# Public policy and Stata

# Public Policy and Stata

Stephen P. Jenkins
(LSE)

8 July 2014

[Approximately 3,600 words, including References and About the Author]

*Abstract*

This essay considers how public policy has evolved over the past three decades and how Stata has been part of this process.

*Correspondence*

Stephen P. Jenkins, Department of Social Policy, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, United Kingdom. Email: s.jenkins@lse.ac.uk

*Acknowledgements*

I owe personal thanks to Bill Gould and all at StataCorp for their contribution to my research career: Stata has been an ever-present companion in my work over the last quarter century. Specific thanks to Vince Wiggins for suggestions about how to proceed with the essay and to Lucinda Platt for comments on a first draft.

## Introduction

I was invited to write about how public policy has evolved over the past three decades and how Stata has been part of this process. This is an impossible brief, so I am going to be selective in terms of coverage and, even then, all perspectives provided are strongly coloured by my own career as an applied economist which, as it happens, spans roughly the same three decades as Stata's. I limit my scope to the areas of health, education, welfare, and the labour market; to individuals, families, and households, and statistical analysis of survey or administrative data. This is a 'micro' perspective; I'm not discussing macroeconomics or time-series data. It is quantitative policy analysis that is my subject rather than the public policies themselves. I'm focusing on topics researched primarily by social scientists, and mainly those in which economists and econometricians now play an influential role.

## The Credibility Revolution in public policy analysis

Let me begin by putting on the hat of a contemporary mainstream empirical microeconomist. From this perspective, there is a very clear view about what has happened to policy analysis over the last three decades: there has been substantial change in approach, and all for the better. As Angrist and Pischke put it, '[e]mpirical microeconomics has experienced a credibility revolution, with a consequent increase in policy relevance and scientific impact. … the primary engine driving improvement has been a focus on the quality of empirical research designs' (2010: 4).

By empirical research designs, Angrist and Pischke mean methods and data sets that allow analysts to identify causal effects credibly, referring to approaches based on random assignment to treatment and control groups (randomized control trials), and to natural and quasi-experiments. In all the approaches, the researcher seeks data in which there is variation across cases in a key treatment variable and in addition, crucially, that variation can be taken as exogenously given. Regression-based methods in the quasi-experimental approach include instrumental variables, regression discontinuities, or differences in differences. (In the latter case, fixed effects estimators applied to panel data are commonly used to control for time-invariant unobserved confounders.) Nonparametric methods for comparing treated and control cases that account for observable differences are based on covariate adjustment using matching by propensity score, nearest neighbour, or kernel, and related reweighting methods.

What was Stata's role in these developments? My view is that it has been substantial, for two reasons. The first is that, even though Stata is not essential to implement the purer experimental evaluations, it has been used by many researchers anyway because Stata was already their software of choice for data management and it also had the generic statistical tools required. The closer the research design is to a randomized control trial, the closer the estimation of treatment effects is to a simple comparison of means and so the statistical component of the evaluation is a relatively straightforward task compared to getting the empirical research design and data right.

The second reason Stata has played a substantial role is that specialist statistical routines for estimating treatment effects were written and made widely available to users early on during the credibility revolution. Randomized control trials are relatively rare in evaluations of public policy because of perceived ethical or infeasibility problems (it's difficult to randomly assign marital status, or differences in drug or alcohol consumption, say), or because of high set-up costs and lack of specialist know-how. And the events facilitating natural experiments are also relatively rare. External validity is also an issue for both types of experiment. So, quasi-experimental designs based on observational data have been the most prevalent approach and, in this case, statistical analysis takes on a greater role with the overall evaluation. There were a number of Stata modules that became widely available in the last decade or so that provided the requisite specialist tools.

