cemmap

Policy Evaluation Methods

Tutor: Barbara Sianesi, IFS

Resources

These notes set out some essential references in the Programme Evaluation (or Treatment Effect) literature that underlie my corresponding Cemmap Course, and other resources that the interested student might find helpful, in particular software and data.

Whenever possible, there are active links to the papers (via JSTOR or the authors' webpages) and for software downloads.

General overviews

Very comprehensive, but both highly technical:

Imbens, G. and Wooldridge, J. (2009), "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47, 5-86. NBER WP version.

Heckman, J.J., LaLonde, R.J. and Smith, J.A. (1999), "The Economics and Econometrics of Active Labour Market Programmes", in Ashenfelter, O. and Card, D. (eds.), *The Handbook of Labour Economics*, Volume III.

Focuses on the returns education, but the issues and methods are the same:

Card, D. (1999), 'The Causal Effect of Education on Earnings', in O. Ashenfelter and D. Card, *Handbook of Labor Economics*, vol. 3, Amsterdam: Elsevier-North Holland.

Strong statistical slant; clear focus on design issues:

Rosenbaum, P.R. (2010), *Design of observational studies*, Springer Series in Statistics.

A thorough book with a clear emphasis on causal effects:

Angrist, J.D. and Pischke, J.S. (2009), *Mostly Harmless Econometrics*, Princeton University Press.

Less technical references:

Blundell, R., Dearden, L. and Sianesi, B. (2005), "Evaluating the Effect of Education: Models, Methods and Results from the National Child Development Survey", *Journal of the Royal Statistical Society*, Series A, 168, 473-512. IFS WP No.WP03/20.

Blundell, R. and Costa Dias, M. (2002), "Alternative Approaches to Evaluation in Empirical Microeconomics", *Portuguese Economic Journal*, 1, 91-115, Cemmap WP No.CWP10/02 and (2009), *Journal of Human Resources*, 44, 565-640, Cemmap WP No.CWP26/08.

Accessible and entertaining introductions to the concepts and methods of impact evaluation:

Angrist, J.D. and Pischke, J.S. (2014), *Mastering 'Metrics: The Path from Cause to Effect*, Princeton University Press.

Ravallion, M. (1999), "The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation", World Bank Policy Research Working Paper No. 2153.

Specific topics

• *The Evaluation Problem*

Definition of causality, potential outcomes and the Fundamental Problem of Causal Inference

Though the concepts behind the framework have been around since Fisher (1935) and Neyman (1935), Rubin (1974) formally applied them to the study of causation; see however also the work by Roy (1951) and Quandt (1972):

Fisher, R.A. (1935), *The Design of Experiments,* Edinburgh: Oliver&Boyd.

Neyman, J. (1935), "Statistical Problems in Agricultural Experimentation" (with discussion), *Supplement of the Journal of the Royal Statistical Society*, 2, 107-180.

Quandt, R. (1972), "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association*, 67, 306-310.

Roy, A. (1951), "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, 3, 135-146.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non- Randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

SUTVA – stable unit-treatment value assumption

Rubin, D.B. (1980), "Discussion of 'Randomisation Analysis of Experimental Data in the Fisher Randomisation Test'" by Basu, *Journal of the American Statistical Association*, 75, 591-593.

Rubin, D.B. (1986), "Discussion of 'Statistics and Causal Inference'" by Holland, *Journal of the American Statistical Association*, 81, 961-962.

Holland, P.W. (1986), "Rejoinder", *Journal of the American Statistical Association*, 81, 968-970.

Decomposition of selection bias

Heckman, J.J., Ichimura, H., Smith, J.A. and Todd, P. (1998), "Characterising Selection Bias Using Experimental Data", *Econometrica*, 66, 1017-1098.

Bonus paper

Holland, P. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-970.

• *Social Experiments (or Randomised Trials)*

Fisher, R.A. (1935), *The Design of Experiments,* Edinburgh: Oliver&Boyd.

Burtless, G. (1995), "The Case for Randomised Field Trials in Economic and Policy Research", *Journal of Economic Perspectives*, 9, 63-84.

Heckman, J., and J. Smith (1995), "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85-110.

Heckman, J., Smith, J. and Taber, C. (1998), "Accounting for Dropouts in Evaluations of Social Programs", *The Review of Economics and Statistics*, 80, 1-14.

• *Natural Experiments and Instrumental Variables*

Looking for serendipitous causes of randomization: Natural experiments

Rosenzweig, M. and Wolpin, K. (2000), "Natural 'Natural' Experiments in Economics" Journal of Economic Literature, 38, 827-874.

