The Smarter Use of Data

Plenary Address to the NatStats08 Conference

Melbourne

20 November 2008

David Gruen and Anthony Goldbloom

Macroeconomic Group

The Treasury

I am very pleased to be part of the first ever NatStats conference dealing with a most worthy, though often overlooked, subject.

Much of what is done in public policy is based on statistical insights, making our nation's statistics crucially important. However, it's also true that more of what is done in public policy could be based on statistical insights, making improving the statistical base, which is a goal of this conference, just as important.

Professor Adrian Smith, in his 1996 inaugural address as president of Britain's Royal Statistical Society, outlined his vision for an evidence-based society "where decisions about matters of substance… are taken on the basis of the best available evidence".[1] At the Treasury, we deal with matters of substance but we don't always have access to the best available evidence.

Today, I will talk about some of the remarkable achievements of statistical analysis over the past eighty years. These serve to illustrate the importance of statistics and show the power of a good evidence base.

I'll then talk briefly about today's trends, which are conspiring to push evidence into more and more of our policy decisions. Both the technology for collecting data and the appreciation of its value have driven almost universal agreement that policy should be based on the "best available" evidence.

However, as I mentioned, we are still some way from having the "best available" evidence at hand. I'll spend the remainder of the talk discussing how we might make better use of the masses of data that we currently collect.

Statistics played a vital role in last century

In order to motivate the power of a strong evidence base, I will begin by looking back over the past eighty years; at achievements that are linked to the development of statistics.

American economist Richard Froyen, from the University of North Carolina, Chapel Hill, refers with despair to

> "Presidents Hoover and then Roosevelt designing policies to combat the Great Depression… on the basis of such sketchy data as stock price indices, freight car loadings, and incomplete indices of industrial production". [2]

At the same time in Australia, former Commonwealth Statistician (and later Treasury Secretary) Sir Roland Wilson had to publish 'tentative' balance of payments statistics with query marks denoting data gaps.[3] The policymakers trying to steer the economy through the Great Depression were driving with a shattered windscreen and a badly fogged rear view mirror.

---

[1] Smith (1996).

[2] Froyen (2005).

[3] Australian Bureau of Statistics (2005).

Australia's first national accounts were released in 1945 and were followed soon after by the first use of fiscal policy as a stabilisation tool, because for the first time policymakers had a comprehensive picture of macroeconomic conditions. As the century wore on, policymakers and academics could use the national accounts to gauge the effectiveness of different interventions, helping to refine the macroeconomic levers.[4]

This allowed the Australian economy to transition from periods of growth punctuated by depression – which characterised the nature of the macroeconomy before the Second World War – to the much more benign business cycles of recent history.

Today, as we are faced with the worst financial crisis since the Great Depression, it is to our considerable collective benefit that the statistical evidence base has improved so much as to make possible the pre-emptive monetary and fiscal policy actions that we have seen in recent months.

The availability of data has also helped with the design of more effective microeconomic policies. The 1975 Asprey tax review relied on few statistics and consequently gave only limited insight into how the burden of the proposed system might be shared (the review barely mentions the word "distribution"). During the 1970s the development of the household income survey, which presented the distribution of income in Australia, supported the modelling of the 1985 tax reforms.

Then in the 1980s the release of confidentialised unit record files, which show individual responses to surveys, boosted the use of microsimulation modelling, allowing analysts to segment the population into much finer cohorts. Analysts could move from examining the impact of tax policy on the average person, to the impact on the average 30-year-old single male earning less than $60,000 a year.

Unit record files and input-output tables were instrumental in designing the New Tax System in the late 1990s, and distributional analysis using these types of data will also be crucial to the deliberations of the Australia's Future Tax System Review, which was set up by the Federal Government earlier this year.

These data have also allowed Treasury's climate change modellers to tackle calculations such as estimating the least cost path to our emissions target. And advances in data allow more targeted interventions to be rigorously examined; for example, identifying industries that may be at risk of carbon leakage upon the introduction of an emissions trading scheme.

Various factors are conspiring to push data-driven decision-making into more public policy

Let me now turn my attention to the present; to discuss why it is a good time to focus on improving our evidence base.

Given the achievements of statistical analysis over the past eighty years, it's not surprising that evidence-based policy is gaining traction. At the start of last century, the eminent

[4] Romer (1999) and Bernanke (2004) argue that lower amplitude modern business cycles are the result of better macroeconomic management. But without good measures of economic activity (and consumer price inflation), it would not have been possible to develop (or make use of) the necessary macroeconomic policy instruments.

