
  - consistent and efficient

Kernel matching

Estimators used (1)

> Use only 'winner' of Frölich (2004, REStat): *Local (non-)linear Ridge regression* (Seiffert & Gasser, 1996)

- Idea-I: Use local regression instead of just local mean as in NW
- Idea-II: Use Ridge term in denominator to stabilise estimator in small samples (ridge term disappears for large samples)

$$\hat{\theta}_{ridge} = \frac{1}{N_1}\sum_{i=1}^{N} d_i \cdot \left[Y_i - \hat{m}_0(\hat{p}(X_i)) \right]$$

$$\hat{m}_0(\hat{p}(X_i)) = \frac{A_0(\hat{p}(X_i))}{B_0(\hat{p}(X_i))} + \frac{A_1(\hat{p}(X_i)) \cdot (\hat{p}(X_j) - \bar{p}(X_i))}{B_1(\hat{p}(X_i)) + r \cdot h \,|\, \hat{p}(X_j) - \bar{p}(X_i)|}$$

$$A_a(\hat{p}(X_i)) = \sum_{j:D_j=0}^{n} Y_j \cdot \left[\hat{p}(X_j) - \bar{p}(X_i) \right]^a \cdot K\left(\frac{\hat{p}(X_j) - \hat{p}(X_i)}{h} \right)$$

$$\bar{p}(X_i) = \sum_{j:D_j=0}^{n} \hat{p}(X_j) \cdot K\left(\frac{\hat{p}(X_j) - \hat{p}(X_i)}{h} \right) \Bigg/ \sum_{j:D_j=0}^{n} K\left(\frac{\hat{p}(X_j) - \hat{p}(X_i)}{h} \right)$$

$$B_a(\hat{p}(X_i)) = \sum_{j:D_j=0}^{n} \left[\hat{p}(X_j) - \bar{p}(X_i) \right]^a \cdot K\left(\frac{\hat{p}(X_j) - \hat{p}(X_i)}{h} \right)$$

K(): Epanechnikov kernel

Kernel matching

Estimators used (2)

> Local linear vs. local logit (both used for employment)

> Bandwidth choice

- Cross validation for conditional mean function (100%, 33%, 300% of CV value)
- Rule of thumb

> 8 (x2) estimators for employment, 4 (x2) estimators for earnings

Kernel estimators

Monte Carlo performance: Summary

> Bandwidth appears not to matter much for RMSE

- Based on varying the multiplier of the cross-validation bandwidth

> No gain by using local logits instead of local linear regression

> Larger bandwidth leads to superior distributional properties

- Surprising as we would expect that undersmoothing should perform better, as the CV used is the one for estimating E(y|X) not E[E(Y|X)]

Estimators

… for average treatment effect on the treated (ATET)

> Main classes of estimators

- IPW
- Matching
- Kernel
- **Tobit, Probit, OLS**
- Double robust

Tobit, Probit, OLS

Estimation principles

> Use independent variables of propensity score as regressors

> Estimate regressions in ***subsamples*** of (treated and) control

> Estimators used

- OLS
- Tobit (in Heckit version) for earnings (without exclusion restriction)
- Probit for employment

> 2 (x2) estimators for employment, 2 (x2) for earnings

Tobit, Probit, OLS

Asymptotic properties & Mt Carlo performance

> Asymptotic properties of parametric regression models

- Asymptotically efficient if regression model is correctly specified
- Otherwise, inconsistent

> Monte Carlo results *(well known from other studies)*

- OLS dominated by Probit for binary outcome
- Heckit very unstable particularly in small samples

Estimators

… for average treatment effect on the treated (ATET)

> Main classes of estimators

- IPW
- Matching
- Kernel
- Tobit, Probit, OLS
- **Double robust**

Double robust

Estimation principle

- Combine parametric regression modelling with IPW weighting
- Robins and Rotnitzky (1995), Scharfstein, Rotnitzky, and Robins (1999), etc.
- Efficient if both models are correctly specified
- Consistent and asymptotically normally distributed if either p-score ***or*** regression model is correct (double robustness, Robins and various co-authors)