The leading example is the `psmatch2` package by Edwin Leuven and Barbara Sianesi implementing not only Mahalanobis and propensity score matching, but also integrated tools for checking for common support and covariate imbalance. The `psmatch2` module, built on Sianesi's (2001) `psmatch` package, was first released in the Statistical Software Components Archive at Boston College ('SSC') in April 2003 and has been frequently updated ever since. This package is currently ranked number one in terms of total software downloads ever from the SSC (18,257 downloads as of 2014-07-06: http://logec.repec.org/scripts/paperstat.pf?h=RePEc:boc:bocode:s432001). (Less well-known is the suite of programs providing similar functionality that accompanies Becker and Ichino's (2002) *Stata Journal* article. Abadie et al. (2004) provided extensions focusing on nearest neighbour matching.) It was only in 2013 that the number of `psmatch2` downloads began to fall, no doubt because of the release of Stata 13 with its `teffects` suite of estimators, though the number remains relatively large.

Another Stata package important for policy evaluation is `ivreg2` (Baum, Schaffer, and Stillman 2003; 2007), together with its panel data sibling `xtivreg2` (Schaffer 2011).

These do instrumental variables regression for linear models, providing functionality that goes beyond that provided by Stata's built-in commands `ivreg` and its successor `ivregress`. `ivreg2` is fourth ranked in terms of total downloads from SSC ever (first released May 2002; 14,657 downloads as of 2014-07-06), and `xtivreg2` is eighth in the total download ranking (first released November 2005; 7,746). Software for regression discontinuity analysis was provided by Austin Nichols's `rd` package (14th on the SSC total download list with 14,657 downloads since November 2007). Although basic differences-in-differences analysis can be straightforwardly implemented using Stata's built-in commands, some extensions require more specialist estimators, some of which are provided by Juan Miguel Villa's `diff` package (17th on the SSC total download ranking with 4,956 downloads since October 2009).

Although the SSC is not the only source of Stata code for policy analysis and the usefulness of download statistics can be questioned, my summary judgement is that the SSC data provide good prima facie evidence of the Stata's contribution to policy analysis in the post-credibility revolution environment.

**A broader view of what counts as valuable public policy analysis**

However, this is not the only way to assess what counts as public policy analysis. Although the credibility revolution of contemporary mainstream empirical microeconomics is very influential (and rightly so), it incorporates a rather narrow view about what counts as valuable policy analysis. There is also a substantial contribution made by what some researchers label rather disparagingly as 'descriptive' analysis.

I would contend that knowing how things are, and how they have changed or compare with another country, is an essential prerequisite to any discussion about policy options and priorities, let alone any sort of evaluation of specific policy measures. Some of the biggest policy debates are founded on arguments about 'the facts'. Recent examples in the UK include how much intergenerational social mobility there is and how its changed over time, how much real incomes, inequality and poverty rates have changed in the aftermath of the Great Recession in the era of austerity, the educational performance of school children and how this differs by ethnic minority group, social background, and type of school attended; employment rates and earnings of recent migrants compared to native-born UK workers; differences in the inflation rates faced by pensioners and other groups, etc., etc.

As soon as the definition of policy analysis is broadened to include this sort of research, then Stata's role can be counted as even larger. The reason is simply that the people doing this sort of work are increasingly using Stata to do their analysis: the user base widens beyond credible revolutionaries to many other quantitative researchers in universities and increasing numbers of people working in research institutes, and local, national, and international governmental agencies.

How has their descriptive policy analysis changed over the last three decades? One fundamental change is in the nature of the data available and the capacity to analyse it. Policy-relevant quantitative research requires good data. Let me take the UK as an example. Thirty years ago, unit record data from household surveys were only just beginning to become more widely accessible to researchers, facilitated by the work of the UK Data Archive acting as a national data library. But analysis of such data was constrained by hardware. The survey data files were considered 'large' and typically held on magnetic tape and accessed via a mainframe computer. The personal computer had only recently been invented and few were seen in universities initially. Statistical software packages were limited in their functionality and their integration. (I began my career using one package for data management and another for estimation.) The teaching of econometrics often focused on time series analysis, reflecting the availability of such data. There were no national longitudinal surveys. Fast forward to 2014, and the world is totally different.