British statistician Karl Pearson studied the effect of alcohol swigging parents on children. Before beginning the study, Pearson was sympathetic to the temperance movement, which supported the prohibition of alcohol. But to the chagrin of his comrades, his study concluded that children's health and intelligence were not affected by boozing parents. He became a pariah in the movement even though few bothered to read his study.[5]

Today, we'd be less likely to blindly dismiss Pearson's study. More and more social science students are exposed to statistical methods, and data-driven decision-making is becoming mainstream in public policy circles. Last month, Australian academics and policymakers gathered in Canberra for the inaugural *Evidence-Based Policy Development Conference*. On the international scene, over 130 countries sent representatives to last year's OECD World Forum on Statistics, Knowledge and Policy, which aims to foster the development of key indicators that measure the progress of societies. (I'm sure Enrico Giovannini, the OECD's chief statistician, will talk more about this in his session later today.) And attendance at this inaugural NatStats conference is testimony to the importance we all place on our nation's statistics.

Importantly, our political leaders are promoting a greater use of evidence. In an address to senior bureaucrats, Prime Minister Kevin Rudd expressed a need for "facts, not fads". He went on to say that "[g]overnment must receive the best advice, based on the best available information and evidence".[6] Meanwhile, Treasurer Wayne Swan has spoken of "ways in which public information can lead to real improvements in policy outcomes".[7]

The enthusiasm for statistics appears to be shared at all levels of government as the Council of Australian Governments (COAG) agreed to link Commonwealth payments to measured performance outcomes. In December last year, COAG outlined several key areas for reform (including indigenous affairs, health and education, which we will hear more about in the conference sessions that immediately follow this one). Each reform area has a so-called *OOMS* framework (objectives, outcomes and performance measures), which outlines the benchmarks that should be met and the statistics that should be used to measure performance in the sector.

To support the national reform agenda, the Government has set up bodies with the express purpose of gathering data. The May federal budget provided for a National Schools Data and Assessment Centre, which will compile comprehensive data on a school-by-school basis. The data will be used to inform parents, identify underachieving schools and recognise and reward the best teachers.

And the National Housing Supply Council (NHSC), also set up in May, is another data gathering body. The NHSC is promoting consistency across councils, making the data on land available for release comparable across localities. Only with a consistent count is it possible to tackle housing affordability and plan for future housing needs.

Moreover, technological developments are supporting a larger evidence base. Today, Amazon.com's two largest databases are said to hold 42,000 gigabytes of data; storage that would have cost over $30 billion twenty years ago (more than Amazon.com's current

---

[5] Stigler (1999).

[6] Rudd (2008).

[7] Swan (2008).

market capitalisation).[8] At the end of 2007, the world stored 281 billion gigabytes of digital data, but even this enormous amount is expected to grow tenfold over the next five years.[9]

And the proliferation of sensors means that more data can be reliably captured. In the past, retailers had to do a manual count to know what was on their shelves; today's technologies not only mean that they know what is flying out the door in real time but with smart cart technology, retailers can track their customers as they browse the aisles, helping them to tweak their store layout.

And advances in computer power means far more data can be crunched. Modelling the impact of a profile of rising carbon prices on the Australian economy over the next 100 years takes up to 10 hours on Treasury's high-end desktops. Two decades ago, such computations would have taken over a year – and would therefore never have been attempted.

We can improve the evidence base by making better use of the data we collect

So history has demonstrated the power of good statistics and we now have almost universal agreement on the importance of a strong evidence base. We also have the technology to support large scale data collection, storage and analysis. All that remains is to compile the best available evidence base.

I'll spend the rest of the talk discussing five ways that we might make better use of the data that are already collected.

We should prefer data collected as people go about their everyday lives.

First, we should prefer data collected as people go about their everyday lives.

We collect a large amount of data as people go about their daily business, but frequently rely on evidence collected in surveys. While surveys allow us to ask precise questions, this flexibility comes at the cost of potentially inaccurate answers. In responding to official surveys, businesses lack a strong incentive to report accurate figures, while the threat of litigation compels armies of accountants to file accurate business activity statements. Surveys also suffer from the difficulties of putting together a representative sample; they often rely on people accurately recalling past events and it is difficult to frame questions without sometimes subtly prejudicing the answer.

