pscore2: Stata module to enforce covariate balance

Sabrina Dorn

ETH Zurich

UK Stata User Group Meeting, London, September 2012

Outline of points covered

1. Introduction: Short literature review, ATT by stratification on the propensity score, simulation study about limitations of current implementation
2. The `pscore2` algorithm: What is it doing? How does it work?
3. Stata implementation: The command `pscore2`, options, output
4. Examples: NSW-PSID1 data example from Dehejia and Wahba (2002), Fixed Currency Regimes and the Pattern of Time (Dorn and Egger, 2012), Simulation studies on behavior of `pscore2` for different cutoff-levels and in presence of omitted variables and/or nonlinearity
5. Conclusion: There are efficiency gains!!

A short genesis

How to find good comparisons for treatment evaluation?

- The central role of the propensity score in observational studies for causal effects (Rosenbaum and Rubin, 1983)
- Subclassification on the propensity score to reduce the bias of the estimated treatment effect (Rosenbaum and Rubin, 1984)
- Dehejia and Wahba (2002) propose an algorithm to implement subclassification on the propensity score
- Becker and Ichino (2002) provide the Stata implementation `pscore`
- Newer, data-driven approaches e.g., Diamnond and Sekhon's (2012) genetic matching (`GenMatch` in `R`)

Background

- Rubin causal model: $Y = Y^1 T + Y^0(1-T),\ T \in \{0,1\}$
- Y^0 is only observed if $T = 0$, but we want to infer treated subjects' counterfactual outcome
- Parameter of interest could be ATT: $\gamma = \mathbb{E}[Y^1 - Y^0 | T = 1]$
- Though $\mathbb{E}[Y^0|T=0] \neq \mathbb{E}[Y^0|T=1]$ (not mean independent), we can condition on $\boldsymbol{X} = \boldsymbol{x}$ to restore mean independence: $\mathbb{E}[Y^0|T=0, \boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}[Y^0|\boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}[Y^0|T=1, \boldsymbol{X}=\boldsymbol{x}]$
- ATT can be inferred from: $\gamma = \mathbb{E}_{\boldsymbol{X}}\{\mathbb{E}[Y^1 - Y^0 | T=1, \boldsymbol{X}=\boldsymbol{x}]\}$
- *Balancing score*: Iff $\mathbb{E}[Y^0|T=0, \boldsymbol{X}=\boldsymbol{x}] = \mathbb{E}[Y^0|\boldsymbol{X}=\boldsymbol{x}]$, then $\mathbb{E}[Y^0|T=0, \pi(\boldsymbol{X}=\boldsymbol{x})] = \mathbb{E}[Y^0|\pi(\boldsymbol{X}=\boldsymbol{x})]$, $\pi(\boldsymbol{X}=\boldsymbol{x}) : \mathbb{R}^K \to \mathbb{R}$

A prototypical situation

- Suppose one is interested in estimating γ (ATT) using stratification on the propensity score (`atts` and variations on the theme)
- Situation: `pscore` concludes that the balancing property is not satisfied
- Suggestion Dehejia and Wahba (2002), p. 161:
  Algorithm step 4.c.: If a covariate is not balanced for many strata, modify the logit [balancing score model] by adding interaction terms and/or higher-order terms of the covariate and reevaluate.
- Question: Does this really solve the problem?
  $\longrightarrow$ in terms of MSE($\hat{\gamma}$)?
  $\longrightarrow$ in terms of times the null hypothesis of balancing is rejected?

Simulation study

- I simulate $R = 10.000$ samples of sample size $N = 400$ from the model: $T = 1[\boldsymbol{X}\beta_0 + \epsilon_0 > 0]$ and $Y = \boldsymbol{X}\beta_1 + \gamma T + \epsilon_1$ where it is assumed that $(\epsilon_0, \epsilon_1)^T \sim \mathcal{N}(\mathbf{0}, \mathrm{I})$, $(X_1, Z_2)^T \sim \mathcal{N}(\mathbf{0}, \Sigma)$, and $X_2 = 1[Z_2 > 0]$.
- For each simulation $j = 1, \ldots, R$, I estimate $\hat{\gamma}$ using `pscore` followed by `atts`
- I vary the type I error α in $\{0.01, 0.05, 0.1\}$ and collect information on MSE and the number of cases when `pscore` reports failure of the balancing property
- Results: Left: using correct specification, right: second order interactions added

α	MSE($\hat{\gamma}$)			reject H_0	MSE($\hat{\gamma}$)			reject H_0
0.01	0.232	0.177	0.227	9.55%	0.250	0.203	0.247	6.53%
0.05	0.256	0.215	0.249	17.43%	0.276	0.221	0.265	18.55%
0.1	0.268	0.233	0.257	30.89%	0.286	0.253	0.274	34.54%

Can pscore2 beat this?