Double robust

Implementation

> Use independent variables of propensity score as regressors

> Estimate regressions in subsamples of treated and control separately

> OLS for both outcomes (misspecified)

> Tobit (in Heckit version) for earnings, probit for employment

> Combine with IPW weighting

  - Instead of weighting use p-score as additional regressor

> 2 (x2) estimators for employment and 2 (x2) estimators for earnings

Parametric and DR

Monte Carlo Performance: Summary

- OLS (earnings) and Probit (employment) best in smaller samples
- Heckit (earnings) without exclusion restriction
  - Disastrous in small samples
  - Better performance in large sample
- Robust:
  - Worst performance of all estimators in small sample when weights are not trimmed
  - Heckit performs very bad in small sample, but OLS is ok
- No gain (for nonlinear models) if regressions also performed in treated population

Comparative Monte Carlo Results

Features of DGP, trimming, and specification: Summary (1)

> Misspecified p-score

- Fewer (too few) control variables lead to some precision gain in small sample (even for IPW)
- Bias in larger sample

> Selectivity

- The stronger the selection, the less precise the semiparametric estimators
- Parametric estimator least precise without selection (?)

Comparative Monte Carlo Results

Features of DGP and trimming: Summary (2)

> Share of treated

- All semiparametric estimators most precise if share of treated is not larger than share of controls
- Not so clear for parametric estimators

> **Trimming**

- All estimators more precise if trimmed (incl. parametric)
- Small sample: Largest gains for IPW
- Medium sample: Similar (and largest) gains for IPW and matching

Comparative Monte Carlo Results

The competition of the estimators: Choosing a subset

> Goal: Reduce number of estimators to manageable quantity by a pre-selection within any particular 'class of estimators'

> Use RMSE criterion only

- if mean absolute deviation deviates from RMSE (rarely the case), inference will be difficult (check, but do not use for selection)

> A 'good' estimator in its class should be

- among the best in a majority (> 50%) of cases
  - *best* defined as not more than 25% higher RMSE than best *in class*
- never be among the worst estimators
  - *worst* defined as more than double the RMSE of best in class

Trimming for selected estimators

Table 5.3: Comparison of the properties of the selected estimators: trimming

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Propensity score correctly specified														
Without trimming														
RelRMSE	39	16									36	201	62	144
Bias	0.5	1.0									23	5	29	9
Std. dev.	5.1	4.1									117	178	137	216
Skew.	0.1	0.0									-0.4	-0.2	-2.8	-2.8
Kurtosis	3.4	3.0									7.0	3.4	172	174
RelRMSE	11	9									12	73	3	21
Bias	0.3	0.7									10	4	23	6
Std. dev.	4.1	3.9									98	153	86	108
Skew.	0.0	0.0	0.0	0.1	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.3	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.3	3.0	3.0	3.1	3.1	3.5	7.5	3.6	5.2	7.4
Trimming level 4%														
RelRMSE	7	7	14	7	61	B	7	7	4	18	4	63	B	15
Bias	0.2	0.7	1.0	0.7	0.1	0.5	0.5	6	19	27	8	3	22	5
Std. dev.	3.9	3.9	3.9	3.9	5.9	3.6	3.9	94	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	3.1	3.4	8.2	3.7	4.5	7.0

No trimming → 6%: Big gain in RMSE
Small **gain** in bias
Reduced kurtosis

6% → 4% Not much change!

Comparative Monte Carlo Results

The competition of the estimators: The subset fulfilling the criteria

> Matching
  - Only radius matching estimators remain

> Kernel
  - All trimmed estimators fairly close, but some untrimmed (with larger bandwidth) ok as well

> Parametric
  - Employment: All estimators remain
  - Earnings: Only OLS remains

> Very few untrimmed estimators remain

> More estimators remain for employment than for earnings

Comparative Monte Carlo Results

The competition of the estimators: The 6 final contenders

> From these estimators a subset is chosen (for the final presentation of results)
  - by ignoring *similar* estimators which are usually inferior

> Trimmed estimators always better than untrimmed ➔ trimmed only
  - Relative performance is different for trimmed & untrimmed estimators