Meyer, B. (1995), "Natural and Quasi- Experiments in Economics," *Journal of Business and Economic Statistics*, 13, 151-161.

Angrist, J. and Krueger, A. (2001), "Instrumental variables and the search for identification: From supply and demand to natural experiments", *Journal of Economic Perspectives,* 15, 69-85.

Deaton, A. (2009), "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development", The Keynes Lecture, British Academy.

Interpreting the estimand: LATE

Imbens, G.W. and Angrist, J.D. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 446-475.

Angrist, J.D., Imbens, G.W. and Rubin, D.B. (1996), "Identification of Causal Effects Using Instrumental Variables", *Journal of American Statistical Association*, 91, 444-472.

Angrist, J.D. and Imbens, G.W (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity", *Journal of American Statistical Association*, 90, 431-422.

Heckman, J.J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations", *The Journal of Human Resources*, 32, 441-462.

Problems when instruments are weak

Bound, J., Jaeger, D. and Baker, R. (1995), 'Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak', *Journal of the American Statistical Association*, 90, 443–450.

Staiger, D. and Stock, J.H. (1997), "Instrumental Variables Regression with Weak Instruments", *Econometrica,* 65, 557-586.

Bonus paper

Angrist, J. and Krueger, A. (1997), "Does compulsory school attendance affect schooling and earnings?", *Quarterly Journal of Economics,* 106, 979-1014.

• *Regression Discontinuity Design*

The original application

Thistlethwaite, D.L. and Campbell, D.T (1960), "Regression discontinuity analysis: an alternative to the ex post facto experiment", *Journal of Educational Psychology*, 51, 309-317.

The main review and overview paper:

Lee, D.S. and Lemieux, T. (2010), "Regression Discontinuity Designs in Economics", *Journal of Economic Literature*, 48, 281–355.

An intuitive yet rigorous introduction (sections 1 and 2) and an interesting extension of the sharp design (section 3):

Battistin, E. and Rettore, E. (2003), "Another Look at the Regression Discontinuity Design", Cemmap WP CWP01/03.

A highly formal treatment:

Hahn, J., Todd, P. and Van der Klauuw, W. (2001), "Identification and estimation of treatment effects with a regression-discontinuity design", *Econometrica*, 69, 201-209.

The special issue of the *Journal of Econometrics*:

(2008), "The regression discontinuity design: Theory and applications", 142, 611-850. In particular:

Imbens, G.W. and Lemieux, T. (2008), "Regression discontinuity designs: A guide to practice", *Journal of Econometrics*, 142, 615-635.

McCrary, J. (2008), "Manipulation of the running variable in the regression discontinuity design: A density test", *Journal of Econometrics*, 142, 698-714. NBER WP version.

Bonus papers: empirical applications in a fuzzy and sharp design respectively

Angrist, J. and Lavy, V. (1999): "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, 114, 533-575.

Lee, D.S. (2008), "Randomized experiments from non-random selection in U.S. House elections", *Journal of Econometrics*, 142, 675-697.

• *Matching Methods*

An up-to-date and comprehensive review:

Imbens, G. (2004), "Semiparametric estimation of average treatment effects under exogeneity: a review", *Review of Economics and Statistics*, 86, 4-29. Mimeo version.

The propensity score and propensity score matching

Rosenbaum, P.R. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984), "Reducing Bias in Observational Studies Using Sub-Classification on the Propensity Score", *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P.R. and Rubin, D.B. (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *The American Statistician*, 39, 1, 33-38.

Dehejia, R.H. and Wahba, S. (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes", *Journal of American Statistical Association*, 94, 1053-1062.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

Heckman, J.J., Ichimura, H. and Todd, P.E. (1998), "Matching as an Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Mahalanobis-metric matching, bias corrections, practical issues

Rubin, D.B. (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

Rubin, D.B. (1980), "Bias Reduction Using Mahalanobis-Metric Matching", *Biometrics*, 36, 293-298.

Abadie, A. and Imbens, G. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267, mimeo.

Lechner (2002), "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods," *Journal of the Royal Statistical Society*, Series A, 165, 59-82. Discussion paper version.

Caliendo, M. and Kopeinig, S. (2008), "Some Practical Guidance for the Implementation of Propensity Score Matching", *Journal of Economic Surveys*, 22, 31-72. IZA DP No.1588.

Multiple treatments

Imbens, G.W. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions", *Biometrika*, 87, 706-710. NBER Technical WP No.237.