- YES: Left: `pscore2`, right: decrease MSE($\hat{\gamma}$) relative to `pscore`

α	MSE($\hat{\gamma}$)	ΔMSE
0.01	0.140	-0.087 (-38.83%)
0.05	0.124	-0.125 (-50.20%)
0.1	0.098	-0.159 (-61.87%)
0.2	0.071	-0.195 (-73.31%)
0.3	0.059	-0.214 (-78.39%)

- HOW: `pscore2` *enforces* covariate balance on the one hand, and automatically *discards* bad comparisons from the analysis on the other hand
- WHY does this work? The `pscore2` algorithm considers *sufficient conditions* regarding each of the marginal covariate distributions and uses a grid search procedure to map the according partitions into regions of the balancing score

Building blocks of pscore2 algorithm

- Instead of pre-assigning the locations of strata from outside of the model, `pscore2` estimates them from the data *subject to covariate balance*
- In doing so, `pscore2` looks for similar treated and controls by checking each regressor's *marginal distribution* for balancing
- At the same time, *bad comparisons* are identified from the data
- Searching along the balancing score function, reduces the problem to *segments on* $(0; 1)$

$\rightarrow$ Idea of clustering into strict partitions with outliers; similarities to the ideas in Dehejia and Whaba (2002), Becker and Ichino (2002)

Conceptual advantages

- Balancing can be enforced to greatest possible extend given the data
- If observations are not comparable, they are at odds with the model assumptions, and hence should be identified from the data
- But there is a *trade-off*, since the amount of *discarded observations* should not be overly excessive (level of *type I error* for the tests controls for this)
- `pscore2` compares *shrinking* partitions of covariates along $(0; 1)$
- Shrinking means that the partition of the data used for the hypothesis tests is getting smaller until the test breaks down due to a lack of observations (not variation)

The pscore2 algorithm

1. Estimate $\hat{\pi}(\cdot) = \hat{\mathbb{P}}[T = 1|X = x]$ with T the treatment indicator and X data on $k = 1, \ldots, K$ variables
2. Initializing step of pscore2:

   Find the first largest partition of the line segment connecting $[\min \hat{\pi}(\cdot), \max \hat{\pi}(\cdot)]$ *where each of the marginal distributions for the* x_k*'s satisfies* $\mathbb{P}[t(x_k^0) = t(x_k^1)|H_0] > \alpha$.

   - Initialize testing interval: $\lambda^+ = \max\{\hat{\pi}(\cdot)\}$, $\lambda_0^- = \min\{\hat{\pi}(\cdot)\}$
   - Update testing interval: $\tilde{\lambda}^+ = (\lambda^+ - \lambda_0^-)/s$, $s = 1, 2, 3, \ldots$
   - Until: either criterion is satisfied or inference impossible $\rightarrow \lambda_1^- = \tilde{\lambda}^+$

The pscore2 algorithm

3. Update step of pscore2:

   Find the next largest partition of the line segment connecting $[\lambda_r^-, \lambda^+]$ *where each of the marginal distributions for the* x_k*'s satisfies* $\mathbb{P}[t(x_k^0) = t(x_k^1)|H_0] > \alpha$.