Initiatives such as standard business reporting, championed by Treasury, which is using e**X**tensible **B**usiness **R**eporting **L**anguage to establish common reporting definitions, allow us to collect more accurate and useful data. XBRL is a standard reporting format developed by the accounting industry which allows businesses to consistently map their internal financial data to standard reporting definitions. This means that reporting to ASIC, APRA, the Australian Tax Office , ABS and the Australian Stock Exchange will be based on a common set of definitions. For example, data submitted in XBRL format could

[8] Business Intelligence Lowdown (2007).

[9] IDC (2008).

cover 95 per cent of the measures collected by the ABS Quarterly Business Indicators Survey and must be completed accurately.

What's more, policy analysts rarely have the time to understand individual business filings because businesses use subtly different conventions. XBRL standardises conventions in a computer readable format, allowing policy analysts to examine companies' finances electronically and more rigorously; to analyse the likely impact of a change in tax treatments or R&D grants on a firm, industry or even the macroeconomy.

We need to make sure data have clear definitions.

Secondly, I'd like to discuss the need for well-defined data.

A useful evidence base depends heavily on clear and consistent data definitions. Inconsistent data are a big problem, particularly with figures collected across different jurisdictions. Every year since 1996, the Productivity Commission has released the Review of Government Services (RoGs) report that compares the performance of government services across states. And every year since 1996, readers of the RoGs report are overwhelmed by footnotes explaining the differences in data definitions across jurisdictions – we don't even have a consistent definition of an indigenous person.[10]

Thankfully, COAG is making progress on standardising the statistics that are used as performance measures. For the first time this year, schools in all Australian states took the same literacy and numeracy test. (This standardised test caused Queenland's year seven students to drop from the best writers in the country when the state administered its own test in 2005, to the worst performing among states this year.) And as mentioned earlier, the National Housing Supply Council is working towards agreed definitions of land supply by sending people from council to council to help with the count.

Data should be shared

Thirdly, I'd like to reflect on the importance of sharing data.

Having clearly defined administrative data is all very well, but it's next to useless if these data are not shared with those best able to build the evidence base. Our universities and research institutes are teeming with people wanting to draw lessons from agencies' statistics. In many cases it's these researchers who have the time and expertise to build the evidence base. But in many cases these same researchers don't have access to the data. Researchers are often forced to fumble around like the drunk that searches for his keys under a street light – not because his keys are likely to be there, but because it's the only spot where he can see.

A lack of data means that many researchers end up working on international datasets. Microsimulation specialists pour into Nordic countries because of their liberal approach towards sharing statistics. It is only recently, with the advent of the HILDA dataset, that Australia has had longitudinal data, which tracks people through time. This has led to a large increase in research using these Australian data.

[10] In Tasmania, somebody is indigenous if they feature on the state's pre-existing registry, while in other states people qualify if they identify as indigenous.

Sharing data not only helps build a base of academic evidence, but also helps more directly in policy decisions. During the current financial turmoil, macro policymakers must make decisions on the basis of only limited information on how the financial turmoil is affecting the real economy. The latest intensification of the turmoil followed the collapse of Lehman Brothers on September 15th. However, it's only in the past week - with the release of dismal US retail trade figures for October - that we're beginning to see the impact of Lehman's collapse on the US macroeconomy. In Australia, we won't have GDP figures that pick up the impact of Lehman's bankruptcy until March next year. All the while, data on things like tax collections, unemployment benefit recipients and business registrations are collected in almost real time and could give macro-policymakers a faster read on the real economy.

And data are invaluable when calibrating our policy models. The computable general equilibrium models used to model climate change policies depend heavily on elasticities – which must be estimated using available data. The more accurate the elasticities, the more valuable the model insights.

Sharing data also promotes accountability. In a recent speech, Treasurer Swan referred to the case of the New York State Department of Health, where collecting and reporting information on every heart bypass sent mortality rates for cardiac operations down by 40 per cent.[11] Also in New York, restaurant chains were recently compelled to report the calories of items on their menus. Studies suggest that this reporting prompts the average diner to choose meals with around 50 fewer calories, and induces restaurants to prepare lighter options.[12] In a similar vein, school-by-school reporting should help parents make informed decisions, while motivating lagging educators.

Unfortunately, many agencies guard their data with puzzling ferocity, ignoring the benefits they might confer. Some fail to release their stats because they're costly to compile in a usable format. However the cost of releasing data pales in comparison to the potential costs of misdirected policy.