Lechner, M. (2001), Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies*, Heidelberg: Physica/Springer, 43-58. WP version.

Sianesi, B. (2008), "Differential Effects of Swedish Active Labour Market Programmes for Unemployed Adults during the 1990s", *Labour Economics*, 15, 392-421. WP version.

Efficiency/inference issues (these are quite technical papers)

Abadie, A. and Imbens, G. (2011), "Matching on the Estimated Propensity Score", mimeo.

Abadie, A. and Imbens, G. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267, mimeo.

Angrist, J.D. and Hahn, J. (2004), "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects", *Review of Economics and Statistics,* 86, 58-72. NBER Technical WP No.241.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315-331.

Hirano, K., G. Imbens, and G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161-1189. NBER Technical WP No.251.

Bonus paper

Angrist, J. (1998), 'Estimating the labour market impact of voluntary military service using social security data on military applicants', *Econometrica*, 66, 249–88.

• *Interrupted Time Series Analysis*

Shaddish, S.R., Cook, T.D. and Campbell, D.T. (2002), "Experimental and quasi-experimental designs for generalized causal inference", Boston, MA: Houghton Mifflin.

Wagner, A.K., Soumerai, S.B., Zhang, F., and Ross-Degnan, D. (2002), "Segmented regression analysis of interrupted time series studies in medication use research", *Journal of Clinical Pharmacy and Therapeutics,* 27, 299-309.

• *Synthetic Control Method*

Abadie, A. and Gardeazabal, J. (2003), "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93 (1), 113–132.

Abadie, A., Diamond, A. and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies of Aggregate Interventions: Estimating the Effect of California's Tobacco Control Program", *Journal of the American Statistical Association,* 105(490): 493-505.

Abadie, A., Diamond, A. and Hainmueller, J. (2015), "Comparative Politics and the Synthetic Control Method", *American Journal of Political Science*, 59, 495-510.

• *Difference-in-Differences*

Less formal overview and discussion:

Meyer, B. (1994), "Natural and Quasi-Experiments in Economics," NBER Technical WP No. 170

Parametric

The famous 'dip':

Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 6, 47-57.

LaLonde, R.J. (1986), "Evaluating the Econometric Evaluation of Training Programmes with Experimental Data", *The American Economic Review*, 76, 604-620.

Non- (or semi-)parametric

Heckman, J.J., Ichimura, H. and Todd, P.E. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

Abadie, A. (2005), "Semiparametric Difference-in-Differences Estimators", *Review of Economic Studies*, 72, 1-19.

Bonus paper

Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2004), "Evaluating the employment effects of a mandatory job search program", *Journal of the European Economic Association*, 4, 569–606.

Stata software

• *Standard commands*

Social experiments	regress; ttest ,by()
Natural experiments / Instrumental variables	ivregress
Ordinary least squares	regress
Before-after	regress; ttest
Difference-in-differences	regress

• *Specialised software: MATCHING*

STATA 13 `teffects` SUITE

Stata command to perform different types of estimators relying on the CIA:
`teffects ipw` – inverse-probability weighting estimator
`teffects aipw` –augmented inverse-probability-weighted estimator
`teffects ipwra` – inverse-probability-weighted regression-adjustment estimator
`teffects nnmatch` – Mahalanobis nearest neighbour(s) matching
`teffects psmatch` – propensity score nearest neighbour(s) matching.
All provide analytical asymptotic standard errors.

From Stata 14: `tebalance` commands for covariate imbalance testing.

PROPENSITY SCORE AND MAHALANOBIS METRIC MATCHING

Leuven, E. and Sianesi, B. (2003), "PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing".

Performs a variety of propensity score matching methods (nearest neighbour or within caliper, with or without replacement; *k*-nearest neighbours; radius matching; kernel matching; local linear regression matching; and spline matching) and full Mahalanobis matching (via nearest neighbour, kernel or local linear regression) to estimate the average effect of treatment on the treated, and, optionally, on the non-treated and the full sample.
It further provides features for common support graphing (`psgraph`) as well as for covariate imbalance testing, both before and after matching (`pstest`).

To install/update (from within net-aware Stata):

```
ssc install psmatch2, replace
```

Alternatively, download the various components manually from here.

PROPENSITY SCORE MATCHING

Becker, S. and Ichino, A. (2002), "PSCORE, ATTS, ATTR, ATTK, ATTNW, ATTND".

Tests for covariate imbalance when stratifying on the propensity score, and estimates the average treatment effect on the treated using stratification matching, radius matching, nearest neighbour matching and kernel matching.