   - Update according to: $\tilde{\lambda}^+ = (\lambda^+ - \lambda_r^-)/s$, $s = 1, 2, 3, \ldots,$ $r = 2, \ldots, \hat{R}$
   - ...$\rightarrow \lambda_{r+1}^- = \tilde{\lambda}^+$

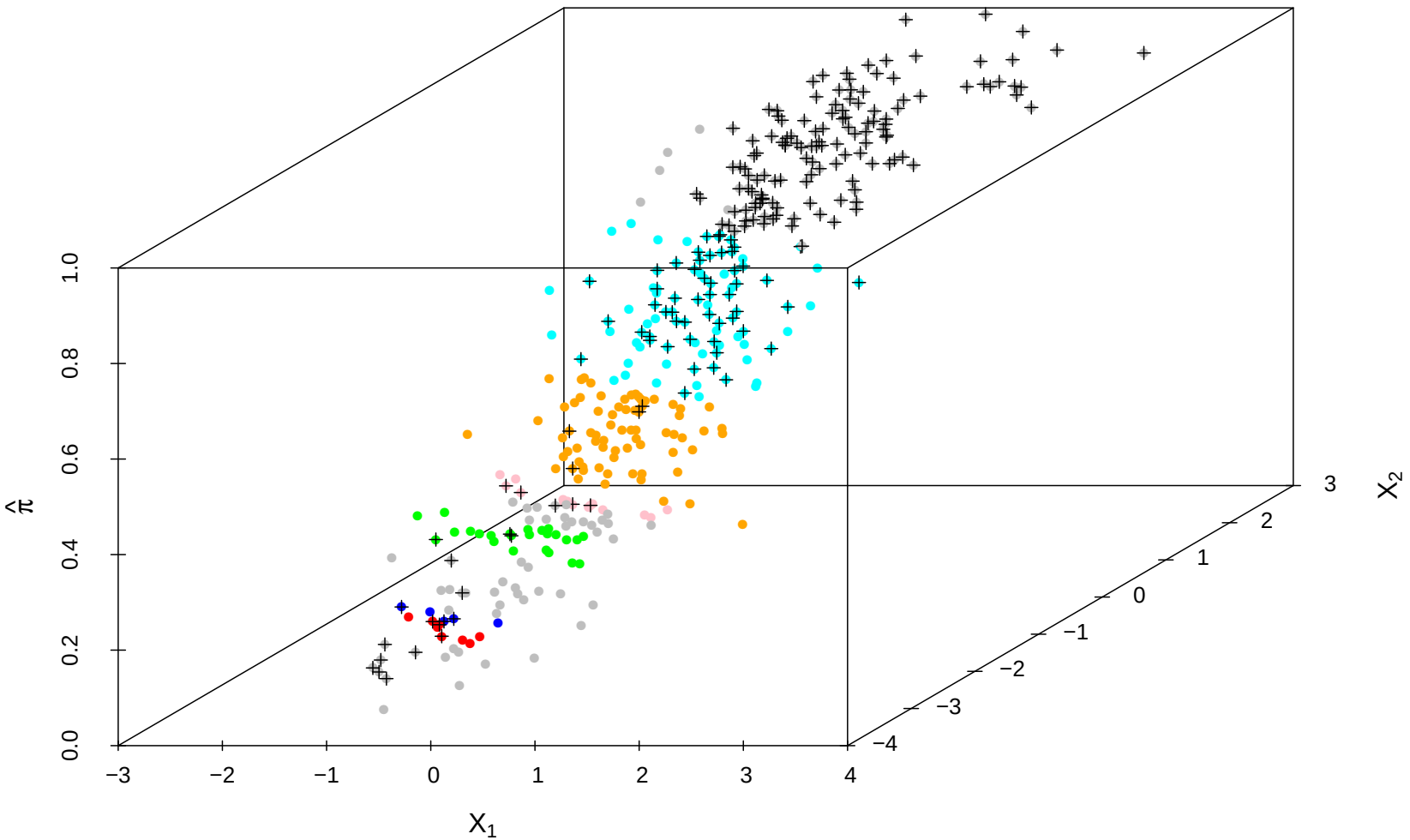
4. Iterate through step 3 until $\lambda_{r+1}^- = \lambda^+$

$\rightarrow [\lambda_0^-, \lambda_1^-), [\lambda_1^-, \lambda_2^-), \ldots, [\lambda_{r-1}^-, \lambda_r^-), [\lambda_r^-, \lambda^+]$

5. Finally, discard all intervals where balancing could not be achieved

Visualization in 3D

Simulated data: $T = 1[X_1\beta_{01} + X_2\beta_{02} + X_1X_2\beta_{03} + \epsilon_0 > 0]$