> Matching
  - Radius (large) matching on linear index & some *X* (Mahalanobis) & logit (employment) / linear regression (earnings) – best or close to best match. estimator
  - Nearest neighbour matching as reference case

> Kernel
  - Employment: Local linear usually slightly better than local logit → local linear only
  - Small and large bandwidths

> Probit (employment) and OLS (earnings)

Comparative Monte Carlo Results

The competition of the estimators: Overall comparison

Table 5.5: Comparison of the properties of the selected estimators: other features

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Correctly specified propensity score														
RelRMSE	7	7	14	7	61	B	7	7	4	18	4	63	B	15
Bias	0.2	0.7	1.0	0.7	0.1	0.5	0.5	6	19	27	8	3	22	5
Std. dev.	3.9	3.9	3.9	3.9	5.9	3.6	3.9	94	88	96	92	145	84	101
Skew.	0.0	0.0	0.0	0.0	0.2	0.0	0.0	-0.1	0.0	-0.1	-0.2	-0.2	0.0	-0.1
Kurtosis	3.0	3.0	3.4	3.0	3.4	3.0	3.1	3.1	3.1	3.4	8.2	3.7	4.5	7.0
Misspecified propensity score														
RelRMSE	18	22	14	B	51	5	7	9	13	7	B	39	8	9
Bias	2.7	2.9	2.4	1.5	2.6	2.1	2.1	66	73	62	53	63	70	62
Std. dev.	3.6	3.6	3.6	3.7	5.3	3.5	3.6	88	85	87	87	129	81	93
Skew.	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.5	-0.2	-0.1	-0.1
Kurtosis	3.0	3.0	3.0	3.0	2.9	3.0	3.0	3.1	3.1	3.2	13.5	3.2	4.6	5.5

Comparative Monte Carlo Results

The competition of the estimators: Overall comparison

> Bias adjusted Matching and parametric regression are best

- OLS/Probit (for controls) are best for the correctly specified model
- Radius matching with bias adjustment is best for the incorrectly specified model (more robust)

> Both have excess kurtosis for earnings

> The other estimators (but NN) are not far away (RMSE)

> Next, consider more detailed results …

Comparative Monte Carlo Results

The competition of the estimators: Correct specification

Table 5.4: Comparison of the properties of the selected estimators: sample size

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
N = 300 (correctly specified score; 50% treated)**														
RelRMSE	1	2	2	5	62	7	10	2	1	3	6	60	0.1	12
Bias	0.3	0.2	0.2	0.6	0.1	1.4	1.4	6	3	2	10	3	9	1
Std. dev.	5.8	5.8	5.9	5.9	9.3	6.0	6.1	136	135	139	142	214	134	150
Skew.	0.1	0.0	0.1	0.1	0.0	0.0	0.0	-0.1	0.0	-0.1	-0.5	-0.5	-0.1	0.2
Kurtosis	3.0	3.1	3.1	3.0	6.6	2.9	2.9	3.0	3.1	3.2	11.9	6.2	4.3	10.9
N = 1200 (correctly specified score; 50% treated)														
RelRMSE	14	14	25	11	67	B	13	15	9	23	1	77	B	11
Bias	0.2	0.3	0.8	0.7	0.2	0.2	0.1	5	7	15	5	1	18	5
Std. dev.	3.3	3.3	3.5	3.1	4.8	2.9	3.3	81	76	85	71	125	66	78
Skew.	0.1	0.0	0.3	0.1	0.2	0.1	0.1	-0.1	0.0	0.1	0.0	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.5	2.9	3.0	3.0	3.0	3.0	3.0	3.5	2.9	3.3	3.0	3.1
N = 4800 (correctly specified score; 50% treated)														
RelRMSE	22	25	40	15	74	B	19	29	15	48	B	87	17	28
Bias	0.1	0.5	0.2	0.5	0.1	0.1	0.1	1	10	11	3	3	21	8
Std. dev.	1.7	1.7	2.0	1.5	2.5	1.4	1.7	46	40	52	36	67	33	45
Skew.	0.0	0.0	0.1	0.0	0.2	0.0	0.0	-0.1	0.0	0.1	0.0	-0.1	0.0	0.0
Kurtosis	3.0	2.9	2.9	3.0	3.0	2.9	2.9	3.1	3.1	3.0	3.2	3.0	3.1	3.4