More information:

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

• *Specialised software: FULLY INTERACTED REGRESSION*

Leuven, E. and Sianesi, B. (2004), "FILM: Stata module to perform 'fully interacted linear matching', that is a fully interacted linear regression model in which the treatment dummy is interacted with each one of the other regressors".

The treatment effect is allowed to vary according to all (or a subset of the) observable variables. An F-test of no heterogeneous effects is displayed. The overall effect – averaged for the treated and optionally also for the non-treated and the population – is then displayed, together with analytical standard errors (allowing for robust standard errors as well as for clustering).
`film` allows the user to perform various steps to increase the flexibility of the regression model until it resembles a matching estimator (see `psmatch2`) – from simple dummy variable OLS regression, to full interactions, to imposition of the common support – while retaining analytical standard errors and related saving in bootstrapping time.

Download ado file
Download help file

See also Stata 13's `teffects ra` – regression adjustment (interacted OLS)

• *Specialised software: INTERRUPTED TIME SERIES ANALYSIS*

Linden, A. (2014), "ITSA: Stata module for conducting interrupted time series analysis for single and multiple groups".

Performs interrupted time series analysis to estimate the effect of an intervention when the outcome variable is ordered as a time series and a number of observations are available in both pre- and post-intervention periods; can estimate ITSA for either a single treatment group or a multiple-group comparison, as well as for multiple treatment periods.

To install/update (from within net-aware Stata):

```
ssc install itsa, replace
```

• *Specialised software: SYNTHETIC CONTROL METHOD*

Hainmueller, J., Abadie, A. and Diamond, A. (2014), "SYNTH".

Implements the synthetic control method for causal inference in comparative case studies.

To install/update (from within net-aware Stata):

```
ssc install synth, replace
```

• *Specialised software: REGRESSION DISCONTINUITY DESIGN*

ROBUST INFERENCE IN REGRESSION DISCONTINUITY DESIGNS

Calonico, S. Cattaneo, M.D and Titiunik, R. (2015), "RDROBUST".

Stata package to implement statistical inference and graphical procedures for Regression Discontinuity designs, including in particular local-polynomial-based point estimators, confidence intervals estimators and bandwidth selectors, as well as automatic RD-plots.

To install/update

```
net install rdrobust, from(http://www-
personal.umich.edu/~cattaneo/software/rdrobust/stata) replace
```

OTHER STATA MODULES FOR SHARP AND FUZZY RDD

Imbens, G. (2009), "rdob".

To download.

Nichols, A. (2011), "rd 2.0: Revised Stata module for regression discontinuity estimation"

To install/update: `ssc install rd, replace`

DENSITY TEST IN REGRESSION DISCONTINUITY DESIGNS

Kovak, B. and McCrary, J. (2009), "DCdensity".

Stata package for the estimation of a discontinuous density function, as outlined in McCrary (2008)

To download.

The NSW Data

These data are very well known in the evaluation literature: they form the basis of a seminal paper in the evaluation literature[1] and are at the heart of an interesting debate[2] on matching methods.

This dataset is great to play with, in that it offers an opportunity to apply all of the above methods (bar regression discontinuity design) to a common dataset, and allows the performance of the different non-experimental estimators to be assessed against the benchmark of the experimental estimate. (They are also used e.g. by Ichino and Becker (2002) and Abadie, Drukker, Leber Herr and Imbens (2001) to illustrate their respective Stata matching programs[3]).

Specifically, the data we use in the Cemmap course combine treatment and control individuals from a randomised evaluation of the National Supported Work (NSW) Demonstration with non-experimental individuals drawn from survey data from the Panel Study of Income Dynamics (PSID):

- nsw.dta
- nsw_psid.dta

The full dataset can be found on Rajeev Dehejia's website.

[1] LaLonde, R.J. (1986), "Evaluating the Econometric Evaluation of Training Programmes with Experimental Data", *The American Economic Review*, 76, 604-620.

[2] Dehejia, R.H. and Wahba, S. (1999), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programmes", *Journal of American Statistical Association*, 94, 1053-1062.

Smith, J. and Todd, P. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?", *Journal of Econometrics* 125, 305-353.

[3] Becker, S. and Ichino, A. (2002), "Estimation of average treatment effects based on propensity scores", *The Stata Journal*, 2, 358-377.

Abadie, A., Drukker, D., Leber Herr, J. and Imbens, G. (2001), "Implementing Matching Estimators for Average Treatment Effects in Stata", *The Stata Journal*, 1, 1-18.