Stata syntax

pscore2 *treatment* [*myscore*] *indepvars* [*if*] [*in*] [*weight*] ,
blockid(*newvar1*) pscore(*newvar2*) [revert logit supplied
comsup wilk median tenforce ksmirnov variance level(#)
detail summary]

Options

- Compulsory options:
  - `blockid(`*newvar1*`)`: Variable name for strata identifier
  - `pscore(`*newvar2*`)`: Variable name for balancing score according to *newvar1*
- Balancing score options:
  - Default: A probit model is estimated internally
  - `supplied`: In this case, the balancing score is supplied externally; if this is specified, the name of the externally supplied balancing score has to be specified as the second element in *varlist*
  - `logit`: Use a logistic regression model to estimate the propensity score internally
  - `comsup`: Restrict computations to common support

Options

- Options how to compare marginal distributions:
  - Default: `pscore2` uses `ttest` for continuous regressors and calls `ranksum` for binary regressors
  - `wilk`: `pscore2` calls `ranksum` for all variables
  - `median`: `pscore2` calls `median` instead of `ttest` for continuous regressors
  - `tenforce`: Compute `ttest` for all variables
  - `ksmirnov`: `pscore2` will use Kolmogorov-Smirnov equality of distributions test
  - `variance`: `pscore2` tests for equal means and variances of each regressor

Options

- Options for algorithm:
  - Default: If nothing is specified, `pscore2` searches into the direction of the minimum estimated propensity score (fixes λ^+ from above) and the default type I error of 0.1 is used
  - `revert`: Search direction to the maximum propensity score (i.e., now fix λ^-)
  - `level(#)`: specifies the desired level of the type I error for the tests
- Summary options:
  - `summary`: If specified, a detailed summary of the p-values and the tests conducted to estimate the strata is displayed
  - `detail`: `pscore2` reports the estimation output of the internally estimated propensity score model or displays a detailed summary of the externally supplied variable

Sample output

```
**************************************************
Propensity score model
**************************************************
Note: The common support option has been selected
=> The region of common support is [.0003, .972]
Note: Searching in direction of minimum propensity score
**************************************************
Initializing and computing grid search
**************************************************
Interval 1 complete
(convergence not achieved - truncating interval)
Interval 2 complete
(convergence achieved)
Interval 3 complete
(convergence achieved)
Interval 4 complete
(convergence achieved)
Interval 5 complete
(convergence achieved)
Interval 6 complete
(convergence achieved)
Interval 7 complete
(convergence achieved)
Interval 8 complete
(convergence achieved)
Interval 9 complete
(convergence not achieved - truncating interval)
**************************************************
Estimation results
**************************************************
myblock2 = 1
--------------------------------------------------
Estimated propensity score in [.0021,.2439)
Number of treated obs. = 19
Number of control obs. = 663
--------------------------------------------------
myblock2 = 2
--------------------------------------------------
Estimated propensity score in [.2462,.2895)
Number of treated obs. = 8
Number of control obs. = 8
--------------------------------------------------
myblock2 = 3
--------------------------------------------------
Estimated propensity score in [.2929,.3741)
Number of treated obs. = 13
Number of control obs. = 16
--------------------------------------------------
myblock2 = 4
--------------------------------------------------
Estimated propensity score in [.3773,.4113)
Number of treated obs. = 8
Number of control obs. = 5
--------------------------------------------------
myblock2 = 5
--------------------------------------------------
Estimated propensity score in [.421,.6803)
Number of treated obs. = 30
Number of control obs. = 24
--------------------------------------------------
myblock2 = 6
--------------------------------------------------
Estimated propensity score in [.6992,.7595)
Number of treated obs. = 4
Number of control obs. = 3
--------------------------------------------------
myblock2 = 7
--------------------------------------------------
Estimated propensity score in [.7643,.972]
Number of treated obs. = 102
Number of control obs. = 7
--------------------------------------------------
Total number of tests conducted = 238
```