Public policy analysts are now awash with accessible data from multiple sources. Household survey data are routinely available from national data archives or downloadable from the internet, and there is a wealth of longitudinal as well as cross-sectional survey data. Administrative data are increasingly part of a researchers' portfolio too, and there are many more possibilities for data combination through linkages across data sources, often using geo-referenced identifiers. We can easily access data from other countries in addition to our own. The internet itself is providing data. The era of Big Data is coming upon us.

Our networked personal computers have the capacity to store data files that are much larger than could have been imagined thirty years ago but also to analyse them using a vastly extended portfolio of statistical tools. For example, methods for analysis of limited dependent variables, panel and survival data, sample selection, and robust standard errors, and survey design effects, are routinely available. The training of graduate social scientists in quantitative methods has improved in parallel.

As a result of these developments, commissioners and funders of research expect more and researchers deliver more, in terms of coverage of data (e.g. drawing on a full time

series of cross-sectional surveys, rather than simply one or maybe two) and statistical sophistication. However, one thing that has not changed is the report style required for non-academic audiences such as national and international agencies and some research foundations. It remains the case that reports must be written in accessible plain English for readers without the quantitative skills of the researchers. Communicating research findings effectively to non-academic audiences is as big a challenge to policy analysts today as it was thirty years ago. Indeed it may be a greater challenge now because the gap between the statistical training and experience of researchers and non-academic research users has probably widened.

Stata has played a significant part in the descriptive policy analysis context, though it is a role that is hard to quantify precisely – especially in a manner that would satisfy credibility revolutionaries! One strong signal of Stata's role is its take-up among researchers who do quantitative policy analysis. We are not privy to StataCorp's sales figures but it is manifest that, whereas the market for general statistical software was dominated by two Goliaths thirty years ago, Stata is the David of today. Moreover, casual observation suggests that it is Stata that is increasingly being used in the training of the policy analysts of tomorrow in the quantitative methods teaching in universities today.

Stata use has increased substantially not only in universities and non-governmental research institutes but in governmental agencies with research capacity. Internet searches on '<agency name> Stata' frequently lead to evidence of Stata use, especially where the agency is North American or an international organisation such as WHO or the World Bank (of which more shortly). National statistical offices have tended to favour common software approaches, standardizing on a single relational database management system interfaced with Some Alternative Software. But even in these environments, Stata is increasingly used for specialist tasks. (For Canadian and UK examples, see McCrosky 2012 and Barnes 2002.)

Early adoption of Stata by researchers with esteem and influence, together with free sharing of resources of multiple complementary kinds, leads to an increase in take-up and use by others – a form of virtuous circle. A leading example of an effect of this kind is the role played by the World Bank's research department and associated researchers from outside the Bank. They were early adopters of Stata, and have long made programs freely available to outside researchers. The collection of stand-alone programs in the 'Poverty Analysis Toolkit' (http://go.worldbank.org/YF9PVNXJY0) is an example. This has recently been substantially redeveloped (extending the topic coverage) and is now part of an integrated software environment, ADePT (Automated DEC Poverty Tables) which is built on top of Numerics by

Stata – a version of Stata that is embedded within applications developed by others. See Poi (2010) for further discussion.

ADePT is freely downloadable from http://go.worldbank.org/UDTL02A390, together with extensive documentation and data sets. The latest version contains eight modules for the analysis of poverty and inequality, social transfers, labour, gender, health equity, education, and food security. Accompanying these are five books, all freely downloadable, introducing the underlying analytical methods with extensive examples and written by leaders in the respective fields. The quality and comprehensiveness of *A Unified Approach to Measuring Poverty and Inequality* (Foster et al. 2013), combined with its unbeatable price, is such that I plan to adopt it as a course textbook in the coming academic year.