Moreover, we're not maximising the value we get from our researchers. In 2007-08, the Australian Research Council doled out $600m worth of grants to support research that benefits the community. But, at the same time, the work of researchers is being constrained by limiting their access to statistics.

The other major reason for withholding data is privacy; made more worrying by misplaced public records in Britain and Japan. Privacy is a valid concern, with a Carnegie Mellon University study showing that 87 per cent of Americans can be identified from data on their gender, birth date and zip code.[13] However, these concerns are surmountable using techniques that obscure the fields that might reveal individual identities; techniques that have been deployed by the Australian Bureau of Statistics in their confidentialised unit record files.

Data becomes exponentially more valuable when they are combined.

---

[11] Swan (2008).

[12] The Economist (2008).

[13] Quoted from Baker (2008).

My fourth suggestion is to combine disparate records.

We can give a big boost to our evidence base by joining separate records. Combining records is the tax office's most powerful weapon against tax evasion and tax avoidance. If somebody's declared income only supports an ascetic life, but a road authority registers their new sports car, and a state revenue office collects stamp duty on their expensive property, these discrepancies are identified by the tax office prompting it to investigate. Linking data might also make it easier to identify disadvantaged children who often have parents on income support, poor school attendance records and a criminal record – but these data reside with different agencies.

Combining data could potentially create one big longitudinal data set, allowing policymakers to track people from enrolment in childcare, to school attendance and results, through to university and work, and the receipt of pensions or use of superannuation. By tracing people's lives through their data footprint, we could see where and why people slip through the gaps. Of course, there are privacy concerns with such a comprehensive data set, and these concerns are reasonable and need to be managed carefully.

A data repository would be useful.

And finally, collating data into a central statistical repository would be helpful. Such a repository could carefully categorise and document data, saving analysts and researchers valuable time. It could also facilitate the sharing of data and possibly the linking of disparate records. The repository might form part of a statistics wiki – a suggestion championed by Enrico Giovannini – with organisations posting their statistics to an open access website, just as people post encyclopaedic entries onto Wikipedia.

The repository should also help to overcome some agencies' reservations about releasing their administrative data. The repository could include systems that allow agencies to publish their data cheaply so that setting up the requisite infrastructure doesn't eat into their budgets. And the repository could include a facility that confidentialises data – meaning that agencies don't have to worry about the privacy of the people that use their facilities.

Of course, the funding of these initiatives must be weighed against other spending priorities – and as a treasury representative you would expect me to remind you that budget constraints must be honoured – but they should nevertheless remain a goal.

Concluding remarks

Let me conclude by saying that I think everyone here will agree that statistics have been an important part of the past eighty years. And that they'll be an even more important part of the next eighty years. However, as I've mentioned, there is still much we can do to make the most of this vital resource.

Thank you.

References

Baker Stephan (2008), “the Numerati”, *Houghton Mifflin Company*

Bernanke, Ben (2004), “The Great Moderation”, http://www.federalreserve.gov/BOARDDOCS/SPEECHES/2004/20040220/default.htm

Business Intelligence Lowdown (2007), “Top 10 Largest Databases in the World”, *http://www.businessintelligencelowdown.com/2007/02/top_10_largest.html*

Froyen, Richard T (2005), “Macroeconomics: Theories and Policies”, *Prentice Hall*.

IDC (2008), “The Diverse and Exploding Digital Universe”, *http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf*

Romer, Christina D. (1999), Changes in Business Cycles, Evidence and Explanation”, *NBER Working Paper Number 6948.*

Rudd, K (2008), “Address to Heads of Agencies and Members of Senior Executive Service”, *http://www.pm.gov.au/media/speech/2008/speech_0226.cfm*

Smith, AFM (1996), “Mad cows and ecstasy: Chance and choice in an evidence-based society”, *Journal of the Royal Statistical Society.*

Stigler, Stephen M. (2002), “Statistics on the Table”, *Harvard University Press*.

Swan, W (2008), “Modern Federalism Not Creeping Centralism”, *http://www.treasurer.gov.au/DisplayDocs.aspx?doc=speeches/2008/025.htm&pageID=005&min=wms&Year=&DocType=1*.

The Economist (2008), “**Menu items”,** *http://www.economist.com/business/displaystory.cfm?story_id=12010393*.