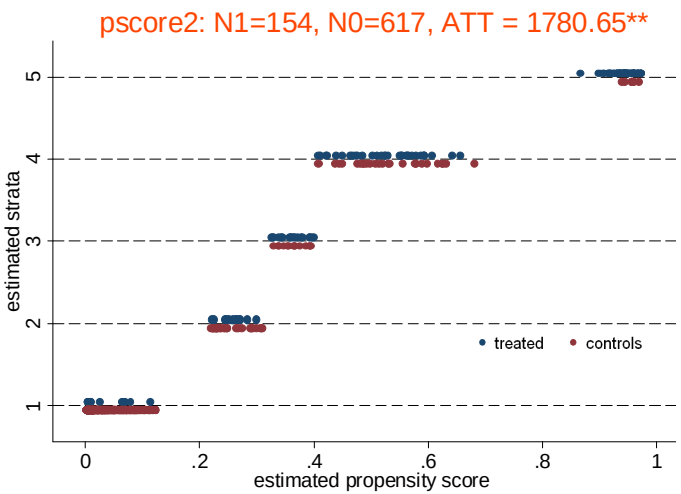
Sample output after summary

```
**************************************************
Estimation results
**************************************************
myblock2 = 1
--------------------------------------------------
Estimated propensity score in [.0021,.2439)
Number of treated obs. = 19
Number of control obs. = 663
--------------------------------------------------
p-value mean comparison test age = .7562
p-value mean comparison test age2 = .722
p-value mean comparison test educ = .7734
p-value mean comparison test educ2 = .9501
p-value Wilcoxon rank-sum test marr = .9827
p-value Wilcoxon rank-sum test black = .5416
p-value Wilcoxon rank-sum test hisp = .4222
p-value mean comparison test RE74 = .1578
p-value mean comparison test RE75 = .1885
p-value mean comparison test RE742 = .7053
p-value mean comparison test RE752 = .7244
p-value Wilcoxon rank-sum test blackU74 = .4696

myblock2 = 2
--------------------------------------------------
Estimated propensity score in [.2462,.2895)
Number of treated obs. = 8
Number of control obs. = 8
--------------------------------------------------
p-value mean comparison test age = .816
p-value mean comparison test age2 = .9373
    .
    .
    .
p-value mean comparison test RE742 = .3614
p-value mean comparison test RE752 = .2971
p-value Wilcoxon rank-sum test blackU74 = 1

myblock2 = 3
--------------------------------------------------
Estimated propensity score in [.2929,.3741)
Number of treated obs. = 13
Number of control obs. = 16
--------------------------------------------------
    .
    .
    .
myblock2 = ...
    .
    .
    .
myblock2 = 7
--------------------------------------------------
Estimated propensity score in [.7643,.972]
Number of treated obs. = 102
Number of control obs. = 7
--------------------------------------------------
p-value mean comparison test age = .5062
p-value mean comparison test age2 = .3944
p-value mean comparison test educ = .6103
p-value mean comparison test educ2 = .6592
p-value Wilcoxon rank-sum test marr = .4141
p-value Wilcoxon rank-sum test black = .5505
p-value Wilcoxon rank-sum test hisp = .5505
p-value mean comparison test RE74 = .7947
p-value mean comparison test RE75 = .6905
p-value mean comparison test RE742 = .7947
p-value mean comparison test RE752 = .6826
p-value Wilcoxon rank-sum test blackU74 = .5505
--------------------------------------------------
Total number of tests conducted = 238
--------------------------------------------------
```