Earlier World Bank books have had substantial influence in their fields. I refer for example to O'Donnell et al. (2008) on health equity measurement, with a large collection of Stata examples and downloadable resources. The pioneer *par excellence* is Deaton's magisterial *The Analysis of Household Surveys* (1997), a source I still consult and also use in teaching. Deaton, one of the world's leading economists (and 2009 President of the American Economic Association), wrote in his Preface that 'in my experience [Stata] is the most convenient package for working with data from household surveys' (1997: 2), and, unusually for its time, the book provided the Stata code for the analysis, thereby enabling others to implement the methods, many of which were advanced for their time and not easily available elsewhere.

**What is it about Stata?**

So, Stata has made a substantial and growing contribution to quantitative public policy analysis over the last thirty years. In many ways, the features of Stata that underpin this contribution are the same characteristics that make it the software of choice for other forms of quantitative analysis. What are these?

It is tempting to begin by simply pointing to the way Stata integrates tools for data management, statistical analysis, and graphics, but that hardly makes it distinctive among competing software packages and, arguably, my preferences for Stata over them could reflect habit. Nevertheless, I would point to two aspects of Stata that have been particularly important in my own research career. Switching to Stata as my main statistical software in around 1994, I was struck by its emphasis on do and log files and hence the ability to create audit trails and reproducible results. As a researcher and research project team manager, I rate

this capability very highly. As a creator of data released to a wider public (income variables for the British Household Panel Survey), it was essential. Second, there was the shift to the new suite of graphics commands in version 8 (2003). Not only could chart creation be automated using do file code, but there were the substantial improvements in functionality per se. I referred earlier to the need for policy analysts to communicate effectively to research users, and my experience is that well-designed graphs are particularly valuable for this.

There are six factors underpinning Stata's success in addition to its integrated nature. Here follows a reprise of my list from a decade ago (Jenkins 2005), suitably updated. First, there is Stata's extensibility – building in the capacity for users to extend Stata themselves (using ado files and, more recently, Mata), combined with an openness and encouragement to do so from StataCorp. Second, there was the early exploitation of the internet, with seamless integration of the ability to download free software updates and user-written programs, and to search for such materials. Third, Stata runs in RAM memory and so is relatively fast. Although memory capacity was once a constraint, it was recognised early on that the ever-falling price of memory meant that this would not remain a practical problem. Fourth, Stata is produced for virtually all operating systems, and it is the same in each case.

Fifth, StataCorp fosters close links with its users: it listens. For example, it sends staff to the independently-run Stata user group meetings held worldwide and developers present scientific papers, run short courses, and host 'wishes and grumbles' sessions with users. Stata developers read and contribute to Statalist. StataCorp provides technical support to users of a quality that is unparalleled.

Sixth, and perhaps of most vital importance to researchers, Stata does not sacrifice academic integrity: Stata is for science. It provides capacity for cutting-edge statistical methods though in a suitably conservative manner. Implementation of methods typically follows scientific acceptance (as with the treatment effects packages cited earlier), often based on consultations with specialist experts in the relevant field (again, as with the treatment effects packages cited earlier), and always after extensive in-house validation and certification exercises. The links with science have been fostered by the development of the *Stata Journal*, and the publication of many excellent '… with Stata' books by world-leading econometricians and statisticians.

What links all these features is a type of integration and connectedness between software, developers, and users that constitutes a virtuous circle that has played out to mutual advantage over the last three decades and shows no sign of abating. Stata will continue to

play a major role in quantitative analysis of all kinds and in policy analysis in particular. Happy 30th birthday!

**References**


Abadie, A., Drukker, D., Leber Herr, J., and Imbens, G. W. (2004). 'Implementing matching estimators for average treatment effects in Stata', *The Stata Journal*, 4 (3), 290–311.

Angrist, J. D. and Pischke, J.-S. (2010). 'The credibility revolution in empirical economics: how better research design is taking the con out of econometrics', *Journal of Economic Perspectives*, 24 (2), 3–30.