Note: RelRMSE: Difference in relative root mean squared error in % compared to best estimator. Bias and standard

Comparative Monte Carlo Results

The competition of the estimators: Correct specification

Table 5.4: Comparison of the properties of the selected estimators: sample size

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Selection: Random														
RelRMSE	0.5	4	0.1*	7	48	2	2	0.5	3	0.1*	11	49	3	5
Bias	0.0	0.1	0.1	0.4	0.1	0.3	0.2	1	2	2	4	2	1	1
Std. dev.	3.1	3.2	3.1	3.2	4.5	3.1	3.1	70	72	70	78	104	72	74
Skew.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.3	-0.1	-0.2	-0.1
Kurtosis	3.0	3.0	3.0	3.0	3.3	3.0	3.0	3.1	3.1	3.1	17.1	3.4	5.7	6.3
Selection: Normal														
RelRMSE	6	5	5	B	46	1	3	4	3	5	B	45	1	6
Bias	1.3	1.6	1.4	0.9	1.2	1.2	1.2	31	43	37	26	28	39	31
Std. dev.	3.5	3.4	3.5	3.6	5.1	3.5	3.5	88	82	87	87	129	82	90
Skew.	0.0	0.0	0.0	0.0	0.1	0.0	0.0	-0.1	0.0	-0.1	-0.6	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.1	3.0	3.2	3.1	3.0	3.0	3.0	3.3	12.3	3.3	3.3	4.0
Selection: Strong														
RelRMSE	22	25	24	B	62	0.2	9	17	19	25	B	57	10	21
Bias	3.1	3.6	3.7	2.0	2.8	2.5	2.4	76	92	95	62	69	98	69
Std. dev.	4.7	4.5	4.7	4.5	7.1	4.1	4.6	116	106	119	103	178	93	128
Skew.	0.1	0.0	0.0	0.1	0.2	0.1	0.1	-0.2	0.0	-0.2	-0.1	-0.4	0.0	-0.2
Kurtosis	3.0	3.0	3.6	3.0	3.0	3.0	3.1	3.2	3.1	3.6	3.2	3.8	4.6	8.5

Comparative Monte Carlo Results

The competition of the estimators: Correct specification

Table 5.4: Comparison of the properties of the selected estimators: sample size

	Employment							Earnings						
	IPW	Kernel		Matching		Probit		IPW	Kernel		Matching		OLS	
		high	low	logit	pair		DR		high	low	OLS	pair		DR
Share of treated: 10%														
RelRMSE	10	20	10	10	53	0.1*	5	0.1	10	6	4	45	0.2*	4
Bias	1.1	1.7	1.3	1.0	1.1	0.8	0.8	24	39	36	27	24	38	27
Std. dev.	3.4	3.5	3.4	3.5	4.9	3.2	3.3	83	87	84	85	126	77	85
Skew.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.1	0.0	0.0
Kurtosis	3.0	3.0	3.0	2.9	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.1	3.1
Share of treated: 50%														
RelRMSE	12	12	15	B	53	2	7	11	8	13	B	53	7	13
Bias	1.4	1.4	1.4	0.9	1.3	1.3	1.2	34	34	35	27	32	43	33
Std. dev.	3.4	3.5	3.6	3.4	5.0	3.2	3.4	84	82	86	79	125	75	87
Skew.	0.0	0.1	0.1	0.1	0.1	0.0	0.1	-0.1	0.0	0.0	-0.1	-0.2	0.0	0.0
Kurtosis	3.0	3.0	3.1	3.0	3.4	3.0	3.0	3.1	3.1	3.2	4.6	3.5	3.3	5.0
Share of treated: 90%														
RelRMSE	18	16	18	B	56	6	9	8	8	12	B	41	3	7
Bias	2.2	2.6	2.7	1.4	2.0	2.0	1.9	55	74	69	41	50	59	46
Std. dev.	4.1	3.6	3.8	3.9	6.0	3.8	4.1	97	80	95	98	143	87	103
Skew.	0.0	0.0	-0.2	0.0	0.2	0.0	0.0	-0.1	0.0	-0.4	-1.1	-0.3	-0.2	-0.3
Kurtosis	3.1	3.0	3.6	3.1	2.9	3.2	3.2	3.1	3.1	3.5	30.4	3.6	6.6	6.9