NSW-PSID1

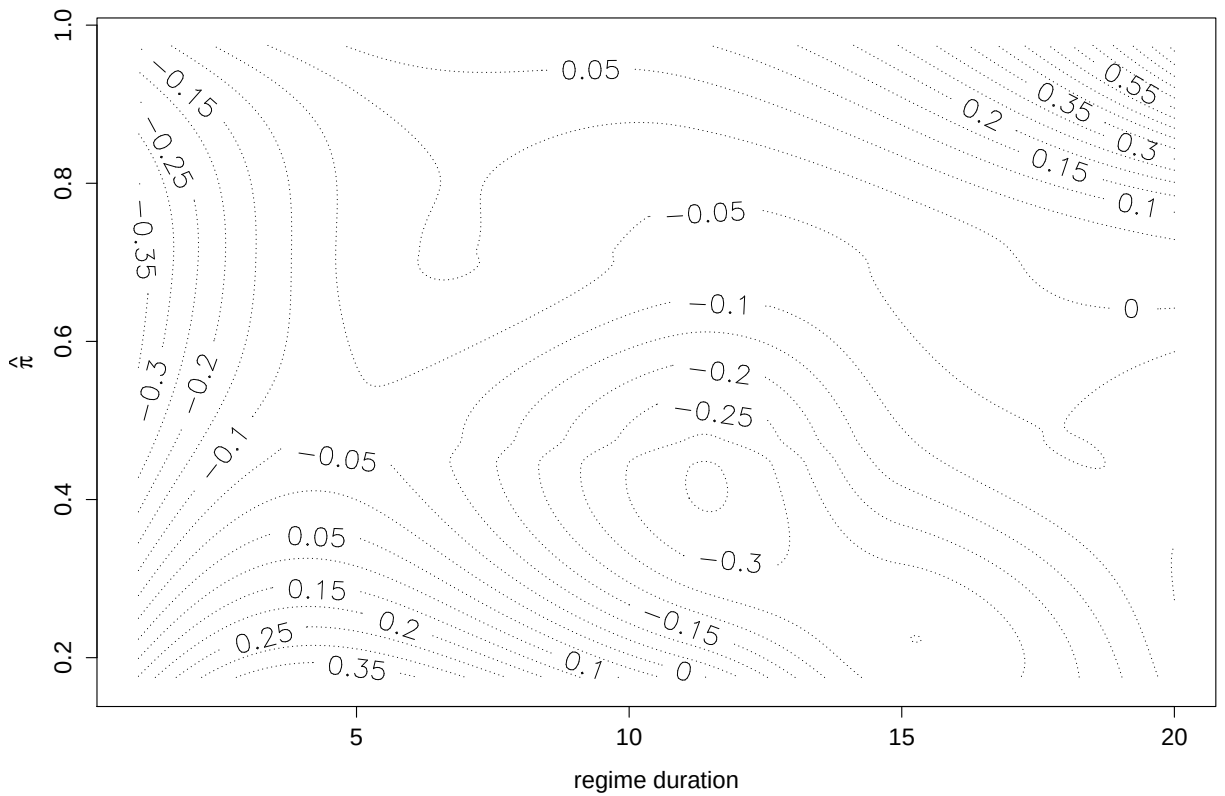
Data example: Deheija and Whaba (2002) with 185 treated observations where the non-experimental control group is used (2490 observations); however their estimate for ATT using the experimental control group is equal to $\hat{\gamma} = 1794$

pscore2	$\hat{\gamma}$	se	t	N_1	N_0
default	2067.18	755.58	2.74	184	664
tenflogit	1812.59	870.24	2.08	168	734
tenflogit02	1857.53	936.35	1.98	166	730
ksm02	1780.65	856.82	2.08	154	617
var02	-1082.39	1890.06	-.57	34	22
rev02	1953.76	886.13	2.2	145	163
ksmrev02	1953.76	992.63	1.97	145	163
medianrev02	2090.22	908.37	2.3	151	190
atts	2210.32	877.51	2.52	185	1154
attk	1540.15	842.04	1.83	185	1154
attnd	1446.93	1177.07	1.23	185	58
attr	-6023.44	4443.65	-1.36	26	69



Fixed Currency Regimes and the Pattern of Time

Data example: (Dorn and Egger, 2012, work in progress) Disaggregation of duration-specific ATTs for annual growth of bilateral trade into different regions of the estimated propensity to receive treatment using `pscore2`



What is a good choice of the type I error α?