Barnes, M. (2002). 'Using Stata at the Office for National Statistics'. Presentation at the 8th UK Stata User Group meeting, London. http://www.stata.com/meeting/8uk/ONSstatausergroup.pdf

Baum, C. F., Schaffer, M. E., and Stillman, S. (2003). 'Instrumental variables and GMM: estimation and testing'. *The Stata Journal*, 3 (1), 1–31.

Baum, C. F., Schaffer, M. E., and Stillman, S. (2007). 'Enhanced routines for instrumental variables/generalized method of moments estimation and testing', *The Stata Journal*, 7 (4), 465–506. (`ivreg2`, updated on SSC, February 2014.)

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

Deaton, A. (1997). *The Analysis of Household Surveys: a Microeconometric Approach to Development Policy*. Baltimore MD: The Johns Hopkins University Press for the World Bank. http://documents.worldbank.org/curated/en/1997/07/694690/analysis-household-surveys-microeconometric-approach-development-policy with Stata code downloadable from http://ideas.repec.org/c/boc/bocode/s360701.html

Foster, J. E., Seth, S., Lokshin, M., and Sajaia, Z. (2013). *A Unified Approach to Measuring Poverty and Inequality*. Washington, DC: The World Bank. http://go.worldbank.org/SKL1X5BT20

Jenkins, S. P. (2005). 'Oration for the award of honorary doctorate to William Gould'. Colchester UK: University of Essex. http://www.essex.ac.uk/honorary_graduates/or/#2005

Leuven, E. and Sianesi, B. (2003). '`psmatch2`: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance

testing', SSC Archive, Boston College. (Updated on SSC, February 2014.) http://ideas.repec.org/c/boc/bocode/s432001.html

McCrosky, J. (2012). 'Custom Stata commands for semi-automatic confidentiality screening of Statistics Canada data'. Presentation at 2012 Stata Conference, San Diego. http://www.stata.com/meeting/sandiego12/materials/sd12_mccrosky.pdf

Nichols, A. (2007). '`rd`: Stata module for regression discontinuity estimation', SSC Archive, Boston College. http://ideas.repec.org/c/boc/bocode/s456888.html (Updated on SSC, June 2014.)

O'Donnell, O., van Doorslaer, E., Wagstaff, A., and Lindelow, M. (2008). *Analyzing Health Equity Data Using Household Survey Data*. Washington, DC: World Bank. http://go.worldbank.org/LVSSZJX9O0

Poi, B. (2010). 'Stata makes a difference at the World Bank: automated poverty analysis', *The Stata News*, 25 (2), 2–3. http://www.stata.com/news/statanews.25.2.pdf

Schaffer, M. E. (2005), '`xtivreg2`: Stata module to perform extended IV/2SLS, GMM and AC/HAC, LIML and k-class regression for panel data models', SSC Archive, Boston College. http://ideas.repec.org/c/boc/bocode/s456501.html. (Updated on SSC, November 2013.)

Sianesi, B. (2001). 'Implementing propensity score matching estimators with Stata', Presentation at the Seventh UK Stata Users Group meeting, London, May 2001. http://fmwww.bc.edu/repec/usug2001/psmatch.pdf

Villa, J. M. (2009). '`diff`: Stata module to perform differences in differences estimation', SSC Archive, Boston College. http://ideas.repec.org/c/boc/bocode/s457083.html (Updated on SSC, May 2014.)

**About the author**

Stephen P. Jenkins is Professor of Economic and Social Policy at the London School of Economics and Political Science, and was formerly at the Institute for Social and Economic Research, University of Essex, where he was Director 2006–2009. Stephen has been a Stata user since version 2.1, is the author of a number of frequently-downloaded modules on SSC, and a regular presenter of Stata-based short courses on Survival Analysis and Statistical Graphics. He is an associate editor of the *Stata Journal*, and has co-organised the UK Stata User Group meetings in London every second year since 1999. He has had a long career in

policy research, including projects for UK government departments and agencies, New Zealand Treasury, and the OECD. Stephen is currently the Editor-in-Chief of the *Journal of Economic Inequality*.