Comparative Monte Carlo Results

Estimators in detail: Inverse probability weighting (IPW)

- Usually not the best
- Least biased in correctly specified model
- Lack of robustness to misspecification of p-score
- Excessive RMSE in a few cases (+30%)
- Excess kurtosis in a few cases
- May need even more trimming (more sensitive to outliers than the other estimators?)

Comparative Monte Carlo Results

Kernel matching

> Usually not the best, and not the worst estimator within the subset of 'good' estimators

> Larger bandwidth gives more stable results

> Low bandwidth has bad performance in a few cases

- high RMSE (+40%)
- excess kurtosis

Comparative Monte Carlo Results

Matching

- Nearest neighbour matching should not be used
- Bias adjustment by logit or OLS reduces bias and adds variance
  - Bias reduction is dominating compared to radius matching
  - Without regression adjustment radius matching does have bias that is too large to be highly competitive
- Regression (logit/OLS) adjusted matching is in almost all cases
  - The best estimator by RMSE, or very close to the best estimator
  - Low bias (*in particular for misspecified p-score*) ➔ robustness
  - Employment:
    - excess kurtosis in smaller samples (inference?)
    - no problem for radius matching without regression (but too large RMSE)

Comparative Monte Carlo Results

Probit and OLS (in subsample of non-treated)

> Competitive (if trimmed), if sample is not too large
  - Almost always smallest variance
  - ***OLS: Bias becomes dominant when N increases***

> Never the worst estimator (but it will be for some large enough *N*)

> Employment
  - Excess kurtosis for smaller samples

> Performance similar to matching with regression adjustment
  - but with smaller variance and larger bias
  - For larger *N*, OLS becomes uncompetitive because of the bias
  - Probit almost unbiased in our setting

> Unattractive because estimators are inconsistent!

Comparative Monte Carlo Results

Which is the best estimator? (1)

> Only trimmed estimators relevant!

- Optimal trimming level ?

  (not much further gain when moving from 6 to 4%)?

Comparative Monte Carlo Results

Which is the best estimator? (2)

> Bias-adjusted matching estimator seems to win the competition

- Best in RMSE in many cases
- Fairly robust to functional misspecification of p-score
- Excess kurtosis in smaller samples for continuous outcome

> Parametric estimators (in subsample of treated) are most precise

- Best in RMSE in several cases
- Excess kurtosis in smaller samples (needs larger samples)
- Inconsistent (bias dominates in larger samples)

Comparative Monte Carlo Results

Which is the best estimator? (3)

> IPW is the most easy to compute

- It does not perform badly on average, but sometimes it has a too large RMSE
- Trimming seems to be the most important issue here
- There is also a lack of robustness to misspecification of *p(x)*

> Kernel-ridge regression with large bandwidth also an important competitor

- Never really top, never really bad, but always close to the top
- Distribution always normal

Conclusions

> Non-optimized estimators perform badly
  - Bias adjustment for matching estimators
  - Ridge regressions for kernel matching
  - All estimators should be trimmed (incl. probit and OLS)

> No estimator is superior in all scenarios

> IPW is asymptotically efficient, but never the best in finite samples

> Optimized matching is *probably* the one to go for in an application
  - Good large control sample properties (incl. consistency)
  - Very good in the smaller control samples

Further research

> Investigate trimming more thoroughly

- How to chose optimal level?
- Any role for estimator specific trimming levels?

> Inference!

- Next paper!

Thank you for your attention!

Michael Lechner
SEW – University of St. Gallen
April 2011

Swiss Institute for
Empirical Economic Research
University of St.Gallen