Simulated data: Same simulation set-up as in introductory example

α	MSE($\hat{\gamma}$)	rel. MSE	Bias($\hat{\gamma}$)	$\frac{\text{Bias}(\hat{\gamma})^2}{\text{MSE}(\hat{\gamma})}$	Var($\hat{\gamma}$)	$\frac{\text{Var}(\hat{\gamma})}{\text{MSE}(\hat{\gamma})}$	$\gamma \in$ 0.95CI
0.01	0.143	100%	0.128	11.55%	0.126	88.45%	99.14%
0.05	0.122	85.50%	0.078	4.96%	0.116	95.04%	98.96%
0.08	0.106	74.48%	0.058	3.21%	0.103	96.79%	99.12%
0.1	0.098	68.51%	0.046	2.18%	0.096	97.82%	99.15%
0.15	0.081	56.96%	0.030	1.07%	0.080	98.93%	99.19%
0.2	0.071	49.99%	0.023	0.71%	0.071	99.29%	99.24%
0.25	0.066	46.23%	0.017	0.43%	0.066	99.57%	99.36%
0.3	0.060	42.30%	0.013	0.28%	0.060	99.72%	99.41%
0.4	0.055	38.64%	0.008	0.10%	0.055	99.90%	99.50%
0.5	0.053	36.96%	0.007	0.09%	0.053	99.91%	99.57%

→ The MSE of the estimated ATT ($\hat{\gamma}$) *decreases* with α increasing but there is a decreasing efficiency gain

→ Moreover, *bias-variance-trade-off*

Nasty data

Simulated data: Simulation study with $R = 10.000$ samples of size $N = 400$; in the left panel there are 3 regressors and one is omitted, in the panel in the center there are two regressors and an omitted interaction term, finally the data-design for the outer right panel combines both complications

	Omitted regressor			Omitted nonlinearity			Both problems		
α	MSE($\hat{\gamma}$)	$\frac{\text{Bias}(\hat{\gamma})^2}{\text{MSE}(\hat{\gamma})}$	$\gamma \in$ 0.95CI	MSE($\hat{\gamma}$)	$\frac{\text{Bias}(\hat{\gamma})^2}{\text{MSE}(\hat{\gamma})}$	$\gamma \in$ 0.95CI	MSE($\hat{\gamma}$)	$\frac{\text{Bias}(\hat{\gamma})^2}{\text{MSE}(\hat{\gamma})}$	$\gamma \in$ 0.95CI
0.01	0.242	44.07%	96.70%	0.157	24.26%	98.52%	0.286	8.29%	94.77%
0.05	0.190	34.90%	97.05%	0.108	9.44%	98.89%	0.186	3.51%	96.41%
0.08	0.161	32.10%	97.25%	0.086	5.32%	99.13%	0.145	2.12%	97.26%
0.1	0.145	31.18%	97.45%	0.074	3.85%	99.35%	0.127	1.62%	97.50%
0.15	0.119	29.07%	97.78%	0.058	1.78%	99.44%	0.097	0.95%	98.02%
0.2	0.105	29.06%	97.90%	0.051	1.00%	99.50%	0.085	0.73%	98.23%
0.25	0.095	28.96%	97.99%	0.048	0.70%	99.50%	0.078	0.61%	98.29%
0.3	0.089	28.46%	97.95%	0.047	0.63%	99.54%	0.074	0.55%	98.33%
0.4	0.081	27.98%	98.08%	0.045	0.26%	99.50%	0.072	0.52%	98.43%
0.5	0.077	28.11%	98.34%	0.044	0.13%	99.63%	0.070	0.50%	98.57%

→ Data partitions estimated by `pscore2` allow for *reliable inference about ATT* (γ) also in case of *misspecification* of the propensity score model

→ Omitted nonlinearity less problematic than omitted regressors

Concluding remarks

- The program `pscore2` implements a data-driven distinction between good comparisons and partitions of the covariate-space that do not satisfy the identifying support conditions for ATT, ATE etc.
- Moreover, for real data the estimated balancing score might be more or less sparsely populated with comparable observations, a data-driven approach to estimate strata seams natural
- The program `pscore2` uses a simple grid search procedure, but there are substantive efficiency gains!!!
- And finally, it is also quick since the dimensionality reducing feature of the propensity score allows to map a high-dimensional problem into a search problem on (0;1)
- Still, the `pscore2` algorithm is greedy and therefore the result depends on the search direction

References

Becker, S.O. and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal* 2(4): 358-377.

Dehejia, R.H. and S. Wahba. 2002. Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* 84(1): 151-161.

Rosenbaum, P.R. and D.B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrica* 70(1): 41-55.

Rosenbaum, P.R. and D.B. Rubin. 1984. Reducing the bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79(387): 516-524.

Diamond, A. and J.S. Sekhon. 2012. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics forthcoming*.